

PERFIS DE RENTABILIDADE NO ECOSSISTEMA CINEMATOGRAFICO: Construção e Normalização de um Dataset Com Dados Abertos da Web¹

Marcela C. Caram Peito²

Danilo Boechat Seufitelli³

Daniel Hasan Dalip⁴

Submetido em: 19/06/2026 | Aprovado em: 19/06/2026

RESUMO

A indústria cinematográfica global opera sob uma severa assimetria financeira, tornando a alocação de recursos e a mitigação de riscos desafios complexos para estúdios e investidores. O presente trabalho tem como objetivo aplicar técnicas de Engenharia de Dados e Aprendizado de Máquina Não Supervisionado para mapear a estrutura latente deste mercado, visando a geração de inteligência de negócios. A metodologia baseou-se na construção e higienização de uma base de dados estruturada, aplicando tratamento de outliers e transformações logarítmicas em variáveis determinantes, como lucro estimado, avaliação crítica e engajamento do público. Na etapa de clusterização, conduziu-se uma análise comparativa entre os modelos DBSCAN e K-Means. Os resultados do DBSCAN sugerem uma estrutura densa de difícil segmentação, demonstrando eficácia na detecção de anomalias (ruído), mas apresentando limitações severas para a segmentação estratégica. Em contrapartida, o algoritmo K-Means forçou o particionamento do mercado em quatro quadrantes (Titãs Comerciais, Os Injustiçados, Fracassos Absolutos e Nicho de Baixo Orçamento), evidenciando o distanciamento sistêmico entre a aclamação crítica e a rentabilidade. Conclui-se que categorizar produções em perfis estritos é um desafio de altíssima complexidade. A pesquisa demonstra as restrições da clusterização tradicional, provando que o mercado resiste a divisões engessadas. Assim, a principal contribuição do estudo reside em mapear as abordagens que falham nesta categorização, apontando a necessidade imperativa de novas arquiteturas analíticas para compreender a dinâmica do setor audiovisual. Adicionalmente, a própria base de dados construída e estruturada também se configura como uma importante contribuição deste trabalho.

Palavras-chave: Aprendizado de Máquina. Engenharia de Dados. Clusterização. K-Means. DBSCAN.

¹ Artigo científico elaborado para o trabalho de conclusão do curso de Engenharia de Computação do campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.)

² Graduanda em Engenharia de Computação no campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.). marcelaccaram@gmail.com

³ Orientador do trabalho. Docente do Departamento de Computação do campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.). danilobs@cefetmg.br

⁴ Coorientador do trabalho. Docente do Departamento de Computação do campus Nova Gameleira do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG.). hasan@cefetmg.br

ABSTRACT

The global film industry operates under severe financial asymmetry, making resource allocation and risk mitigation complex challenges for studios and investors. This study aims to apply Data Engineering and Unsupervised Machine Learning techniques to map the latent structure of this market to generate business intelligence. The methodology relied on building and cleaning a structured database, applying outlier treatment and logarithmic transformations to key variables such as estimated profit, critical acclaim, and audience engagement. During the clustering stage, a comparative analysis between the DBSCAN and K-Means models was conducted. The DBSCAN results suggest a dense structure that is difficult to segment, demonstrating effectiveness in anomaly (noise) detection but presenting severe limitations for strategic segmentation. In contrast, the K-Means algorithm forced the market's partitioning into four quadrants, highlighting the systemic disconnect between critical acclaim and financial profitability. It is concluded that categorizing film productions into strict profiles is a highly complex challenge. The research demonstrates the constraints of traditional clustering, proving that the market resists rigid divisions. Therefore, the study's main contribution lies in mapping the approaches that fail in this categorization, emphasizing the imperative need for new analytical architectures to understand the dynamics of the audiovisual sector. Additionally, the constructed and structured database itself also serves as an important contribution of this work.

Keywords: Machine Learning. Data Engineering. Clustering. K-Means. DBSCAN.

1 INTRODUÇÃO

A indústria cinematográfica se destaca como um dos setores mais influentes e lucrativos do entretenimento, movimentando um volume financeiro que alcança bilhões de dólares anualmente. No entanto, por trás destas produções, encontra-se um desafio significativo: a previsão do sucesso comercial de um filme antes de sua estreia. A receita de bilheteria, principal métrica de desempenho financeiro, é o resultado de uma interação complexa de fatores, que incluem o prestígio do elenco e da direção, o orçamento, o gênero, a data de lançamento e até mesmo o volátil contexto sociocultural do público. Nesse cenário de incertezas, a capacidade de estimar o potencial de arrecadação com antecedência representa um diferencial estratégico, oferecendo a produtores e investidores uma base mais sólida para a tomada de decisões.

Com o avanço da ciência de dados, o uso de técnicas de aprendizado de máquina para modelar e prever o sucesso em domínios de entretenimento tem se tornado uma área de intensa pesquisa. No campo específico da previsão de bilheteria, abordagens recentes demonstram alta sofisticação. [Ni et al. \(2022\)](#), por exemplo, alcançaram um erro de previsão de apenas 14,49% ao combinar dados tradicionais com fontes dinâmicas, como tendências de busca e o impacto da pandemia, usando um modelo de *stacking*. De forma complementar, outros estudos focam em arquiteturas de *deep learning* com redes LSTM para capturar a evolução diária da receita, mostrando como a integração de características estáticas do filme melhora significativamente a acurácia ([Ru et al., 2018](#)). De fato, abordagens baseadas em dados históricos também já são utilizadas para prever a popularidade de forma geral ([Fernandes; Oliveira, 2021](#)). Em um domínio análogo ao do cinema, o trabalho de [Oliveira e Santos \(2022\)](#) demonstra a eficácia de modelos preditivos para estimar o sucesso de músicas com base em suas características. Adicionalmente, a análise computacional também revela

a complexidade do próprio objeto de estudo, pois a percepção e a descrição de um mesmo filme podem variar significativamente entre diferentes culturas e idiomas, como evidenciado por [Pereira et al. \(2024\)](#) em sua análise de artigos da Wikipédia.

Alinhado a esse panorama, este trabalho concentrou seu esforço metodológico na expansão e no refinamento de uma base de dados estruturada. Partindo de um conjunto original de 33.600 registros, no qual apenas 9.406 filmes (28,0%) apresentavam o par financeiro completo de orçamento e bilheteria, a pesquisa aplicou técnicas de raspagem de dados (*web scraping*) e consumo de APIs para recuperar informações ausentes. O processo de higienização resultou em uma base consolidada de 11.828 filmes, gerando um aumento absoluto de 25,7% (2.422 filmes) na quantidade de registros com dados financeiros válidos. Uma etapa central do projeto é a higienização da base, que envolveu a limpeza de valores inconsistentes, a padronização de formatos de gênero e equipe técnica, e o ajuste financeiro das moedas para corrigir distorções temporais e inflacionárias. Após a estruturação da base, o trabalho aplicou o algoritmo não supervisionado K-Means para agrupar as obras e observar padrões de rentabilidade. O modelo segmentou a amostra em quatro perfis distintos, o que permitiu analisar de forma comparativa o comportamento financeiro e de recepção em diferentes escalas da indústria, englobando desde produções de menor orçamento até grandes franquias.

O objetivo deste trabalho é compreender a estrutura latente do mercado cinematográfico e observar as variáveis que estão associadas ao sucesso financeiro e a recepção das obras. Especificamente, os objetivos de pesquisa (*Research Goals* - RG) selecionados para conduzir esta análise são:

- RG1: Construir e higienizar um dataset enriquecido, aumentando o número de dados orçamentários e de bilheteria presentes.
- RG2: Analisar o perfil da base consolidada, mapeando as relações entre investimento, rentabilidade e engajamento do público.
- RG3: Identificar e segmentar grupos de filmes a partir da aplicação de algoritmos de clustering, testando a viabilidade de categorizar as dinâmicas de mercado em nichos estruturados.

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados e o embasamento teórico da pesquisa. A Seção 3 detalha a metodologia aplicada, descrevendo o fluxo de extração, limpeza, engenharia de atributos e a configuração dos algoritmos de aprendizado de máquina não supervisionado. A Seção 4 expõe os resultados obtidos, avaliando o impacto numérico do enriquecimento da base de dados e interpretando os perfis de mercado identificados pela clusterização. Por fim, a Seção 5 consolida as conclusões do estudo e sugere perspectivas analíticas para trabalhos futuros.

2 TRABALHOS RELACIONADOS

A predição de sucesso no mercado cinematográfico é um desafio que tem evoluído de análises estatísticas simples para *pipelines* complexos de aprendizagem de máquina. Um dos pontos fundamentais nessa discussão reside na definição da métrica de sucesso. Enquanto abordagens tradicionais focam na receita bruta, [Lash e Zhao \(2016\)](#) argumentam que o faturamento isolado é um indicador incompleto, uma vez que receitas elevadas

não garantem lucratividade. Os autores defendem o uso do Retorno sobre Investimento (ROI) como métrica primária, demonstrando que a correlação entre a bilheteria bruta e a rentabilidade real é baixa. Esta perspectiva fundamenta a decisão deste trabalho em priorizar métricas de rentabilidade relativa frente ao capital investido.

Para mitigar as limitações de bases de dados isoladas, a literatura aponta para a eficácia de modelos híbridos. [Zhang e Skiena \(2009\)](#) provaram que a inclusão de variáveis quantitativas provenientes de análises de notícias eleva a precisão das predições de forma estatisticamente significativa em comparação ao uso exclusivo de metadados estruturados do IMDb. Esse conceito de “buzz” é expandido por [Krauss *et al.* \(2008\)](#), que utilizam a intensidade e a positividade de discussões em fóruns online para antecipar não apenas o desempenho comercial, mas também o reconhecimento crítico em premiações como o Oscar. Tais estudos validam a integração de fontes heterogêneas e o uso de métricas de engajamento público como dimensões essenciais para a caracterização dos filmes.

Do ponto de vista da Ciência de Dados, a literatura ressalta que o sucesso dos modelos depende do rigor no pré-processamento. [Dhir e Raj \(2018\)](#) destacam que o escalonamento de atributos (*feature scaling*) é mandatório devido à disparidade de ordens de magnitude entre variáveis financeiras e notas de avaliação. A necessidade de tratar dados ruidosos e inconsistentes é corroborada por [Ankit *et al.* \(2020\)](#), que demonstram como a conversão de metadados qualitativos em numéricos e o preenchimento criterioso de lacunas elevam a estabilidade de algoritmos baseados em similaridade espacial. Esse arcabouço teórico dá suporte à metodologia aplicada nesta pesquisa, justificando o uso do algoritmo K-Means sobre um espaço vetorial normalizado para a identificação de perfis de mercado distintos.

3 METODOLOGIA

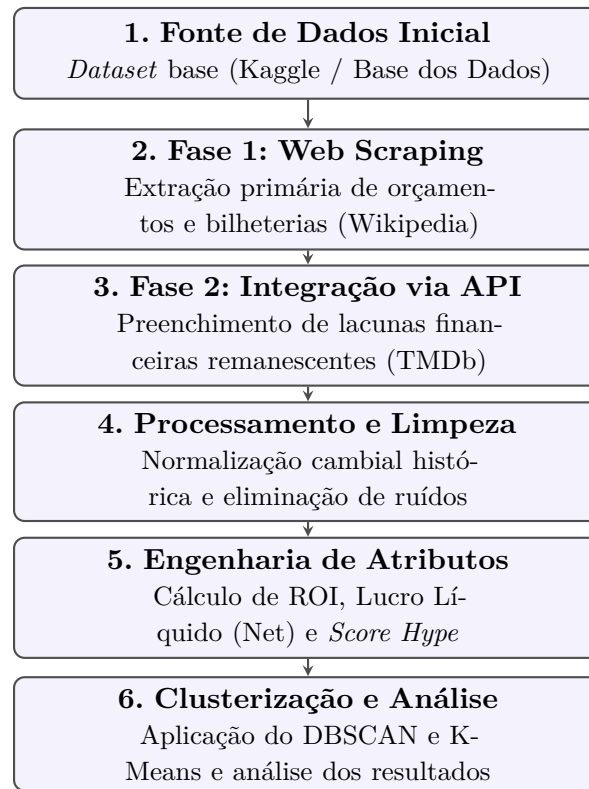
Esta metodologia foi elaborada para criar uma base de dados financeiramente consistente para a análise descritiva do mercado cinematográfico, buscando diminuir a escassez de informações de orçamento e bilheteria em catálogos públicos. O fluxo de trabalho, ilustrado detalhadamente na Figura ??, envolveu a extração, o enriquecimento via automação (*web scraping* e APIs), o tratamento rigoroso de ruídos e, por fim, a aplicação de algoritmos de aprendizado de máquina não supervisionado para a segmentação mercadológica.

3.1 Fonte de Dados e Caracterização Inicial

O ponto de partida deste trabalho foi a tabela `top_movies_per_year`, parte do conjunto de dados `imdb-movies`, disponível publicamente na plataforma Base dos Dados ([Base dos Dados, 2023](#)). Esse conjunto foi utilizado como base inicial, fornecendo informações importantes como identificador único (`id`), título original (`title`), ano de lançamento (`year`) e gênero (`genre`).

Apesar de sua riqueza qualitativa, a base apresentava limitações financeiras críticas. Uma parcela majoritária dos registros possuía apenas o orçamento (`budget`), apenas a bilheteria mundial (`gross_world_wide`) ou nenhum dos dois dados, inviabilizando cálculos de rentabilidade. Adicionalmente, identificou-se que parte dos valores financeiros brutos não correspondia ao dólar americano (USD), apresentando cifras em moedas locais (como a rupia indiana ou a antiga lira italiana). Essa fragmentação e heterogeneidade justificaram a necessidade de um *pipeline* completo de enriquecimento e normalização de dados.

Figura 1 – Fluxograma detalhado das etapas metodológicas da pesquisa



Fonte: Elaborado pela autora (2026).

3.2 Processo de Enriquecimento dos Dados

Para sanar a carência de dados financeiros, foi conduzido um processo de enriquecimento de dados estruturado em duas fases sequenciais e hierárquicas. O procedimento teve como pilar uma coleta de dados em larga escala (*web scraping*) da Wikipedia, complementada ao final por consultas à API do The Movie Database (TMDb). Um passo fundamental, prévio a estas etapas, foi a normalização dos títulos dos filmes, que consistiu na limpeza e padronização dos nomes para garantir a correspondência precisa entre o *dataset* original e as informações a serem coletadas.

3.2.1 Fase 1: Extração e Integração de Dados da Wikipedia

Esta fase foi dedicada a extrair sistematicamente dados financeiros da Wikipedia e foi dividida em duas subetapas principais:

- 1. Mapeamento e Coleta de Links:** O ponto de partida não foram as páginas individuais dos filmes, mas sim a página-índice da Wikipedia⁵. A partir dela, o *script* de coleta navegou para as listas de filmes organizadas alfabeticamente (ex: `/wiki/List_of_films:_A`). Nessas páginas, os links dos filmes eram apresentados no formato `[[Altar (2014 film)|Altar]]`. O *script* foi programado para extrair a primeira parte do link (`Altar_(2014_film)`) para construir a URL canônica de cada filme. Este procedimento resultou em uma lista de URLs de filmes a serem analisados.

⁵ <https://en.wikipedia.org/wiki/Lists_of_films>

2. **Extração de Dados Financeiros:** Com a lista de URLs em mãos, um segundo *script* foi executado para visitar cada página de filme individualmente. O objetivo foi analisar a tabela de resumo (*infobox*) e extrair os valores de orçamento e bilheteria. Este processo foi executado ao longo de vários dias. Os resultados (nome do filme, link, orçamento e bilheteria) foram consolidados em um arquivo JSON para servirem como uma base de dados intermediária.

Após a conclusão da extração, iniciou-se a integração dos dados. Para evitar a confusão entre filmes com o mesmo nome (como remakes ou homônimos), o *script* realizou a correspondência entre o *dataset* principal e os dados da Wikipedia utilizando tanto o título normalizado quanto o ano de lançamento da obra. As colunas oficiais `budget` e `gross_world_wide` foram então atualizadas para os filmes que antes não possuíam essas informações. Adicionalmente, o processo incluiu duas lógicas de enriquecimento importantes:

- **Registro de Moeda:** A informação sobre a moeda original dos dados financeiros (ex: USD, EUR) foi extraída e armazenada sempre que disponível.
- **Preenchimento de Dados Parciais:** O *script* também buscou preencher dados para filmes que já possuíam uma das duas informações financeiras (orçamento ou bilheteria), mas não ambas, completando o par sempre que o dado faltante era encontrado.

3.2.2 Fase 2: Enriquecimento Complementar via API do TMDb

A API do The Movie Database (TMDb) atuou como recurso complementar de resgate para lacunas remanescentes. O algoritmo de consulta foi parametrizado para interpretar retornos com valor zero como “dado indisponível”, impedindo a corrupção do *dataset* com dados falsos.

3.2.3 Consolidação Hierárquica da Base

A unificação dos dados operou sob uma lógica de precedência para garantir máxima fidelidade. As colunas financeiras foram preenchidas adotando o primeiro valor não nulo encontrado na seguinte hierarquia: 1) Base dos Dados, 2) Wikipedia e 3) TMDb. Após a integração transversal, as colunas de armazenamento temporário foram expurgadas, gerando a `baseDeDadosResultante.csv`.

3.3 Processamento, Normalização e Limpeza

Com os valores nominais preenchidos, o *pipeline* avançou para a etapa de qualificação matemática e correção de ruídos.

3.3.1 Normalização Cambial e Fator Temporal

A padronização monetária para o Dólar Americano (USD) era mandatória para viabilizar a execução correta dos cálculos espaciais e das métricas de rentabilidade. Para rastrear as moedas originais, criou-se uma coluna temporária no *dataset* dedicada exclusivamente ao armazenamento do indicativo cambial extraído junto aos dados brutos.

O escopo de verificação adotou um filtro geográfico estrito, analisando todos os registros cuja lista de países de origem não incluísse os Estados Unidos. Adicionalmente, realizou-se uma varredura para identificar “valores suspeitos” (cifras de orçamento ou bilheteria com magnitudes numéricas atípicas que pudessem mascarar a presença de moedas desvalorizadas frente ao dólar). Na etapa de conversão, evitou-se a adoção de taxas contemporâneas fixas. Para garantir o rigor econômico das correções, aplicou-se a cotação cambial histórica correspondente ao ano de lançamento de cada obra, utilizando como fonte primária a série temporal de taxas de câmbio oficiais (*Official exchange rate*) disponibilizada pelo Banco Mundial ([Banco Mundial, 2026](#)).

3.3.2 Eliminação de Ruídos e Filtros de Integridade

A base foi submetida a filtros lógicos restritivos para garantir a viabilidade da clusterização subsequente:

- **Integridade do Par Financeiro:** Exclusão de qualquer registro que não possuisse orçamento e bilheteria preenchidos e válidos simultaneamente.
- **Eliminação de Orçamentos Simbólicos e Limiares:** Obras com orçamentos documentados em valores ínfimos (como \$1 ou \$5) geram um Retorno sobre Investimento (ROI) extremamente elevado. Estabeleceu-se um piso de segurança de \$500,00, assegurando que os dados refletissem operações cinematográficas comerciais.

3.4 Construção de Variáveis Analíticas

Com o objetivo de preparar os dados para a clusterização, os dados brutos foram transformados em indicadores estruturados de mercado.

3.4.1 Tratamento Vetorial de Gêneros e Multi-origens

Textos longos contendo vários gêneros misturados foram separados em listas estruturadas. Para padronizar a base e reduzir o excesso de categorias, o mapeamento foi limitado apenas aos dois primeiros gêneros listados para cada filme (os predominantes). Em seguida, essa dupla inicial foi ordenada alfabeticamente (ex: Action-Adventure), garantindo que o algoritmo não tratasse combinações idênticas com ordens invertidas como se fossem categorias diferentes. Filmes que possuíam apenas um gênero receberam a marcação nula para o segundo espaço (ex: Drama-None).

Essa mesma lógica de separação em listas foi aplicada às colunas de países e idiomas, permitindo realizar a contagem correta da origem das obras e identificar co-produções internacionais de forma precisa.

3.4.2 Geração de Indicadores Contínuos

Para viabilizar comparações justas e permitir que os algoritmos de aprendizado de máquina avaliassem o desempenho dos filmes de maneira proporcional, foram projetadas quatro novas dimensões contínuas:

- **ROI Real (*Return on Investment*):** Calculado pela razão simples entre a bilheteria mundial bruta e o orçamento de produção (Gross/Budget). O ROI atua como uma métrica de normalização da rentabilidade, revelando a real capacidade

da obra de multiplicar o capital investido. Diferentemente da análise exclusiva da arrecadação nominal, que naturalmente favorece superproduções, o ROI permite comparar de forma equivalente a eficiência comercial de filmes de micro-orçamento com mega-*blockbusters*.

- **Net (Lucro Líquido Estimado):** Estima o montante financeiro que efetivamente retorna aos estúdios após o repasse para as salas de exibição. Na dinâmica da indústria cinematográfica, a receita bruta de bilheteria não é repassada integralmente aos produtores; dados de mercado apontam que as redes de cinema retêm, em média, cerca de 52% do valor arrecadado com a venda de ingressos (G1, 2020). Para fins de simplificação e modelagem neste trabalho, adotou-se uma penalização conservadora e arredondada de 50% sobre o faturamento bruto antes de subtrair o custo inicial de produção, sendo o indicador definido pela seguinte equação: $\text{Net} = \text{Gross} \times 0.5 - \text{Budget}$.
- **Índice de Sucesso:** Um indicador híbrido concebido para cruzar a performance financeira com a aceitação qualitativa da obra. Calculado pelo produto do Retorno sobre Investimento pelo grau de aprovação do público (normalizado para uma escala decimal), a equação $(\text{Gross}/\text{Budget}) \times (\text{Nota IMDb}/10)$ atua como um avaliador bidimensional. Dessa forma, filmes altamente lucrativos, porém com rejeição crítica severa, sofrem uma penalização matemática, enquanto obras que aliam rentabilidade financeira a uma alta qualidade percebida pelo espectador têm seu índice maximizado.
- **Score Composto de Repercussão (Hype):** Métrica desenvolvida para quantificar o engajamento popular e a relevância cultural de um filme. O volume de votos na internet apresenta uma distribuição estatística de “cauda longa”, onde poucas obras concentram milhões de avaliações, enquanto a maioria possui apenas algumas dezenas. Para impedir que os algoritmos de *clustering* fossem enviesados por esses volumes absolutos extremos, aplicou-se um nivelamento matemático utilizando o logaritmo de base 10. A constante de adição (+1) garante a robustez algébrica da equação, prevenindo indefinições matemáticas em obras sem engajamento prévio ($\log_{10}(0)$). A métrica é finalmente ponderada pela nota crítica do filme:

$$\text{Score Hype} = \log_{10}(\text{Votos} + 1) \times \text{Nota IMDb}$$

3.4.3 Concepção de Atributos Categóricos por Regras de Negócio

Com base nos indicadores contínuos gerados, foram estruturadas três colunas categóricas ordinais para discretizar o perfil de mercado das obras cinematográficas:

- **Classificação Comercial:** Mapeia o desempenho econômico real cruzando as variáveis *net* e ROI. Os registros são indexados sob quatro limiares econômicos estruturados:
 - *Fracasso Crítico (Flop):* Obras com $\text{Net} < 0$, indicando incapacidade nominal de recuperar o custo de produção após a divisão de receita com as salas exibidoras.
 - *Pagou as Contas (Breakeven):* Obras com $\text{Net} \geq 0$ que, contudo, apresentam um $\text{ROI} < 2.5$, indicando que o filme operou no azul, mas não superou o limiar de mercado estimado para a cobertura de custos de marketing e distribuição.
 - *Lucrativo:* Obras situadas no intervalo de $2.5 \leq \text{ROI} < 5.0$.

- *Alta Rentabilidade (Hit)*: Produções de altíssima eficiência que atingiram um $\text{ROI} \geq 5.0$.
- **Perfil de Hype**: Estratificado em faixas percentílicas (50%/40%/10%), dividindo o engajamento em: Baixo Impacto, Repercussão Moderada e Alta Repercussão.

3.5 Clusterização e Análise de Dados

Para identificar padrões de comportamento e segmentar o mercado cinematográfico em perfis, utilizou-se a abordagem de aprendizado de máquina não supervisionado (*Unsupervised Learning*). A fim de garantir a confiabilidade dos resultados, o processo de agrupamento (*clustering*) foi estruturado para testar diferentes modelos matemáticos.

3.5.1 Tratamento de Outliers e Escalonamento (StandardScaler)

Algoritmos de agrupamento são altamente sensíveis à diferença de escala entre as variáveis e à presença de valores atípicos nos dados. Por isso, a etapa de pré-processamento exigiu duas ações corretivas fundamentais.

Primeiramente, realizou-se a remoção de outliers financeiros extremos (filmes com orçamentos ou bilheterias excessivamente fora do padrão de distribuição da indústria). Essa limpeza foi necessária para evitar que pontos isolados distorcessem os cálculos de distância e puxassem a formação dos grupos de maneira artificial.

Em seguida, aplicou-se a padronização dos dados utilizando o *StandardScaler*. Essa técnica estatística ajusta todas as variáveis contínuas para que passem a ter uma média igual a zero e desvio padrão igual a um. Na prática, esse escalonamento foi necessário para evitar que os atributos com valores na casa dos milhões (como dólares) dominem a matemática do modelo e anulem a influência de variáveis com escalas muito menores (como a nota do IMDb, que varia apenas de 0 a 10).

3.5.2 Avaliação Exploratória via DBSCAN

A primeira abordagem testada foi o algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*). Sua principal vantagem é não exigir a definição prévia do número de grupos (K). O objetivo inicial foi verificar se o mercado audiovisual — que possui concentrações irregulares de dados, com muitos filmes de baixo orçamento de um lado e poucos *blockbusters* isolados do outro — seria melhor segmentado pela densidade matemática. Os parâmetros de raio de vizinhança (ϵ) e o número mínimo de amostras (*min_samples*) foram ajustados iterativamente com o intuito de formar grupos lógicos e evitar que muitos filmes fossem descartados como ruído.

3.5.3 Agrupamento Definitivo via K-Means e Otimização

Como alternativa ao modelo de densidade e buscando formar nichos de mercado mais delimitados, aplicou-se o algoritmo K-Means. Esse modelo agrupa os dados minimizando as distâncias euclidianas internas dentro de cada grupo (inércia).

Para determinar o número ideal de grupos (K), utilizou-se o Método do Cotovelo (*Elbow Method*), analisando o comportamento da variância para valores de K entre 1 e 10. O ponto de inflexão da curva apontou $K = 4$ como a divisão ideal. Essa escolha foi validada em seguida pelo Coeficiente de Silhueta (*Silhouette Score*), uma métrica que

quantifica a coesão interna dos filmes dentro de um mesmo *cluster* e o grau de separação em relação aos grupos vizinhos.

3.6 Validação e Interpretação Analítica

Após a execução do algoritmo e a divisão dos dados, a última etapa metodológica consistiu em traduzir os agrupamentos matemáticos em perfis de mercado reais. Para que a análise fizesse sentido prático, os dados numéricos de orçamento e retorno tiveram sua padronização revertida para as escalas financeiras originais. Isso permitiu extrair as médias reais de investimento e lucro de cada *cluster*, além de verificar se as categorias criadas na fase de engenharia de atributos (como o Perfil de Hype) conversavam organicamente com as divisões feitas pela máquina.

A fim de garantir a transparência, fomentar a reprodutibilidade metodológica e auxiliar futuras pesquisas na área de análise de dados cinematográficos, a base de dados consolidada resultante deste trabalho — intitulada *Cinematic Profitability Dataset* — foi disponibilizada publicamente. O conjunto de dados, juntamente com o seu respectivo dicionário de variáveis, encontra-se hospedado em repositório aberto no GitHub ([Caram et al., 2026](#)).

4 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados do trabalho divididos em duas etapas principais. Inicialmente, realiza-se uma avaliação da base de dados resultante em comparação com o conjunto original, com o intuito de validar o processo de limpeza e entender o perfil das obras cinematográficas que possuem informações financeiras disponíveis. Em um segundo momento, são discutidos os resultados da tentativa de clusterização desses dados, detalhando a avaliação do modelo escolhido e a interpretação das características que definem cada grupo de filmes formado.

4.1 Engenharia de Dados: Impacto do Enriquecimento e Limpeza da Base

A base de dados original, obtida do repositório Kaggle, apresentava um volume inicial de 33.600 registros. No entanto, uma análise de completude demonstrou uma alta defasagem nas variáveis financeiras, essenciais para o cálculo de retorno financeiro (ROI). Na base original, apenas 9.406 filmes (28,0% do total) apresentavam o par financeiro completo (orçamento e arrecadação mundial). O processo de *web scraping*, aliado às regras de negócio de eliminação de ruídos, buscou alterar este panorama.

A base enriquecida consolidou-se com 11.828 filmes. Embora represente uma redução de 64,8% em relação ao volume bruto da base do Kaggle (devido à exclusão de filmes que permaneceram sem dados financeiros após as buscas e a remoção de orçamentos simbólicos), a integração entre técnicas de *web scraping* e o consumo de APIs gerou um aumento absoluto de 25,7% (2.422 filmes novos) na quantia de registros com dados válidos comparado aos originais. Assim, enquanto na base bruta apenas 28,0% do total de filmes possuía o par financeiro (orçamento e bilheteria), a nova amostra consolidada de 11.828 obras atingiu 100% de completude nesses atributos, tornando-a integralmente apta para a aplicação de algoritmos de clusterização.

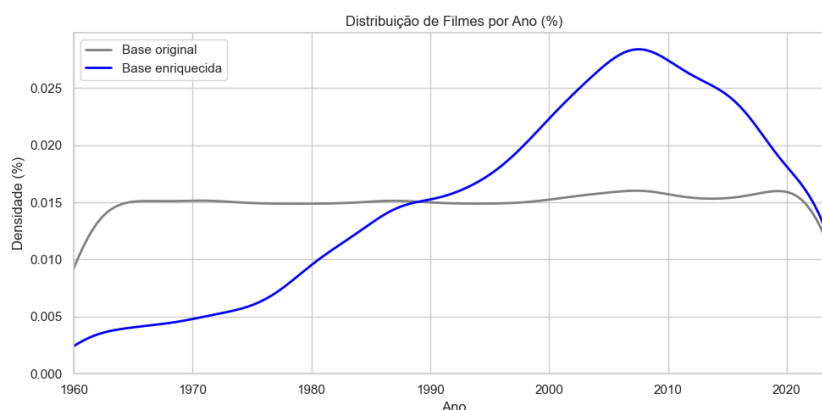
O impacto qualitativo dessa recuperação de dados fica evidente ao se observar a alteração na média orçamentária das obras. A base original válida apresentava um orçamento

médio de \$102.488.657, um valor irreal que evidenciava um forte viés de preenchimento voltado apenas para os mega-*blockbusters* de Hollywood. Após o enriquecimento, a média orçamentária sofreu uma correção drástica, caindo para \$23.846.169. Este novo valor sugere que as técnicas utilizadas foram bem-sucedidas em resgatar as informações de filmes de médio porte e produções independentes, democratizando a amostra e aproximando-a da realidade econômica da indústria cinematográfica global.

4.1.1 Perfil e Distribuição Temporal da Amostra

Para facilitar a análise das diferenças entre os conjuntos de dados, foram gerados gráficos comparativos normalizados em proporções percentuais. Isto é, o eixo Y dos gráficos de distribuição reflete a densidade em porcentagem (%), o que permite uma comparação justa das curvas independentemente da discrepância no número absoluto de filmes entre a base original e a enriquecida.

Figura 2 – Distribuição de Filmes por Ano



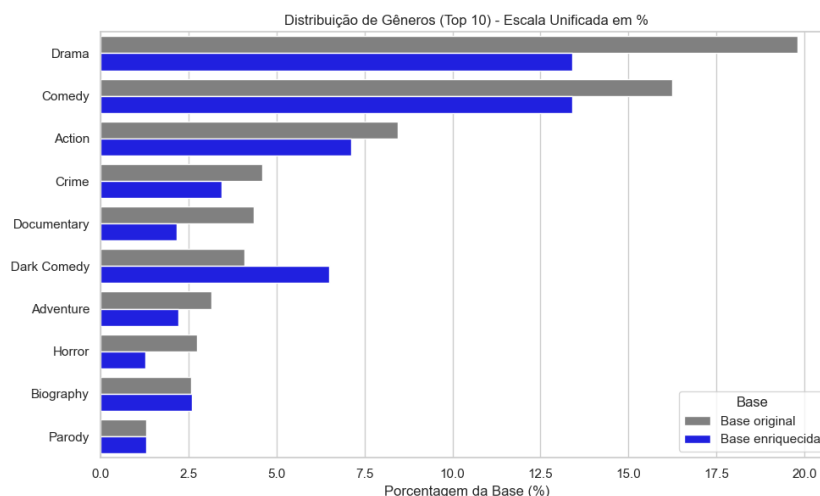
Fonte: Elaborado pela autora (2026).

A análise da distribuição temporal (Figura 2) evidencia as diferenças estruturais na composição das amostras. A curva da base original apresenta uma constância ao longo das décadas (próxima a 0,015%), refletindo um fato sobre a estruturação desse conjunto de dados, que possui uma distribuição de obras deliberadamente uniforme por ano.

Em contrapartida, a base enriquecida apresenta um comportamento alinhado à realidade do mercado cinematográfico. Observa-se um pico expressivo na densidade de registros a partir da década de 2000. Este fenômeno ocorre porque a base enriquecida restringe-se estritamente aos filmes que possuem transparência financeira, dados que historicamente são muito mais difíceis de obter para produções anteriores à era digital.

Destaca-se também uma queda vertiginosa na densidade de ambas as curvas logo após os anos 2010. Uma possível explicação dessa queda é a consolidação das plataformas de *streaming* no mercado audiovisual. Filmes lançados diretamente nestas plataformas frequentemente possuem seus orçamentos declarados, contudo, não geram bilheteria tradicional de cinema (**gross_worldwide**). Diante da ausência de arrecadação formal para o cálculo de rentabilidade, essas produções foram metodologicamente descartadas do funil de enriquecimento, resultando na queda representada no gráfico recente.

Figura 3 – Distribuição Comparativa de Gêneros (Top 10)



Fonte: Elaborado pela autora (2026).

4.1.2 Distribuição Comparativa de Gêneros

A análise comparativa dos gêneros cinematográficos (Figura 3) também revela mudanças importantes após o processo de enriquecimento.

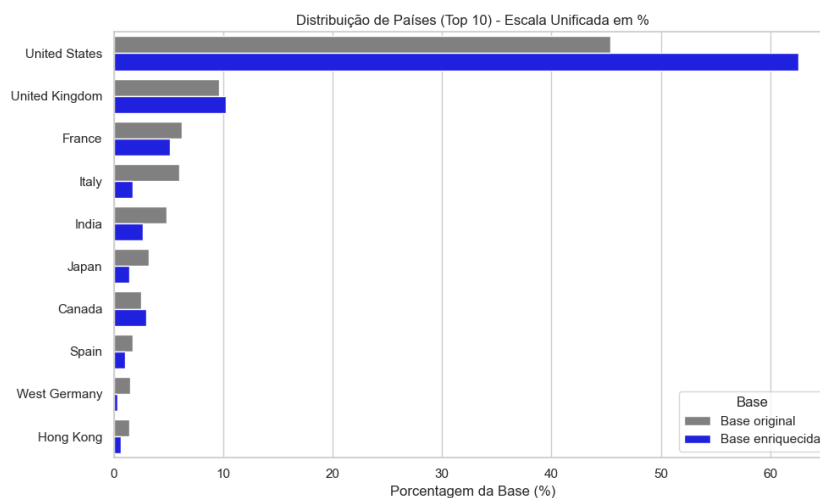
Embora Drama e Comédia permaneçam como os pilares dominantes em ambos os conjuntos de dados (apesar de uma queda proporcional na base enriquecida), a grande alteração estrutural ocorre em categorias de nicho. O destaque é o crescimento do gênero de Comédia de Humor Negro (*Dark Comedy*), que foi o único a apresentar um aumento de representatividade, saltando de aproximadamente 4% na base original para quase 6,5% na enriquecida. Esse salto sugere que obras com essa classificação, que frequentemente operam com orçamentos mais contidos e fora do circuito dos mega-*blockbusters*, careciam de cobertura financeira no *dataset* do Kaggle, sendo resgatadas de forma eficiente pelo processo de enriquecimento.

4.1.3 Análise Geográfica e Disparidade de Transparência

A análise da origem das obras cinematográficas (Figura 4) evidencia um viés estrutural na obtenção de dados públicos sobre a indústria. Observa-se que, ao restringir a base enriquecida apenas aos filmes com o par financeiro completo (orçamento e bilheteria), houve um aumento na proporção de obras provenientes dos Estados Unidos em relação ao conjunto original.

Simultaneamente, países com indústrias cinematográficas robustas, como Índia, Japão e Itália, sofreram uma defasagem percentual. Este fenômeno pode ser interpretado não como uma redução na produção destes países, mas sim como uma carência de transparência financeira globalizada. Os algoritmos de *scraping* e as APIs demonstram alta eficácia na extração de dados orçamentários padronizados de produções faladas em inglês e vinculadas a Hollywood, porém encontram dificuldades na captura de relatórios financeiros de obras asiáticas ou europeias, resultando na sua exclusão do *dataset* final por insuficiência de dados.

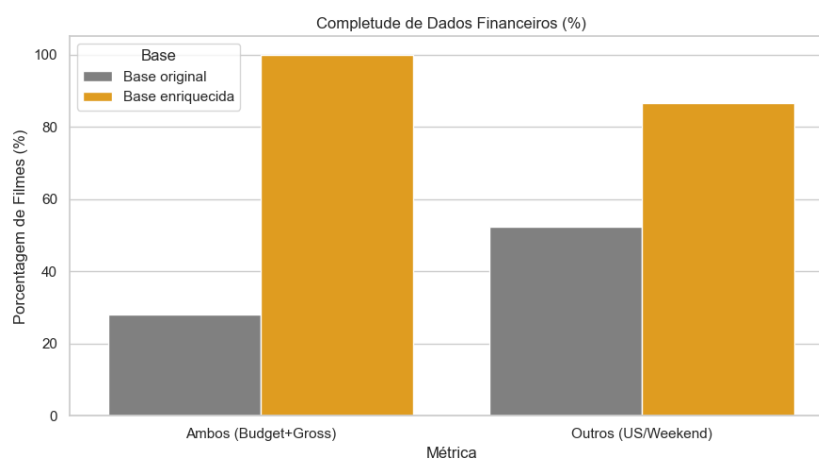
Figura 4 – Distribuição de Países



Fonte: Elaborado pela autora (2026).

4.1.4 Completude do Funil Monetário

Figura 5 – Completude de Dados Financeiros



Fonte: Elaborado pela autora (2026).

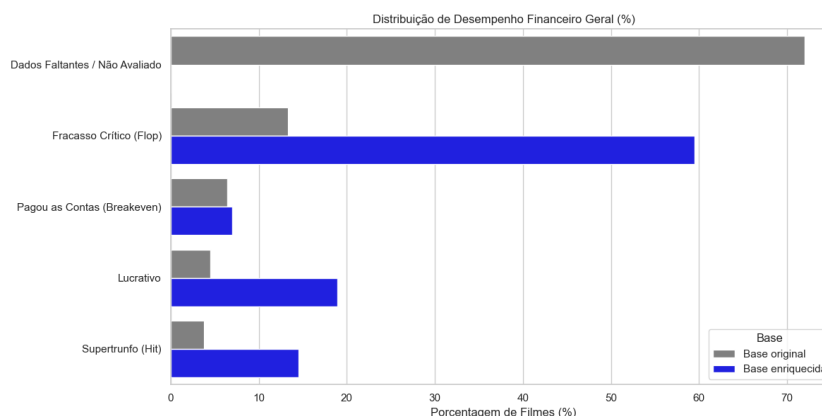
A Figura 5 ilustra o impacto direto do cruzamento de dados na recuperação das variáveis. Enquanto a base original possuía apenas 28% dos filmes com as informações de orçamento e bilheteria mundial preenchidas simultaneamente, a base enriquecida alcançou 100% de completude. Esse preenchimento era um passo obrigatório para que o cálculo do Retorno sobre Investimento (ROI) fosse possível.

Para as variáveis financeiras secundárias (a bilheteria nos EUA e Canadá (`gross_us_canada`) e a arrecadação do fim de semana de estreia (`gross_opening_weekend`)) o preenchimento saltou da faixa dos 25% para mais de 80%. O fato de não atingir 100% de completude nestes casos específicos é um comportamento esperado. Por serem métricas muito detalhadas, elas nem sempre são divulgadas e consolidadas publicamente pelos estúdios, o que reflete uma limitação de transparência da própria indústria cinematográfica global.

4.1.5 Distribuição de Desempenho Financeiro

Para além da análise de valores absolutos de orçamento e arrecadação, a metodologia deste estudo aplicou as regras de negócio de desempenho sobre todo o universo das duas bases. A comparação proporcional entre elas evidencia que o conjunto de dados inicial apresentava uma forte distorção de seleção: havia uma tendência de que apenas produções de sucesso ou de grande orçamento possuísem dados financeiros documentados, ocultando o volume real de fracassos de bilheteria.

Figura 6 – Distribuição de Desempenho Financeiro Geral



Fonte: Elaborado pela autora (2026).

A Figura 6 consolida a maior contribuição do processo de engenharia de dados deste trabalho. Na base original, a categoria “Dados Faltantes / Não Avaliado” obstruía a utilidade do *dataset*, correspondendo a cerca de 72% de toda a base. Nas poucas obras que continham dados válidos, os cenários dividiam-se de forma tímida: aproximadamente 13% de Flops, 7% de Breakeven, 5% de Lucrativos e apenas 3% de Hits.

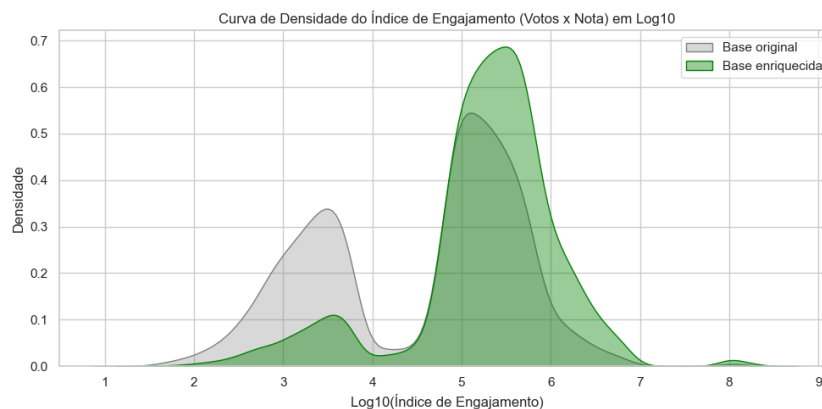
Com o enriquecimento e a eliminação completa das lacunas (0% de dados faltantes), uma nova distribuição econômica da indústria cinematográfica pôde ser observada. O cenário de fracasso comercial ganha o protagonismo real que estava oculto: a categoria “Fracasso Crítico (Flop)” salta para quase 60% de representatividade na base enriquecida. Em contrapartida, a fatia de filmes que apenas “Pagou as Contas (Breakeven)” mantém-se estável em 6,5%.

Após o processo de enriquecimento e a exclusão das obras que permaneceram com dados financeiros faltantes, observa-se uma alteração estrutural na distribuição de rentabilidade da amostra. O cenário de prejuízo financeiro passa a ser a classe majoritária: a categoria “Fracasso Crítico (Flop)” salta para quase 60% da base tratada. Em contrapartida, a fatia de filmes que apenas “Pagou as Contas (Breakeven)” mantém-se estável em 6,5%.

O dado mais revelador encontra-se nas faixas superiores: os filmes classificados como “Lucrativo” expandem-se para quase 20% e os “Alta Rentabilidade (Hit)” consolidam-se em 15% da base enriquecida. Essa configuração demonstra que, embora o fracasso (Flop) seja o resultado estatisticamente mais provável para uma produção isolada, o *pipeline* de captura conseguiu resgatar uma quantidade expressiva de obras economicamente bem-sucedidas que estavam completamente sem documentação financeira na base original, fornecendo ao modelo de *Machine Learning* um volume robusto de exemplos de sucesso para o aprendizado.

4.1.6 Comportamento e Densidade do Engajamento

Figura 7 – Curva de Densidade do Índice de Engajamento



Fonte: Elaborado pela autora (2026).

A assinatura estatística do impacto do *web scraping* sobre o comportamento do público pode ser observada na curva de densidade do Índice de Engajamento (Figura 7). Ambas as bases exibem o seu pico principal de concentração exatamente na mesma região do espectro logarítmico, o que atesta a manutenção da consistência identitária do *dataset*. Contudo, há uma mudança estrutural crucial na base das curvas.

Na base original (cinza), observa-se um segundo pico menor, mas bastante proeminente, situado na faixa de baixo engajamento (entre 3.0 e 4.0). Na base enriquecida (verde), esse pico de baixa repercussão foi drasticamente reduzido, enquanto a densidade do pico principal expandiu-se, tornando-se mais alta e larga.

Este comportamento indica que o processo de enriquecimento agiu como um filtro natural de relevância comercial. Filmes posicionados no vale de engajamento de 3.0 a 4.0 representam obras de apelo popular quase nulo e sem circulação de mercado significativa. Consequentemente, tais produções raramente possuem registros de orçamento ou bilheteria arquivados de forma pública e estável na Wikipedia ou no TMDb. Ao exigir o preenchimento do par financeiro para a modelagem do ROI, o processo de filtragem limitou involuntariamente a representatividade de produções independentes, regionais ou de nicho na base final. Isso se deve à escassez histórica de dados financeiros públicos para essas obras, concentrando a amostra resultante no circuito comercial de maior escala, cujas métricas de orçamento, bilheteria e engajamento são documentadas e abertas.

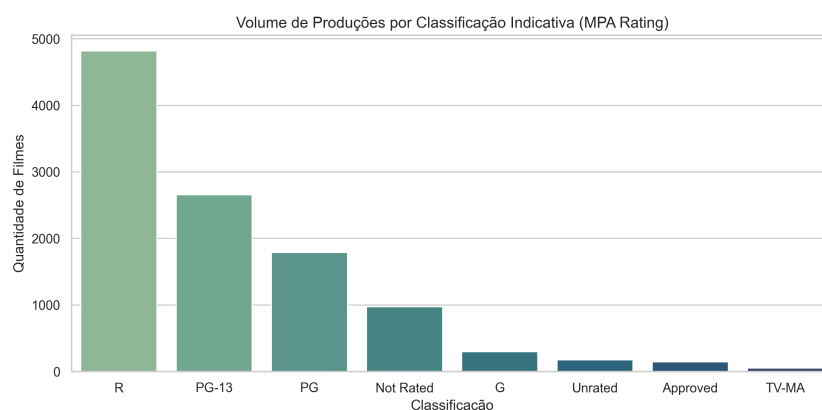
4.2 Perfil da Base Resultante

Após o processo de enriquecimento e a aplicação dos filtros necessários, a base de dados consolidou-se como um *dataset* focado em obras com integridade de informações financeiras. Antes de avançarmos para a etapa de clusterização, é importante compreender as características deste conjunto final, delineando o perfil da amostra e as relações numéricas entre suas variáveis.

4.2.1 Características Demográficas e Hegemonia Geográfica

A distribuição das variáveis categóricas revela aspectos estruturais do mercado cinematográfico. No que tange à restrição de público, observada através da classificação indicativa norte-americana (MPA), a amostra é fortemente dominada por produções voltadas ao público adulto.

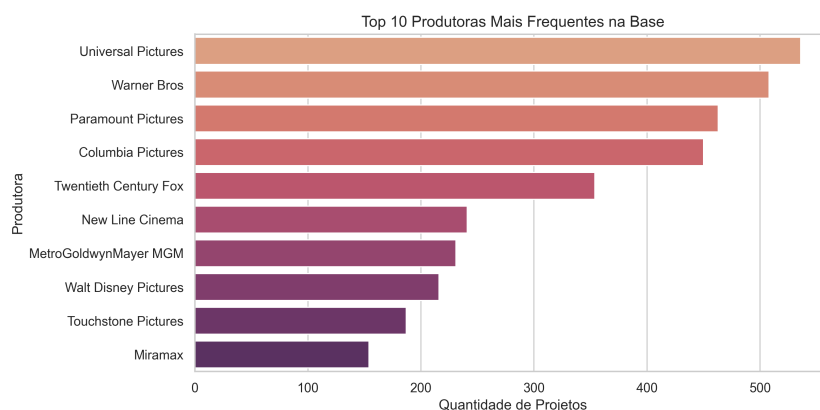
Figura 8 – Distribuição de Filmes por Classificação Indicativa (MPA)



Fonte: Elaborado pela autora (2026).

Na Figura 8, a classificação “R” (*Restricted*) representa quase 50% dos filmes do *dataset*, seguida pela “PG-13” (aproximadamente 20%). Essa prevalência de classificações restritivas alinha-se ao fato de que, conforme demonstrado na análise comparativa de gêneros anterior, a base consolidada mantém uma forte estrutura de dramas e obras de ação, além de ter ganhado uma parcela de comédias de humor negro (*Dark Comedy*), gêneros que frequentemente abordam temáticas maduras e linguagens voltadas ao público adulto.

Figura 9 – Distribuição por Produtoras (Top 10)



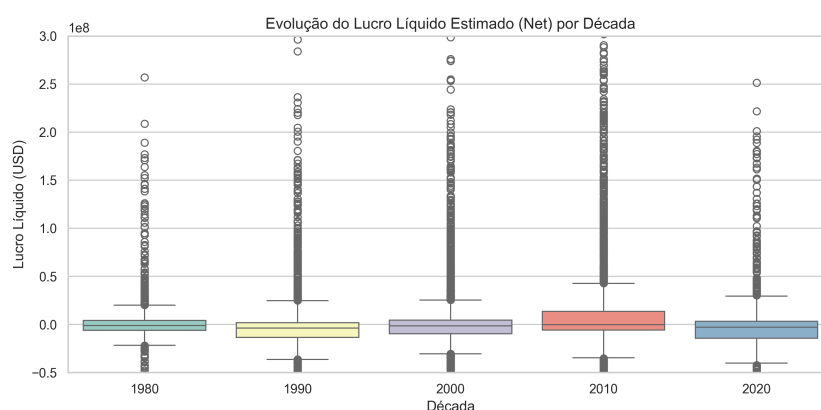
Fonte: Elaborado pela autora (2026).

Em relação aos estúdios produtores (Figura 9), constata-se uma notável pulverização do mercado. Mesmo produtoras gigantes como Universal Pictures e Warner Bros. detêm parcelas individuais de apenas 4% a 5% do volume total de 11.828 filmes.

Como evidenciado nos gráficos comparativos da seção anterior, essa distribuição demográfica é acompanhada por uma forte hegemonia geográfica, onde os Estados Unidos respondem por quase 50% da base final, refletindo a maior transparência financeira do eixo hollywoodiano. Da mesma forma, a análise prévia do perfil de arrecadação confirmou a polarização comercial extrema do setor: a maioria dos registros enquadra-se como "Fracasso Crítico (Flop)", evidenciando que, uma vez superado o limiar de prejuízo, os filmes tendem a saltar diretamente para patamares consolidados de lucro.

4.2.2 Evolução Financeira e Análise de Correlação

Figura 10 – Lucro Líquido Médio por Década de Lançamento

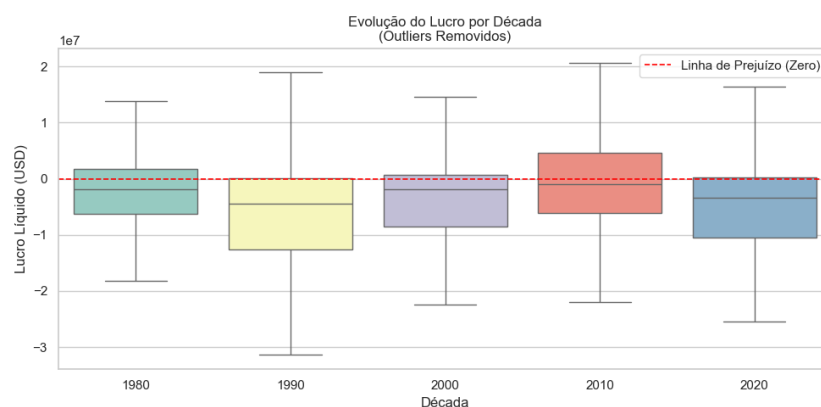


Fonte: Elaborado pela autora (2026).

Avançando para a evolução histórica do lucro líquido médio ao longo das décadas (Figura 10), identifica-se um padrão de crescimento contínuo desde 1960.

Nesta primeira visualização, nota-se um aparente crescimento desde 1980, atingindo a sua maior amplitude na década de 2010. Contudo, é preciso considerar o impacto dos *outliers* nessa distribuição. Os *outliers* representam os *blockbusters* e superproduções com lucros muito acima do normal. Esses valores extremos puxam a distribuição e as médias gerais para cima, mascarando o verdadeiro comportamento da maior parte do mercado.

Figura 11 – Lucro Líquido por Década (Sem Outliers)



Fonte: Elaborado pela autora (2026).

Para tentar observar a realidade da maioria dos filmes sem essa distorção, gerou-se uma segunda visualização (Figura 11), removendo os *outliers* por meio do método do Intervalo Interquartil (IQR).

Ao analisar o gráfico sem os valores atípicos, percebe-se que aquele crescimento anterior era mascarado pelos sucessos extremos. A mediana e a concentração principal dos lucros parecem ficar na mesma faixa de valor ao longo de praticamente todas as décadas analisadas, próximo a linha de zero. Apenas na década de 2020 observa-se uma contração dessa linha central, que cai para a faixa de prejuízo e regride aos patamares financeiros dos anos 1980 e 1990. Este decréscimo recente pode refletir tanto o impacto da pandemia de COVID-19, que paralisou a operação tradicional dos cinemas, quanto a consolidação das plataformas de *streaming*, que alteraram os hábitos de consumo e a forma como produções de alto orçamento são lançadas.

Figura 12 – Mapa de Correlação de Pearson



Fonte: Elaborado pela autora (2026).

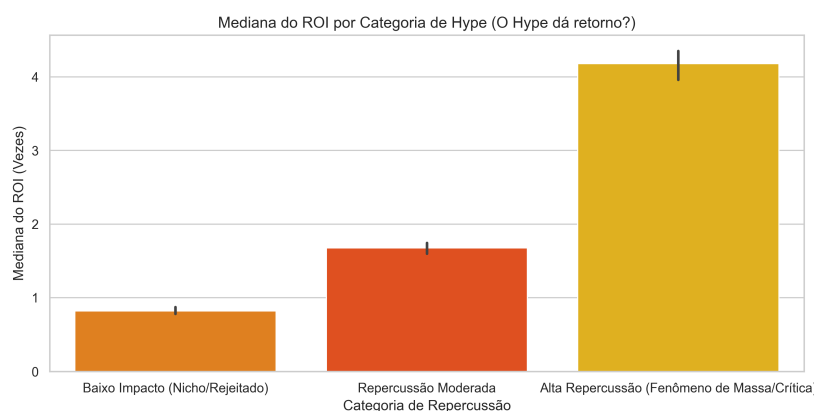
Para investigar a dependência linear entre as variáveis numéricas, utilizou-se a Matriz de Correlação de Pearson (Figura 12).

Os resultados demonstram relações estruturais importantes. A correlação forte (0.72) entre o orçamento (**budget**) e a bilheteria (**gross**) indica que filmes com maiores orçamentos tendem a apresentar maiores arrecadações. Da mesma forma, o número de avaliações correlaciona-se moderadamente com o orçamento (0.56) e a bilheteria (0.58), confirmando que filmes caros e lucrativos atraem maior engajamento do público.

Entretanto, duas ausências de correlação chamam a atenção. A primeira é a nota do IMDb (**nota_imdb**), que apresenta associação quase nula com as métricas financeiras, demonstrando que a aprovação do público não se traduz necessariamente em sucesso comercial direto. A segunda é a variável de Retorno sobre Investimento (ROI). O ROI possui correlações muito baixas com todas as outras métricas da base, incluindo o próprio orçamento. Isso indica que injetar mais capital na produção de um filme não garante que ele terá um fator multiplicador de lucro (ROI) maior, reforçando a alta complexidade de

prever a rentabilidade no setor audiovisual.

Figura 13 – ROI Médio por Classificação de Hype



Fonte: Elaborado pela autora (2026).

Essa característica complexa do ROI fica ainda mais clara ao ser cruzada com a métrica de Hype (Figura 13).

O cruzamento evidencia que os filmes de “Baixo Impacto (Nicho)” concentram a maior média de rentabilidade (ROI superior a 7.5x), enquanto as produções de “Alta Repercussão” apresentam o menor multiplicador médio (próximo a 3.0x). Esse comportamento é um reflexo da escala do investimento inicial. Filmes de nicho operam com orçamentos reduzidos, fazendo com que arrecadações absolutas menores elevem o multiplicador de forma exponencial. Já as obras de grande repercussão exigem orçamentos massivos. Nessas faixas de investimento, alcançar múltiplos elevados torna-se exponencialmente mais difícil, visto que dobrar ou triplicar um custo gigantesco demanda uma penetração de mercado global. Portanto, embora o ROI proporcional seja menor nas superproduções, ele indica um volume absoluto de sucesso comercial mais complexo de ser atingido.

4.3 Resultados da Clusterização

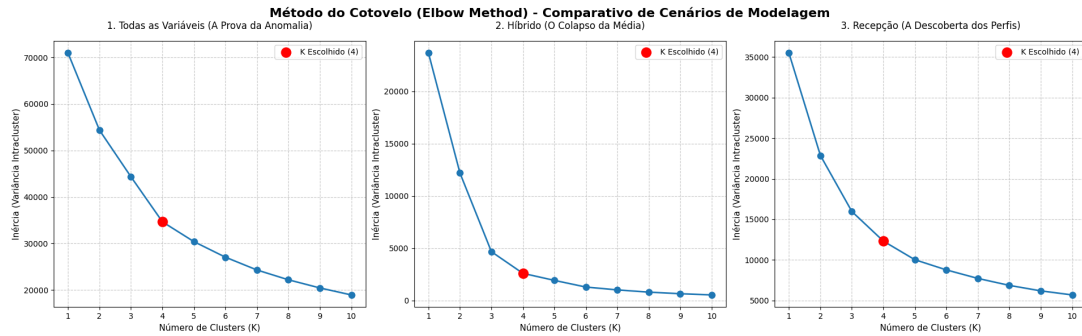
A aplicação do algoritmo K-Means teve como objetivo segmentar a indústria cinematográfica em perfis comportamentais e financeiros claros. No entanto, o cinema é um mercado caracterizado por alta assimetria, onde milhares de filmes independentes convivem com *blockbusters* multibilionários. Para garantir que o agrupamento não fosse uma mera coincidência matemática, a metodologia estabeleceu o teste de diferentes subconjuntos de variáveis (cenários), buscando encontrar a combinação que melhor explicasse a realidade da indústria.

4.3.1 A Definição do Hiperparâmetro K

A primeira etapa crítica na extração dos resultados foi a definição do número de *clusters* (K). Para assegurar a validade e a comparabilidade metodológica, o Método do Cotovelo (*Elbow Method*) foi aplicado simultaneamente aos três cenários finais de modelagem, conforme ilustra a Figura 14.

A análise visual das três curvas de inércia revela um comportamento consistente: a queda abrupta da variância estabiliza-se, transformando-se em um declínio suave, na

Figura 14 – Método do Cotovelo Aplicado aos Três Cenários de Modelagem



Fonte: Elaborado pela autora (2026).

região próxima ao número quatro. Optou-se por fixar $K = 4$ para todas as abordagens por duas razões fundamentais. Primeiro, a padronização permite comparar as falhas e sucessos de cada modelo na mesma escala. Segundo, $K = 4$ traduz perfeitamente a matriz clássica do mercado de entretenimento (Alto e Baixo Risco *versus* Alto e Baixo Retorno), permitindo que o algoritmo busque organicamente esses quatro quadrantes.

A seguir, são detalhados os três cenários testados, evidenciando as limitações reais dos dados e, por fim, a descoberta do perfil definitivo de mercado.

4.3.2 Cenário 1: Todas as Variáveis

O primeiro cenário testou a robustez do algoritmo ao fornecer todas as variáveis simultaneamente (orçamento, bilheteria, lucro líquido, ROI, nota do IMDb e avaliações).

Tabela 1 – Médias dos Agrupamentos - Cenário Todas as Variáveis

Cluster	Participação	Orçamento	Bilheteria	Lucro (Net)	ROI	Nota
0	23,1% (2.732)	\$ 2,8 M	\$ 2,7 M	-\$ 1,4 M	5,22x	6.1
1	34,5% (4.081)	\$ 32,7 M	\$ 147,0 M	\$ 40,7 M	8,82x	6.6
2	42,4% (5.014)	\$ 28,0 M	\$ 24,7 M	-\$ 15,7 M	0,80x	6.2
3 (Anomalia)	0,0% (1)	\$ 15 mil	\$ 193,3 M	\$ 96,6 M	12.890x	6.3

Fonte: Elaborado pela autora (2026).

A Tabela 1 comprova uma limitação estrutural na análise de dados do cinema: a existência de outliers tão extremos que corrompem a métrica de distância euclidiana. O modelo reservou um *cluster* inteiro (Grupo 3) para isolar uma única obra, identificada na base de dados como o filme *Atividade Paranormal*. Produções independentes de terror como esta, feitas com orçamentos caseiros (\$15 mil) mas que faturam centenas de milhões de dólares globalmente, quebram por completo a proporcionalidade matemática do Retorno sobre Investimento (gerando um ROI irreal de 12.890x). Este cenário prova que tentar modelar a indústria utilizando proporções e valores absolutos ao mesmo tempo inviabiliza a segmentação lógica da massa de filmes comuns.

4.3.3 Cenário 2: Híbrido

Para contornar o problema dos multiplicadores irreais, o segundo cenário utilizou apenas os indicadores compostos desenvolvidos na etapa de Engenharia de Atributos: o Índice de Sucesso (que pondera ROI e nota) e o Score de Engajamento. A hipótese era que índices balanceados gerariam grupos mais sensatos.

Tabela 2 – Médias dos Agrupamentos - Cenário Híbrido

Cluster	Participação	Orçamento	Bilheteria	Lucro (Net)	ROI	Nota
0	11,2% (1.325)	\$ 8,5 M	\$ 17,4 M	\$ 219 mil	3,00x	5.8
1 (Massa)	50,6% (5.985)	\$ 15,1 M	\$ 20,6 M	-\$ 4,8 M	3,55x	6.1
2	0,6% (71)	\$ 79,9 M	\$ 569,1 M	\$ 204,5 M	10,46x	8.4
3	37,6% (4.447)	\$ 39,2 M	\$ 123,0 M	\$ 22,2 M	9,27x	6.8

Fonte: Elaborado pela autora (2026).

Contudo, como observado na Tabela 2, esta abordagem gerou uma super-simplificação. Ao fundir variáveis qualitativas e financeiras antes da modelagem, o algoritmo aglutinou mais de 50% de toda a base de dados no Cluster 1. Esta “massa cinzenta” misturou filmes que deram prejuízo real com obras de rentabilidade moderada, simplesmente porque ambas possuíam índices médios parecidos. Isso comprova que o mercado cinematográfico não pode ser resumido a um índice de “nota média”, pois o comportamento do caixa (*box office*) opera de forma desvinculada do comportamento crítico.

4.3.4 Cenário 3: Recepção e Lucro Absoluto

Com as limitações mapeadas, o cenário definitivo abandonou o ROI e focou em dimensões diretas: o Dinheiro Real gerado (Net logarítmico), a Qualidade Percebida (Nota do IMDb) e o Tamanho do Público (Volume de Avaliações). Ao isolar essas variáveis, o algoritmo finalmente conseguiu dividir a indústria em quatro perfis comercialmente coerentes, ignorando as anomalias numéricas.

Tabela 3 – Médias dos Agrupamentos - Cenário de Recepção (Modelo Selecionado)

Perfil de Mercado	Fatia	Orçamento	Bilheteria	Lucro (Net)	Nota
0. Os Fracassos Absolutos	18,3% (2.165)	\$ 22,9 M	\$ 21,3 M	-\$ 12,3 M	5.0
1. Titãs Comerciais	34,9% (4.128)	\$ 30,8 M	\$ 139,5 M	\$ 38,9 M	6.7
2. Nicho de Baixo Orçamento	11,0% (1.301)	\$ 8,3 M	\$ 17,6 M	\$ 506 mil	5.9
3. Os Injustiçados	35,8% (4.234)	\$ 22,2 M	\$ 20,3 M	-\$ 12,0 M	6.8

Fonte: Elaborado pela autora (2026).

A leitura das médias revela dinâmicas de mercado que confirmam empiricamente o que produtores experienciam na prática:

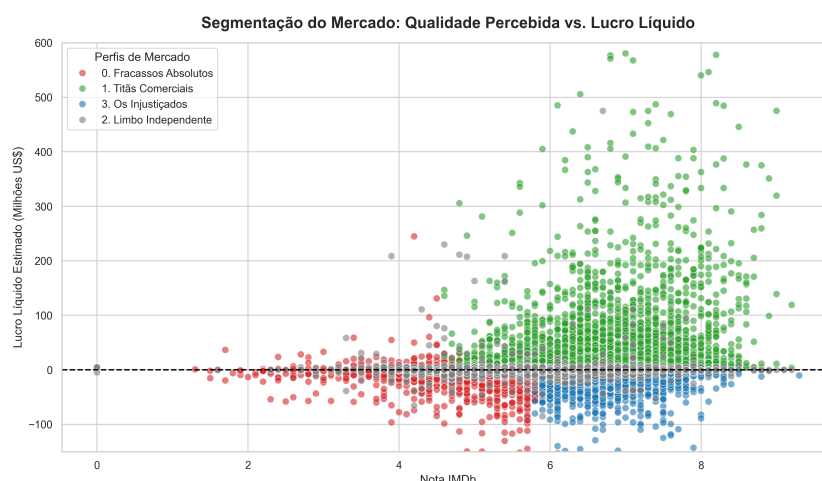
- **Titãs Comerciais (Cluster 1):** Concentrando quase 35% da amostra, são as obras que combinam aprovação de público (nota média 6.7) com engajamento massivo (mais de 360 mil votos). Essa tração popular converte-se em sustentabilidade financeira, garantindo o maior lucro absoluto da base (média de \$38,9 milhões de Net).

- **Os Injustiçados (Cluster 3):** O agrupamento mais revelador do estudo (35,8% da base). O algoritmo separou os filmes que registraram a mais alta aprovação do público no *dataset* (nota média 6.8). Trata-se de obras com boa percepção de qualidade por parte dos espectadores, mas que operam em prejuízo financeiro (média de -\$12 milhões de Net). Representam produções que agradam o seu nicho, mas que falham em converter essa aceitação em viabilidade comercial nas bilheterias.
- **Fracassos Absolutos (Cluster 0):** Obras com o menor índice de aceitação da amostra. Além do grave prejuízo financeiro (média de -\$12,3 milhões de Net), essas produções recebem as piores avaliações por parte do público (nota média 5.0), evidenciando projetos que não obtiveram êxito de engajamento nem retorno comercial.
- **Nicho de Baixo Orçamento (Cluster 2):** Uma fatia minoritária (11% da base). São obras com orçamentos restritivos (média de \$8,3 milhões) e que operam com baixo volume de avaliações (apenas 525 votos em média). Contudo, por possuírem custos contidos, conseguem atingir o limiar neutro do *breakeven* (média de \$506 mil de lucro), garantindo a recuperação do investimento sem atingir repercussão massiva no mercado.

4.3.5 Análise Qualitativa: Dissecando os Agrupamentos de Mercado

Para além das médias financeiras e das notas isoladas, a verdadeira riqueza da modelagem por K-Means reside na capacidade de observar a distribuição física do mercado e as características qualitativas de seus componentes. A análise gráfica e o cruzamento com os metadados textuais revelam a “identidade” de cada *cluster* mapeado.

Figura 15 – Segmentação do Mercado: Qualidade Percebida vs. Lucro Líquido

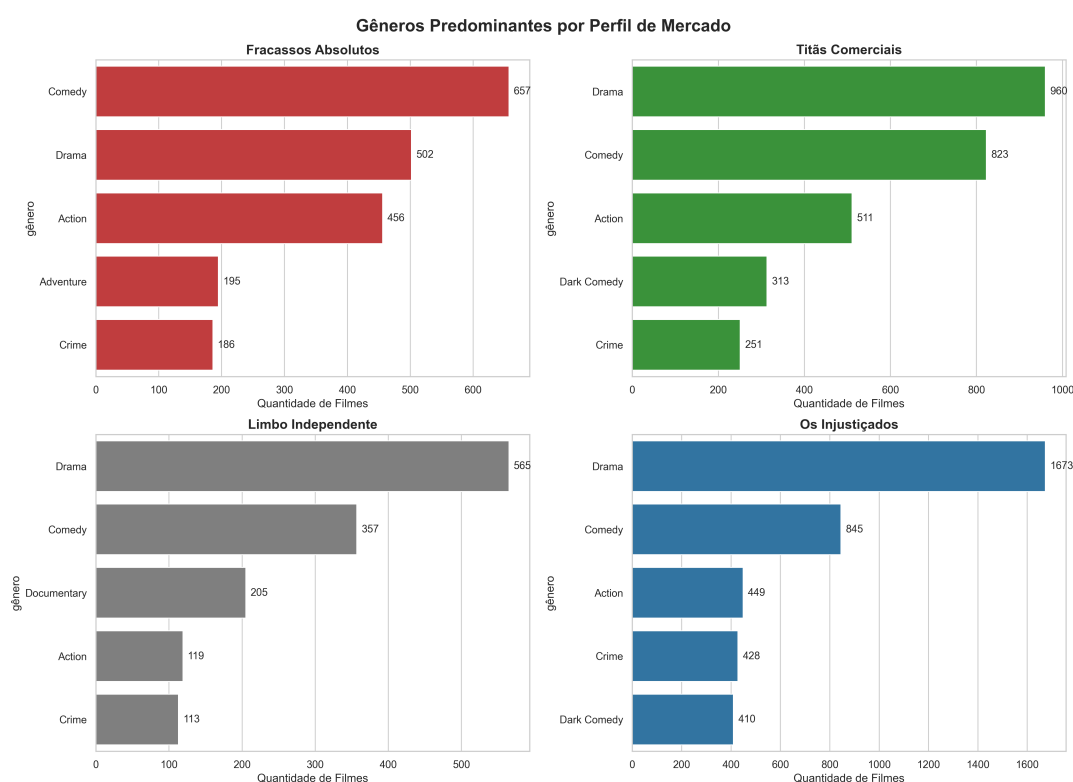


Fonte: Elaborado pela autora (2026).

A Figura 15 apresenta o mapeamento espacial do mercado cinematográfico. A linha tracejada demarca a fronteira do *breakeven* financeiro (lucro zero). A estrutura ilustra a gravidade da indústria: uma base massiva e densa de filmes que operam abaixo da linha da rentabilidade, contrastando com o espalhamento vertical de produções altamente lucrativas.

É possível observar que o quadrante de alto lucro e alta nota (onde a maior parte dos pontos verdes dos Titãs Comerciais está alocada) é consideravelmente mais esparsa e alongado no eixo vertical do que os agrupamentos de prejuízo, provando que o sucesso financeiro absoluto é uma raridade que descola dos padrões da massa. Além disso, a mancha azul (Os Injustiçados) destaca uma anomalia estrutural da indústria: obras com qualidade acima da média (notas consistentemente acima de 6.0), mas que afundam no eixo financeiro, misturando-se especialmente aos Fracassos Absolutos (pontos vermelhos) que possuem avaliações drasticamente inferiores.

Figura 16 – Distribuição dos Cinco Gêneros Predominantes por Perfil de Mercado



Fonte: Elaborado pela autora (2026).

O cruzamento dos agrupamentos matemáticos com os dados de produção confirmou que o sucesso e o fracasso comercial carregam assinaturas de gênero muito específicas, conforme ilustrado na Figura 16.

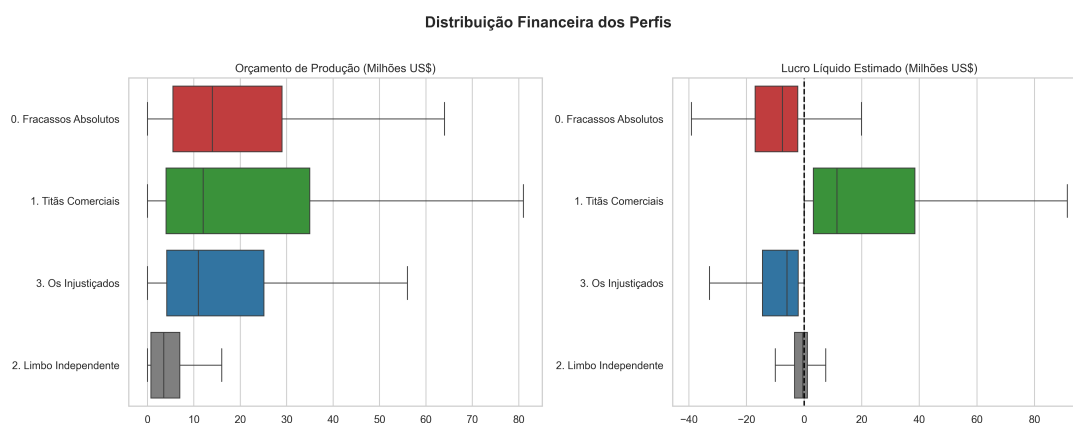
Os Titãs Comerciais e Os Injustiçados lideram a amostra com um grande volume de filmes em gêneros massivos (Comédia, Ação e Drama, superando largamente a marca de 500 a 1000 títulos em suas categorias principais). Contudo, a diferença reside no fato de que filmes focados primariamente em Drama ou Crime, muitas vezes presentes na elite crítica de premiações (característicos d'Os Injustiçados), falham mais frequentemente em atingir a sustentabilidade financeira quando comparados à fluidez da Aventura e Comédia, que marcam forte presença no espectro dos Titãs.

Um ponto de divergência notável reside no Nicho de Baixo Orçamento. Diferente dos demais agrupamentos dominados por ficção comercial, este perfil tem como gênero principal o Documentário (*Documentary*), acompanhado por volumes substancialmente menores de Drama e Comédia. O formato documental demanda aportes financeiros radicalmente

inferiores, o que explica a média orçamentária restrita apontada na modelagem anterior (US\$ 8,3 milhões). Ao operarem num espectro de custos mínimos, esses projetos evitam o risco de um fracasso retumbante na bilheteria, mas encontram-se presos a um nicho, fenômeno atestado pelo baixíssimo volume médio de avaliações do grupo (apenas 525 votos).

A distribuição exata do risco financeiro entre os agrupamentos está evidenciada na Figura 17.

Figura 17 – Distribuição de Orçamento e Lucro Líquido por Perfil



Fonte: Elaborado pela autora (2026).

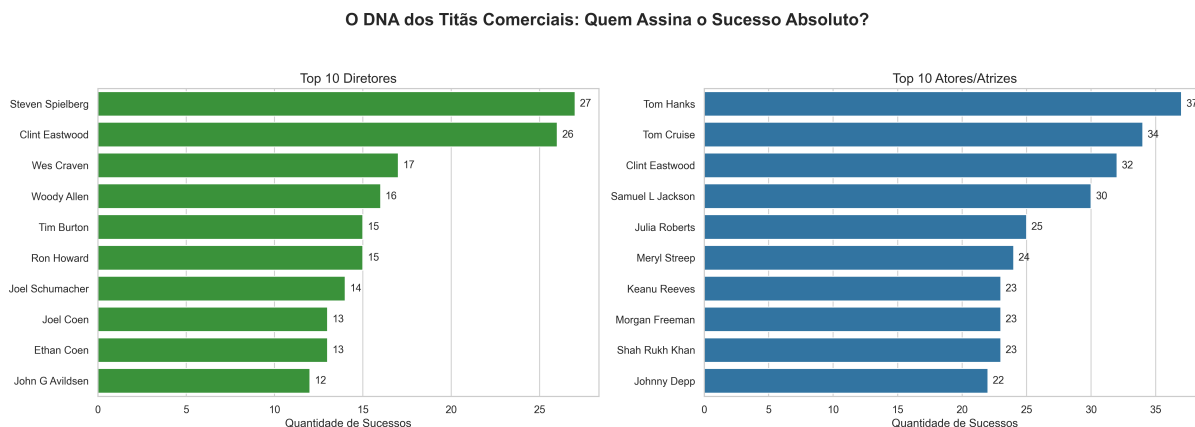
Observa-se que, apesar da distribuição de orçamentos (*boxplot* à esquerda) possuir medianas relativamente próximas entre Titãs, Injustiçados e Fracassos, o reflexo desse gasto na receita líquida (*boxplot* à direita) é radicalmente distinto. Os Titãs apresentam uma variância positiva extrema, com *outliers* encostando nos US\$ 600 milhões de lucro, enquanto os blocos centrais de Fracassos e Injustiçados são esmagados abaixo da linha neutra. Isso demonstra de forma contundente que um alto orçamento de produção não oferece garantia estatística contra o prejuízo comercial.

Por fim, o algoritmo isolou o grupo de excelência comercial absoluta (Titãs Comerciais) para uma varredura de metadados, revelando os padrões estruturais das obras de alta rentabilidade, detalhados na Figura 18.

A análise gráfica, ancorada no relatório textual do agrupamento, corrobora a premissa de que a indústria de entretenimento atua fortemente sob a força de talentos estabilizados. O ranking dos diretores mais lucrativos da base de dados atesta uma acentuada concentração criativa: a liderança isolada pertence a Steven Spielberg, atrelado a 27 projetos vitoriosos neste agrupamento de excelência. Em seguida, surgem cineastas fundamentados em grandes marcas ou com assinaturas de gênero altamente massificadas, como Clint Eastwood (26 filmes), Wes Craven (17 filmes) e Tim Burton (15 filmes). Na dimensão do elenco principal, a garantia de conversão em bilheteria repousa em figuras históricas de Hollywood, como Tom Hanks (37 projetos lucrativos) e Tom Cruise (34 projetos).

Esta combinação da clusterização numérica com a decodificação dos atributos demonstra que, embora o sucesso comercial costume vir acompanhado de boa recepção, a qualidade técnica isolada não é determinante de lucro. O cinema opera como um ecossistema de alto risco que privilegia a segurança atuarial de astros renomados e gêneros

Figura 18 – O DNA dos Titãs Comerciais: Principais Diretores e Atores



Fonte: Elaborado pela autora (2026).

de apelo global, punindo de forma severa produções financeiramente robustas que falham em engajar a massa consumidora.

4.3.6 Validação Algorítmica: A Complexidade do Mercado via DBSCAN

Para atestar a robustez da escolha do K-Means como modelo principal de segmentação, a pesquisa aplicou o algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) como método de contraprova. Diferente do K-Means, que força a alocação dos dados em K agrupamentos baseados na variância e distância até um centroide, o DBSCAN mapeia as conexões a partir da densidade da distribuição vetorial. Esta característica o torna implacável na detecção de geometrias irregulares e na identificação de ruídos (outliers), mas também extremamente sensível à natureza das variáveis.

Para testar a topologia do mercado cinematográfico, o algoritmo foi submetido a duas abordagens distintas, que revelaram a complexidade inerente aos dados da indústria:

Na primeira execução, o DBSCAN foi alimentado exclusivamente com variáveis contínuas padronizadas (logaritmo do lucro, volume de avaliações e nota do IMDb). O resultado evidenciou a natureza contínua do mercado: o algoritmo isolou cerca de 4% da base como ruído absoluto (projetos com comportamentos estatísticos atípicos, limpando os extremos) e aglutinou a esmagadora maioria dos dados — mais de 11 mil filmes — em um único e massivo *cluster* central, acompanhado por pequenos grupos periféricos irrelevantes. Matematicamente, isso ocorre porque os dados financeiros e críticos do cinema formam uma nuvem densa e contínua; não existem “abismos” ou espaços vazios claros entre um filme que lucrou US\$ 10 milhões e outro que lucrou US\$ 11 milhões. Como a densidade não se rompe, o DBSCAN conecta toda a massa.

Na segunda execução, a abordagem foi invertida. Submeteu-se a base a uma matriz de alta dimensionalidade, incorporando as variáveis categóricas ordinais (classificações de retorno e repercussão) e a fragmentação de gêneros via One-Hot Encoding. Ao parametrizar o raio de busca espacial (*eps*) em 2.0, o cenário mudou drasticamente: o algoritmo estilhou a base em 36 microagrupamentos densos. Este fenômeno, atrelado à “Maldição da Dimensionalidade”, ocorre porque a inserção de dezenas de colunas binárias (0 e 1

para os gêneros) transforma o espaço vetorial. O DBSCAN passou a encontrar “ilhas” isoladas de filmes que possuem combinações categóricas exatas e inflexíveis (por exemplo, um grupo estrito apenas de filmes de “Ação + Alto Hype + Sucesso Comercial”).

5 CONSIDERAÇÕES FINAIS

A presente pesquisa adaptou seu escopo analítico durante o desenvolvimento prático. Inicialmente focada na predição de bilheterias contínuas, a etapa de análise exploratória revelou a alta assimetria e volatilidade dos dados de receita bruta, dificultando a formulação de previsões exatas consistentes. Consequentemente, o escopo transicionou para o Aprendizado de Máquina Não Supervisionado, estabelecendo como meta central segmentar a estrutura latente do mercado para fins de inteligência de negócios. Ao final deste processo, foi possível atestar o atingimento dos três objetivos de pesquisa propostos.

O primeiro objetivo da pesquisa (RG1) visava construir e higienizar um conjunto de dados enriquecido para mitigar a escassez de informações financeiras. Este objetivo foi alcançado com a entrega do próprio *dataset* consolidado. O processo de Engenharia de Dados corrigiu as inconsistências estruturais da base original do Kaggle — como avaliações corrompidas e lacunas de orçamento — utilizando técnicas de *web scraping* e consumo de APIs. Adicionalmente, a etapa de enriquecimento englobou o tratamento de outliers, a transformação logarítmica de métricas de volume e a criação de novas variáveis analíticas essenciais para o estudo, tais como a formulação do Lucro Líquido (*Net Profit*), o Retorno sobre Investimento (ROI) e os índices numéricos de engajamento e sucesso.

No que tange ao segundo objetivo (RG2), que buscava analisar o perfil geral da base para mapear as relações entre investimentos, rentabilidade e engajamento, a pesquisa documentou de forma empírica a complexidade da indústria. O mapeamento evidenciou que a relação entre orçamentos e lucros absolutos não obedece a um comportamento estritamente linear, sendo afetada pela lei de rendimentos decrescentes em superproduções e por multiplicadores exponenciais em filmes de baixíssimo orçamento.

Por fim, o terceiro objetivo (RG3) consistia em identificar e segmentar grupos de obras através de *clustering*, testando a viabilidade de categorizar o setor em nichos estruturados. Para tal, conduziu-se uma análise comparativa. O algoritmo DBSCAN confirmou matematicamente a topologia densa e hiperfragmentada da indústria, mas apresentou limitações práticas para a formulação de estratégias de mercado. Em contrapartida, o K-Means consolidou-se como a arquitetura ideal, alcançando o objetivo ao segmentar o mercado em quatro agrupamentos distintos, baseados em tendências de lucro, nota do público e engajamento:

- **Titãs Comerciais:** O grupo com maior indicativo de sustentabilidade da base, registrando as maiores médias conjuntas de lucro financeiro e avaliações massivas do público.
- **Os Injustiçados:** Agrupamento de produções detentoras das maiores médias de nota técnica (6.8), mas que tendem a operar na faixa de prejuízo financeiro.
- **Fracassos Absolutos:** Filmes que apresentaram simultaneamente menor qualidade percebida (nota média 5.0) e déficits no balanço financeiro.

- **Nicho de Baixo Orçamento:** Uma fatia composta predominantemente por produções de custo reduzido, cujo impacto mais restrito (votos em torno de 525) permite atingir um *breakeven* técnico, ainda que sem gerar ampla tração comercial.

A validação do RG3 também indicou que resumir um ecossistema tão diverso em apenas quatro categorias não captura toda a variância dos dados. Os perfis destacam as principais tendências de rentabilidade, mas produções intermediárias tendem a apresentar sobreposição espacial. Esse comportamento responde à provocação do RG3 ao concluir que existe uma limitação natural na tentativa de segmentar o cinema utilizando exclusivamente variáveis numéricas contínuas, apontando a necessidade de abordagens híbridas no futuro.

5.1 Trabalhos Futuros

Com base nas observações levantadas por esta pesquisa e nas limitações inerentes ao modelamento puramente contínuo, sugerem-se as seguintes direções para a continuidade do estudo:

1. **Modelos de classificação:** Aproveitar os grupos formados pelo K-Means como rótulos para treinar algoritmos supervisionados (como Random Forest ou Redes Neurais). A ideia é testar se é viável estimar o risco financeiro de um projeto utilizando apenas os dados da sua fase de pré-produção.
2. **Dados de streaming:** Incorporar métricas de audiência de plataformas de *streaming* ao *dataset*. Isso ajudaria a entender se os filmes do grupo “Os Injustiçados” conseguem recuperar o prejuízo das bilheteiras ao longo do tempo no mercado digital.
3. **Processamento de Linguagem Natural:** Aplicar análise de sentimento nas críticas textuais do público em vez de usar apenas a nota final. Explorar os textos pode trazer perspectivas sobre a recepção dos filmes que os números absolutos acabam escondendo.
4. **Novas métricas e expansão categórica:** Propor e testar novos atributos de produção (como diretores, estúdios ou chaves de gênero) no algoritmo para avaliar se é possível expandir a quantidade de grupos, descobrindo subperfis que o modelo exclusivamente numérico atual não conseguiu mapear.

REFERÊNCIAS

Ankit; Lakshmi, M.; Shastry, K. A.; Sandilya, A.; Shekhar, R. A comparative analysis of machine learning approaches for movie success prediction. In: IEEE. **Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)**. [S.l.], 2020. p. 684–689. Citado na página 3.

Banco Mundial. **Official exchange rate (LCU per US\$, period average)**. 2026. Disponível em: <<https://data.worldbank.org/indicator/PA.NUS.FCRF>>. Acesso em: 15 mar. 2026. Citado na página 6.

Base dos Dados. **IMDB Movies: Tabela Melhores Filmes por Ano (top_movies_per_year)**. 2023. Disponível em: <<https://basedosdados.org/dataset/br-basedosdados-imdb-movies>>. Acesso em: 8 jun. 2026. Citado na página 3.

Caram, M.; Seufitelli, D. B.; Dalip, D. H. **Cinematic Profitability Dataset**. [S.l.]: [s.n.], 2026. GitHub. Disponível em: <<https://github.com/MarcelaCaram/cinematic-profitability-dataset>>. Acesso em: 24 jun. 2026. Citado na página 9.

Dhir, R.; Raj, A. Movie success prediction using machine learning algorithms and their comparison. In: IEEE. **2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)**. [S.l.], 2018. p. 385–390. Citado na página 3.

Fernandes, F. A.; Oliveira, L. N. d. Uma abordagem de aprendizado de máquina para prever popularidade com base em dados históricos. In: **Brazilian Symposium on Multimedia and the Web (WebMedia)**. [s.l.]: [s.n.], 2021. Disponível em: <<https://sol.sbc.org.br/index.php/brasnam/article/view/16130>>. Acesso em: 8 jun. 2026. Citado na página 1.

G1. **Acha caro ir ao cinema? Entenda como se calcula o valor dos ingressos e a divisão da bilheteria**. 2020. Disponível em: <<http://g1.globo.com/pop-arte/cinema/noticia/2020/08/26/acha-car0-ir-ao-cinema-entenda-como-se-calcula-o-valor-dos-ingressos-e-a-divisao-da-bilheteria.ghml>>. Acesso em: 8 jun. 2026. Citado na página 7.

Krauss, J.; Nann, S.; Simon, D.; Fischbach, K.; Gloor, P. Predicting movie success and academy awards through sentiment and social network analysis. In: **Proceedings of the 16th European Conference on Information Systems (ECIS)**. [S.l.]: [s.n.], 2008. p. 1–13. Citado na página 3.

Lash, M. T.; Zhao, K. Early predictions of movie success: the who, what, and when of profitability. **Journal of Management Information Systems**, Taylor & Francis, v. 33, n. 3, p. 874–903, 2016. Citado na página 2.

Ni, Y.; Dong, F.; Zou, M.; Li, W. Movie box office prediction based on multi-model ensembles. **Information**, v. 13, n. 6, 2022. ISSN 2078-2489. Disponível em: <<https://www.mdpi.com/2078-2489/13/6/299>>. Acesso em: 8 jun. 2026. Citado na página 1.

Oliveira, L. O. d.; Santos, E. Análise e predição do sucesso de músicas utilizando aprendizado de máquina. In: **Simpósio Brasileiro de Banco de Dados - Trabalhos Estendidos (SBBT)**. [s.l.]: [s.n.], 2022. Disponível em: <https://sol.sbc.org.br/index.php/sbbt_estendido/article/view/21841>. Acesso em: 8 jun. 2026. Citado na página 1.

Pereira, C.; Vasconcelos, M.; Silva, A. P. Análise comparativa dos artigos de filmes indicados ao oscar em três versões da wikipédia. In: **Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web**. Porto Alegre, RS, Brasil: SBC, 2024. p. 37–40. ISSN 2596-1683. Disponível em: <https://sol.sbc.org.br/index.php/webmedia_estendido/article/view/30473>. Acesso em: 8 jun. 2026. Citado na página 2.

Ru, Y.; Li, B.; Liu, J.; Chai, J. An effective daily box office prediction model based on deep neural networks. **Cognitive Systems Research**, v. 52, p. 182–191, 2018. ISSN 1389-0417. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1389041718300731>>. Acesso em: 8 jun. 2026. Citado na página 1.

Zhang, W.; Skiena, S. Improving movie gross prediction through news analysis. In: IEEE. **2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology**. [S.l.], 2009. v. 3, p. 301–304. Citado na página 3.

Agradecimentos

Quero começar agradecendo aos meus pais, Ricardo e Patrícia, pela base que me deram, pelo apoio incondicional e por sempre acreditarem em mim. Agradeço também ao meu padrasto, Duarte, e à minha madrastra, Viviane, que estiveram por perto durante a maior parte da minha vida e compartilharam toda essa jornada junto aos meus pais. Um abraço especial ao meu irmão, Victor, por me lembrar o motivo pelo qual vale a pena lutar por um mundo melhor. Agradeço a toda a minha família, tios, tias e primos, por construírem um ambiente tão acolhedor e seguro na minha vida.

Aos meus amigos do CEFET-MG, muito obrigada pela trajetória incrível que vivemos juntos. Sem vocês não seria possível chegar até aqui. Vocês tornaram esse caminho especial e memorável. Agradeço em especial ao Lucas, à Lara e ao Vinícius pelo vínculo pra vida toda e por toda força dada nos momentos mais difíceis. Ao Pedro Vitor, meu muito obrigada por viver ao meu lado a maior aventura da minha vida no intercâmbio.

Faço um agradecimento separado à minha amiga irmã, Lívia. Muito obrigada por sempre estar ao meu lado e ter me mantido de pé tantas vezes.

Deixo também um agradecimento sincero aos meus orientadores, Prof. Daniel Hasan e Prof. Danilo Boechat, sem os quais este trabalho sequer existiria. Obrigada por acreditarem em mim e aceitarem fazer parte de tudo isso.

Quero fazer uma menção saudosa aos meus avós e às minhas tias, que infelizmente não puderam vivenciar este momento comigo. Vô Jamil, Vó Teresinha, tia Marília, tia Betinha e tia Tete, eu carrego vocês comigo em cada conquista.

Por fim, agradeço principalmente a Deus, sem o qual nada disso aconteceria.



MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS
GERAIS
DEPARTAMENTO DE COMPUTAÇÃO - NG



FOLHA DE APROVAÇÃO DE TCC Nº 11 / 2026 - DECOM (11.56.03)

Nº do Protocolo: 23062.031835/2026-20

Belo Horizonte-MG, 19 de junho de 2026.

Marcela Camarano Caram Peito

**Perfis de rentabilidade no ecossistema cinematográfico: construção e
normalização de um dataset com dados abertos da Web**

Trabalho de Conclusão de Curso apresentado
ao Curso de **Engenharia de Computação** do
Centro Federal de Educação Tecnológica de
Minas Gerais, do Campus **Nova Gameleira**,
como parte do requisito para obtenção do grau
de Bacharel em **Engenharia de Computação**.

Aprovado em 19 de junho de 2026.

Banca Examinadora:

Prof. Dr. Daniel Hasan Dalip - Orientador
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Danilo Boechat Seufitelli - Coorientador
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Gabriel Pereira de Oliveira
Pontifícia Universidade Católica de Minas Gerais

Prof. Dra. Mariana de Oliveira Santos Silva

Pontifícia Universidade Católica de Minas Gerais

**Belo Horizonte
2026**

(Assinado digitalmente em 23/06/2026 14:59)
DANIEL HASAN DALIP
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 2312571

(Assinado digitalmente em 23/06/2026 14:31)
DANILO BOECHAT SEUFITELLI
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 1371545

(Assinado digitalmente em 23/06/2026 14:51)
GABRIEL PEREIRA DE OLIVEIRA
ASSINANTE EXTERNO
CPF: ###.###.836-##

(Assinado digitalmente em 23/06/2026 14:21)
MARIANA DE OLIVEIRA SANTOS SILVA
ASSINANTE EXTERNO
CPF: ###.###.716-##

Visualize o documento original em <https://sig.cefetmg.br/public/documentos/index.jsp>
informando seu número: **11**, ano: **2026**, tipo: **FOLHA DE APROVAÇÃO DE TCC**, data de emissão:
19/06/2026 e o código de verificação: **4cf4598ee5**