

RELATÓRIO EXECUTIVO - PROJETO LIGHTHOUSE

Sumário Executivo

Este estudo usa dados históricos do cinema para reduzir incertezas na escolha de **quais filmes produzir**. Analisamos ~45 mil títulos (Kaggle) e a base Lighthouse (~1 mil), construindo protótipos para **prever gêneros pela sinopse (H9)** e **estimar a nota do IMDb (H11)**.

Em negócios, os resultados são consistentes: **orçamento e engajamento** (popularidade, votos) explicam grande parte da **receita bruta**, enquanto **gênero e janela** modulam o risco e a estabilidade.

Em portfólio, **Animation/Family/Adventure sustentam bilheteria** com consistência e alto custo; **Horror/Crime/Mystery entregam ROI** superior por exigir menos investimento, porém com maior volatilidade. A recomendação é **balancear tratores de bilheteria com apostas eficientes de médio/baixo orçamento**, ancoradas por métricas prévias de engajamento.

1. Modelo de negócio do estúdio e alinhamento com dados

Estúdios como a PProductions monetizam por **bilheteria, licenciamento/streaming, VOD, TV e merchandising**.

Os custos se concentram em **produção e marketing/distribuição**, e as decisões de “greenlight” são essencialmente **gestão de portfólio de risco**. Nesse contexto, **dados ajudam em três frentes**: (i) identificar **determinantes de receita e ROI**; (ii) calibrar **janela e posicionamento**; (iii) priorizar **projetos** com base em sinais precoces (engajamento, força de gênero, histórico de talentos).

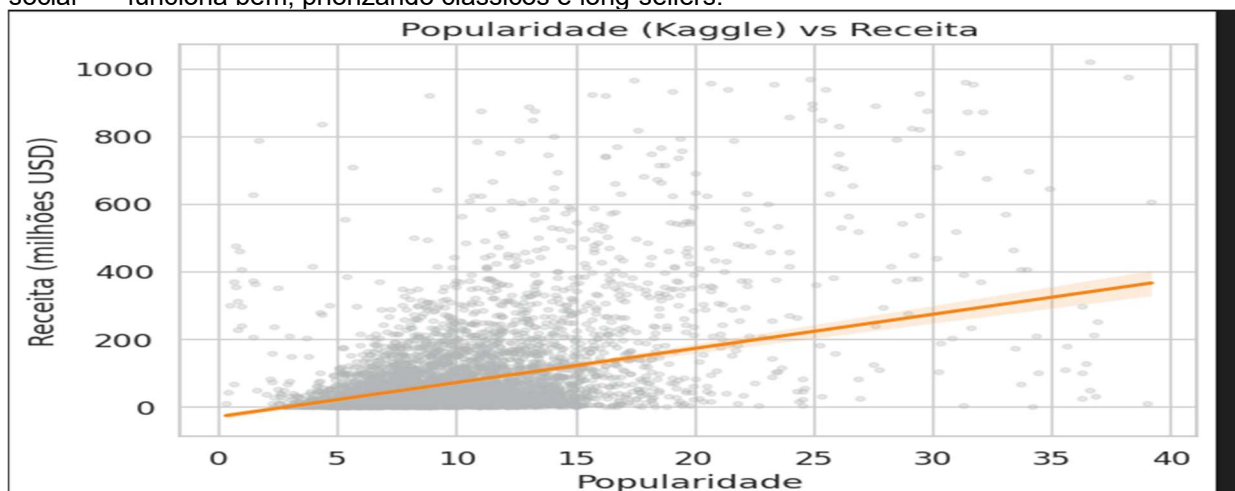
As descobertas deste trabalho se encaixam nessa lógica: orçamento e engajamento elevam a barra de faturamento; gêneros e janela ajustam o **perfil de risco**; texto de sinopse contém **sinal semântico** suficiente para agilizar a **curadoria**.

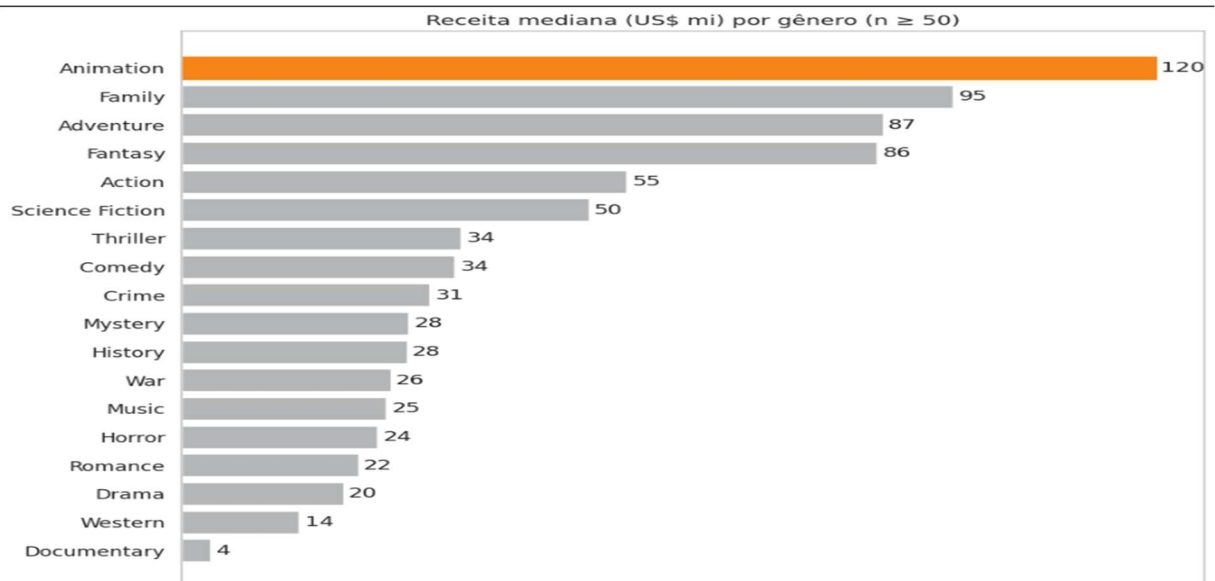
2. O que os dados mostram (EDA → visão de negócios)

A análise confirma que **orçamento** cresce junto com **receita bruta**, mas **não garante ROI**; **popularidade e votos** funcionam como **leading indicators** de performance comercial; **nota do IMDb** favorece longevidade/atração, porém **não assegura bilheteria**. O papel do **gênero** é **nuançado**:

- **Animation/Family/Adventure** sustentam **altas medianas de receita** (apelo global) com **alto custo e ROI moderado**.
- **Horror/Crime/Mystery** tendem a **maior ROI** por custos menores, com **maior variância** de resultado.
- **Idioma e janela** influenciam magnitude (inglês e verão costumam favorecer alcance).

Para recomendação “fria” (sem perfil do usuário), a regra **alta nota + muitos votos** — “consenso social” — funciona bem, priorizando clássicos e long-sellers.





3. Respostas ao Tim de Negócios (síntese)

- A. **Filme recomendado para uma pessoa desconhecida:** Recomendar títulos com **alta nota e muitos votos** (consenso social). Exemplos naturais: *The Shawshank Redemption* (1994) e *Dilwale Dulhania Le Jayenge* (1995).
- B. **Principais fatores ligados à alta expectativa de faturamento:** **Orçamento** (escala), **popularidade/votos** (engajamento antecipado), **janela/sazonalidade** e **gênero/talentos**.

Observação central: **ROI** depende do **equilíbrio de custo**; blockbusters lideram receita absoluta, enquanto gêneros de menor custo (Horror/Crime/Mystery) tendem a melhor retorno proporcional.

- C. **Sobre a coluna Sinopse (overview) é possível inferir gênero? O que ela revela:** Sim. A sinopse carrega **sinal semântico** suficiente para **inferir gêneros**. O protótipo **H9** (TF-IDF + One-vs-Rest(LogReg)) acerta pelo menos um gênero relevante em **~57%** dos casos (F1-micro ~0,57), sendo adequado para **pré-tagueamento** com revisão editorial

Sinopse: A group of teenagers spend the night in a haunted house where strange events begin to unfold.
Gêneros previstos (prob.):
- Horror: 0.914
- Thriller: 0.536

Sinopse: A heartfelt story about a family overcoming challenges during a road trip across the country.
Gêneros previstos (prob.):
- Drama: 0.758
- Comedy: 0.494

Sinopse: A young wizard begins his journey at a school of magic
Gêneros previstos (prob.):
- Family: 0.503
- Fantasy: 0.474
- Drama: 0.451

Top termos - Drama
['drama', 'true story', 'affair', 'prostitute', 'relationship', 'lives', 'love', 'mother', 'friendship', 's'

Top termos - Comedy
['comedy', 'hilarious', 'comic', 'comedic', 'bumbling', 'funny', 'comedian', 'spoof', 'wedding', 'satire',

Top termos - Horror
['horror', 'vampire', 'zombie', 'blood', 'zombies', 'evil', 'terrifying', 'vampires', 'supernatural', 'kill









Top termos - Action
['assassin', 'cop', 'action', 'martial', 'fight', 'cia', 'warrior', 'mercenary', 'agent', 'yakuza', 'ruthle

Top termos - Documentary
['documentary', 'interviews', 'filmmaker', 'footage', 'look', 'history', 'portrait', 'film', 'journey', 'd

- D. **Como prever a nota do IMDb:** Variáveis, tipo de problema, modelo, métricas
Problema de **regressão** (alvo contínuo 0–10). Usamos **numéricos** (ex.: log_budget,

log_revenue, roi, runtime, popularity, vote_count...), **categóricos** com *top-K* (original_language, genre_primary, director, star1..4) e **texto** (overview via TF-IDF). O melhor equilíbrio veio com **HistGradientBoosting/Random Forest**, medidos por **RMSE/MAE/R²**.

- E. **Nota do IMDB para o exemplo (Shawshank)**: O modelo retornou $\approx 6,3$ versus **8,5** no dataset. A diferença está alinhada à **regressão à média** e à ausência de alguns preditores causais (premiações, telas, franquia, marketing).

H	Status	Resumo
0	H1  parcial	Orçamento ↑ = Receita ↑; ROI não garantido.
1	H2 	Popularidade antecipa bilheteria (bom leading indicator).
2	H3 	Nota IMDb ↑ ajuda longevidade/atração, mas não garante bilheteria.
3	H4  fraco/indefinido	Runtime alto (>150m) com evidência não robusta sobre nota.
4	H5  com nuances	Animation/Family/Adventure sustentam receita/consistência (alto orçamento/ROI moderado); Horror/Crime/Mystery tendem a maior ROI (mais voláteis). Escolha depende de receita × ROI × prestígio.
5	H6  parcial	Diretor/Elenco ajudam receita; ROI depende do custo.
6	H7 	Recomendação p/ desconhecido: alta nota + muitos votos (ex.: DDLJ (1995), Shawshank (1994)).
7	H8 	Pós-2010: engajamento ↑, mediana de receita ↓ (fragmentação/streaming).

4. Recomendações de portfólio

Sustentar bilheteria com 1–2 títulos **Animation/Family/Adventure** por ciclo, apoiados por campanhas robustas. **Equilibrar ROI** adicionando pelo menos 1 projeto **Horror/Crime/Mystery** de médio/baixo orçamento — volatilidade é maior, mas o risco financeiro é menor e o upside, relevante.

Janela: priorizar **verão** onde fizer sentido; **idioma inglês** facilita alcance internacional.

Engajamento pré-lançamento (busca, social, trailers) deve orientar investimento incremental em marketing, dada sua correlação com performance.

5. Riscos e Limitações

As relações identificadas são **associativas**; não implicam causalidade. Há **lacunas e zeros** em dados financeiros (especialmente budget/revenue) e **viés de fonte** (Kaggle). Categorias de alta cardinalidade foram reduzidas por *top-K*, perdendo granularidade em nomes específicos. O poder do texto de sinopse depende do vocabulário; mudanças de **gosto** e **janela de consumo** (pós-2010, streaming) geram **deriva temporal** e pedem recalibração periódica.

6. Próximos Passos

H9 (gêneros): calibrar *thresholds* por classe, reforçar balanceamento e avaliar **embeddings** (BERT); meta de **+3–5 p.p. F1-micro**.

H11 (nota IMDb): ampliar representação textual (TF-IDF mais profundo ou embeddings), criar decade e sinal de **franquia/continuação**, refinar *top-K* de director/stars e testar **blending**; meta de **reduzir RMSE $\geq 20\%$ e elevar R² em +0,03–0,05**.

Negócio: medir **uplift** por **mix gênero × orçamento × janela × elenco** e incorporar **sinais de marketing e premiações** para decisões de greenlight.

7. Anexo e Artefatos

- **Modelos**: models/h9_multilabel_pipeline.joblib, models/h11_imdb_rating_model.pkl.
- **Notebook final**: 00_final_movie_analytics.ipynb (síntese de ponta a ponta).
- **Relatórios/figuras**: reports/ (inclui hypotheses_summary.csv).