RELATÓRIO FINAL – CASE INDICIUM

Autor: Leonardo Nassib Fonseca
Projeto: Análise e Modelagem de Dados Cinematográficos
Objetivo: Apoiar PProductions na escolha do próximo filme a ser produzido

1. Resumo Executivo

Este projeto teve como objetivo analisar um conjunto de dados cinematográficos para fornecer ao estúdio PProductions uma base analítica que apoie a decisão sobre qual filme produzir. Foram conduzidas análises exploratórias, engenharia e preparação de dados, além de modelagem preditiva para estimar a nota IMDB a partir das características dos filmes.

Apesar de o modelo não ter atingido desempenho preditivo satisfatório, o projeto gerou insights relevantes sobre padrões de notas, faturamento e elementos associados ao desempenho crítico e comercial dos filmes.

2. Problema de Negócio

A PProductions deseja identificar o tipo de filme que apresenta maior potencial de sucesso, tanto em avaliação crítica quanto em atração de público, antes de investir em uma nova produção. A empresa busca respostas para perguntas centrais: quais gêneros tendem a gerar maior faturamento? Quais características influenciam as notas? Quais perfis de direção e elenco são mais promissores?

3. Dados Utilizados

O dataset utilizado contém informações estruturadas sobre filmes, incluindo:

- Ano de lançamento
- Diretor e elenco principal

- Certificação indicativa
- Gêneros
- Duração
- Nota IMDB
- Número de votos
- Faturamento
- Metascore

Trata-se de uma base abrangente em atributos descritivos, mas limitada em profundidade. O dataset não inclui informações críticas para modelagem robusta, como orçamento, investimento em marketing, histórico de desempenho dos diretores/atores, contexto de lançamento ou indicadores de popularidade.

Esse caráter restrito faz com que o conjunto seja adequado para análises exploratórias e modelagens básicas, mas insuficiente para capturar plenamente a complexidade do desempenho real de um filme no mercado.

4. Estratégia de Solução

A solução foi estruturada em seis etapas principais:

1. Descrição dos dados: Identificação da estrutura da base, tipos de variáveis, valores ausentes, duplicidades e aderência ao problema de negócio.
2. Engenharia de variáveis: Criação e transformação de variáveis para aumentar o poder explicativo, incluindo recodificações, derivação de atributos e padronizações.
3. Filtragem das variáveis: Remoção de atributos redundantes, não disponíveis em produção ou que não agregam valor ao modelo.
4. Análise exploratória: Investigação de distribuições, correlações, tendências temporais e relações entre variáveis, validando hipóteses e orientando a modelagem.
5. Preparação dos dados: Aplicação de codificação de variáveis categóricas, escalonamento robusto, tratamento de valores ausentes e separação em treino e teste.
6. Modelagem e avaliação: Testes com diversos algoritmos de regressão, comparação de métricas e escolha do modelo com melhor equilíbrio entre desempenho e robustez.

5. Principais Insights

1. Ano de lançamento vs. Nota IMDb

Não foram encontradas evidências de que filmes recentes recebam notas maiores. Observou-se leve tendência de maior avaliação para filmes antigos, indicando que obras clássicas tendem a se consolidar como referências críticas ao longo do tempo (figura 1).

2. Fama do elenco vs. Faturamento

Contrariando o senso comum, filmes com elencos menos famosos apresentaram medianas de faturamento superiores aos filmes com atores amplamente reconhecidos. Isso indica que o sucesso comercial depende de um conjunto de fatores mais amplo, gênero, marketing, direção e apelo narrativo, não apenas da popularidade do elenco (figura 2).

6. Resultados da Modelagem

Diversos algoritmos foram avaliados, incluindo Regressão Linear, Gradient Boosting e Random Forest. Após testes e ajustes, o **RandomForestRegressor** apresentou o melhor desempenho:

- **RMSE ≈ 0,27** (erro pontual baixo na escala 0–10)
- **R^2 ≈ 0,04** (explica apenas 4% da variabilidade real)

Embora o erro absoluto seja baixo, o coeficiente de determinação mostra que o modelo não consegue capturar relações estruturais relevantes entre as variáveis explicativas e a nota IMDb.

A análise dos resíduos revelou distribuição aleatória, sem padrões identificáveis, indicando ausência de aprendizado significativo. Esse resultado reforça que a limitação não está no algoritmo, mas na **ausência de dados profundos e historicamente relevantes no dataset**.

7. Conclusões

O estudo permitiu identificar os elementos que tornam um filme mais promissor tanto em avaliação crítica quanto em potencial de faturamento, atendendo ao objetivo do estúdio de selecionar o próximo projeto.
As análises mostram que filmes de gêneros **Família, Ação e Aventura**, preferencialmente combinados com **Comédia ou Ficção Científica**, apresentam maiores medianas de arrecadação. Do ponto de vista crítico, produções com diretores de histórico consolidado tendem a alcançar melhores notas. Além disso, classificações indicativas mais amplas (Livre e 12+) aumentam o público potencial, contribuindo diretamente para o desempenho financeiro.

Embora o modelo preditivo tenha apresentado baixo poder explicativo devido à limitação dos dados, os padrões observados permitem orientar a estratégia do estúdio. Assim, recomenda-se priorizar a produção de filmes que combinem **gêneros de alta arrecadação**, **direção de reputação comprovada**, **classificação indicativa ampla** e **elenco com bom histórico comercial**, ainda que não necessariamente composto por celebridades de primeira linha.

Portanto, o problema de negócio é respondido ao indicar que o estúdio deve direcionar seus esforços para projetos que maximizem alcance, apelo familiar e histórico comercial consolidado, aumentando assim a probabilidade de sucesso crítico e financeiro.

8. Recomendações

Para aumentar a precisão dos modelos e ampliar a qualidade das análises, recomenda-se:

1. Enriquecimento da base de dados

- Histórico de faturamento do diretor e dos principais atores
- Orçamento detalhado e investimento em marketing
- Engajamento em redes sociais e buscas online
- Participação em festivais e premiações
- Tendências de bilheteria por região

2. Melhor tratamento de variáveis categóricas

Atores e diretores possuem alta cardinalidade, e técnicas como One-Hot Encoding se tornam ineficientes. Recomenda-se:

- Uso de embeddings
- Frequency Encoding aprimorado
- Indicadores históricos (ex.: média de faturamento por diretor)

3. Processo contínuo de atualização

A base deve ser enriquecida e revisada periodicamente, garantindo dados atualizados, padronizados e alinhados às necessidades de decisão da empresa.

9. Limitações

As principais limitações identificadas foram:

- Variáveis essencialmente descritivas e ausência de fatores subjetivos fundamentais
- Falta de contexto histórico sobre diretores, atores e franquias
- Alta cardinalidade em variáveis categóricas
- Baixo poder explicativo dos dados, refletido no R^2 reduzido
- Dependência de atributos que não representam o que realmente impulsiona notas e faturamento

Essas limitações restringem o desempenho dos modelos e a profundidade das conclusões.

Anexos – Gráficos

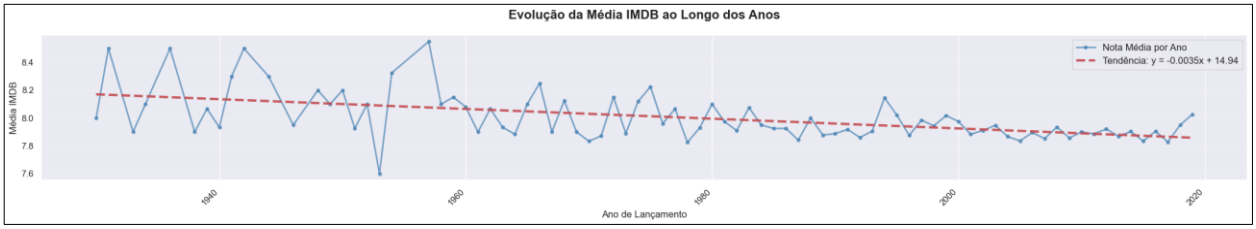

Figura 1 – Média IMDB ao longo dos anos.

O gráfico mostra que, apesar da variação anual das notas, existe uma tendência de queda gradual na avaliação média dos filmes ao longo das décadas. Isso sugere que filmes antigos são mais bem avaliados em média, enquanto produções recentes apresentam notas ligeiramente menores.

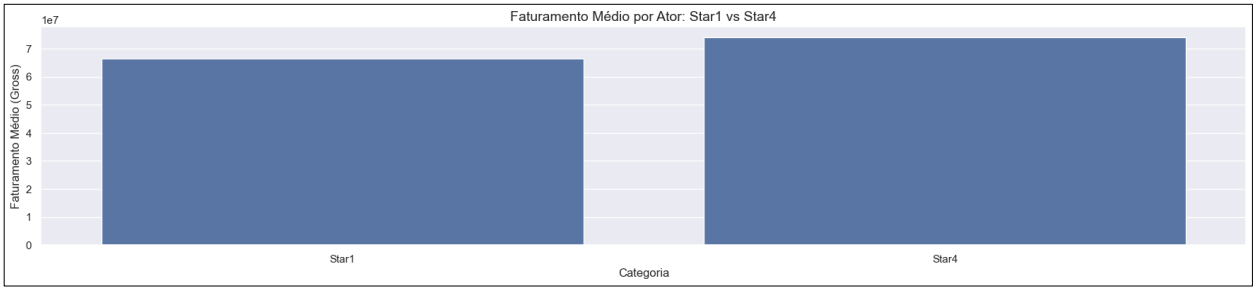

Figura 2 – Faturamento médio por tipo de elenco.

Filmes com atores menos famosos (Star4) têm faturamento médio semelhante ou até superior aos filmes com atores mais conhecidos (Star1). Isso indica que a presença de grandes estrelas, por si só, não garante maior retorno financeiro, e que o sucesso comercial depende de fatores mais amplos do que apenas o elenco.

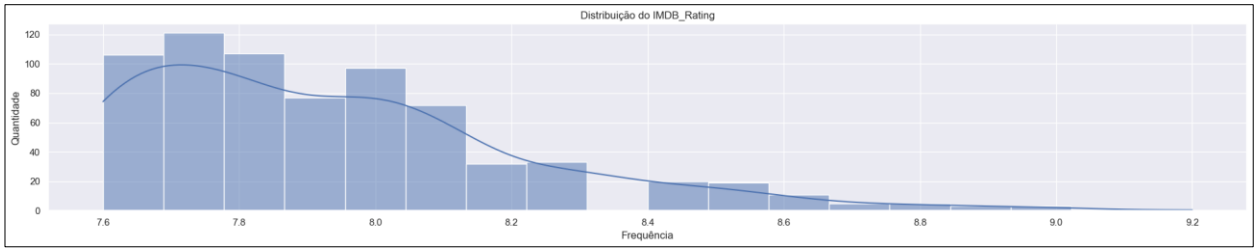

Figura 3 – Distribuição IMDB.

A distribuição da variável IMDB é assimétrica à direita (assimetria positiva), ou seja, os dados estão mais concentrados no lado esquerdo da escala. Observa-se que a maioria dos filmes possui avaliação entre 7.6 e 8.1, enquanto poucos alcançam notas mais elevadas, acima de 9.0. A partir desse gráfico, é possível inferir que a variabilidade das notas é relativamente baixa, o que pode levar o modelo a produzir previsões próximas à média.

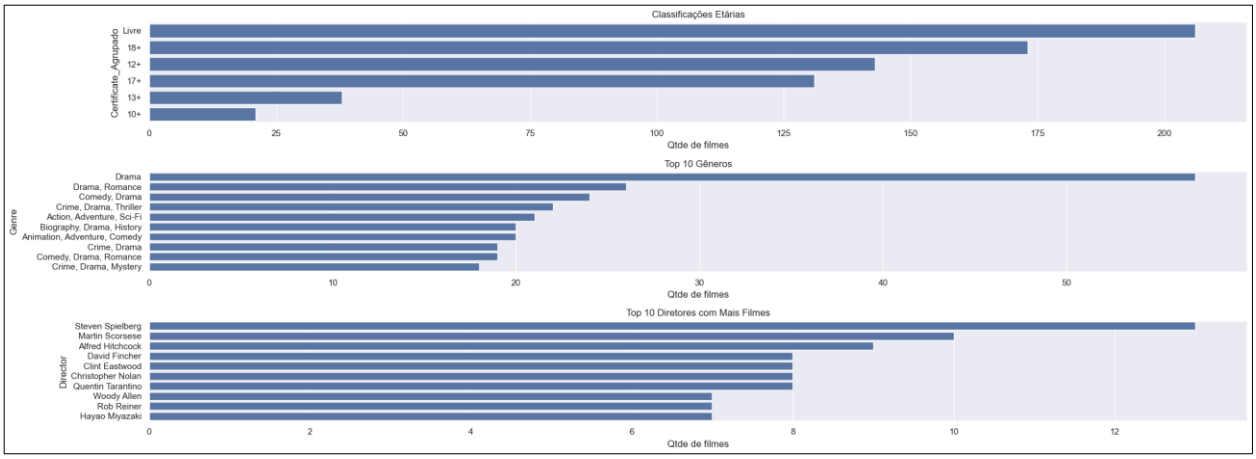

Figura 4 – Classificação etária, gêneros e diretores por qtde de filmes.

A classificação etária livre seguida de 18+ e 12+ concentram a maior parte dos filmes, ou seja, uma classificação mais voltada para a família de um modo geral, a respeito do gênero, drama é disparado o mais comum, sobre o diretor, o nome Steven Spielberg lidera com 13 filmes, sugerindo que diretores renomados possivelmente tem maior impacto em notas ou bilheterias.

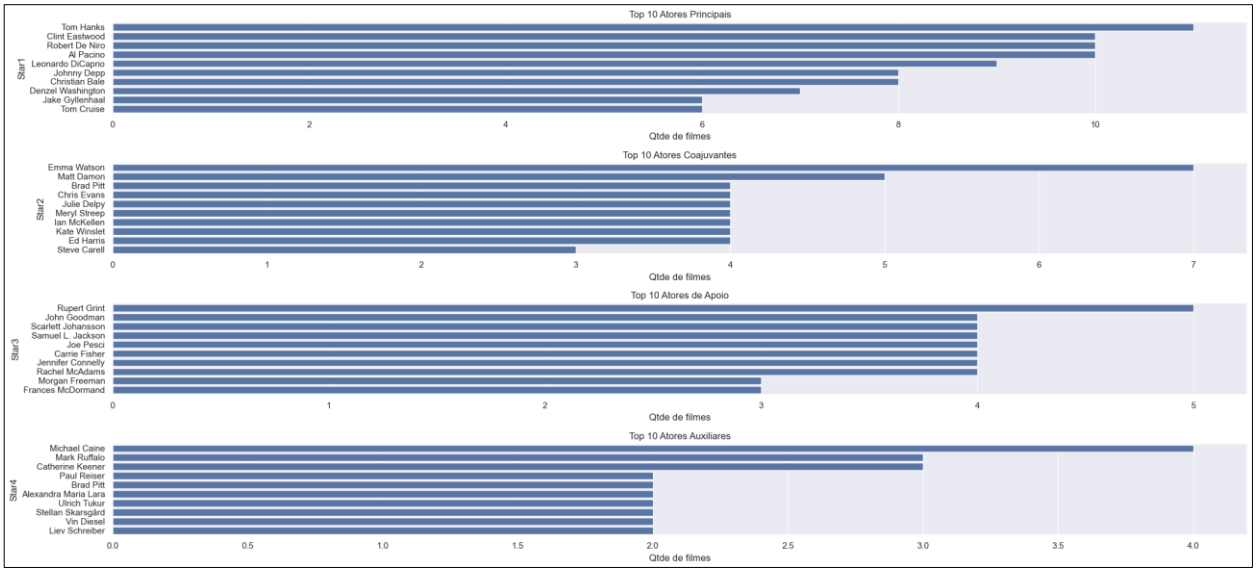

Figura 5 – Atores protagonistas, coadjuvantes, secundários e auxiliares por qtde de filmes.

Entre os protagonistas, os atores masculinos predominam, com Tom Hanks (11), Clint Eastwood (10), Robert De Niro (10) e Al Pacino (10) aparecendo com maior frequência. Já entre os atores coadjuvantes, destaca-se a atriz Emma Watson (7), que lidera com expressiva vantagem. Nos papéis de apoio e auxiliares, as diferenças são menos acentuadas: Rupert Grint (5) aparece mais frequentemente, enquanto Michael Caine (4) lidera.

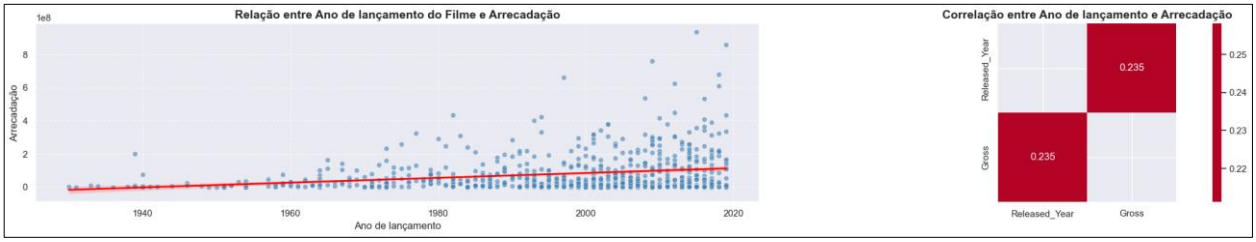

Figura 6 – Relação entre Ano de lançamento do filme e arrecadação.

Existe uma correlação positiva (0,235) considerada fraca, ou seja, filmes mais recentes tendem a arrecadar um pouco mais, mas somente esta variável não é um fator relevante.

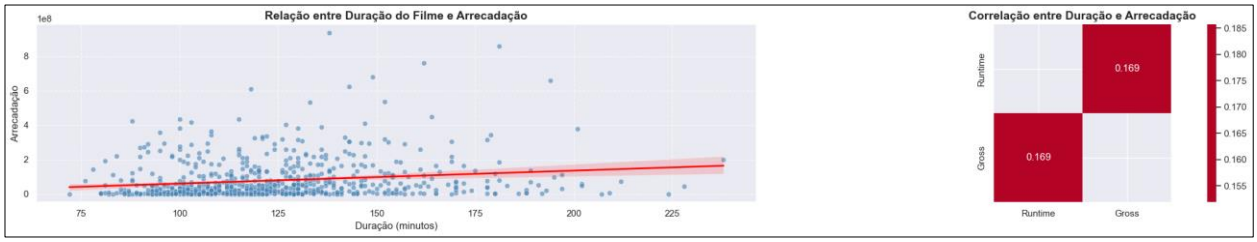

Figura 7 – Relação entre Duração do filme e arrecadação.

Existe uma correlação positiva (0,169) considerada fraca, ou seja, filmes mais longos tendem a arrecadar levemente mais, mas a duração não é um fator relevante isoladamente.

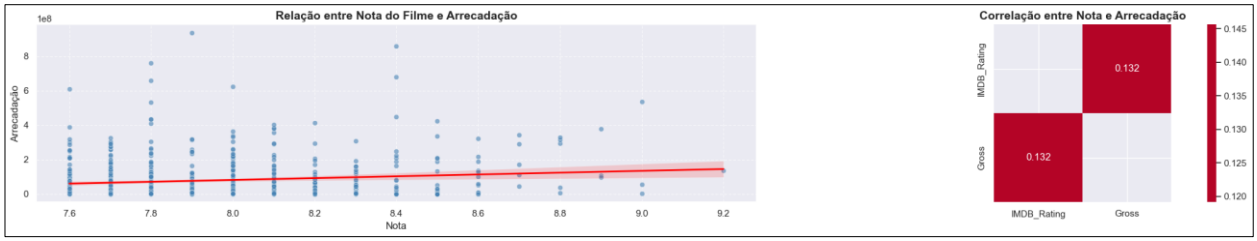

Figura 8 – Relação entre Nota do filme e arrecadação.

A relação entre nota do filme e arrecadação existe, mas é positiva (0,132) fraca, isso significa que filmes com avaliações mais altas tendem a arrecadar levemente mais, contudo não garante sucesso na arrecadação.

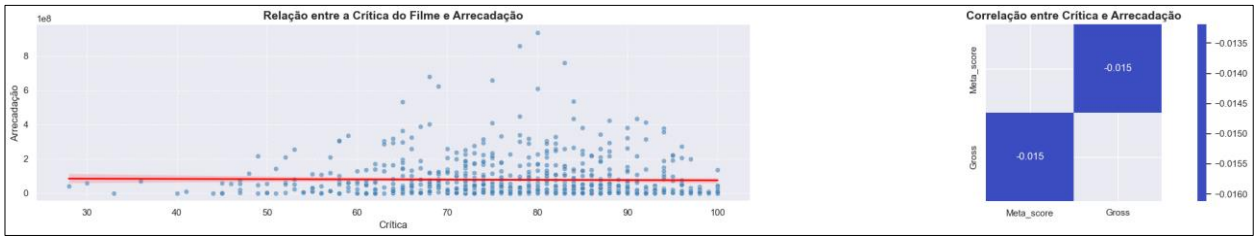

Figura 9 – Relação entre Crítica do filme e arrecadação.

A relação entre crítica e arrecadação é praticamente inexistente (-0,015), isso evidência que a avaliação da crítica não é um fator que influencia no sucesso de bilheteria.

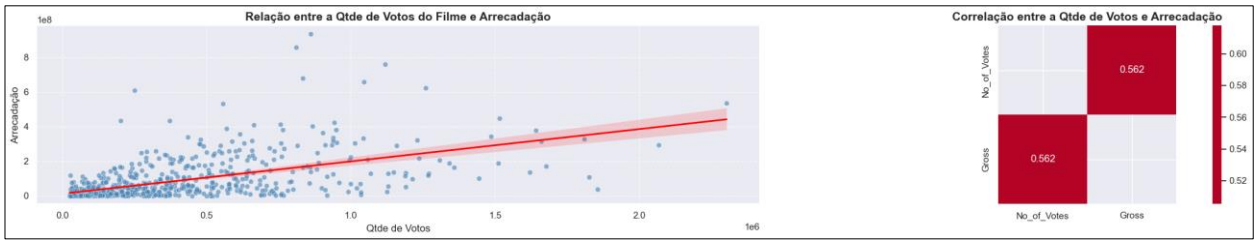

Figura 10 – Relação entre Qtde de votos do filme e arrecadação.

A quantidade de votos apresenta a correlação mais alta com a arrecadação (0,562), mostrando uma relação moderada a forte, em outras palavras, filmes com maior bilheteria tendem a receber mais votos, refletindo maior alcance e engajamento do público.

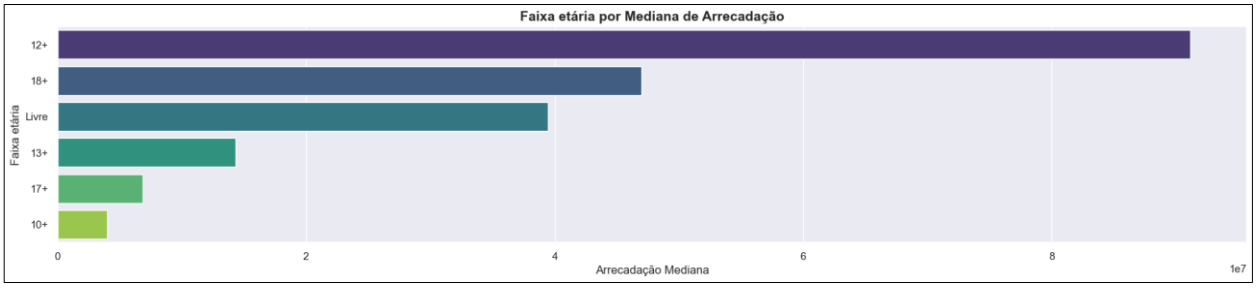

Figura 11 – Distribuição por Faixa etária e mediana de arrecadação.

Filmes das faixas etárias 12+, 18+ e livre respectivamente tendem a ter mediana de arrecadação maior.

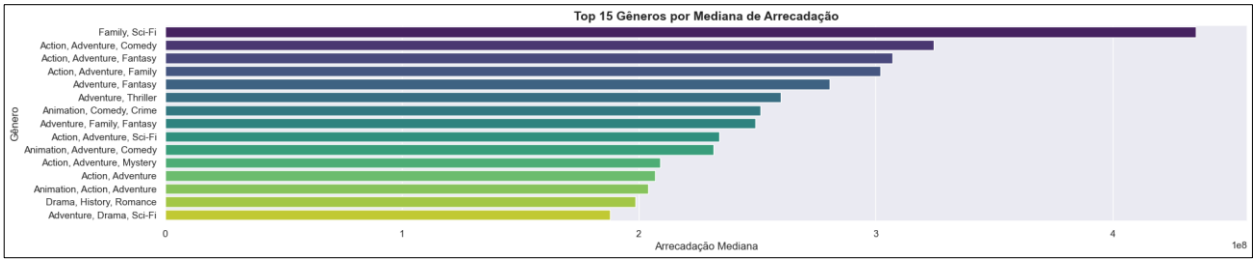

Figura 12 – Distribuição por Gêneros dos filmes e mediana de arrecadação.

Filmes do gênero familiar, ação e aventura, especialmente quando combinados com comédia ou ficção científica, tendem a ter a maior mediana de arrecadação.

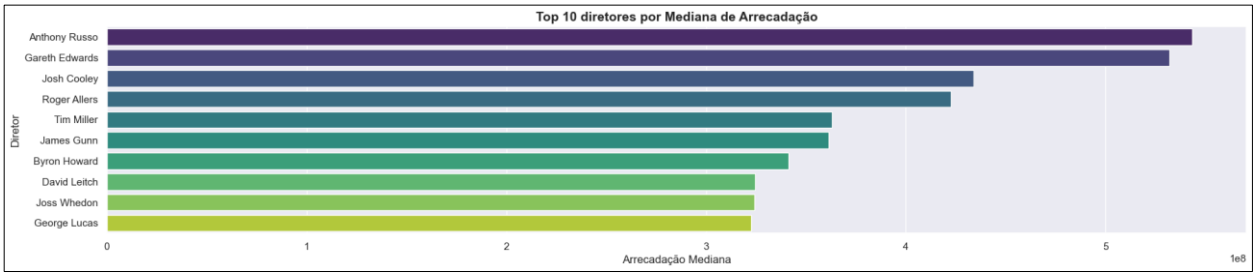

Figura 13 – Distribuição por Diretores e mediana de arrecadação.

Antony Russo seguido de Gareth Edwards são os diretores respectivamente com tendência das maiores medianas de arrecadação, significativamente à frente de todos os outros na lista.

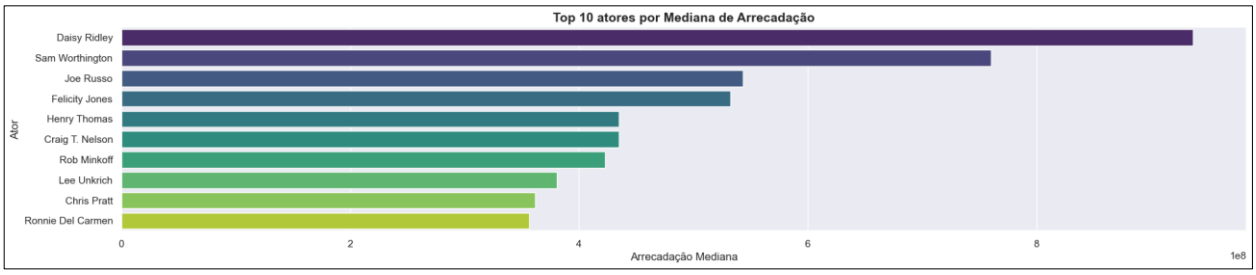

Figura 14 – Distribuição por Atores e mediana de arrecadação.

Daisy Ridley e Sam Wothington são os atores respectivamente com tendência das maiores medianas de arrecadação, bem à frente de todos os outros na lista.