Desafio Cientista de Dados

Introdução

Este relatório tem como finalidade indicar à PProductions qual o melhor filme a ser produzido, analisando um histórico de filmes produzidos, para chegarmos a conclusão de qual o melhor Gênero a ser produzido, qual o melhor diretor a ser contratado para uma maior expectativa de sucesso.

Análise Exploratória de Dados (EDA)

1. Distribuição das notas do público (IMDb)

As notas se distribuem de forma assimétrica, com dados se acumulando à esquerda. O que mostra que esses dados não estão dentro da distribuição normal. Podemos perceber que as notas variam de 7.6 a 9.2, com a maioria dos filmes com notas em torno de 7.7 a 8.2, como podemos observar no gráfico 1.

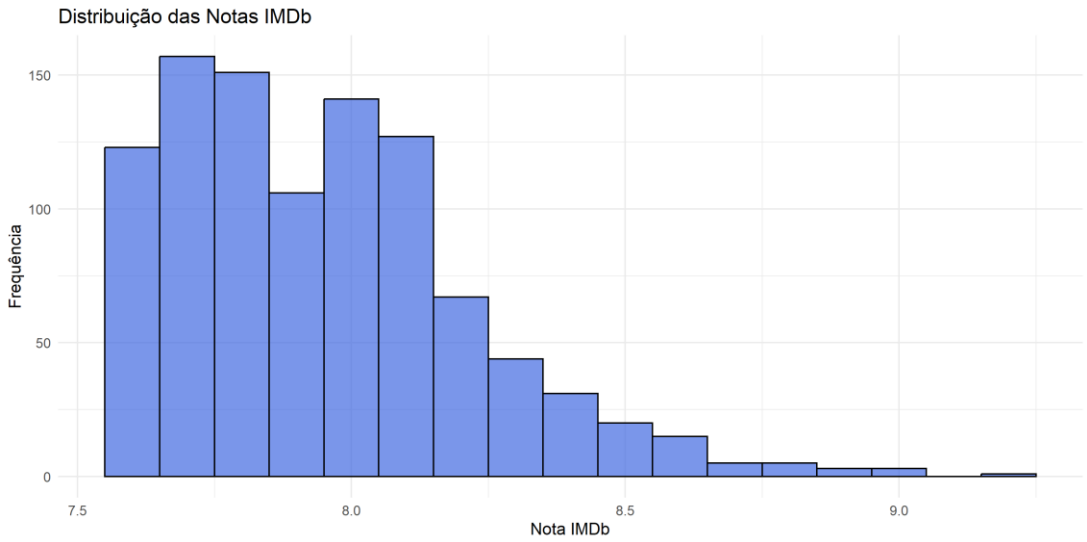


Gráfico 1.

1.1 Distribuição das notas da crítica (Meta score)

As notas se distribuem de forma assimétrica à direita, com a maioria das notas entre o 65 e 85. Os não se enquadram dentro da distribuição normal como podemos observar no gráfico 2.

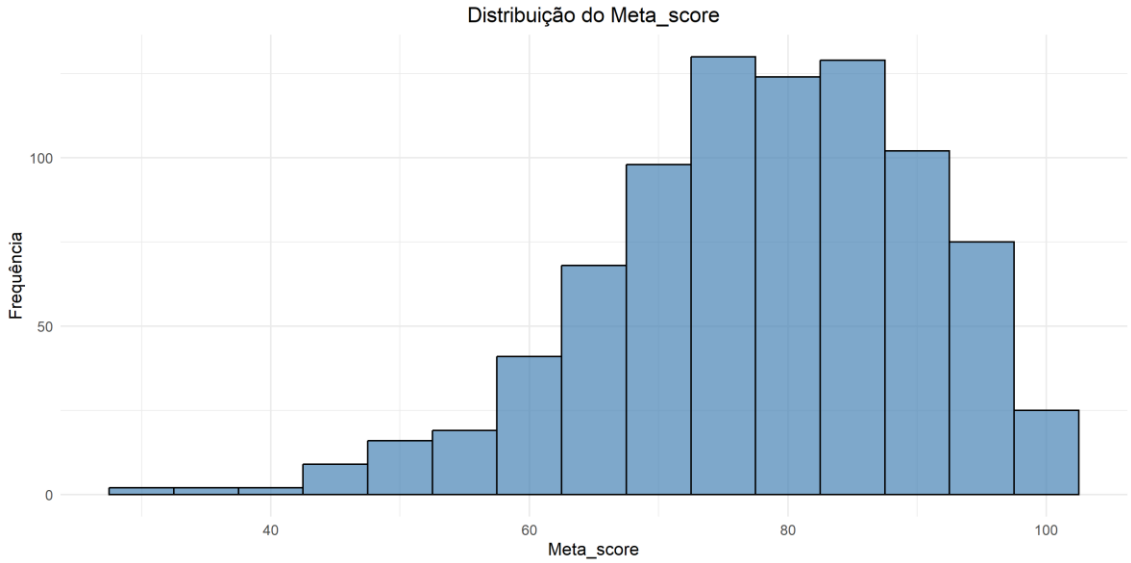


Gráfico 2.

1.2 Distribuição do faturamento bruto (Gross)

A distribuição do faturamento se dá de forma muito assimétrica, a grande maioria dos filmes arrecadam valores menores, apenas poucos filmes demonstram valores extremamente altos, podendo indicar presença de outliers. Filmes de grande faturamento como Star Wars, Vingadores: Ultimato e Avatar fazem com que a média de faturamento suba, como podemos observar no gráfico 3.

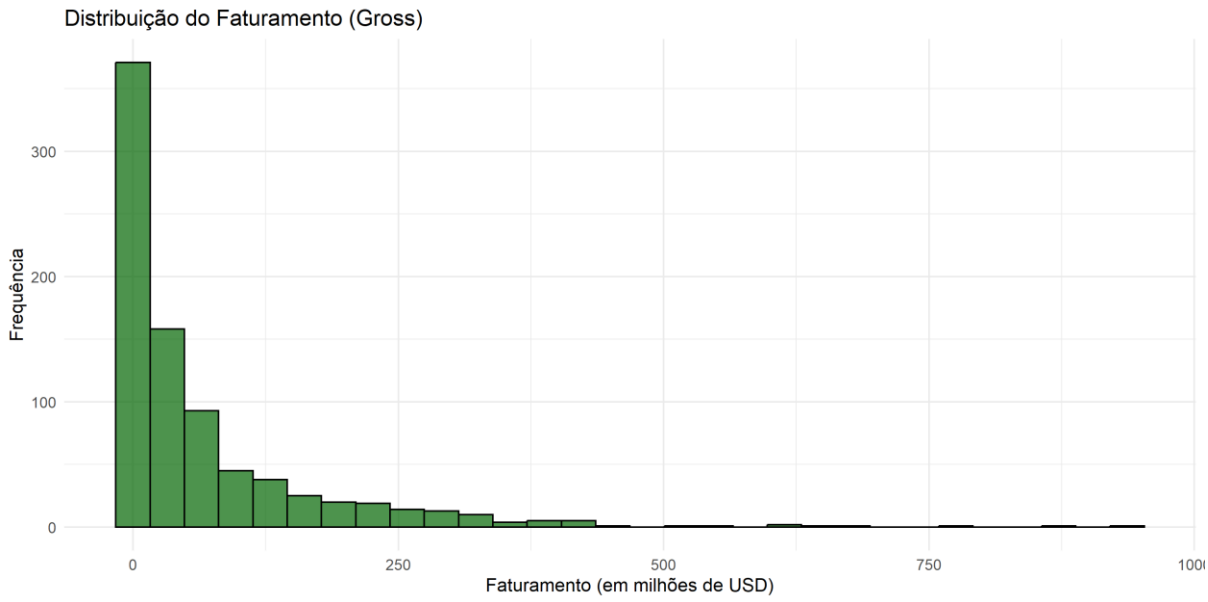


Gráfico 3.

1.3 Diretores com mais filmes

Alguns diretores estão muito presentes no conjunto, e vários deles possuindo filmes com ampla aceitação de público. O que pode trazer um indicativo de que o diretor é um fator importante para a aceitação do público e da crítica especializada. No gráfico 4 podemos observar os 10 diretores com mais filmes no conjunto de dados.

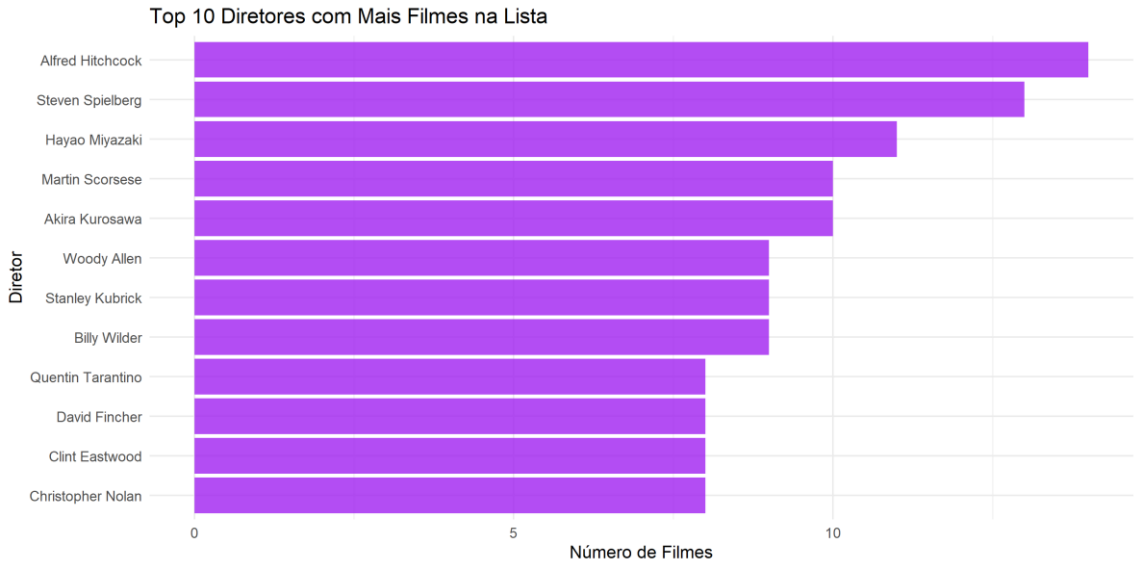


Gráfico 4.

1.4 Gêneros mais comuns

Filmes de drama aparecem com a maior frequência no conjunto de dados. Comédia, crime e aventura aparecem logo em seguida. A grande maioria dos filmes mostrados pelo conjunto possui mais de um gênero. Como pode ser observado no gráfico 5.

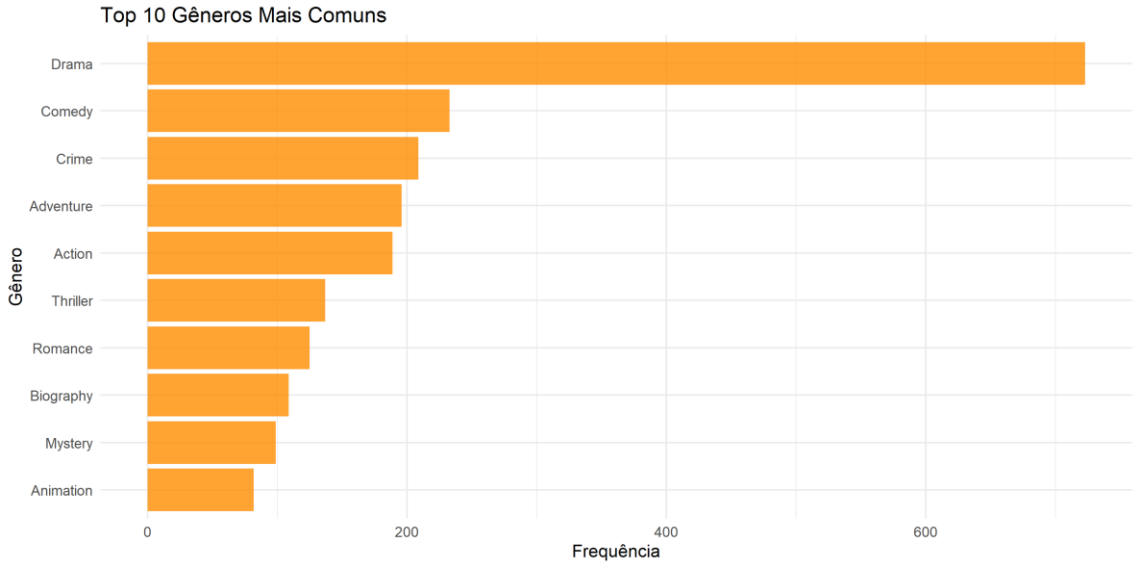


Gráfico 5.

Recomendação de Filme

Baseado nos dados, A recomendação seria O Poderoso Chefão, pois se trata de um filme extremamente popular, baseado no número de votos, na alta aceitação de público, com uma nota IMDb de 9.2, além da aceitação da crítica, com uma nota 100.

2.1. A coluna Overview

A coluna Overview nos mostra a sinopse do filme. Dessa coluna podemos tirar várias informações sobre a estrutura do filme, qual a história em foco, e uma ideia prévia do roteiro, com isso podemos identificar palavras-chave para e correlacionar com a aceitação do filme pelo público. Podemos também fazer uma avaliação de palavras de tom negativo e positivo para também correlacionar com a aceitação do filme, com a hipótese de que filmes negativos ou extremamente dramáticos teriam menor popularidade.

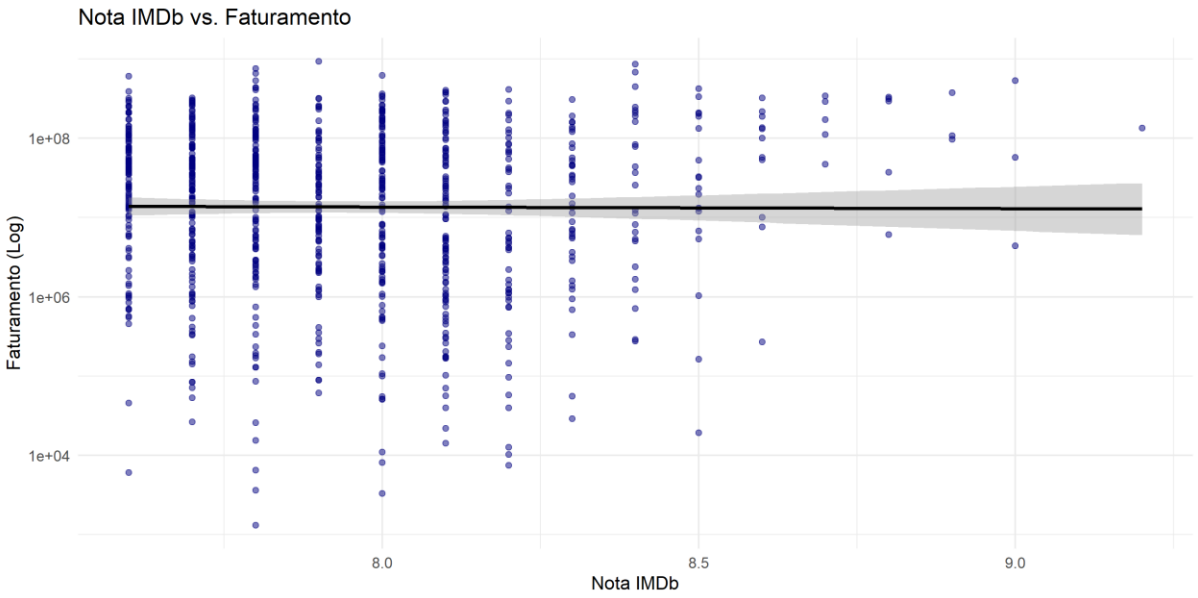
2.2. Inferindo o gênero de um Filme

Através das palavras-chave fornecidas no overview podemos prever o gênero de um filme, podemos utilizar modelos de classificação multi-rótulo, como uma regressão logística multinominal, rede neurais simples ou modelos bayesianos.

Análise Estatística e Desenvolvimento de Hipóteses

3. Há relação entre a aceitação do público (Nota do IMDb) e o faturamento bruto (Gross) ?

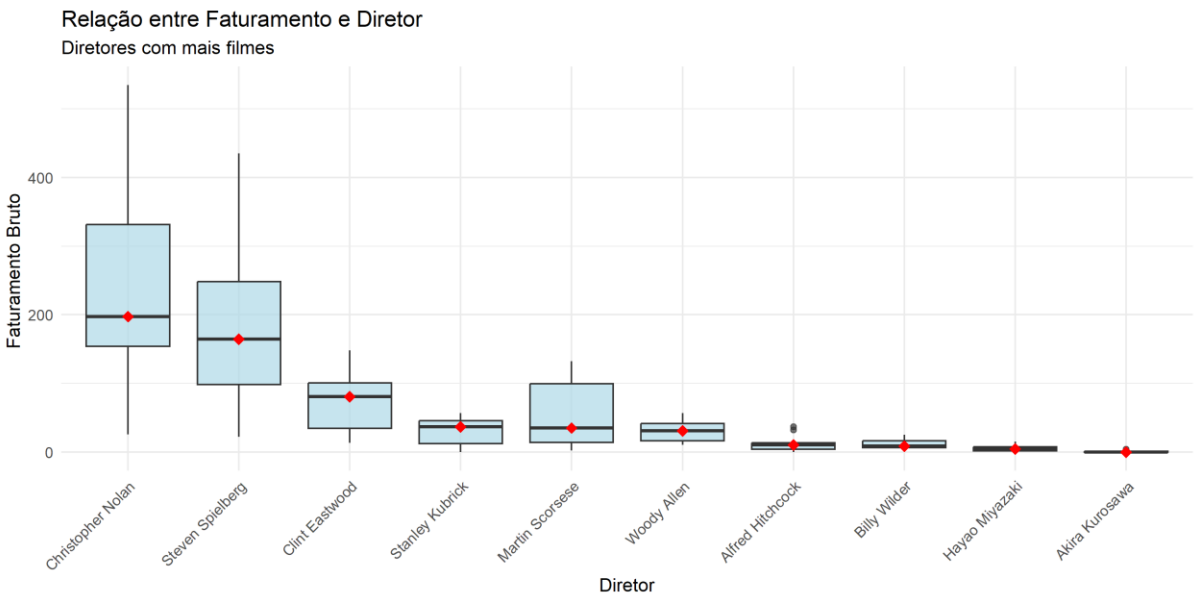
A aceitação do Público pode ser um fator importante para podermos prever o faturamento de um filme. Pode existir a tendencia de um filme que agrade ao grande público seja rentável.



Existe uma correlação significativa entre a nota do público e o faturamento ($p < 0.0001$), porém essa correlação é fraca (cor = 0.099). Existe também uma relação linear entre as variáveis. Existe a tendencia de que o faturamento aumente de acordo com a nota ($p<0.05$), no entanto, essa variável sozinha não é um bom indicador do faturamento, dado o baixo valor de R^2.

3.1 Há relação entre o Faturamento (Gross) e o Diretor?

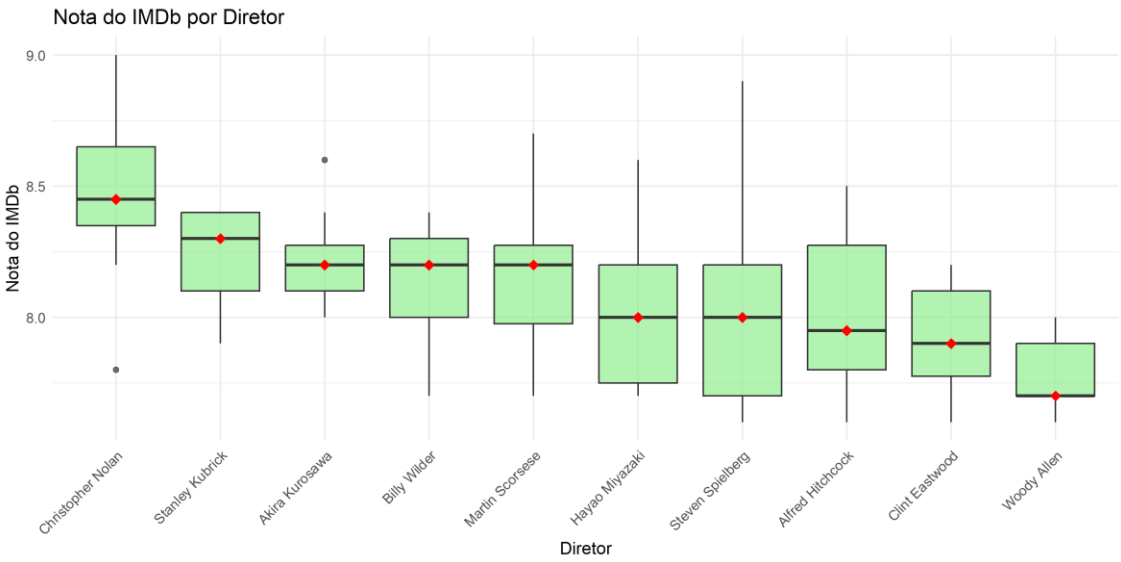
Observamos que há diretores que aparecem com maior frequência no conjunto de dados, isso pode ser um indicativo de que os diretores mais populares produzem com frequência filmes que são sucesso de público, rendendo uma margem de faturamento maior.



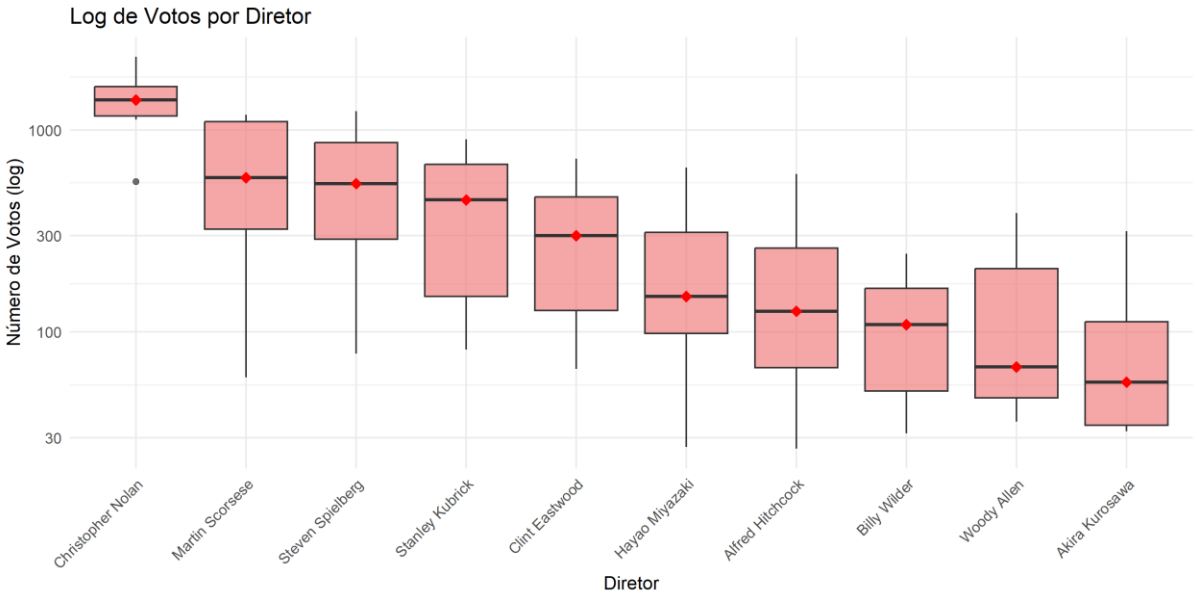
Existe uma diferença significativa no faturamento médio ($p<0.001$), mostrando que a escolha do diretor influencia diretamente no faturamento bruto.

3.2 Há relação entre a aceitação do público (nota IMDB e Número de votos) e o Diretor do filme?

Diretores mais populares produzem com frequência filmes que são sucesso de público, rendendo uma margem de faturamento maior.

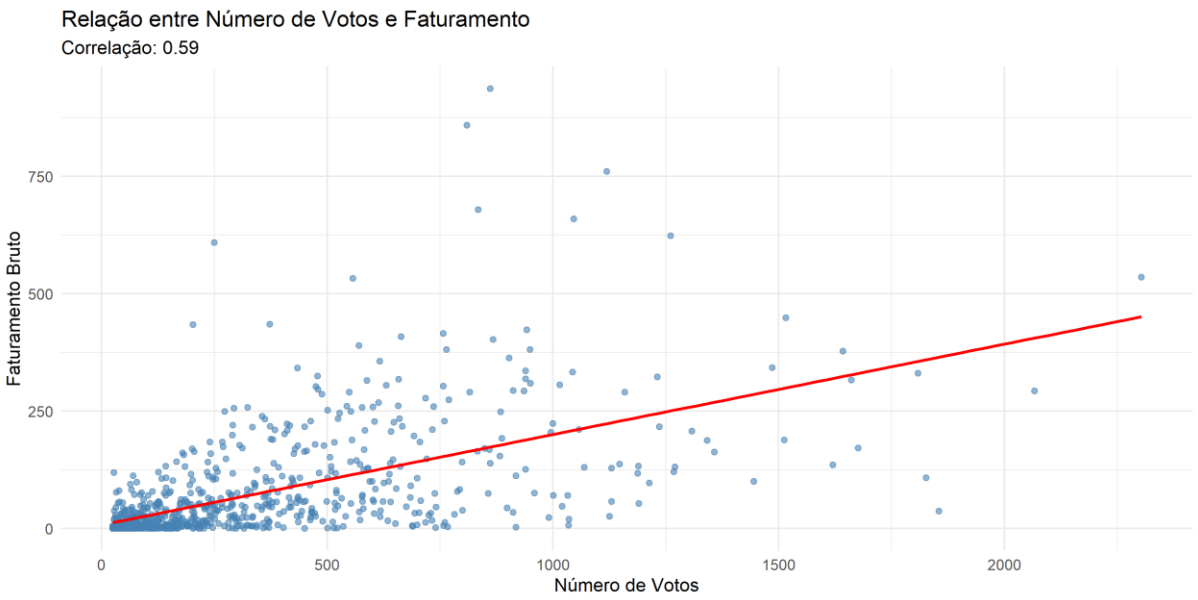


A escolha do diretor influencia significativamente ($p<0.05$) na nota do IMDb, mostrando que a escolha do diretor também influencia na aceitação do público.



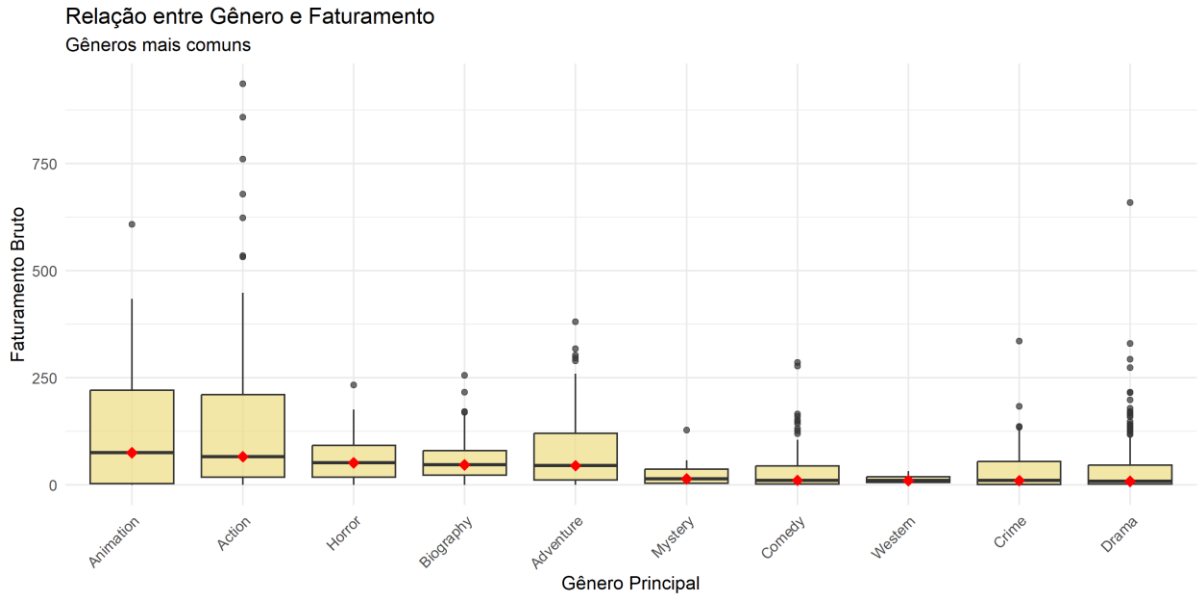
existe também a relação entre escolha do diretor e número de votos ($p>0.001$), mostrando que a escolha do diretor não impacta somente a aceitação do público, mas também o número de pessoas que assistem aos filmes.

3.3 Há relação entre o número de votos e o Faturamento?



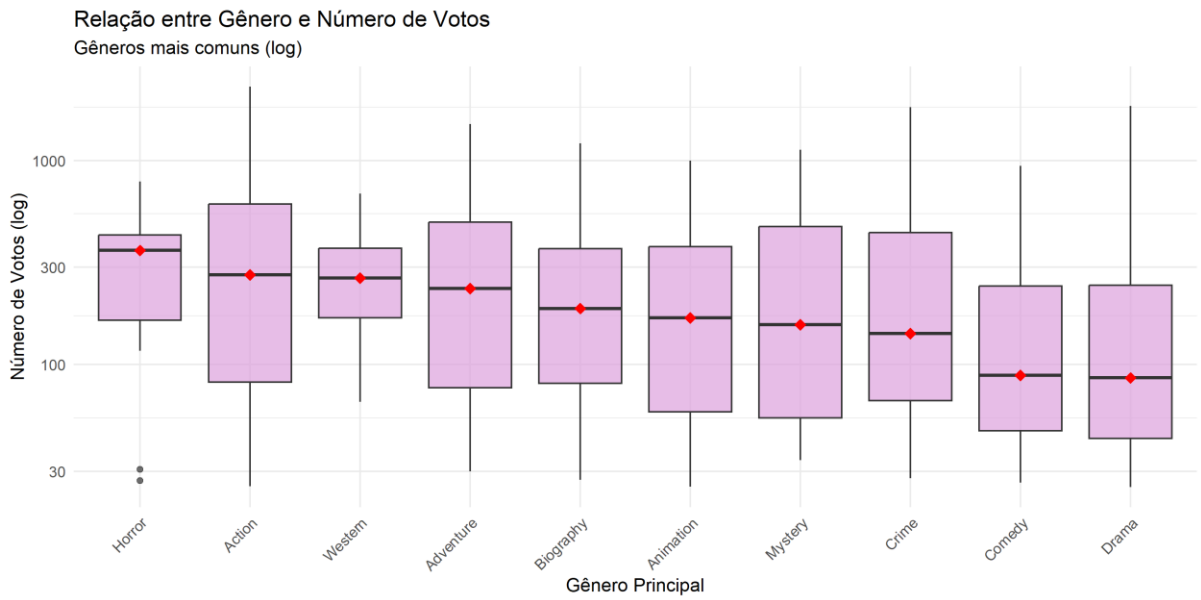
Existe uma correlação significativa (P<0.0001) relativamente forte (cor = 0.51) entre os dados, mostrando que filmes com mais votos tem tendência a ter maior faturamento bruto, com uma relação linear significativa (P<0.0001)

3.4. Há relação entre o gênero e o Faturamento?



A análise mostra que existem diferenças significativas entre o faturamento médio por gênero produzido (p < 0.001). Filmes de animação, ação e horror tendem a possuir maior faturamento médio.

3.5. Há relação entre o gênero e o Número de votos?



Existe também uma diferença significativa entre o gênero e o número de expectadores médio ($p < 0.001$), em média filmes de horror, ação, faroeste, e aventura tendem a atrair mais o público.

Prevendo a nota do IMDb

Para responder o desafio, utilizaremos um modelo de regressão, pois estamos lidando com uma variável numérica contínua, com valores de 0 a 10. O modelo escolhido foi o Gradient-boosting, por ser um modelo robusto que é capaz de lidar com dados ausentes e não ajustados a distribuição normal. As variáveis a utilizadas foram a nota da crítica (Meta_score), ano de lançamento (Released_Year), duração (Runtime), número de votos (No_of_votes), com uma transformação logarítmica, como forma de ajuste para normalizar a distribuição dos dados, faturamento bruto (Gross) também com transformação logarítmica para normalização dos dados. Utilizei também as variáveis categóricas diretor e gênero, para essas variáveis utilizei o target-encoding, substituindo o nome do diretor e o gênero do filme pela média da nota do público (IMDB_Rating). Para medida da performance do modelo foi utilizado o RMSE (Root mean squared error) pela facilidade de interpretação do erro.

Nota do IMDb do filme "'The Shawshank Redemption"

O filme mostrado no exemplo é uma das maiores notas presentes no IMDb, com nota 9.3, porém o modelo treinado previu que a nota seria 8.9, o que está dentro do esperado pelo erro de aproximadamente 0.2 fornecido pelo RMSE.