# Desafio Cientista de Dados - Renato Susin

## Introdução

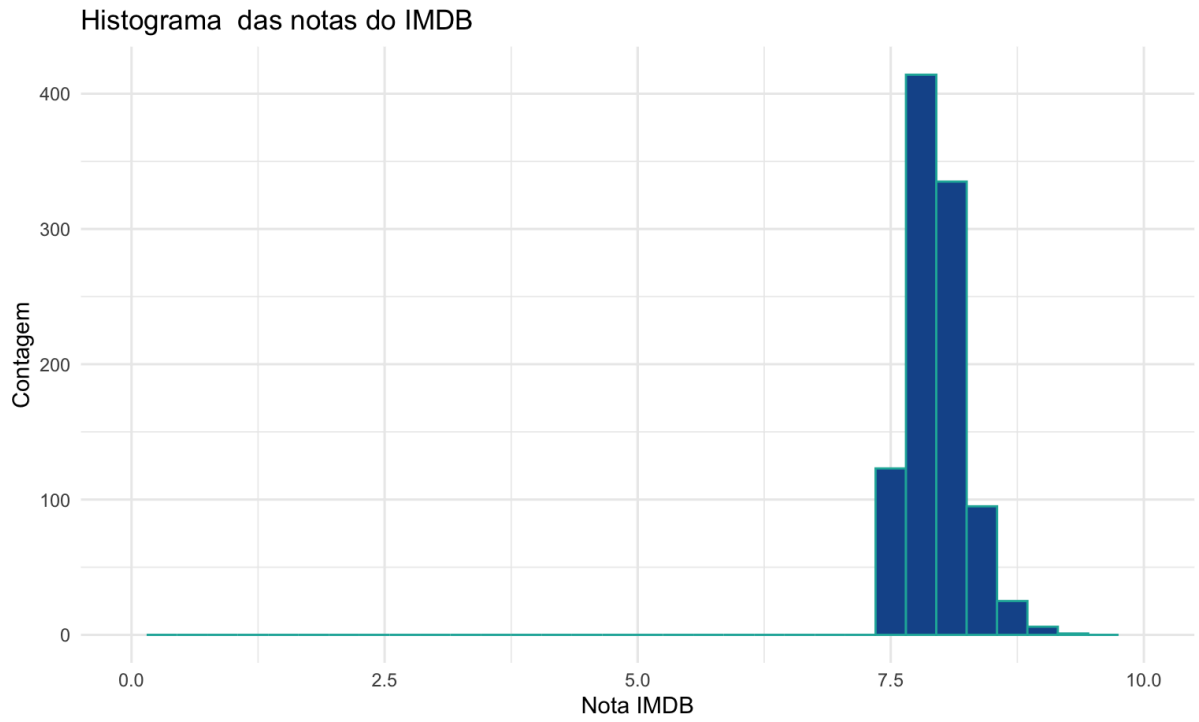
Surgido na virada do séculos XIX para o XX, o cinema exerce um papel de suma influência na nossa sociedade. Com o nome derivado do grego *kinema*, cujo nome significa movimento, a sétima arte foi um grande pilar na popularização do acesso ao cinema pelo mundo. Os conhecidos *nickelodeons* (com esse nome por custar um níquel) deram acesso amplo e generalizado das pessoas à arte, que antes era inviável em decorrência dos altos custos de ir ao teatro ou escutar recitais/concertos, que era de exclusividade das classes mais altas.

Diante disso, surge em meados dos anos 90, com o surgimento da internet, o que hoje é conhecido como IMDb (*Internet Movie Database*), qual é consistido por uma base de dados contendo informações sobre cinema, TV, música e games.

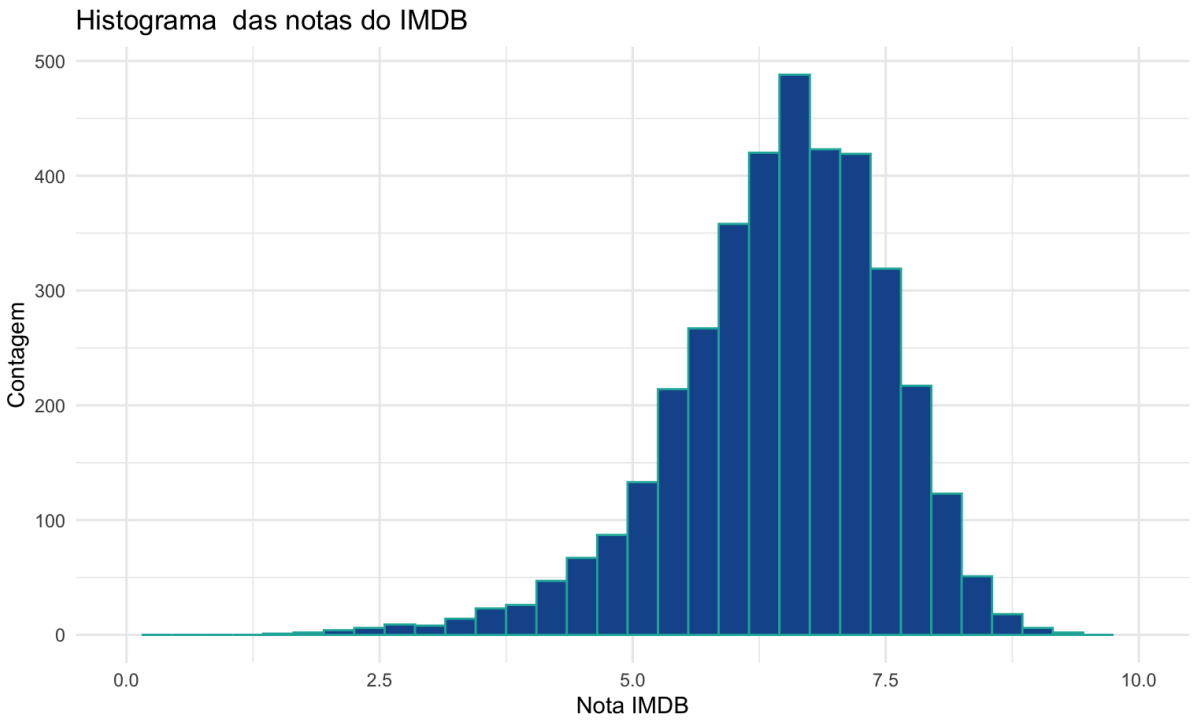
Tendo como ferramenta inicial uma base de dados de filmes presentes no IMDb, temos como objetivo realizar inferências, análises e previsões baseados em aplicações de técnicas de *machine learning* .

## Análise Exploratória dos dados

A base de dados fornecida, embora extensa, mostrava algumas carências. Inicialmente, foi analisado, por meio de um histograma de frequência, o comportamento das notas do IMDb dos filmes que seriam analisados, as quais se portaram da seguinte forma :



Nota-se uma alta concentração em torno de notas altas, o que pode enviesar as conclusões obtidas, por isso, obteve-se uma base de dados mais extensa e mais heterogênea[1], a qual apresentou o seguinte comportamento:



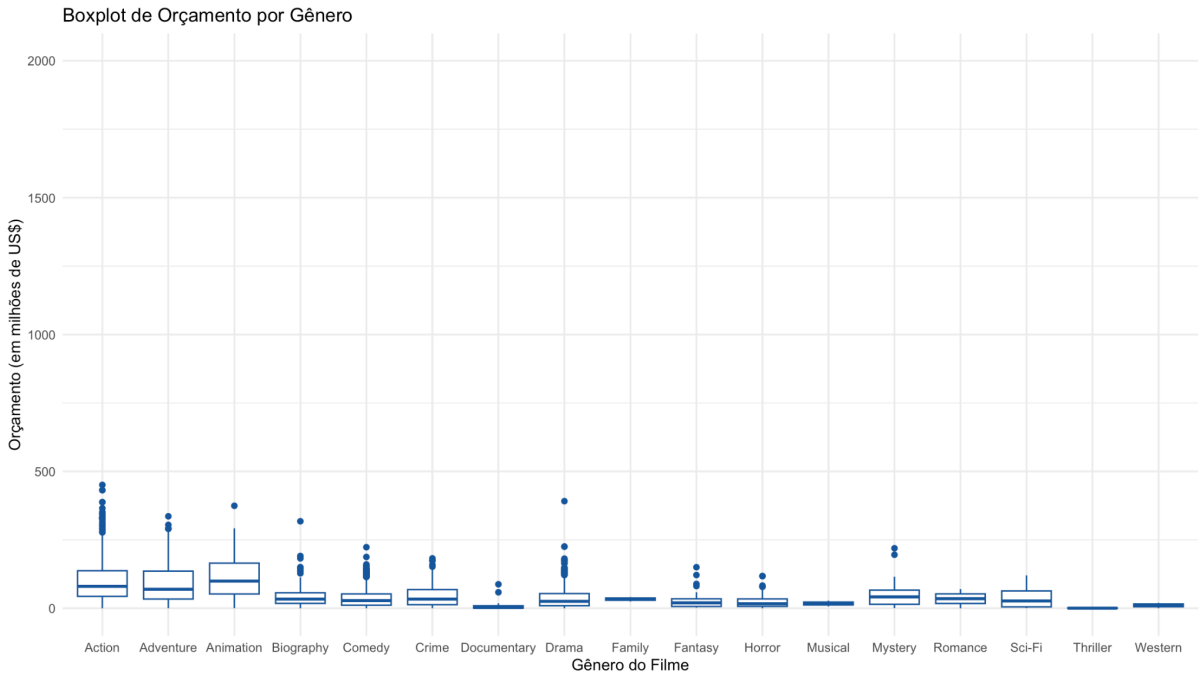

O próprio gráfico já mostra o comportamento mais amplo que essa base apresenta, assemelhando - se a uma distribuição normal ainda tendo os dados acerca de orçamento dos filmes, os quais permitem uma análise ainda mais detalhada.

Ainda, tanto os dados de faturamento e orçamento são apenas disponibilizados na sua forma nominal, isto é, sem levar em conta a inflação, o que também enviesa as análises de dados monetários ao longo do tempo. Por isso, as análises feitas neste relatório serão sempre feitas tendo como base os valores reais (corrigidos pela inflação americana)[2]

[1] link para a base de dados: https://www.kaggle.com/datasets/ramakrushnamohapatra/movies?resource=download)
[2] Podem ser obtidos no link: https://inflationdata.com/Inflation/Consumer_Price_Index/HistoricalCPI.aspx?reloaded=true#Table

## Análise de Orçamento, Faturamento e Lucro da base de dados

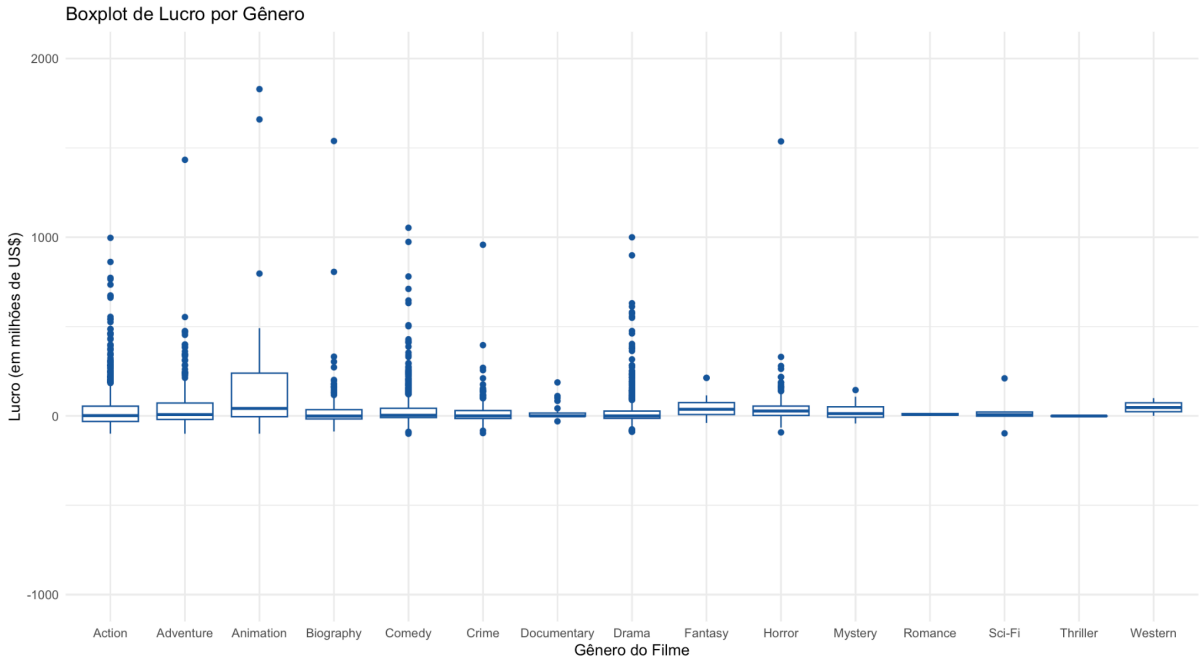
Os dados de orçamento, faturamento e lucro nos permite algumas considerações importantes acerca da produção cinematográfica, a seguir, um gráfico em formato boxplot considerando o orçamento de cada gênero de filme:




Com base nessa figura nota-se um evidente custo adicional, bem como maior variabilidade no valor despendido para produzir um filme dos gêneros Ação, Aventura ou Animação, o que pode ser explicado pelos altos custos de produção envolvendo dublês (no caso de ação e aventura) e de efeitos computacionais, como o CGI (Imagens geradas por computador), por exemplo.

Essa análise inicial já mostra que a produção de um filme em um desses três gêneros pode ser difícil ou inviável para estúdios menores devido aos altos custos e riscos, tornando-o mais atrativos para os produtores mais consolidados dentro desse mercado.

O gráfico a seguir é semelhante ao anterior, no entanto, discriciona os gêneros com base no lucro que eles obtiveram em suas produções:




Aqui, nota-se um maior lucro dentro dos filmes de gênero de animação, o qual foi em média 206 milhões de dólares (em valores de 2024) mostrando que os altos investimentos despendidos são rentáveis, além disso, para um estúdio de maior tamanho, pode ser o gênero mais vantajoso, pois outros gêneros de custo de produção elevada não foram capazes de oferecer o mesmo retorno de forma sistemática.

Além do lucro gerado, calculou-se a probabilidade de cada filme ser lucrativo, dividindo o número de filmes que conseguiram obter retornos maiores dos que os custos de produção pelo número total de filmes, as probabilidades de lucro por gênero são apresentadas na tabela a seguir:

| Gênero | Filmes Lucrativos | Filmes Totais | Probabilidade de Lucro |
|---|---|---|---|
| Ação | 472 | 912 | 52% |
| Aventura | 186 | 332 | 56% |
| Animação | 45 | 67 | 67% |
| Biografia | 101 | 207 | 49% |
| Comédia | 575 | 1022 | 56% |
| Crime | 126 | 250 | 50% |
| Documentário | 20 | 38 | 53% |
| Drama | 340 | 684 | 50% |
| Fantasia | 27 | 35 | 75% |
| Terror | 122 | 161 | 76% |
| Mistério | 15 | 25 | 60% |

A tabela permite verificar uma maior probabilidade de lucro nos gêneros Fantasia e Terror. Outrossim, pode-se afirmar que para um **estúdio de menor tamanho e poder econômico**, **os gêneros a serem preteridos devem ser Fantasia e Terror**, tendo em vista a maior probabilidade de lucro, enquanto para os **estúdios mais tradicionais, o gênero a ser priorizado deve ser Animação, por oferecer retornos mais volumosos.**

Uma outra análise que foi feita com o fim de investigar o comportamento dos dados acerca do mercado cinematográfico foi um **PCA**[3]. Dentro desse modelo, por aceitar apenas variáveis numéricas, considerou-se o número de filmes que cada diretor/ator participou (dentre os presentes no conjunto de dados) como uma variável *proxy* para talento ou experiência do artista. O resultado obtido foi que **as**

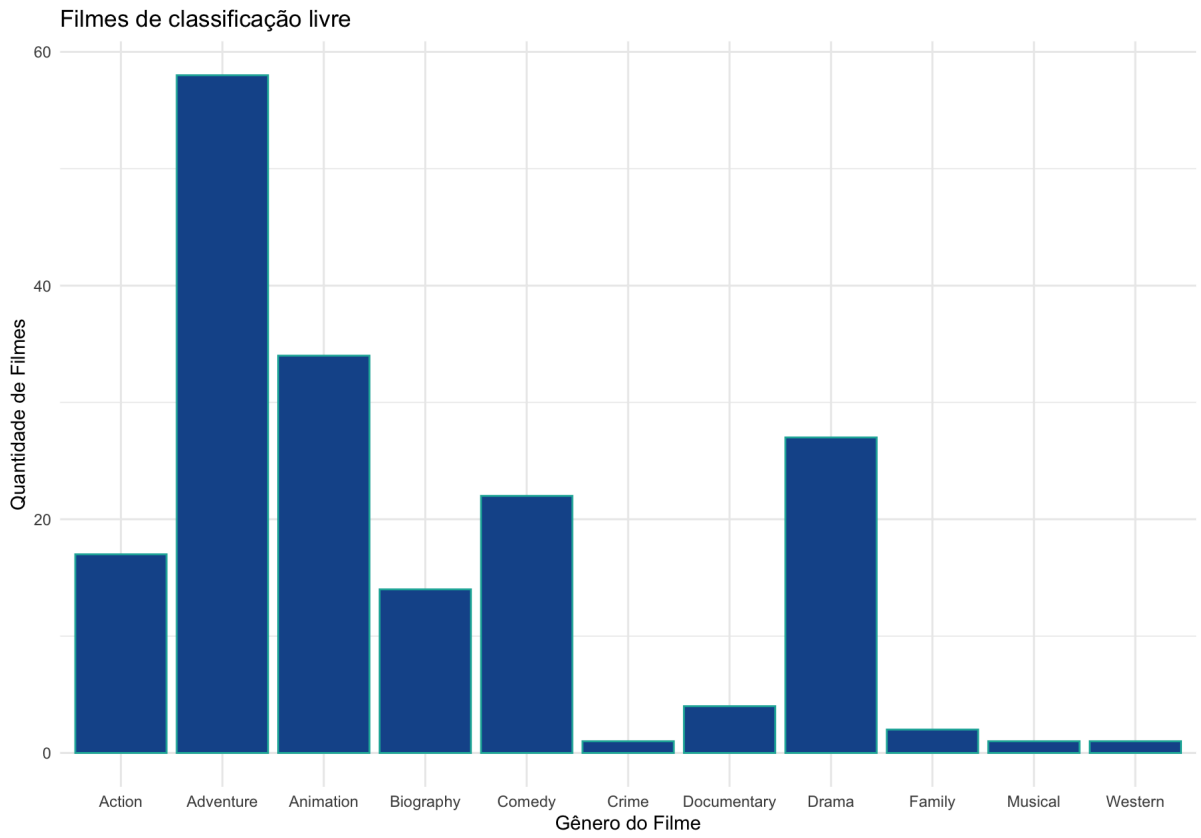
[3] Análise dos Componentes Principais é uma técnica de análise multivariada que objetiva a redução de dimensionalidade, bem como observar quais são as variáveis que mais influem na variabilidade do conjunto de dados

**variáveis financeiras (lucro, faturamento e orçamento) tem um maior efeito na variabilidade, enquanto a experiência de atores e diretores influencia pouco no comportamento dos dados acerca da obra cinematográfica.**

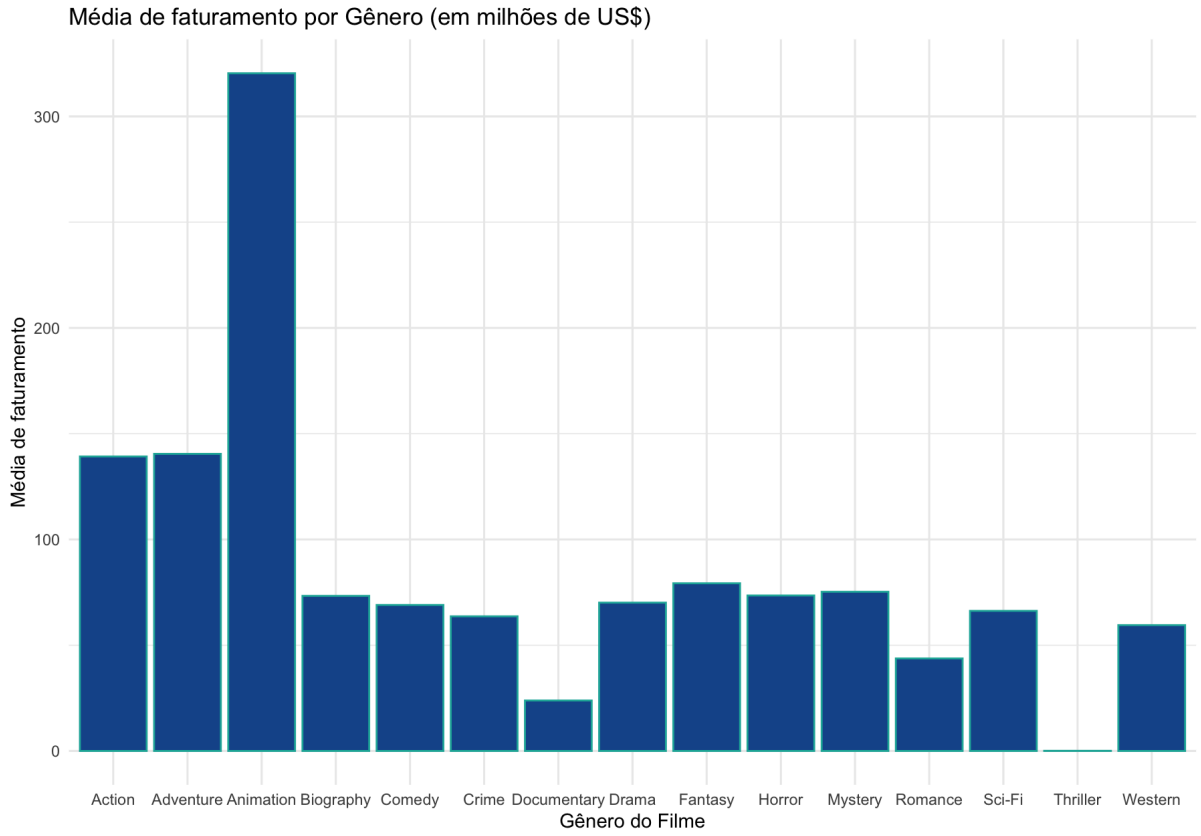
## Recomendação de filmes a pessoas desconhecidas

Recomendar filmes a pessoas que não conhecemos previamente pode ser uma tarefa ingrata, no entanto, há alguns indicadores que podemos nos basear para efetuar a recomendação da melhor maneira possível.

Inicialmente, por desconhecer a idade da pessoa, tampouco se ela gostaria de assistir com a família, priorizaria-se filmes de classificação livre, com poucas cenas impróprias para menores. Logo, os gêneros com maior número de títulos de classificação livre podem ser observados no seguinte gráfico de barras:



Nota-se uma soberania dos gêneros de Aventura e Animação, avaliando o faturamento destes (por conseguinte, maior sucesso em bilheteria), podemos observar o seguinte comportamento:




Evidencia-se um maior faturamento, e consequentemente maior interesse das pessoas, pelo gênero animação. Logo um filme com maior probabildade de agradar a uma pessoa escolhida ao acaso seria do gênero um filme que combine os dois gêneros.

Logo, uma boa opção encontrada dentro desses critérios é o filme de animação japonesa A Viagem de Chihiro (2001) dirigida por Hayao Miyazaki, que se enquadra entre esses dois gêneros mencionamos e ainda é bem avaliado dentro do IMDb.

## Fatores que mais influenciam o faturamento de um filme

Para determinar os fatores predominantes do faturamento de um filme, utilizou-se do mesmo conjunto de dados numéricos que compuseram o PCA, envolvendo as *proxys* para experiência dos artistas e também as variáveis monetárias foram postas no modelo em sua forma logarítmica, para não permitir a diferença de ordem de grandeza interferir no resultado.

As correlações das variáveis com o logaritmo do faturamento são denotadas na tabela a seguir:

| **Variável** | **Coeficiente de correlação**[4] |
|---|---|
| Nota do IMDb | 0.13 |
| N° de filmes do diretor | 0.20 |
| N° de filmes do ator | 0.27 |
| N° de votos (em log) | 0.66 |
| Orçamento (em log) | 0.64 |

Com base nessa tabela de correlações, verificamos que os filmes de maior orçamento tendem a ter também melhor bilheteria, e mais pessoas avaliando a obra no IMDb em casos de maior sucesso financeiro.

Em um modelo feito de regressão linear, o p-valor mostrou que as variáveis referentes ao Número de filmes do diretor e do ator não foi significativa, evidenciando que **a experiência dos atores e diretores envolvidos na obra não tem impacto no faturamento de um filme.**

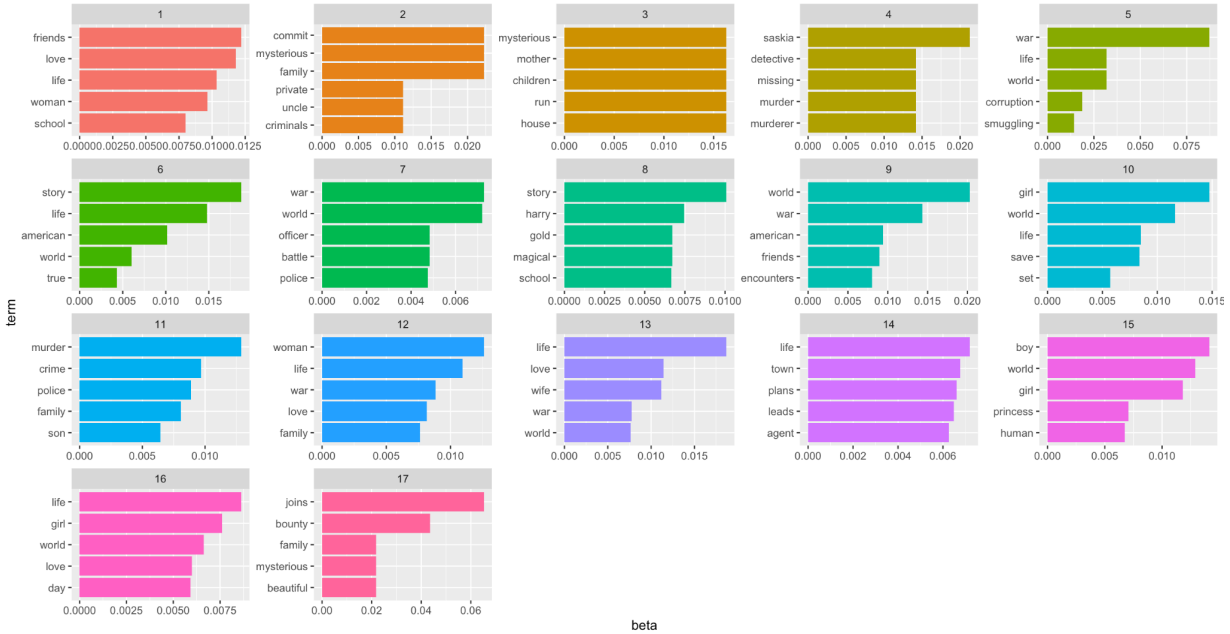
[4] Utilizou-se do coeficiente de correlação de Pearson

O valor de R - quadrado do modelo foi de 0,56, um valor razoável, explicando relativa aderência da Nota do IMDb, número de votos e do orçamento ao faturamento que o filme obterá.

## Modelo de LCA[5] para detectar o gênero com base na sinopse

O modelo LCA (Latent Dirichlet Allocation (LDA)) é um modelo comumente utilizado para agrupar diferentes vetores de texto com base em suas palavras em comum. Para a análise deste, usamos o *dataframe* original, em decorrência de uma sinopse mais bem detalhada do que o *dataframe* encontrado na base de dados.

Utilizando - se do agrupamento por tópicos com o auxílio do dos pacotes "*topicmodels*" e "*quanteda*", conseguimos agrupar as palavras dentre os seguintes tópicos, cada um deles representando um gênero diferente:




[5] Latent Dirichlet Allocation (LDA) é um modelo probabilístico de machine learning usado para análise de tópicos em coleções de textos

Portanto, de acordo com essa divisão em tópicos, consegue-se realizar algumas inferências, como o tópico 11 que está relacionado a filmes policiais, ou o tópico 1, relacionando palavras muito associadas à filmes de comédia.

Logo, podemos sim fazer inferências acerca do gênero tendo como base a sinopse graças a sofisticados modelos de *Machine Learning*.

## Previsão de nota do IMDb

Modelos de previsão são amplamente utilizados por profissionais da área de dados em todo o mundo, no âmbito do mercado cinematográfico, eles podem prever o êxito de um filme antes do lançamento  com base em dados já obtidos, além de poderem metrificar a viabilidade de projetos financeiros, etc.

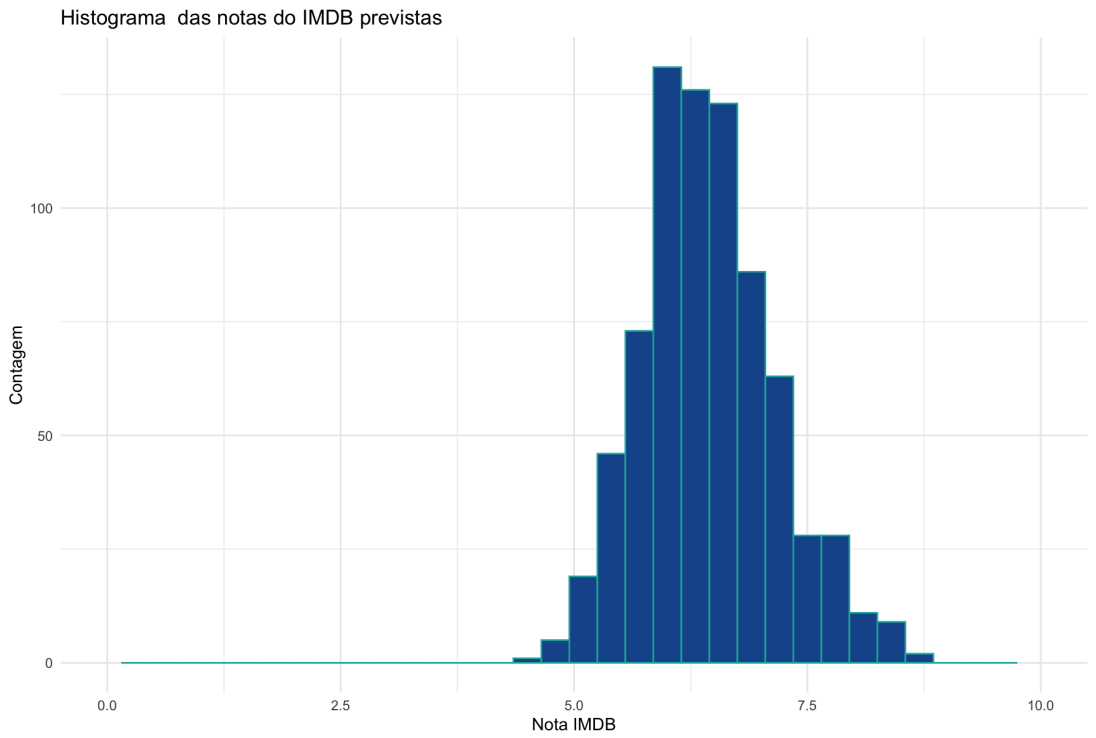
Um dos modelos mais tradicionais e amplamente utilizados é o de regressão linear, que também já foi descrito neste relatório. No modelo implementado com as variáveis numéricas do conjunto, o resultado obtido foi considerado insatisfatório.

Tendo a Nota do IMDb como variável independente e as demais numéricas como dependentes, o as variáveis não significativas foram novamente as de número de filmes dos  diretores e dos atores, enquanto as outras, apesar de significativas, foram muito pouco aderentes ao modelo, denotado pelo baixo valor de R-quadrado.

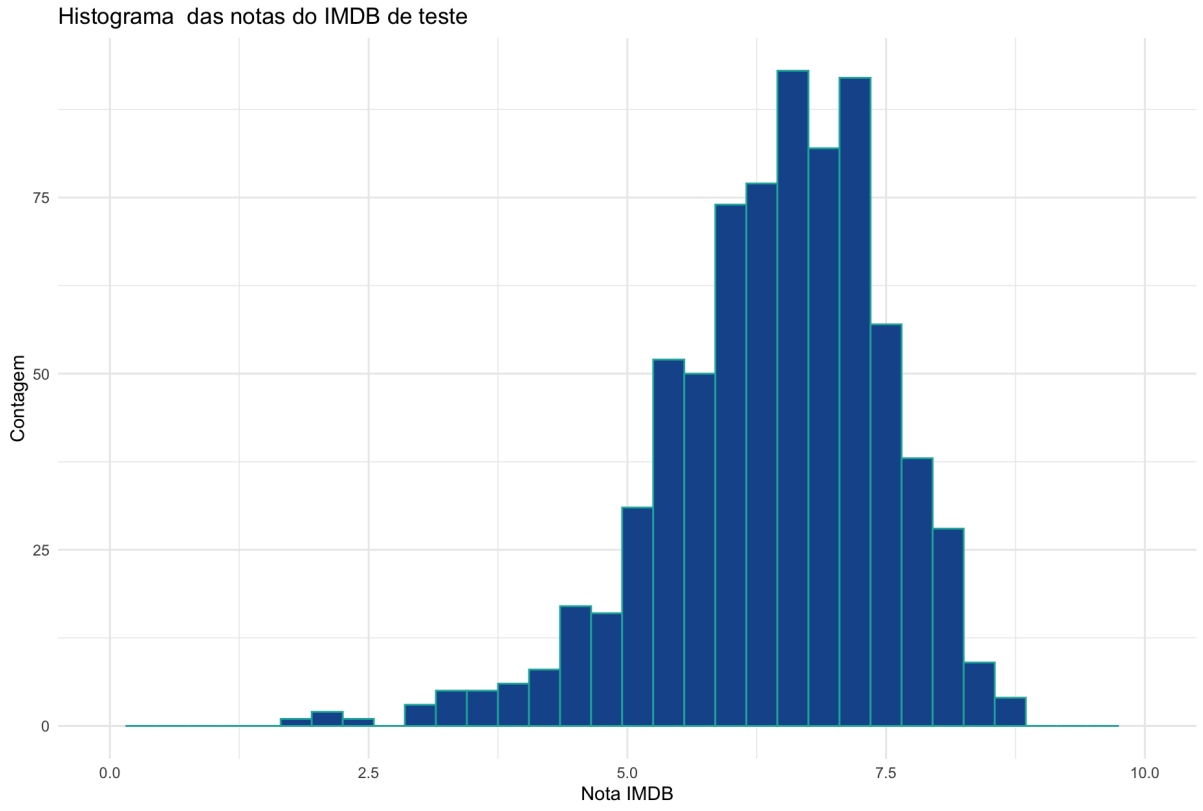
Tendo isso em vista, utilizou-se de outro modelo, o *Random Forest*[6], que é uma técnica de *machine learning* que consiste no uso de diferentes árvores de decisão que pode ser tanto de classificação, no caso de variáveis categóricas, quanto de regressão, para variáveis quantitativas, como é o caso da nota do IMDb.

Com uma amostra de 80% separados para treino do modelo e 20% para teste, obteve-se o seguinte resultado:

[6] Random Forest é um modelo de machine learning que constrói múltiplas árvores de decisão a partir de amostras aleatórias dos dados (bootstrap) e subconjuntos de features, combinando suas previsões para melhorar a acurácia e reduzir o overfitting. Para classificação, usa votação majoritária; para regressão, calcula a média das previsões. No contexto de filmes, por exemplo, pode prever a nota IMDB com base em variáveis como votos e faturamento, sendo robusto e fácil de interpretar, implementado no R com o pacote randomForest




A seguir, o histograma mostra o comportamento dos dados de fato observados dentro do conjunto de teste:



Nota - se certa semelhança do formato, embora, a variância dos dados previstos seja um pouco menor,  o que indica uma maior regressão à média vinda dos dados previstos pelo modelo.

Dentro desse modelo, as variáveis de ator e diretor não puderam entrar na análise pois o código não permitiu variáveis categóricas com tantos fatores distintos, no entanto, foi utilizado o número de filmes que cada diretor/ator participou novamente como *proxy* para a experiência destes.

No entanto, a comparação de ambos os gráficos permite inferir que o **modelo de *Random Forest* é um modelo eficiente para previsão dentro desse contexto.**

## Previsão de nota do IMDb no filme “Um sonho de Liberdade”

Tendo em vista o modelo desenvolvido, sua eficácia é posta à prova ao analisar os dados referentes ao filme “Um sonho de Liberdade” (1994) de Frank Darabont.

A nota obtida foi de 8,79, apenas 0,5 abaixo da nota verídica encontrada no IMDb.

Outrossim, confirma-se a **eficácia da técnica de *Random Forest* para a previsão de dados,** com ampla usabilidade em diversas áreas de conhecimento e possibilidade de contribuir muito com o desenvolvimento científico, econômico e social.

## Conclusão

Concluindo, após aplicação de técnicas de estatística e *machine learning* para gerar *insights* a respeito do mercado cinematográfico, podemos aferir o maior investimento necessário para filmes de Ação Aventura e Animação, sendo este último o de melhor retorno, e por conseguinte um caminho mais indicado para ofertantes de melhor poder aquisitivo. Já para seus pares de menor imponência financeira, deve-se privilegiar os gêneros de Fantasia e Terror por terem maior probabilidade de serem lucrativos.

Também destaca-se a eficácia dos modelos de aprendizado de máquinas quanto para a classificação de gênero por meio da sinopse através de um LDA, quanto para a previsão de notas do IMDb com um *Random Forest*, mostrando o potencial e a abrangência desse tipo de técnica para o desenvolvimento das mais amplas áreas do conhecimento, transcendendo a sua utilidade dentro da Sétima Arte.