Relatório de Análise Exploratória e Visualização de Dados

Introdução

Para elaboração deste trabalho da disciplina de Análise Exploratória e Visualização de Dados, foi utilizado o conjunto de dados: IMDB Dataset 2023, disponível no Kaggle. Foi feita uma análise exploratória de dados utilizando Python e suas respectivas bibliotecas de análise e visualização de dados. O notebook foi feito no Colab, de onde foram extraídas as informações contidas neste relatório.

Este relatório apresenta uma análise exploratória de um conjunto de dados extraído do IMDB, contendo informações sobre filmes lançados após 1970, com mais de 50.000 avaliações e valores financeiros em dólar. O objetivo da análise é explorar características dos filmes e identificar possíveis padrões e correlações entre as variáveis.

Descrição do Conjunto de Dados

O dataset possui 3.348 observações e 12 atributos, que incluem identificadores, títulos, gêneros, classificação, tempo de execução, orçamento e receita de bilheteria, entre outros. Uma prévia da estrutura e das variáveis foi realizada para garantir que os dados estão completos e prontos para análise.

Conjunto de Dados contendo Informações sobre Filmes do site IMDB

Os dados foram obtidos por meio de web scraping em Python e combinados com um repositório compartilhado pelo IMDB. O conjunto foi pré-processado para incluir apenas filmes lançados após 1970 e que atualmente tenham mais de 50.000 avaliações. Além disso, foram selecionados apenas filmes cujos orçamentos e arrecadações estão denominados em USD, para evitar discrepâncias.

O conjunto de dados contém 3.348 observações descritas por 12 atributos.

Atributos

- **id** - ID do filme usado pelo repositório IMDB
- **primaryTitle** - título em inglês
- **originalTitle** - título original na língua nativa
- **isAdult** - classificação etária
- **runtimeMinutes** - duração total em minutos
- **genres** - gêneros

- **averageRating** - avaliação final, baseada em todas as avaliações
- **numVotes** - número total de votos (avaliações)
- **budget** - orçamento total em USD
- **gross** - receita mundial total em USD
- **release_date** - data de lançamento, primeira exibição
- **directors** - diretores

Última Atualização: 12 de novembro de 2023

Análise das Variáveis

Variáveis Categóricas

- **Gêneros**: A variável de gêneros foi analisada para identificar os tipos de filmes mais comuns. Como será mostrado neste trabalho, os gêneros mais frequentes são Drama, Comédia e Ação, indicando uma maior produção ou popularidade desses estilos entre o público.
- **Diretores**: A análise dos diretores revela que alguns têm vários filmes no dataset, sugerindo uma correlação entre o número de produções e a popularidade ou a receptividade de seus filmes.
- **Classificação Etária (isAdult)**: A coluna de classificação etária destaca a proporção de filmes voltados para o público adulto em comparação com os filmes para todos os públicos, o que pode refletir as preferências de conteúdo.

Variáveis Numéricas

- **Duração (runtimeMinutes)**: A duração dos filmes varia, mas a maioria está entre 90 e 120 minutos. Essa análise ajuda a identificar a duração padrão de filmes populares.
- **Avaliação Média (averageRating)**: A maioria dos filmes tem uma avaliação média entre 6 e 8, e filmes com uma média superior a 8 são mais raros, possivelmente indicando produções de alta qualidade ou com grande aprovação do público.
- **Número de Votos (numVotes)**: Alguns filmes possuem um número de votos muito alto, indicando popularidade ou impacto cultural significativo.
- **Orçamento (budget)**: O orçamento dos filmes varia consideravelmente. A análise mostra que filmes com orçamentos maiores tendem a ter maior visibilidade, embora nem todos sejam grandes sucessos de bilheteria.
- **Receita (gross)**: A receita global reflete o sucesso financeiro dos filmes. Filmes que alcançam altas receitas costumam ser os mais populares ou altamente divulgados.
- **Data de Lançamento (release_date)**: A variável de data foi utilizada para explorar a distribuição de lançamentos ao longo dos anos, identificando possíveis tendências, como aumento de lançamentos em certos períodos ou preferência por datas específicas para grandes lançamentos.

- **Títulos (primaryTitle e originalTitle)**: Os títulos foram analisados para identificar variações linguísticas e como o título em inglês pode diferir do original. Isso pode refletir estratégias de adaptação cultural, visando aumentar o apelo dos filmes em mercados específicos.

A imagem a seguir mostra as principais características do conjunto de dados:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3348 entries, 0 to 3347
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   id              3348 non-null   object 
 1   primaryTitle    3348 non-null   object 
 2   originalTitle   3348 non-null   object 
 3   isAdult         3348 non-null   int64  
 4   runtimeMinutes  3348 non-null   int64  
 5   genres          3348 non-null   object 
 6   averageRating   3348 non-null   float64
 7   numVotes        3348 non-null   int64  
 8   budget          3348 non-null   int64  
 9   gross           3297 non-null   float64
 10  release_date    3343 non-null   object 
 11  directors       3348 non-null   object 
dtypes: float64(2), int64(4), object(6)
memory usage: 314.0+ KB
```

Número de Entradas: O DataFrame contém 3.348 registros, o que é um tamanho razoável para uma análise detalhada de filmes. Esse volume permite análises estatísticas e visuais consistentes.

Colunas e Tipos de Dados:

- **Objetos (object)**: As colunas id, primaryTitle, originalTitle, genres, release_date e directors são categóricas ou textuais. Isso inclui o índice do filme neste conjunto de dados, título do filme em inglês e no idioma original, o gênero, a data de lançamento e os diretores.
- **Inteiros (int64)**: As colunas isAdult, runtimeMinutes, numVotes e budget contêm valores inteiros. A coluna isAdult indica se o filme é voltado para adultos (1) ou não (0), enquanto runtimeMinutes representa a duração do filme em minutos. numVotes e budget registram o número de votos e o orçamento, respectivamente.

- **Floats (float64)**: As colunas averageRating e gross estão em formato de ponto flutuante. averageRating representa a nota média dos filmes e gross registra a receita bruta mundial em USD.

Valores Nulos:

- **gross**: A coluna **gross** apresenta 51 valores ausentes. Para lidar com esses valores, optou-se por preenchê-los com a **mediana** da coluna, uma vez que a mediana é menos sensível a outliers e representa melhor distribuições assimétricas.
- **release_date**: A coluna release_date foi convertida para tipo data e os valores ausentes foram tratados usando interpolação linear. Esse método permite estimar as datas ausentes com base nas datas existentes, aplicando uma distribuição linear, o que garante uma continuidade razoável sem introduzir grandes discrepâncias temporais. Por se tratar apenas 5 valores faltantes, poderíamos também substituir as datas manualmente, porém a substituição linear, embora possa ser imprecisa, forneceu uma boa aproximação neste caso; conseguindo estimar com relativa precisão ao menos o ano de lançamento dos filmes, o único valor incorretamente estimado foi para o filme Shark Tale ('O Espanta Tubarões'), lançado em 2004, que teve ano de lançamento estimado em 2003. Para manter maior integridade dos dados, e por ter apenas 5 valores faltantes, foi optado por substituir as datas por seus valores reais.

```
# Fixando o tipo dos dados e lidando com valores nulos
df['release_date'] = pd.to_datetime(df['release_date'], errors='coerce')

# Preenchendo valores ausentes em 'release_date' usando o método de interpolação com distribuição l
df['release_date'] = df['release_date'].interpolate(method='linear', limit_direction='forward')

# Preenchendo valores ausentes em 'gross' com a mediana (mediana é menos sensível à outliers e repr
df['gross'] = df['gross'].fillna(df['gross'].median())
```

Análise das Correlações

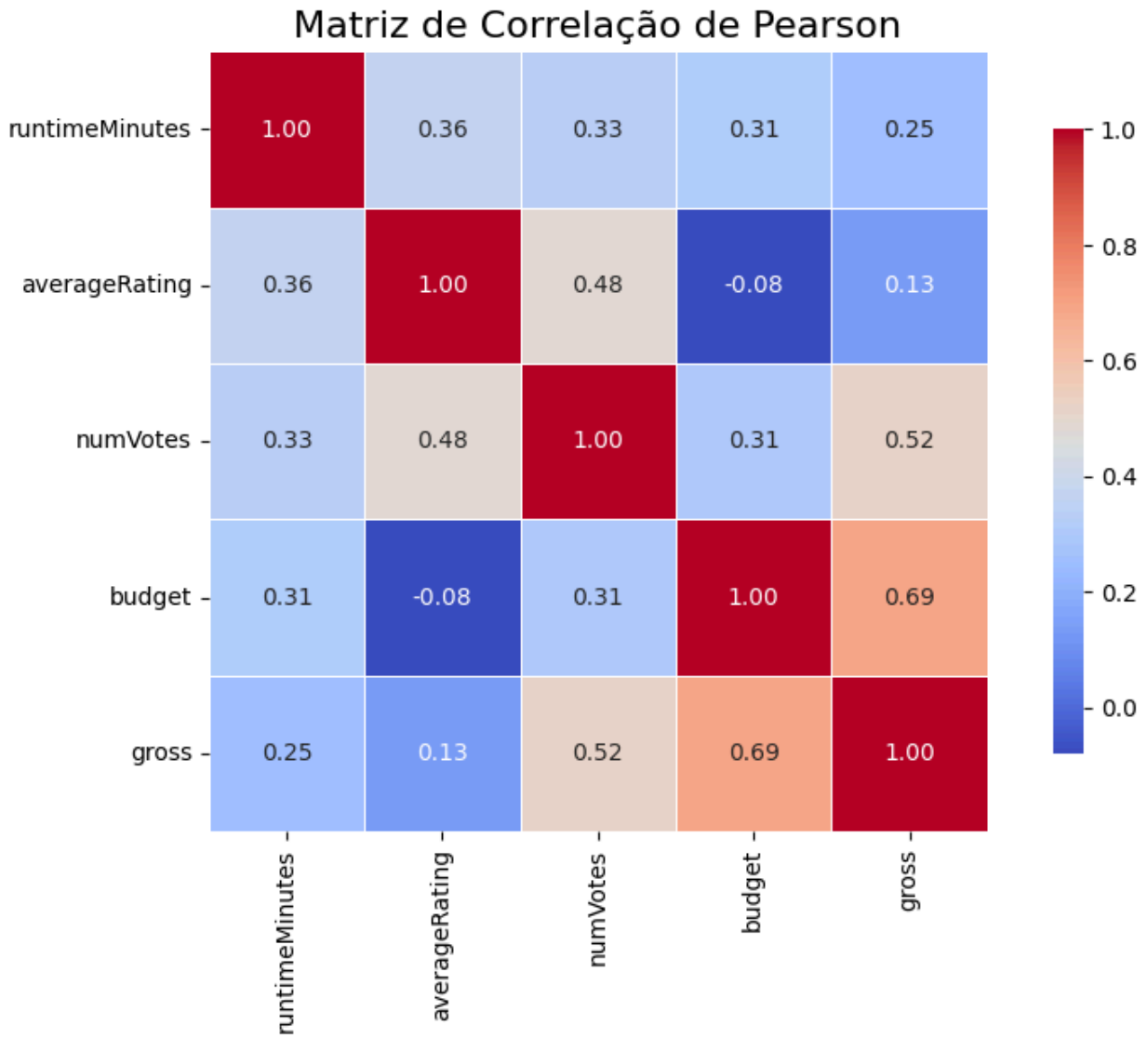
Para termos melhor entendimento sobre a correlação entre diferentes variáveis, podemos utilizar a matriz de correlação de Pearson.

A **matriz de correlação de Pearson** é uma ferramenta estatística que mede o grau de associação linear entre pares de variáveis numéricas. Os valores da correlação variam de -1 a 1, onde:

- **+1** indica uma correlação positiva perfeita: conforme uma variável aumenta, a outra também aumenta na mesma proporção.

- **-1** indica uma correlação negativa perfeita: conforme uma variável aumenta, a outra diminui proporcionalmente.
- **0** indica ausência de correlação: as variáveis não têm nenhuma relação linear significativa.

Na matriz de correlação, cada célula representa o coeficiente de correlação entre um par específico de variáveis. Valores próximos de +1 ou -1 sugerem uma forte associação, enquanto valores próximos de 0 indicam uma relação fraca ou inexistente. Essa matriz é útil para identificar relações potenciais e entender a estrutura dos dados, auxiliando na escolha de variáveis relevantes para modelos preditivos e análises mais detalhadas.

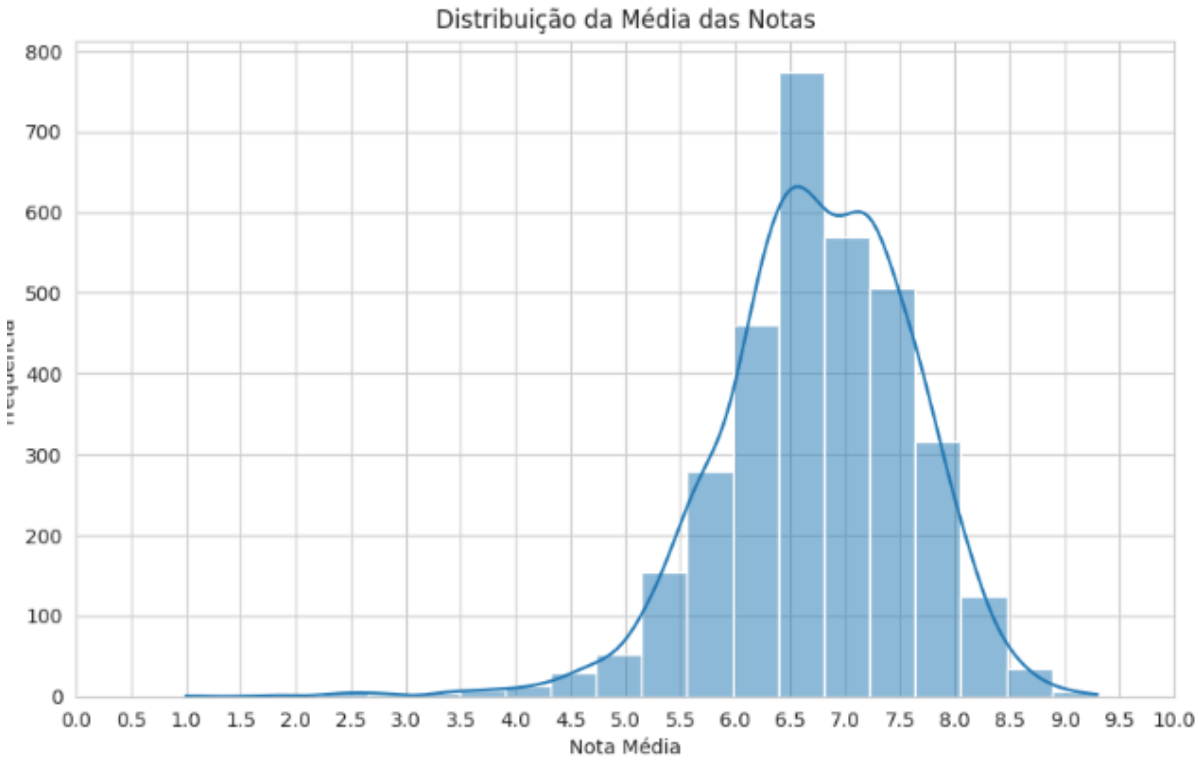


1. **Correlação Budget e Gross (0.69):** A correlação mais forte é entre o **budget** e o **gross** (0.69), indicando uma relação moderada a forte. Isso sugere que filmes com

orçamentos maiores tendem a gerar receitas mais altas. No entanto, a correlação não é perfeita, o que significa que um alto orçamento não garante uma alta receita, mas aumenta a probabilidade.

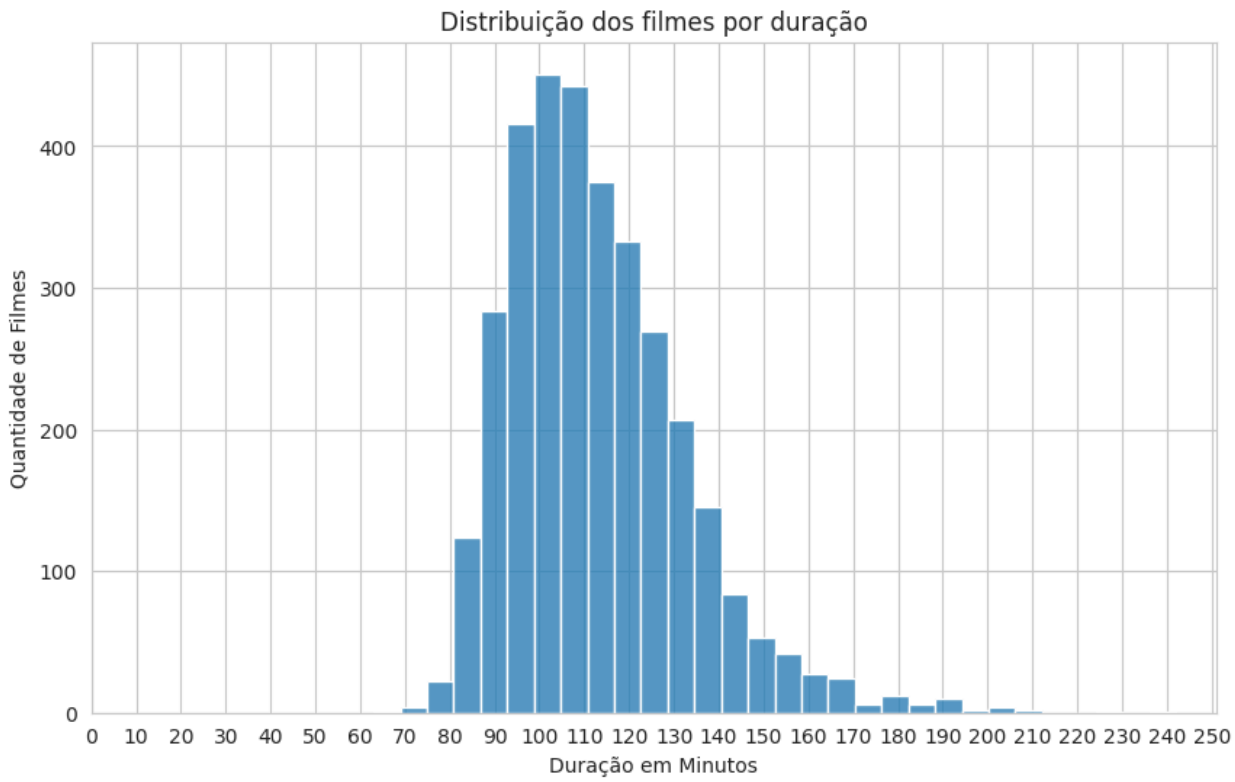
2. **Correlação NumVotes e Gross (0.52):** Há uma correlação positiva entre **numVotes** e **gross** (0.52), indicando que filmes com mais votos (geralmente mais populares) tendem a ter maior receita. Isso pode refletir o impacto do interesse público na performance de bilheteria.
3. **Correlação AverageRating e NumVotes (0.48):** Existe uma correlação moderada entre **averageRating** e **numVotes** (0.48), sugerindo que filmes com avaliações médias mais altas tendem a receber mais votos, possivelmente devido à popularidade impulsionada por uma boa recepção.
4. **Correlação RuntimeMinutes com AverageRating (0.36) e NumVotes (0.33):** A duração dos filmes tem uma correlação moderada com a média de avaliações (0.36) e com o número de votos (0.33), indicando que filmes mais longos podem ser ligeiramente mais populares ou bem avaliados.
5. **Outras Correlações:** A relação entre **averageRating** e **budget** (-0.08) é praticamente nula, sugerindo que o orçamento não tem grande influência na avaliação média do público.

A seguir, faremos algumas análises acerca de nosso conjunto de dados. Iniciaremos com a análise da distribuição da **AverageRating,** para isso foi usado o histograma a seguir:



Conforme apresentado no gráfico, as médias das notas se concentram entre 6 - 8.

De igual maneira, podemos visualizar a distribuição dos filmes por duração:

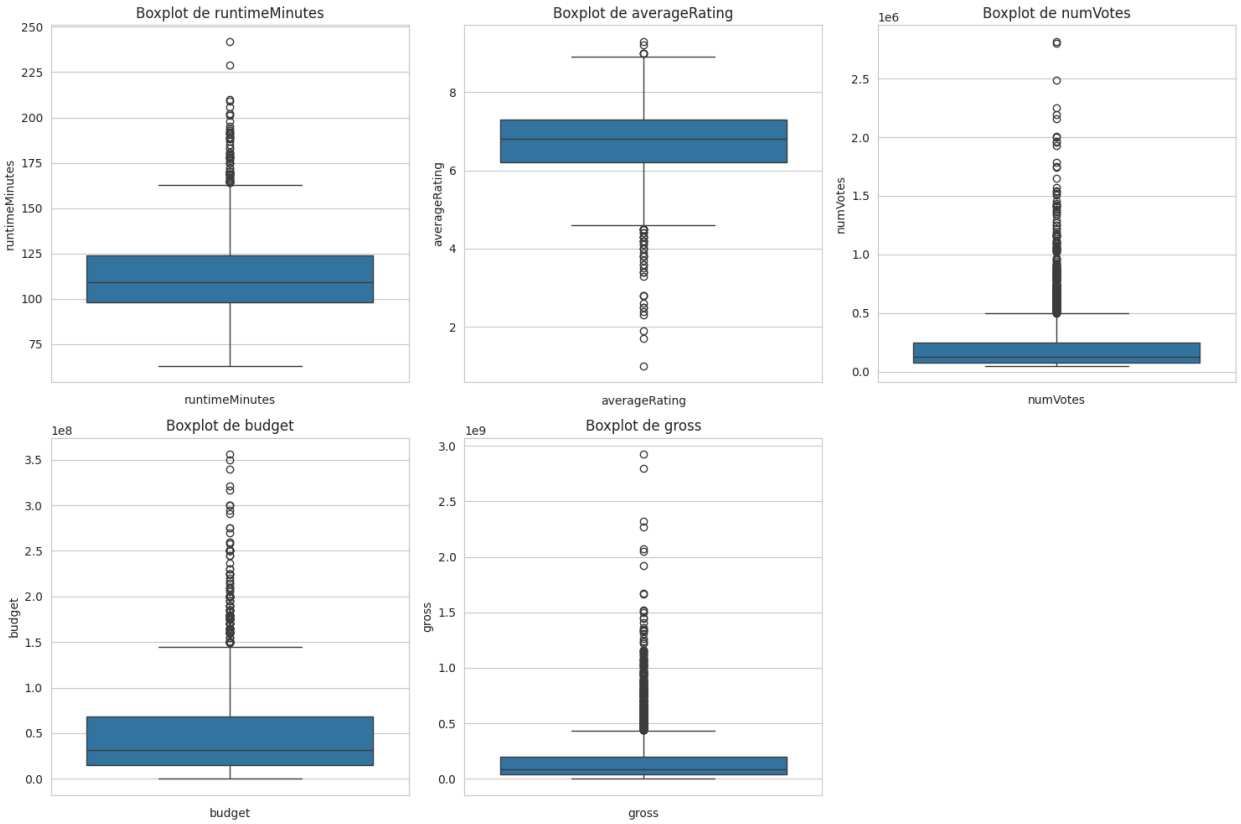


Mostrando que a maioria dos filmes possuem uma duração entre 90min e 120min.

A análise da distribuição das variáveis foi realizada utilizando boxplots. Essa ferramenta visual permite compreender melhor as características de cada variável no dataset de filmes, revelando informações importantes sobre a distribuição dos dados.

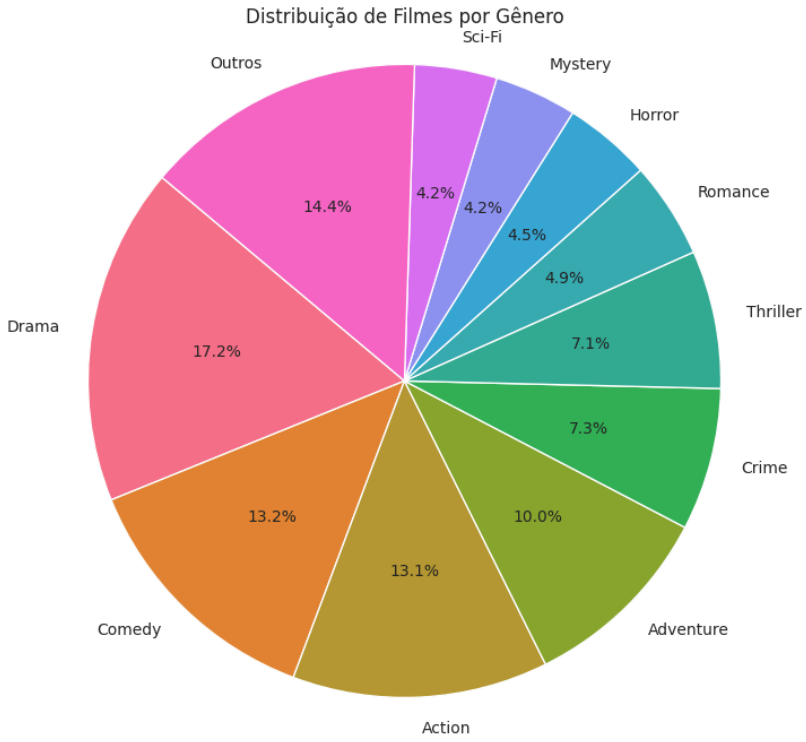
boxplots são usados principalmente para representar variáveis numéricas. Eles são eficazes para resumir a distribuição de dados quantitativos, permitindo visualizar a mediana, os quartis e possíveis outliers. Através dos boxplots, é possível analisar a dispersão, simetria e assimetrias de um conjunto de dados, facilitando a comparação entre diferentes grupos ou categorias.

A seguir, apresentaremos os boxplots correspondentes a cada variável, que ajudarão a ilustrar suas características e comportamentos.



runtimeMinutes: A maioria dos filmes tem uma duração entre aproximadamente 90 e 120 minutos, conforme observado anteriormente. Há alguns filmes com duração significativamente maior, indicando a presença de outliers na parte superior.
averageRating: Como esperado, a média dos votos se concentra entre 6 e 8, com pouca variação. A distribuição parece relativamente simétrica.
numVotes: A maioria dos filmes tem um número de votos relativamente baixo, com alguns poucos filmes com um número muito alto de votos, indicando a presença de outliers na parte superior.
budget: A maioria dos filmes possui um orçamento relativamente baixo, com alguns filmes com orçamentos muito altos, indicando a presença de outliers na parte superior.
gross: A maioria dos filmes tem uma bilheteria relativamente baixa, com alguns filmes com bilheterias muito altas, indicando a presença de outliers na parte superior.

Foi feita também a análise sobre a distribuição de filmes lançados por gênero, conforme a seguir:

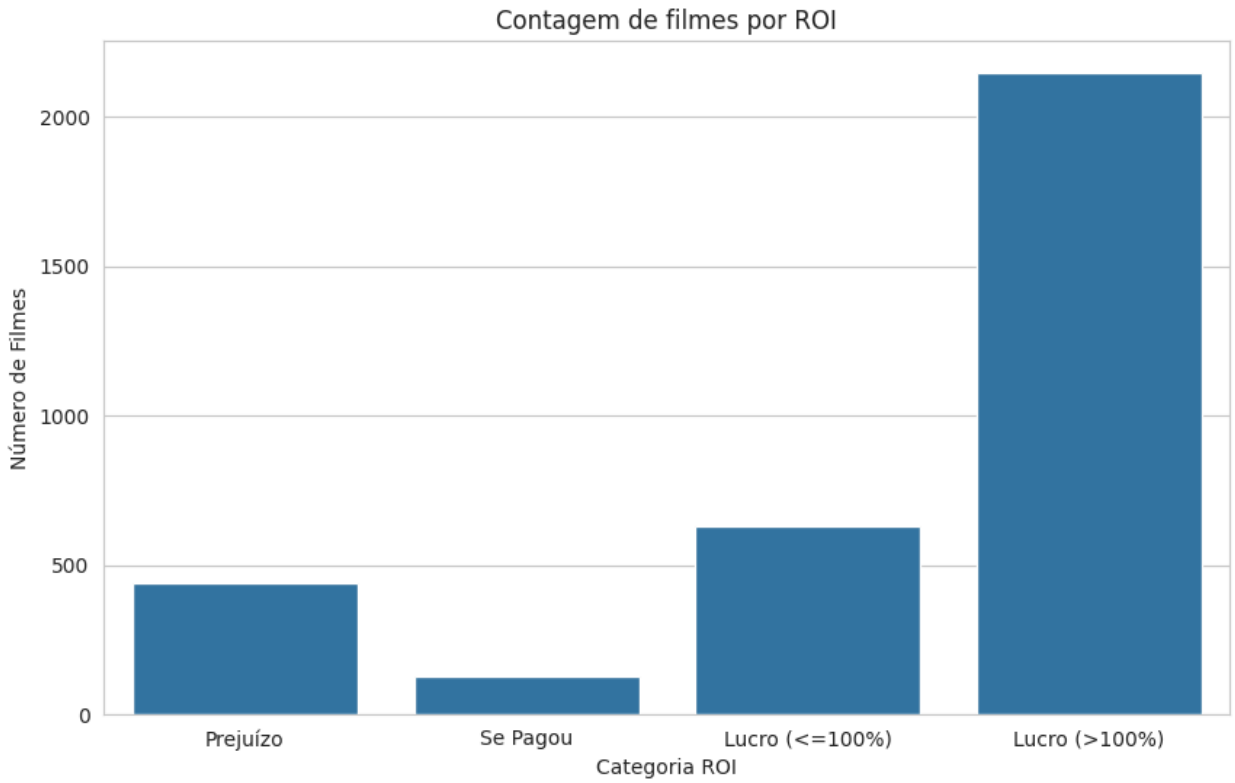


Os gêneros "Drama" e "Comédia" se destacam como os mais frequentes no conjunto de dados, os gêneros de "Ação" e "Aventura" também possuem uma representatividade considerável, mas apesar da dominância de alguns gêneros, o gráfico demonstra uma boa diversidade de categorias, abrangendo desde dramas e comédias até filmes de terror, mistério e ciência ficção.

Uma análise interessante que podemos fazer é do Retorno sobre Investimento (ROI), que é uma ferramenta essencial para avaliar a performance financeira dos filmes, permitindo entender se um projeto foi lucrativo ou se resultou em prejuízo. O ROI é calculado como a diferença entre a receita bruta e o orçamento, dividido pelo orçamento, expressando-se como uma porcentagem.

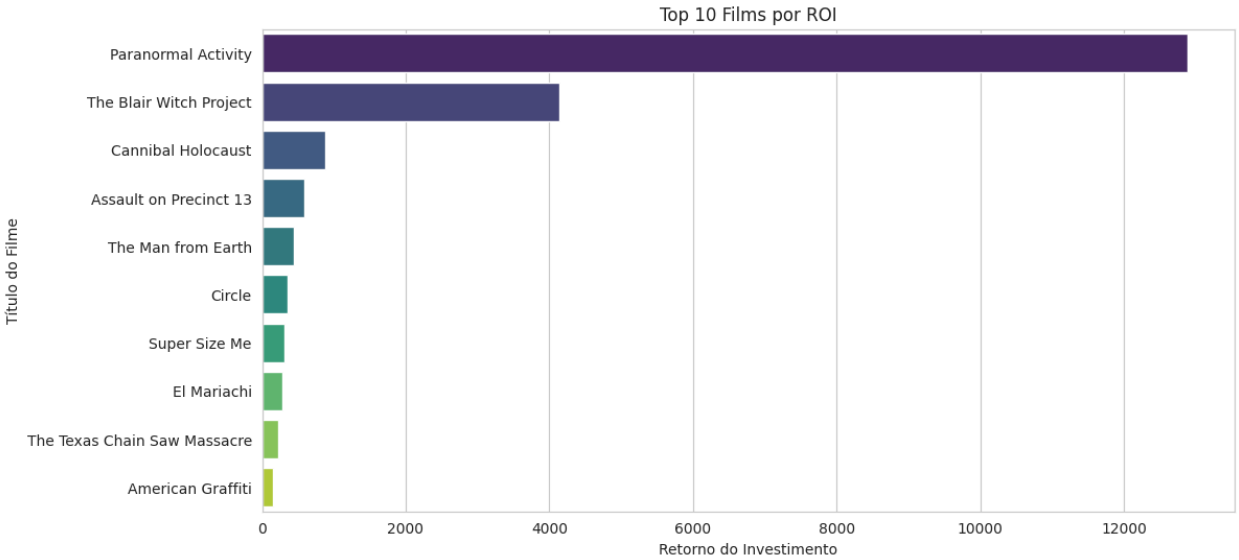
Neste trabalho foram definidas categorias de ROI da seguinte forma: prejuízo ocorre quando o ROI é menor que -10%, indicando uma perda superior a 10% do investimento; a categoria "Se Pagou" abrange ROIs entre -10% e 10%, o que significa que o filme pelo menos recuperou o investimento; "Lucro (<=100%)" corresponde a ROIs entre 10% e 100%, representando lucros de até 100% do investimento; e "Lucro (>100%)" refere-se a ROIs superiores a 100%, indicando lucros que ultrapassam o dobro do investimento inicial.

A seguir, apresentaremos os resultados da análise, que mostram a contagem de filmes em cada uma dessas categorias de ROI.



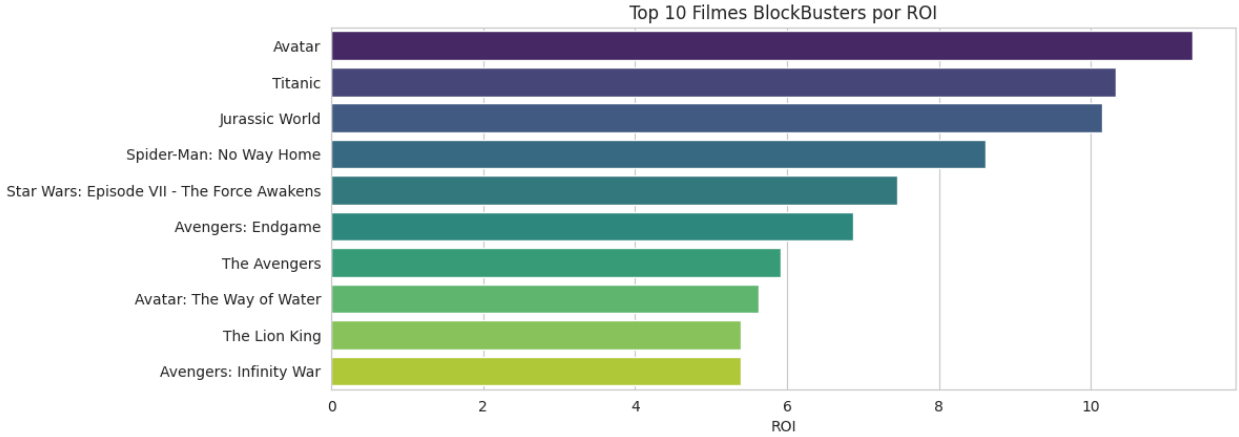
Esta distribuição mostra que a maioria dos filmes do nosso conjunto de dados tiveram um ROI maior que 100%.

A seguir os filmes com maior ROI:

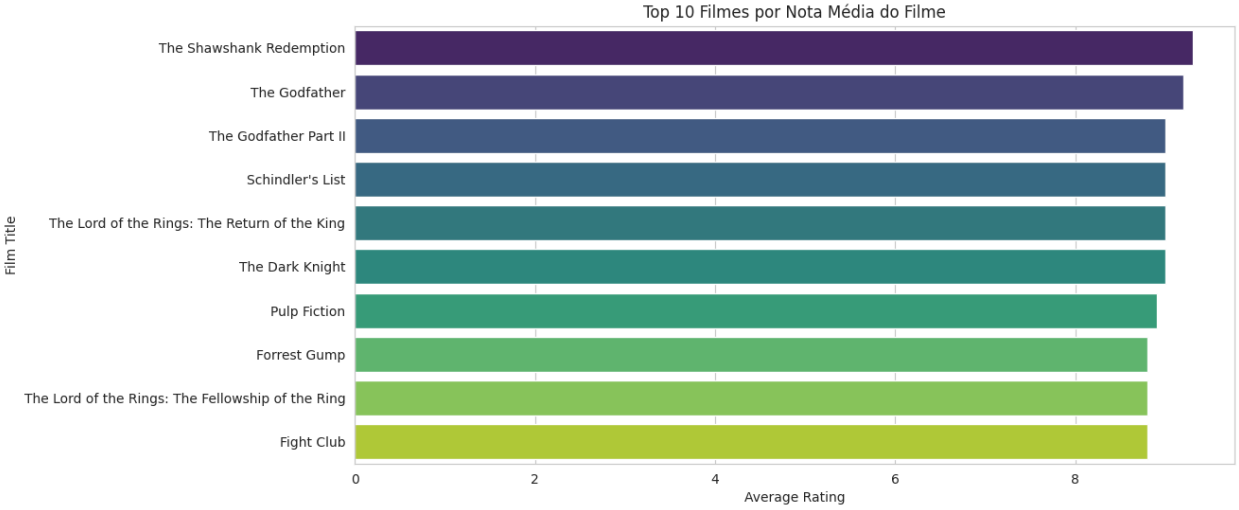


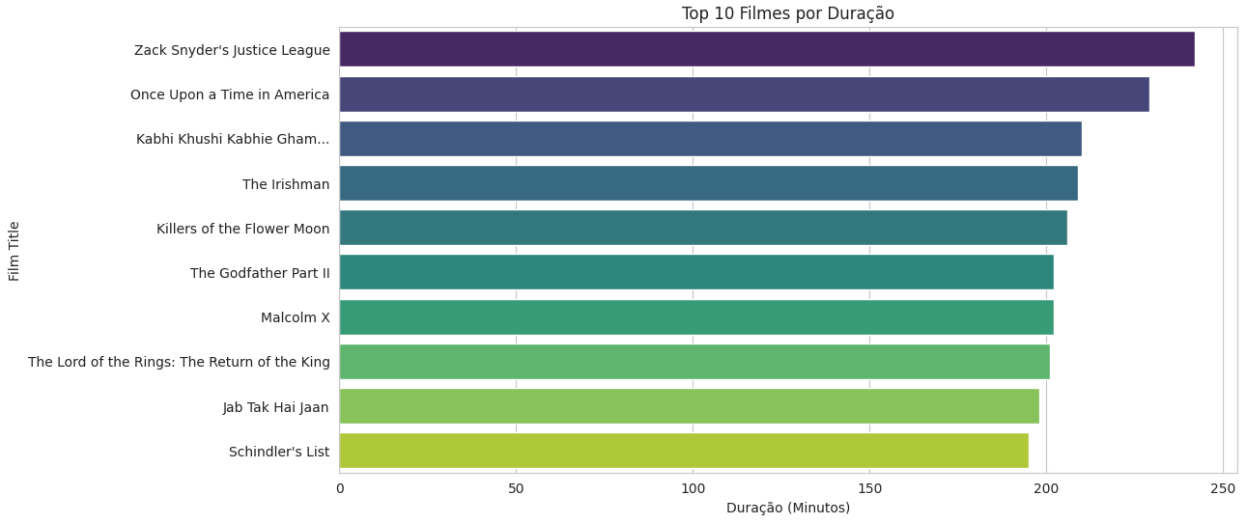
No geral, são filmes com baixo orçamento mas que foram amplamente aclamados pela crítica e tiveram excelente recepção do público, e por isso trouxeram enormes retornos do investimento.

Podemos também analisar a performance de filmes de alto orçamento, os chamados ‘BlockBusters’, filmes produzidos por grandes estúdios e que trazem alto retorno financeiro, os BlockBusters com maior retorno são os mostrados a seguir:

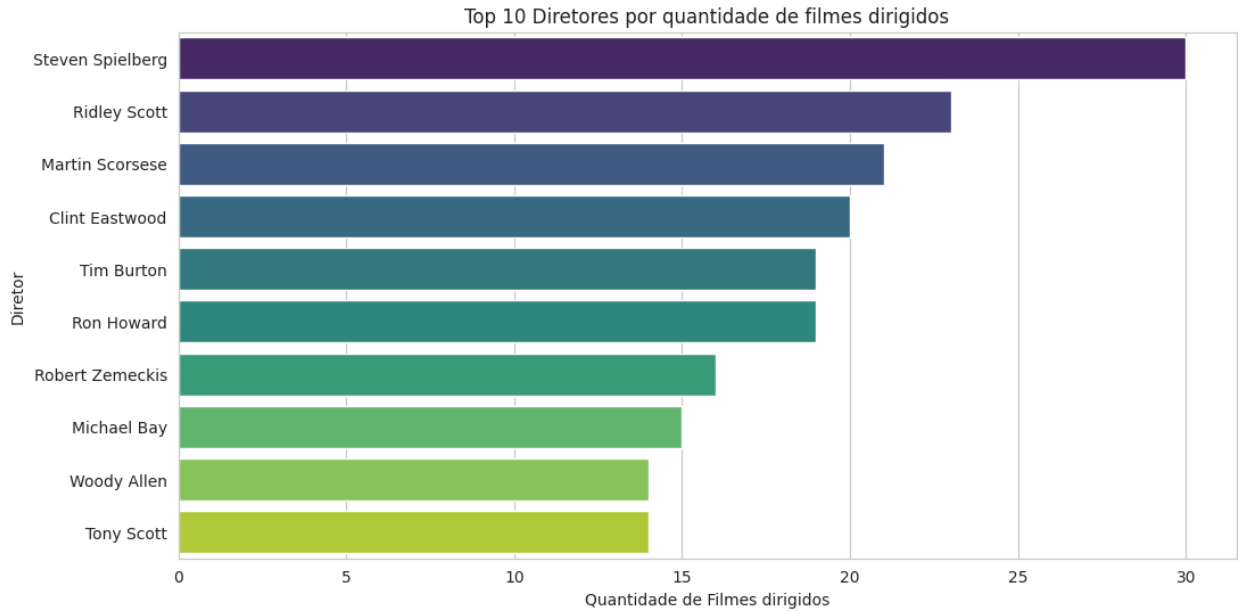


Por fim, também foi feita uma análise sobre os filmes melhor avaliados pelo público, e filmes com maior duração, conforme mostrado a seguir:

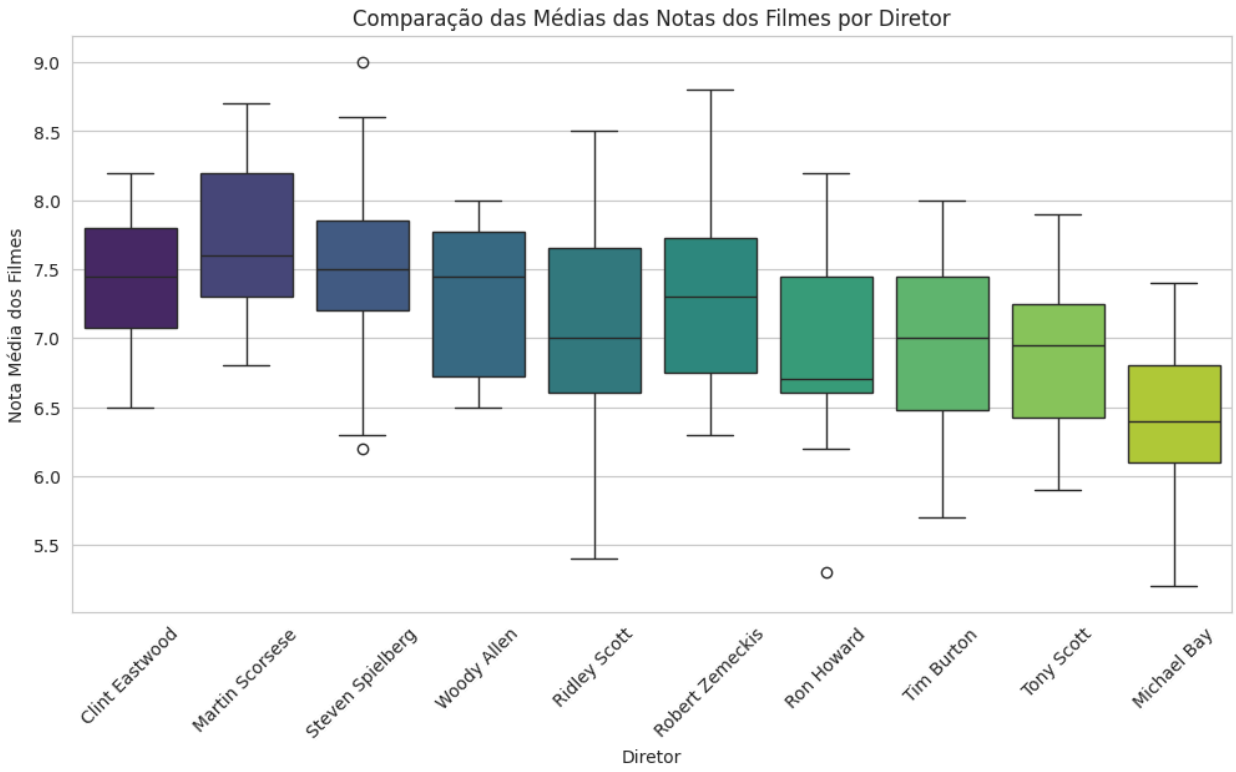




Podemos também avaliar a performance por diretor, para isso vamos considerar o top 10 de diretores com maior quantidade de filmes dirigidos,



e suas respectivas distribuições das notas médias de seus filmes, representadas pelos boxplots no comparativo a seguir:



Conclusão

A análise do conjunto de dados do IMDb proporcionou perspectivas valiosas sobre o desempenho financeiro dos filmes, evidenciando as relações entre orçamento, receita e retorno sobre investimento (ROI), recepção do público, bem como visualizar características importantes de nosso conjunto de dados como a distribuição dos filmes por duração, gênero, recepção do público, performance de diferentes diretores entre outras.

Utilizando ferramentas e bibliotecas do Python, foi possível identificar e tratar variáveis faltantes e outliers, bem como gerar gráficos e figuras que auxiliam no entendimento e distribuição das informações contidas no conjunto de dados escolhido.

É importante ressaltar que, devido à extensão do relatório, algumas análises adicionais não foram incluídas, mas podem ser acessadas no notebook do Google Colab. Essas análises complementares oferecem uma visão mais abrangente sobre os dados, bem como mostram os métodos utilizados para chegar nos resultados apresentados, podendo ser úteis para aprofundar a compreensão dos fatores que influenciam o desempenho dos filmes no mercado.