1. Análise Exploratória dos Dados

Características da Base de Dados

Número de Variáveis: Existem 15 variáveis (colunas) no total.
Número de Entradas: A base contém 999 entradas (linhas), o que significa que há dados de 999 filmes.

a. Tipos de Variáveis
As variáveis podem ser divididas em dois tipos principais: quantitativas e qualitativas.

Variáveis Quantitativas (Numéricas)

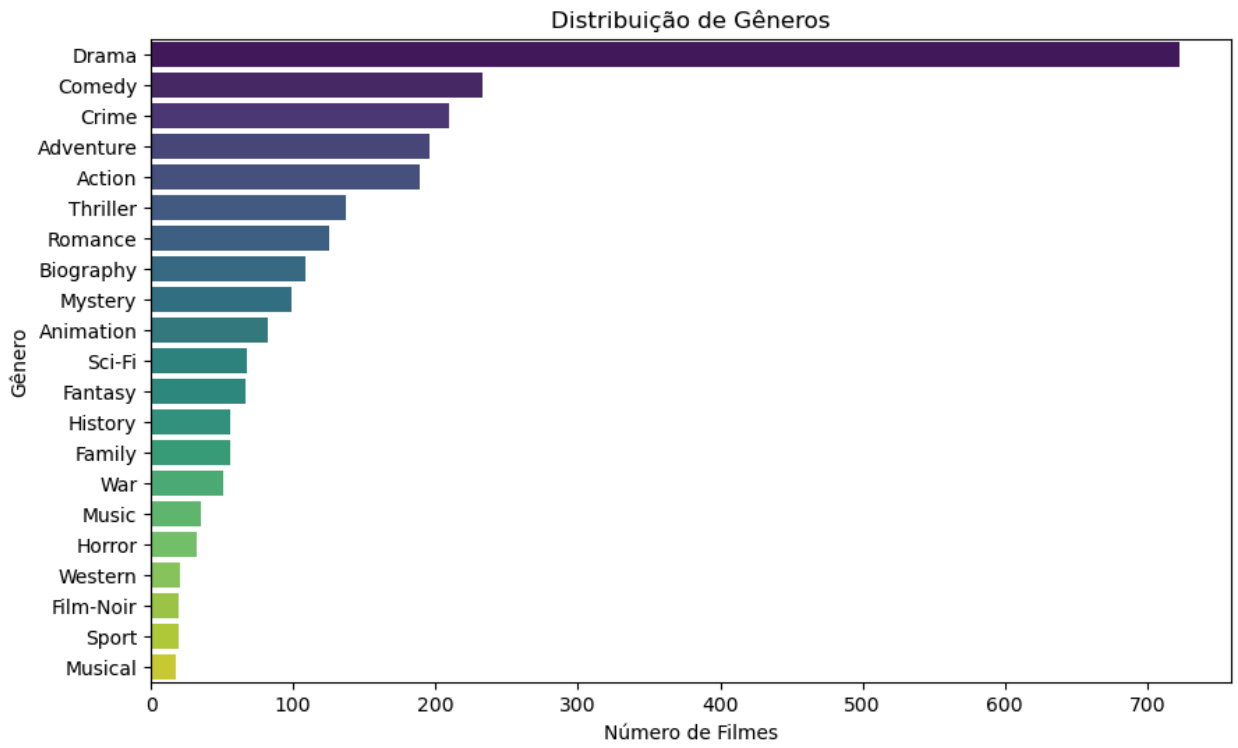
Essas variáveis representam valores numéricos e podem ser usadas para cálculos e análises estatísticas.

- `IMDB_Rating` (nota do IMDB)
- `Meta_score` (nota da crítica)
- `No_of_Votes` (número de votos)
- `Gross` (faturamento bruto)
- `Runtime` (duração do filme em minutos)

Variáveis Qualitativas (Categóricas)

Essas variáveis representam categorias ou rótulos e são usadas para agrupar e descrever os dados.

- `Series_Title` (título do filme)
- `Released_Year` (ano de lançamento)
- `Certificate` (certificação de idade)
- `Genre` (gênero do filme)
- `Overview` (sinopse)
- `Director` (diretor)
- `Star1` (ator principal)
- `Star2` (segundo ator)
- `Star3` (terceiro ator)
- `Star4` (quarto ator)



O gênero mais presente na base de dados é Drama, seguido com Comédia, Crime e Aventura.

Os gêneros menos presentes são Musical, Esportes e Filme-Noir

b. Valores Faltantes (Ausentes)

Dados ausentes são um problema comum e precisam ser tratados para que os modelos e análises geradas a partir dos dados não provoquem erros:

- A coluna Certificate possui 101 valores faltantes.
- A coluna Meta_score possui 157 valores faltantes.
- A coluna Gross possui 169 valores faltantes.

c. Duplicatas

É importante verificar se há filmes duplicados na base. Uma verificação rápida sugere que não há duplicatas de filmes, já que o ID e o Título do filme são únicos.

d. Valores de Outliers

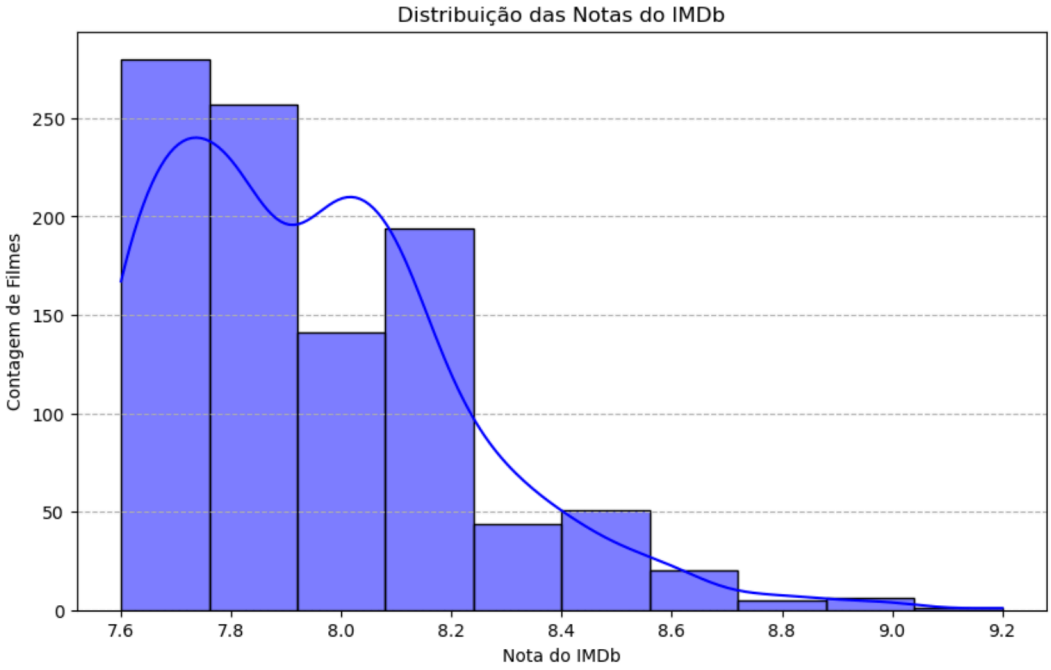
Outliers podem distorcer a análise e o treinamento do modelo. Por exemplo:

- No_of_Votes: Um filme com um número de votos muito superior à média pode influenciar o modelo, principalmente porque a nota do IMDB é uma média ponderada.

- Gross: Filmes com um faturamento excepcionalmente alto podem ser outliers.
-

e. Análise da Variância

- As notas do IMDB_Rating estão concentradas entre 7.6 e 9.2, o que indica que a base de dados inclui apenas filmes bem avaliados, sem notas muito baixas.

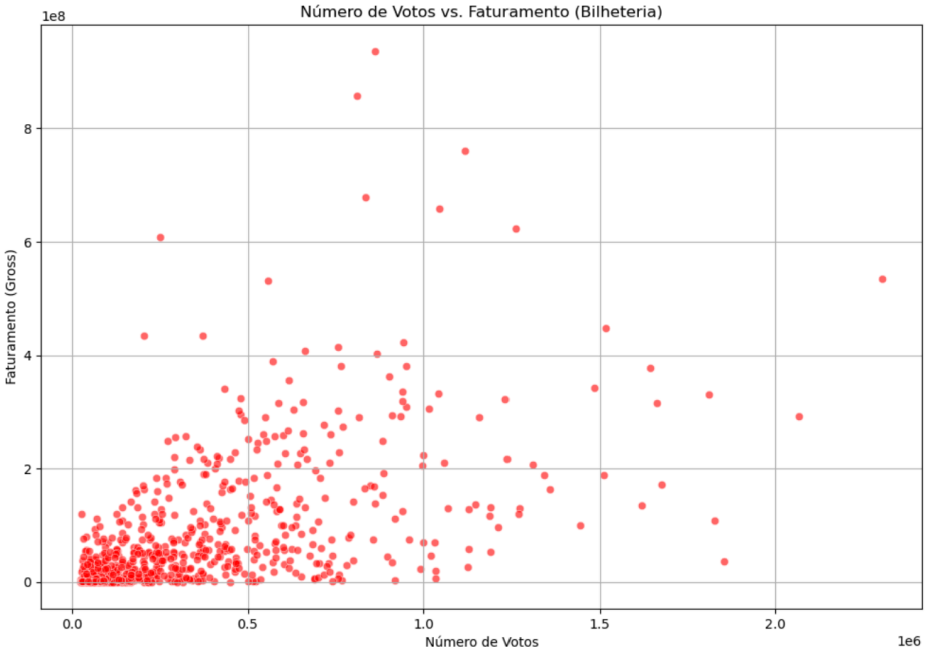


- O Runtime varia de 45 a 321 minutos. A maioria dos filmes tem entre 100 e 150 minutos.

Características entre as variáveis

a. Relação entre Variáveis Numéricas
   i. *IMBD_Rating e Meta_score*
      1. Hipótese: A avaliação do púlico (IMDB_Rating) é similar à avaliação da crítica (Meta_score).
      2. Análise:
   ii. *No_of_Votes e Gross*
      1. Hipótese: Filmes com maior quantidade de votos terão uma arrecadação maior

2. Análise:



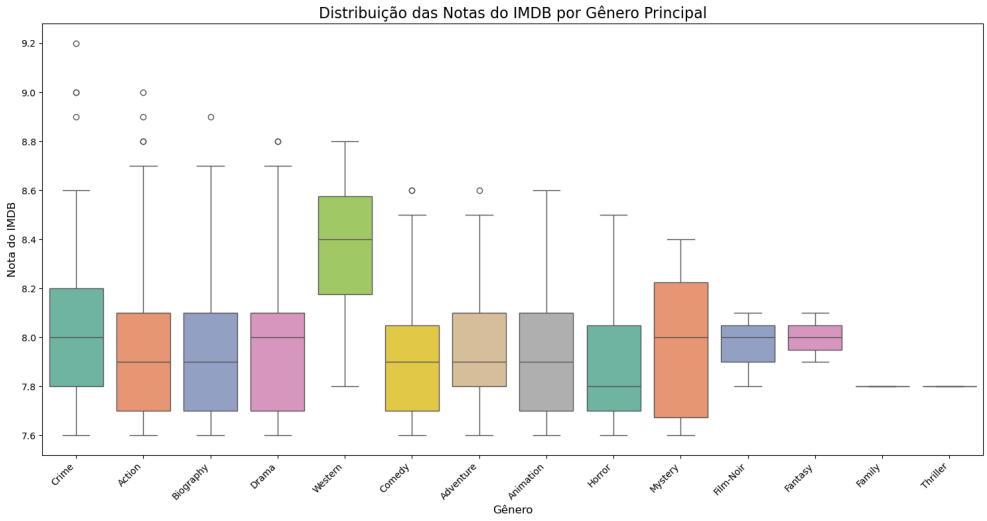
Há uma correlação de 0.62 entre as variáveis, indicando uma relação entre as variáveis, porém não forte o suficiente para tirar outras conclusões

b. Relação entre Variáveis Categóricas e a Nota (IMDB_Rating)

i. IMDB_Rating por Genre:

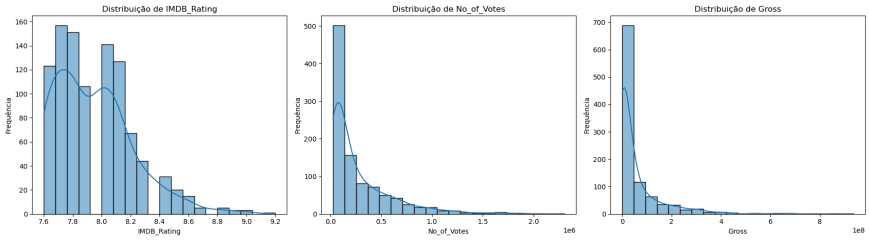
1. Hipótese: A nota média do público muda dependendo do gênero.

2. Análise:



Análise visual do box plot valida a sua hipótese de que o gênero do filme tem uma relação direta com a nota. Os resultados sugerem que Western e Biography são os gêneros com maior probabilidade de ter notas altas.

ii. Histogramas



2. Perguntas

a. Qual filme você recomendaria para uma pessoa que você não conhece?

Podemos abordar a questão de duas maneiras:

1. Fazendo perguntas para entender melhor o que a pessoa busca, como gênero, diretor ou ator favorito. Assim, a recomendação seria personalizada e, dentro da base, poderia-se buscar filmes com maior avaliação dentro dos critérios de resposta.
2. Utilizando apenas os filmes com maiores notas, a partir de um número grande de votos, por exemplo:

   Critérios: Nota maior que 8.5 e mais de 1 milhão de votos.
   Filme Recomendado: The Dark Knight
   Nota: 9.0
   Votos: 2.303.232
   Gênero: Ação, Crime, Drama

   Se o usuário quisesse um filme de outro gênero, como Drama, a recomendação seria The Godfather, com 9.2 de nota e mais de 1.6 milhão de votos.

b. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?

A popularidade (No_of_votes)é a variável mais importante para o faturamento de um filme, de acordo com o modelo utilizado.
A segunda variável mais importante, porém com uma importância bem menor, é o Gênero do filme.

```
 R-quadrado (R²): 0.60
Erro Quadrático Médio (MSE): 3979386579770856.50
```

```
--- Importância das Variáveis (Feature Importance) ---
           Variável  Importância
1       No_of_Votes     0.485071
5     Genre_encoded     0.152317
3  Director_encoded     0.110276
0       IMDB_Rating     0.091543
4     Star1_encoded     0.084739
2        Meta_score     0.076054
```

Aqui, utilizamos Random Forest Regressor uma vez que temos variáveis categóricas e relações não lineares entre as variáveis, usamos o RFR ao in´ves de regressão linear, que comparativamente teve um R^2 maior (0.6 vs 0.4)

c. Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

Por meio de um algoritmo de NLP seria possível inferir o gênero do filme por meio de keywords. Por exemplo, filmes de guerra com sinopses contendo "batalha", "soldado", ou filmes de romance contendo "amor", "relacionamento".

Outros insights incluem a análise de padrões narrativos, como o clássico "Jornada do Herói", "Luta pela sobrevivência", entre outros , ou o impacto da localização ou contexto na nota ou bilheteria (por exemplo, filmes ambientados nos anos 90, ou filmes distópicos)

3. Previsão da nota

Estamos resolvendo um problema de regressão porque a variável que queremos prever (IMDB_Rating) é um valor numérico contínuo, que pode variar de 1.0 a 10.0. A regressão busca encontrar a melhor relação entre as variáveis de entrada e essa variável numérica de saída.

Selecionamos e transformamos as variáveis com maior potencial para prever a nota, conforme nossa análise exploratória:

No_of_Votes: Mantida como variável numérica. É um forte indicador de popularidade, o que tende a se correlacionar com notas mais altas.

Gross: Convertida para numérica, com a remoção de caracteres especiais, e os valores ausentes foram preenchidos com a média. É uma medida de sucesso comercial, que costuma estar ligada à nota.

Runtime: Mantida como variável numérica. A duração do filme pode influenciar a experiência do público.

Meta_score: Mantida como numérica, com valores ausentes preenchidos pela média. A nota dos críticos é um forte indicador da qualidade do filme.

Director e Star1: Essas variáveis categóricas foram transformadas em numéricas usando o OrdinalEncoder. Isso permitiu que o modelo entendesse o impacto do diretor e do ator principal, sem o risco de erros com nomes desconhecidos.

Certificate, Genre e Released_Year: Também foram transformadas com o OrdinalEncoder para que pudessem ser usadas no modelo. A análise de dados mostrou que esses fatores influenciam a nota do filme.

O Random Forest Regressor foi o modelo escolhido porque apresentou o melhor desempenho em relação a Regressão Linear Múltipla. Esse fato pode ser explicado pela existência de variáveis não lineares e categóricas

A principal medida de performance escolhida foi o R-quadrado (R^2), que nos dá uma porcentagem clara do quanto a nossa previsão se aproxima dos dados reais. Além disso, também utilizamos o Erro Quadrático Médio (MSE) para complementar a análise, pois ele quantifica o erro médio das nossas previsões.

Não testamos modelos mais complexos, como Gradient Boosting, por uma restrição de tempo para a entrega do projeto.

4. Nota IMDB para Shawshank Redemption

Usando o modelo RFR, presente no arquivo modelo_previsao_imdb.pkl:
A nota do IMDb prevista para The Shawshank Redemption é: 8.77