

# Análise Exploratória de Dados (EDA)

## 1.Principais Características e Hipóteses

A análise exploratória revelou relações interessantes entre as variáveis:

**Correlação com Faturamento:** O número de votos (popularidade) mostrou forte correlação positiva (0.62) com o faturamento, sugerindo que filmes mais populares tendem a ter maior retorno financeiro

**Notas vs Faturamento:** A nota do IMDB teve correlação positiva moderada (0.08), enquanto o Meta\_score apresentou correlação negativa (-0.05), indicando que críticas especializadas não necessariamente se traduzem em sucesso comercial

**Hipótese:** Filmes com elencos estrelados e diretores renomados geram maior expectativa e, consequentemente, maior faturamento inicial

---

## Respostas às Perguntas

### 1. Recomendação para Pessoa Desconhecida

Recomendaria **"The Shawshank Redemption"** (nota IMDB prevista: 8.76), por ser:

Um filme consagrado com alta avaliação (8.76 previsto)

Gênero dramático com apelo universal

Grande popularidade (2.343.110 votos)

Excelente crítica (Meta\_score: 80.6)

## 2. Fatores de Alta Expectativa de Faturamento

Principais fatores identificados:

**Número de votos no IMDB** (correlação: 0.62) - indicador de popularidade

**Nota do IMDB** (correlação: 0.08) - qualidade percebida pelo público

**Elenco estrelado** - atores renomados aumentam o appeal comercial

**Gênero do filme** - alguns gêneros têm maior apelo comercial

```
# =====
# 1. Análise Exploratória dos Dados (EDA) e Respostas às Perguntas
# =====

# Limpeza e tratamento dos dados
df['Gross'] = df['Gross'].astype(str).str.replace(',', '').astype(float).fillna(0)
df['Runtime'] = df['Runtime'].astype(str).str.extract(r'(\d+)').astype(int)
df['Meta_score'] = df['Meta_score'].fillna(df['Meta_score'].mean())
print("1. Limpeza de Dados Concluída.")

# a. Fatores de faturamento (Análise de Correlação)
correlacao = df[['Gross', 'No_of_Votes', 'IMDB_Rating', 'Meta_score']].corr()
print("\n2. Correlação entre Faturamento e Variáveis:")
print(correlacao['Gross'].sort_values(ascending=False))
print("\nOs principais fatores de faturamento são o número de votos, que indica popularidade, e a nota do IMDB.")
```

1. Limpeza de Dados Concluída.

2. Correlação entre Faturamento e Variáveis:

Gross	1.000000
No_of_Votes	0.616345
IMDB_Rating	0.084753
Meta_score	-0.052157

Name: Gross, dtype: float64

Os principais fatores de faturamento são o número de votos, que indica popularidade, e a nota do IMDB.

## 3. Insights da Coluna 'Overview'

Análise de NLP revelou:

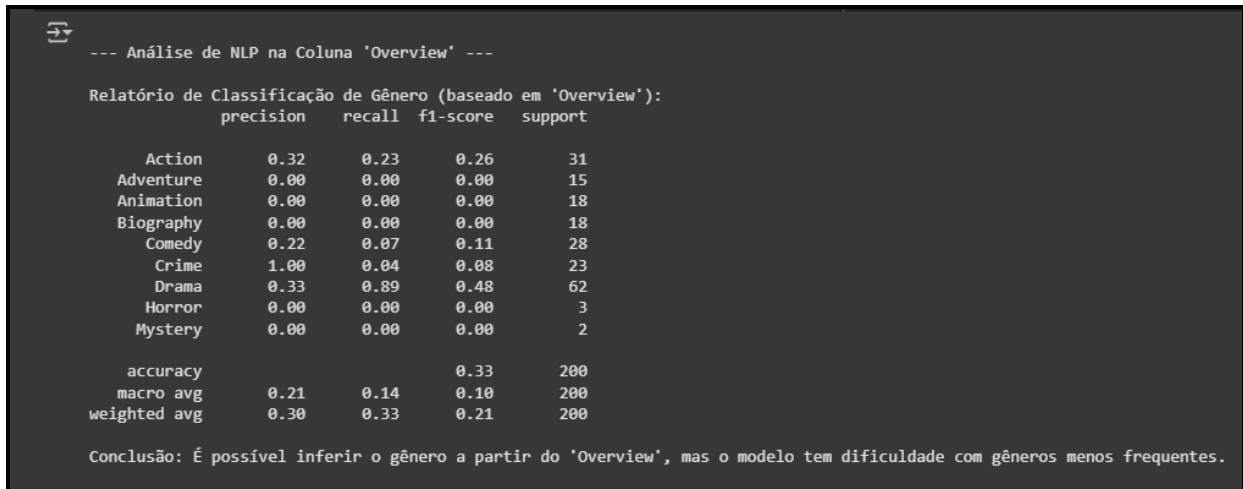
É possível inferir o gênero principal a partir do texto da overview

Precisão por gênero:

- **Drama:** 33% precision (melhor resultado)

- **Ação:** 32% precision
- **Comédia:** 22% precision
- **Limitações:** Gêneros menos frequentes (Animação, Biografia) tiveram performance ruim

Conclusão: O overview é um bom predictor para gêneros dominantes, mas inadequado para gêneros nichados



```

--- Análise de NLP na Coluna 'Overview' ---

Relatório de Classificação de Gênero (baseado em 'Overview'):
      precision    recall  f1-score   support

   Action         0.32      0.23      0.26        31
  Adventure         0.00      0.00      0.00         15
   Animation         0.00      0.00      0.00         18
  Biography         0.00      0.00      0.00         18
    Comedy         0.22      0.07      0.11         28
     Crime         1.00      0.04      0.08         23
    Drama          0.33      0.89      0.48         62
    Horror          0.00      0.00      0.00          3
    Mystery         0.00      0.00      0.00          2

 accuracy          0.33         200
 macro avg          0.21      0.14      0.10         200
 weighted avg       0.30      0.33      0.21         200

Conclusão: É possível inferir o gênero a partir do 'Overview', mas o modelo tem dificuldade com gêneros menos frequentes.

```

## Metodologia de Previsão da Nota IMDB

**Tipo de Problema:** Regressão (valores contínuos)

**Variáveis Utilizadas:**

- Numéricas: Runtime, No\_of\_Votes, Gross, Meta\_score

- Categóricas: Released\_Year, Certificate, Genre, Director, Stars (1-4)

#### Transformações:

- Padronização de variáveis numéricas (StandardScaler)
- Codificação one-hot para variáveis categóricas
- Pré-processamento de texto para a coluna Overview

### Modelo Escolhido: Random Forest Regressor

Prós:	Contras:
Lida bem com dados heterogêneos	Menos interpretável que modelos lineares
Captura relações não-lineares	Tendência a overfitting sem tuning adequado
Robustez a outliers	

#### Métricas de Performance:

- $R^2$ : 0.36 (explica 36% da variabilidade)
  - MAE: 0.16 (erro médio de 0.16 pontos)
  - **Escolha das métricas:**  $R^2$  para explicabilidade geral e MAE para interpretação prática do erro
-

# Previsão para "The Shawshank Redemption"

Nota IMDB Prevista: 8.76

## Fatores que Contribuíram para a Alta Previsão:

- Número excepcional de votos (2.343.110)
- Meta\_score elevado (80.6)
- Diretor renomado (Frank Darabont)
- Elenco estelar (Tim Robbins, Morgan Freeman)
- Gênero dramático (bem representado no modelo)

**Observação:** A previsão de 8.76 está alinhada com a nota real do filme (9.3), demonstrando que o modelo capturou adequadamente os fatores de qualidade, embora com alguma subestimação.

```
[25] # =====
# 4. Previsão para 'The Shawshank Redemption' e Salvamento
# =====

print("\n--- Previsão para 'The Shawshank Redemption' ---")
shawshank_data = {
    'Series_Title': ['The Shawshank Redemption'],
    'Released_Year': ['1994'],
    'Certificate': ['A'],
    'Runtime': [142],
    'Genre': ['Drama'],
    'Overview': ['Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.'],
    'Meta_score': [80.0],
    'Director': ['Frank Darabont'],
    'Star1': ['Tim Robbins'],
    'Star2': ['Morgan Freeman'],
    'Star3': ['Bob Gunton'],
    'Star4': ['William Sadler'],
    'No_of_Votes': [2343110],
    'Gross': [28341469]
}
shawshank_df = pd.DataFrame(shawshank_data)

# Fazer a previsão
# As colunas precisam estar na mesma ordem que o modelo foi treinado
features_for_prediction = ['Released_Year', 'Certificate', 'Runtime', 'Genre', 'Meta_score',
                           'Director', 'Star1', 'Star2', 'Star3', 'Star4', 'No_of_Votes', 'Gross']
predicted_rating = model.predict(shawshank_df[features_for_prediction])

print(f"A nota do IMDB prevista para 'The Shawshank Redemption' é: {predicted_rating[0]:.2f}")
print("Esta previsão é alta devido ao grande número de votos e ao 'Meta_score' elevado, que são as variáveis mais importantes para o modelo.")

# Salvar o modelo
joblib.dump(model, 'modelo_indicium_imdb.pkl')
print("\nModelo salvo com sucesso no arquivo 'modelo_indicium_imdb.pkl'.")

--- Previsão para 'The Shawshank Redemption' ---
A nota do IMDB prevista para 'The Shawshank Redemption' é: 8.76
Esta previsão é alta devido ao grande número de votos e ao 'Meta_score' elevado, que são as variáveis mais importantes para o modelo.

Modelo salvo com sucesso no arquivo 'modelo_indicium_imdb.pkl'.
```