Nome: Beatriz Conceição da Costa
Email: beatrizdacosta1@gmail.com

Relatórios das análises estatísticas e EDA

Estatísticas descritivas:

	Released_Year	Runtime	IMDB_Rating	Meta_score	No_of_Votes
count	998.000000	998.000000	998.000000	998.000000	9.980000e+02
mean	1991.214429	122.854709	7.948297	65.704409	2.716239e+05
std	23.308539	28.110078	0.272203	30.595425	3.210735e+05
min	1920.000000	45.000000	7.600000	0.000000	2.508800e+04
25%	1976.000000	103.000000	7.700000	63.000000	5.541675e+04
50%	1999.000000	119.000000	7.900000	76.000000	1.381685e+05
75%	2009.000000	136.750000	8.100000	85.750000	3.735062e+05
max	2020.000000	321.000000	9.200000	100.000000	2.303232e+06

	Gross
count	9.980000e+02
mean	5.644759e+07
std	1.032710e+08
min	0.000000e+00
25%	4.387472e+05
50%	1.065580e+07
75%	6.144663e+07
max	9.366622e+08

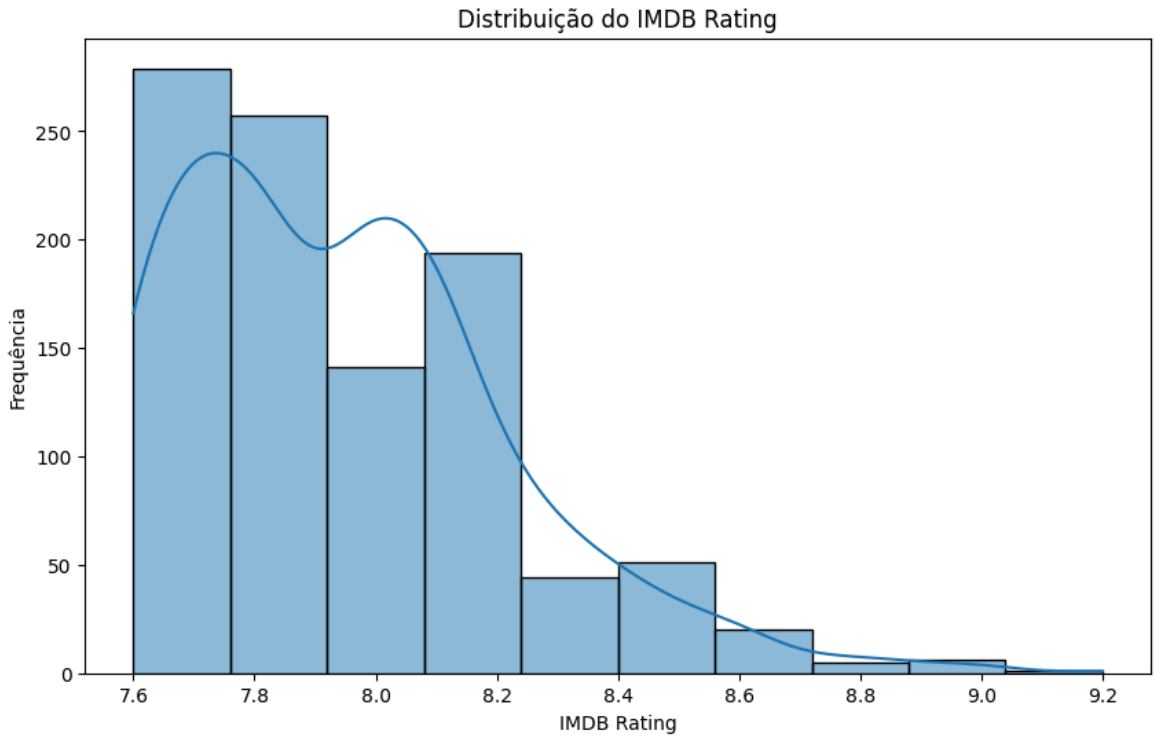

Gráfico 1 - Frequência da coluna Nota do IMDB desse banco de dados

No gráfico 1, podemos analisar que a maior frequência dos filmes desse dataset (mais de 250) possuem nota do IMDB entre 7.6 e 7.8. O que também pode ser inferido quando analisamos as estatísticas descritivas e 25% (primeiro quartil) dessa coluna corresponde a 7.7.

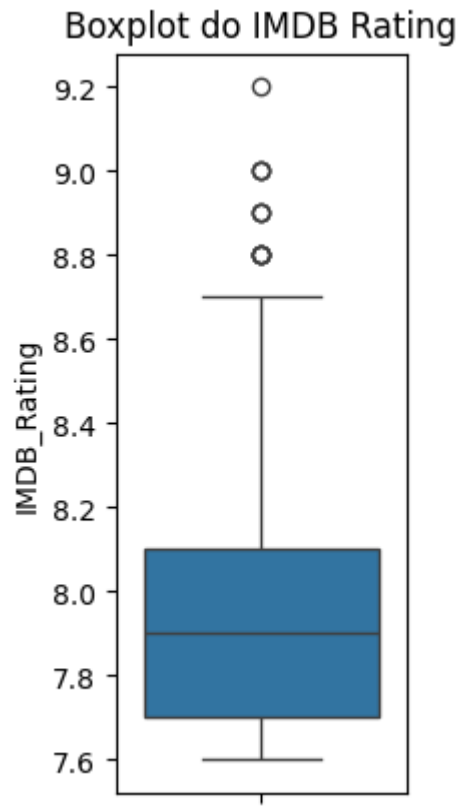

Gráfico 2 - Boxplot da coluna IMDB Rating

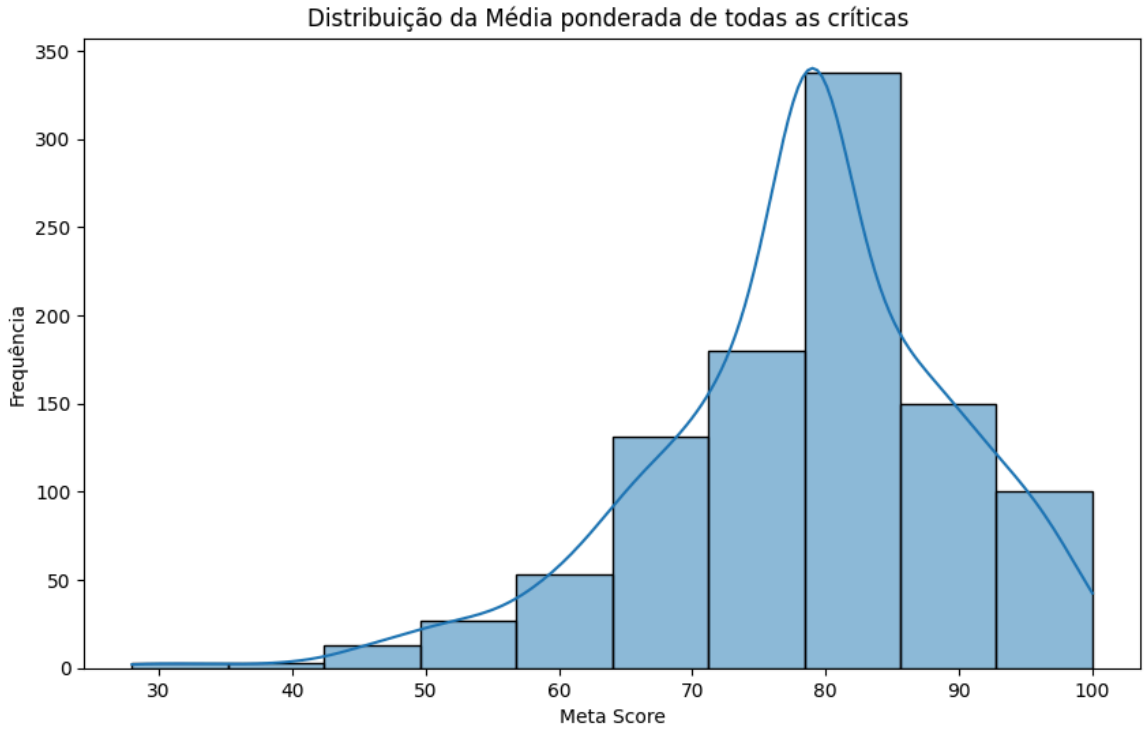

Gráfico 3 - Distribuição da coluna média ponderada desse banco de dados

No gráfico 2, podemos analisar que a maior parte dos filmes desse dataset (em torno de 350) possuem Média ponderada de todas as críticas entre 78 e 83.

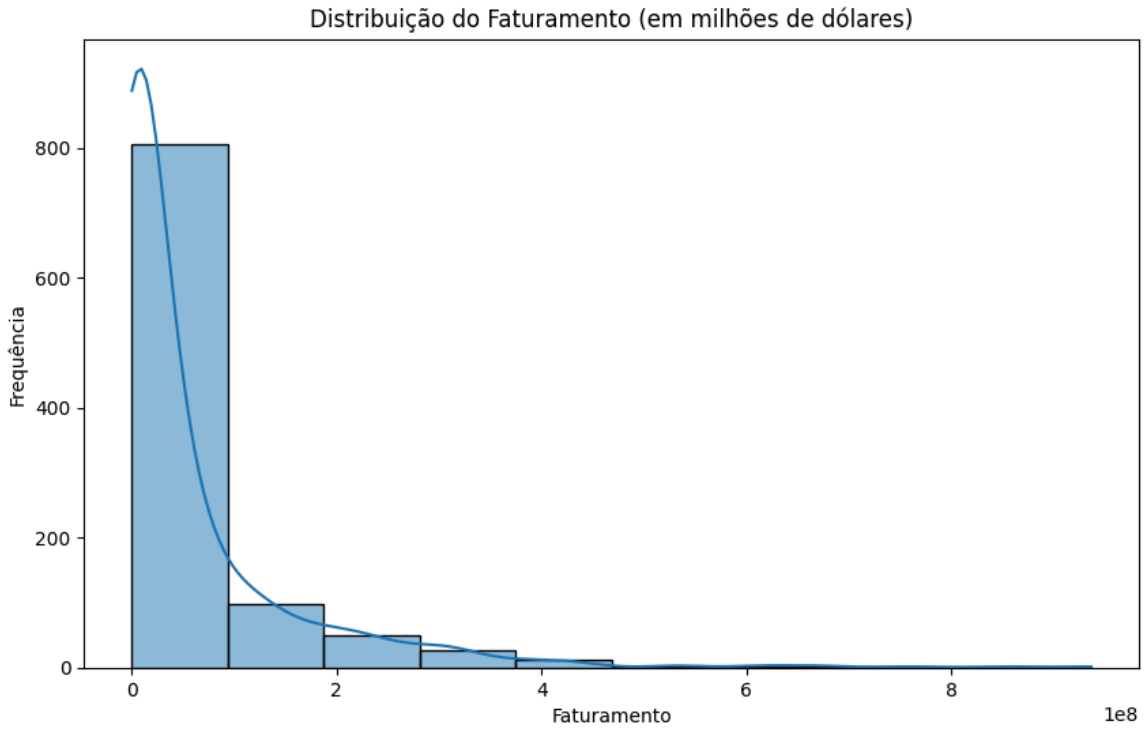

Gráfico 4 - Distribuição da coluna faturamento do banco de dados

No gráfico 3, podemos analisar que a maior parte dos filmes desse dataset (em torno de 800) possuem faturamento de até 1 milhão de dólares e alguns esporádicos em até mais de 8 milhões. Sendo eles:

Filmes com Receita Bruta Acima de 8 Milhões de Dólares:

	Series_Title	Gross
58	Avengers: Endgame	858373000
476	Star Wars: Episode VII - The Force Awakens	936662225

Contagem de cada tipo de certificado:

Certificate	Count
U	233
A	196
UA	175
R	146
X	101
PG-13	43
PG	37
Passed	34
G	12
Approved	11
TV-PG	3

```
      GP       2
   TV-14       1
      16       1
   TV-MA       1
 Unrated       1
     U/A       1
```

O certificado do tipo *X*, foi adicionado por mim para substituir os dados faltantes.

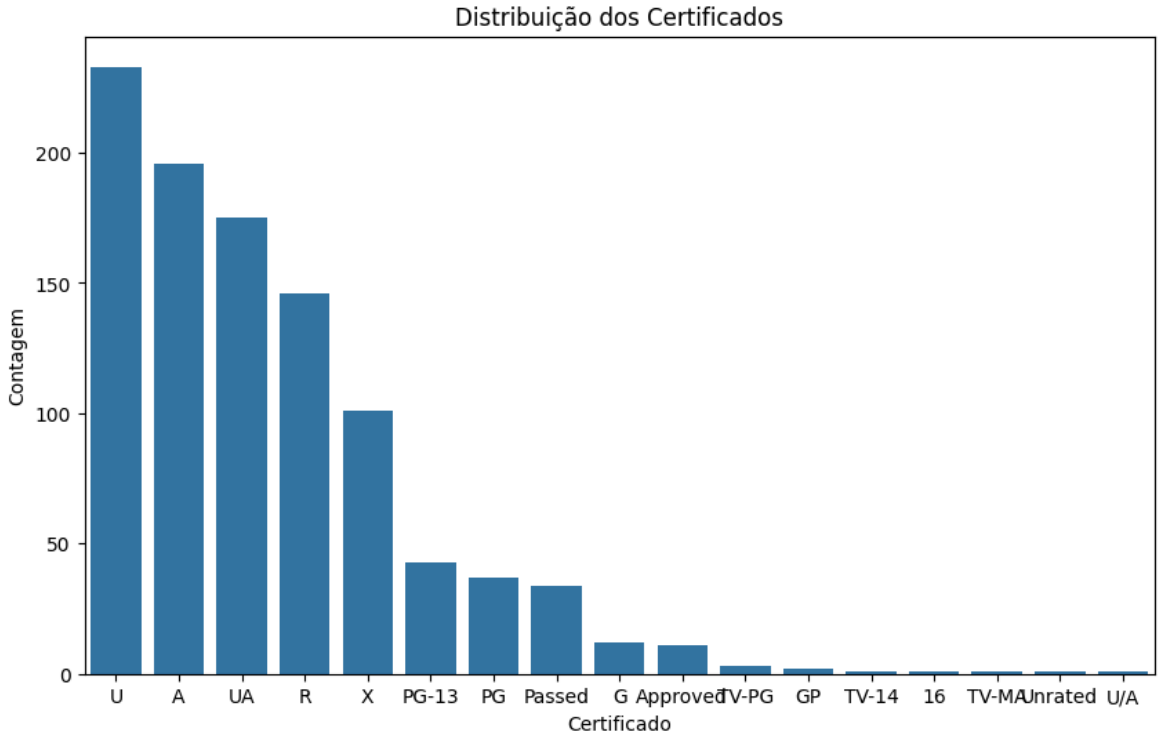

Gráfico 5 - Contagem da coluna certificados do banco de dados

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

Acredito que primeiro indicaria filmes que são clássicos, aqueles que todo mundo deve assistir pelo menos uma vez na vida. Ainda poderia indicar meus filmes favoritos, considerando que não conheço os gêneros preferidos dessa pessoa. E por último ainda poderia indicar os filmes da última década que tem melhores pontuações no IMDB e Rotten Tomatoes.

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

Acredito que os fatores que mais impactam no faturamento de um filme pode ser a abrangência do gênero do filme, ou seja, gêneros mais populares, que atraem maior público como ação, comédia e aventura por exemplo.

Depois as estrelas (atores, atrizes mas também diretores e diretoras) podem influenciar o faturamento dos filmes, por possuírem fãs fieis por exemplo. Ou ainda por terem uma boa reputação.

Exemplo com esse dataset:

```
Atores dos Filmes com Faturamento Acima de 7 Milhões de Dólares:
['Joe Russo' 'Daisy Ridley' 'Sam Worthington' 'Robert Downey Jr.'
 'John Boyega' 'Zoe Saldana' 'Chris Evans' 'Oscar Isaac'
  'Sigourney Weaver' 'Mark Ruffalo' 'Domhnall Gleeson' 'Michelle
Rodriguez']
Número de Filmes em que Cada Ator Está Envolvido:
Joe Russo: 4
Daisy Ridley: 1
Sam Worthington: 2
Robert Downey Jr.: 7
John Boyega: 1
Zoe Saldana: 4
Chris Evans: 6
Oscar Isaac: 2
Sigourney Weaver: 4
Mark Ruffalo: 7
Domhnall Gleeson: 4
Michelle Rodriguez: 1
Selecionando os atores que mais fizeram filmes: Robert Downey Jr., Mark
Ruffalo e Chris Evans.
Filmes com o ator Robert Downey Jr. ordenados por Faturamento:
             Series_Title  Released_Year     Gross
                   Zodiac           2007  33080084
           Sherlock Holmes           2009 209028679
                  Iron Man           2008 318412101
Captain America: Civil War           2016 408084349
              The Avengers           2012 623279547
    Avengers: Infinity War           2018 678815482
         Avengers: Endgame           2019 858373000
Filmes com o ator Mark Ruffalo ordenados por Faturamento:
          Series_Title  Released_Year     Gross
           Dark Waters           2019         0
                Zodiac           2007  33080084
             Spotlight           2015  45055776
```

```
        Shutter Island           2010 128012934
        Thor: Ragnarok           2017 315058289
Avengers: Infinity War           2018 678815482
     Avengers: Endgame           2019 858373000
Filmes com o ator Chris Evans ordenados por Faturamento:
                      Series_Title  Released_Year     Gross
                             Gifted           2017  24801212
                         Knives Out           2019 165359751
Captain America: The Winter Soldier           2014 259766572
          Captain America: Civil War           2016 408084349
                       The Avengers           2012 623279547
                   Avengers: Endgame           2019 858373000
```

Poderíamos supor que ao longo da carreira desses atores o faturamento dos filmes em que eles estão envolvidos vai aumentando, então poderíamos supor que ter atores com uma carreira consolidada produz filmes com mais faturamento. Por outro lado, é importante destacar que esses 3 atores podem levar a um enviesamento pois estão envolvidos em uma grande franquia da Marvel, com altos investimentos na produção e marketing.

Também acredito que o marketing sobre os filmes é um grande decisor sobre questão de faturamento, primeiro pois permite que as pessoas conheçam sobre a história do filme antes mesmo do lançamento por exemplo, enfoque nos artistas participantes, parcerias que possam fortalecer a imagem do filme, ex. de parceiros: alimentos (Coca-Cola, McDonalds, Burger King), passagens aéreas/viagens (American Airlines), smartphones (Samsung), esses parceiros podem lançar produtos com a temática do filme por exemplo. Informações sobre valores de marketing investido poderia ser uma informação útil para análise desse fator.

Referência:

https://www.agenciaslim.com.br/marketing-no-cinema/#:~:text=Como%20o%20cinema%20trabalha%20as%20a%C3%A7%C3%B5es%20de%20marketing,lucraram%20muito%20com%20o%20marketing%20do%20cinema%20 Acesso em: 04/07/2024

Outro fator são os filmes concorrentes na mesma época de lançamento, por exemplo: Barbie x Oppenheimer, os dois filmes em questão lançaram juntos mas nesse caso o marketing acabou sendo positivo para os dois, pois gerou uma certa rivalidade entre os fãs nas redes sociais, o que aumentou (na minha percepção) a

visibilidade dos dois. Mas podem ocorrer casos em que um filme fique ofuscado por produções maiores.

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

Com a coluna Overview é possível descobrir sobre o enredo do filme, supor o gênero e atmosfera e até mesmo o público-alvo. Sim, é possível inferir o gênero do filme a partir dessa coluna, por meio de técnicas de machine learning e análise de linguagem natural. Referências: Movie Genre Prediction Using Multi Label Classification (analyticsvidhya.com) Acesso em: 04/07/2024

Explique como você faria a previsão da **nota do imdb** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

O primeiro passo seria selecionar as variáveis preditoras, eu plotei um gráfico de correlação entre algumas variáveis e a nota do IMDB, como podemos visualizar a correlação foi baixa. Mas selecionei algumas como: `'Released_Year'`, `'Runtime'`, `'Meta_score'`, `'Gross'`. Depois separei os dados em treinamento e teste: `X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)`.

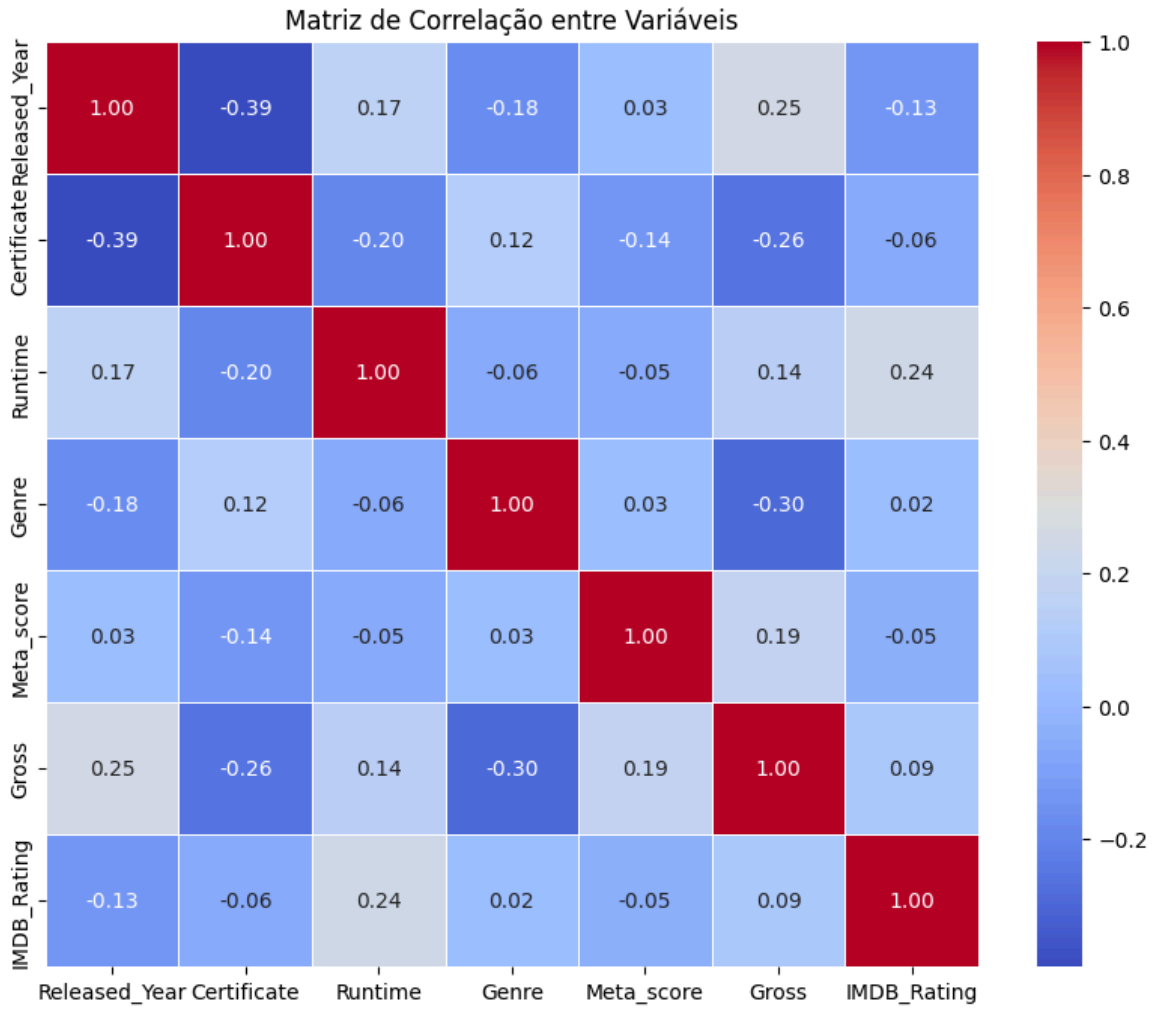

Gráfico 6 - Gráfico da correlação entre as variáveis

Escolhi utilizar nesse primeiro momento uma regressão linear, pois é mais simples e de fácil interpretação. Modelos lineares podem capturar relações lineares entre variáveis preditoras e a variável de resposta (IMDB_Rating). Outros modelos Não-Lineares podem ser úteis para análises futuras visto que a relação entre nota IMDB e as outras variáveis não está clara,como: modelos como árvores de decisão, random forests, ou redes neurais.

Nesse momento escolhi o Erro Quadrático Médio (MSE) para medir a performance do modelo de regressão. MSE penaliza erros grandes mais severamente, o que é importante quando estamos interessados em prever valores numéricos precisos como a nota do IMDb.

1. Supondo um filme com as seguintes características:

{'Series_Title': 'The Shawshank Redemption',
 'Released_Year': '1994',
 'Certificate': 'A',
 'Runtime': '142 min',
 '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': '28,341,469'}

Qual seria a nota do IMDB?

```
Mean Squared Error (MSE): 0.05838240952791825
Predicted IMDb Rating: 7.978811404253857
```