

## ✓ Desafio Cientista de Dados

# Análise exploratória de dados e Machine Learning

Relatório feito por: Ian Périgo

[LinkedIn](#) | [GitHub](#) | [WhatsApp](#)

[Repositório do código](#)

## Objetivo:

Fazer uma análise de um banco de dados cinematográfico para orientar qual tipo de filme deverá ser o próximo a ser produzido pelo estúdio.

## Entregas:

- 1- Análise exploratória dos dados, demonstrar principais características e apresentar algumas hipóteses
- 2- Responder as seguintes perguntas:
  - A- Qual filme seria recomendado para uma pessoa que não conheço
  - B- Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme ?
  - C- Quais insights podem ser tirados com a coluna Overview? é possível inferir o gênero do filme a partir dela?
- 3- Fazer a previsão da nota do IMDB a partir dos dados
  - A- Quais variáveis e transformações foram utilizadas
  - B- Qual tipo de problema está sendo resolvido
  - C- Qual modelo melhor se aproximou dos dados, prós e contras
  - D- Qual medida de performance foi escolhida
- 4- Com o modelo de Machine Learning treinado, avaliar qual será a nota do IMDB para um filme com essas 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' }
```

## Mostrar código

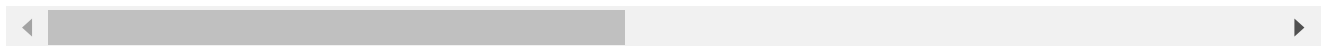
```
↩ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 999 entries, 0 to 998
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0             999 non-null    int64
1   Series_Title           999 non-null    object
2   Released_Year          999 non-null    object
3   Certificate             898 non-null    object
4   Runtime                999 non-null    object
5   Genre                  999 non-null    object
6   IMDB_Rating            999 non-null    float64
7   Overview               999 non-null    object
8   Meta_score             842 non-null    float64
9   Director               999 non-null    object
10  Star1                  999 non-null    object
11  Star2                  999 non-null    object
12  Star3                  999 non-null    object
13  Star4                  999 non-null    object
14  No_of_Votes            999 non-null    int64
15  Gross                  830 non-null    object
dtypes: float64(2), int64(2), object(12)
memory usage: 125.0+ KB
```

```
df.head()
```



Unnamed:  
0

		Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating
0	1	The Godfather	1972	A	175 min	Crime, Drama	9.2
1	2	The Dark Knight	2008	UA	152 min	Action, Crime, Drama	9.0
2	3	The Godfather: Part II	1974	A	202 min	Crime, Drama	9.0
3	4	12 Angry Men	1957	U	96 min	Crime, Drama	9.0
4	5	The Lord of the Rings: The Return of the King	2003	U	201 min	Action, Adventure, Drama	8.9



Primeiras observações:

Nosso dataset contém mais features tipo object.

Há alguns valores nulos em Gross, Meta\_score e Certificate.

As features Gross, Runtime e Released\_Year são objects porém podem ser convertidas para int e float com mais facilidade.

As demais features precisarão de outras abordagens.

[Mostrar código](#)



```
[ '1972' '2008' '1974' '1957' '2003' '1994' '1993' '2010' '1999' '2001'
  '1966' '2002' '1990' '1980' '1975' '2020' '2019' '2014' '1998' '1997'
  '1995' '1991' '1977' '1962' '1954' '1946' '2011' '2006' '2000' '1988'
  '1985' '1968' '1960' '1942' '1936' '1931' '2018' '2017' '2016' '2012' ]
```

```
'2009' '2007' '1984' '1981' '1979' '1971' '1963' '1964' '1950' '1940'
'2013' '2005' '2004' '1992' '1987' '1986' '1983' '1976' '1973' '1965'
'1959' '1958' '1952' '1948' '1944' '1941' '1927' '1921' '2015' '1996'
'1989' '1978' '1961' '1955' '1953' '1925' '1924' '1982' '1967' '1951'
'1949' '1939' '1937' '1934' '1928' '1926' '1920' '1970' '1969' '1956'
'1947' '1945' '1930' '1938' '1935' '1933' '1932' '1922' '1943' 'PG']
```

```
#Vamos analisar as porcentagens de valores nulos
```

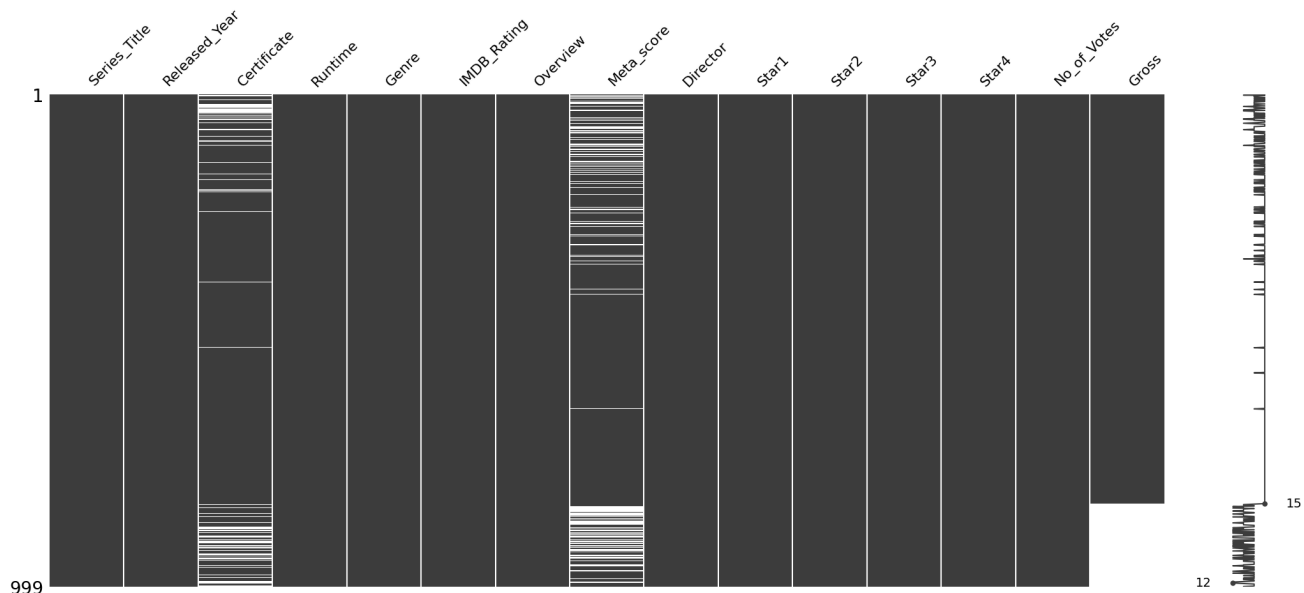
```
df_null_percent = df.isnull().sum() / len(df) * 100
df_null_percent
```

```
Series_Title      0.000000
Released_Year      0.000000
Certificate       10.110110
Runtime           0.000000
Genre             0.000000
IMDB_Rating       0.000000
Overview          0.000000
Meta_score        15.715716
Director          0.000000
Star1             0.000000
Star2             0.000000
Star3             0.000000
Star4             0.000000
No_of_Votes       0.000000
Gross            16.916917
dtype: float64
```

[Mostrar código](#)



<Axes: >



### Mostrar código



Número de linhas com valores faltantes em Gross, Meta\_score e Certificate: 286  
Porcentagem de linhas com valores faltantes em Gross, Meta\_score e Certificate: 28.62

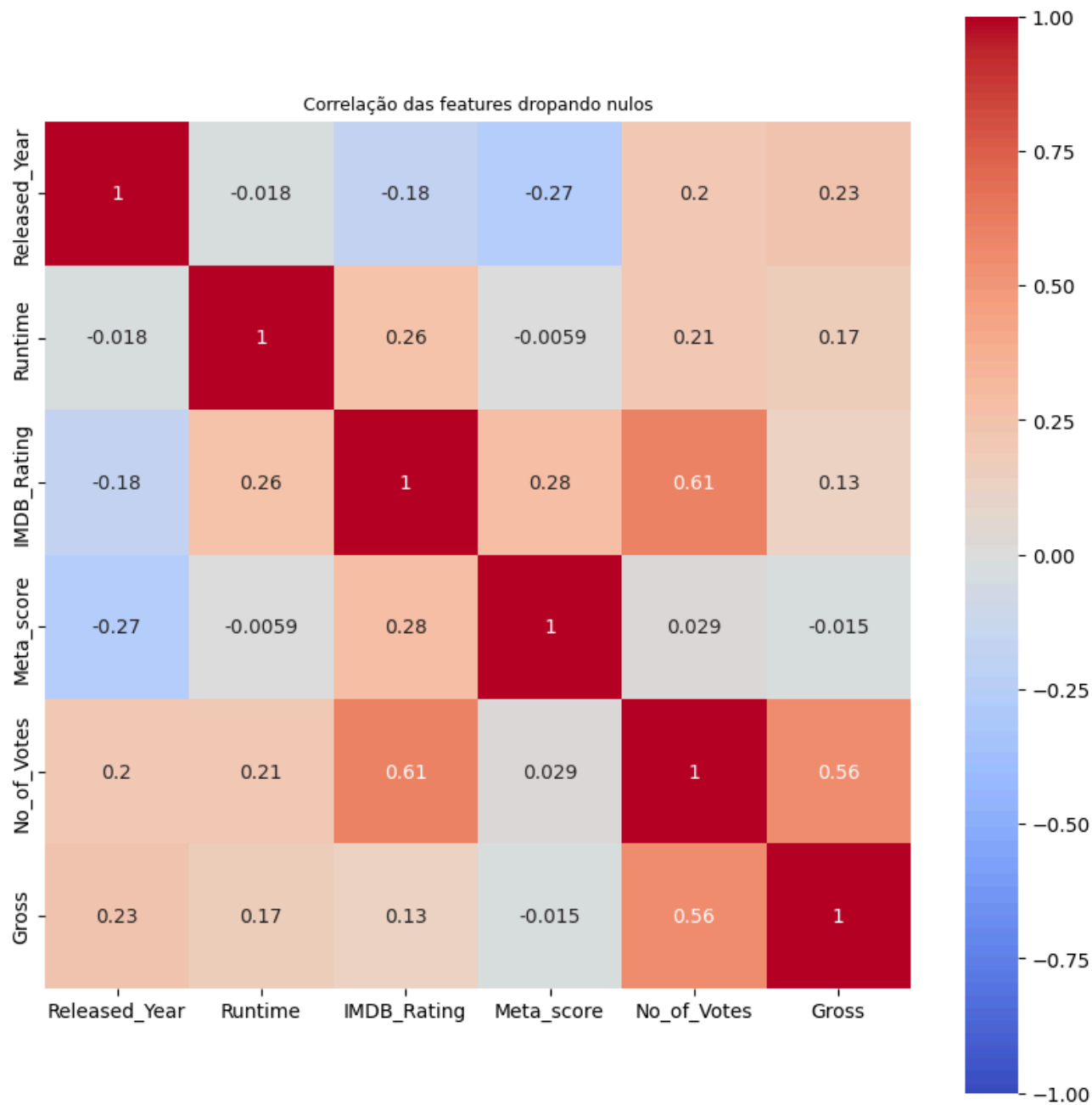


Notamos uma forte correlação dos dados faltantes na feature Gross, certificate e Meta\_score. Temos 28% de valores de ao menos um valor faltante. Vamos verificar qual será a melhor abordagem, dropar ou fazer input data.

Vamos dividir o data set em numéricos e categoricos para começar algumas manipulações dos dados para ter fazer uma análise exploratória.

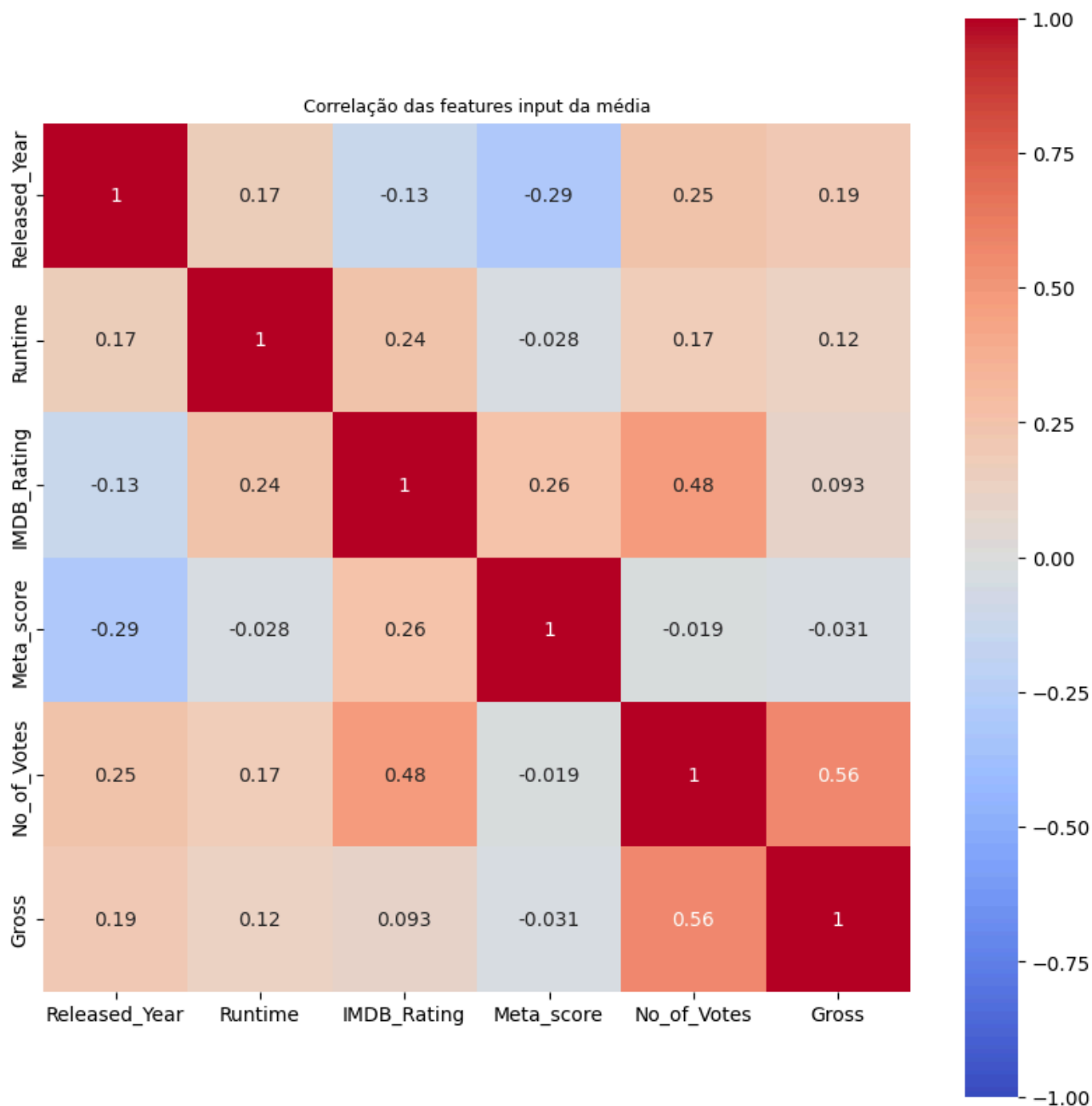
[Mostrar código](#)

```
<Axes: title={'center': 'Correlação das features dropando nulos'}>
```



[Mostrar código](#)

➡ <Axes: title={'center': 'Correlação das features input da média'}>



Notamos que os número de votos e o ano de lançamento apresentam uma maior correlação com o faturamento do filme

Dropando os nulos:

tivemos uma maior correlação do IMDB\_Rating com Gross, comparado aos outros métodos de replace.

Tivemos Número de votos, Meta Score e Runtime com maior correlação com IMDB\_Rating que será nosso alvo.

Métodos de Replace:

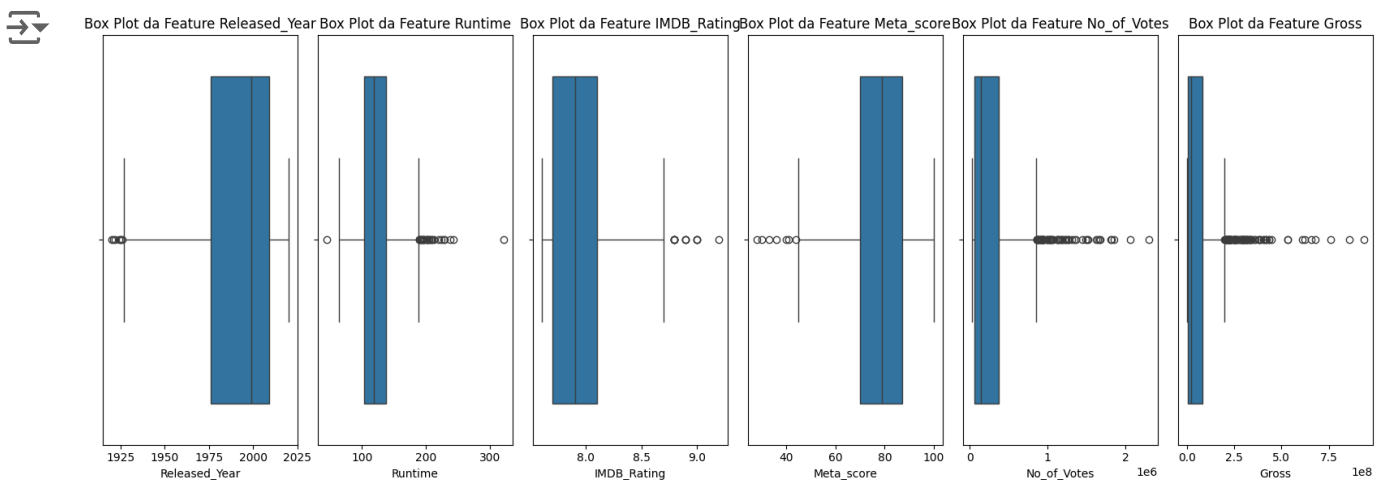
Obtivemos melhores correlações pela mediana comprado a média.

porém dropando os nulos obtivemos correlações mais alta.

Vamos analisar como estão a distribuição gerando um boxplot.

Com isso também queremos analisar a presença de outliers e investigar se precisaremos dropar

[Mostrar código](#)



Considerações

Released\_Year:

A maioria dos filmes foi lançada entre aproximadamente 1970 e 2015. Existem alguns outliers, especialmente filmes lançados antes de 1970. Runtime:

A maioria dos filmes tem uma duração entre aproximadamente 100 e 200 minutos. Existem outliers para filmes com durações muito curtas e muito longas.



IMDB\_Rating:

As avaliações IMDb estão concentradas entre 7.5 e 8.5. Existem alguns outliers com avaliações acima de 8.5.

Meta\_score:

As pontuações Meta estão concentradas entre 50 e 80. Existem vários outliers com pontuações abaixo de 50 e acima de 80.

No\_of\_Votes:

A maioria dos filmes tem até 1 milhão de votos. Existem muitos outliers com mais de 1 milhão de votos.

Gross:

A maioria dos filmes tem uma receita bruta de até 250 milhões. Existem outliers com receitas brutas significativamente maiores, chegando a até 750 milhões.

As variáveis No\_of\_Votes e Gross têm uma escala significativamente diferente das outras variáveis, o que pode exigir normalização ou padronização

[Mostrar código](#)

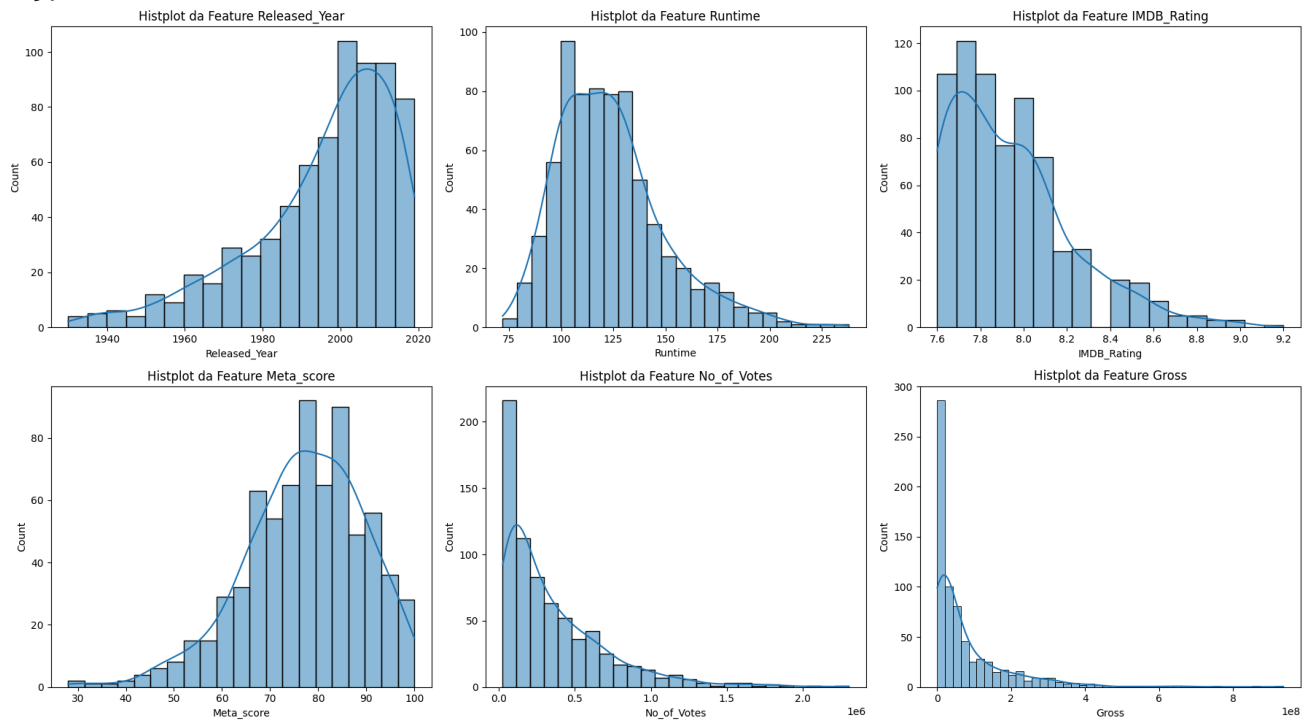


Skewness:

Released\_Year -1.144282  
Runtime 1.009530  
IMDB\_Rating 1.114826  
Meta\_score -0.583372  
No\_of\_Votes 1.819175  
Gross 2.916618  
dtype: float64

Kurtosis:

Released\_Year 0.917191  
Runtime 1.355109  
IMDB\_Rating 1.252506  
Meta\_score 0.476433  
No\_of\_Votes 4.202182  
Gross 12.108811  
dtype: float64



Vamos analisar a caudose e curtose com os valores nulos inputados pela mediana

[Mostrar código](#)

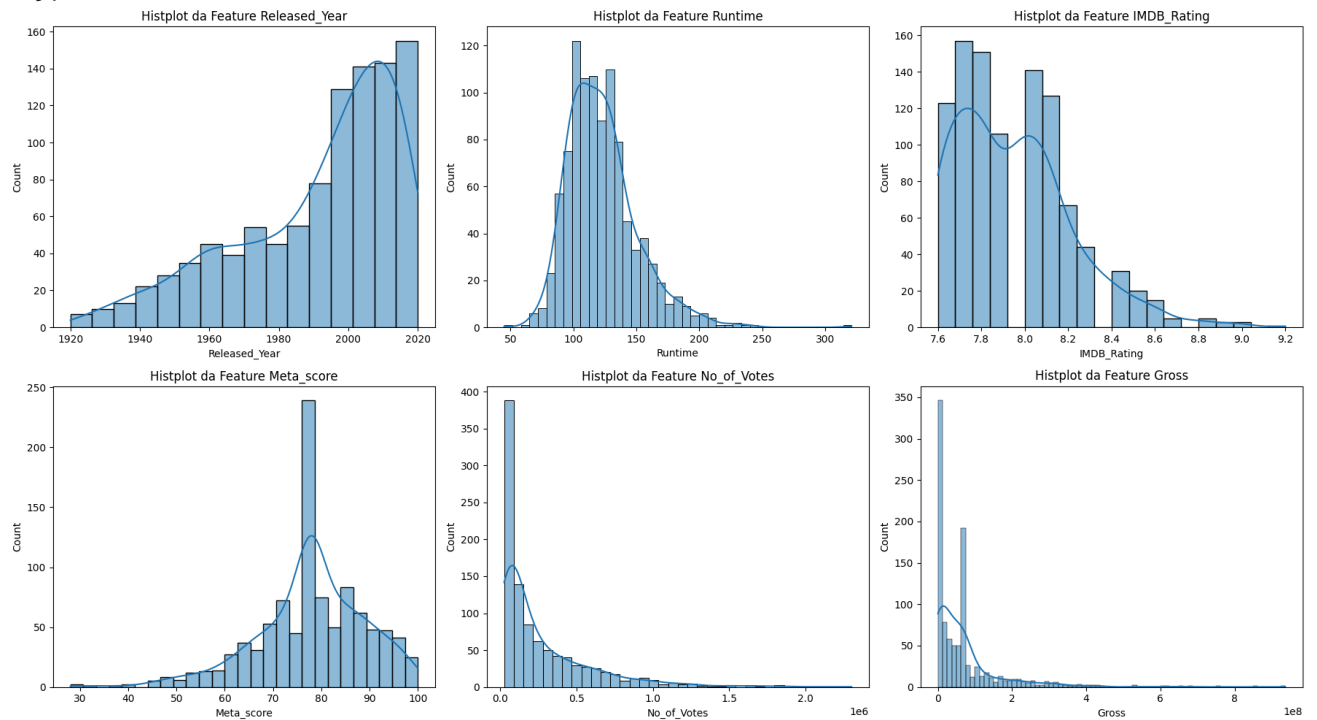


### Skewness:

Released\_Year -0.938045  
Runtime 1.208060  
IMDB\_Rating 0.945271  
Meta\_score -0.657077  
No\_of\_Votes 2.191055  
Gross 3.425225  
dtype: float64

### Kurtosis:

Released\_Year -0.027235  
Runtime 3.405769  
IMDB\_Rating 1.047107  
Meta\_score 1.042189  
No\_of\_Votes 6.005129  
Gross 17.224666  
dtype: float64



Released\_Year: Apresenta assimetria negativa moderada e kurtosis próxima de zero, indicando uma distribuição levemente distorcida para a esquerda, com caudas normais.

Runtime: Alta assimetria positiva e kurtosis acima de 3 indicam uma distribuição fortemente distorcida para a direita, com algumas caudas pesadas.

IMDB\_Rating: Assimetria positiva moderada e kurtosis acima de 1, sugerindo uma leve distorção para a direita e algumas caudas mais pesadas.

Meta\_score: Moderada assimetria negativa e kurtosis levemente acima de 1, indicando uma distribuição levemente distorcida para a esquerda com algumas caudas mais pesadas.

No\_of\_Votes: Alta assimetria positiva e kurtosis muito elevada indicam uma distribuição muito distorcida para a direita com muitas caudas pesadas.

Gross: Muito alta assimetria positiva e kurtosis extremamente elevada indicam uma distribuição extremamente distorcida para a direita com muitas caudas pesadas.

### Comparação Entre Métodos de Imputação

Os valores de skewness e kurtosis são idênticos para ambas as técnicas de imputação (média e mediana), sugerindo que a escolha entre média e mediana não afetou a distribuição das variáveis. No entanto, é importante considerar as correlações com a variável alvo para determinar qual método é mais apropriado para a modelagem.

### Escolha do Método de Imputação para Modelo de Regressão

Resistência a Outliers: A mediana não é influenciada por valores extremos, o que é especialmente relevante para variáveis como Gross e No\_of\_Votes, que apresentam alta assimetria positiva e kurtosis elevada.

Vamos verificar como será o desempenho de ML com sem imputação, média e mediana na regressão.

### Mostrar código



```
Fitting 3 folds for each of 50 candidates, totalling 150 fits
Melhores parâmetros encontrados: {'subsample': 0.5, 'reg_lambda': 1.0, 'reg_alpha': 0
Mean Squared Error on teste: 0.033854938232578895
R² Score on teste: 0.5270660509905415
Mean Absolute Error teste: 0.1472445547997535
MAPE: 1.85%
Accuracy: 98.15 %
```

### Mostrar código

⇒ Fitting 3 folds for each of 50 candidates, totalling 150 fits  
Melhores parâmetros encontrados: {'subsample': 0.5, 'reg\_lambda': 1.0, 'reg\_alpha': 0  
Mean Squared Error on teste: 0.03235306012841006  
R<sup>2</sup> Score on teste: 0.5279940165454902  
Mean Absolute Error teste: 0.1447999739646912  
MAPE: 1.82%  
Accuracy: 98.18 %

### Mostrar código

⇒ Fitting 3 folds for each of 50 candidates, totalling 150 fits  
Melhores parâmetros encontrados: {'subsample': 0.7, 'reg\_lambda': 0.1, 'reg\_alpha': 0  
Mean Squared Error on teste: 0.031893310213657246  
R<sup>2</sup> Score on teste: 0.5347014102138088  
Mean Absolute Error teste: 0.14077048015594484  
MAPE: 1.77%  
Accuracy: 98.23 %

Nosso melhor modelo até agora foi :

Com inputação pela média

Melhores parâmetros encontrados: {'subsample': 0.7, 'reg\_lambda': 0.1, 'reg\_alpha': 0.1,  
'n\_estimators': 200, 'max\_depth': 3, 'learning\_rate': 0.1, 'gamma': 0.1, 'colsample\_bytree': 1.0}

Mean Squared Error on teste: 0.031893310213657246

R<sup>2</sup> Score on teste: 0.5347014102138088

Mean Absolute Error teste: 0.14077048015594484

MAPE: 1.77%

Accuracy: 98.23 %

Palavras mais frequentes para o gênero 'Crime': young: 16 murder: 15 crime: 13 family: 11 man: 11

Palavras mais frequentes para o gênero 'Action': must: 19 young: 15 man: 14 world: 12 former: 12

Palavras mais frequentes para o gênero 'Biography': story: 22 life: 17 man: 10 war: 7 world: 6

Palavras mais frequentes para o gênero 'Drama': life: 42 man: 41 young: 40 woman: 34 love: 32

Palavras mais frequentes para o gênero 'Western': joins: 3 bounty: 2 hunting: 1 scam: 1 men: 1

Palavras mais frequentes para o gênero 'Comedy': young: 24 man: 17 life: 17 love: 16 friends: 14

Palavras mais frequentes para o gênero 'Adventure': world: 10 story: 8 war: 7 find: 7 man: 7

Palavras mais frequentes para o gênero 'Animation': young: 23 girl: 14 world: 13 new: 12 boy: 9

Palavras mais frequentes para o gênero 'Horror': young: 3 becomes: 3 run: 2 mother: 2 mysterious: 2

Palavras mais frequentes para o gênero 'Mystery': man: 3 saskia: 3 wifes: 2 murderer: 2 murder: 2

Palavras mais frequentes para o gênero 'Film-Noir': pulp: 1 novelist: 1 holly: 1 martins: 1 travels: 1

Palavras mais frequentes para o gênero 'Fantasy': hypnotist: 1 dr: 1 caligari: 1 uses: 1  
somnambulist: 1

Palavras mais frequentes para o gênero 'Family': troubled: 1 child: 1 summons: 1 courage: 1 help: 1

Palavras mais frequentes para o gênero 'Thriller': recently: 1 blinded: 1 woman: 1 terrorized: 1  
trio: 1



Mostrar código

```
➡ DataFrame com variáveis dummy para gêneros:
```

	Series_Title	Certificate
0	the godfather	a
1	the dark knight	ua
2	the godfather: part ii	a
3	12 angry men	u
4	the lord of the rings: the return of the king	u
..	...	...
994	breakfast at tiffany's	a
995	giant	g
996	from here to eternity	passed
997	lifeboat	u

```

                                Genre \
0      crime, drama
1      action, crime, drama
2      crime, drama
3      crime, drama
4      action, adventure, drama
..
994    comedy, drama, romance
995      drama, western
996      drama, romance, war
997      drama, war
998    crime, mystery, thriller

```

```

                                Overview                                Director \
0      an organized crime dynasty's aging patriarch t... francis ford coppola
1      when the menace known as the joker wreaks havo... christopher nolan
2      the early life and career of vito corleone in ... francis ford coppola
3      a jury holdout attempts to prevent a miscarria... sidney lumet
4      gandalf and aragorn lead the world of men agai... peter jackson
..
994    a young new york socialite becomes interested ... blake edwards
995    sprawling epic covering the life of a texas ca... george stevens
996    in hawaii in 1941, a private is cruelly punish... fred zinnemann
997    several survivors of a torpedoed merchant ship... alfred hitchcock
998    a man in london tries to help a counter-espion... alfred hitchcock

```

```

                                Star1                                Star2                                Star3                                Star4 \
0      marlon brando                                al pacino                                james caan                                diane keaton
1      christian bale                                heath ledger                                aaron eckhart                                michael caine
2      al pacino                                robert de niro                                robert duvall                                diane keaton
3      henry fonda                                lee j. cobb                                martin balsam                                john fiedler
4      elijah wood                                viggo mortensen                                ian mckellen                                orlando bloom
..
994    audrey hepburn                                george peppard                                patricia neal                                buddy ebsen
995    elizabeth taylor                                rock hudson                                james dean                                carroll baker
996    burt lancaster                                montgomery clift                                deborah kerr                                donna reed
997    tallulah bankhead                                john hodiak                                walter slezak                                william bendix
998    robert donat                                madeleine carroll                                lucie mannheim                                godfrey tearle

```

```

                                action ... horror music musical mystery romance sci-fi sport \
0      0 ... 0 0 0 0 0 0 0
1      1 ... 0 0 0 0 0 0 0
2      0 ... 0 0 0 0 0 0 0
3      0 ... 0 0 0 0 0 0 0

```

Vamos continuar a exploração e tratamentos das features categoricas

vamos verificar a coluna certificate que contém a classificação indicativa dos filmes

Foi realizado uma pesquisa sobre os tipos de Certificates, alguns a maioria deles em nossa features são do padrão Grã-bretanha e dos EUA. Porém temos como agrupar para diminuir uma melhor padronização e diminuir a quantidades para gerar melhores insights.



➞ Certificate

u	335
a	196
ua	175
r	146
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

Name: count, dtype: int64

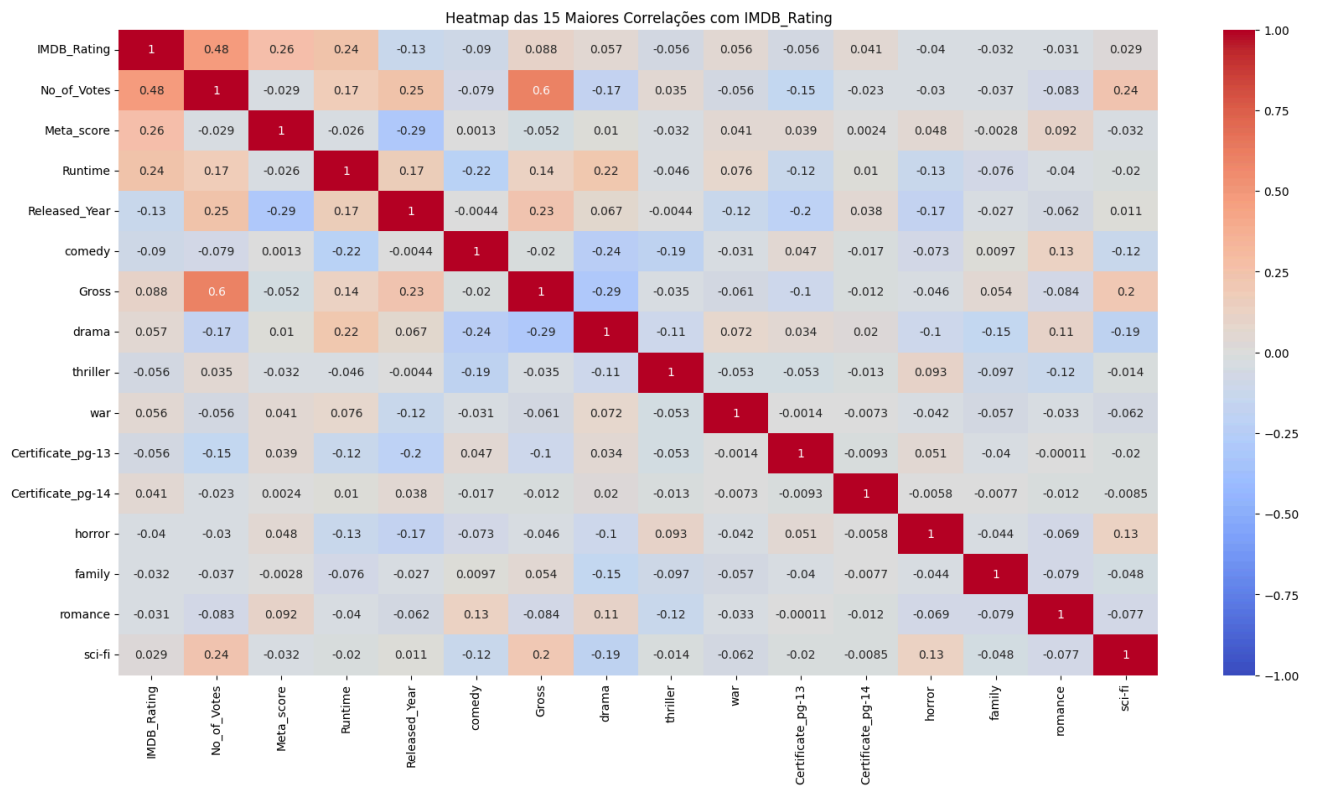
[Mostrar código](#)

➞ Certificate\_Agrupado

adult	343
pg-12	176
pg-13	80
pg	39
general audience	23
pg-14	1
pg-16	1
unrated	1

Name: count, dtype: int64

[Mostrar código](#)



Notamos que os gêneros que apresentaram maior correlação positiva com a nota do IMDB foi drama e war, já comedy e thriller apresentaram uma correlação mais fortes que as demais porém negativa. Isso indica que os filmes com as maiores notas do IMDB são de drama e comédia.

As classificações indicativa que apresenta maior correlação com a nota do IMDB é PG-13 e PG-14

#### Mostrar código

```
⇒ Fitting 3 folds for each of 50 candidates, totalling 150 fits
Melhores parâmetros encontrados: {'subsample': 0.7, 'reg_lambda': 0.1, 'reg_alpha': 0
Mean Squared Error on teste: 0.03320164268439449
R² Score on teste: 0.5156138570709294
Mean Absolute Error teste: 0.14425987148284913
MAPE: 1.81%
Accuracy: 98.19 %
```

No modelo com as features numéricas nulas inputadas com a média se demonstrou mais sensível quando fizemos o treino com as features de Gênero e Classificação indicativa, diminuindo o  $R^2$  para 0.49 já o modelo com mediana ficou em 0.51, se mostrando mais estável. Porém não tivemos melhoras comparado aos primeiros modelos testados.

#### Mostrar código

```
⇒ Top 20 Diretores:
Director
Steven Spielberg      13
Martin Scorsese       10
Alfred Hitchcock      9
David Fincher         8
Clint Eastwood        8
Quentin Tarantino     8
Christopher Nolan     8
Woody Allen           7
Rob Reiner            7
Hayao Miyazaki        7
Stanley Kubrick       6
Richard Linklater     6
Wes Anderson         6
Ridley Scott          6
Joel Coen             6
James Cameron         5
Alfonso Cuarón        5
Denis Villeneuve      5
Francis Ford Coppola  5
Ron Howard            5
Name: count, dtype: int64

Top 20 Atores:
Robert De Niro      16
Tom Hanks           14
Al Pacino           13
Brad Pitt           12
Matt Damon          11
Christian Bale      11
Leonardo DiCaprio  11
```

Clint Eastwood	11
Johnny Depp	9
Denzel Washington	9
Scarlett Johansson	9
Ethan Hawke	9
Harrison Ford	8
Ian McKellen	7
Jake Gyllenhaal	7
Robert Downey Jr.	7
Emma Watson	7
Edward Norton	7
Russell Crowe	7
Bruce Willis	7

Name: count, dtype: int64

Temos a lista dos diretores e atores com maior frequência nos filmes do nosso data set. Steven Spielberg e Robert De Niro lideram o ranking

#### Mostrar código

```

Feature Correlation
0      No_of_Votes    0.479308
1      Meta_score    0.261010
2      Runtime       0.242751
3  Director_Christopher Nolan    0.194498
4  Director_Francis Ford Coppola    0.135251
5      Released_Year   -0.133355
6      Star_Harrison Ford    0.120726
7      Director_Stanley Kubrick    0.105733
8      Director_Martin Scorsese    0.096967
9      Star_Ian McKellen    0.096225
10     comedy          -0.090209
11  Director_Quentin Tarantino    0.088451
12     Gross           0.088139
13  Star_Denzel Washington   -0.087735
14  Star_Robert De Niro      0.086457
Fitting 3 folds for each of 50 candidates, totalling 150 fits
Melhores parâmetros encontrados: {'subsample': 0.7, 'reg_lambda': 0.1, 'reg_alpha': 0
Mean Squared Error no teste: 0.03640133981120897
R² Score no teste: 0.4852678939010877
Mean Absolute Error no teste: 0.15243037986755367
MAPE: 1.92%
Accuracy: 98.08%
```

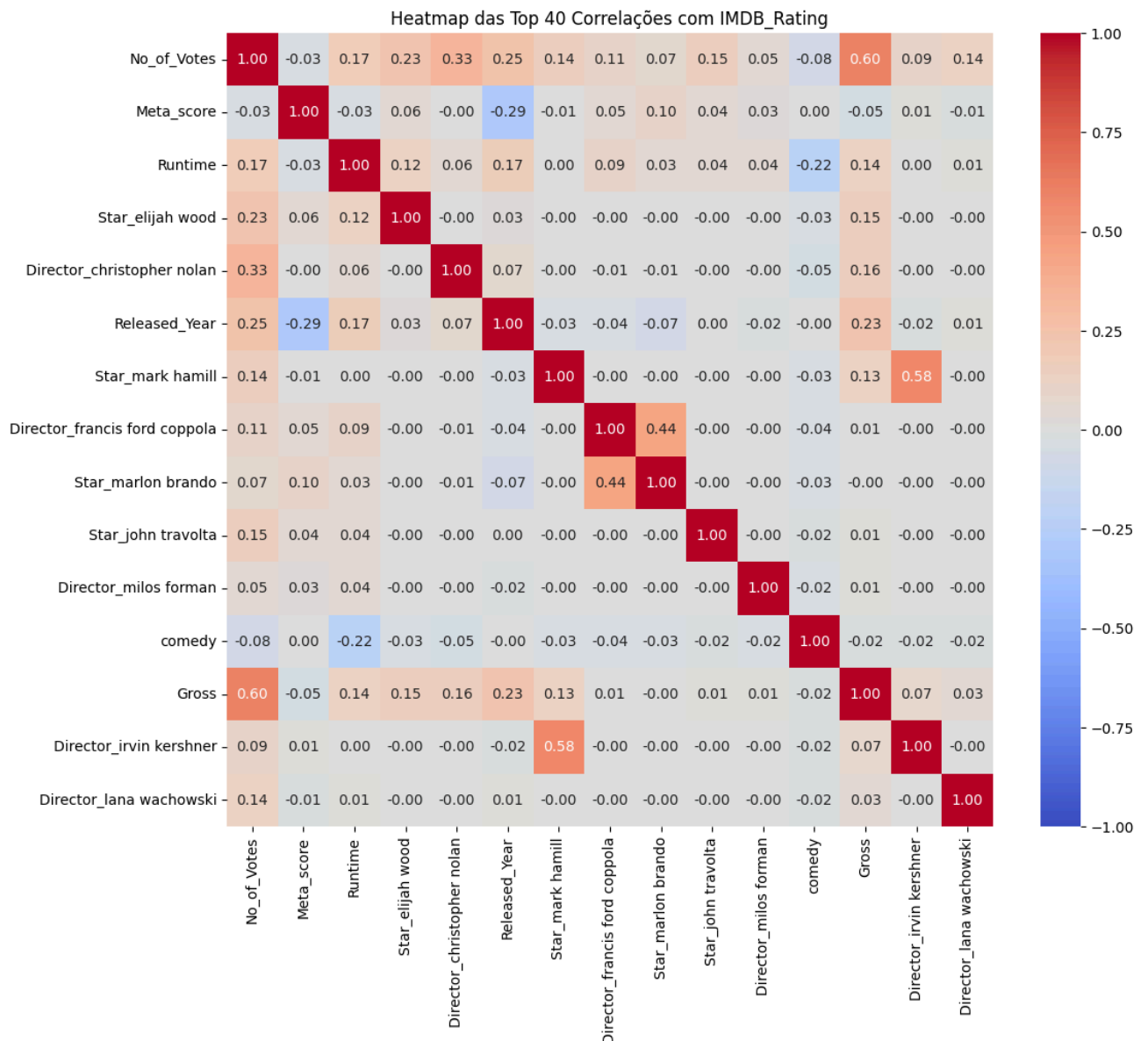
Os diretores Christopher Nolan e Francis Ford Coppola lideram o ranking o primeiro ator que aparece com maior correlação é Harrison Ford.

Vamos verificar agora se os atores e diretores filtrados por estarem nos filmes com maior nota do IMDB terão maior correlação.

[Mostrar código](#)



	Feature	Correlation
0	No_of_Votes	0.479308
1	Meta_score	0.261010
2	Runtime	0.242751
3	Star_elijah wood	0.171824
4	Director_christopher nolan	0.169873
5	Released_Year	-0.133355
6	Star_mark hamill	0.118048
7	Director_francis ford coppola	0.117806
8	Star_marlon brando	0.111140
9	Star_john travolta	0.110734
10	Director_milos forman	0.090852
11	comedy	-0.090209
12	Gross	0.088139
13	Director_irvin kershner	0.087472
14	Director_lana wachowski	0.087472



Fitting 3 folds for each of 50 candidates, totalling 150 fits

Melhores parâmetros encontrados: {'subsample': 0.5, 'reg\_lambda': 1.0, 'reg\_alpha': 0}

Mean Squared Error no teste: 0.0320988156782857

R<sup>2</sup> Score no teste: 0.5317032453245453

Mean Absolute Error no teste: 0.14338677501678468

MAPE: 1.80%

Accuracy: 98.2%

Nosso melhor modelo de regressão foi treinado com os atores com as maiores notas do IMDB. Fizemos a seleção das features com maior valor absoluto da correlação com nosso alvo.

A métricas são :

Mean Squared Error no teste: 0.0320988156782857

R<sup>2</sup> Score no teste: 0.5317032453245453

Mean Absolute Error no teste: 0.14338677501678468

MAPE: 1.80%

Accuracy: 98.2%

A acurácia está alta, o que significa que uma boa parte das previsões no conjunto de teste estão corretas. No entanto, queremos aumentar o R<sup>2</sup> score para garantir que nosso modelo capture melhor a variação nos dados e, assim, possa fazer previsões mais precisas para uma variedade maior de conjuntos de dados.

Vamos importar dois data sets, o primeiro de filmes ganhadores do oscar e o segundo de diretores/diretoras, atores e atrizes ganhadores do oscar.

Nosso intuito será encontrar features que tenham correlação maior com a nota do IMDB, parece intuitivo que filmes que ganharam o oscar e atores ganhadores do oscar estejam em filmes com maiores notas. Vamos verificar por meio de correlações, heatmaps e treinando alguns modelos.

```

oscar_movies = pd.read_csv('./drive/MyDrive/Colab Notebooks/oscars/oscar_movies.csv')
oscar_actors = pd.read_csv('./drive/MyDrive/Colab Notebooks/oscars/oscar_actors.csv')
df_numerical_features = df.copy()
df_numerical_features = df_numerical_features.select_dtypes(include=['number'])
df_median = df_numerical_features.median()
df_numerical_features = df_numerical_features.fillna(df_median)

```

### Mostrar código

```

↩️ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 1360 entries, 0 to 1359
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Unnamed: 0   1360 non-null   int64
1   Film         1359 non-null   object
2   Year         1360 non-null   int64
3   Award        1360 non-null   int64
4   Nomination   1360 non-null   int64
dtypes: int64(4), object(1)
memory usage: 53.2+ KB
Unnamed: 0    0
Film          1
Year          0
Award         0
Nomination    0
dtype: int64

```

### Mostrar código

```

↩️ Proporção de filmes indicados ao Oscar: 27.23%
Proporção de filmes com indicações que ganharam algum Oscar: 27.03%
Proporção de filmes indicados que não ganharam Oscar: 0.20%
Proporção de filmes que não foram indicados ao Oscar: 72.77%
Filmes que foram indicados mas não ganharam Oscar:
   Series_Title  Indicacoes_Oscar  Oscars_Ganhos
100         toy story                3              0
425  planet of the apes                2              0

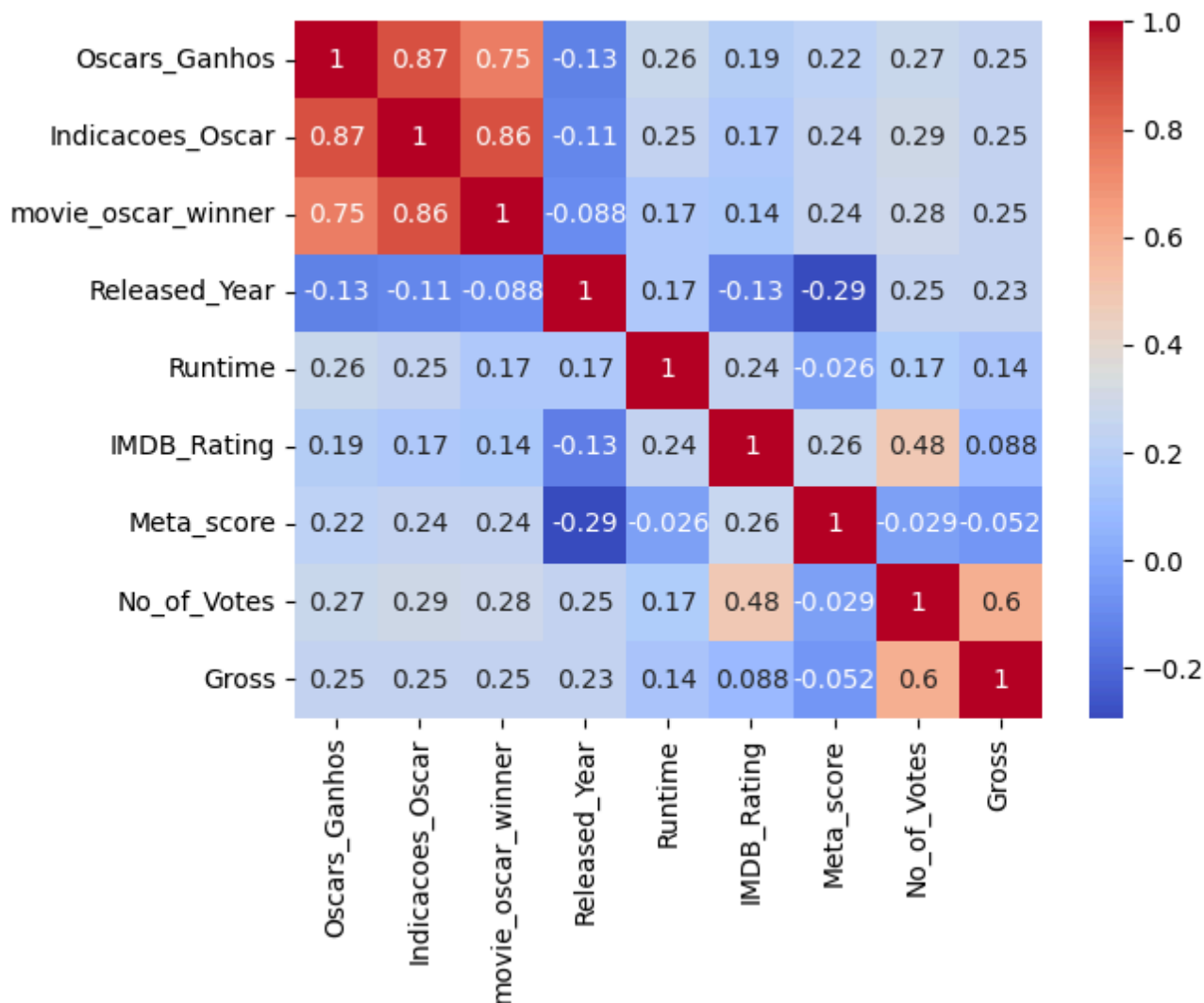
```

Fizemos alguns tratamentos necessário no data set e concatenação com nosso data set do IMDB, vamos investigas algumas correlações

### Mostrar código



↔ <Axes: >



Nossas features da quantidade de oscars ganhos, indicações ao oscar e se o filme foi ganhador do oscar tiveram correlação mais fortes com a nota do IMDB do que muitas de nossas features do data set original.

As correlações com o faturamento foram até maior do que a nota do IMDB, indicado que os filmes ganhadores de oscar e a quantidade de oscars influencia sim o faturamento do filme.

Oscars Ganhos e indicações oscars apresentam uma forte multicolinearidade.

[Mostrar código](#)

↔ Mean Squared Error: 0.0323459618200779  
R<sup>2</sup> Score: 0.5280975753430779  
Mean Absolute Error: 0.14009504222869876  
MAPE: 1.76%  
Accuracy: 98.24 %

Nosso modelo que está apresentando bom desempenho tem as features dos filmes que ganharam o oscar, os top 10 diretores e atores com maiores notas do IMDB e as demais features ( gross, ano de lançamento, meta score, quantidade de avaliações)

Agora vamos verificar nosso data set com atores e diretores ganhadores do oscar.

Foi verificado que muitos dos valores nulos estão deslocados uma coluna para direita, vamos relocar.

[Mostrar código](#)



	category	name	film	winner
0	actor	richard barthelmess	the noose	False
1	actor	emil jannings	the last command	True
2	actress	louise dresser	a ship comes in	False
3	actress	janet gaynor	7th heaven	True
4	actress	gloria swanson	sadie thompson	False
5	art direction	rochus gliese	sunrise	False
6	art direction	william cameron menzies	the dove	True
7	art direction	harry oliver	7th heaven	False
8	cinematography	george barnes	the devil dancer	False
9	cinematography	charles rosher	sunrise	True
10	cinematography	karl struss	sunrise	True
11	directing (comedy picture)	lewis milestone	two arabian knights	True
12	directing (comedy picture)	ted wilde	speedy	False
13	directing (dramatic picture)	frank borzage	7th heaven	True
14	directing (dramatic picture)	herbert brenon	sorrell and son	False
15	directing (dramatic picture)	king vidor	the crowd	False
17	engineering effects	roy pomeroy	wings	True
19	outstanding picture	the caddo company	the racket	False
20	outstanding picture	fox	7th heaven	False
21	outstanding picture	paramount famous lasky	wings	True

[Mostrar código](#)

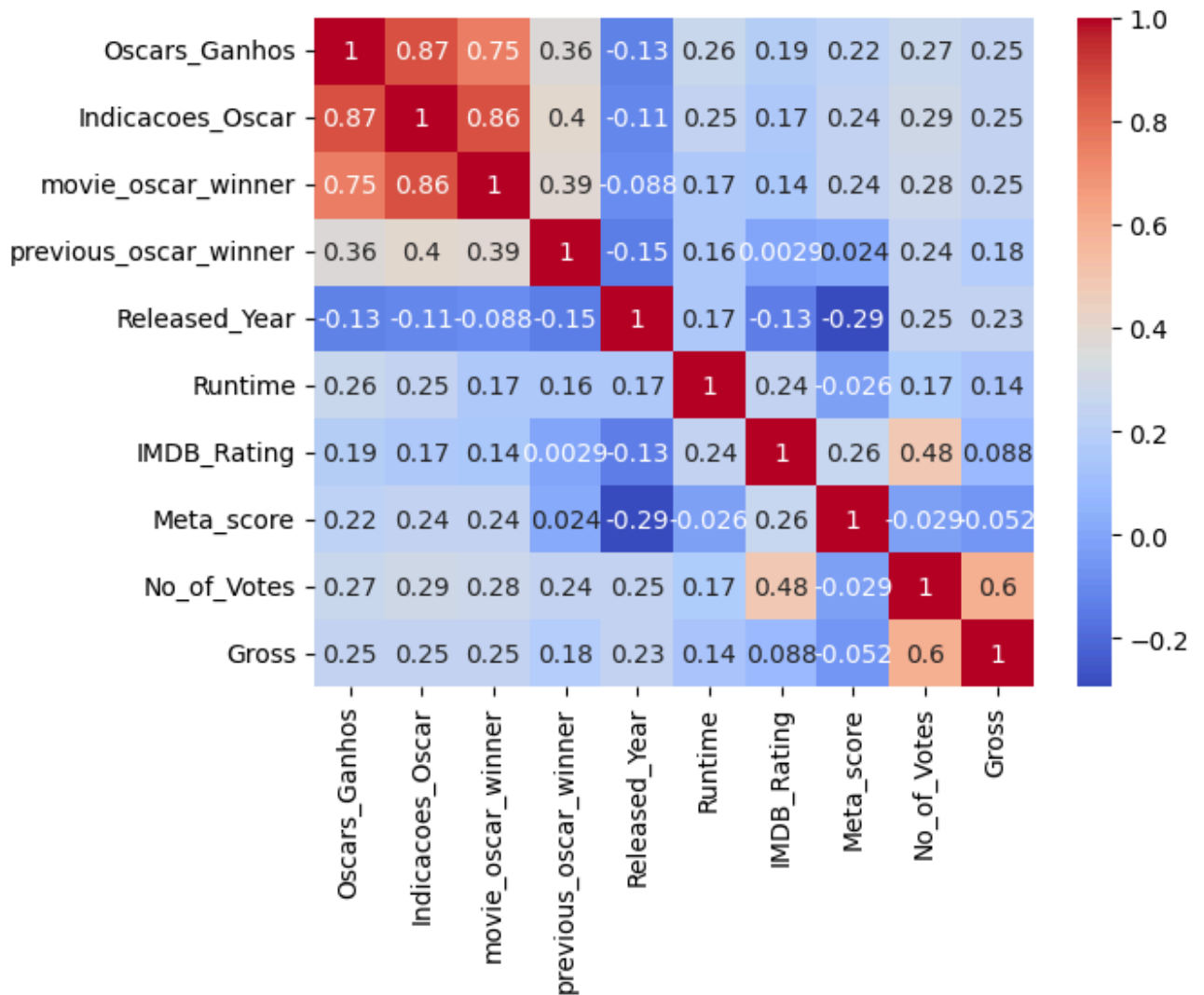
```
➡ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 999 entries, 0 to 998
Data columns (total 13 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Certificate                            999 non-null    object
1   Genre                                999 non-null    object
2   Overview                              999 non-null    object
3   Director                             999 non-null    object
4   Star1                                999 non-null    object
5   Star2                                999 non-null    object
6   Star3                                999 non-null    object
7   Star4                                999 non-null    object
8   Series_Title_Normalizado              999 non-null    object
9   Oscars_Ganhos                         999 non-null    int64
10  Indicacoes_Oscar                      999 non-null    int64
11  movie_oscar_winner                    999 non-null    int64
12  previous_oscar_winner                 999 non-null    bool
dtypes: bool(1), int64(3), object(9)
memory usage: 94.8+ KB
```

[Mostrar código](#)

```
➡ Porcentagem de ganhadores do Oscar no nosso dataset: 48.25%
Porcentagem de não ganhadores do Oscar no nosso dataset: 51.75%
```

[Mostrar código](#)

↩ <Axes: >



[Mostrar código](#)

↩ Mean Squared Error: 0.03207678685276501  
R<sup>2</sup> Score: 0.532024628755138  
Mean Absolute Error: 0.13978295660018922  
MAPE: 1.75%  
Accuracy: 98.25 %

Nosso modelo de regressão teve um desempenho levemente maior com o dataset original com as features numéricas concatenado com os datasets de filmes, atores e diretores ganhadores do oscar.

Vamos concatenar o dataset com as features de gênero e classificação para verificar o desempenho.

[Mostrar código](#)

↩

	Feature	Correlation
0	No_of_Votes	0.479308
1	Meta_score	0.261010

2	Runtime	0.242751
3	Oscars_Ganhos	0.187774
4	Star_elijah wood	0.171824
..	...	...
65	adventure	0.007430
66	history	0.005035
67	previous_oscar_winner	0.002864
68	action	0.001294
69	musical	-0.000430

[70 rows x 2 columns]

Fitting 3 folds for each of 50 candidates, totalling 150 fits

Melhores parâmetros encontrados: {'subsample': 0.7, 'reg\_lambda': 10.0, 'reg\_alpha':

Mean Squared Error no teste: 0.03273865078069816

R<sup>2</sup> Score no teste: 0.522368548836354

Mean Absolute Error no teste: 0.1412326216697693

MAPE: 1.77%

Accuracy: 98.23%

## Mostrar código



	Feature	Correlation
0	No_of_Votes	0.479308
1	Meta_score	0.261010
2	Runtime	0.242751
3	Oscars_Ganhos	0.187774
4	Star_elijah wood	0.171824
..	...	...
65	adventure	0.007430
66	history	0.005035
67	previous_oscar_winner	0.002864
68	action	0.001294
69	musical	-0.000430

[70 rows x 2 columns]

Model: RandomForestRegressor

Mean Squared Error: 0.0335382999999999965

R<sup>2</sup> Score: 0.510702288684235

Mean Absolute Error: 0.144579999999999996

MAPE: 1.81%

Accuracy: 98.19%

Model: GradientBoostingRegressor

Mean Squared Error: 0.03231566047629319

R<sup>2</sup> Score: 0.528539648380878

Mean Absolute Error: 0.14372255338905357

MAPE: 1.80%

Accuracy: 98.2%

Model: AdaBoostRegressor

Mean Squared Error: 0.04289630050488991

R<sup>2</sup> Score: 0.3741763398575375

Mean Absolute Error: 0.1706391976289605

MAPE: 2.16%

Accuracy: 97.84%

Nosso modelo com Gradiente XGBoost apresentou melhores resultados comparado ao AdaBoost, GradienteBoost e RandomForest

Porém nosso modelo com a combinação do dataset com onehot de diretores , atores , genre e certificate não apresentou melhores resultados do que o modelo com as features numéricas do dataset original combinado com os ganhadores do oscars.

Vamos retomar com nosso modelo com melhor desempenho e verificar o desempenho dele com o input do filme a ser predito.

Também vamos serializar o modelo.

```
{'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'}
```

#### Mostrar código

```
➦ Mean Squared Error: 0.03210785932190307
R² Score: 0.5315713055982045
Mean Absolute Error: 0.14130054235458378
MAPE: 1.77%
Accuracy: 98.23 %
{'Director_Won_Oscar': 1, 'Star1_Won_Oscar': 1, 'Star2_Won_Oscar': 1, 'Star3_Won_Oscar': 1, 'Star4_Won_Oscar': 1}
Predicted IMDB Rating: 7.823731899261475
```

Fizemos um treino no nosso modelo com filme, atores e diretores que ganharam o oscar para verificar se a função de checar se o input do dicionário estava retornando corretamente o valor boelano

#### Mostrar código

```
➦ Mean Squared Error: 0.03210785932190307
R² Score: 0.5315713055982045
Mean Absolute Error: 0.14130054235458378
MAPE: 1.77%
Accuracy: 98.23 %
{'Director_Won_Oscar': 0, 'Star1_Won_Oscar': 1, 'Star2_Won_Oscar': 1, 'Star3_Won_Oscar': 1, 'Star4_Won_Oscar': 1}
Processed Input:
Oscars_Ganhos  Indicacoes_Oscar  movie_oscar_winner  previous_oscar_winner  \
0              0              0              0              True

Released_Year  Runtime  Meta_score  No_of_Votes  Gross
```

0 1994 142 80.0 2343110 28341469.0  
Predicted IMDB Rating: 8.738204956054688

### Mostrar código

```
↔ Feature Correlation
0 No_of_Votes 0.479308
1 Meta_score 0.261010
2 Runtime 0.242751
3 Star_elijah wood 0.171824
4 Director_christopher nolan 0.169873
5 Released_Year -0.133355
6 Star_mark hamill 0.118048
7 Director_francis ford coppola 0.117806
8 Star_marlon brando 0.111140
9 Star_john travolta 0.110734
10 Director_milos forman 0.090852
11 comedy -0.090209
12 Gross 0.088139
13 Director_irvin kershner 0.087472
14 Director_lana wachowski 0.087472
Mean Squared Error no teste: 0.03339910788627613
R² Score no teste: 0.5127329933615227
Mean Absolute Error no teste: 0.14445956707000734
MAPE: 1.81%
Accuracy: 98.19%
Predicted IMDB Rating: 8.63430404663086
```

### Mostrar código

```
↔ Mean Squared Error on test: 0.03237910359808707
R² Score on test: 0.5276140625791984
Mean Absolute Error on test: 0.14328102159500122
MAPE: 1.80%
Accuracy: 98.2 %
Predicted IMDB Rating: 8.761387825012207
```

Obtivemos melhores resultados com nosso modelo de regressão treinado com o dataset de ganhadores do oscar. Tivemos um R² Score um pouco maior e um MSE e MAPE um pouco menor. A diferença foi pouca, mas vamos fazer alguns testes com filmes mais recentes que não estão no dataset e verificar qual modelo de aproximou mais do nosso alvo.

### Mostrar código

```
↔ Mean Squared Error: 0.03210785932190307
R² Score: 0.5315713055982045
Mean Absolute Error: 0.14130054235458378
MAPE: 1.77%
Accuracy: 98.23 %
{'Director_Won_Oscar': 0, 'Star1_Won_Oscar': 0, 'Star2_Won_Oscar': 0, 'Star3_Won_Osca
Processed Input:
Oscars_Ganhos Indicacoes_Oscar movie_oscar_winner previous_oscar_winner \
```

```
Mean Squared Error: 0.03210785932190307
R² Score: 0.5315713055982045
Mean Absolute Error: 0.14130054235458378
MAPE: 1.77%
Accuracy: 98.23 %
{'Director_Won_Oscar': 0, 'Star1_Won_Oscar': 0, 'Star2_Won_Oscar': 0,
'Star3_Won_Oscar': 0, 'Star4_Won_Oscar': 0}
Processed Input:
  Oscars_Ganhos  Indicacoes_Oscar  movie_oscar_winner
previous_oscar_winner \
0                0                0                0
False

Released_Year  Runtime  Meta_score  No_of_Votes  Gross
0             2023     150         64.0        388000  1.500000e+09
Predicted IMDB Rating: 7.836304664611816
```

```
Mean Squared Error on test: 0.03237910359808707
R² Score on test: 0.5276140625791984
Mean Absolute Error on test: 0.14328102159500122
MAPE: 1.80%
Accuracy: 98.2 %
Predicted IMDB Rating: 8.021060943603516
```



Elenco e equipe · Avaliações de usuários · Curiosidades · Perguntas frequentes

IMDbPro

Todos os tópicos

# Guardiões da Galáxia Vol. 3

Título original: Guardians of the Galaxy Vol. 3

2023 · 14 · 2 h 30 min

AVALIAÇÃO DA IMDb

7,9/10  
388 mil

SUA AVALIAÇÃO

Avaliar

POPULARIDADE

294 ▲18

+



MAIS UMA VEZ COM EMOÇÃO

4 DE MAIO NOS CINEMAS



Reproduzir trailer com som 0:30

10 4

42 VÍDEOS

99+ FOTOS

Nosso modelo Chegou bem próximo da avaliação do IMDB de um filme mais recente que não estava no data set. Foi utilizado o meta score e quantidade de avaliações informadas no próprio site do IMDB e o faturamento do filme informado no google. Porém o outro modelo treinado se aproximou também, vamos verificar com outro filme que não tem no data set.

Página 2 de Anotações Rápidas

Elenco e equipe · Avaliações de usuários · Curiosidades · Perguntas frequentes
IMDbPro
Todos os tópicos

# Cruella

2021 · 12 · 2 h 14 min

AVALIAÇÃO DA IMDb
7,3/10  
269 mil

SUA AVALIAÇÃO
Avaliar

POPULARIDADE
626 · 592

+ EMMA STONE

Disney  
Cruella  
EM BREVE 2021

Reproduzir trailer com som 1:01

1 3

65 VÍDEOS

99+ FOTOS

Aventura Comédia Policial

```

Star1': 'Emma Stone',
'Star2': 'Emma Thompson',
'Star3': 'Joel Fry',
'Star4': 'Paul Walter Hauser',
'No_of_Votes': 269000,
'Gross': '233,000,000'

# Preprocessar o novo input
processed_input = preprocess_input(new_input)

# Fazer a previsão
prediction = regressor.predict(processed_input)
print(f'Predicted IMDB Rating: {prediction[0]}')

Mean Squared Error: 0.03210785932190307
R² Score: 0.5315713055982045
Mean Absolute Error: 0.14130054235458378
MAPE: 1.77%
Accuracy: 98.23 %
{'Director_Won_Oscar': 0, 'Star1_Won_Oscar': 1, 'Star2_Won_Oscar': 1, 'Star3_Won_Oscar': 0, 'Star4_Won_Oscar': 0}
Processed Input:
Oscars_Ganhos  Indicacoes_Oscar  movie_oscar_winner  previous_oscar_winner \
0              0              0              0              True

Released_Year  Runtime  Meta_score  No_of_Votes  Gross
0             2021     134         59.0       269000  233000000.0
Predicted IMDB Rating: 7.782958984375

```

Página 3 de Anotações Rápidas

```


'Runtime': '134 min',
'Genre': 'Comedy, Crime',
'Overview': 'A live-action prequel feature film following a young Cruella de Vil.',
'Meta_score': 59.0,
'Director': 'Craig Gillespie',
'Star1': 'Emma Stone',
'Star2': 'Emma Thompson',
'Star3': 'Joel Fry',
'Star4': 'Paul Walter Hauser',
'No_of_Votes': 269000,
'Gross': '233,000,000'

}

# Preprocessar o novo input
processed_input = preprocess_input(new_input)

# Fazer a previsão
prediction = regressor.predict(processed_input)
print(f'Predicted IMDB Rating: {prediction[0]}')

```

 Mean Squared Error on test: 0.03237910359808707  
 R<sup>2</sup> Score on test: 0.5276140625791984  
 Mean Absolute Error on test: 0.14328102159500122  
 MAPE: 1.80%  
 Accuracy: 98.2 %  
 Predicted IMDB Rating: 7.808229923248291

Nosso modelo se aproximou mais do filme cruella de 2021. Porém a diferença do outro modelo que só usa as features numéricas do data set original não foi tão grande. Notamos que R<sup>2</sup> e o MSE do nosso modelo tem desempenho ligeiramente melhor. Como o modelo já foi treinado e apresentou melhores métricas vamos seguir com ele, porém comparado o desempenho com a complexidade de importar os dois data sets e fazer várias modelagens é provável que não compense por um desempenho tão similar.

Mean Squared Error on test: 0.03237910359808707  
 R<sup>2</sup> Score on test: 0.5276140625791984  
 Mean Absolute Error on test: 0.14328102159500122  
 MAPE: 1.80%  
 Accuracy: 98.2 %  
 Predicted IMDB Rating: 7.808229923248291

[91] [Mostrar código](#)

 Top 10 Filmes Recomendados para Todos os Públicos:

	Series Title	IMDB_Rating \
9	The Lord of the Rings: The Fellowship of the Ring	8.8
4	The Lord of the Rings: The Return of the King	8.9
35	The Prestige	8.5
46	Back to the Future	8.5
65	WALL·E	8.4
43	Terminator 2: Judgment Day	8.5
108	Star Wars: Episode VI - Return of the Jedi	8.3
241	Finding Nemo	8.1
42	The Lion King	8.5
249	The Truman Show	8.1

	No_of_Votes	Genre Certificate
9	1661481	Action, Adventure, Drama U
4	1642758	Action, Adventure, Drama U
35	1190259	Drama, Mystery, Sci-Fi U

Mean Squared Error on test: 0.10523751055800707  
 R² Score on test: 0.5276140625791984  
 Mean Absolute Error on test: 0.14328102159500122  
 MAPE: 1.80%  
 Accuracy: 98.2 %  
 Predicted IMDB Rating: 7.808229923248291

[91] [Mostrar código](#)

Top 10 Filmes Recomendados para Todos os Públicos:

	Series_Title	IMDB_Rating
9	The Lord of the Rings: The Fellowship of the Ring	8.8
4	The Lord of the Rings: The Return of the King	8.9
35	The Prestige	8.5
46	Back to the Future	8.5
65	WALL·E	8.4
43	Terminator 2: Judgment Day	8.5
108	Star Wars: Episode VI - Return of the Jedi	8.3
241	Finding Nemo	8.1
42	The Lion King	8.5
249	The Truman Show	8.1

	No_of_Votes	Genre	Certificate
9	1661481	Action, Adventure, Drama	U
4	1642758	Action, Adventure, Drama	U
35	1190259	Drama, Mystery, Sci-Fi	U

4	1642758	Action, Adventure, Drama	U
35	1190259	Drama, Mystery, Sci-Fi	U
46	1058081	Adventure, Comedy, Sci-Fi	U
65	999790	Animation, Adventure, Family	U
43	995506	Action, Sci-Fi	U
108	950470	Action, Adventure, Fantasy	U
241	949565	Animation, Adventure, Comedy	U
42	942045	Animation, Adventure, Drama	U
249	939631	Comedy, Drama	U

Já que a recomendação é para uma pessoa desconhecida devemos analisar com cautela pois além da popularidade do filme baseado na nota do IMDB com a quantidade de votos, devemos fazer uma filtragem para que o filme tenha classificação indicativa para todo o público já que não sabemos a idade da pessoa.

Agora vamos serializar nosso modelo em formato .pkl para poder ser carregado e utilizado em outro notebook retratando de forma fiel o nosso modelo

Fizemos a serealização do nosso modelo com melhor desempenho, tabém geramos o readme.txt com as bibliotecas utilizadas para a nossa EDA e Machine Learning. Abaixo temos as respostas para as perguntas, mas também elas se encontrarão em um pdf a parte no repositório.

## 2- Responder as seguintes perguntas:

A- Qual filme seria recomendado para uma pessoa que não conheço



The Lord of the Rings: The Fellowship of the Ring.

Esse filme foi escolhido selecionando a melhor relação entre popularidade(número de votos) e a avaliação do IMDB mais alta. Também foi levado em conta que já que não conheço a pessoa devo escolher um filme com classificação indicativa Livre.

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

Baseado no dataset do IMDB os principais fatores são:

Número de votos, Ano de lançamento, se o filme foi dirigido pelo diretor Christopher Nolan, Se o ator Elijah Wood atuou, Tempo de duração do filme e se o ator Mark Hamill atuou no filme.

Baseado na combinação com os data sets do oscar temos outros fatores como:

Quantidade de indicações do filme ao oscar, quantidade de oscars ganhado pelo filme, se o ator/atriz ou diretor/diretoras já foram premiados pelo oscar anteriormente

**C- Quais insights podem ser tirados com a coluna Overview? é possível inferir o gênero do filme a partir dela?**

Sim nós temos algumas correlações por exemplo, filmes do Gênero "Crime" possuem palavras como young,murder,crime,family com frequência. Já filmes do gênero "Action": must, man, world, former. Filmes de Biografia: story, life, war. Comédia tem friends com frequência já drama tem love e woman.

Acesse a nuvem de palavras completa:

[https://github.com/lanPerigoVianna/IMDB\\_PREDICTOR/tree/main/WordCloud](https://github.com/lanPerigoVianna/IMDB_PREDICTOR/tree/main/WordCloud)

### 3- Fazer a previsão da nota do IMDB a partir dos dados

**A- Quais variáveis e transformações foram utilizadas**

Foram utilizadas as variáveis numéricas (Faturamento, número de votos, duração do filme, ano de lançamento, média ponderada das críticas) Foi combinado com dois data sets de filmes, atores/atrizes, diretores/diretoras ganhadores do oscars e foram utilizado as variáveis (Quantidade de oscars ganho pelo filme, quantidade de indicações, se alguém do elenco já havia ganhado algum oscar)

**B- Qual tipo de problema está sendo resolvido**

Um problema de regressão onde queremos prever um valor contínuo (Nota do IMDB)

**C- Qual modelo melhor se aproximou dos dados, prós e contras?**

O modelo que mais se aproximou dos dados foi o regressor da biblioteca XGBoost. Esse modelo foi treinado com 9 features no eixo X e nosso alvo (IMDB\_Rating) no eixo Y.

As features foram: Oscars\_Ganhos, Indicações\_Oscar, movie\_oscar\_winner, previous\_oscar\_winner, Released\_Year, Runtime, Meta\_score, No\_of\_Votes, Gross.

Prós:

Maior acurácia: Leva em conta se o elenco e o filme já ganharam o Oscar, o que pode ser um forte indicador de qualidade.

Capacidade de lidar com dados complexos: XGBoost pode capturar relações não lineares entre as features e a variável alvo, resultando em melhores previsões.

Eficiência computacional: XGBoost é altamente otimizado para velocidade e desempenho, utilizando técnicas de paralelismo e processamento distribuído.

Contras:

Necessidade de atualizações constantes: O data set com os ganhadores do Oscar precisa ser atualizado regularmente. Filmes recentes que ainda não foram atualizados na lista do Oscar podem afetar o desempenho do modelo.

Complexidade do modelo: XGBoost pode ser mais difícil de interpretar em comparação com modelos mais simples, o que pode dificultar a compreensão dos fatores que influenciam as previsões.

Dependência de tuning de hiper parâmetros: A eficácia do XGBoost pode depender significativamente da escolha de hiper parâmetros, o que requer tempo e recursos para otimização adequada.

#### D- Qual medida de performance foi escolhida

A medida de performance escolhida foi a Acurácia, calculada pela subtração da porcentagem da Média Absoluta do Erro (MAPE) da perfeição (100%) e o  $R^2$  Score.

Acurácia: Indicador que mostra o quão próximo o modelo está das previsões corretas. Uma acurácia alta significa que o modelo tem um bom desempenho geral.

MAPE (Mean Absolute Percentage Error): Mede a precisão do modelo, expressando o erro absoluto médio como uma porcentagem das observações reais. É útil por ser uma métrica intuitiva e fácil de interpretar.

$R^2$  Score: Avalia a proporção da variabilidade total dos dados explicada pelo modelo. Um  $R^2$  próximo de 1 indica um modelo que explica bem a variabilidade dos dados, enquanto um valor próximo de 0 indica o contrário.