🎬 01 — EDA & Hipóteses (PProductions)

Este notebook realiza uma **Análise Exploratória de Dados (EDA)** na base cinematográfica entregue em CSV. O objetivo é entender as principais características dos filmes, levantar hipóteses e orientar a **PProductions** sobre quais tipos de filmes tendem a performar melhor.

> **Entrada esperada**: `data/movies.csv` contendo as colunas: `Series_Title, Released_Year, Certificate, Runtime, Genre, IMDB_Rating, Overview, Meta_score, Director, Star1, Star2, Star3, Star4, No_of_Votes, Gross`.

1. Carregamento dos dados

```
Shape: (999, 16)
```

	Unnamed: 0	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating	Overview	Meta_score	Director	Star1	Star2	Star3	Star4	No_of_Vot
0	1	The Godfather	1972	A	175 min	Crime, Drama	9.2	An organized crime dynasty's aging patriarch t...	100.0	Francis Ford Coppola	Marlon Brando	Al Pacino	James Caan	Diane Keaton	16203
1	2	The Dark Knight	2008	UA	152 min	Action, Crime, Drama	9.0	When the menace known as the Joker wreaks havo...	84.0	Christopher Nolan	Christian Bale	Heath Ledger	Aaron Eckhart	Michael Caine	23032
2	3	The Godfather: Part II	1974	A	202 min	Crime, Drama	9.0	The early life and career of Vito Corleone in ...	90.0	Francis Ford Coppola	Al Pacino	Robert De Niro	Robert Duvall	Diane Keaton	11299

```
<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
```

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	999.0	500.000000	288.530761	1.0	250.5	500.0	749.5	999.0
IMDB_Rating	999.0	7.947948	0.272290	7.6	7.7	7.9	8.1	9.2
Meta_score	842.0	77.969121	12.383257	28.0	70.0	79.0	87.0	100.0
No_of_Votes	999.0	271621.422422	320912.621055	25088.0	55471.5	138356.0	373167.5	2303232.0

2. Limpeza & Parsing de colunas

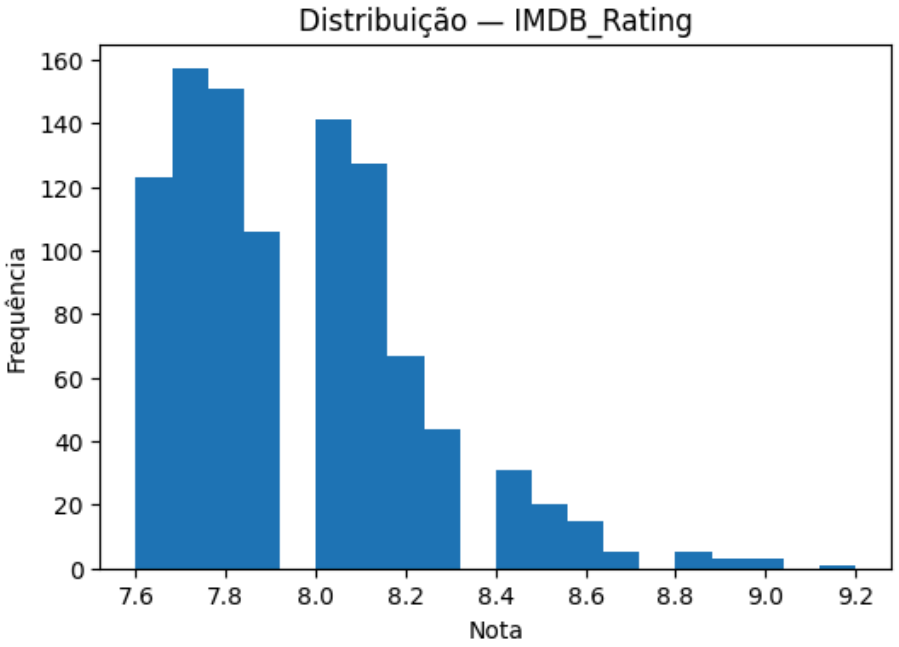
Transformações principais:

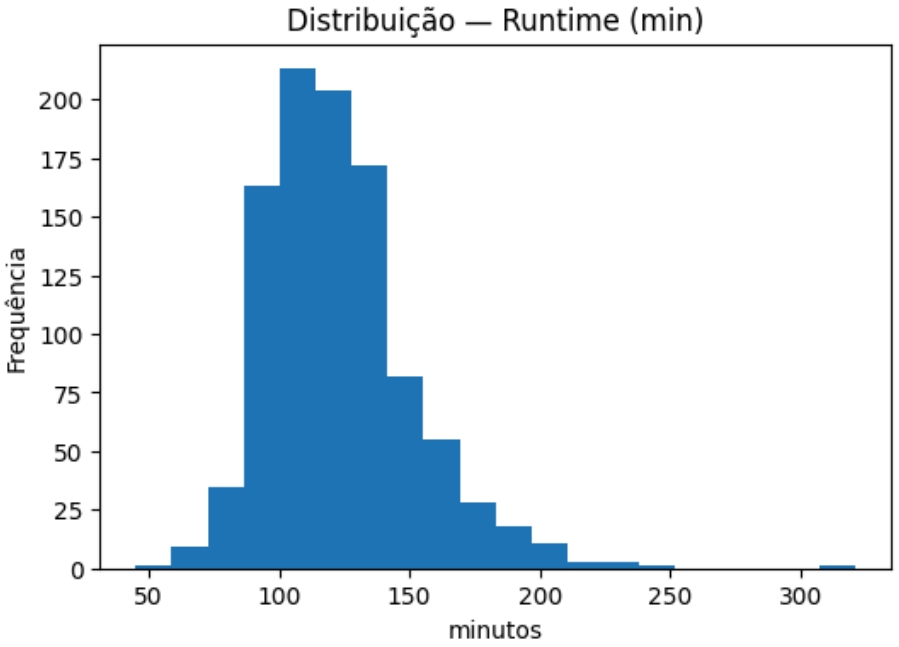
- **Runtime** → minutos (inteiro)
- **Gross** → número (float), removendo separadores
- **Released_Year** → inteiro
- **Genre** → lista de gêneros + gênero principal (`Genre_Primary`)

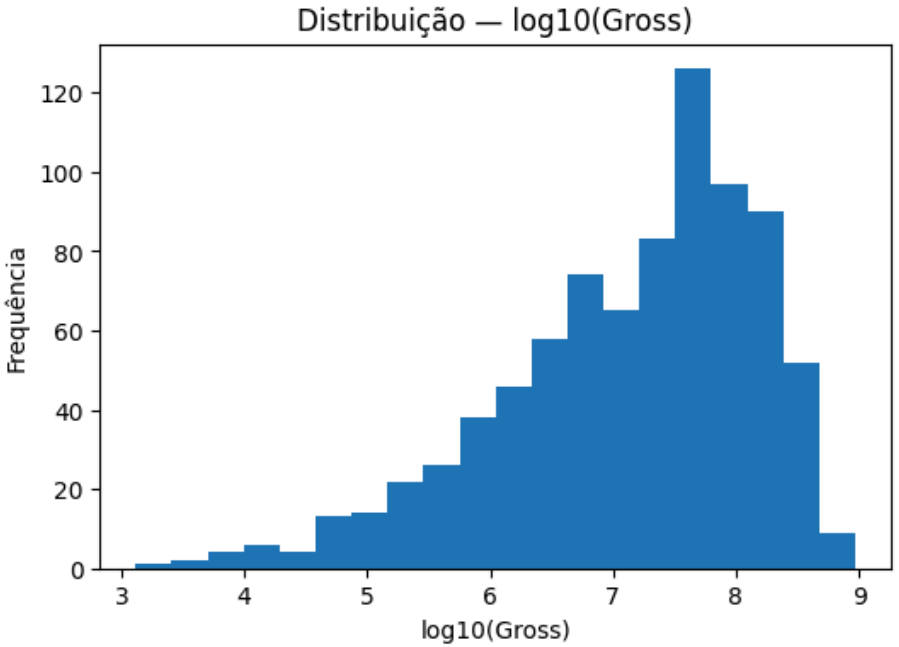
	Unnamed: 0	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating	Overview	Meta_score	Director	Star1	Star2	Star3	Star4	No_of_Vot
0	1	The Godfather	1972	A	175 min	Crime, Drama	9.2	An organized crime dynasty's aging patriarch t...	100.0	Francis Ford Coppola	Marlon Brando	Al Pacino	James Caan	Diane Keaton	16203
1	2	The Dark Knight	2008	UA	152 min	Action, Crime, Drama	9.0	When the menace known as the Joker wreaks havo...	84.0	Christopher Nolan	Christian Bale	Heath Ledger	Aaron Eckhart	Michael Caine	23032
2	3	The Godfather: Part II	1974	A	202 min	Crime, Drama	9.0	The early life and career of Vito Corleone in ...	90.0	Francis Ford Coppola	Al Pacino	Robert De Niro	Robert Duvall	Diane Keaton	11299

	missing_ratio
Gross_num	0.169169
Gross	0.169169
Meta_score	0.157157
Certificate	0.101101
Released_Year_int	0.001001
Unnamed: 0	0.000000
Star3	0.000000
Genre_List	0.000000
Runtime_min	0.000000
No_of_Votes	0.000000
Star4	0.000000
Star1	0.000000
Star2	0.000000
Series_Title	0.000000
Director	0.000000

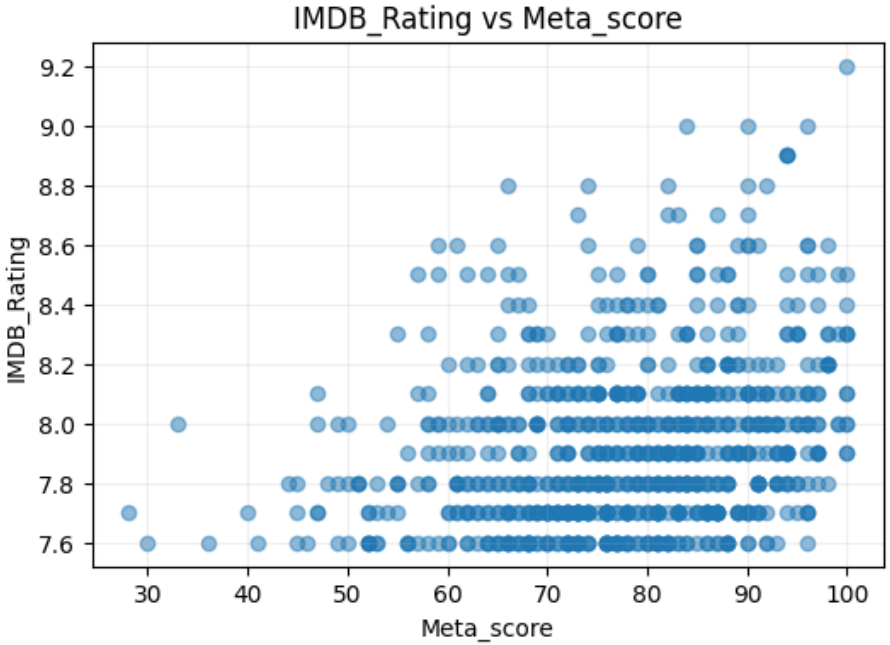
3. Distribuições básicas

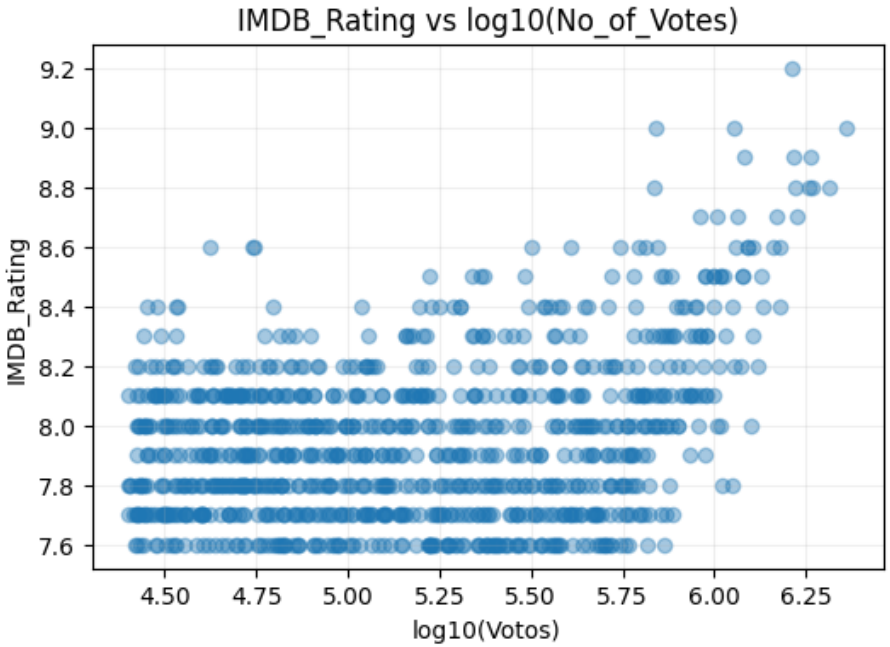

Distribuição — IMDB_Rating
Frequência
Nota
0
20
40
60
80
100
120
140
160
7.6
7.8
8.0
8.2
8.4
8.6
8.8
9.0
9.2

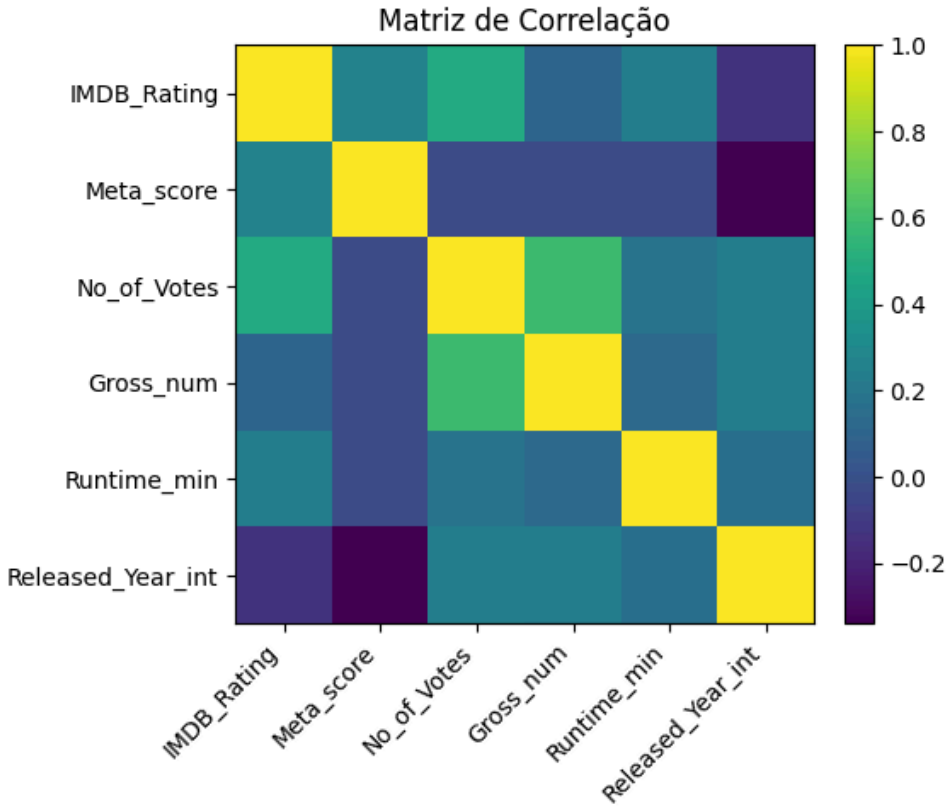

Distribuição — Runtime (min)
Frequência
minutos
0
25
50
75
100
125
150
175
200
50
100
150
200
250
300



4. Relações entre variáveis

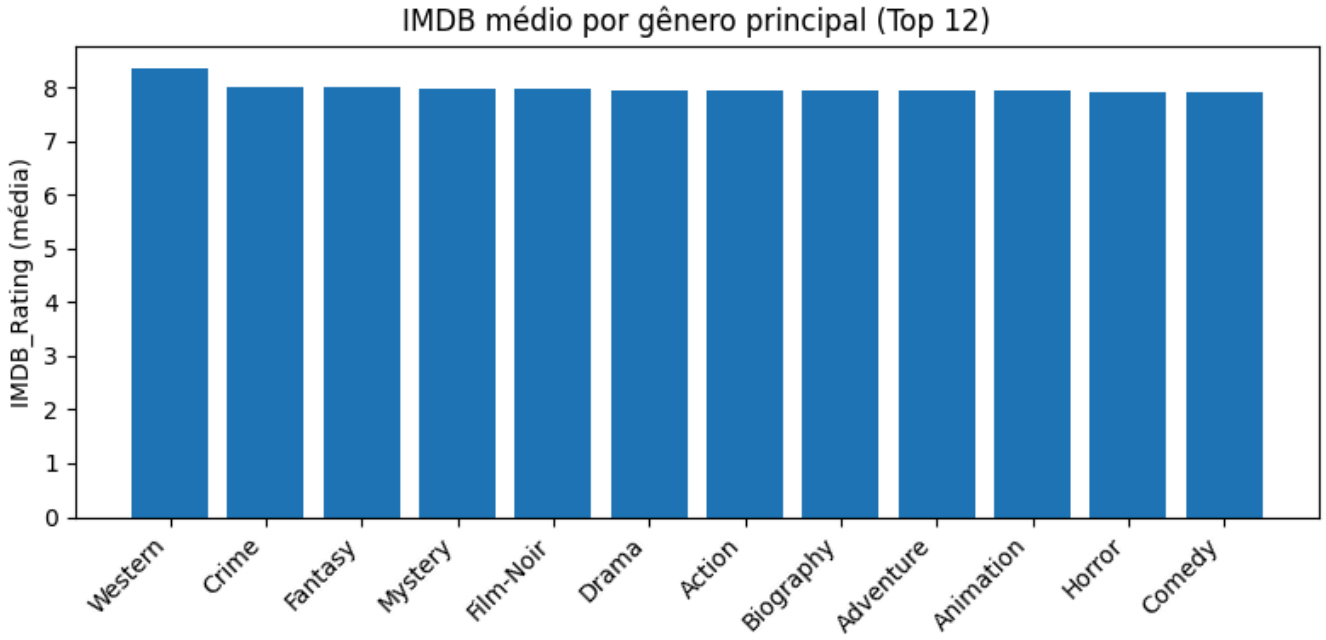

IMDB_Rating vs Meta_score
IMDB_Rating
9.2
9.0
8.8
8.6
8.4
8.2
8.0
7.8
7.6
30
40
50
60
70
80
90
100
Meta_score


IMDB_Rating vs log10(No_of_Votes)
IMDB_Rating
9.2
9.0
8.8
8.6
8.4
8.2
8.0
7.8
7.6
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
log10(Votos)

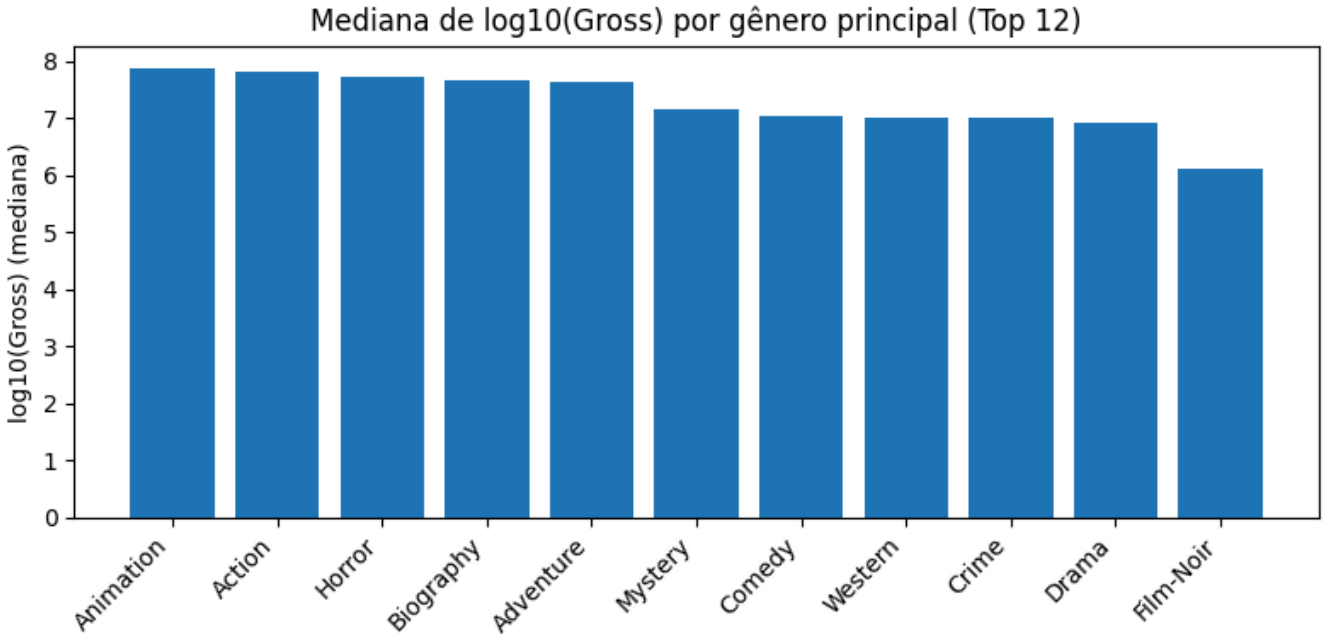


	IMDB_Rating	Meta_score	No_of_Votes	Gross_num	Runtime_min	Released_Year_int
IMDB_Rating	1.000000	0.271374	0.479308	0.099393	0.242751	-0.133257
Meta_score	0.271374	1.000000	-0.020091	-0.030480	-0.031604	-0.339291
No_of_Votes	0.479308	-0.020091	1.000000	0.589527	0.172483	0.246005
Gross_num	0.099393	-0.030480	0.589527	1.000000	0.140002	0.233270
Runtime_min	0.242751	-0.031604	0.172483	0.140002	1.000000	0.165765
Released_Year_int	-0.133257	-0.339291	0.246005	0.233270	0.165765	1.000000

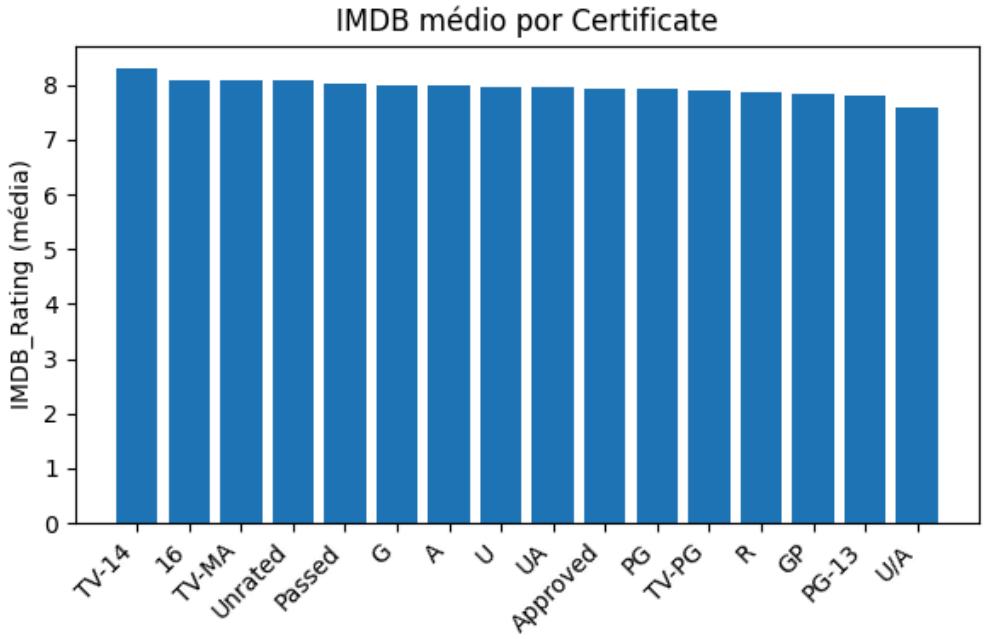
5. Cortes por Gênero e Certificado




```
Genre_Primary
Western      8.350000
Crime        8.016822
Fantasy      8.000000
Mystery      7.975000
Film-Noir    7.966667
Drama        7.952778
Action       7.949419
Biography    7.938636
Adventure    7.937500
Animation    7.930488
Horror       7.909091
Comedy       7.901290
Name: IMDB_Rating, dtype: float64
```




```
Genre_Primary
Animation    75082668.0
Action       66208183.0
Horror       51500184.5
Biography    46836394.0
Adventure    44824144.0
Mystery      14378331.0
Comedy       10728127.0
Western      10550000.0
Crime        10095170.0
Drama         8264530.0
Film-Noir     1278625.5
Fantasy             NaN
Name: Gross_num, dtype: float64
```




```
Certificate
TV-14       8.300000
16          8.100000
TV-MA       8.100000
Unrated     8.100000
Passed      8.020588
G           8.000000
A           7.992347
U           7.976923
UA          7.957143
Approved    7.945455
PG          7.927027
TV-PG       7.900000
R           7.869863
GP          7.850000
PG-13       7.797674
U/A         7.600000
Name: IMDB_Rating, dtype: float64
```

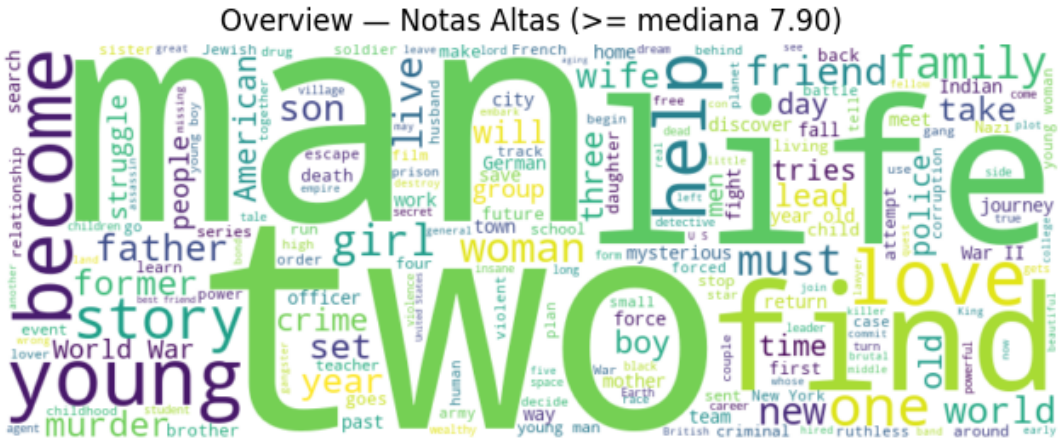
6. Diretores e Elenco

Director	n_filmes	imdb_mean
Milos Forman	2	8.500000
Christopher Nolan	8	8.462500
Francis Ford Coppola	5	8.400000
Peter Jackson	5	8.400000
Charles Chaplin	6	8.333333
Fritz Lang	2	8.300000
Lee Unkrich	2	8.300000
Nitesh Tiwari	2	8.300000
Sergio Leone	6	8.266667
Andrew Stanton	2	8.250000
Damien Chazelle	2	8.250000
Stanley Kubrick	9	8.233333
Akira Kurosawa	10	8.220000
Frank Capra	4	8.200000
Michael Curtiz	2	8.200000

Star	n	imdb_mean
Elijah Wood	3	8.800000
Orlando Bloom	4	8.600000
Mark Hamill	3	8.533333
Madhavan	3	8.466667
Marlon Brando	4	8.425000
Lee J. Cobb	3	8.366667
Charles Chaplin	6	8.333333
James Caan	3	8.333333
Kevin Spacey	5	8.300000
Carrie Fisher	4	8.300000
Takashi Shimura	4	8.300000
Henry Fonda	4	8.275000
Robert Duvall	4	8.275000
Paresh Rawal	4	8.275000
Ralph Fiennes	3	8.266667

7. Overview: nuvem de palavras

A ideia é inspecionar termos mais frequentes em **filmes com notas altas** vs **notas baixas**.



8. NLP rápido: é possível inferir o gênero a partir da Overview?

Sim. Abaixo, um baseline com **TF-IDF + Regressão Logística** prevendo o **gênero principal** a partir da sinopse.

```
Acurácias (cv=5): [0.345 0.34  0.335 0.34  0.337]
Acurácia média: 0.339
```

9. Hipóteses iniciais

- **H1**: Filmes do gênero **Drama** tendem a ter **notas IMDb mais altas**, mas **bilheteria mediana** menor do que gêneros de ação/aventura.

- **H2**: **No_of_Votes** (popularidade) está **positivamente** relacionado com **Gross** e com a **nota IMDb** (efeito de sobrevivência e boca a boca).
- **H3**: **Meta_score** tem **correlação positiva** com **IMDB_Rating** (crítica especializada influencia percepção do público).
- **H4**: **Certificate** do tipo **PG-13** tende a **maximizar bilheteria** por atingir um público mais amplo.
- **H5**: É possível **inferir o gênero** do filme a partir da **Overview** com acurácia superior ao acaso usando TF-IDF (e potencialmente melhor com embeddings).

10. Próximos passos

1. Refinar limpeza de dados (tratamento robusto de `Gross`, `Runtime`, outliers).
2. Enriquecer com dados externos (orçamento, prêmios, janela de lançamento / sazonalidade).
3. Testar diferentes modelos para prever **IMDB_Rating** (regressão): Regressão Linear, Random Forest, Gradient Boosting.
4. Medir performance com **RMSE/MAE** e validação cruzada.
5. Gerar **relatório** consolidado (exportar HTML/PDF) para stakeholders da PProductions.