Desafio Indicium IMDB (Ciência de Dados)

Por Yuri Gonzaga

Questão 1

O conjunto de dados fornecidos possui 999 linhas e 16 colunas (variáveis). As variáveis abaixo, de natureza numérica, possuem as seguintes estatísticas.

Released_Year	**Runtime** (em minutos)	**IMDB_Rating**	**Meta_score**
Ocorrências: 999	999	999	842
Média: 1991	122	7,95	77,97
Mínimo: 1920	45	7,6	28
25%: 1976	103	7,7	70
50%: 1999	119	7,9	79
75%: 2009	137	8,1	87
Máximo: 2020	321	9,2	100

No_of_Votes	**Gross**
999	830
271.621,42	$ 68.082.574,10
25.088,0	$ 1.305,00
55.471,5	$ 3.245.338,50
138.356,0	$ 23.457.439,50
373.167,5	$ 80.876.340,25
2.303.232,0	$ 936.662.225,00

As variáveis Series_Title e Overview são dados textuais. Já Certificate, Genre, Director, Star1, 2, 3 e 4 são dados categorizados, que também podem ser avaliados visualmente, conforme gráficos obtidos através do código Python. Dessa forma, faço as seguintes observações.

- *Certificate*: A maior parte das ocorrências está em U, A, UA e R (nesta ordem).
- *Director*: Possui uma grande quantidade de diretores, alguns deles se destacam por maior frequência. O top 5 é Alfred Hitchcock, Steven Spielberg, Hayao Miyazaki, Martin Scorsese e Akira Kurosawa.
- *Genre*: Cerca de 20 gêneros diferentes aparecem, dos quais Drama é o grande campeão.
- *Stars*: Robert De Niro é o que se destaca como mais presente, seguido de Tom Hanks, Al Pacino, Brad Pitt e Clint Eastwood (pra citar o top 5 apenas).

Para visualizar o relacionamento entre as variáveis numéricas, o gráfico de pairplot foi muito útil, além dos dados de correlação obtidos. Assim, destaco as seguintes observações.

- No_of_Votes tem correlação positiva com IMDB_Rating, o que deve acontecer porque notas maiores no IMDB leva a maior participação de votantes.
- No_of_Votes tem correlação positiva com Gross, o que deve acontecer porque filmes de maior sucesso (e maior arrecadação) atrai mais votantes.
- Meta_score tem correlação negativa com Released_Year, o que deve acontecer porque existem mais avaliações de filmes mais novos. Aqueles mais antigos que se

mantêm ativos neste contexto tendem a ser clássicos e com isso possuírem notas melhores.

Para as variáveis categorizadas, foi feito um levantamento gráfico de suas relações especialmente com Gross e IMDB_Rating, que são de notável interesse desse desafio. Com isso, observamos que:

- Gross é impactada pelo tipo de Certificate, onde a grande campeã é a categoria UA (12 anos), seguida de A (18 anos) e U (livre para todas as idades).
- Gross é impactada por Genre, com os campeões Family, Adventure e Sci-fi.
- Utilizando um top 50 tanto de diretores, como de atores/atrizes, não parece haver grande diferença no Gross. Claro que são todos grandes nomes do cinema mundial.
- As variáveis categorizadas não parecem impactar significativa IMDB_Rating.

Questão 2

a) Utilizando os dados, recomendaria The Dark Knight, por ter sido aquele que dentre um top 5 de número de votantes (logo, considero mais eclético, por ter envolvido mais gente) é o que tem a maior nota IMDB.
b) O ano de lançamento parece impactar no crescimento da arrecadação. Não significa que o simples passar dos anos vai fazer o faturamento crescer, porém os tempos modernos estão associados a maiores investimentos, avanço de tecnologia, estratégias de marketing, etc. Outro fator importante envolvido é a quantidade de votantes, o que mostra como é fundamental a relação com o público, com os fãs, meios de divulgação e assim por diante. Penso que quanto mais um filme está na boca do povo, mais atrai gente pra impactar nas bilheterias. Além disso, a temática do filme (roteiro) também tem papel crucial, vide a influência da classificação indicativa (temas jovens e adultos) e do gênero.
c) O texto presente na coluna Overview pode ser utilizado para obtenção de palavras-chaves e relacioná-las com outras variáveis. Analisar essas relações pode ser muito útil na avaliação deste conjunto de dados. Dessa forma, é possível que seja criado um modelo de aprendizado a partir do texto e usá-lo para inferir gênero.

Questão 3

É possível prever a nota do IMDB através de um modelo preditivo de regressão, já que estamos pensando num valor numérico. Buscando maior precisão do modelo, foi utilizada a regressão polinomial. As variáveis escolhidas foram No_of_Votes, Meta_score e Runtime, por possuírem as maiores correlações com IMDB_Rating. No_of_Votes foi submetida a uma transformação logarítmica para que seus valores fiquem mais normalizados em relação aos demais. A métrica avaliativa para o modelo em questão foi a MAE (Mean Absolute Error) pela sua simplicidade, dado a dinâmica do desafio.

Questão 4

Com base no modelo descrito na questão anterior e codificado em Python, a resposta é 9,1303068.