Análise de dados sobre os filmes ao passar dos anos

Analysis of data about films over the years

Luiz Guilherme Teixeira Bim[1], Fernando José Antonio[2]

RESUMO

Usando um banco de dados do Kaggle, esta pesquisa investiga dados relativos a filmes encontrados no IMDb. Foi feito um recorte de tempo entre 1930 e 2019, além de uma limpeza de filmes sem avaliação. Utilizando a plataforma Anaconda e a linguagem Python, variáveis individuais e sua interação foram analisadas através de vários métodos. Além disso, foi empregado o método *bootstrap* para aumentar a confiabilidade dos parâmetros importantes, permitindo estimativas de intervalos de confiança. Ao longo das décadas, houve estabilidade nas classificações atribuídas aos filmes. Isto sugere a existência potencial de um fenômeno em que as preferências do público permanecem consistentes. As classificações dos filmes são distribuídas de uma forma que se assemelha a uma curva gaussiana, e o desvio padrão ratifica que as classificações estão distribuídas de forma aleatória. Além disso, a análise do número de filmes lançados por ano mostrou um aumento constante ao longo do tempo.
PALAVRAS-CHAVE: análise de dados; IMDb; python.

ABSTRACT

Using a Kaggle database, this research investigates data relating to films found on IMDb. A time frame was made between 1930 and 2019, besides cleaning unrated films. Using the Anaconda platform and the Python language, individual variables and their interaction were analyzed using several methods. Furthermore, the bootstrap method was used to increase important parameter reliability, allowing confidence interval estimates. Over the decades, there has been stability in the ratings given to films. This suggests the potential existence of a phenomenon in which public preferences remain consistent. Film ratings are distributed similarly to a Gaussian curve, and the standard deviation confirms that the ratings are distributed randomly. Furthermore, analysis of the number of movies released yearly showed a steady increase over time.
KEYWORDS: data analysis; IMDb; python.

INTRODUÇÃO

Diversas abordagens para identificar e quantificar padrões em dados de várias áreas do conhecimento foram criadas pelo uso dos conceitos e técnicas da física estatística. Em particular, vários campos, como física, biologia e ciências sociais, têm estudado sistemas complexos, que são compostos por várias partes interagindo entre si e que podem demonstrar comportamentos coletivos em várias escalas sobrepostas (Antonio, Mendes, Thomaz, 2011).

Neste estudo, colocamos nosso foco na investigação de informações associadas a filmes da base de dados do IMDb. Esses dados foram obtidos através da plataforma Kaggle, uma fonte online que compartilha várias coleções de dados de forma gratuita. O conjunto de dados engloba detalhes como título, categoria, diretor, elenco, avaliação média e desempenho de bilheteria dos filmes. Realizamos um procedimento de limpeza na base de dados, onde procedemos à eliminação de duplicatas e filmes sem avaliação, o

[1] Voluntário. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: lbim@alunos.utfpr.edu.br. ID Lattes: https://lattes.cnpq.br/3503661492001604
[2] Docente no Departamento Acadêmico de Ciências da Natureza. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: fjantonio@utfpr.edu.br. ID Lattes: http://lattes.cnpq.br/2833172211868473.

que nos permitiu obter um conjunto de dados final para posterior análise. Além disso, também restringimos a nossa pesquisa aos filmes lançados entre 1930 e 2019.

MÉTODOS

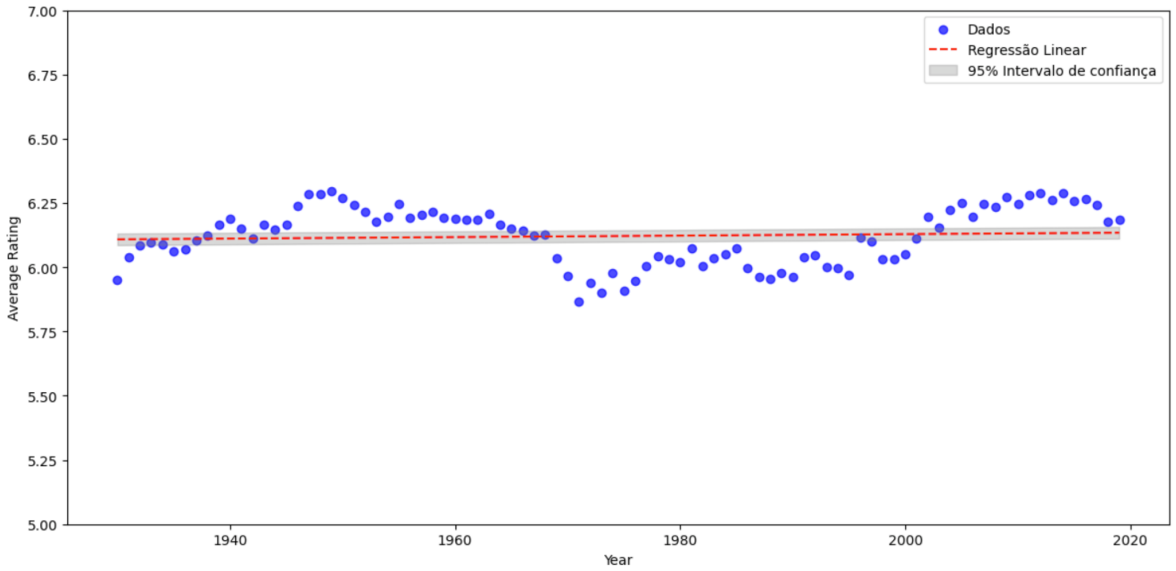
Para a análise dos dados, foi utilizada a linguagem de programação Python, juntamente com a plataforma de ciência de dados Anaconda (Python, 2023; Anaconda, 2023). Primeiramente, examinamos cada variável separadamente, sempre comparando os dois conjuntos de dados disponíveis. Em seguida, investigamos a possível correlação entre pares de variáveis dentro do mesmo conjunto de dados.

Outra ferramenta que empregamos foi o método de *bootstrap*, uma abordagem estatística que tem como objetivo a obtenção de intervalos de confiança para parâmetros de interesse. Isso é alcançado através da reamostragem do conjunto de dados original. Essa técnica estima a distribuição de amostragem ao coletar uma quantidade significativa de amostras com reposição a partir de uma única amostra aleatória, chamada de reamostra (Efron e Tibshirani, 1993).

RESULTADOS

A primeira parte da análise foi entender investigar como a classificação média dos filmes lançados em cada ano evoluiu ao longo do tempo. A Figura 1 revela uma tendência global de estabilidade nessas classificações médias dos filmes lançados anualmente de 1930 a 2019. Apesar de diferentes tendências locais, um modelo de regressão linear para todo o conjunto de dados, levou a um coeficiente de inclinação baixo, de 0,0003, indicando que as médias das classificações se mantiveram ao longo desse período.

Figura 1 – Evolução temporal da classificação média dos filmes. A linha tracejada é um ajuste linear aos dados, e tem inclinação 0,0003. A região em cinza é um intervalo de confiança de 95%.



Fonte: Autoria própria (2023)

O método de *bootstrap* foi usado para calcular o intervalo de confiança de 95% em torno do valor médio reforça a consistência desses resultados, sugerindo que essa estabilidade não é devido a flutuações aleatórias. Logo, pode-se concluir que se destaca a estabilidade das classificações médias dos filmes ao longo das décadas, indicando a possível presença de um fenômeno de estabilidade nas preferências do público ou na qualidade média das produções cinematográficas ao longo do tempo.

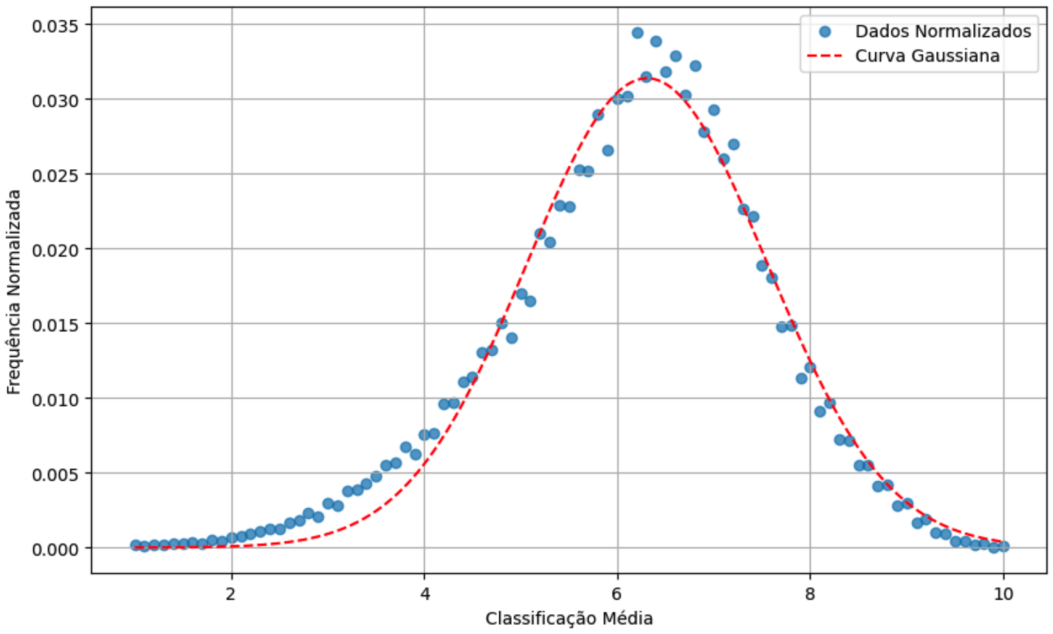
Tendo em vista que a classificação média se mantém distribuída em torno de um valor médio, o próximo passo foi investigar o padrão dessas flutuações. O resultado está disposto na Figura 2, em que os pontos azuis representam uma distribuição das classificações de filmes. A presença de um único pico na distribuição reforça que a distribuição das notas é uma distribuição aleatória em torno do valor médio. Essa distribuição exibe características próximas a uma curva gaussiana, também conhecida como distribuição normal, cuja densidade de probabilidade é dada pela Eq. (1)

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} Exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right], \quad (1)$$

em que μ e σ são a média e o desvio padrão.

O valor médio das classificações calculado foi $\mu = 6,31$, que é uma métrica central que indica o ponto em que a maioria das classificações se concentra. Isso sugere que a maioria dos filmes tende a receber uma classificação próxima a esse valor médio. O desvio padrão calculado foi $\sigma = 1,24$, que é uma medida de dispersão que indica o grau de variabilidade nas classificações dos filmes. Um desvio padrão relativamente baixo sugere que as classificações estão relativamente próximas do valor médio, o que é consistente com a forma estreita da curva gaussiana.

Figura 2 – Comparação da distribuição das notas de avaliação dos filmes

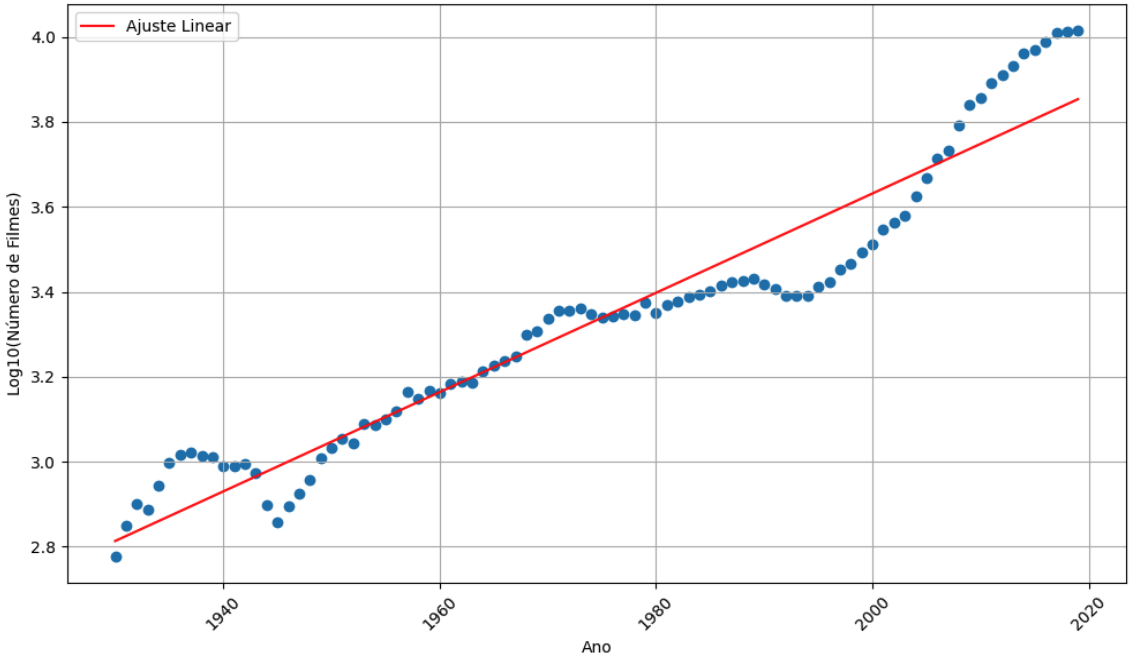


Fonte: Autoria própria (2023)

Usando os valores da média e desvio padrão dos dados e o método da máxima verossimilhança, obtivemos uma curva gaussiana, ilustrada na Figura 2, em comparação com os dados. Tendo em vista que os dados são exibidos sem normalização a Eq. (1) foi multiplicada pelo total de filmes no período.

O terceiro passo da análise foi investigar como o volume de títulos lançados anualmente varia. A Figura 3 deste artigo apresenta o número de filmes lançados no ano apresentado em escala logarítmica de base 10. A escolha de utilizar uma escala logarítmica pode ser justificada por é frequentemente utilizada quando se lida com dados que abrangem várias ordens de magnitude. A transformação logarítmica comprime os valores extremos, tornando-os mais visíveis e facilitando a análise de tendências em diferentes escalas. Na Figura 3, percebe-se diversas tendências locais. Porém, um ajuste linear aos dados levou a uma reta cujo coeficiente de inclinação é 0,0117. A regressão linear aplicada sugere que houve um aumento global quase exponencial no número de filmes lançados por ano. Em particular, após o final da década de 1990, esse crescimento tornou-se ainda mais pronunciado e mais lento a partir da década de 2010.

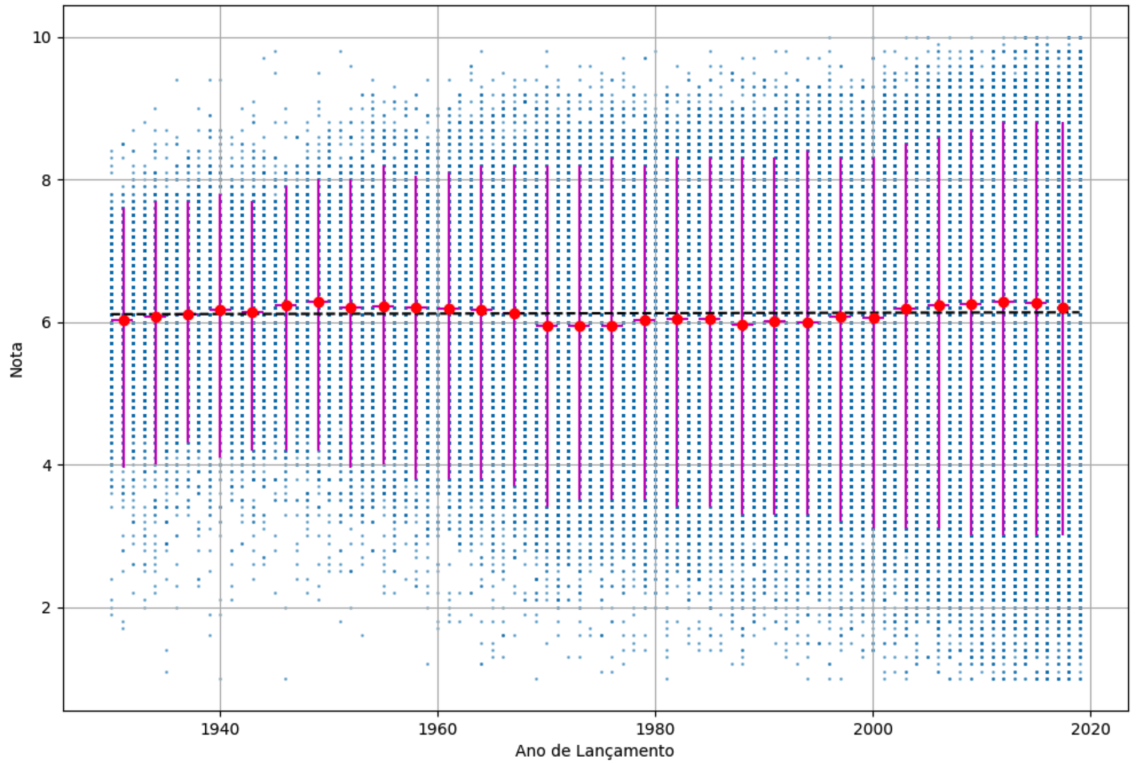
Figura 3 – Evolução do número de filmes lançados por ano.



Fonte: Autoria própria (2023)

Por fim, a última análise foi investigar a flutuação nas investigações médias dos filmes considerando janelas temporais menores. A Figura 4 deste artigo apresenta a avaliação dos filmes em função do ano de lançamento agrupado em janelas temporais equidistrandes e a avaliação média foi computada dentro de cada uma dessas janelas, sendo computado um intervalo de confiança de 95% por meio do método de *bootstrap,* o intervalo de confiança para a inclinação foi de ±0,001576 e para o intercepto foi de ±3,111889.

Figura 4 – Avaliação dos filmes em função do ano de lançamento.



Fonte: Autoria própria (2023)

A análise identificou uma tendência mínima de aumento nas avaliações médias dos filmes ao longo do tempo, representada por uma reta de inclinação 0,000296, obtida por regressão linear. Naturalmente, esse resultado é consistente com o resultado obtido a partir da Figura 1. Um ponto a se destacar é que os intervalos de confiança em torno dos valores médios têm aumentado com o passar dos anos, indicando que o desvio padrão na distribuição das avaliações continua aleatória, mas está ficando mais larga. Isso pode ser associado a um padrão menos consistente no padrão de qualidade dos longa-metragens lançados com o passar do tempo.

CONCLUSÃO

Neste estudo, podemos observar uma estabilidade nas médias de classificação dos filmes ao longo das décadas. Isso sugere que a qualidade de produção dos filmes e as preferências dos avaliadores permaneceram consistentes.

A análise da distribuição das avaliações revela uma semelhança com uma curva gaussiana, com uma tendência de crescimento ao longo do tempo. Além disso, observou-se um aumento quase exponencial no número de filmes lançados, especialmente entre as décadas de 1990 e 2010. Esses resultados constituem uma base para pesquisas futuras.

Agradecimentos

Agradeço ao meu orientador Fernando José Antonio, pelo apoio e orientação que me guiaram durante o desenvolvimento deste artigo. As ideias neste artigo foram moldadas e refinadas por sua experiência e conhecimento. Além disso, gostaria de agradecer à UTFPR – Universidade Tecnológica Federal do Paraná – pela oportunidade de estudos proporcionada.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

SANGWAN, Ashirwad. IMDb Dataset. Disponível em: https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset?select=title.ratings.tsv.gz. Acesso em: 20 ago. 2023.

ANACONDA. 2023. Disponível em: https://www.anaconda.com/. Acesso em: 10 abr. 2023.

EFRON, Bradley; TIBSHIRANI, Robert J. An introduction to the bootstrap. New York: John Wiley & Sons, 1993.

PYTHON. 2023. Disponível em: https://www.python.org/. Acesso em: 10 abr. 2023.

ANTONIO, FJ; Mendes, RS; Thomaz, SM. Identifying and modeling patterns of tetrapod vertebrate mortality rates in the Gulf of Mexico oil spill. Aquatic Toxicology, v. 105, p. 177-179, 2011.