ISSN: 2317-0840

Os filmes com maior arrecadação são os mais bem avaliados?

Bruno M. Silva

Universidade Federal de Sergipe. E-mail: brunomeloslv@gmail.com.

Resumo: A paixão global pelo cinema e a importância da indústria tanto para o mercado financeiro quanto para a população são evidentes. Portanto, é necessário realizar estudos nesse setor, buscando insights relevantes. Nesse contexto, este estudo se concentra em tirar algumas conclusões sobre a indústria bilionária do cinema. O objetivo é investigar se os filmes de maior arrecadação são também os melhor avaliados, uma vez que esses costumam atrair um grande público. Para alcançar esse objetivo, utilizamos um conjunto de dados que contém os 100 melhores filmes, de acordo com a avaliação da população no site IMDb, abrangendo o período de 1932 a 2015. Esses dados foram obtidos publicamente no site Kaggle. A análise foi realizada utilizando o método de análise de correspondência, também conhecido como ANACOR, em conjunto com o software R-Studio, que permitiu realizar os cálculos necessários. Esse método consiste em examinar a relação entre as variáveis analisadas no estudo: arrecadação do filme, número de votos e média das notas atribuídas aos filmes no site. Para tornar as variáveis numéricas em categorias, utilizamos os percentis, pois a ANACOR é adequada para esse tipo de variável. Através da aplicação dessa técnica estatística, é possível obter evidências que sustentam a crença de que filmes com maior arrecadação tendem a receber mais votos, entretanto, arrecadar mais não significa ter notas melhores.

Palavras-chave: Coleta de dados. Ciência de dados. Análise de dados. Análise de correspondência. Estatística aplicada.

Are the films with the highest grossing the best rated?

Abstract: *The global passion for cinema and the importance of the industry both for the financial market and the population are evident. Therefore, it is necessary to conduct studies in this sector, seeking relevant insights. In this context, this study focuses on drawing some conclusions about the billion-dollar film industry. The objective is to investigate whether the highest-grossing films are also the best-rated, as they usually attract a large audience. To achieve this goal, we used a dataset containing the top 100 films according to the population's ratings on the IMDb website, spanning from 1932 to 2015. This data was publicly obtained from the Kaggle website. The analysis was conducted using the correspondence analysis method, also known as ANACOR, in conjunction with the R-Studio software, which allowed us to perform the necessary calculations. This method involves examining the relationship between the variables analyzed in the study: film revenue, number of votes, and average ratings given to films on the website. To categorize the numerical variables, we used percentiles, as ANACOR is suitable for this type of variable. Through the application of this statistical technique, it is possible to obtain evidence that supports the belief that higher-grossing films tend to receive more votes; however, earning more does not necessarily mean better ratings.*

Keywords: Data collection. Data science. Data analysis. Correspondence analysis. Applied statistics.

Introdução

A indústria de entretenimento é uma das que mais cresce dentro do mercado mundial, é o maior opioide da população que necessita de uma distração para manter o discernimento necessário para o cotidiano, e esse fato demonstra um possível motivo desse crescimento.

Entre os temas estudados no campo do lazer, a história desse fenômeno é certamente um dos mais negligenciados. Além do número limitado de trabalhos sobre o assunto, também há certa fragilidade empírica nos resultados apresentados (DIAS, 2018). Dentro do universo do entretenimento, existem várias opções, incluindo shows, filmes, televisão e esportes, como afirmado pelo Cambridge Dictionary. Neste artigo em particular, vamos focar nos filmes e no mercado cinematográfico como um todo.

A Motion Picture Association – MPA emitiu em 2021 o seu relatório anual falando sobre o cenário do entretenimento audiovisual e informou que em 2021 arrecadou U$ 99,7 bilhões, isso significa um crescimento de 24% em relação ao ano anterior. O crescimento que mais chama atenção é do digital, em 2019 arrecadou U$ 28,7 bilhões e em 2021 U$ 71,9 bilhões que corresponde a um crescimento de 151%.

O crescimento exponencial dos serviços de streaming é evidente, e para fornecer uma perspectiva mais clara sobre o aumento de 151%, a Netflix merece destaque como a força motriz por trás desse crescimento. Fundada em 1997, a Netflix atualmente conta com mais de 200 milhões de assinantes, o que, para se ter uma ideia, é como se todos os brasileiros tivessem assinado um de seus pacotes. Hoje em dia, existem inúmeras plataformas de streaming de todos os tipos, desde esportes de combate até futebol, filmes, educação e muito mais.

Visto a paixão mundial e a importância da indústria, seja para o mercado financeiro ou população, fica evidente a necessidade de estudos quanto a esse mercado e tentar encontrar insights. Deste modo, esse estudo está concentrado em tirar algumas conclusões em volta dessa indústria bilionária chamada de cinema.

Diante da paixão global e da importância da indústria tanto para o mercado financeiro quanto para a população, fica evidente a necessidade de estudos nesse campo e a busca por insights. Assim, este estudo visa tirar algumas conclusões sobre essa indústria bilionária chamada de cinema.

No contexto brasileiro, a produção cinematográfica brasileira foi intensificada durante as décadas de 1970 e 1980, graças à intensa e direta atuação do Estado. Principalmente porque o regime militar, dentro de seus princípios de centralização política e administrativa, estabeleceu um projeto de institucionalização cultural em âmbito nacional (AMANCIO, 2007).

Após uma longa crise que durou praticamente toda a década de 1980, a Embrafilme, principal instrumento das ações do governo federal no campo cinematográfico, foi fechada em um dos primeiros atos do presidente Fernando Collor de Mello (AUTRAN, 2009).

Segundo Autran, Esta situação perdurou até 1993, quando já no governo Itamar Franco foi aprovada a Lei do Audiovisual, instrumento que bem ou mal permitiu o início do reaquecimento da produção de longas-metragens, a qual em alguns anos voltou a um patamar significativo em termos numéricos.

Observa-se a evolução do cinema brasileiro, e podem ser mencionados filmes que obtiveram sucesso tanto no Brasil quanto no exterior, como Central do Brasil, Tropa de Elite e Cidade de Deus. Este último, aliás, foi considerado o filme mais visto em uma pesquisa do site americano Internet Movie Database - IMDb, e é relevante mencionar que também recebeu quatro indicações ao Oscar.

Material e Métodos

Os dados foram obtidos do Kaggle, uma plataforma on-line de cientistas de dados e profissionais de aprendizado de máquina, subsidiária da Google LLC (WIKIPEDIA, 2022). Lá, é possível fazer o download de diversos conjuntos de dados e até participar de competições em que os proprietários dos conjuntos de dados oferecem prêmios em dinheiro para a melhor solução do caso. O conjunto de dados em questão foi carregado pelo usuário Mrityunjay Pathak, que realizou um web scraping utilizando Python e disponibilizou os dados publicamente.

O banco de dados contém informações sobre os 100 filmes mais bem avaliados na plataforma IMDb, abrangendo o período de 1931 a 2015. Entre as variáveis presentes, temos o índice, nome do filme, ano de lançamento, categoria, duração, gênero, nota, número de votantes e valor arrecadado.

O objetivo é realizar uma análise de correspondência simples entre a variável "votos" e as notas, também conhecida como ANACOR. Essa técnica de análise envolve o estudo da associação entre as categorias das duas variáveis e a intensidade dessa associação, utilizando uma tabela de dados cruzados, também conhecida como tabela de contingência ou tabela de correspondência. Essa associação pode ser expressa pela seguinte fórmula:

$$N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

Para a confirmação da existência de associação entre as duas variáveis categóricas e entre as suas categorias é utilizado o teste qui-quadrado e a análise dos residuos. Sendo assim para dado número de graus de liberdade e determinado nível de significância, caso o valor da estatística qui-quadrado for maior que seu valor crítico, pode-se afirmar que existe associação estatisticamente significante entre duas variáveis categóricas. O cálculo realizado para conseguir essa conclusão é:

$$x^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left[n_{ij} - \left(\frac{\sum c_j \sum l_i}{N}\right)\right]^2}{\left(\frac{\sum c_j \sum l_i}{N}\right)}, com\ (I-1)(J-1) graus\ de\ liberdade$$

com as hipóteses:

H_0: *As duas variáveis categóricas se associam de forma aleatória*

H_1: *A associação entre as duas variáveis categóricas não se dá de forma aleatória*

Os gráficos com os componentes principais representam uma técnica de visualização que utiliza a análise de componentes principais para reduzir a dimensionalidade dos dados e permitir a representação visual das relações entre as variáveis. Esses gráficos são construídos utilizando projeções dos dados em um espaço de menor dimensão, onde os eixos principais representam as combinações lineares das variáveis originais. (JAMES et al., 2013)

Para o experimento será utilizado o software estatístico R Studio, que agora faz parte da Posit, ele é famoso no meio estatístico e principalmente com avanço da ciência de dados. É uma ferramenta de código aberto que irá auxiliar com a análise dos dados.

A análise de correspondência é uma técnica estatística utilizada para explorar e visualizar a associação entre variáveis categóricas em uma tabela de contingência. Ela permite identificar padrões, tendências e relações entre as categorias das variáveis, através da representação gráfica dos dados (GREENACRE).

Outra mudança que será feita é a alteração do tipo de variável, a análise de correspondência é para variáveis categórica e as variáveis utilizadas aqui são numéricas, bem como o número de votos com a nota média dos votantes. Para o estudo e com o objetivo de testar se a análise de correspodência teria um bom resultado dentro dessa base, será realizado uma alteração fazendo uma divisão percentual.

Tabela 1. Divisão das Variáveis Categóricas

	Menores	Médias	Maiores
%	x <= 25%	25% < x <= 75%	x > 75%

Fonte: Do autor.

Ambas as variáveis foram divididas em 3 categorias, como pode ser observado na tabela anterior, menores_notas/menores_votos que irá conter os menores 25% dos valores, depois as notas_medias/votos_medios que conterá as os valores maiores que os 25% e menor que 75% geral, a última serão as 25% maiores notas, para facilitar observe a tabela abaixo.

Resultados e Discussão

Feitas as alterações necessárias para iniciar, significa que pode partir para a tabela de contingência que tem o objetivo de demonstrar alguns itens de suma importância para a sequência da análise. Ela irá demonstrar a importância da categoria para o modelo, o peso de cada linha ou coluna para dimensão e a importância da dimensão para o modelo. Além do teste valor p do teste qui-quadrado indicará se faz sentido continuar o experimento até plotar os dados ou não.

Tabela 2. Tabela de contingência Notas x Votos

notas	*Votos*			***Total***
	maiores_votos	menores_votos	votos_medios	
maiores_notas	14	0	4	18
	5	5	9	18
	77.8 %	0 %	22.2 %	100 %
	56 %	0 %	8.2 %	18.2 %
menores_notas	0	12	14	26
	7	7	13	26
	0 %	46.2 %	53.8 %	100 %
	0 %	48 %	28.6 %	26.3 %
notas_medias	11	13	31	55
	14	14	27	55
	20 %	23.6 %	56.4 %	100 %
	44 %	52 %	63.3 %	55.6 %
Total	25	25	49	99
	25	25	49	99
	25.3 %	25.3 %	49.5 %	100 %
	100 %	100 %	100 %	100 %

χ^2=39.261 · df=4 · Cramer's V=0.445 · p=0.000

Fonte: Do autor.

O primeiro valor da tabela (em preto) representa os valores observados, enquanto o segundo valor representa os valores esperados. As porcentagens correspondem aos percentuais por linha e coluna, respectivamente. A estatística que deve ser analisada com mais detalhes é o p-valor, que foi menor que 0,05. Isso indica que o experimento pode prosseguir, pois está no cenário "ideal". A segunda estatística a ser observada é o V Cramer, que retornou o valor de 0,445. Segundo Acastat (2023), o grau de associação pode ser mensurado através da tabela.

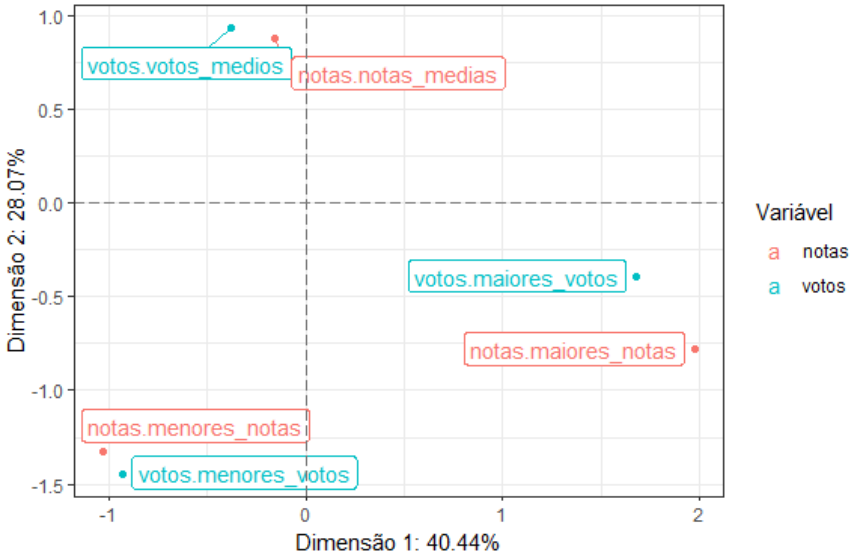
Outra estatística importante a ser observada é o teste qui-quadrado. Com um valor crítico de 5%, o valor tabelado é de 9,488, enquanto o valor calculado foi de 39,261. Portanto, se o valor calculado é maior que o valor crítico, rejeita-se a hipótese nula e aceita-se a hipótese alternativa. Não há evidências para rejeitar a hipótese alternativa, indicando que a associação entre as duas variáveis categóricas não ocorre de forma aleatória.

Assim, nossos dados apresentam uma associação de nível médio e justificam a continuação dos estudos. Os dados podem ser plotados para visualizar de forma mais evidente como essa relação entre as variáveis estudadas funciona.

Pode ser observado no gráfico acima que as menores notas ficaram mais próximas dos menos votados, as notas médias dos que tiveram uma votação média e por último os mais votados próximo das maiores notas.

É um resultado interessante visto que quanto maior o número de votantes esperamos uma variabilidade maior e consequentemente uma nota mais ao meio, mas também estatisticamente falando espera-se uma distribuição mais próxima da normal quanto maior o número de votos.

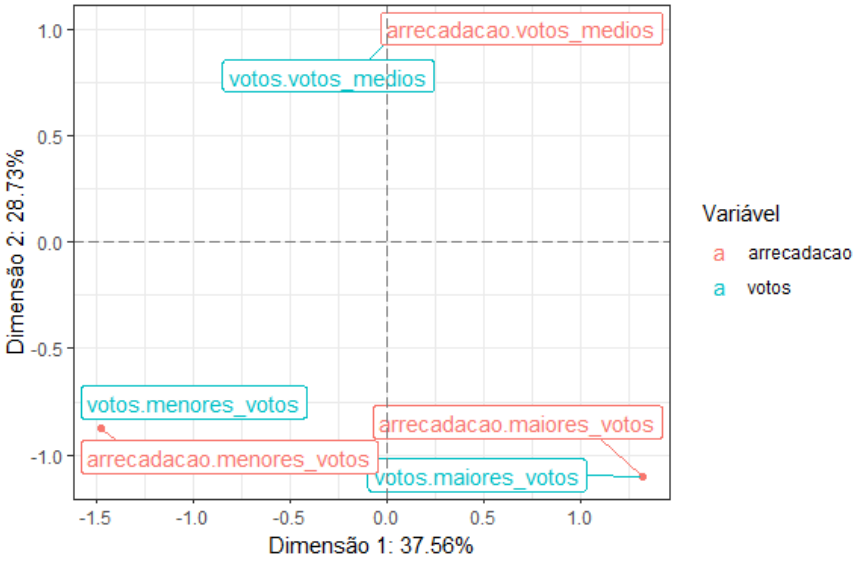
Imagem 1. Mapa Perceptual Notas x Votos



Fonte: Do autor.

Então pode ser teorizar que as pessoas que mais gostam dos filmes são as que mais vão até a plataforma votar e consequentemente os filmes que atraem um menor público acabam por sofrer as consequências de um número de votos menor e podemos comprovar através dos dados que quanto maior a arrecadação maior é o número de votantes como pode ser visto na imagem abaixo.

Imagem 2. Mapa Perceptual Arrecadação x Votos



Fonte: Do autor.

Para validar o gráfico pode ser avaliado as estatísticas dentro da tabela de contingência entre essas duas variáveis e como pode ser observado abaixo as estatísticas determinem a rejeição da hipótese nula e aceitação da alternativa que indica que os dados não tem uma associação aleatória, ou seja, existe uma relação mesmo que baixa entre as variáveis.

Caso seja necessário confirmar a relação entre a arrecadação e a nota, é possível executar o script para obter a tabela de contingência entre as variáveis em questão. Ao analisar a tabela a seguir, pode-se observar que a hipótese nula (H0) é aceita com base no teste qui-quadrado, uma vez que o valor calculado é menor que o valor crítico.

Tabela 3. Tabela de contingência Arrecadação x Votos

arrecadacao	*votos*			***Total***
	maiores_votos	menores_votos	votos_medios	
maiores_votos	13	1	11	25
	6	6	12	25
	52 %	4 %	44 %	100 %
	52 %	4 %	22.4 %	25.3 %
menores_votos	1	14	10	25
	6	6	12	25
	4 %	56 %	40 %	100 %
	4 %	56 %	20.4 %	25.3 %
votos_medios	11	10	28	49
	12	12	24	49
	22.4 %	20.4 %	57.1 %	100 %
	44 %	40 %	57.1 %	49.5 %
Total	25	25	49	99
	25	25	49	99
	25.3 %	25.3 %	49.5 %	100 %
	100 %	100 %	100 %	100 %

χ^2=27.180 · df=4 · Cramer's V=0.371 · p=0.000

Fonte: Do autor.

Tabela 4. Tabela de contingência Votos x Arrecadação

Votos	*arrecadacao*			***Total***
	maiores_votos	menores_votos	votos_medios	
maiores_votos	13	1	11	25
	6	6	12	25
	52 %	4 %	44 %	100 %
	52 %	4 %	22.4 %	25.3 %
menores_votos	1	14	10	25
	6	6	12	25
	4 %	56 %	40 %	100 %
	4 %	56 %	20.4 %	25.3 %
votos_medios	11	10	28	49
	12	12	24	49
	22.4 %	20.4 %	57.1 %	100 %
	44 %	40 %	57.1 %	49.5 %
Total	25	25	49	99
	25	25	49	99
	25.3 %	25.3 %	49.5 %	100 %
	100 %	100 %	100 %	100 %

χ^2=27.180 · df=4 · Cramer's V=0.371 · p=0.000

Fonte: Do autor.

Portanto, como pode ser observado na imagem 5, estatisticamente falando, não há uma correspondência entre a arrecadação e a média das notas dos filmes.

Conclusão

O estudo por se tratar de váriaveis númericas existiam "n" possibilidades com modelos supervisionados e não supervisionados, mas usando a ideia de pensar fora da caixa e até por uma questão de estudo foi utilizado a ANACOR simples e por esse fato pode ser levantado vários questionamentos do melhor modelo.

Nesse estudo, ceteris paribus, o que pode ser observado é que as notas tem associação com o número de votos quanto maior as notas também é maior o número de votantes e assim podendo supor que quanto mais pessoas assistem e gostam mais elas vão votar. Sendo o inverso verdadeiro, quanto menor o número de pessoas menor o número de votantes e joga a media do filme para baixo.

Isso pode ser confirmado pela a arrecadação, ou seja, quanto maior a arrecadação maior o número de votos porque significa que teve uma maior exposição as grandes massas, seja através das mídias sociais ou televisivas e entre outras que podem acontecer a depender da epoca de lançamento do filme.

Entretanto o que trouxe maior curiosidade foi que arrecadar mais não significa ter notas melhores e foi comprovado isso com a tabela 4, em que estatisticamente não há associação entre as variaveis citadas, ou seja, arrecadar bem não significa sucesso de um filme e isso pode prejudicar ou ser percebido na sequência de uma franquia.

Referências

ACASTAT. acastat. [S.l.], 23 jan. 2023. Disponível em: https://shorturl.at/nyGQX. Acesso em: 13 jun. 2023.

AMANCIO, T. Pacto cinema-Estado: os anos Embrafilme. ALCEU, v. 8, n. 16, p. 173-184, jul./dez. 2007.

ANÁLISE de correspondência para avaliação do perfil de mulheres na pós-menopausa e o uso da terapia de reposição hormonal. Cad. Saúde Pública, Rio de Janeiro, v. 20, n. 1, p. 100-108, jan./fev. 2004.

ANÁLISE de Correspondência: Uma Aplicação do Método à Avaliação de Serviços de Vacinação. Cad. Saúde Públi., Rio de Janeiro, v. 8, n. 3, p. 287-301, jul./set. 1992.

AUTRAN, A. O cinema brasileiro contemporâneo diante do público e do mercado exibidor. Significação: revista de cultura audiovisual, vol. 36, núm. 32, p. 119-135, jul./dez. 2009.

CANAL TECH. Canal Tech. [S.l.], 21 jan. 2023. Disponível em: https://canaltech.com.br/empresa/netflix/. Acesso em: 13 jun. 2023.

DIAS, C. História e historiografia do lazer. Revista de História do Esporte, v. 3, n. 1, p. 1-26, jan./jun. 2018.

GONÇALVES, M. T.; SANTOS, S. R. dos. Aplicação da análise de correspondência à avaliação institucional da FECILCAM. In: ENCONTRO DE PRODUÇÃO CIENTÍFICA E TECNOLÓGICA, 20-23 out. 2009.

JAMES, G. et al. An Introduction to Statistical Learning: with Applications.

MOTION PICTURE ASSOCIATION - MPA. Theme Report. [S.l.], 21 jan. 2023. Disponível em: https://www.motionpictures.org/wp-content/uploads/2022/03/MPA-2021-THEME-Report-FINAL.pdf. Acesso em: 13 jun. 2023.

PEREIRA, C. Análise de Dados Qualitativos Aplicados às Representações Sociais. Psicologia, v. 15, p. 177-204, 2001.

R CORE TEAM. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2023. ISBN 3-900051-07-0. Disponível em: http://www.R-project.org/. Acesso em: 13 jun. 2023.