Correlação entre o Retorno do Investimento em um Filme e a Nota de Críticos e do Público

Cleiton Pereira da Silva – São Paulo – SP cleitonps.mkt@hotmail.com;
Leila Rabello de Oliveira

Correlação entre o Retorno do Investimento em um Filme e a Nota de Críticos e do Público

Resumo

O estudo avaliará a correlação entre o retorno percentual do investimento financeiro em um filme e as diversas outras variáveis focando em analisar a correlação com as notas atribuídas pelos críticos especializados e pelo público em geral. A avaliação técnica de um filme é correlacionada com diversos fatores, como elenco, direção, roteiro, gênero, orçamento e duração, entre outros. Com o avanço das tecnologias de coleta e análise de dados, é possível explorar amplamente as relações entre essas variáveis e as notas atribuídas pelos críticos e o público para traçar metas que ajudem os filmes a conseguir um retorno percentual do investimento financeiro.

Aqui utilizaremos técnicas de data analytics e machine learning, que nos permitirão examinar um conjunto de dados abrangente de uma ampla lista de filmes, contendo informações como retorno percentual do investimento, elenco, diretor, gênero, orçamento e duração, juntamente com as notas atribuídas pelos críticos e pelo público. Faremos uso de métodos estatísticos e algoritmos de regressão para identificar quais variáveis têm maior influência na determinação de um retorno satisfatório do investimento total.

Este estudo tentará contribuir para uma melhor compreensão das relações entre as variáveis de um filme e as avaliações de críticos e público, fornecendo possíveis insights para a indústria cinematográfica. Além disso, os resultados obtidos podem auxiliar na tomada de decisões, como a seleção de elenco, direção e roteiro, visando alcançar uma recepção mais positiva tanto pelos críticos especializados quanto pelo público em geral.

Palavras-chave: Correlação entre variáveis, Nota de Críticos de cinema, Nota do Público de cinema, Retorno do investimento em um Filme, Data Analytics, Machine Learning.

Abstract

The study will assess the correlation between the percentage return on financial investment in a film and various other variables, with a focus on analyzing the correlation with ratings given by specialized critics and the general public. The technical assessment of a film is correlated with several factors, such as cast, direction, screenplay, genre, budget, and duration, among others. With the

advancement of data collection and analysis technologies, it's now possible to extensively explore the relationships between these variables and the ratings provided by critics and the public, in order to set goals that assist films in achieving a percentage return on their financial investment.

In this study, we will employ data analytics and machine learning techniques, allowing us to examine a comprehensive dataset of a wide range of films. This dataset will encompass information such as percentage return on investment, cast, director, genre, budget, duration, along with ratings from critics and the public. We will utilize statistical methods and regression algorithms to identify which variables have the greatest influence in determining a satisfactory return on the total investment.

This study aims to contribute to a better understanding of the relationships between film variables and assessments by critics and the public, providing potential insights for the film industry. Additionally, the findings obtained can assist in decision-making processes, such as casting, direction, and screenplay selection, with the aim of achieving a more positive reception from both specialized critics and the general audience.

Keywords: Correlation between variables, Movie Critics Rating, Movie Audience Rating, Return on Investment in a Movie, Data Analytics, Machine Learning.

Introdução

A indústria cinematográfica tem despertado um interesse crescente no estudo das relações entre as características de um filme, o retorno do investimento feito nele e as avaliações feitas pelos críticos especializados e pelo público em geral. Diversos fatores, como elenco, direção, roteiro, gênero, orçamento e duração, são considerados na avaliação técnica de um filme. Com o avanço das tecnologias de coleta e análise de dados, é possível explorar de forma abrangente as relações entre essas variáveis e as notas atribuídas pelos críticos e pelo público.

Nesta monografia, utilizaremos técnicas de data analytics e machine learning para examinar um conjunto de dados abrangente de uma ampla lista de filmes. Esse conjunto de dados contém informações como elenco, diretor, gênero, orçamento e duração, além das notas atribuídas pelos críticos e pelo público. Por meio de

métodos estatísticos e algoritmos de regressão, buscaremos identificar quais variáveis possuem maior influência na determinação da nota média.

O objetivo deste estudo é contribuir para uma melhor compreensão das relações entre as características de um filme e as avaliações realizadas por críticos e público. Ao fornecer possíveis insights para a indústria cinematográfica, os resultados obtidos poderão auxiliar na tomada de decisões, como a seleção de elenco, direção e roteiro, visando alcançar um melhor percentual do retorno no investimento e uma recepção mais positiva tanto pelos críticos especializados quanto pelo público em geral.

Por meio da análise aprofundada dessas relações, esperamos contribuir para o avanço do conhecimento no campo do cinema, possibilitando uma melhor compreensão dos fatores que influenciam nestas variáveis importantes.

Material e Métodos

A base de dados utilizada

Um conjunto de dados de filmes foi obtido no sistema Data.World contendo informações abrangente sobre os filmes, como elenco, direção, roteiro, gênero, orçamento, arrecadação mundial, duração, notas atribuídas pelos críticos especializados (Indicada pelo site Metacritics) e pelo público em geral (Indicada pelo site IMDB) e a partir destas variáveis foi possível calcular o retorno do investimento percentual correlacionando a variável orçamento e a de arrecadação de bilheteria global.

Após filtragem das observações que indicavam as variáveis de maior interesse neste estudo, foi obtido uma base de dados com 6760 observações com 13 variáveis pertinentes entre qualitativas e quantitativas que tiveram seus impactos analisados como é descrito abaixo na análise exploratória de cada uma das variáveis, procurando melhor compreender os dados disponíveis, para somente após passar para a seleção do Método Supervisionado de regressão apropriado.

Variáveis qualitativas

Título original (Titulo_original)

A variável título original foi uma variável nominal que apresentou a qual filme os dados se referem e foi classificada no R como Character. Na lista de nomes apresentada não havia respostas não disponíveis [NA].

Gênero (gênero)

A variável gênero foi uma variável nominal que apresentou a qual filme os dados se referem e foi classificada no R como Character. Na lista de nomes apresentada não havia respostas não disponíveis [NA].

A lista original de possibilidade de gêneros se mostrou extensa e redundante em algumas subclassificações como mostra a tabela 1 abaixo:

Tabela 1 – frequência das observações de cada gênero, 25 maiores ocorrências.

Genero	Frequencia	Porcentagem.Freq
Drama	405	5.99112426
Comedy, Drama, Romance	312	4.615384615
Comedy, Drama	292	4.319526627
Comedy	259	3.831360947
Drama, Romance	222	3.284023669
Comedy, Romance	205	3.032544379
Action, Crime, Drama	170	2.514792899
Animation, Adventure, Comedy	162	2.396449704
Crime, Drama, Thriller	139	2.056213018
Action, Adventure, Sci-Fi	120	1.775147929
Action, Adventure, Comedy	107	1.582840237
Crime, Drama, Mystery	105	1.553254438
Horror, Mystery, Thriller	100	1.479289941
Action, Crime, Thriller	99	1.464497041
Action, Comedy, Crime	97	1.434911243
Biography, Drama, History	95	1.405325444
Horror, Thriller	89	1.316568047
Action, Adventure, Fantasy	88	1.301775148
Drama, Thriller	88	1.301775148
Crime, Drama	87	1.286982249
Comedy, Crime, Drama	80	1.183431953
Biography, Drama	77	1.139053254
Horror	76	1.124260355
Action, Adventure, Drama	73	1.079881657
Comedy, Crime	68	1.00591716

Fonte: Banco de dados original

Para possibilitar a transformação destas categorias em dummys, uma reclassificação foi aplicada para reduzir os gêneros avaliados, assim a primeira indicação de gênero de cada observação foi considerada a principal e única e uma nova tabela de frequências foi estabelecida como na tabela 2 para os gêneros compilados:

Tabela 2 - frequência das observações de cada gênero após a redução dos gêneros a suas indicações principais, juntamente com sua porcentagem de ocorrências.

Genero compilado	Frequencia	Porcentagem
Comedy	1808	26.74556213
Drama	1460	21.59763314
Action	1455	21.52366864
Crime	483	7.144970414
Biography	431	6.375739645
Adventure	349	5.162721893
Horror	333	4.926035503
Animation	326	4.822485207
Fantasy	33	0.48816568
Mystery	26	0.384615385
Thriller	15	0.221893491
Romance	11	0.162721893
Sci-Fi	9	0.133136095
Family	8	0.118343195
Western	5	0.073964497
Musical	4	0.059171598
War	2	0.029585799
Film-Noir	1	0.014792899
Music	1	0.014792899

Fonte: Banco de dados original

Tabela 3 - Variação da variavel ROI_percentual para cada gênero principal encontrado no dataframe, onde 100 indica que houve 100% de lucro sobre o orçamento para as gravações e -100 significa que todo o valor investido não foi recupero pela bilheteria nos cinemas.

genero_principal	ROI_percentual.max	ROI_percentual.mean	ROI_percentual.min
Action	29056	194	-100
Adventure	10834	259	-100
Animation	12237	369	-100
Biography	3945	165	-100
Comedy	36552	297	-100
Crime	4002	121	-100
Drama	17291430	12170	-100
Family	7452	1003	-82

Fantasy	895	137	-98
Film-Noir	-100	-100	-100
Horror	3.86E+08	1164277	-100
Music	-56	-56	-56
Musical	6516	1571	-100
Mystery	3603	320	-88
Romance	416	24	-100
Sci-Fi	563	73	-100
Thriller	901	0	-100
War	-52	-70	-87
Western	2400	993	-99

Tabela 4 – 1º e 3º quatis da variavel ROI_percentual para cada gênero principal encontrado no dataframe, onde 100 indica que houve 100% de lucro sobre o orçamento para as gravações e -100 significa que todo o valor investido não foi recupero pela bilheteria nos cinemas.

genero_principal	ROI_percentual.Q1.25%	ROI_percentual.Q3.75%
Action	-25	243
Adventure	-32	305
Animation	36	389
Biography	-59	197
Comedy	-50	251
Crime	-79	159
Drama	-83	191
Family	-72	288
Fantasy	-64	217
Film-Noir	-100	-100
Horror	6	727
Music	-56	-56
Musical	-80	1586
Mystery	21	368
Romance	-99	93
Sci-Fi	-78	-4
Thriller	-100	-87
War	-78	-61
Western	-98	2004

Fonte: Banco de dados original

Língua (lingua)

A variável língua é categórica e indica as línguas usadas durante os diálogos nos filmes, esta variável foi classificada no R como numérica. Na lista de minutos por filme apresentada não havia respostas não disponíveis [NA].

Tabela 5 – frequência das trinta observações mais encontradas de cada língua utilizada nos diálogos dos filmes.

Lingua	Frequencia	Porcentagem.Freq
English	3986	58.96449704
English, Spanish	361	5.340236686
English, French	187	2.766272189
English, German	80	1.183431953
English, Italian	80	1.183431953
English, Russian	72	1.065088757
French	49	0.724852071
Spanish	41	0.606508876
English, Japanese	39	0.576923077
French, English	36	0.532544379
English, Mandarin	35	0.517751479
English, Arabic	32	0.473372781
English, Latin	29	0.428994083
English, French, Spanish	24	0.355029586
English, Ukrainian	24	0.355029586
Spanish, English	24	0.355029586
English, American Sign Lan	22	0.325443787
English, Hebrew	21	0.310650888
Japanese	20	0.295857988
English, French, German	19	0.281065089
English, Cantonese	17	0.25147929
English, German, French	16	0.236686391
English, Portuguese	16	0.236686391
Mandarin	16	0.236686391
English, French, Italian	15	0.221893491
German	15	0.221893491
Korean	13	0.192307692
English, Italian, Spanish	12	0.177514793
English, Spanish, French	12	0.177514793
Portuguese	11	0.162721893

Fonte: Banco de dados original

Diretor (diretor)

A variável diretor é categórica e indica os profissionais que comandaram as filmagens em cada um dos filmes, esta variável foi classificada no R como Factor. Na lista de diretores por filme apresentada não havia respostas não disponíveis [NA].

Tabela 6 – frequência dos sessenta diretores presentes em observações mais encontradas de cada trabalho responsável nos filmes.

Diretor	Frequencia	Porcentagem.Freq
Clint Eastwood	34	0.502959
Woody Allen	34	0.502959
Steven Spielberg	31	0.45858
Steven Soderbergh	26	0.384615
Martin Scorsese	24	0.35503
Ridley Scott	24	0.35503
Ron Howard	22	0.325444
Brian De Palma	18	0.266272
Renny Harlin	18	0.266272
Robert Zemeckis	18	0.266272
Tim Burton	18	0.266272
Barry Levinson	17	0.251479
Oliver Stone	17	0.251479
Spike Lee	17	0.251479
Francis Ford Coppola	16	0.236686
Joel Schumacher	16	0.236686
John Carpenter	16	0.236686
Rob Reiner	16	0.236686
Richard Linklater	15	0.221893
Tyler Perry	15	0.221893
David Cronenberg	14	0.207101
Ivan Reitman	14	0.207101
Richard Donner	14	0.207101
Sam Raimi	14	0.207101
Tony Scott	14	0.207101
Wes Craven	14	0.207101
Gus Van Sant	13	0.192308
Michael Bay	13	0.192308
Mike Nichols	13	0.192308
Robert Altman	13	0.192308
Robert Rodriguez	13	0.192308
Roland Emmerich	13	0.192308
Stephen Frears	13	0.192308
Antoine Fuqua	12	0.177515
Dennis Dugan	12	0.177515
Garry Marshall	12	0.177515

Kevin Smith	12	0.177515
Paul Schrader	12	0.177515
Stephen Herek	12	0.177515
Walter Hill	12	0.177515
Ang Lee	11	0.162722
Bobby Farrelly, Peter Farrelly	11	0.162722
Chris Columbus	11	0.162722
Danny Boyle	11	0.162722
David Gordon Green	11	0.162722
Edward Zwick	11	0.162722
Jon Turteltaub	11	0.162722
M. Night Shyamalan	11	0.162722
Marc Forster	11	0.162722
Michael Mann	11	0.162722
Peter Weir	11	0.162722
Phillip Noyce	11	0.162722
Roger Donaldson	11	0.162722
Roman Polanski	11	0.162722
Shawn Levy	11	0.162722
Sidney Lumet	11	0.162722
Terry Gilliam	11	0.162722

Fonte: Banco de dados original

Escritor (escritor)

A variável escritor é categórica e indica os profissionais que redigiram a história de cada um dos filmes, esta variável foi classificada no R como Factor. Na lista de diretores por filme apresentada não havia respostas não disponíveis [NA].

Tabela 7 – frequência dos trinta escritores presentes em observações mais encontradas de cada trabalho responsável nos filmes.

Escritor	Frequencia	Porcentagem.Freq
Woody Allen	29	0.428994
Joel Coen, Ethan Coen	13	0.192308
John Hughes	11	0.162722
Kevin Smith	10	0.147929
M. Night Shyamalan	10	0.147929
Jon Lucas, Scott Moore	9	0.133136
Christopher Markus, Stephen McFeely	8	0.118343
Luc Besson, Robert Mark Kamen	8	0.118343
Patrick Melton, Marcus Dunstan	8	0.118343
Tyler Perry	8	0.118343
Cinco Paul, Ken Daurio	7	0.10355

Lilly Wachowski, Lana Wachowski	7	0.10355
Mike Leigh	7	0.10355
Sean Anders, John Morris	7	0.10355
Steven Knight	7	0.10355
Ehren Kruger	6	0.088757
Fran Walsh, Philippa Boyens	6	0.088757
Jason Friedberg, Aaron Seltzer	6	0.088757
John Sayles	6	0.088757
John Waters	6	0.088757
Lars von Trier	6	0.088757
Lowell Ganz, Babaloo Mandel	6	0.088757
Luc Besson	6	0.088757
Paul Thomas Anderson	6	0.088757
Paul W.S. Anderson	6	0.088757
Phil Hay, Matt Manfredi	6	0.088757
Robert Rodriguez	6	0.088757
Scott Neustadter, Michael H. Weber	6	0.088757
Tyler Perry, Tyler Perry	6	0.088757
Abby Kohn, Marc Silverstein	5	0.073964

Fonte: Banco de dados original

Atores principais e coadjuvantes

Fora divididas quatro variáveis indicando os Atores principais e coadjuvantes relacionados as filmagens de cada um dos filmes, esta variável foi classificada no R como Factor. Na lista de diretores por filme apresentada não havia respostas não disponíveis [NA].

Tabela 8 – frequência dos trinta atores principais mais encontrados em observações de cada filme descrito.

Ator_principal	Frequencia	Porcentagem
Nicolas Cage	45	0.665680473
Robert De Niro	36	0.532544379
Bruce Willis	35	0.517751479
Clint Eastwood	33	0.48816568
Johnny Depp	32	0.473372781
Tom Hanks	32	0.473372781
Tom Cruise	31	0.458579882
Adam Sandler	30	0.443786982
Denzel Washington	28	0.414201183
John Travolta	28	0.414201183
Sylvester Stallone	28	0.414201183
Arnold Schwarzenegger	26	0.384615385
Al Pacino	25	0.369822485
Mel Gibson	25	0.369822485

Robin Williams	25	0.369822485
Mark Wahlberg	24	0.355029586
Eddie Murphy	23	0.340236686
Harrison Ford	23	0.340236686
Jim Carrey	22	0.325443787
Kevin Costner	22	0.325443787
Liam Neeson	22	0.325443787
Matthew McConaughey	22	0.325443787
Brad Pitt	21	0.310650888
John Cusack	21	0.310650888
Matt Damon	21	0.310650888
Sandra Bullock	21	0.310650888
Will Ferrell	21	0.310650888
Dwayne Johnson	20	0.295857988
George Clooney	20	0.295857988
Keanu Reeves	20	0.295857988

Fonte: Banco de dados original

Tabela 9 – frequência dos trinta atores coadjuvantes mais encontrados em observações de cada filme descrito.

Ator_coadjuvante	Frequencia	Porcentagem
Gene Hackman	14	0.207192541
Morgan Freeman	14	0.207192541
Robert Downey Jr.	14	0.207192541
Annette Bening	13	0.192393074
Danny DeVito	13	0.192393074
Diane Keaton	13	0.192393074
Samuel L. Jackson	13	0.192393074
Julianne Moore	12	0.177593607
Nicole Kidman	12	0.177593607
Robert De Niro	12	0.177593607
Ashley Judd	11	0.162794139
Cate Blanchett	11	0.162794139
Michelle Pfeiffer	11	0.162794139
Tommy Lee Jones	11	0.162794139
Aaron Eckhart	10	0.147994672
Angelina Jolie	10	0.147994672
Bruce Willis	10	0.147994672
Charlize Theron	10	0.147994672
Eddie Murphy	10	0.147994672
Kate Winslet	10	0.147994672
Owen Wilson	10	0.147994672
Susan Sarandon	10	0.147994672
Alec Baldwin	9	0.133195205
Brad Pitt	9	0.133195205
Channing Tatum	9	0.133195205

Christopher Lloyd	9	0.133195205
Christopher Plummer	9	0.133195205
Ed Harris	9	0.133195205
Ewan McGregor	9	0.133195205
Gary Oldman	9	0.133195205

Fonte: Banco de dados original

Variáveis quantitativas

Duração (duração)

A variável duração esta em minutos e foi classificada no R como numérica. Na lista de minutos por filme apresentada não havia respostas não disponíveis [NA].

Tabela 10 – Analise descritiva de 6760 observações da variável duração em minutos

Variavel Duração em minutos	Valores
Min.	63
1st Qu.	95
Median	104
Mean	107.6419
3rd Qu.	117
Max.	321

Fonte: Banco de dados original

Nota média IMDB (Nota_média_IMDB)

A variável nota média IMDB mostra a nota média de 0 a 10 dadas pelo publico em geral ao filme, quanto maior o número de votos na variável número de votos IMDB maior a estabilidade desta média e a garantia que ela representa a opinião de diferentes espectadores de diferentes culturas e regiões geográficas.

Tabela 11 – Analise descritiva de 6760 observações da nota média IMDB dos filmes da base de dados original.

	Nota_média_Imdb
Minimo	1.4
1° quartil	5.8
Média	6.4
3° quartil	7.1
Maximo	9.3

Fonte: Banco de dados original

Número votos IMDB (Número_votos_IMDB)

A variável número votos IMDB mostra o número de votos IMDB recebidos por diferentes espectadores em cada filme da base. Quanto maior o número de votos recebidos, maior a estabilidade da nota média de um filme e maior a garantia que ela representa a opinião de diferentes espectadores de diferentes culturas e regiões geográficas.

Tabela 12 – Analise descritiva de 6760 observações da nota média IMDB dos filmes da base de dados original

	Número_votos_IMDB
Minimo	100
1° quartil	10,166
Média	90,120
3° quartil	98,731
Maximo	2,159,628

Fonte: Banco de dados original

Nota média Metacritcs (Nota_média_metascore)

A variável nota média Metacritics mostra a nota média de 0 a 100 dadas por críticos de cinema ao filme, quanto maior o número de votos na variável número avaliações críticos de cinema maior a estabilidade desta média e a garantia que ela representa a opinião de diferentes profissionais do ramo cinematografico de diferentes culturas e regiões geográficas.

Tabela 13 – Analise descritiva de 6760 observações da nota média Metacritics dos filmes da base de dados original.

	Nota_média_metascore
Minimo	1
1° quartil	41
Média	55
3° quartil	68
Maximo	100

Fonte: Banco de dados original

Número avaliações críticos de cinema (Número_avaliações_criticoscinema)

A variável número avaliações críticos de cinema mostra o número de votos no sistema Metacritics, especializado em identificar avaliações de filmes feitas por profissionais do ramo, avaliações preparadas por diferentes criticos para cada filme da base. Quanto maior o número de avaliações recebidas, maior a estabilidade da nota média de um filme e maior a garantia que ela representa a opinião de diferentes profissionais do ramo cinematografico de diferentes culturas e regiões geográficas.

Tabela 14 – Analise descritiva de 6760 observações do número de avaliações de críticos de cinema para os filmes da base de dados original.

	Número_avaliações_criticoscinema
Minimo	1
1° quartil	55
Média	144
3° quartil	192
Maximo	987

Fonte: Banco de dados original

Número avaliações Metacritics (Número_avaliações_Metacritics)

A variável número de avaliações Metacritics mostra o número de avaliações indicadas no sistema por diferentes espectadores em cada filme da base. Por não ser a especialidade do site Metacritics, é possível perceber que quanto maior a popularidade do filme em questão, maior o número de avaliações. Estas avaliações não contam para montar a média Metascore pois ela utiliza somente opiniões de profissionais do ramo.

Tabela 15 – Analise descritiva de 6760 observações da nota média IMDB dos filmes da base de dados original

	Número_avaliações_Metacritics
Minimo	1
1° quartil	69
Média	288
3° quartil	327
Maximo	8302

Fonte: Banco de dados original

Orçamento USD USA (Orçamento_USD_USA)

A variável orçamento USD USA indica qual foi o montante utilizado para a a produção de divulgação de cada filme, está classificada como variável numérica no R e auxiliou a criar a variável ROI percentual.

Tabela 15 – Analise descritiva de 6760 observações do valor do orçamento de cada filme indicado no banco de dados.

	Orçamento_USD_USA
Minimo	$2.00
1° quartil	$5,000,000.00
Média	$29,414,267.18
3° quartil	$35,000,000.00
Maximo	$356,000,000.00

Fonte: Banco de dados original

Renda bruta mundial USD (Renda_bruta_mundial_USD)

A variável renda bruta mundial USD indica qual foi o montante arrecadado em bilheteria após a comercialização de cada filme, a mesma está classificada como variável numérica no R e auxiliou a criar a variável ROI percentual.

Tabela 16 – Analise descritiva de 6760 observações do valor arrecadado em bilheteria de cada filme indicado no banco de dados.

	Renda_bruta_mundial_USD
Minimo	$77.00
1° quartil	$3,231,421.50
Média	$84,368,142.36
3° quartil	$86,298,548.50
Maximo	$2,797,800,564.00

Fonte: Banco de dados original

ROI percentual (ROI_percentual)

A variável ROI percentual indica a porcentagem de lucro ou prejuízo após a comercialização de cada filme, a mesma está classificada como variável numérica no R e foi criada a partir das variáveis Renda Bruta mundial e Orçamento USD USA pois pode-se estimar aproximadamente o ROI (Return on investment) de um projeto a partir das despesas e receitas totais do mesmo.

Tabela 17 – Analise descritiva de 6760 observações do ROI percentual calculado de cada filme indicado no banco de dados.

	ROI_percentual
Minimo	-100
1° quartil	-56
Média	60.158
3° quartil	249
Maximo	385.711.840

Fonte: Banco de dados original

Resultados e Discussão

Correlação Variáveis Numéricas

Com a função Cor que faz parte da base do R studio é possível entender qual a correlação entre as variáveis numéricas e como seus comportamentos estão ou não interligados. O calculo apresenta números de -1 a 1 onde 1 são correlações diretamente proporcionais e -1 é apresentado para correlações inversamente proporcionais. Os cálculos que apresentam números abaixo de 0,5, sejam positivos ou negativos, indicam que não há uma correlação significativa entre as variáveis, já cálculos que apresentam números entre 0,5 e 0,7, sejam positivos ou negativos, indicam uma correlação moderada para estas variáveis.
Ainda que uma correlação forte seja algo significativo entre variáveis, não podemos atribuir o peso de causa e efeito as variáveis correlacionadas pois muito frequentemente o comportamento de uma variável é ditado pelo comportamento de um conjunto de outras variáveis que influenciam o resultado final, assim, precisamos inclui-las ao cálculo de um modelo de regressão para entendermos qual a magnitude do impacto de cada variável independente no resultado final de uma variável de interesse.

Após o calculo foi obtido as seguintes matrizes de correlação, onde as correlações fortes e moderadas foram destacadas em vermelho.

Tabela 18 – Matrix de correlação 1

	Nota_media_Imdb	Numero_votos_IMDB
ROI_percentual	0.007566378	0.020786152
Duração	0.393848365	0.322905304
Nota_media_Imdb	1	0.429060951
Numero_votos_IMDB	0.429060951	1
Orçamento_Usd_USA	0.03638646	0.445612336
Renda_bruta_mundial	0.187968565	0.60745703
Nota_media_metascore	0.737522875	0.294331177
Numero_avaliações_Metacritics	0.285854836	0.740869616
Numero_avaliações_criticoscinema	0.321291417	0.607526423

Tabela 19 – Matrix de correlação 2

	Duração	Numero_avaliações_Metacritics
ROI_percentual	-0.020449276	0.05914556
Duração	1	0.318545567
Nota_media_Imdb	0.393848365	0.285854836
Numero_votos_IMDB	0.322905304	0.740869616
Orçamento_Usd_USA	0.276429593	0.493707867
Renda_bruta_mundial	0.25113298	0.621451303
Nota_media_metascore	0.277902281	0.200792944
Numero_avaliações_Metacritics	0.318545567	1
Numero_avaliações_criticoscinema	0.25564011	0.608298185

Tabela 20 – Matrix de correlação 3

	Nota_media_ metascore	Numero_avaliações_ criticoscinema
ROI_percentual	0.024015817	0.028568535
Duração	0.277902281	0.25564011
Nota_media_Imdb	0.737522875	0.321291417
Numero_votos_IMDB	0.294331177	0.607526423
Orçamento_Usd_USA	-0.039414471	0.505259314

Renda_bruta_mundial	0.121124117	0.53785828
Nota_media_metascore	1	0.307194319
Numero_avaliações_Metacritics	0.200792944	0.608298185
Numero_avaliações_criticoscinema	0.307194319	1

Tabela 21 – Matrix de correlação 4

	Orçamento_Usd_USA	Renda_bruta_mundial
ROI_percentual	-0.016100116	0.018882105
Duração	0.276429593	0.25113298
Nota_media_Imdb	0.03638646	0.187968565
Numero_votos_IMDB	0.445612336	0.60745703
Orçamento_Usd_USA	1	0.748274837
Renda_bruta_mundial	0.748274837	1
Nota_media_metascore	-0.039414471	0.121124117
Numero_avaliações_Metacritics	0.493707867	0.621451303
Numero_avaliações_criticoscinema	0.505259314	0.53785828

Tabela 22 – Matrix de correlação 5

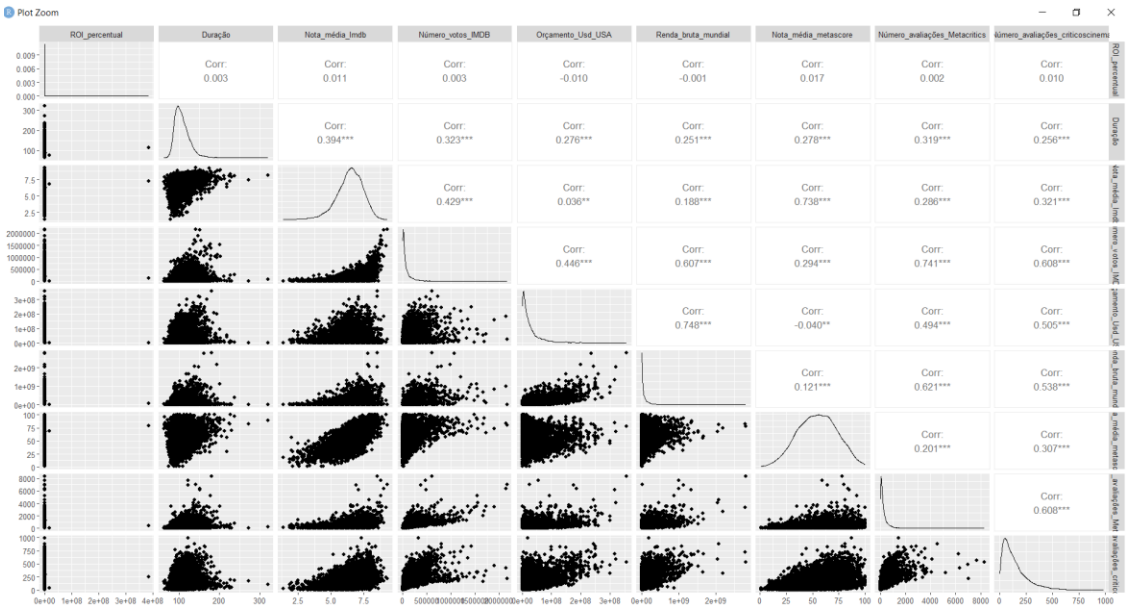
	ROI_percentual
ROI_percentual	1
Duração	-0.020449276
Nota_media_Imdb	0.007566378
Numero_votos_IMDB	0.020786152
Orçamento_Usd_USA	-0.016100116
Renda_bruta_mundial	0.018882105
Nota_media_metascore	0.024015817
Numero_avaliações_Metacritics	0.05914556
Numero_avaliações_criticoscinema	0.028568535

Cálculos de significância estatística

Ao utilizar o pacote Ggally do R studio indicando para cálculo as variáveis numéricas do banco de dados, foi obtida a figura 1, que indicou o grau de significância estatística de cada correlação entre as variáveis numéricas, bem com os gráficos de dispersão entre elas e o comportamento linear proporcional entre algumas delas.
Os asteriscos indicados em cada correlação mostram se esta correlação tem baixa probabilidade de ser obtida por acaso, onde um “*” significa até 5% de chance de ser

uma correlação ao acaso, dois “**” significam até 1% e três “***” significam menos de 1% de acaso na correlação.

Figura 1 - Gráficos de dispersão, de frequência e significância estatística das variáveis numéricas:



Modelagem de uma regressão linear simples

Após avaliar o comportamento das variáveis de interesse ROI_percentual, Nota_média_imdb e Nota_média_metascore, percebeu-se que apenas o comportamento das duas ultimas variáveis tem correlação estatisca entre si e podem ser modelados afim de estimar os resultados de uma com dados do outra, assim foi modelada uma estimativa para a Nota_média_imdb, que apresenta a recepção do publico em geral, a partir da variável Nota_média_metascore, que apresenta a recepção média dos profissionais da área, pois a variável independente é composta com em média 144 opiniões de profissionais do ramo distribuídos em diferentes culturas de regiões geográficas, assim uma média mais fácil de se obter antes de disponibilizar o filme para o publico geral e gerar receita a partir das bilheterias.
O modelo calculado retornou os seguintes detalhes:

Call:
lm(formula = Nota_média_Imdb ~ Nota_média_metascore, data = Dadosfiltrados)
(um modelo de regressão linear simples com "Nota_média_Imdb" como a variável de resposta e "Nota_média_metascore" como a variável independente.)

Residuals:

```
    Min      1Q  Median      3Q     Max
-4.5181 -0.3801  0.0210  0.4243  2.5472
```

(Esta parte fornece estatísticas resumidas sobre os resíduos do modelo)

Coefficients:

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.225449   0.025819  163.66   <2e-16 ***
Nota_média_metascore 0.040222   0.000448   89.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Aqui, você encontra informações sobre os coeficientes do modelo, que representam a relação entre a variável independente e a variável de resposta.

Estimate: Este é o valor estimado do coeficiente. Neste caso, o coeficiente estimado para "Nota_média_metascore" é 0.040222.
Std. Error: O erro padrão do coeficiente estimado.
t value: O valor t é a estatística t, que mede o quão distante o coeficiente estimado está de zero em termos de erros padrão.
Pr(>|t|): Este é o valor-p associado ao teste t. Indica a probabilidade de observar um valor t tão extremo quanto o observado, assumindo que não há relação entre as variáveis. Quanto menor o valor-p, mais significativo é o coeficiente.
Signif. codes: São códigos que indicam o nível de significância estatística. No seu caso, "***" significa altamente significativo (p-value < 0.001).)

Residual standard error: 0.6717 on 6758 degrees of freedom
Multiple R-squared: 0.544,
Adjusted R-squared: 0.5439
F-statistic: 8062 on 1 and 6758 DF,
p-value: < 2.2e-16

Neste caso, o modelo parece ser altamente significativo, com valores de p muito baixos (menos de 2.2e-16), indicando que a variável "Nota_média_metascore" é altamente significativa na previsão da "Nota_média_Imdb". Além disso, o R^2 múltiplo de 0.544 sugere que cerca de 54.4% da variabilidade na "Nota_média_Imdb" pode ser explicada pela "Nota_média_metascore".

Ao calcular o erro percentual em relação à variável de resposta real foi obtido o valor de erro percentual médio de 7.96%, este valor foi calculado a partir do erro do Erro Médio Absoluto (MAE - Mean Absolute Error), onde ele foi dividido pela média da variável resposta e multiplicado por 100. Assim, pode-se indicar que o modelo erra a predição em apenas 8% de suas tentativas, podendo ser utilizado como uma prévia da recepção do público em geral do filme em estudo.

Resultados e Discussão

Em análise.

Conclusão(ões) ou Considerações Finais

Pendente.

Agradecimentos

Gostaria de agradecer a Cleverson Silva, meu irmão, graduando em Ciências da computação, que me inspirou a ser persistente no aprendizado e na utilização das ferramentas e das técnicas de machine learning. Sem o exemplo dele, não teria me mantido engajado para completar este passo de minha pós-graduação.

Referências

Introduction to Econometrics with R - Christoph Hanck, Martin Arnold, Alexander Gerber, and Martin Schmelzer

Data Science for Business and Decision Making - Por Favero, Luiz Paulo - Belfiore, Patricia - 2019

Handbook of Regression Modeling in People Analytics - With Examples in R and Python - Por McNulty, Keith· 2021

Fávero, L. P.; Belfiore, P. 2017. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier, Rio de Janeiro, RJ, Brasil.