ANÁLISE EXPLORATÓRIA NO CONJUNTO DE DADOS *IMDB 5000 MOVIE DATASET*

Anderson A. dos Santos, Ederson R. da Costa, Higor H. P. Nucci

[1]Faculdade de Computação – Universidade Federal de Mato Grosso do Sul (UFMS)
Av. Costa e Silva, s/n – 79.070-900 – Campo Grande – MS – Brazil

{anderson.asantos3, higornucci}@gmail.com, edersondacosta@hotmail.com

Abstract. *This article displays the exploited results of the IMDB 5000 Movie Dataset data combinations. The visualization techniques employed from a correlation matrix enabled a better analysis and comparison of several data. A scatter plot with the most closely related values was generated. Another analysis made it possible to detail the missing values as well as techniques for filling in these values, graphs were generated that exhibit such absence of values in a dataset, which facilitates in the decision making, on which filling technique should be performed.*

Resumo. *Este artigo exibe os resultados explorados das combinações de dados do IMDB 5000 Movie Dataset. As técnicas de visualizações empregadas a partir de uma matriz de correlação possibilitaram uma melhor análise e comparação de diversos dados. Um gráfico de dispersão com os valores que mais se relacionavam foi gerado. Outra análise possibilitou detalhar os valores ausentes, bem como técnicas de preenchimento de tais valores, foram gerados gráficos que exibem tal ausência de valores em um dataset, o que facilita na tomada de decisão, sobre qual técnica de preenchimento deve ser efetuada.*

1. Introdução

Uma análise exploratória, basicamente, consiste em detectar padrões, tendências e relações em dados, o que também está associado com a visualização destes dados [Andrienko and Andrienko 2006]. Assim sendo, o objetivo deste artigo é responder a alguns questionamentos da Atividade II, da disciplina de Mineração de Dados, do Programa de Pós-Graduação da Faculdade de Computação (Facom) da Universidade Federal de Mato Grosso do Sul (UFMS).

O conjunto de dados (*dataset*) explorado é o *Internet Movie Database* IMDB 5000 [Zhang 2017]. O IMDB é uma base de dados online de informação sobre música, cinema, filmes, programas e comerciais para televisão e jogos de computador, hoje pertencente à empresa Amazon. O *dataset* explorado contém 28 atributos, os quais representam informações como, por exemplo, nome do filme, nome do diretor, ator principal, duração do filme, entre outras.

A análise dos dados foi realizada utilizando a linguagem de programação Python[1], a biblioteca científica Numpy[2], a biblioteca de *Machine Learning* scikit-learn[3], a bib-

[1]Disponível em: `https://www.python.org/`.
[2]Disponível em: `https://www.numpy.org/`.
[3]Disponível em: `https://scikit-learn.org/stable/`.

lioteca de análise de dados Pandas[4] e a biblioteca de plotagem Matplotlib[5]. Com base nas informações descritas, cada uma das questões da atividade é abordada como uma seção neste artigo. Desta maneira, a Seção 2 descreve as informações sobre os dados, a Seção 3 mostra a correlação entre dados e, por fim, a Seção 4 apresenta as abordagens utilizadas para preencher dados ausentes do *dataset*.

2. Entendimento da Base

2.1. A fim de entender a base de dados, descreva o que cada um dos atributos representa.

O *dataset* IMDB 5000 Movie, possui diversas informações sobre filmes. Uma visão geral do conjunto de dados é apresentada na Tabela 1, em que é possível observar informações como o nome, o tipo e a descrição de cada atributo.

Table 1. Informações Gerais do Conjunto de Dados

Atributo	Tipo	Descrição
color	Nominal	Colorido ou preto e branco
director_name	Nominal	Nome do diretor
num_critic_for_reviews	Racional	Quantidade de cometários críticos
duration	Racional	Duração do filme em minutos
director_facebook_likes	Racional	Quantidade de curtidas na página do Facebook do diretor
actor_3_facebook_likes	Racional	Quantidade de curtidas na página do Facebook do ator 3
actor_2_name	Nominal	Nome do ator 2
actor_1_facebook_likes	Racional	Quantidade de curtidas na página do Facebook do ator 1
gross	Racional	Lucro bruto do filme em dólares
genres	Nominal	Categoria do filme (ex. animação)
actor_1_name	Nominal	Nome do ator 1
movie_title	Nominal	Título do filme
num_voted_users	Racional	Quantidade de votos dos usuários
cast_total_facebook_likes	Racional	Quantidade de curtidas no Facebook de todo o elenco do filme
actor_3_name	Nominal	Nome do ator 3
facenumber_in_poster	Racional	Quantidade de rostos no pôster do filme
plot_keywords	Nominal	Palavras-chave do enredo do filme
movie_imdb_link	Nominal	Link IMDB do filme
num_user_for_reviews	Racional	Quantidade de usuários que fizeram uma revisão
language	Nominal	Idioma do filme
country	Nominal	País em que o filme foi produzido
content_rating	Nominal	Classificação indicativa do filme
budget	Racional	Orçamento do filme em dólares
title_year	Racional	Ano de lançamento do filme
actor_2_facebook_likes	Racional	Quantidade de curtidas na página do Facebook do ator 2
imdb_score	Racional	Pontuação do filme no IMDB
aspect_ratio	Racional	Proporção da tela em que o filme foi gravado
movie_facebook_likes	Racional	Quantidade de curtidas na página do Facebook do filme

Compreender os atributos e o tipo de dados de um *dataset* possibilita uma melhor análise dos dados. Uma outra maneira de observar os dados é com análises estatísticas e gráficos. Na Figura 1, os dados dos atributos quantitativos do IMDB 5000 são representados por gráficos de violino e boxplot.

[4]Disponível em: `https://pandas.pydata.org/`.
[5]Disponível em: `https://matplotlib.org/`.

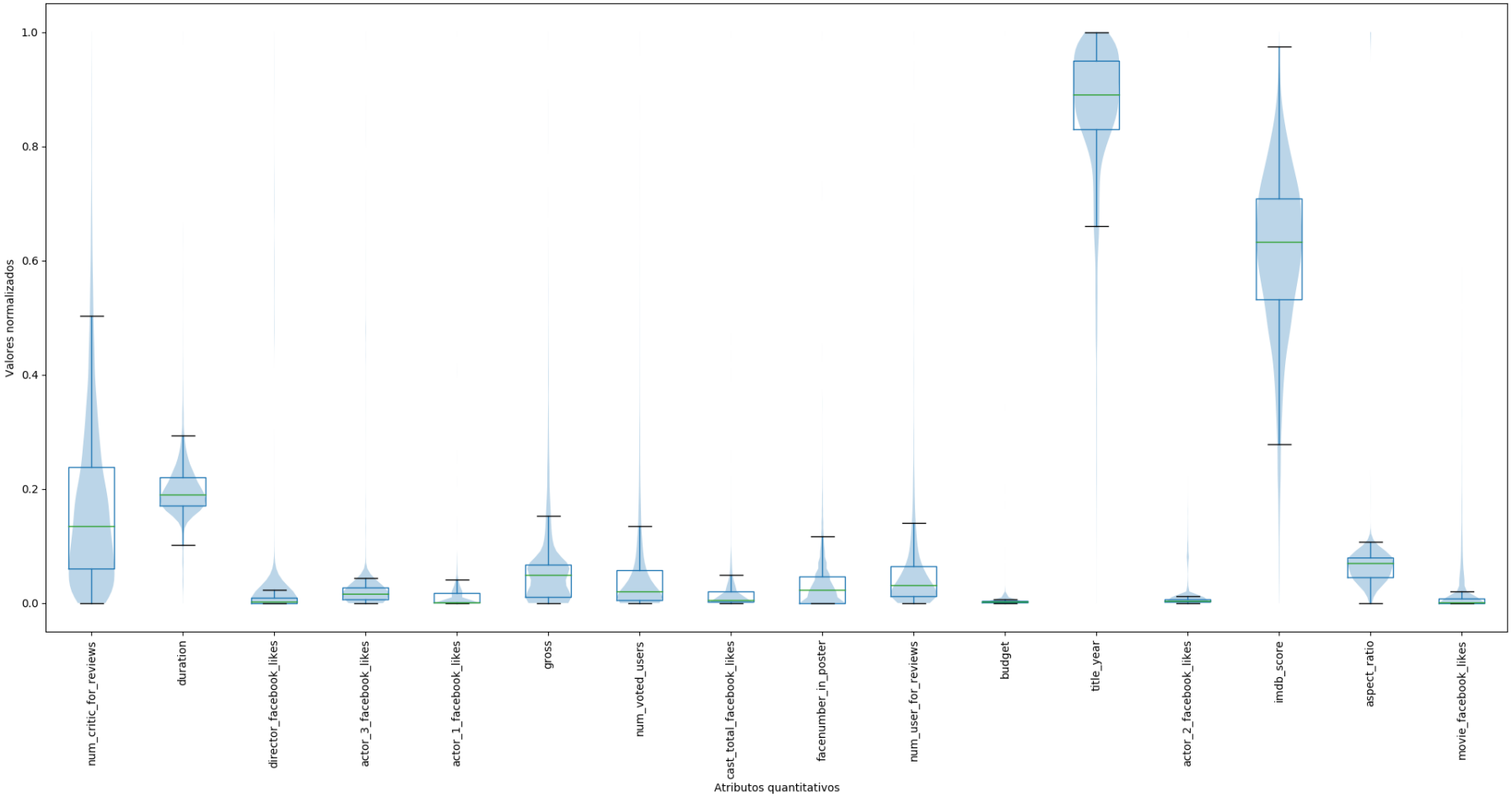

Figure 1. Atributos quantitativos representados em gráficos de violino e boxplot. Para esta representação, os dados ausentes foram preenchidos com a média do atributo correspondente, normalizados em uma escala entre 1 e 0 e os *outliers* foram ocultados.

Observando o gráfico 1, verifica-se que os valores da maioria dos atributos se encontram na parte inferior do gráfico, ou seja, próximos do valor mínimo de cada atributo. Atributos como title_year e imdb_score indicam que, respectivamente, a maioria dos filmes do conjunto de dados são mais recentes e o *score* geral dos filmes está acima de 50%.

3. Correlação entre os Atributos

3.1. O que significa a correlação entre dois atributos? O que é uma matriz de correlação? Quais atributos da base IMDB Movie Dataset fariam sentido em uma matriz de correlação?

O estudo de correlação tem por finalidade verificar a força de associação entre os atributos. Quando se identifica de forma clara a relação entre os atributos, essa correlação é conhecida como supostamente lógica. Já quando não se identifica nenhuma correlação razoável entre os atributos essa correlação é conhecida como ilusória. Por exemplo: imagine que a densidade da população de um determinado inseto possa estar correlacionado com o porte de alguma espécie de erva ou, possa simplesmente estar correlacionada com o tempo. A densidade dos insetos pode também não estar correlacionado com questões ecológicas, tal população pode depender de outras variáveis [Callegari-Jacques 2009]. Algumas correlações podem ser:

- Comprimento do braço e da perna de determinados mamíferos;
- Quantidade de colesterol no sangue e peso de pessoas de mesmo sexo e idade;
- Altura dos pais e dos filhos de pessoas de mesma etnia;
- Rendimento e gastos por faixa salarial;
- Valor e demanda;
- Produtividade agrícola e adubo.

A matriz de correlação é empregada para análise estatística de dados, fornecendo de forma visual os atributos que estão relacionados entre si. Cada linha e coluna de uma matriz de correlação representa um atributo. Assim sendo, a matriz é quadrada e sua diagonal principal representa a correlação máxima de um atributo com ele mesmo. O grau de relação pode ser determinado pelo coeficiente de Pearson, também conhecido como coeficiente de correlação. Este coeficiente possui um valor entre -1 e 1 que representa a intensidade de dependência linear entre dois atributos quantitativos. Neste caso, quando o coeficiente é negativo, indica que o valor de um atributo diminui com o aumento do valor de outro atributo. Por outro lado, quando o coeficiente é positivo, os valores dos atributos aumentam uns com os outros [Johnson et al. 2002].

Atributos da base IMDB Movie que fariam sentido em uma matriz de correlação são os atributos quantitativos, por exemplo, o orçamento do filme (budget) com o lucro bruto obtido (gross), o ano de lançamento (title_year) com a quantidade de comentários (num_critic_for_reviews) ou curtidas (movie_facebook_likes) ou quais atributos estão mais correlacionados com a pontuação do filme no IMDB (imdb_score).

3.2. Gere uma matriz de correlação entre os atributos. Quais atributos estão mais correlacionados? Gere um gráfico que mostre uma relação entre esses atributos.

Para uma melhor análise dos atributos foi gerada uma matriz de correlação, utilizando os valores de correlação de Pearson, que efetuam a medição do grau de relação linear entre

pares de atributos. A correlação entre os atributos é exibida na Figura 2. Para gerar esta matriz, primeiramente os dados ausentes foram preenchidos. Quando numéricos, os valores ausentes foram preenchidos com a média do respectivo atributo. Quando nominais, foi criado um novo atributo para representar a ausência. Posteriormente, os atributos nominais foram codificados usando LabelEncoder()[6] e então foi gerada a matriz de correlação com o método corr()[7].

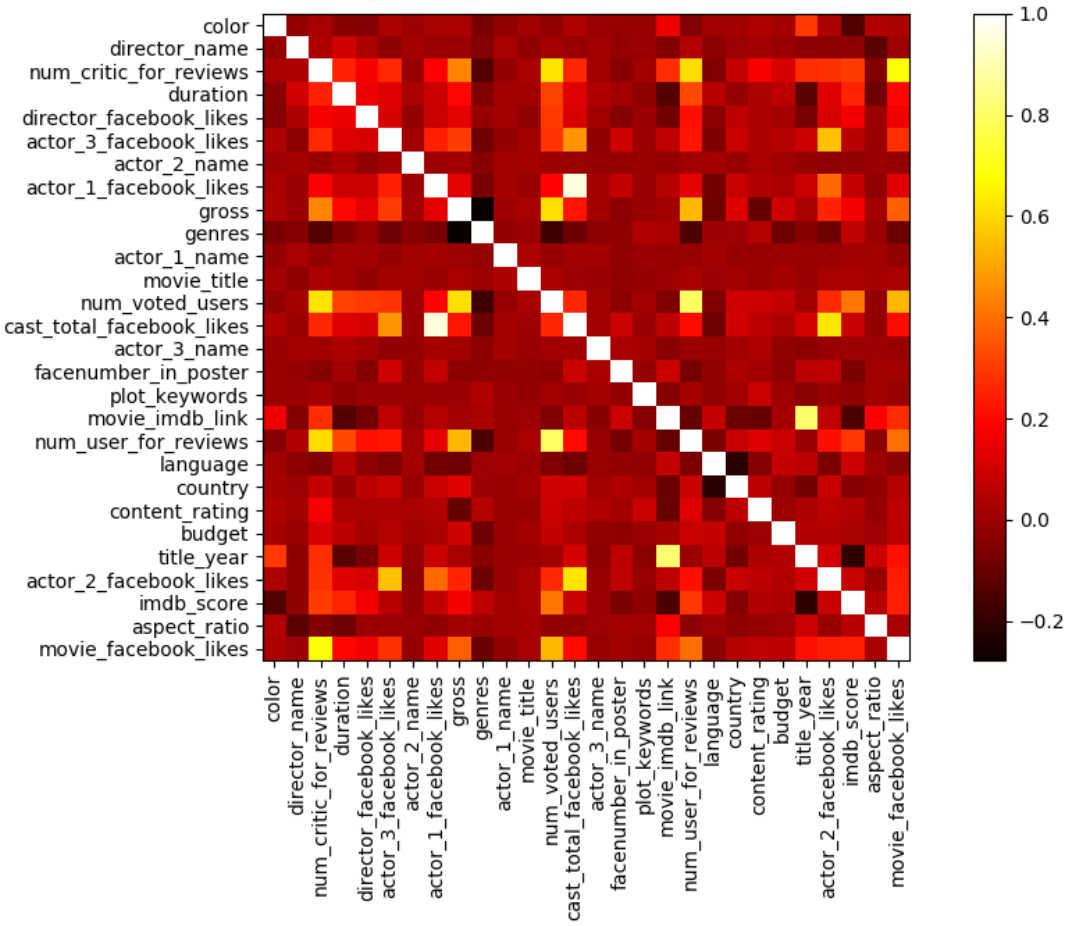

Figure 2. Matriz de correlação entre todos os atributos do *dataset*.

É possível perceber na Figura 2, que nem todos os atributos possuem correlação significativa, na maioria dos casos, os atributos qualitativos possuem uma correlação baixa. Por exemplo, os atributos color e director_name, possuem uma correlação baixa com a maioria dos outros atributos. Outros atributos que apresentam uma baixa correlação são atributos de nomes (actor_1_name, actor_2_name, actor_3_name), movie_title, plot_keywords, movie_imdb_link, language, country, content_rating, budget e aspect_ratio.

Com o objetivo de explorar os atributos que mais se relacionam, foi gerado uma matriz de dispersão com os atributos que possuem uma coloração mais amarelada na

[6]Disponível em: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html`.

[7]Disponível em: `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html`.

Figura 2. A Figura 3 exibe a matriz de dispersão. Para gerar esta matriz, os dados foram normalizados com MinMaxScaler()[8] e então a matriz foi gerada com o método scatter_matrix()[9].

[8]Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[9]Disponível em: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.plotting.scatter_matrix.html.

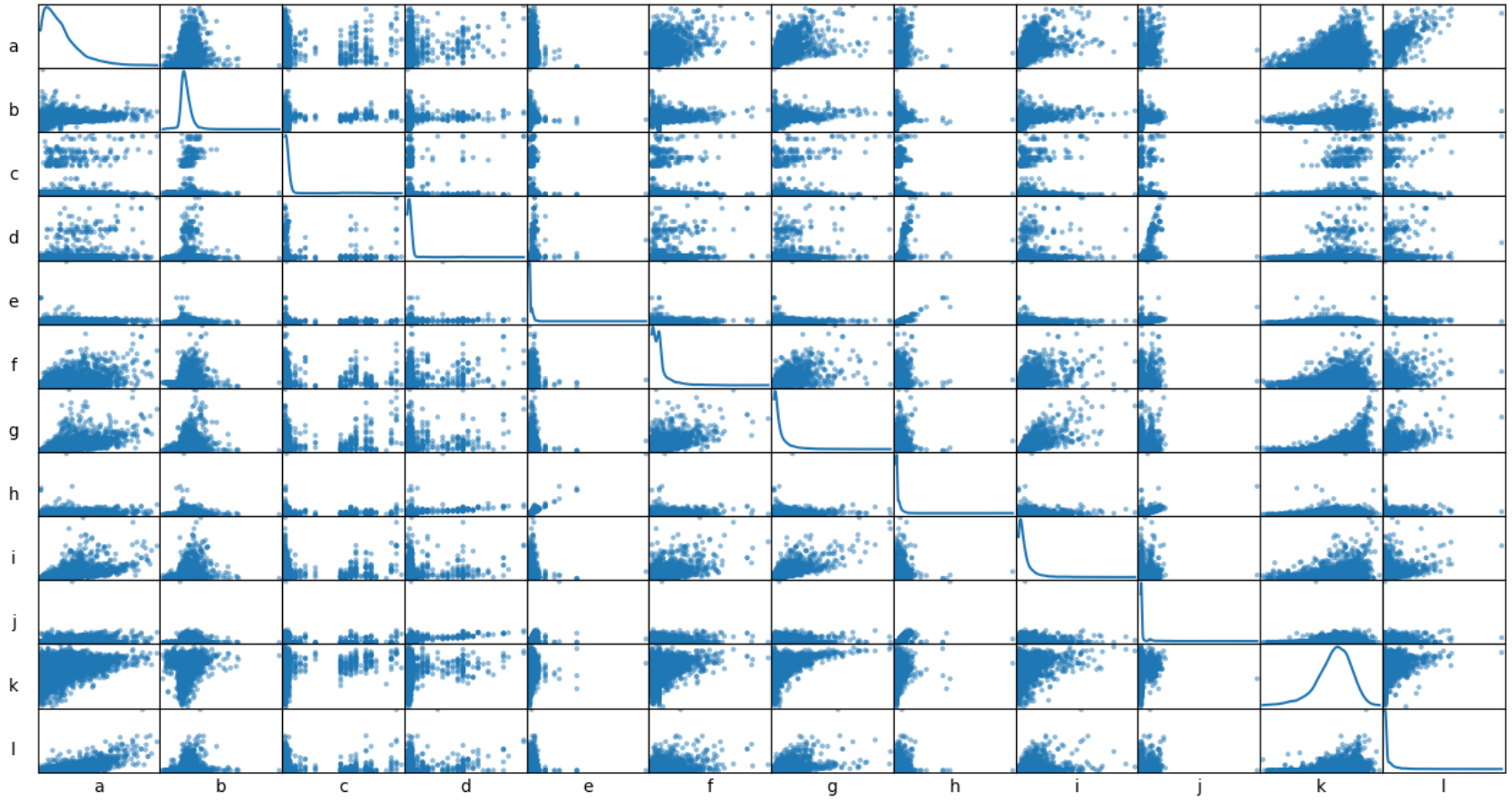

Figure 3. Matriz de dispersão entre os atributos mais correlacionados: (a) num_critic_for_reviews; (b) duration; (c) director_facebook_likes; (d) actor_3_facebook_likes; (e) actor_1_facebook_likes; (f) gross; (g) num_voted_users; (h) cast_total_facebook_likes; (i) num_user_for_reviews; (j) actor_2_facebook_likes; (k) imdb_score; (l) movie_facebook_likes.

A Figura 3 exibe a dispersão de pares de atributos. A correlação pode ser avaliada de quatro formas: correlação positiva forte, correlação positiva média, correlação negativa forte, correlação negativa média [Pereira 2017]:

- Correlação positiva forte: há uma tendência clara nos valores. Quando a variável X aumenta, é provável que conjuntamente exista um aumento na variável Y. A pouca dispersão dos valores sugere que essa tendência é forte [Pereira 2017]. Essa correlação forte é exibida entre os itens "a" e "l" da Figura 3.
- Correlação positiva média: quando a variável X aumenta, a variável Y tende a aumentar também. Porém, a dispersão maior dos valores indica que outras variáveis podem estar envolvidas [Pereira 2017]. Tal situação é vista na Figura 3 entre os atributos "a" e "k".
- Correlação negativa forte: semelhante à correlação positiva forte, porém quando X aumenta, Y tende a diminuir [Pereira 2017]. Essa cenário ocorreu entre os valores de "a" e "b" na Figura 3. .
- Correlação negativa média: idêntica à correlação positiva média, no entanto, quando X aumenta, Y tende a diminuir de forma dispersa [Pereira 2017]. Essa situação ocorreu entre os valores de "f" e "l" da Figura 3.

4. Valores Ausentes

4.1. O que são valores ausentes? Por quais motivos uma base pode possuir valores ausentes? Quais técnicas são comuns para preencher valores ausentes?

A ausência de dados em um conjunto é um problema comum, quando se analisam dados, e geralmente ocorre devido a diversos fatores, tais como: falha de sensores, dados contaminados ou amostras ausentes [Costa 2018].

Diversas técnicas vêm sendo utilizadas no decorrer dos anos para solucionar o problema de valores ausentes, sendo mais comumente utilizadas técnicas de subtração dos valores ausentes e, substituindo por valores estimados com base nos demais dados, tal técnica também é conhecida como imputação. Porém para se identificar qual a melhor técnica para lidar com os dados ausentes, se faz necessária a compreensão das características individuais de cada atributo. Pesquisas recentes propõem a utilização de métodos de visualização dos dados, pois tal método possibilita uma análise aprofundada, direcionando na escolha de qual dado poderá sofrer uma imputação [Costa 2018].

A remoção de amostras que contenham valores nulos, embora seja uma solução rápida e possa funcionar em alguns casos, quando a proporção de valores omissos for relativamente baixa ($< 10\%$), na maioria das vezes, uma grande quantidade de dados será perdida [Hawthorne et al. 2005]. Em muitos casos, simplesmente por causa da falta de valores em um dos atributos, toda a observação precisa ser abandonada, mesmo que o restante dos recursos esteja perfeitamente preenchido e informativo.

Existem formas mais criativas para lidar com valores ausentes, que consistem em dividir o tipo de valores ausentes por seu tipo de dados pai [Hawthorne et al. 2005]. Diferentes abordagens podem ser utilizadas, de acordo com o tipo dos dados de um atributo:

- **Valores ausentes numéricos:** Uma abordagem padrão e geralmente muito boa é substituir os valores ausentes por média, mediana ou moda. Para valores numéricos recomenda-se o uso da média e se houver *outliers* recomenda-se o uso da mediana. Visto que a mediana é menos sensível a *outliers*;

- **Valores ausentes categóricos:** Os valores categóricos podem ser um pouco mais complicados, portanto, deve-se observar as métricas de desempenho do modelo após a edição (comparar antes e depois, por exemplo). O padrão a ser feito é substituir a entrada faltante pela mais frequente.

4.2. Verifique quais atributos da base IMDB Movie possuem valores ausentes. Tente preencher os valores de algum atributo. Reporte o que fez.

A matriz de nulidade é uma exibição de dados densos que permite visualmente identificar padrões no preenchimento de dados [Bilogur 2018]. Ao observar a Figura 4, pode-se notar nas colunas com "riscos" brancos a inexistência de um ou mais dados. Assim sendo, nota-se que os atributos *gross*, *content_rating*, *budget* e *aspect_ratio* possuem a maior quantidade de valores ausentes, enquanto os atributos *genres*, *movie_title*, *num_voted_users*, *cast_total_facebook_likes*, *movie_imdb_link*, *imdb_score* e *movie_facebook_likes* estão com todos os valores preenchidos.

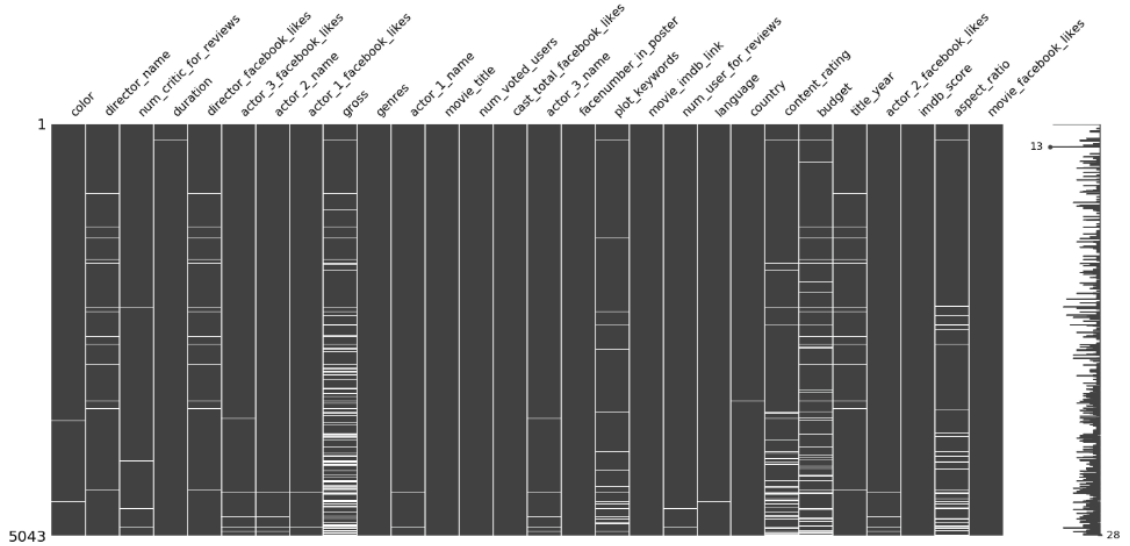

Figure 4. Matriz de nulidade entre todos os atributos do *dataset* [Bilogur 2018]

Na Figura 4, o minigráfico à direita resume a forma geral da integridade dos dados e indica as linhas com a máxima e mínima nulidade no conjunto de dados. A Tabela 2 mostra a quantidade de dados ausentes por atributo. O atributo *gross* é o que tem mais dados ausentes, chegando a 884 de 5043 amostras.

Existe uma grande quantidade de dados ausentes no conjunto de dados. Uma das dúvidas mais comuns é a de quantos dados ausentes devem ser preenchidos. O gráfico que pode ajudar nesse tipo de tarefa é o dendrograma, que pode ser visualizado na Figura 5. O dendrograma usa um algoritmo de *clustering* hierárquico para variáveis entre si por sua correlação de nulidade (medida em termos de distância binária). Em cada etapa da árvore, as variáveis são divididas com base em qual combinação minimiza a distância dos *clusters* restantes. Quanto mais monótono é o conjunto de variáveis, mais próxima a distância total é zero e quanto mais próxima a distância média (o eixo y) é zero.

Table 2. Quantidade de dados ausentes por atributo.

Atributo	Total	Atributo	Total
color	19	actor_3_name	23
director_name	104	facenumber_in_poster	13
num_critic_for_reviews	50	plot_keywords	153
duration	15	movie_imdb_link	0
director_facebook_likes	104	num_user_for_reviews	21
actor_3_facebook_likes	23	language	12
actor_2_name	13	country	5
actor_1_facebook_likes	7	content_rating	303
gross	884	budget	492
genres	0	title_year	108
actor_1_name	7	actor_2_facebook_likes	13
movie_title	0	imdb_score	0
num_voted_users	0	aspect_ratio	329
cast_total_facebook_likes	0	movie_facebook_likes	0

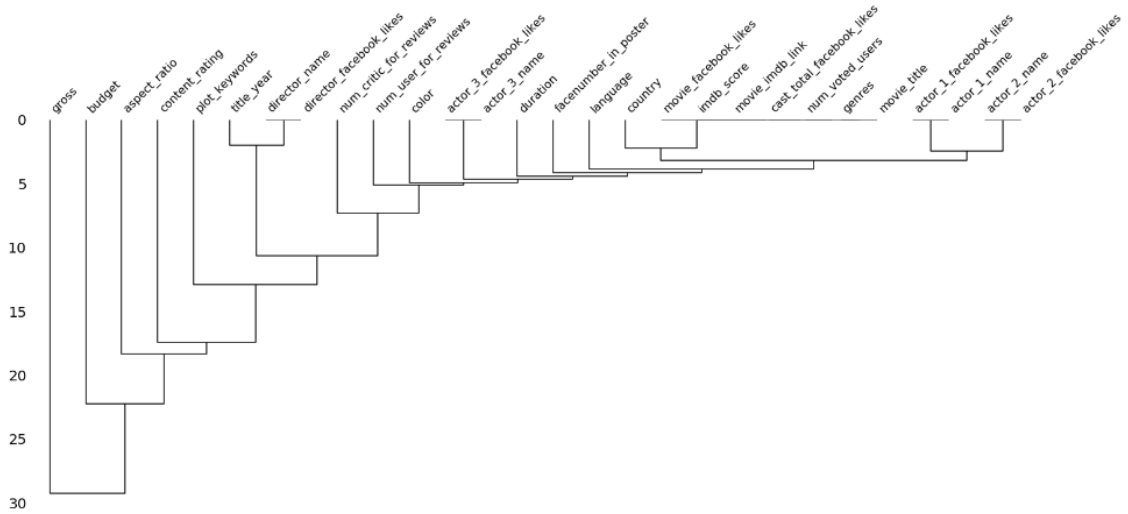

Figure 5. O dendrograma permite correlacionar mais completamente a conclusão das variáveis, revelando tendências mais profundas nas correlações de nulidade.

O gráfico da Figura 5 deve ser lido de cima para baixo. As folhas do *cluster* que estão unidas a uma distância de zero predizem totalmente a presença um do outro, ou seja, uma variável pode estar sempre vazia quando outra é preenchida ou ambas podem estar sempre cheias ou vazias, e assim por diante. Neste exemplo específico, o dendrograma une as variáveis que são necessárias e, portanto, presentes em cada registro.

As folhas que se dividem perto de zero, mas não nele, preveem um ao outro muito bem, mas ainda imperfeitamente. Se a sua própria interpretação do conjunto de dados é que essas colunas são realmente ou devem ser coincidentes em nulidade (por exemplo, como *director_name* e *director_facebook_likes*), a altura da folha do *cluster* informa, em termos absolutos, com que frequência os registros são "incompatíveis" ou incorretamente arquivados, isto é, quantos valores teriam que preencher ou descartar.

Os valores ausentes do conjunto de dados foram preenchidos de acordo com o tipo de dados de cada atributo. Para tipos de dados numéricos, os valores ausentes foram preenchidos com a média do respectivo atributo. Para tipos de dados nominais, os valores ausentes foram preenchidos com a criação de um novo valor, com objetivo de representar a ausência de um dado.

5. Conclusão

Pôde-se perceber que as técnicas de visualização utilizadas neste artigo a partir da matriz de correlação, bem como nos gráficos de valores ausentes e relação entre pares de dados foram fundamentais para auxiliar na análise e compreensão dos dados. Foi possível visualizar as mais variadas relações entre os dados, o que possibilitou visualizar situações claras de comparações. Também foi possível entender o funcionamento de técnicas de preenchimentos de valores ausentes, problema recorrente em um conjunto de dados.

References

Andrienko, N. and Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media.

Bilogur, A. (2018). Missingno: a missing data visualization suite. *Journal of Open Source Software*, 3(22):547.

Callegari-Jacques, S. M. (2009). *Bioestatística: princípios e aplicações*. Artmed Editora.

Costa, C. F. G. d. (2018). Exploração de dados em falta: Uma abordagem visual. Master's thesis.

Hawthorne, G., Hawthorne, G., and Elliott, P. (2005). Imputing cross-sectional missing data: Comparison of common techniques. *Australian & New Zealand Journal of Psychiatry*, 39(7):583–590. PMID: 15996139.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

Pereira, H. I. (2017). Elementos de probabilidades e estatística. *Física, Licenciaturas*, page 35.

Zhang, Y. (2017). Imdb 5000 movie dataset. `https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset`. Acessado em 11/04/2019.