Uso de visualização de dados para auxiliar na análise e pré-processamento de dados categóricos

Lucas Bertoglio Ciocari, Alessandra Maciel Paz Milani, Isabel Harb Manssour
PUCRS - Pontifícia Universidade Católica do Rio Grande do Sul, Escola Politécnica
Porto Alegre, RS, Brasil
Email: lucas.ciocari@acad.pucrs.br, alessandra.paz@acad.pucrs.br, isabel.manssour@pucrs.br

***Resumo*—A quantidade de dados disponíveis sobre diferentes assuntos cresceu muito nos últimos anos, gerando novos desafios para os cientistas de dados e para diversas áreas de pesquisa, tais como mineração e visualização de dados. Com o objetivo de auxiliar na análise e pré-processamento de dados categóricos, esse trabalho propõe a inclusão de duas novas visualizações para dados categóricos à biblioteca *Pandas Profiling*. A primeira corresponde a um Diagrama de Cordas Bidimensional que possibilita ver a relação entre dados categóricos. A segunda é uma visualização do tipo *Display* Tabular que permite fazer uma análise de todo conjunto de dados. Estas visualizações visam auxiliar cientistas de dados na etapa de pré-processamento, principalmente com o objetivo de ajudar a entender o volume de dados a ser analisado.**

***Abstract*— The amount of data available on different subjects has grown in recent years, creating new challenges for data scientists and several research areas, such as data mining and data visualization. This work proposes the inclusion of two new visualizations for categorical data in the Pandas Profiling library to aid in the analysis and preprocessing of categorical data. The first is a Bidimensional Chord Diagram that allows users to see the relationship between categorical data. The second is a Tabular Display-type of visualization that enables users to analyze the entire dataset. These visualizations are intended to assist data scientists in the preprocessing step, primarily to help to understand the volume of data to be analyzed.**

I. Introdução

Segundo Wu et al. [1] (p. 1) "Em 2014, todo dia, 2.5 quintilhões de bytes de dados são criados e 90% dos dados disponíveis até 2014 haviam sido criados nos últimos dois anos". Estes dados possuem diversos formatos, estão em todo o lugar e oferecem diferentes informações. Eles podem ser obtidos de diferentes formas, através de pesquisas científicas ou pesquisas de satisfação, ou ainda podem ser coletados por sensores ou gerados por sistemas, como ao registrar o fluxo de caixa de uma empresa.

De acordo com Larose [2] (p. 2), "mineração de dados é o processo da descoberta de padrões e tendências úteis em grandes massas de dados". A melhoria da qualidade dos dados é feita na primeira etapa do processo de mineração, o pré-processamento. Han et al. [3] (p. 61-96) cita quatro técnicas para transformação de dados nesta etapa, sendo elas: (a) Limpeza de dados, que visa remover *missing values*, diminuir ruído e identificar *outliers*; (b) Integração de dados, que combina dados de diferentes fontes em uma fonte única; (c) Redução de dados, para extrair um *subset* do conjunto de dados para análise; (d) Discretização de dados, que divide dados contínuos em intervalos.

Os dados podem ser classificados em: categóricos (qualitativos), isto é, descritos por palavras; ou numéricos (quantitativos), normalmente resultantes de contagens, medições ou processamentos, conforme dito por McEvoy [4]. Podemos citar como exemplos de dados categóricos os nomes de fabricantes de carro ou o gênero de uma pessoa. Já dados numéricos podem ser discretos, como o número de andares de um prédio, ou contínuos, como a temperatura.

Considerando este contexto, este trabalho tem o objetivo de auxiliar na análise e pré-processamento de dados categóricos através da extensão da biblioteca *Pandas Profiling*[1]. Esta biblioteca recebe um conjunto de dados como entrada e disponibiliza um relatório geral destes dados. Sua extensão foi feita através da inclusão de duas novas visualizações à biblioteca: O Diagrama de Cordas Bidimensional e um gráfico do tipo *Display* Tabular. Por meio de um estudo de caso, foi possível mostrar que estas visualizações trazem os seguintes benefícios: (a) disponibilizam mais informações para o usuário na etapa de exploração e pré-processamento dos dados; e (b) facilitam a interpretação e análise destes dados apresentando a relação entre eles, informação que não seria facilmente perceptível com o uso de gráficos mais triviais (como o de barras ou de dispersão).

A Seção II apresenta alguns trabalhos relacionados. A Seção III descreve a modificação na *Pandas Profiling*. Um estudo de caso com dados do ENEM é apresentado na Seção IV, seguido pelas conclusões.

II. Trabalhos Relacionados

Alguns trabalhos que usam visualização de dados para auxiliar no pré-processamento de dados já foram desenvolvidos e alguns deles são abordados a seguir. O *Wrangler*, desenvolvido por Kandel et al. [5], é um exemplo de sistema que auxilia na transformação de dados de forma interativa. Já o *Attribute Radviz* proposto por Artur e Minghim [6], modifica o *RadViz* criando uma tabela de correlação entre âncoras e dados, e quanto maior a relação, mais perto o dado estará da âncora. O *Visplause*, descrito em Arbesser et al. [7], é uma ferramenta que busca exibir problemas de qualidade dos dados e, principalmente, a implausibilidade deles. Por fim, o

[1]https://github.com/pandas-profiling/pandas-profiling

trabalho de Milani [8] tem como principal objetivo a proposta de um modelo de pré-processamento para análise visual. Para isso, faz a extensão do modelo de análise visual proposto por Keim et al. [9], adicionando uma etapa de pré-processamento de mesma importância que as demais. Ainda, apresenta uma discussão de mecanismos para auxiliar os analistas de dados durante a etapa de pré-processamento por meio da visualização de informações, como a identificação de problemas de qualidade dos dados e a comparação entre os impactos da escolha de diferentes estratégias para a transformação desses dados.

Apesar dos benefícios trazidos por esses trabalhos, existem algumas funcionalidades que eles não fornecem. Por exemplo, o *Wrangler* não oferece técnicas de visualização de dados diferente da representação por tabela, não se aproveitando das vantagens de visualizações mais robustas para a análise de dados. O *Attribute-RadViz* precisa de um volume de dados já corrigido para gerar a visualização, o que o faz ser, por exemplo, uma boa forma de visualização para definir atributos para um algoritmo de aprendizado de máquina, mas não auxilia tanto quando o volume de dados apresenta problemas. O *Visplause* é um trabalho focado em plantas energéticas, por isso só avalia dados numéricos e não categóricos. Para terminar, o trabalho de Milani não apresenta visualizações com foco em dados categóricos.

III. Ferramenta para Análise de Dados

Com o intuito de auxiliar na análise e pré-processamento de dados categóricos, a *Pandas Profiling* foi modificada para inclusão de duas novas visualizações: diagrama de cordas bidimensional e *Display* Tabular. O diagrama de cordas bidimensional foi escolhido por ser uma visualização autocontida, ocupando sempre o mesmo espaço independente do número de nodos, e que possibilita o uso de dados categóricos. Portanto, considerando o *layout* do relatório da *Pandas Profiling*, este diagrama mostra-se adequado para o objetivo proposto. Já o *Display* Tabular foi escolhido por permitir uma análise visual do conjunto de dados como um todo. A inclusão destas visualizações na *Pandas Profiling* busca apoiar o usuário (tipicamente um cientista de dados) durante o processo de análise e pré-processamento de dados, facilitando a obtenção de *insights* sobre os dados sendo trabalhados para auxiliar a tomada de decisão sobre o processamento dos mesmos. Uma descrição destas visualizações é feita a seguir.

A. Contextualização

O trabalho de Milani [8] apresenta o resultado de uma pesquisa com profissionais da área de ciência de dados e uma das constatações foi de que a maior parte dos entrevistados ainda utilizam dados tabulares nas suas análises, portanto, este trabalho considera dados tabulares como entrada. A *Pandas Profiling* é uma extensão da biblioteca *Pandas*[2] e tem como objetivo principal apresentar de maneira mais completa informações sobre os dados que estão sendo trabalhados, resultando em uma melhoria sobre a função *describe*[3], que é responsável por gerar estatísticas descritivas. Para isso, a *Pandas Profiling* adiciona à *Pandas* a função *profile_report*, que oferece não apenas mais informações sobre o volume de dados, mas também apresenta várias visualizações como, por exemplo, dendrograma, gráfico de barras e de dispersão, entre outros. Esses gráficos apresentam informações sobre o volume de dados de uma forma mais intuitiva.

O funcionamento da *Pandas Profiling* pode ser dividido em duas etapas. A primeira utiliza a linguagem *Python* com o ambiente de desenvolvimento *Jupyter notebook* e faz o primeiro processamento em cima do volume de dados. É nessa etapa que as bibliotecas *Numpy*[4] (para computação científica) e *Pandas* (para estrutura e análise de dados) foram utilizadas. Já a segunda etapa da *Pandas Profiling* tem como objetivo construir o relatório utilizando *templates* em HTML. É nessa etapa que a biblioteca D3[5] foi utilizada para criação das visualizações em formato SVG[6]. Para que os dados processados na primeira etapa estejam disponíveis para a segunda etapa, é utilizada a biblioteca em *Python Jinja*[7], que implementa os *templates* HTML mencionados anteriormente.

B. Diagrama de Cordas Bidimensional

A primeira visualização desenvolvida e incluída na *Pandas Profiling* foi inspirada no trabalho desenvolvido por Humayoun et al. [10]. Nesse trabalho foi proposta uma extensão do diagrama de cordas, popularizado por Krzywinski et al. [11], que foi nomeado de diagrama de cordas bidimensional. Conforme afirmam Humayoun et al. [10] (p. 1), "diagramas de cordas são compactos e utilizam bem o espaço disponível, tendo alta densidade de informação". Essa visualização permite explorar a relação entre dois atributos de um volume de dados, bem como todas as categorias presentes nesses atributos. Cada categoria é representada por um nodo na visualização, e cada relação é representada por uma corda.

Como a *Pandas Profiling* é implementada em duas linguagens diferentes, assim também foram implementadas as visualizações propostas nesse trabalho. Primeiramente, se processa os dados com a linguagem *Python* para a visualização diagrama de cordas bidimensional. Nesse processamento, criou-se um dicionário que contém o nome do atributo, as categorias nele presentes e as posições que se encontram esses atributos na matriz de dados que será utilizada para registrar o número de vezes que a combinação de atributos acontece. Depois, foi feita a contagem de vezes em que acontecem todas as possíveis combinações de valores categóricos, e essa contagem é registrada na matriz de dados mencionada anteriormente. Essas informações são, então, inseridas no código *Javascript* por meio de *templates* da biblioteca *Jinja*. A implementação da visualização foi baseada em um exemplo da D3[8]. A

[2]https://pandas.pydata.org

[3]https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html

[4]https://numpy.org/

[5]https://d3js.org/

[6]https://developer.mozilla.org/pt-BR/docs/Web/SVG

[7]https://jinja.palletsprojects.com/en/2.11.x/

[8]https://www.d3-graph-gallery.com/graph/chord_colors.html

partir desse código, foram adicionadas as funcionalidades presentes na visualização, como a possibilidade de escolher quais colunas comparar, filtros ao passar o mouse pelas cordas, distinção de dados faltantes (*missing values*) utilizando cores diferentes. Além disso, foi feita uma modificação para ser bidimensional, pois o exemplo de base permite visualizar apenas uma dimensão por vez. Exemplos desta visualização são apresentados nas Figuras 1 e 2.

C. Display Tabular

A segunda visualização desenvolvida é baseada na técnica de *table lens*, apresentada por Rao e Card [12]. Conforme dito por Ward et al. [13] (p. 303), "uma visualização *table lens* disponibiliza mecanismos de visualização panorâmica dos dados bem como possibilita aplicar *zoom* em certas partes. Ainda, o ordenamento das colunas permite rapidamente identificar tendências, correlações e *outliers* no volume de dados".

Da mesma forma que o diagrama de cordas, esta visualização foi implementada em *Javascript* e D3. Novamente, os dados em forma de tabela são processados e são coletadas algumas características, sendo para dados numéricos coletado o maior e menor número e para dados categóricos coletado os valores únicos de cada coluna, valores que serão posteriormente utilizados para as legendas. Após o processamento, os dados são transferidos por *template* para o código em *javascript*. Na visualização, o comprimento da linha corresponde ao valor do dado numérico. Já para dados categóricos, cada valor único possui sua cor, destacando-se dos demais. Caso o dado seja faltante, ele será colorido como preto, em ambos os tipos de dados. Funcionalidades da visualização incluem comprimir ou expandir a visualização, exibir ou esconder a legenda, trocar as colunas de ordem, e ordenar as colunas por ordem alfabética (ascendente ou descendente). Um exemplo pode ser visto na Figura 3.

D. Uso das Visualizações

O código das visualizações propostas está disponível no *GitHub*[9]. A pasta *examples* possui um arquivo do *Jupyter Notebook* que exemplifica o uso da *Pandas Profiling* com as visualizações e de como utilizá-las de forma independente. É necessário baixar o repositório e instalá-lo, seguindo as instruções contidas na própria documentação.

IV. Estudo de Caso: Dados do ENEM 2018

O objetivo desta seção é mostrar o uso das visualizações desenvolvidas para obter *insights* sobre o volume de dados dos participantes do Exame Nacional do Ensino Médio (ENEM) do ano de 2018. Estes dados foram disponibilizados na internet[10] pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP). Estes dados foram escolhidos por serem reais, e também pelo seu aspecto social, permitindo extrair conhecimento sobre a situação social e a educação no Brasil.

[9]https://github.com/LucasCiocari/pandas-profiling

[10]http://inep.gov.br/microdados

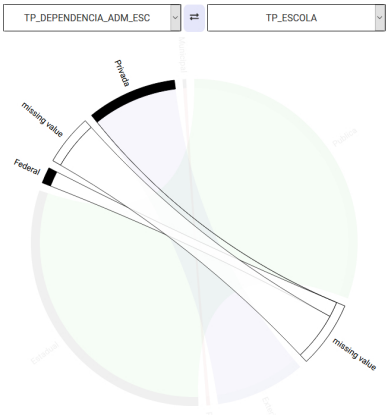

Fig. 1. Diagramas de corda bidimensional: *missing value* conectando em "Federal".

A. Conhecendo o volume de dados

Ao se cadastrar para o ENEM, o participante deve responder um questionário com informações pessoais sobre sua família, escola que frequentou, entre outras, sendo possível optar por não responder algumas perguntas. Essas informações são disponibilizadas junto com as notas e a presença do candidato. Estas informações foram utilizadas como um um estudo de caso para avaliar as visualizações desenvolvidas. Porém, nem todos os dados disponibilizados foram utilizados e outros foram agrupados a fim de gerar outras informações.

B. Primeiro exemplo: Diagrama de Cordas Bidimensional

Para o primeiro exemplo foram utilizados os registros dos candidatos que compareceram a todas as provas, iriam completar o ensino médio em 2018 e não fizeram a prova apenas como treinamento, o que resultou em aproximadamente um milhão e duzentos mil candidatos. A primeira comparação feita foi entre os atributos TP_ESCOLA, que é o tipo de escola que o candidato estuda, e TP_DEPENDENCIA_ADM_ESC, que é o tipo de administração da escola, como municipal, federal ou estadual. Ao escolher esses atributos foi possível notar que uma corda do nodo *missing value* do atributo TP_ESCOLA se conecta com o nodo "Federal", e outra corda se conecta ao nodo "Privada" do atributo TP_DEPENDENCIA_ADM_ESC, como mostra a Figura 1. Essas conexões indicam a possibilidade de classificar estes valores de *missing value* como "Pública" e como "Privada", respectivamente.

Além de demonstrar os problemas nos dados, também é importante mostrar a relação entre eles, pois a escolha dos atributos também faz parte da etapa de pré-processamento. A próxima análise utiliza os atributos REGIAO_RESIDENCIA, que é a região onde o candidato reside, e NU_IDADE, que é a idade do candidato, que foi agrupada. Assim, foi possível identificar que a maioria dos candidatos entre 16 e 18 anos reside na região sudeste, enquanto a maioria entre 19 e 20 anos reside na região nordeste, o que pode indicar um maior índice de reprovação de candidatos no ensino médio ou fundamental nessa região. Essa análise pode ser vista na Figura 2, na qual

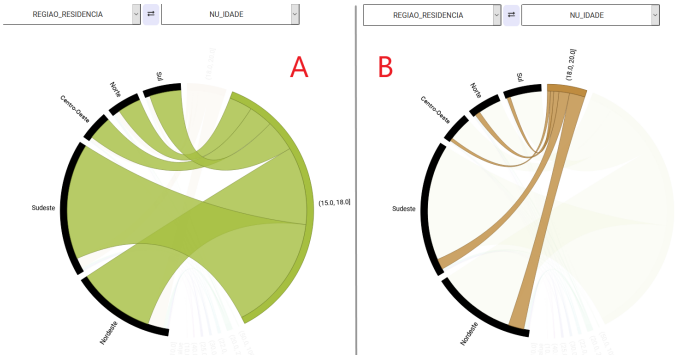

Fig. 2. Idades por região: 16 a 18 anos (A) e 19 a 20 anos (B).

é possível identificar que tanto a corda para região sudeste dos candidatos entre 16 e 18 anos (quarto nodo da Figura 2A) quanto a corda para região nordeste de candidatos entre 19 e 20 anos (quinto nodo da Figura 2B) são maiores.

C. Segundo exemplo: Display Tabular

O volume de dados separado para este exemplo utilizando o *Display* Tabular possui todos os candidatos formandos no ensino médio em 2018, que não declararam estar fazendo a prova como treinamento e que são do município de Porto Alegre, totalizando 3.338 candidatos. Uma primeira observação sobre estes dados pode ser feita ao olhar para a quarta e sexta coluna da Figura 3A, na qual há muitas linhas em preto, representando dados faltantes.

A Figura 3A mostra que ao ordenar o *Display* Tabular pela coluna NOTA_FINAL (sétima coluna) e examinar os dados, é possível notar que na primeira coluna, NU_IDADE, há dois valores que fogem do padrão dos demais, por serem os únicos que possuem o comprimento quase total da célula. Ao perceber esse detalhe, ordenamos os dados por essa coluna, do maior para o menor como na Figura 3B. Esta figura mostra que rapidamente os valores de NU_IDADE decrescem, o que significa que os primeiros valores são *outliers*.

V. Conclusões e Trabalhos Futuros

As visualizações implementadas e descritas neste trabalho visam auxiliar na análise e pré-processamento de dados categóricos. Assim, espera-se auxiliar os cientistas de dados a obterem *insights* a partir da análise de diferentes volumes de dados. O estudo apresentado mostra que as visualizações implementadas tornam a exploração do volume de dados mais simples, e ao mesmo tempo oferecem uma grande quantidade de informação de forma condensada.

Como trabalho futuro, será feita uma análise mais detalhada das visualizações adicionadas com especialistas de domínio. Assim, as visualizações propostas serão apresentadas para cientistas de dados para verificar os seus benefícios e identificar possíveis melhorias que podem ser implementadas. Como as visualizações também possuem versões separadas da *Pandas Profiling*, também é possível incluí-las em outras ferramentas, ou também adicionar novas visualizações a *Pandas Profiling*, como *dense pixel display*, descrito no livro de Ward et al. [13].

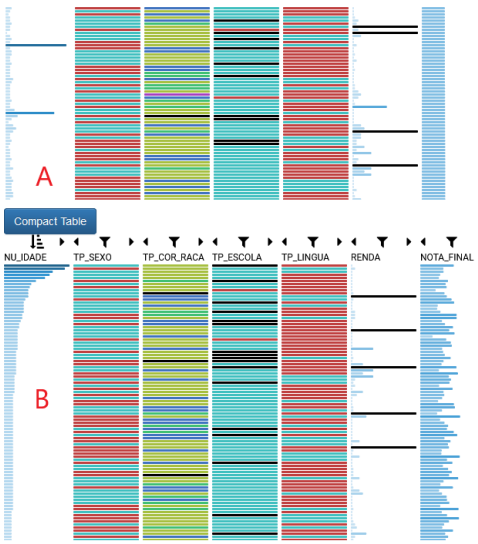

Fig. 3. Identificação de *outlier* utilizando o *Display* Tabular.

References

[1] X. Wu, X. Zhu, G. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, Jan 2014.

[2] D. T. Larose and C. D. Larose, *Discovering Knowledge In Data: An Introduction to Data Mining*. John Wiley and Sons, Inc., 2014.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2011.

[4] D. M. McEvoy, *A Guide to Business Statistics*. John Wiley and Sons, Inc., March 2018.

[5] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. ACM, 2011, pp. 3363–3372. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979444

[6] E. Artur and R. Minghim, "A novel visual approach for enhanced attribute analysis and selection," *Computers and Graphics*, vol. 84, pp. 160–172, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0097849319301438

[7] C. Arbesser, F. Spechtenhauser, T. Mühlbacher, and H. Piringer, "Visplause: Visual data quality assessment of many time series using plausibility checks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 641–650, January 2017.

[8] A. M. P. Milani, "Preprocessing profiling model for visual analytics," Master's thesis, Escola Politécnica – PUCRS, 2019.

[9] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, *Mastering the Information Age Solving Problems with Visual Analytics*. Eurographics Association, 2010.

[10] S. R. Humayoun, K. Bhambri, and R. AlTarawneh, "Bid-chord: An extended chord diagram for showing relations between bi-categorical dimensional data," in *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, ser. AVI '18. ACM, 2018, pp. 65:1–65:3. [Online]. Available: http://doi.acm.org/10.1145/3206505.3206570

[11] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Proceedings of the Genome Research*, vol. 19, p. 1639–1645, 2009.

[12] R. Rao and S. K. Card, "The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '94. New York, NY, USA: Association for Computing Machinery, 1994, p. 318–322. [Online]. Available: https://doi.org/10.1145/191666.191776

[13] M. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundation, Techniques and Applications*. CRC Press, October 2014.