Preprocessing Profiling Model for Visual Analytics

Alessandra Maciel Paz Milani *
School of Technology,
Pontifical Catholic University of
Rio Grande do Sul (PUCRS),
Porto Alegre, Brazil
Email: alessandra.paz@acad.pucrs.br

Fernando V. Paulovich
Faculty of Computer Science,
Dalhousie University,
Halifax, Canada
Email: paulovich@dal.ca

Isabel Harb Manssour
School of Technology,
Pontifical Catholic University of
Rio Grande do Sul (PUCRS),
Porto Alegre, Brazil
Email: isabel.manssour@pucrs.br

***Abstract*—Analyzing and managing raw data are still a challenging part of the data analysis process, mainly regarding data preprocessing. Although we can find studies proposing design implications or recommendations for visualization solutions in the data analysis scope, they do not focus on challenges during the preprocessing phase. Likewise, the current Visual Analytics processes do not consider preprocessing an equally important stage in their process. Thus, with this study, we aim to contribute to the discussion of how we can use and combine methods of visualization and data mining to assist data analysts during the preprocessing activities. To achieve that, we introduce the Preprocessing Profiling Model for Visual Analytics, which contemplates a set of features to inspire the implementation of new solutions. In turn, these features were designed considering a list of insights we obtained during an interview study with thirteen data analysts. Our contributions can be summarized as offering resources to promote a shift to a visual preprocessing.**

I. Introduction

Moving towards a data-driven society triggers new demands for data analysis. Although we have evolved in our data analysis capabilities, data preparation is still a challenging part of this process. This activity is mentioned as laborious and time-consuming [2]–[8], accounting up to 80% of the data analysis workflow [9].

We can observe variations in which tasks are considered part of the data preparation and how it is indicated in a data analysis workflow. However, a broad definition can be explained as "transforming the raw input data into an appropriate format for subsequent analysis" [3]. Also, several different methods are used for data understanding, e.g., similarity and dissimilarity between data objects, and for data transformations, e.g., aggregations and normalization or standardization of variables. This set of activities is identified in this work as preprocessing.

Moreover, the term *Preprocessing Profiling* was coined to indicate the activity of creating informative reviews while performing preprocessing activities. It is inspired by the concept of Data Profiling, i.e., creating informative summaries of a database [10].

Most of the visualization studies are supporting just the last phases of the data analysis process, and even though we can find studies proposing visualization methods to assist with preprocessing, they are predominantly focused on data transformation activities [4], [11], or limited to particular scenarios or data types, e.g., time series data [12]. Thus, we can still observe opportunities, such as (a) alternative visualizations to explore data quality issues; (b) visualizations to support the evaluation of preprocessing impacts in further phases; and (c) creating a list of guidelines and features to support novel visualizations in the context of preprocessing.

Additionally, for many Visual Analytics (VA) processes, such as [13] and [14], preprocessing is not typically considered as an equally important phase, such as Data, Visualization, and Models phases. Furthermore, the preprocessing is described as part of a batch or waterfall approach, and its activities, when detailed, are regarding to data transformation. However, the preprocessing activities should be considered part of the entire cycle, not only because it requires multiple interactions through the whole data analysis workflow but also due to its impacts on the other phases [7], [8].

In this scenario, we built our study efforts to answer the research question *How can we assist the preprocessing activities with visualization techniques during a visual analytics process?* Thus, our main objective was to explore visualization techniques to support the activities performed by data analysts during the preprocessing phase. In particular, the raw data understanding and the evaluation of the preprocessing strategies and its impacts to further phases of the process. To achieve that, different actions were executed, and they are explained in the next subsection.

A. Research Design

Figure 1 depicts the main activities performed in our study. We also correlate the research activities that produced the most relevant outcomes, i.e., 6, 7, 8, and 10. These contributions are described in Subsection I-B.

Based on the research problem (1), the next activities comprised the search and literature review of the state-of-the-art (2) and the evaluation of tools and systems (3). Afterward, the research design was revised (4), and we decided to interview data analysts to capture relevant information about their processes and needs (6). The analysis resulting from this activity contributed as requirements to design the Preprocessing Profiling Model (7) as close as possible to attend real situations. However, to proceed with the interviews (6A), an additional activity was added (5), we followed the Research Ethics Committee (REC) protocol to validate

* This paper relates to a M.Sc. dissertation of the first author [1].

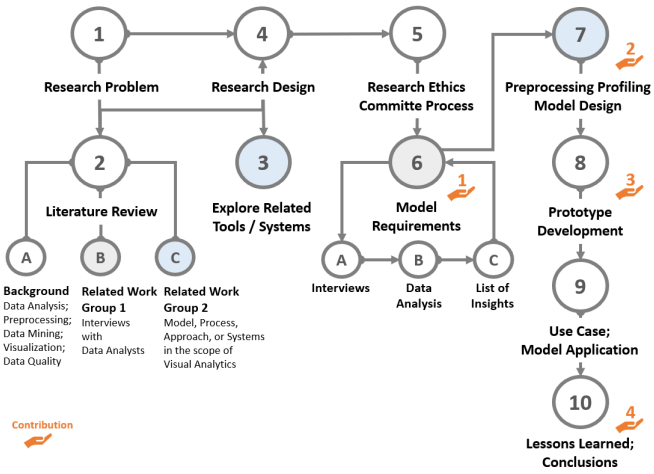

Fig. 1. Research Design: list of the main activities and contributions.

our study and get their approval. The complete documentation is available on Plataforma Brasil [15] (CAAE number 89239418.0.0000.5336). While waiting for the REC process, we performed a complementary review on related work (2B, 2C, and 3).

With the completion of the Preprocessing Profiling Model Design (7), we developed a prototype (8) to support the application and validation of our proposal (9). Finally, based on prior activities, we identified a set of lessons learned (10).

B. Contributions

We list four contributions with our study for the established researchers and newcomers in the VA research area.

1) A list of ten insights obtained from the interview process with data analysts. These insights can be used as a set of requirements for future visualization research studies applied to preprocessing in data mining (DM) workflows. Furthermore, we provide details on the profile of the data analysts, the main challenges they face, and the opportunities that arise while they are engaged in DM projects in diverse organizational areas.
2) A novel conceptual Model for the VA process considering *Preprocessing Profiling* as a new phase. This raises awareness of the importance of preprocessing activities, which is not handled to a sufficient extent in the existing VA processes.
3) A prototype design and usage scenarios, to showcase the possibilities and foster the discussion of the applicability of the Preprocessing Profiling Model for VA.
4) The discussion of lessons learned and research opportunities in the scope of preprocessing, visualization, and VA as possible directions for advancing the area.

This paper is structured to present these contributions as follows. To begin, the first contribution is described in Section II; the second in Section III; the third in Section IV; and the fourth is summarized in Section V. To conclude, we present in Section VI our final considerations, as well as future work and the current impact of our study.

II. Insights for New Visualizations

We developed a semi-structured questionnaire to guide the interviews with data analysts. Most of the questions were open-ended to capture as much information as possible during the interviews. Some questions covered the participant's profile with a few demographic items. Others were intended to encourage the participants to describe their working practices to provide an overview of their data exploration processes. Besides, some questions were phrased to address the visualization strategies as part of the preprocessing activities. Figure 2 shows a portion of the data collected in this process. The detailed information of this interview study is in [1]-Chapter 4.

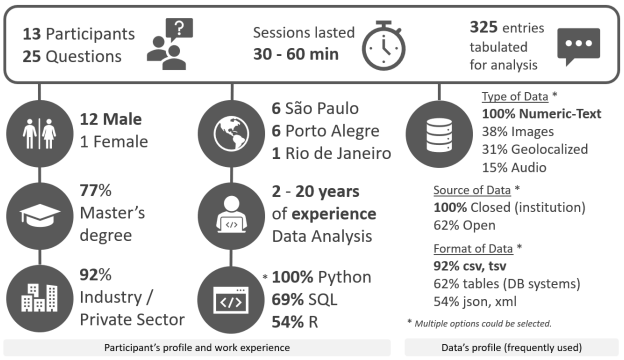

Fig. 2. Overview of the interview study protocol, participant's profile and work experience, and their data's profile.

During our interviews, only one participant mentioned visualization was not a differential for the activities they were performing during preprocessing. Two other participants expressed they felt confident with their set of tools. However, the ten remaining participants demonstrated an interest in different ways to explore their data with visualization techniques. The discussion about the challenges and opportunities based on the responses of the interviewees resulted in a list of insights. Additionally, this list was compared with the closest related work [16]–[18], improving the reliability of our findings and providing background, as a consolidated set of requirements. The ten insights for new visualizations in the scope of preprocessing are explained as follows.

1) Keep it simple: In most of the cases, the existing visualizations or more traditional charts should fulfill the demand.

2) Keep the context: New solutions should prevent the data analysts from losing the context under investigation while alternating among several different tools. Furthermore, they should allow the evaluation of multiple rows and attributes on the same view, avoiding the change blindness effect [19].

3) Save the time: The new solutions should consider intuitive features and little need for configuration and/or coding, aiming to keep the agility in the working process.

4) Think BIG: Even though not all participants mentioned this item as critical in their scope (5 of 13), Big Data is a growing demand, and the development of techniques that can handle this scenario is urged.

5) Allow interaction: It is vital to provide more than static reports and allow the data analyst to perform flexible data manipulation within visualization tools.

6) Tables are OK: Most of the participants are still using tabular data during their analysis. Therefore, aligned with the Insight 1, the tabular format is considered the appropriate choice for visual representation.

7) Pay attention to the work scopes: During our interviews, two work scopes were indicated as lacking attention by current visualizations solutions. One concerns the creation of new variables/features. The other is related to the deep learning scope for visual interpretation of why each decision was made.

8) Preprocessing is part of the entire cycle: Multiple interactions are required among preprocessing activities and all the other stages during the same cycle. Except for confirmatory analysis, where most of the process is already automated and little interaction is needed, but for most of the other cases, especially for the initial data exploration, multiple back and forth in the raw data can be expected.

9) Allow comparison: Considering allow the comparison of data before and after its transformation is essential to support the preprocessing decision.

10) Capture metadata: Besides the two previous insights, if automatic exploratory tasks or data transformations are needed, it is important to present the logic underneath them. The data analysts desired to continue working with the control and visibility of what the tool was doing.

III. Preprocessing Profiling Model

We identified the opportunity to explore the preprocessing activities as part of the VA process. Firstly, motivated by the insights obtained from the data analysts (Section II), preprocessing should not be seen as part of a batch or waterfall approach, but as an activity being constructed during the whole cycle. Also, the preprocessing decisions at this phase may have significant impacts on the next steps of the process. Secondly, although we can find related work proposing VA Models or visualization methods to assist with preprocessing activities, we can still observe opportunities to be explored. Hence, to continue this discussion and support filling these gaps, we proposed the Preprocessing Profiling Model for VA, presented in Subsection III-A. Moreover, we outline nine features to be observed while designing new solutions in compliance with our Model, introduced in Subsection III-B.

A. Process

The Preprocessing Profiling Model for VA is formalized as an extension of the VA process proposed by Keim et al. [13]. This is illustrated in Figure 3, we include a new phase called Preprocessing Profiling, and new possible transitions among the phases:Dataset Understanding, Data Preparation Understanding, Visualization of Preprocessing, Model Testing, and another Feedback Loop.

Despite the description of some preprocessing activities in the original Data phase (e.g., data cleaning, normalization, and other tasks as part of the Transformation transition), by adding

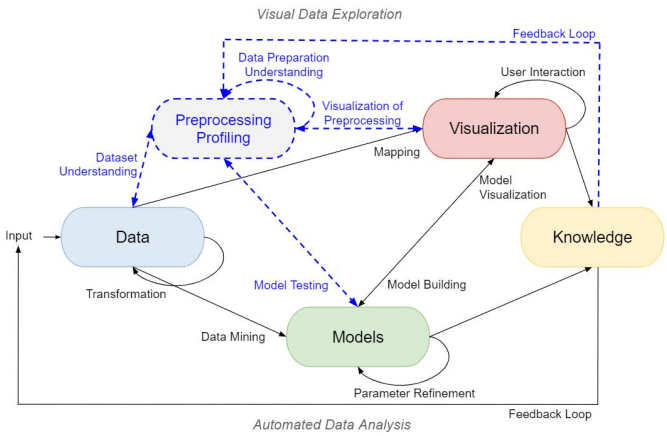

Fig. 3. Overview of the Preprocessing Profiling Model for Visual Analytics. Each node (represented through a rounded rectangle) corresponds to a different phase, and its transitions are represented through arrows. The new objects are represented in blue color for the text font and dashed lines.

Preprocessing Profiling as a phase, we put activities such as the data profiling and the evaluation of preprocessing strategies prior to Model Building in the critical path, i.e., as an equally important phase. Further details are available in [1]-Chapter 5.

B. Features

Based on the literature review (Fig. 1-2C) and, especially, in the list of the insights (Fig. 1-6C), we propose nine features to be considered as part of the Preprocessing Profiling Model for VA's implementation. They are briefly explained here.

1) Unified: The integration with the most used tools for data analysis, such as Python and R.

2) Large Scale: The ability to work with scenarios dealing with huge volumes of data. Evaluate how to produce partial results while the data are being processed.

3) Metadata: The ability to generate informative summaries of the preprocessing activities. The data computation of other features (e.g., 4 and 5) can be used as input. It should be a decisive output of the Preprocessing Profiling process.

4) Data Mining: The use of DM methods to support preprocessing activities. The data quality assessment can benefit from the use of DM algorithms, e.g., identifying data errors and recommendations on data transformation. Additionally, supporting the validation of the preprocessing strategies and DM model testing.

5) Statistics: The use of statistical methods to generate a detailed description of the data and to support preprocessing activities. A thorough review of the characteristics of the variables is relevant to make decisions on data transformation demands, not only to fix data issues but also to better integrate it with the planned DM model.

6) Comparison: The ability to compare the data prior and after transformations and the impacts of the preprocessing decisions. Preprocessed data should be compared to the original data. Moreover, when combined with Feature 4, this feature can support the evaluation of the DM model based on different preprocessing strategies.

7) Recommendation: The use of recommendation systems to propose visualizations according to the type and volume of data under investigation. Also, taking into consideration the particularities of the DM scope or data quality issues.

8) Template: The ability to generate automatically initial visualizations or basic templates based on the data under analysis. This feature, when combined with Feature 7, should avoid some inappropriate uses.

9) Interaction: The use of visualization interaction techniques to support flexible data exploration.

IV. Application

To assist in the validation of the Preprocessing Profiling Model, we developed a prototype solution, described in Subsection IV-A, which is used during the usage scenario description presented in Subsection IV-B.

A. Prototype Design

The developed prototype generates two dynamic reports: *Data Profiling*[1] and *Preprocessing Profiling*[2]. It was written mainly in Python and Javascript. When considering an example of integration with Jupyter Notebook, few lines of code are required to import the library and call the method with an input dataset; then, it will generate the report (HTML) in the output cell. Additional details in [1]-Chapter 5.4.

The *Data Profiling* report supports the dataset understanding. It was developed as an extension of the Pandas Profiling [20]. The main sections are identified as: Overview (Fig. 4-a), Variables (Fig. 4-b), Correlations (Fig. 4-c), Missing Values, and Sample.

The *Preprocessing Profiling* report supports the evaluation of data transformation impacts on the DM model. For this first version, we considered one DM problem (classification), one data issue (missing values) to perform the data transformations, and one type of dataset (tabular data). Overall, the report performs the following tasks: (a) reads a dataset informed and split the data into training and testing. Next, for each imputation strategy, (b) does the data transformations; (c) trains the classification model; (d) runs the testing to predict the classes; (e) creates metadata of preprocessing; and (f) generates the visualizations and the HTML document.

Regarding the task (b), five different strategies of data imputation are considered. One removes all the rows with at least one missing value, and it is named *Baseline (no missing)*. Other replaces all missing values by zero, named *Constant(=zero)*. A third and fourth replace missing values by mean and median values computed based on all records on the same column. The fifth replaces missing values by the *Most Frequent* value on the column.

B. Usage Scenario

In this hypothetical usage scenario, we present a persona named Tim, a biology student. In Figure 5, we illustrate the pathways performed by Tim during his activities.

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

Fig. 4. *Data Profiling* Report.

Tim is searching for strategies on how to solve the taxonomic problems of his current research. He has collected data about a group of Iris flowers, and he is interested in identifying Iris species by the attributes measured from a morphological variation of the flowers. Tim's dataset contains 186 samples (36 more than the original Iris dataset [21]) from three different species of Iris, i.e., Setosa, Virginica, and Versicolor. For each sample, four attributes were measured in centimeters: sepal length, petal length, sepal width, and petal width. Additionally, a fifth attribute informs the corresponding class of each sample. However, Tim was not able to get all the data for the new samples, then his dataset has data quality problems, i.e., it contains missing values and outliers.

Tim is familiar with the Python programming environment. To begin, he tries to run a classification model using his dataset without any transformation. However, he could not move forward since an error message is returned informing him the classification algorithm cannot proceed due to missing values in the dataset (Pathway 1A). Therefore, he replaces all the missing values by the number zero. He reruns the classification model and visualizes the model results, but he is not confident about the results obtained. Due to the uncertainty of his previous results (Pathway 1B), Tim decides to use the Preprocessing Profiling Model for VA to guide his activities.

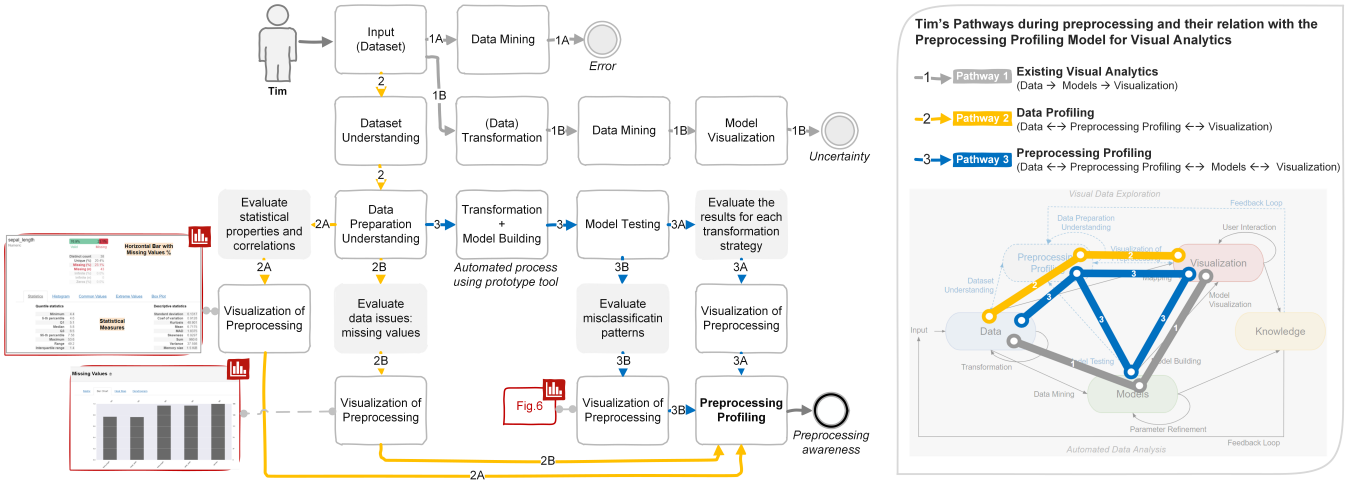

Fig. 5. Usage scenario - The pathways took by Tim: **1 (connection lines in gray)** considering the existing VA process, and so not focusing in preprocessing activities, except by elementary data transformation; **2 (yellow)** focusing on the dataset understanding; **3 (blue)** concentrate on the impacts of preprocessing strategies. On the right of this figure, each pathway is related to the Preprocessing Profiling Model process. We describe the paths as sequential steps to facilitate the usage scenario explanation, but the idea is to allow multiple backward and forward between the phases.

He starts by running descriptive statistics in Python. However, many lines of code and outputs with plain text would be required to generate all the information he wants. Then, he decides to use an alternative solution (the prototype, Subsection IV-A). With the *Data Profiling* report, he got an overview of the dataset. Based on that, he realizes the petal columns are highly correlated with each other. Even though Tim had previously generated the covariance and correlation matrix, he considered it was challenging to observe the relation between two variables just by looking at the plain text. Also, still part of **Pathway 2A**, he explores each variable of his dataset, which allowed him to confirm the presence of data issues.

Additionally, he explores the *Missing Values* section of the report, and despite the observation of the total amount of 10% of missing values, no significant pattern in relation to these occurrences is noted for his dataset (**Pathway 2B**).

Next, Tim wants to evaluate the impacts of the preprocessing strategies on his classification problem, and he informs his dataset as input to the *Preprocessing Profiling* report. Since all the data transformation and model building are done automatically, Tim takes advantaged of the time saved and he runs multiple rounds (of training and testing) to evaluate the results of classification and other possibilities.

Although the classification results varied in each round, Tim is still able to notice differences among the imputation strategies for all rounds performed. For example, the class of Iris Setosa was initially clear to classify, but, with the presence of data issues and the need to perform data transformation, the classification results are negatively impacted. Tim observes a significant variation on the accuracy metric for the *Mean* imputation strategy compared to the others. With that, it is clear to him that he needs to identify outliers, e.g., using visualizations such as Boxplot, and remove them before continuing, or, for this particular case, he could use the *Median* imputation strategy to avoid data with high magnitude to dominate results. These activities correspond to **Pathway 3A**.

Furthermore, while comparing the *Flow of Classes* visualization for different rounds (**Pathway 3B**), he notes that, even for a classification resulting in the same accuracy, there is variation in each group of classes being misclassified. For instance, when he ran a round using the four variables (Fig. 6-a), four imputation strategies resulted the same accuracy (91.1%). However, he could notice an additional flow of classes from actual class 2 (Versicolor) to predicted class 3 (Virginica) during *Constant* and *Most Frequent* imputation strategies. While for *Mean* and *Median*, the misclassification occurred only from actual class 3 (Virginica) to predicted class 2 (Versicolor). Likewise, when observing the results for another round, which considered only two variables (Fig. 6-b), he noticed variations on the misclassifications.

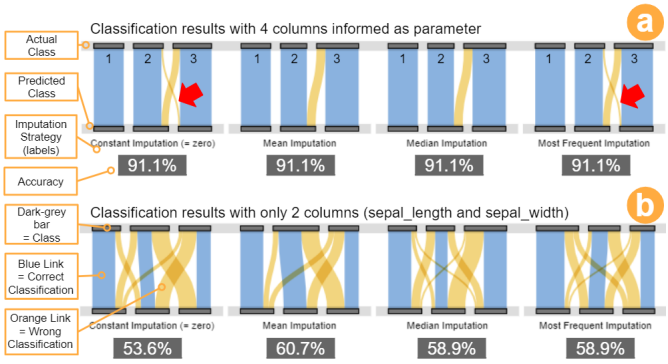

Fig. 6. *Preprocessing Profiling* report. Classification results for different missing values imputation strategies using the *Flow of Classes* Visualization.

In conclusion, Tim takes these insights as reinforcement of the importance of exploring preprocessing strategies before moving to further phases in the VA process or any DM workflow, mainly when the dataset has data quality issues.

This process is shown in Fig. 5 as **Pathway 3**, which, when combined with **Pathway 2**, promotes awareness of the preprocessing profiling.

V. Lessons Learned

In addition to the interview study, an extensive literature review was done to build the proposed Preprocessing Profiling Model. Thus, the list of insights (Section II) and the features description (Subsection III-B) summarize most of what we have learned in this process.

As an important remark, one participant of our interview stated that despite preprocessing activities being fundamental and at some level performed by all data analysts, few people are truly proficient at them. Hence, this visual support could help the data analysts adopt visualization as part of their daily strategies, and so promote awareness of the preprocessing.

Through practicing on the developed prototype, three main advantages can be mentioned. First, we can generate detailed and relevant information to support preprocessing activities by executing only one command line to import the library and another to call the report (considering the dataset is already loaded in the Python programming environment). Consequently, we are contributing to simplify the working procedures of data analysts, which is a concern since preprocessing is reported as one of the most laborious tasks [9]. Second, as the reports present several metrics and visualizations by default, metrics that could be neglected by the data analyst due to unawareness, difficulties in applying, or limitation of time, can now be incorporated as part of their analysis. Third, the detailed information about the dataset and data preparation can be used as metadata for preprocessing, helping to build the principle of transparency on activities performed as part of the preprocessing phase.

Furthermore, although most of the visualizations used are simple, they still demonstrate more benefits to understand the data when compared to viewing the plain text. Also, the simplicity should favor understanding since it does not require a prior explanation, i.e., most of the visualizations are already part of the data analysts' culture.

VI. Final Considerations

We presented the results obtained from the interview process with thirteen data analysts to understand their data analysis practices and how they use visualization during preprocessing. The main output of this process was the organization of the challenges and opportunities identified during our analysis of the responses, which resulted in a list of insights.

Based on the list of ten insights and the review of the related work, we proposed the Preprocessing Profiling Model for VA as an alternative to support the data analysts during the preprocessing phase. We advocate that the quality of the data and the decision making on data preparation strategies can be improved by enabling better methods for data understanding and evaluation of preprocessing impacts. Moreover, we presented the design of a prototype solution that was used during our Model validation.

Within this study, we raise awareness of the relevance of the preprocessing activities and its current opportunities in the VA processes. Ultimately, it contributes as a source of requirements to fill the visualization gap during the preprocessing and to instigate a visual preprocessing approach during the data analysis workflow.

A. Future Work

The Preprocessing Profiling Model for VA was designed to be extensible to a variety of scenarios. During our usage scenarios we explored a limited number of combinations, thus, further studies can extend our Model to explore various application domains, DM problems, data types, data quality issues, data transformation strategies, and visualization techniques.

Regarding the developed prototype, multiple enhancements could be listed such as the scalability to handle big data volumes (Feature 2) and recommendation systems (Feature 7). In the scope of new visual metaphors, the understanding of misclassification patterns (resulting from the different data transformation strategies) remains an excellent opportunity. Finally, future work should consider conducting in-depth interviews with the participation of domain experts to validate the process and the usability of the visualizations.

B. Current Impact

As a result of this study, we published on the Information Visualization Journal (2019 Impact Factor: 1.325) the study [22], which covers the scope presented in Section II. Also, the content presented in Sections III and IV has been extended, and its submission to a Journal is in progress.

Besides the publications, as expansion of this study, we can mention four main items: (i) open-sourced the prototype tool developed. (ii) Scientific Initiation Project approved and financed by PUCRS (BPA Student Research Scholarship Program, General Call 1/2019 and 1/2020), which allowed the engagement of an undergraduate student. (iii) This study has already been extended as part of an undergraduate thesis [23].

Finally, (iv) we were awarded the Emerging Leaders in the Americas Program (ELAP 2018-2019), administered by the Canadian Bureau for International Education, which allowed the first author of this paper to work as a visiting researcher in Canada during the first semester of 2019. As part of that, the content of this study was discussed with researchers from different countries and the informal feedback received on it brought confidence in our contributions to the data analysis, visualization, and the visual analytics research areas.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Also, this study was achieved in cooperation with Hewlett Packard Brasil LTDA. and HP Brasil Indústria e Comércio de Equipamentos Eletrônicos LTDA. using incentives of Brazilian Informatics Law (Law nº 8.2.48 of 1991). Lastly, with support of the Government of Canada.

REFERENCES

[1] A. M. P. Milani, "Preprocessing profiling model for visual analytics," Master's thesis, School of Technology, PUCRS, Porto Alegre, 2019. [Online]. Available: http://tede2.pucrs.br/tede2/handle/tede/9007

[2] W. Kim, B. Choi, E. Hong, S. Kim, and D. Lee, "A taxonomy of dirty data," *Data Mining and Knowledge Discovery*, vol. 7, pp. 81–99, 2003.

[3] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson Education, 2006.

[4] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the Conference on Human Factors in Computing Systems*, 2011, pp. 3363–3372.

[5] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. Van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling: Visualizations and transformations for usable and credible data," *Information Visualization*, vol. 10, pp. 271–288, 2011.

[6] H. Wickham, "Tidy Data," *Journal of Statistical Software, Articles*, vol. 59, pp. 1–23, 2014.

[7] S. Krishnan, D. Haas, M. J. Franklin, and E. Wu, "Towards reliable interactive data cleaning: A user survey and recommendations," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, 2016, pp. 1–5.

[8] C. Turkay, N. Pezzotti, C. Binnig, H. Strobelt, B. Hammer, D. Keim, J.-D. Fekete, T. Palpanas, Y. Wang, and F. Rusu, "Progressive data science: Potential and challenges," *arXiv preprint*, vol. 1812.08032, pp. 1–10, 2018.

[9] T. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, 2003.

[10] T. Johnson, "Data profiling," *Encyclopedia of Database Systems*, pp. 604–608, 2009.

[11] S. Kandel, R. Parikh, A. Paepcke, J. Hellerstein, and J. Heer, "Profiler: Integrated statistical analysis and visualization for data quality assessment," in *Proceedings of the Conference on Advanced Visual Interfaces*, 2012, pp. 547–554.

[12] T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, and N. Suchy, "TimeCleanser: A visual analytics approach for data cleansing of time-oriented data," in *Proceedings of the 14th international conference on knowledge technologies and data-driven business*, 2014, pp. 1–8.

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

[14] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, pp. 1604–1613, 2014.

[15] DATASUS, "Plataforma Brasil." [Online]. Available: http://plataformabrasil.saude.gov.br/login.jsf

[16] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Enterprise data analysis and visualization: An interview study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, pp. 2917–2926, 2012.

[17] A. Batch and N. Elmqvist, "The interactive visualization gap in initial exploratory data analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 278–287, 2018.

[18] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst, "Futzing and moseying: Interviews with professional data analysts on exploration practices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, pp. 22–31, 2019.

[19] R. Rensink, "Seeing, sensing, and scrutinizing," *Vision Research*, vol. 40, pp. 1469–1487, 2000.

[20] Pandas-profiling, "Create HTML profiling reports from pandas dataframe objects." [Online]. Available: https://github.com/pandas-profiling/pandas-profiling

[21] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[22] A. M. P. Milani, F. V. Paulovich, and I. H. Manssour, "Visualization in the preprocessing phase: Getting insights from enterprise professionals," *Information Visualization*, vol. 19, pp. 273–287, 2020.

[23] L. Ciocari, "Uso de visualização de dados para auxiliar no pré-processamento de dados categóricos," Undergraduation thesis, School of Technology, PUCRS, Porto Alegre, 2019.