Info Vis

Article

Information Visualization
2021, Vol. 20(2-3) 101–122

DOI: 10.1177/14738716211021591
journals.sagepub.com/home/ivi
SAGE

PrAVA: Preprocessing profiling approach for visual analytics

Alessandra Maciel Paz Milani[1,2], Lucas Angelo Loges[2], Fernando Vieira Paulovich[3] and Isabel Harb Manssour[2]

Abstract
To accommodate the demands of a data-driven society, we have expanded our ability to collect and store data, develop sophisticated algorithms, and generate elaborated visual representations of the data analysis process outcomes. However, data preprocessing, as the activity of transforming the raw data into an appropriate format for subsequent analysis, is still a challenging part of this process. Although we can find studies that address the use of visualization techniques to support the activities in the scope of preprocessing, the current Visual Analytics processes do not consider preprocessing an equally important phase in their processes. Hence, with this paper, we aim to contribute to the discussion of how we can incorporate the preprocessing as a prominent phase in the Visual Analytics process and promote better alternatives to assist the data analysts during the preprocessing activities. To achieve that, we are introducing the Preprocessing Profiling Approach for Visual Analytics (PrAVA), a conceptual Visual Analytics process that includes Preprocessing Profiling as a new phase. It also contemplates a set of guidelines to be considered by new solutions adopting PrAVA. Moreover, we analyze its applicability through use case scenarios that show resourceful methods for data understanding and evaluation of the preprocessing impacts. As a final contribution, we indicate a list of research opportunities in the scope of preprocessing combined with visualization and Visual Analytics to stimulate a shift to visual preprocessing.

Keywords
Information visualization, visual analytics, preprocessing, data preparation profiling

Introduction

Moving toward 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 frequently mentioned as laborious and time-consuming.[1–8] According to Dasu and Johnson[9] (p. IX), "the tasks of exploratory data mining and data cleaning constitute 80% of the effort that determines 80% of the value of the ultimate data mining results."

We can observe variations in which tasks are considered part of the data preparation and how they are indicated in a data analysis process.[8] However, in general, data preparation is the process "to transform the raw input data into an appropriate format for subsequent analysis" (Tan et al.,[2] p.3). As part of this process, several different strategies, methods, and techniques are used for data understanding, for example, similarity and dissimilarity between data objects, and for data transformations, for example, aggregations and normalization or standardization of

[1]University of Victoria, Victoria, BC, Canada
[2]Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil
[3]Dalhousie University, Halifax, NS, Canada

Corresponding author:
Alessandra Maciel Paz Milani, University of Victoria, Department of Computer Science, Office ECS 542, 3800 Finnerty Road, Victoria, BC, V8P 5C2, Canada.
Email: amilani@uvic.ca

variables. This set of activities is identified in this work as preprocessing, but this term is also referenced in the literature as data wrangling,[3] data cleaning, or scrubbing.[10]

Data quality problems are present in most datasets, due to misspellings during data entry, missing information, or other invalid data. Moreover, when multiple data sources need to be integrated, the need for preprocessing increases.[10] Although automated processes are fundamental and accessible in this context, the data analyst's participation in the decision of how this data should be transformed is still critical in many cases.[1,4,6,11,12] To support the cases when the "human in the loop" is vital to data preprocessing, the use of visualization techniques can play an essential role in data analysis while providing meaningful insights[4,13,14] since one of the strengths of visualization is enabling users to quickly identify erroneous data.[15]

Nevertheless, most of the works in the scope of visualization are focused on supporting just the last phases of the data analysis process. Even though we can find studies proposing visualization methods to assist with preprocessing, they are predominantly focused on data transformation activities, for example, Kandel et al.,[3,16] or limited to particular scenarios or data types, for example, time series data Bernard et al.[17] and Gschwandtner et al.[18] Thus, we can still observe opportunities, such as (a) alternative visualizations to explore data quality issues; (b) visualizations to support the evaluation of the preprocessing impacts in further phases; and (c) creating a list of guidelines to support novel visualizations in the context of preprocessing.

Additionally, for many Visual Analytics (VA) processes, such as in Keim et al.[19] and Sacha et al.,[20] the preprocessing phase is not acknowledged as important as Data, Visualization, Models, or Knowledge phases. Furthermore, the preprocessing is described as part of a batch or waterfall approach inside one of the existing phases, and its activities, when detailed, are basically with regards to data transformation. However, as discussed by Krishnan et al.[6] and Milani et al.,[8] preprocessing activities should be considered part of the entire process, not only because these activities require multiple interactions through the whole data analysis process but also due to their impact on the other phases.

This paper aims to raise awareness of these issues seeking to answer the research question: *How preprocessing activities can be effectively incorporated into the VA process?* Based on an extensive literature review around the topic, we derived nine different guidelines for consolidating preprocessing in the VA workflow, discussing their purpose, and presenting examples found in the literature. As a result, we extend the VA process to accommodate our findings and acknowledge the preprocessing phase's importance, aiming at enabling data analysts to increase their ownership of the data under analysis, master the impacts of preprocessing activities, and contributing to more trustworthy knowledge discovery. The main contributions of this paper are:

- A list of nine guidelines to be considered by VA solutions to incorporate preprocessing in the analysis life-cycle, presenting different examples found in the literature;
- A conceptual process, named **Preprocessing Profiling Approach for Visual Analytics (PrAVA)**, extending the existing VA process to raise awareness of the importance of preprocessing activities and accommodate the derived guidelines;
- Further research opportunities in the scope of preprocessing, visualization, and VA for advancing the area.

We use the term *Preprocessing Profiling* to indicate the activity of creating informative summaries while performing the data preprocessing activities. This term was inspired by the concept of Data Profiling, defined by Johnson[21] as the activity of generating informative summaries of a database (e.g. the total number of missing records in a table).

The structure of this paper follows the order of steps taken in the development of this work. First, in the *Related work* section, we present an extensive literature review involving preprocessing activities in VA scenarios

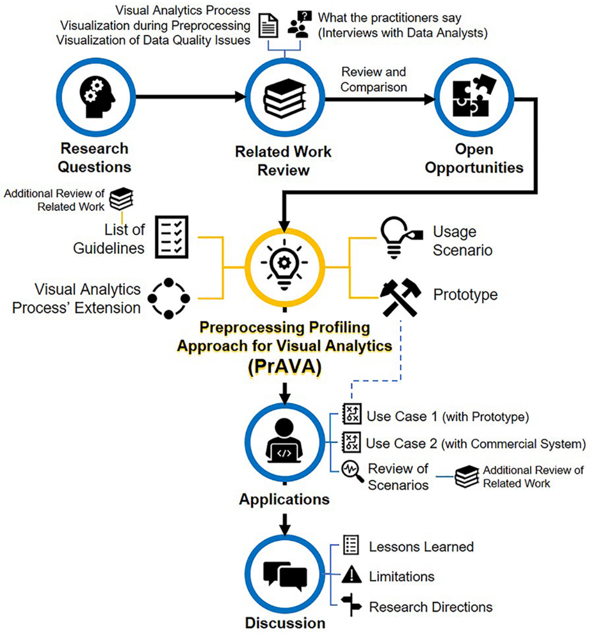

Visual Analytics Process
Visualization during Preprocessing
Visualization of Data Quality Issues
What the practitioners say
(Interviews with Data Analysts)
Research Questions
Related Work Review
Review and Comparison
Open Opportunities
Additional Review of Related Work
List of Guidelines
Visual Analytics Process' Extension
Usage Scenario
Prototype
Preprocessing Profiling Approach for Visual Analytics (PrAVA)
Use Case 1 (with Prototype)
Use Case 2 (with Commercial System)
Review of Scenarios
Additional Review of Related Work
Applications
Lessons Learned
Limitations
Research Directions
Discussion

Figure 1. Overview of this work.

that serve as background and motivation for this work. Then, we describe the guidelines derived from the literature and the PrAVA process in the *Preprocessing profiling approach* section. The following sections present a potential *Usage scenario* and *Applications* as part of the validation of our proposal. In the *Discussion* section, we explain the lessons learned and limitations of this work and research opportunities. In the last section, we outline our *Conclusions*. Figure 1 presents an overview of these steps.

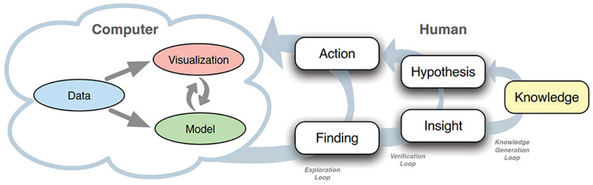


Figure 3. The knowledge generation model for VA proposed by Sacha et al.[20]

Related work

This section covers related work that serve as background and, at the same time, influenced the Preprocessing Profiling Approach for Visual Analytics (PrAVA). These works are grouped in four subsections according to their focus on Visual Analytics process, visualization during preprocessing, visualization of data quality issues, or interviews with practitioners. Finally, we present a review and comparison of the selected related work.

Visual analytics process

As part of the Visual Analytics (VA) discussion, Keim et al.[19] contribute with an overview of the different phases in the VA process. Their process (Figure 2) combines automatic and visual analysis methods with human interaction to gain insights and promote knowledge generation. Despite their notorious relevance to the VA area, their process does not detail the importance of the preprocessing activities. Also, the representation of their process such as a waterfall flow does not allow interactions related to data preprocessing.

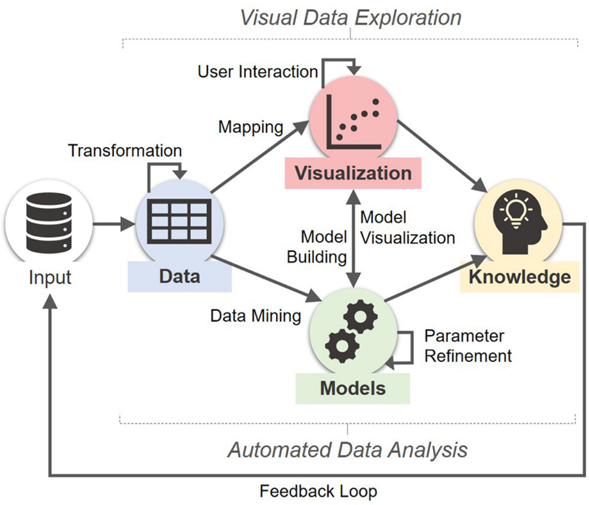


Figure 2. The Visual Analytics process based on Keim et al.[19]. Each node (colored rectangle) corresponds to a different phase, and their transitions are represented through arrows.

As an extension of Keim et al.,[19] Sacha et al.[20] presents a new model for Knowledge Generation (Figure 3) that includes a high-level description of the human work process in the visual analytics integrating this model with different frameworks. Next, other works emerged inspired by these previous works, such as Ribarsky and Fisher[22] addressing the human-machine interaction loop complementary to Sacha et al.[20] and Federico, Wagner et al.[23] explaining the role of explicit knowledge in the analytical reasoning process when proposing a conceptual model for knowledge-assisted visualizations. These three references share the focus on the “Human” side, that is, cognitive science and knowledge generation aspects. Thus, despite Sacha et al.[20] also being one of the works that most describes the “Computer” side, the discussion about the data profiling and preprocessing challenges are still existent.

Although limited to a subarea of VA, we can identify studies that contribute toward our discussion by showing preprocessing activities as part of their VA process description. For instance, Lu et al.[15] and Lu et al.[24] while introducing the Predictive Visual Analytics pipeline, and Sacha et al.[7] during their proposal of an ontology for VA assisted Machine Learning.

Visualization during preprocessing

In the existing literature, we observed few visualization studies concerned with data preparation activities. Also, the use of VA for the preprocessing phase is least reported in general. The same observations are also reported by other authors, for example, Kandel et al.,[4] Sacha et al.,[7] Seipp et al.,[25] Lu et al.,[15] Bernard et al.,[17] and Lu et al.[24]

Some studies in the context of VA and preprocessing can be found, for example, Bernard et al.[17] and Gschwandtner et al.,[18] but they are focusing in time series data and do not provide a comprehensive discussion for preprocessing with different types of data.

Likewise, we can find studies explaining how they are handling preprocessing during a VA process, for example, Krause et al.[26] and Sacha et al.[27] However, these studies are still not entirely dedicated to cover preprocessing problems. Nevertheless, their observation of how shifting the attention from visual analysis to visual preprocessing can improve the analytical processes contributes to our discussion's relevance.

In this context, few relevant works can be cited with a broader coverage in visualization in preprocessing. One of them is the Predictive Interaction framework for interactive systems, developed by Heer et al.,[28] that covers general design considerations for data transformations. As the main discussion, the authors propose that the data analyst can decide the next steps of data transformation by highlighting guidelines of interest in visualizations, instead of specifying details of their data transformations. With that, they expect to avoid a variety of data-centric problems related to the technical challenges of data analysts during programming. Similarly, Wrangler[3] is introduced as a system for interactive data transformations, which includes an interface language to support data transformation with a mixed interface of suggestions and user interaction on visual resources. Both papers provide primordial techniques in the scope of preprocessing, but they are limited to the data transformation activities.

Regarding visual data profiling, von Zernichow and Roman[29] propose an approach to use visual data profiling in tabular data cleaning and transformation processes to improve data quality. As part of their study, they also evaluate the usability of their implemented software prototype, which brings considerations under the usability issues and suggestions for further research, such as exploring visual recommender systems.

One of the most comprehensive proposals about preprocessing is Profiler,[16] an integrated statistical analysis, and visualization tool for assessing data quality issues. Profiler uses data mining methods to support anomaly detection. However, there is still the opportunity to explore different ways to view frequent data issues, for example, missing values in a dense-pixel display.

Visualization of data quality issues

There is comprehensive literature available on how to diagnose and handle data errors, for example, Kim et al,[1] Wickham,[5] Rahm and Do,[10] Chandola et al.,[30] and Wang et al.[31] Among the different types of data quality issues, the missing data are one of the most frequently referenced.[4,6,8]

Templ et al.[32] criticize that no matter how well the classification mechanism for missing data has been planned, they still have limitations such as the difficulty to accurately identify the cause of the value being missing while working with multivariate data. Subsequently, they argue for the importance of visualization to solve the related questions, and they introduce Visualization and Imputation of Missing Values (VIM). In an empirical study to evaluate the best design for interpretation of graphs with missing data, Eaton et al.[33] observe that data interpretation is negatively impacted when there is a poor indication of the missing values. Additionally, more recent studies such as Sjöbergh and Tanaka[34] and Song and Szafir[35] endorse the importance of developing different ways of visualizing missing values as an attempt to avoid misleading interpretations resulting from the way the visualization procedure was developed. Similarly, McNutt et al.[36] claim that dirty data or bad user choices can cause errors in all stages of the VA process, and a superficial visualization without a closer re-examination can lead to misleading or unwarranted conclusions from data (what they call visualization mirage).

What the practitioners say

In addition to the research related to visualization techniques and the VA process, it is also important to understand the current practice of enterprise professionals with data preprocessing and how visualization supports this process. However, few works can be found sharing the experiences of the practitioners in the scope of data analysis and visualization, for example, Batch and Elmqvist,[37] Kandogan et al.,[12] Kandel et al.,[38] and Milani et al.[8] At the same time, other interview studies are focusing on interactive data cleaning, such as Krishnan et al.[6] When combined, these works bring light on practitioners' reality on different perspectives, supporting a broader view of the practice and the current needs.

In the most recent of these works, Milani et al.,[8] we interviewed thirteen enterprise data analysts and compiled a list of 10 insights for new visualizations in preprocessing scope. We compared our findings to the other interview studies to compile the final list, which brings confidence that this list of insights can be used as a consolidated set of requirements based on what the practitioners report. Moreover, these insights improved the reliability of our findings and provided background, helping in the definition of the guidelines presented in the next section.

Review and comparison

To better organize our discussion on the related work and to facilitate the comparison with the scope of our work, we defined six items to guide this effort. The

Table 1. Is the work presenting details on the following items? **(1)** Process or model or workflow or pipeline; **(2)** Preprocessing is considered an explicit phase on the process; **(3)** Preprocessing activities and strategies; **(4)** Preprocessing impacts in the next phases; **(5)** Specifications or guidelines for solutions in preprocessing; **(6)** Visualizations for data quality issue.

Section	Related work	(1)	(2)	(3)	(4)	(5)	(6)
Visual Analytics Process	Keim et al.[19]	✓					
	Sacha et al.[20]	✓					
	Ribarsky and Fisher[22]	✓					
	Federico, Wagner et al.[23]	✓					
	Lu et al.[15]	✓	✓	✓			
	Lu et al.[24]	✓	✓	✓			
	Sacha et al.[7]	✓		✓	✓		
Visualization during preprocessing	Heer et al.[28]					✓	
	Kandel et al.[3]			✓			✓
	von Zernichow and Roman[29]			✓		✓	✓
	Kandel et al.[16]			✓		✓	✓
Visualization of data quality issues	Templ et al.[32]			✓			✓
	Eaton et al.[33]						✓
	Sjöbergh and Tanaka[34]						✓
	Song and Szafir[35]			✓		✓	✓
	McNutt et al.[36]	✓	✓	✓	✓		✓
What the practitioners say	Milani et al.[8]			✓	✓	✓	✓
	Frequency *(... of 17)*	8	3	10	3	5	9

results are summarized in Table 1, and further comments for each item are provided. We did not add all related work to the table, but only those we considered closer or more relevant to our discussion.

Regarding **Item 1** (*Process or model or workflow or pipeline*) and **Item 2** (*Preprocessing is considered an explicit phase on the process*), we evaluated if the related work addresses our central problem regarding the indication of preprocessing as an equally important phase in the process representation. We can observe that the studies in the scope of Predictive Visual Analytics (Lu et al.[15,24]) present preprocessing formally as a phase during their pipeline and discussion. However, they address data mining problems in the scope of Predictive tasks,[2] which does not cover the Descriptive tasks as in the initial VA processes.[19,20,22,23] Also, even though Sacha et al.[7] show preprocessing (as *Prepare-Data*) in evidence on their VIS4ML ontology, preprocessing is classified as a process and not an entity such as Data or Model phases (as in Figure 2).

Next, **Item 3** (*Preprocessing activities and strategies*) is related to the discussion of the activities and strategies covered as part of the preprocessing phase. We did not expect a complete taxonomy under discussion. On the contrary, we recognized if the related work was at least considering the existence of the complexity in selecting different strategies. The Predictive Visual Analytics's related work[15,24] and Sacha et al.[7] contribute with a high-level discussion on the topic. Other few studies in visualization during preprocessing[3,16,29] cover that aspect as well. Finally, even if focused on only one data issue, Templ et al.[32] and Song and Szafir[35] also mention the complexity of handling missing values.

Complementing the previous, **Item 4** (*Preprocessing impacts in the next phases*) considers the effects that the decisions made during the preprocessing may cause in later stages, similar to the discussion promoted by Crone et al.[39] Even though the related work selected as part of Subsections *Visualization during preprocessing* and *Visualization of data quality issues* recognize the importance of preprocessing and its impacts on the overall process, most of them are concerned about how to enhance the capabilities of the data analysts while performing the cleaning and transformation tasks. Therefore, only Sacha et al.,[7] McNutt et al.,[36] and Milani et al.[8] mention this topic, at least in an explicit manner. To illustrate, Sacha et al.[7] present examples of pathways in the Machine Learning workflow, and during the Evaluate-Model process, they explain that a model-developer may wish to make some changes to what was set in the previous steps, which includes data preparation tasks. In Milani et al.,[8] there is a discussion calling attention to the fact that multiple interactions among preprocessing and the other stages should be expected in the data analysis process.

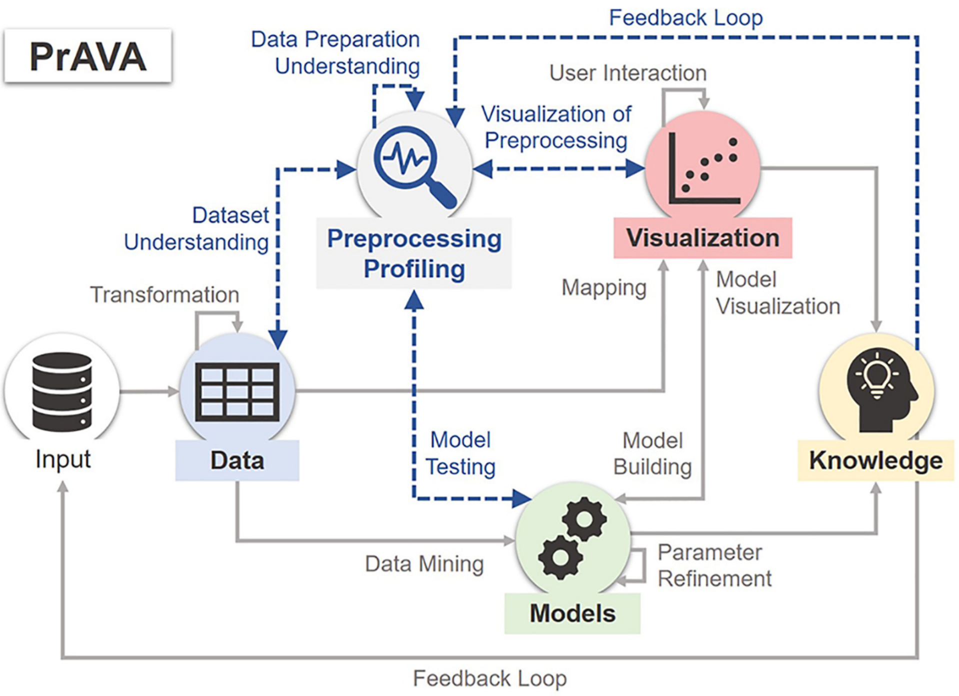


Figure 4. The Preprocessing Profiling Approach for Visual Analytics (PrAVA) is an extension of the VA process proposed by Keim et al.[19]. We added the Preprocessing Profiling phase and new transition options: Dataset Understanding, Data Preparation Understanding, Visualization of Preprocessing, Model Testing, and another Feedback Loop. The new objects are represented in blue color for the text font and dashed lines.

While evaluating **Item 5** (*Specifications or guidelines for solutions in preprocessing*), we were looking for detailed descriptions in support to design new visualizations or systems for any preprocessing activity. Only Heer et al.,[28] Song and Szafir,[35] and Milani et al.[8] address this item. The majority of the other related work that could contribute to this item was designed as Systems. However, Kandel et al.[16] and von Zernichow and Roman[29] were added to this list because they provide valuable insights during their system architecture and usability suggestions.

Multiple works[3,8,16,29,32–36] cover the content of **Item 6** (*Visualizations for data quality issue understanding*). We acknowledge that a complementary investigation is required to include different data quality issues. Still, we are confident the currently selected studies should support us with an overall understanding of the efforts developed in this scope.

In conclusion, besides the relevant contributions of these works, we can still observe opportunities to be discussed. From that, the following items receive less attention than the others:

- Preprocessing as an equally important phase in the VA process.
- Alternative visualizations to cover the same data quality issue by different perspectives.
- Visualizations to support the evaluation of the preprocessing impacts in further phases.
- List of guidelines to support novel visualizations in the context of preprocessing in a data analysis process.

To continue this discussion and support filling these gaps, we are proposing the Preprocessing Profiling Approach for Visual Analytics, which is described in the next sections.

Preprocessing profiling approach

In this section, we present the Preprocessing Profiling Approach for Visual Analytics (PrAVA), illustrated in Figure 4. First, we outline the nine guidelines that we identified as important to be observed while planning new solutions in compliance with our proposed approach and considering preprocessing an equally important phase in the VA workflow. Second, we explain the PrAVA process and its relation to the guidelines.

Guidelines

We identify nine guidelines for consolidating preprocessing in the VA process, composing the foundation for the proposed PrAVA extension. These guidelines

Table 2. List of nine guidelines to be considered as part of the Preprocessing Profiling Approach for Visual Analytics (PrAVA). For each guideline, we describe their meaning, motivation, and some examples of implementations in the context of VA or Visualization.

Guideline	Meaning	Motivation	Examples of implementation
G1 Unified	Integration with the most used tools for data analysis.	To build an uninterrupted work environment, preventing the data analysts from losing the context under investigation while alternating among several different tools. Also, as an approach to simplify and save time during the analysis activities.	(a) Dataset understanding: Pandas Profiling[40] *(for Python)*. (b) Missing values: VIM[32] *(for R)*; Missingno[41] *(for Python)*. (c) Model validation: Yellowbrick[42] *(for Python)*.
G2 Large Scale	Ability to work with scenarios dealing with huge volumes of data.	To attend the crescent demand for Big Data, evaluate how to produce partial results while the data are being processed. Hence, data analysts can visualize huge volumes of data in a continuously flow.	(a) A training dynamics analysis module that samples the time series to preserve outliers and reduce visual clutter caused by a large amount of time series data – Liu et al.[43]
G3 Metadata	Ability to generate informative summaries of the preprocessing activities.	The data computation of other guidelines, for example, **G4** and **G5**, should be the source of this guideline, which should result in a critical output of the Preprocessing Profiling process. Also, this metadata can be used as input for new visualizations of the dataset under analysis, and generally for documentation purposes.	(a) Visualization of metadata by combining analysis of (time series) clusters and additional metadata attributes – Sacha et al.[27] *Note: this work is not discussing the generation of metadata based on preprocessing activities outputs, but how to use metadata.*
G4 Data mining	Use of data mining methods to support preprocessing activities.	Data quality assessment can benefit from the use of Machine Learning algorithms, for example, the identification of data errors and recommendations on data transformation. Additionally, supporting the validation of the preprocessing strategies and model testing.	(a) Identification of quality issues in the training data of Convolutional Neural Networks in image classification – Alsallakh et al.[44] (b) Dataset understanding, hidden states in Recurrent Neural Networks – Strobelt et al.[45]
G5 Statistics	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 for decision making on data transformation demands, not only to fix data issues but to better integrate with the planned model. Later, this information should be combined with visualization techniques.	(a) Visual diagnostics of binary classifiers using instance-level explanations (e.g. aggregate statistics to see how data distributes across correct or incorrect decisions) – Krause et al.[46]
G6 Comparison	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 **G4**, this guideline can support the evaluation of the model based on different preprocessing strategies.	(a) Classification analysis and selecting training data: TreePOD – Mühlbacher et al.[47] (b) Performance Analysis (Model Classification): Squares – Ren et al.[48]
G7 Recommendation	Use of recommendation systems to propose visualizations.	Visualization techniques can be proposed according to the type and volume of data under investigation. Also, taking into consideration the particularities of the data mining scope or data quality issues.	(a) Exploratory visual data analysis: Voyager – Wongsuphasawat et al.[49] (b) List of requirements and design considerations for a visualization recommendation system – Vartak et al.[50]
G8 Template	Ability to generate automatically initial visualizations or basic templates.	This guideline refers to a solution that generates initial visualizations or template options based on the data under analysis. This guideline, when combined with **G7**, should avoid some inappropriate uses.	(a) Commercial solutions such as Tableau Prep[51] and Trifacta[52] present visualizations based on data under analysis, but they do not allow customization. (b) Thus, as closer example is Tableau[53]
G9 Interaction	Use of visualization interaction techniques to support flexible data exploration.	Interaction is fundamental to data visualization, and this should allow the data analysts to perform flexible data manipulation instead of static reports.	(a) Iterative cluster refinement: SOMFlow – Sacha et al.[27] (b) Visual query tool to interactively create cohort populations with temporal constraints: COQUITO – Krause et al.[26]

were identified based on the current relevant literature (*Related Work* section), on the research directions in data wrangling raised by Kandel et al.,[4] Krishnan et al.,[6] and in our previous study that we interviewed enterprise data analysts.[8] In Table 2, we present a description of the meaning and motivation for each guideline: **G1** Unified, **G2** Large Scale, **G3** Metadata, **G4** Data Mining, **G5** Statistics, **G6** Comparison, **G7** Recommendation, **G8** Template, and **G9** Interaction.

We also indicate additional work or software solutions that we consider related to each guideline. In other words, that can illustrate its possible implementations. It is pertinent to note that some of the suggested references may cover more than one guideline, or they may not fully cover even one guideline. Moreover, some of them do not have the preprocessing as an ultimate purpose. However, in their presentation, we can observe how they use the VA or Visualization during preprocessing tasks.

The structured list of guidelines aims to guide the design of new solutions in adherence to the PrAVA. At the same time, the insights gained during the examination of these guidelines supported us in devising the PrAVA process, which is explained in the next subsection.

Process

PrAVA is formalized as an extension of the VA process (see Figure 2), in which we include a new phase called Preprocessing Profiling, and new possible transitions among the phases. An overview of the PrAVA process is shown in Figure 4. Even though we recognize the importance of human cognitive activities in the VA process (see Figure 3), we decided to continue using Keim et al.[19] representation aiming for simplicity to illustrate the VA process; therefore, this decision allowed us to focus on the Preprocessing Profiling transitions.

By adding Preprocessing Profiling as a phase, we put activities such as the data profiling and the evaluation of preprocessing strategies before Model Building in the critical path, that is, as an equally important phase. However, preprocessing activities planned in the original Data phase as part of the Transformation transition (Data ↔ Data) can still occur since, for example, the dataset input may require data cleaning and normalization before proceeding with any analysis. Also, the other four original phases and their transitions remain the same. Next, we focus on explaining only the new transitions. Furthermore, we are indicating how the guidelines presented in Table 2 can be associated with this process.

The new transition **Dataset Understanding** (Data ↔ Preprocessing Profiling) intends to explore the dataset, its data types, value distribution, and other descriptive statistics (**G5**) that will be important to create the data profiling, that is, metadata (**G3**). Consequently, this process supports the data analyst's decisions while they are progressing to further activities.

Data Preparation Understanding (Preprocessing Profiling ↔ Preprocessing Profiling) intends to allow the creation of metadata for the data preparation strategies developed during the preprocessing (**G3**). Additionally, with the **Visualization of Preprocessing** (Preprocessing Profiling ↔ Visualization), the data analyst should be able to explore these different data preparation strategies with the support of visualization techniques. These visualization techniques can be recommended based on the data under analysis (**G7**), or even initial visualizations as templates can be presented to support this activity (**G8**).

Another new transition is **Model Testing** (Preprocessing Profiling ↔ Models), which considers the validation of the model during the Preprocessing Profiling phase. With the support of data mining methods (**G4**), it is an opportunity to evaluate and compare the impacts of the chosen preprocessing strategies that can be used as input for Model Building transition (**G6**).

All the transitions leaving the Preprocessing Profiling phase have a way back on the same connection (i.e. the arrows in Figure 4). Different from the original VA process (see Figure 2), which can be read as one-way direction, such as a waterfall approach, PrAVA considers the possibility of multiple interactions between two phases during the same process. Thus, we also added a new **Feedback Loop** (Knowledge → Preprocessing Profiling).

However, the model proposed by Sacha et al.[20] (see Figure 3) better describes the different loops in this scope of knowledge generation and should be used as a reference for the subject. In summary, they define three different usage loops: (1) the exploratory loop, where finds are discovered; (2) the verification loop, where insights are generated by interpreting the findings; and finally (3) the knowledge generation loop, where insights are converted into verified hypotheses and data is transformed into knowledge. Our proposed **Feedback Loop** stresses that after deriving knowledge from the process, the user can choose to return to the Data phase or go to the Preprocessing Profiling phase using the acquired knowledge to do a new data preparation. In some cases, it is better to go back to the Preprocessing Profiling phase since the produced knowledge may be influenced (positively or negatively) by the employed techniques. Subsection *Cervical cancer dataset* exemplifies how an imputation decision affects the analysis at hand. This approach is similar to what we can see described in the data mining literature. For instance, the Cross Industry Process for Data Mining (CRISP-DM)[54] shows explicit phases as "Data Understanding" and "Data Preparation" in interactively process with its "Modeling" phase.

Big Data scenarios are the concern behind **G2**, that is, huge number of records and high-dimensional data. In these cases, during a flow as Data → Preprocessing Profiling, we can consider an alternative such as the Progressive Paradigm[55,56] to produce partial results while the entire dataset is still being processed. Also, for a flow as Preprocessing Profiling → Visualization, aggregation techniques[57] could be used to support generating visual representations more efficiently.

In reference to **G1** and **G9**, they should be considered as part of the entire process. The combination of these features should attend an urgent demand mentioned by Heer and Kandel[58] (p.53) "interactive tools for data analysis should make technically proficient users more productive while also empowering users with limited programming skills." Moreover, despite **G9** may seem evident to visualization practitioners, it requires significant efforts to design and implement bolder interactions, according to Dimara and Perin,[59] and therefore, deserves attention.

The VA process described in PrAVA includes cases in which data adjustments are identified in several phases of the data analysis process. These are not limited to the first time data are selected and transformed. We also advocate the advantage of using visualization techniques during the preprocessing, and not only to generate the final visualizations. Ultimately, our proposal with PrAVA considers the Preprocessing Profiling as a prominent phase, which deserves to have its transitions explicitly extended in the VA process.

Among our rationale for this novel approach, we can indicate a couple of reasons. First, even though Keim et al.[19] covered Data activities, as previously explained, it was not covering all the preprocessing activities as we are proposing in this work. We also do not consider Preprocessing Profiling a sub-phase of Data because we understand that the complexity related to data preparation has evolved over the years. These processes have been overlooked by the visualization research community as reported in our *Related Work* section and other references such as Crisan and Munzner,[60] which corroborates with this need for a revisited approach. Second, similar to what Munzner[61] explains during their nested model for visualization design and validation, the intellectual value of separating in explicit stages is that we can separately analyze whether each phase has been addressed correctly, no matter what order they were undertaken. Furthermore, the author conjectures that many experienced practitioners (visualization designers) carried out methodologies, albeit implicitly or subconsciously. Conversely, newcomers do not have that tacit knowledge, so we consider conceptual models fundamental to this audience. Moreover, even though these experienced practitioners have these internal processes that they can implicitly follow, as indicated by Munzner[61] (p.922), "sometimes designers cut corners by making assumptions rather than engaging with any target users." Thus, our proposed approach aims to make these subconscious activities more explicit to provide a model that can be used to help guide the VA process itself. To conclude, PrAVA should enable the practitioners (data analysts or visualization designers) to increase their ownership of the data under analysis, master the impacts of preprocessing activities to the model building, and contributing to more trustworthy knowledge discovery in the VA process.

Usage scenario

In this section we present a usage scenario with PrAVA. We implemented a prototype solution, first, to assist with this usage scenario, and later, with other possible applications of PrAVA. This solution is described in Subsection *Prototype*, and the usage scenario is presented in Subsection *Tim and the Iris Dataset*.

Prototype

Since our primary goal is to describe a conceptual VA process (PrAVA), and not a system, we introduce in this subsection just the information that we consider relevant to the prototype's overall understanding as it is referenced in the next subsections. The developed prototype solution generates two dynamic reports: *Data Profiling* (https://github.com/DAVINTLAB/pandas-profiling) and *Preprocessing Profiling* (https://github.com/DAVINTLAB/preprocessing-profiling).

The *Data Profiling* report supports the dataset understanding. This report was developed as an extension of Pandas-profiling.[40] The main sections are identified as *Overview*, information about the dataset such as the total number of rows and columns, variable types, and Warnings; *Variables*, descriptive statistics and visual representations to support a detailed view of each variable (or attribute) of the dataset; *Missing Values*, visualizations to help the identification of particular patterns related to the missing values occurrences; and *Correlations*, visual heatmap to present the values of the correlation coefficient of all pairs of variables.

The *Preprocessing Profiling* report supports the evaluation of data transformation impacts on the model.

For this first version, we considered one data mining problem (Classification), one data issue to perform the data transformations (Missing Values), and one type of dataset (tabular data). Overall, the report performs the following tasks (a) reads an informed dataset and splits the data into training and testing; (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. Regarding task (b), five different strategies of data imputation are considered. One strategy removes all the rows with at least one missing value, and this data imputation strategy is named *Baseline (no missing)*. Another strategy replaces all missing values by zero, named *Constant(=zero)*. A third and fourth replace missing values by mean and median values computed, respectively, based on all records on the same column. The fifth strategy replaces missing values by the most frequent value on the column.

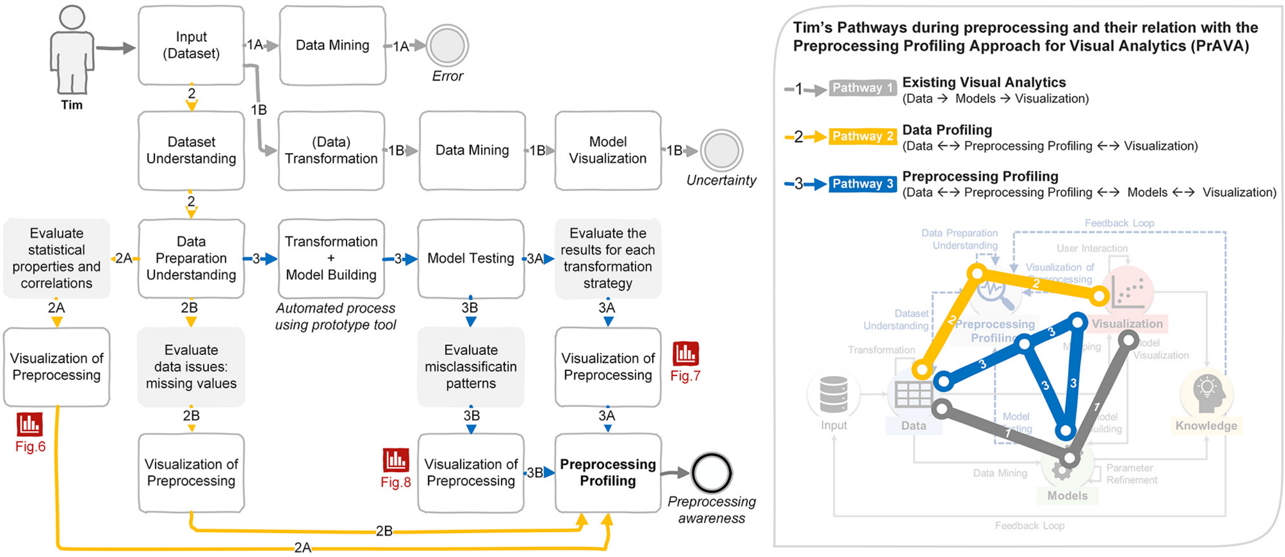
As a final observation, the developed prototype is functional, but it cannot be considered an end-to-end VA System. Additionally, not all the guidelines were implemented. **G2** (Large Scale) and **G7** (Recommendation) were out of scope since the beginning of the prototype project, due to their complexity to be implemented and for not being our primary focus with this paper.

Tim and the Iris dataset

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.

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 the 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)[62] from three different species of Iris, namely, Iris Setosa, Iris Virginica, and Iris 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; as a result, his dataset has data quality problems, that is, the dataset contains outliers and missing values.

Tim is familiar with the Python programming development environment. To begin, he tries to run a classification model using his dataset without any data transformation. However, he could not move forward since an error message is returned informing him the classification algorithm cannot proceed due to missing



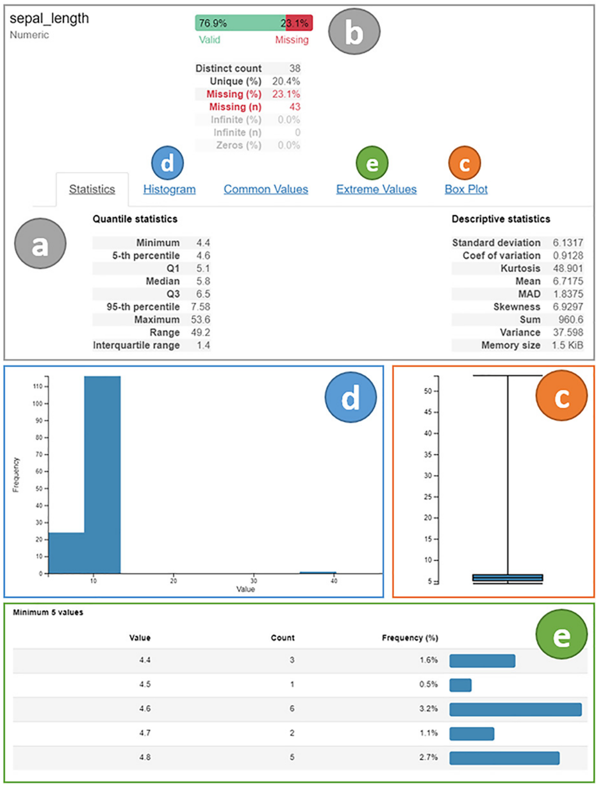
Tim
Input (Dataset)
1A
Data Mining
1A
Error
2
1B
Dataset Understanding
(Data) Transformation
1B
Data Mining
1B
Model Visualization
1B
Uncertainty
2
Evaluate statistical properties and correlations
2A
Data Preparation Understanding
3
Transformation + Model Building
3
Model Testing
3A
Evaluate the results for each transformation strategy
2A
2B
Automated process using prototype tool
3B
3A
Visualization of Preprocessing
Evaluate data issues: missing values
Evaluate misclassificatin patterns
Visualization of Preprocessing
Fig.7
Fig.6
2B
3B
3A
Visualization of Preprocessing
Fig.8
Visualization of Preprocessing
3B
Preprocessing Profiling
Preprocessing awareness
2B
2A
Tim's Pathways during preprocessing and their relation with the Preprocessing Profiling Approach for Visual Analytics (PrAVA)
1 Pathway 1 Existing Visual Analytics (Data → Models → Visualization)
2 Pathway 2 Data Profiling (Data <-> Preprocessing Profiling <-> Visualization)
3 Pathway 3 Preprocessing Profiling (Data <-> Preprocessing Profiling <-> Models <-> Visualization)
Feedback Loop
Data Preparation Understanding
User Interaction
Visualization of Preprocessing
Dataset Understanding
Preprocessing Profiling
Visualization
Transformation
Mapping
Model Visualization
Input
Data
Model Testing
Model Building
Knowledge
Data Mining
Parameter Refinement
Models
Feedback Loop
1
2
3

Figure 5. Usage scenario – The pathways took by Tim: **1** (connection lines in gray) considering the existing VA process (Figure 2), and so not focusing in preprocessing activities, except by elementary data transformation; **2** (connection lines in yellow) focusing on the dataset understanding; **3** (connection lines in blue) concentrate on the impacts of preprocessing strategies. On the right of this figure, each pathway is related to the PrAVA process (Figure 4). We describe the paths as sequential steps to facilitate the usage scenario explanation, but the idea behind PrAVA is to allow multiple backward and forward between the phases.

values in the dataset. This attempt is shown in Figure 5 as **Pathway 1A**. Therefore, he transforms the missing values by replacing all of them 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 (Figure 5 – **Pathway 1B**), Tim decides to use PrAVA to guide his analysis to perform his activities. First, he explores his dataset for a better understanding (Subsection *Dataset Profiling*). Next, he evaluates the impacts of his decisions on data transformation to the model building (Subsection *Preprocessing Profiling*).

Dataset profiling. Tim starts by running descriptive statistics using Python. However, many lines of code and outputs with plain text would be required to generate all the information he wants. Consequently, he decides to use PrAVA's prototype integrated into his development environment to create the first report for his analysis. With *Data Profiling* report information, he got an overview regarding the number of records, the dataset size, and variable types distribution. By reading the messages under *Warnings* subsection of the *Overview*, and by viewing the *Correlations* section of the report, Tim realizes the petal_length and petal_width columns are highly correlated with each other. Even though he had previously generated the covariance and correlation matrix, when he was executing his initial set of code, he considers it was challenging to observe the relation between two variables just by looking at the output with plain text.

Tim decides to explore each variable of his dataset (still part of Figure 5 – **Pathway 2A**). Figure 6 shows an example of what he sees for the sepal_length. Based on that, Tim confirms the value distribution and the presence of data issues. Additionally, he explores the *Missing Values* section of the *Data Profiling* report (Figure 5 – **Pathway 2B**), and despite the observation of the total amount of 10% missing values (entire dataset), no significant pattern in relation to these occurrences is noted, for example, he did not identify a column that has been highlighted with missing data. Up to this point, Tim completes the activities related to understanding the data (Figure 5 - **Pathway 2**).



sepal_length
Numeric

Distinct count	38
Unique (%)	20.4%
Missing (%)	23.1%
Missing (n)	43
Infinite (%)	0.0%
Infinite (n)	0
Zeros (%)	0.0%

Statistics | Histogram | Common Values | Extreme Values | Box Plot

Quantile statistics

Minimum	4.4
5-th percentile	4.6
Q1	5.1
Median	5.8
Q3	6.5
95-th percentile	7.58
Maximum	53.6
Range	49.2
Interquartile range	1.4

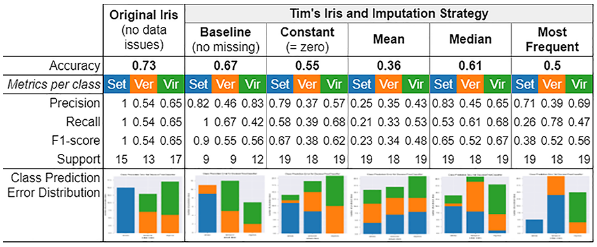
Descriptive statistics

Standard deviation	6.1317
Coef of variation	0.9128
Kurtosis	48.901
Mean	6.7175
MAD	1.8375
Skewness	6.9297
Sum	960.6
Variance	37.598
Memory size	1.5 KiB

Minimum 5 values

Value	Count	Frequency (%)
4.4	3	1.6%
4.5	1	0.5%
4.6	6	3.2%
4.7	2	1.1%
4.8	5	2.7%

Figure 6. *Data Profiling* report. *Variables* section for sepal_length: (a) statistical measures, (b) horizontal barplot with valid and missing values distribution, (c) boxplot, (d) histogram, and (e) list of extreme values.



	Original Iris (no data issues)			Tim's Iris and Imputation Strategy														
				Baseline (no missing)			Constant (= zero)			Mean			Median			Most Frequent		
Accuracy	0.73			0.67			0.55			0.36			0.61			0.5		
Metrics per class	Set	Ver	Vir	Set	Ver	Vir	Set	Ver	Vir	Set	Ver	Vir	Set	Ver	Vir	Set	Ver	Vir
Precision	1	0.54	0.65	0.82	0.46	0.83	0.79	0.37	0.57	0.25	0.35	0.43	0.83	0.45	0.65	0.71	0.39	0.69
Recall	1	0.54	0.65	1	0.67	0.42	0.58	0.39	0.68	0.21	0.33	0.53	0.53	0.61	0.68	0.26	0.78	0.47
F1-score	1	0.54	0.65	0.9	0.55	0.56	0.67	0.38	0.62	0.23	0.34	0.48	0.65	0.52	0.67	0.38	0.52	0.56
Support	15	13	17	9	9	12	19	18	19	19	18	19	19	18	19	19	18	19
Class Prediction Error Distribution																		

Figure 7. Classification results for one round of testing using the attributes sepal_length and sepal_width and different preprocessing strategies. First column refers to Original Iris dataset (without data issues). From second to sixth column refer to Tim's Iris dataset and the corresponding imputation strategies performed. The classes are identified as "Set" in blue for Iris Setosa, as "Ver" in orange for Iris Versicolor, and "Vir" in green for Iris Virginica. In the last row, the Barplots also follow this order (Set,Ver,Vir).

Preprocessing profiling. Tim moves to the analysis of the impacts of the preprocessing strategies on his classification problem after the completion of his activities in understanding the data. Tim informs his dataset as input to the *Preprocessing profiling* report. Since all the data transformation and model building are done automatically, Tim takes advantage of the time saved, and he runs multiple rounds (of training and testing)

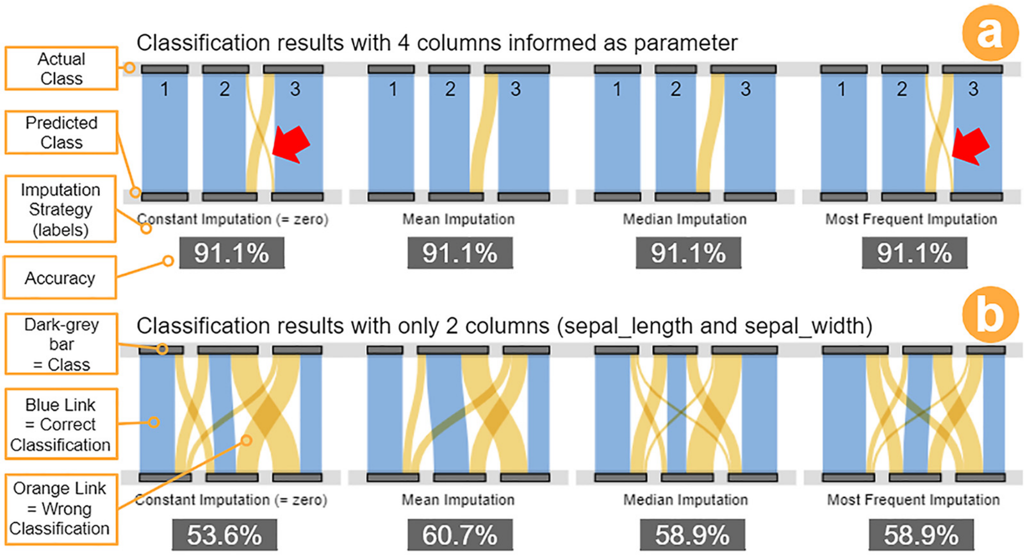


Figure 8. Preprocessing profiling report. Classification results for different missing values imputation strategies: (a) four columns informed into the classification model, and (b) only the two columns related to Sepal attribute were informed. The visualization Flow of Classes is used in both representations (inspired by Sankey diagrams).

to evaluate the results of classification. Figure 7 shows an overview of the results for one round where he used only the variables related to sepal attributes.

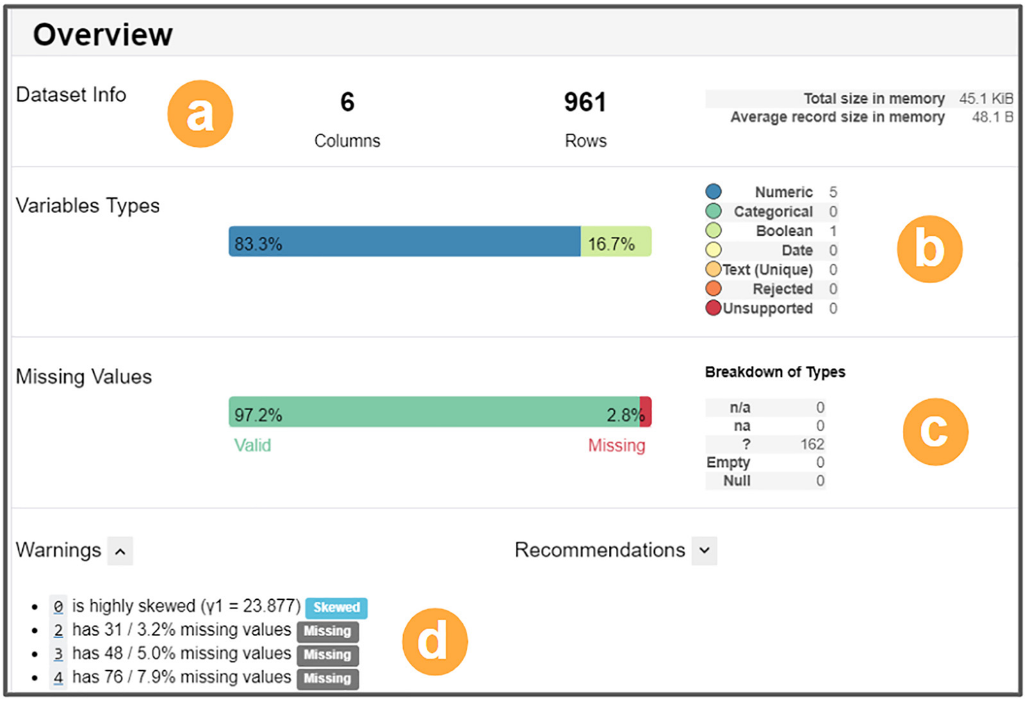
Although the classification results varied in each round, Tim is still able to notice differences among the imputation strategies for all rounds performed. For example, the class of Iris Setosa was initially clear to classify (Figure 7, first column, class in blue). However, with the presence of data issues and the need to perform imputation strategies, the classification results are negatively impacted. Tim also observes a significant variation on the accuracy metric for the *Mean* imputation strategy (Figure 7, fourth column) compared to the others. With that, it is clear to him that he needs to identify outliers, for example, using visualizations such as Boxplot (Figure 6-c), 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 Figure 5 as **Pathway 3A**.

Furthermore, while comparing the *Flow of Classes* visualization for different rounds, he can observe new situations that were not possible with the prior perspectives. 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 runs a round using the four variables (Figure 8-a), four imputation strategies result in the same accuracy (91.1%). However, he can notice an additional flow of classes from actual class 2 (Versicolor) to predicted class 3 (Virginica) during *Constant* and *Most Frequent* imputations. While for *Mean* and *Median* strategies, the misclassification occurs only from actual class 3 (Virginica) to predicted class 2 (Versicolor). Likewise, when observing the results for another round, which considered only two variables (Figure 8-b), he can notice more variations among the possible combination flows.

Under these circumstances, he considers it essential to have different views for the same classification results, mainly when using a dataset with data quality issues. In conclusion, Tim takes these insights as reinforcement of the importance of exploring data transformation strategies before moving to further phases in the VA process or any data mining process. This process is shown in Figure 5 as **Pathway 3B**, which, when combined with **Pathway 2**, promotes awareness of Preprocessing profiling and is in line with to what has been reported in the literature as a promising approach to understanding data quality issues (e.g. Gleicher et al.[63]).

Applications

To showcase the possible advantages of using PrAVA, we created two application scenarios to describe the efforts made to understand datasets with tabular data. We looked into online repositories for open datasets that could be used in the scope of classification problems, and we selected two datasets that we did not have any previous knowledge of. In Subsection *Mammographic mass dataset*, we are using the developed prototype (described in Subsection *Prototype*) to explore one dataset, while in Subsection *Cervical cancer dataset*, we are using commercial software to explore a second dataset. To conclude, in Subsection *Review of*



Overview

Dataset Info

6	961
Columns	Rows

Total size in memory	45.1 KiB
Average record size in memory	48.1 B

Variables Types

83.3% 16.7%

●	Numeric	5
●	Categorical	0
○	Boolean	1
○	Date	0
○	Text (Unique)	0
●	Rejected	0
●	Unsupported	0

Missing Values

97.2% 2.8%

Valid Missing

Breakdown of Types

n/a	0
na	0
?	162
Empty	0
Null	0

Warnings ^ Recommendations ˅

- 0 is highly skewed (γ1 = 23.877) Skewed
- 2 has 31 / 3.2% missing values Missing
- 3 has 48 / 5.0% missing values Missing
- 4 has 76 / 7.9% missing values Missing

Figure 9. *Data Profiling* Report. Overview of the Mammographic Masses dataset: (a) dataset information with columns, rows, and size, (b) variable types, (c) missing values and breakdown of types, and (d) list of warnings.

scenarios, we present a discussion of how preprocessing is being perfomed by other studies using the same datasets, and we relate the guidelines (described in Subsection *Guidelines*) to the tools used during our applications.

Mammographic mass dataset

We selected a dataset from the UCI Machine Learning Repository related to the breast cancer screening method.[64] This dataset contains the discrimination of benign and malignant mammographic masses based on BI-RADS variables and the patient's age. We decided to start by running our prototype to collect information about the dataset for understanding it.

First, while reading the information available on the *Data Profiling* report, we could confirm the number of columns and rows (Figure 9-a), as well as the distribution of variable types (Figure 9-b), predominantly numeric. We could observe the presence of missing values and the information on which character was used in the original dataset to represent the not informed values (Figure 9-c). Also, in the *Warnings* (Figure 9-d), we could confirm which were the columns with missing values, and a highlight regarding the highly skewed distribution for one column. The original downloaded dataset did not contain headers, so the columns appear named as numbers in this report.

We explored the *Variables* section of the *Data Profiling* report. Consequently, we confirmed that the first variable, column 0 (*BI-RADS*), presented high positive Skewness. Also, we noticed a possible outlier value (55.0). Next, we continued the dataset understanding by evaluating the *Missing Values* section. For column 4 (*Margin*), we could observe the higher percentage of missing values (7.9%), as initially listed in the *Warnings*.

Additionally, we explored the *Correlations* section to evaluate the relationship between each pair of variables with a visualization of the Spearman's rank correlation coefficient. Based on that, we saw a strong connection between columns 2 (*Shape*) and 3 (*Margin*). We considered this useful in case we needed to remove columns to avoid potential bias in the classification.

As a final step, we consulted the documentation available for the dataset to confirm some of our findings and assumptions. For the *BI-RADS* variable, the value identified as a potential outlier, in fact, could be considered bad data since the expected values were ranging from 1 to 5. We also confirmed that column 5 (*Severity*) contains the class of each instance, this was the only variable without missing values.

We completed the initial understanding of the dataset, and we decided to move to the evaluation of the missing value imputation strategies. We used the entire original dataset, except column 0 (*BI-RADS*), and we

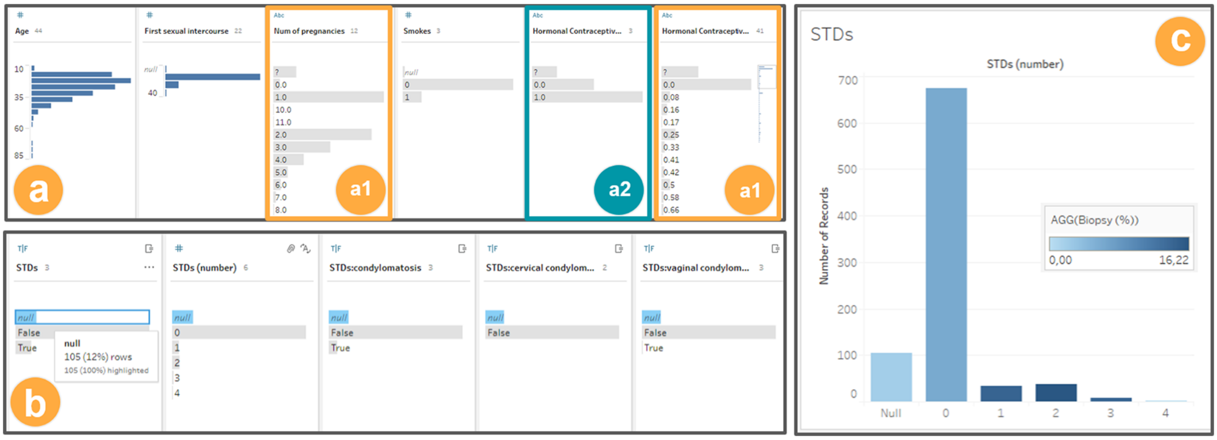


Figure 10. (a) and (b) Exploring the dataset with Tableau prep builder, and (c) histogram generated with Tableau. Both are using the cervical cancer dataset.

ran multiple comparison rounds using the *Preprocessing Profile* report. For all rounds performed, we could observe some variation in the classification results. The maximum variation in accuracy noted was 6.4% between *Baseline (no missing)* and *Mean* imputation strategies. We want to note that rather than evaluating the better imputation strategy performance, our concern remained in observing if the visual resources developed helped to evaluate any possible impacts on the different cleaning or transformation strategies.

Through this scenario, we show some capabilities of using PrAVA, mainly during the data understanding of a new dataset, facilitated when accessing summarized information at a glance, and details on demand. Within minutes, we acquired an overview of the dataset. Furthermore, PrAVA effectively supported the comparison of the results for the different preprocessing strategies, not only because *Preprocessing Profile* report automated part of the work, but primarily because this set of activities performed increased the awareness of the preprocessing impacts. Finally, this approach brought confidence to move forward with the model building after knowing the possible influence of the preprocessing decisions in the final solution.

Cervical cancer dataset

In this second application, we describe the efforts made to understand the cervical cancer dataset that has been acquired from the UCI Machine Learning Repository.[65] We want to know, based on the dataset, which conditions suggest a higher probability of a patient having cervical cancer. To help in the task, we use Tableau,[53] Tableau Prep Builder,[51] and Python programming. Note that when we perform an action that represents a new transition introduced by PrAVA (i.e. the blue dashed lines in Figure 4), we highlight in parentheses the transition that was made.

We decide to load the dataset in Tableau Prep Builder, which should allow us to analyze the missing values and find other issues to address the simpler ones quickly. The visualizations provided by Tableau Prep Builder (Figure 10-a) show the distinct values of every column and, for each value, the number of rows with the same value. Immediately, we can notice that not all variable types were inferred correctly (**Visualization → Preprocessing Profiling**). There are numeric columns shown as string (Figure 10-a1) and boolean columns shown as numeric (Figure 10-a2). Also, in the original dataset, the missing values are represented by the string "?" instead of "null" (**Preprocessing Profiling → Data**). Thus, we replace the string "?" with "null" and change the variable types to the right ones.

After correcting simple problems, we evaluate strategies to deal with the missing values. We examine again the visualizations provided by Tableau Prep Builder, as shown in (Figure 10-b). When a value is selected, for example, "null," the same value is highlighted in the other columns. This helps us to observe that there is missing value correlation between several columns (**Visualization → Knowledge → Preprocessing Profiling**).

Wondering what the meaning of the discovered correlation might be, we transition from Tableau Prep Builder to Tableau and create a histogram of the *STDs (number)* column with the positive Biopsy ratio coded to color (**Preprocessing Profiling → Visualization**). The histogram shown in (Figure 10-c), reveals that, for the *STDs (number)* column, "null" rows have 1.9% positive biopsies and rows with 0, 1, 2, 3, and 4 have 6.08%, 14.71%, 16.22%, 14.29% and

0%, respectively. There are only eight rows with three or four, which means that the sample size is too small to evaluate these scenarios precisely.

After analyzing the histogram, we reach a few conclusions. There is a positive correlation between *STDs (number)* and the biopsy, that is, a bigger number of STDs tends to be correlated to a bigger number of positive biopsies, identified by a dark color. Moreover, since the "null" rows have a lower positive biopsy ratio than any other group, mixing them with another group might result in loss of information, hindering the perception that the percentage of positive biopsies is lower among them (**Knowledge** → **Preprocessing Profiling**). This observation would have been impossible after deciding the imputation of missing values.

To validate this hypothesis, we choose the practical approach of using the Machine Learning Python library.[66] We create a second version of the dataset (**Preprocessing Profiling** → **Data**) where all the missing values in the *STDs (number)* column are replaced with −1. Subsequently, for both, this new version and the original one, we apply a series of different imputation strategies, each creating a new version of the dataset. The five imputation strategies used, considering all the columns of the dataset, were the replacement of missing values by mean, median, most frequent value, zero, and removing rows with missing values. At the end, we created ten different datasets, two for each imputation strategy (**Data** → **Preprocessing Profiling**).

We proceed to train and test using a decision tree model with each dataset (**Preprocessing Profiling** → **Models**). We repeated this process three times, saving the details about the best and the worst result for each dataset. As expected, since only 0.7% of the rows do not have missing values, the strategy of removing rows with missing values resulted in the worst performance. The best results presented an accuracy of 94% for both of the replacement techniques that were tested for the *STDs (number)* column. The worst varied between 83% and 89%, but this variation is probably caused by the small sample size rather than the effectiveness of a particular strategy. All the other tests had similar results, with accuracy 95%–97%.

These results contradicted our expectations because no significant improvement of the results is noticed when changing the *STDs (number)* column. This probably means that the information we thought that we would lose in some of the scenarios was either irrelevant or maintained by some other property of the dataset (**Models** → **Preprocessing Profiling**).

As an alternative visualization for this case, we generated the Nullity Matrix in Python based on Bilogur,[41] which allows us to confirm the correlation among columns with missing values (**Preprocessing Profiling** → **Data** → **Visualization**). The Nullity Matrix is a data-dense display that supports the identification of patterns for the missing values (Figure 11-a). The records are shown in dark gray for valid records and white for the missing values. Even without prior information, we observe patterns quickly. As proof, three patterns are observed for this dataset: (Figure 11-a1) no occurrence of missing values is noticed in the first and the last eight columns; (Figure 11-a2) there are two columns with high nullity; (Figure 11-a3) many columns seem to be nullity correlated, that is, when one column has a missing value for a particular row, there is a high chance of the other columns in

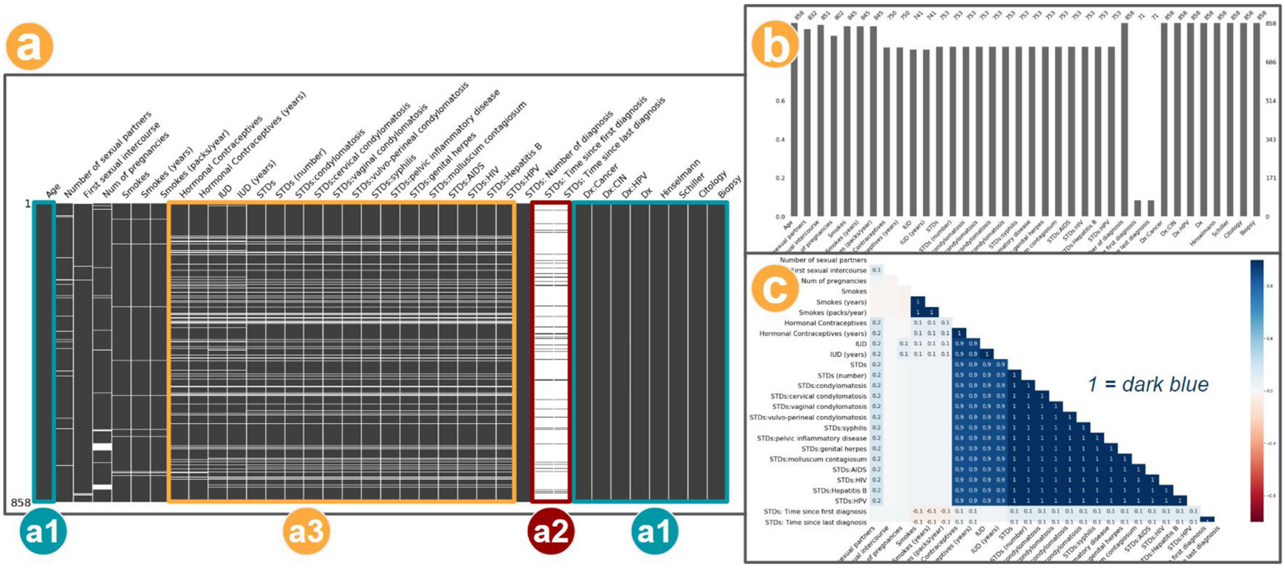


Figure 11. Three visualizations to explore the missing values: (a) matrix (a data-dense display), (b) barplot, and (c) heatmap for variables correlation. This output was generated based on cervical cancer dataset, and using Missingno.[41]

this row having missing values as well. This last pattern was also identified using the Tableau Prep Builder (Figure 10-b), reinforcing our confidence about the property.

Moreover, other visualizations provided by the same library consolidate the observed patterns (**Visualization** → **Knowledge** → **Preprocessing Profiling**). The first and second patterns before mentioned can be confirmed when looking at the Barplot (Figure 11-b), which shows the total count of valid values and allows seeing the proportion of missing values per column. Furthermore, the third pattern can be confirmed when looking at the Heatmap (Figure 11-c), which shows the relationships within pairs of variables having missing values.

Finally, after some additional testing, using combinations of different imputation strategies (**Preprocessing Profiling** ↔ **Data**) and different Machine Learning algorithms (**Preprocessing Profiling** ↔ **Models**), we discovered the combination that results in the best accuracy. More than that, we acquired much deeper knowledge about the dataset (**Knowledge** → **Preprocessing Profiling**).

This use case serves as representation of how the use of PrAVA supports the process of data analysis. This is an example of how the use of a variety of visualization techniques promotes a better understanding of the data under analysis and the impacts of preprocessing. Also, we were able to save information of this process (metadata), which enhances the data preparation understanding itself, that is, the Preprocessing Profiling (**Preprocessing Profiling** → **Preprocessing Profiling**).

Review of scenarios

In this subsection, we present a discussion on how other studies are reporting preprocessing activities as part of their process. To conclude, we summarize how the PrAVA's guidelines are related to the tools used during the application scenarios presented in this section.

How is preprocessing reported? We did an exploratory search for recent works citing the two datasets used in this section. A total of 20 papers were considered: 11 for the mammographic mass dataset, and 9 for the cervical cancer dataset. This exercise supported us to validate our use cases process choices described in this section. We present in this subsection some points observed on the processes involving the preprocessing activities of these works.

The works using the mammographic mass dataset tend not to describe the preprocessing steps in detail. This may happen because of the influence of the work (Elter et al.)[67] for which the dataset was created that used a model capable of handling missing values. Two exceptions are Shobha and Savarimuthu[68] and Azam and Bouguila,[69] which elaborate automatic preprocessing techniques.

Other works that use the cervical cancer dataset tend to describe the preprocessing step in more detail, for example, Ahishakiye et al.,[70] Ahmed et al.,[71] and Ijaz et al.[72] The two primary data quality issues are (a) the missing values and (b) the unbalanced class distribution. The most common preprocessing choices for (a) include removing columns with high missing value ratio, removing rows with missing values, and imputation (mostly with the average and the most frequent value). For (b), the preprocessing strategy is the oversampling.

Most of the other works that use the mammographic mass dataset choose different ways of dealing with data quality issues, including models that accept missing values and automatic preprocessing. These techniques are not the focus of this paper as it is centered around human decision making. Meanwhile, the preprocessing methods we used on the cervical cancer dataset are similar to the mentioned works. Overall, we could not identify any work using visualization to support their process. Therefore, our use of PrAVA exemplifies the possibility of better-informed decisions and a less time-consuming decision process when using the appropriate tools.

As a final remark, we could find observations such as "unstandardized dataset sometimes affects the performance of some of the algorithms" (Ahishakiye et al.,[70] p.10). That supports the value of the preprocessing strategies evaluation and its impacts to further steps of the process.

What is the relation with the guidelines? In Table 3, we present the list of PrAVA's guidelines (Table 2), their status regarding the implementation in each tool used during the application scenarios, and some examples of implementations. In other words, we highlight which guidelines were met by each used tool. The status appears as ✓ to indicate an implemented guideline, ✗ to indicate not implemented guideline, and ⚠ indicates a limited or partially implemented guideline.

Even though Tableau[53] and Tableau Prep[51] are widely used, there are still opportunities to implement further guidelines that should facilitate the preprocessing activities in a VA process, for example, **G1** (Unified), **G4** (Data Mining), and **G6** (Comparison). Consequently, assuming that we do not have access to a solution that covers all the nine guidelines planned to attend PrAVA, we started a parallel effort to

Table 3. List of PrAVA's guidelines and its status of implementation for the used tools: prototype reports (A) *Data Profiling* and (B) *Preprocessing Profiling*; and the commercial software (C) Tableau[53] and (D) Tableau Prep Builder.[51]

Guideline	(A)	(B)	(C)	(D)	Example of implementation
G1 Unified	✓	✓	✗	✗	**(A) (B)** They are integrated with Python Environment (using Jupypter Notebook).
G2 Large Scale	✗	✗	⚠	⚠	**(A) (B)** Aggregation view for the Nullity Matrix, but limited on the number of records and columns the implemented visualization can handle.
G3 Metadata	✓	✓	⚠	⚠	**(A) (B)** HTML Report. **(B)** The information about the results of different strategies ran to build the Preprocessing Profiling understanding.
G4 Data Mining	✗	✓	✗	✗	**(B)** The use of classification model algorithms during the evaluation of missing values imputation strategies.
G5 Statistics	✓	✓	✗	✓	**(A) (D)** A comprehensive presentation of descriptive statistics data.
G6 Comparison	✗	✓	✗	✗	**(B)** Different views to allow the comparison of results during the classification model validation.
G7 Recommendation	✗	✗	✓	⚠	**(C)** Recommendations of visualizations based in the data. **(D)** Recommendations for data transformations, but the tool does not cover advanced possibilities of which visualization can be advised to a particular data issue of interest.
G8 Template	⚠	⚠	✓	⚠	**(A) (B) (D)** Several visualizations generated by default, but they currently do not allow users' customization. **(C)** This tool offers vast options for visualization techniques.
G9 Interaction	⚠	⚠	✓	✓	**(A) (B)** Web user interaction options available, but should be improved for data manipulation flexibility. **(C)** This tool has a user-friendly design with drag and drop properties and other interaction options.

implement a solution to proceed with our intended validation scenarios, mainly for **G4** and **G6**.

It is noteworthy that we do not intend to compare the developed prototype with any commercial software. Rather than that, we aim to show that PrAVA can be used independently of a particular tool. In conclusion, this list of guidelines should be reckoned as a set of practices to evidence the activities executed in the Preprocessing Profiling phase during a VA process. The more these guidelines are considered as part of the developed solution, the more effective the solution will be.

Discussion

In this section, we organize a final discussion of our findings during PrAVA's design and its validation (Subsection *Lessons learned*). We also explain some limitations of this work (Subsection *Limitations*). Finally, we present some topics that can be interpreted as research opportunities in the context of this work (Subsection *Research opportunities*).

Lessons learned

The main findings observed during our literature review were explained in Subsection *Review and comparison*. However, the nine guidelines presented in Table 2 summarize most of what we have learned in this process. To compile this list with some level of confidence in its contribution required the analysis of multiple works. Additionally, we summarize below some of our findings during this process organized as four lessons learned.

Critical but less discussed. Preprocessing is recognized as a critical phase to the data analysis process, due to the data preparation time-consuming nature or its impacts on the final results. Contradictorily, it is still a subject that receives less attention from the VA and visualization communities.

Implementing all the guidelines is not a trivial task. During the scenario coverage planning, we realized that there are many combinations to consider to set up all the required components under a new solution in compliance to PrAVA. We may need definitions of questions as *what is the data mining scope? Which Machine Learning or statistical methods can be used to solve the problem? Which data quality issues are intended to be addressed?* That leads to a chain of other questions, for example, *which data transformation strategies can be used with this particular data issue? Which visualization techniques can be used to support this context?* To sustain our decision on each strategy to use in response to these questions, we considered the references presented in Subsection *What the practitioners say*. Additionally, these decisions impacted how the guidelines could be implemented. To sum up, we acknowledge that implementing all the guidelines, even if aiming to cover a limited scope, is far from a trivial task.

Simplicity of the visualizations. Although most of the visualizations used in the usage scenario (Subsection *Tim and the iris dataset*) and the applications (Section *Applications*) are simple, they still demonstrate more benefits to understand the data when compared to viewing the plain text. The simplicity should favor understanding since it does not require a prior explanation, that is, most of the visualizations used are already part of the data analysts' culture. Thus, since different users have different experiences, expectations, and graph literacy, the use of traditional charts is appropriate for most cases, as suggested by the insights in our previous study.[8] This is also in adherence to the idea of promoting visualization literacy.[73,74]

The value of an integrated tool. Through practicing on a developed prototype, three main advantages can be mentioned. First, considering we have the dataset loaded in the Python programming environment, with one command line to import the library and another to call the report, we can generate detailed and relevant information to support preprocessing activities. Consequently, we contribute to simplify the working procedures of data analysts, which is a big concern since it 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, this detailed information about the dataset and data preparation can be used as metadata for the preprocessing profiling phase. It helps build the principle of transparency on the activities performed, aligned to initiatives such as the European Union General Data Protection Regulation (https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en). As mentioned earlier, a system nor a tool is the focus of this work; however, during the usage scenario, the value of an integrated tool in this process was evidenced, which is aligned with **G1** (Unified).

Awareness-raising. The actual VA process (Figure 2) can continue as-is since it covers confirmatory analysis cases, or when the dataset is well-known and automated methods for preprocessing are in place. However, its current representation conceals the importance of preprocessing. Thus, PrAVA better positions the critical components of preprocessing efforts. That is especially relevant in scenarios where the decisions made during preprocessing are crucial to the further phases of the process, and active participation of the data analyst is required. Moreover, other studies have explored the role of uncertainty as part of the VA process,[25,75,76] and they emphasize that uncertainty in data can often be propagated during preprocessing activities. Thus, the efforts to develop alternatives to increase the awareness and trust of the data under analysis will contribute to a more reliable VA process.

Limitations

We identified four limitations in the current work that we consider important to explain.

Problem instance. As stated by Munzner[77] (p. 3), "Vis systems are appropriate for use when your goal is to augment human capabilities, rather than completely replace the human in the loop." Hence, our scope considers the cases when the "human in the loop" is vital to the preprocessing. That means, the data analyst is still evaluating and formulating the questions about the data under analysis. For other cases, when the quality of the data is not a concern, the dataset properties are known, or all the needed preprocessing tasks are already mapped; thus, most of this process can be automated, and there will be no applicability to the approach we are discussing.

Guidelines' list. To allow the extension of PrAVA to a variety of scenarios, and to facilitate its adoption, we have tried to design our approach as general and as simple as possible. As a consequence, if on the one hand, PrAVA may cause a first impression that some of the guidelines are quite obvious. On the other hand, it may not explicitly indicate all the complexity behind preprocessing. However, using the guidelines will result in solutions in which preprocessing is consistently considered. It is hard to assert that all potential scenarios are covered and new guidelines may emerge in the literature over time or from different types of applications that were not considered. Overall, we still consider helpful keeping the nine proposed guidelines structured for a consolidated reference.

Usage scenarios. We have not intended to present a detailed description of the types and strategies applied to the preprocessing scope, since we consider it a subject to another dedicated work (see Subsection *Preprocessing + Visualization taxonomy*). Thus, we limited our examples to scenarios that allowed us to encourage a general understanding of the PrAVA process.

Applications. We decided to proceed with the use cases (considering the definition from Ward et al.[14]) to support the PrAVA validation strategy instead of using empirical methods with the participation of data analysts or domain experts. As part of the mitigation for the risks in not covering a realistic scenario, as explained in Subsection *How is preprocessing reported*, we searched for related work using the same datasets selected for our applications. We evaluated how they reported the preprocessing activities, and then we compared their process with the activities we performed. Nevertheless, we still consider important that an extension of PrAVA conduct user-centered experiments to obtain insightful comments to fine-tune this work.

Research opportunities

Interesting research directions in the scope of preprocessing and visualization were introduced by Kandel et al.[4] Although this work contains the perspective of a decade ago, its discussion is still relevant. Shall this be explained by the fact preprocessing as an object of study has received less attention from our community? In any case, to advance the discussion, we are indicating promising directions for further research.

Preprocessing + Visualization taxonomy. A comprehensive and up-to-date taxonomy of data quality issues related to preprocessing strategies and visualization techniques is needed. This effort should include the type of data quality, the issue description, the detection methods, the preprocessing transformation methods, and visualization techniques that can be used to assist in this process. To illustrate, a good start could result in an enhanced combination of the discussion presented in Kandel et al.[16] (preprocessing + visualization) and Kim et al.[1] (taxonomy of data issues).

Complementary to the previous point, the exploration of the preprocessing strategies considering the challenges of application domains, for example, fraud detection or public health, and data mining scope. Moreover, besides the perspective of the data analyst, other perspectives can be explored as well. For instance, in healthcare, the preprocessing tasks are often done by the domain expert. *Is there any particular requirement to attend a domain expert compared to the data analyst in the preprocessing solution?* This new study could be used as a benchmark before planing new solutions.

Visualizing data issues. We can consider two main groups of new visualizations to be explored. One is related to the understanding of data issues in raw data. Providing different views for the same data issue may allow discoveries that could not be noticed using just one visualization.

An alternative is to create a coordinated multiple view framework for different data issues. A similar idea was proposed by Sjöbergh and Tanaka[34] in the scope of missing values. Along with missing values, the outliers are another frequent data issue that requires attention, because *how to differentiate what is noise and what is an outlier?* The second group is the understanding of the impacts of the preprocessing. For instance, *how to support pattern identification on misclassification that is caused by missing values?*

Although the VA and Visualization community have a strong foundation in cognitive human perception and a variety of methods and techniques have been developed to create visual metaphors of the data, in the context of the preprocessing, we still can formulate a question like *What helps the data analyst see a data issue?* One possible way to obtain this answer is through empirical studies with the engagement of data analysis while working on practical problems based on real-world data and scenarios. Based on that, we could get inputs on the most significant elements that support data analysts to identify a data issue. This item is somehow aligned to studies in the area of visualization literacy, for example, Galesic and Garcia-Retamero[78] evaluates the graph literacy applied to the medical domain, and those concerned with visualizing uncertainty, for example, Correat et al.,[75] Sacha et al.,[76] and Seipp et al.[25]

Systems and tools. Despite the fact we can find studies such as Zhang et al.,[79] and its more recent revision Behrisch et al.,[80] evaluating VA commercial systems in Big Data scenarios, we consider it worth to continue a comparative review of the state-of-the-art for open source and systems with special attention to preprocessing. As part of this discussion, it should be evaluated if the planned guidelines of PrAVA are attended or not.

Recommendation. Although multiple works have presented advanced solutions in the scope of data cleaning and transformation recommendations, within **G7**, further investigation is required when considering data issues + preprocessing goals. Possibly more effective recommendations can be built based on the discoveries of the taxonomy studies (see Subsection *Preprocessing + Visualization taxonomy*).

Big data. Regarding data transformation activities in high-dimensional data, Liu et al.[81] provide a comprehensive survey on the topic that can be used as source of inspiration. While the Progressive Visual Analytics, proposed by Stolper et al.[56] indicates an alternative to handle Big Data scenarios, its adoption may cause new challenges, such as whether a current partial outcome is already good enough.[82] In the scope of preprocessing (**G2**), if we share part of the data, we may hide data quality issues that need to be observed and fixed. Subsequently, new questions arise, *how can we share partial data without impacting the data quality issue evaluation? Or which other alternatives do we have?*

Likewise, careful validation of aggregation strategies, as indicated by Elmqvist and Fekete,[57] is needed to allow any visual metaphor to scale while analyzing large and complex datasets. Otherwise, a wrong design decision may introduce data distribution issues that may impair the visual identification of any pattern. For these cases, the resulting visualization is diminished and leads to uncertainty in the data.[25]

Conclusion

A state-of-the-art literature review and the practitioners' testimony in data analysis allowed us to reach the following conclusion: Data preprocessing is seen as one of the most laborious and time-consuming – and even tedious as stated by Kandel et al.[4] – activities of the data analysis process. Notwithstanding, few works in the Visual Analytics and Visualization areas address the challenges related to preprocessing as their research subject. Moreover, some studies do not explicitly consider preprocessing as an equally important activity to the knowledge discovery process's final findings.

Thus, in this paper, we presented the Preprocessing Profiling Approach for Visual Analytics (PrAVA). Our main contributions can be summarized as introducing PrAVA as an alternative to support data analysts during preprocessing activities. By enabling better data understanding and evaluating preprocessing impacts, these methods should promote data quality and provide grounds for decision-making on data preparation strategies. Ultimately, we hope that we encourage a shift to a visual preprocessing.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by PUCRS (Edital Chamada Geral 1/2019 - PUCRS/BPA). Isabel Harb Manssour also would like to thank the financial support of the CNPq Scholarship - Brazil (308456/2020-3). Alessandra Maciel Paz Milani wants to acknowledge that this work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, during her Master's studies, and the current support at the University of Victoria during the review process. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

ORCID iDs

Alessandra Maciel Paz Milani https://orcid.org/0000-0001-8900-4179
Fernando Vieira Paulovich https://orcid.org/0000-0002-2316-760X
Isabel Harb Manssour https://orcid.org/0000-0001-9446-6757

References

1. Kim W, Choi B, Hong E, et al. A taxonomy of dirty data. *Data Min Knowl Disc* 2003; 7: 81–99.
2. Tan P, Steinbach M and Kumar V. *Introduction to data mining*. Pearson Education, 2006. Boston, MA, USA.
3. Kandel S, Paepcke A, Hellerstein J, et al. Wrangler: Interactive visual specification of data transformation scripts. In: *Proceedings of the conference on human factors in computing systems*, Vancouver, BC, Canada, 2011, pp.3363–3372.
4. Kandel S, Heer J, Plaisant C, et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Inform Visual* 2011; 10: 271–288.
5. Wickham H. Tidy data. *J Stat Softw* 2014; 59: 1–23.
6. Krishnan S, Haas D, Franklin MJ, et al. Towards reliable interactive data cleaning: A user survey and recommendations. In: *Proceedings of the workshop on human-in-the-loop data analytics*, San Francisco, CA, USA, 2016, pp.1–5.
7. Sacha D, Kraus M, Keim DA, et al. VIS4ML: An ontology for visual analytics assisted machine learning. *IEEE Trans Vis Comput Gr* 2019; 25: 385–395.
8. Milani A, Paulovich F and Manssour I. Visualization in the preprocessing phase: Getting insights from enterprise professionals. *Inform Visual* 2020; 19: 273–287.
9. Dasu T and Johnson T. *Exploratory data mining and data cleaning*. John Wiley & Sons, 2003. New York, NY, USA.
10. Rahm E and Do H. Data cleaning: Problems and current approaches. *Bull IEEE Comput Soc Tech Committee Data Eng* 2000; 23: 3–13.
11. Tam GKL, Kothari V and Chen M. An analysis of machineand human-analytics in classification. *IEEE Trans Visual Comput Graph* 2017; 23: 71–80.
12. Kandogan E, Balakrishnan A, Haber EM, et al. From data to insight: Work practices of analysts in the enterprise. *IEEE Comput Graph Appl* 2014; 34(05): 42–50.
13. de Oliveira MC and Levkowitz H. From visual data exploration to visual data mining: A survey. *IEEE Trans Visual Comput Graph* 2003; 9: 378–394.

14. Ward M, Grinstein G and Keim D. *Interactive Data Visualization: Foundations, Techniques, and Applications*. AK Peters/CRC Press, 2015. Natick, MA, USA.
15. Lu J, Chen W, Ma Y, et al. Recent progress and trends in predictive visual analytics. *Front Comput Sci* 2017; 11: 192–207.
16. Kandel S, Parikh R, Paepcke A, et al. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the conference on advanced visual interfaces*, Capri Island, Italy, 2012, pp.547–554.
17. Bernard J, Ruppert T, Goroll O, et al. Visual-interactive preprocessing of time series data. In: *Proceedings of SIGRAD 2012; interactive visual analysis of data*, November 29–30, 2012; pp. 39–48. Växjö.
18. Gschwandtner T, Aigner W, Miksch S, et al. 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*, Graz, Austria, 2014, pp.1–8.
19. Keim D, Kohlhammer J and Ellis G. *Mastering the information age: solving problems with visual analytics*. Eurographics Association, 2010. Goslar, Germany.
20. Sacha D, Stoffel A, Stoffel F, et al. Knowledge generation model for visual analytics. *IEEE Trans Visual Comput Graph* 2014; 20: 1604–1613.
21. Johnson T. Data profiling. *Encyclopedia Database Syst* 2009; 1: 604–608.
22. Ribarsky W and Fisher B. The human-computer system: Towards an operational model for problem solving. In: *Proceedings of the Hawaii international conference on system sciences*, Wailea, HI, USA, 2016, pp.1446–1455.
23. Federico P, Wagner M, Rind A, et al. The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics. In: *Proceedings of the IEEE conference on visual analytics science and technology*, Phoenix, AZ, USA, 2017, pp.92–103.
24. Lu Y, Garcia R, Hansen B, et al. The state-of-the-art in predictive visual analytics. *Comput Graph Forum* 2017; 36: 539–562.
25. Seipp K, Ochoa X, Gutiérrez F, et al. A research agenda for managing uncertainty in visual analytics. In: *Mensch and computer 2016 — Workshopband*, Aachen, Germany, 2016, pp. 1–10.
26. Krause J, Perer A and Stavropoulos H. Supporting iterative cohort construction with visual temporal queries. *IEEE Trans Visual Comput Graph* 2016; 22: 91–100.
27. Sacha D, Kraus M, Bernard J, et al. SOMFlow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance. *IEEE Trans Visual Comput Graph* 2018; 24: 120–130.
28. Heer J, Hellerstein J and Kandel S. Predictive interaction for data transformation. In: *Proceedings of the biennial conference on innovative data systems research*, Asilomar, CA, USA, 2015, pp.1–7.
29. von Zernichow B and Roman D. Usability of visual data profiling in data cleaning and transformation. In *Proceedings of the on the move to meaningful internet systems*, Rhodes, Greece, 2017, pp.480–496.
30. Chandola V, Banerjee A and Kumar V. Anomaly detection: A survey. *ACM Comput Surv (CSUR)* 2009; 41: 1–58.
31. Wang X, Dong XL and Meliou A. Data X-Ray: A diagnostic tool for data errors. In: *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, Melbourne, Australia, 2015, pp.1231–1245.
32. Templ M, Alfons A and Filzmoser P. Exploring incomplete data using visualization techniques. *Adv Data Anal Classi* 2012; 6: 29–47.
33. Eaton C, Plaisant C and Drizd T. Visualizing missing data: Classification and empirical study. In: *Proceedings of the conference on human-computer interaction*, Rome, Italy, 2005. pp.861–872.
34. Sjöbergh J and Tanaka Y. Visualizing missing values. In: *Proceedings of the conference information visualisation*, London, UK, 2017, pp.242–249.
35. Song H and Szafir DA. Where's my data? evaluating visualizations with missing data. *IEEE Trans Visual Comput Graph* 2019; 25: 914–924.
36. McNutt A, Kindlmann G and Correll M. Surfacing visualization mirages. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*, Honolulu, HI, USA, 2020, pp.1–16.
37. Batch A and Elmqvist N. The interactive visualization gap in initial exploratory data analysis. *IEEE Trans Visual Comput Graph* 2018; 24: 278–287.
38. Kandel S, Paepcke A, Hellerstein J, et al. Enterprise data analysis and visualization: An interview study. *IEEE Trans Visual Comput Graph* 2012; 18: 2917–2926.
39. Crone SF, Lessmann S and Stahlbock R. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *Eur J Oper Res* 2006; 173: 781–800.
40. Pandas-profiling. Pandas-profiling, 2020. https://github.-com/pandas-profiling/pandas-profiling.
41. Bilogur A. Missingno: a missing data visualization suite. *Journal of Open Source Software* 2018; 3: 1–4.
42. Bengfort B and Bilbro R. Yellowbrick: Visualizing the scikit-learn model selection process. *J Open Source Softw* 2019; 4: 1–5.
43. Liu M, Shi J, Cao K, et al. Analyzing the training processes of deep generative models. *IEEE Trans Visual Comput Graph* 2018; 24: 77–87.
44. Alsallakh B, Jourabloo A, Ye M, et al. Do convolutional neural networks learn class hierarchy? *IEEE Trans Visual Comput Graph* 2018; 24: 152–162.
45. Strobelt H, Gehrmann S, Pfister H, et al. LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Trans Visual Comput Graph* 2018; 24: 667–676.
46. Krause J, Dasgupta A, Swartz J, et al. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE conference on visual analytics science and technology (VAST)*, Phoenix, AZ, USA, 2017, pp.162–172.
47. Mühlbacher T, Linhardt L, Möller T, et al. TreePOD: Sensitivity-aware selection of pareto-optimal decision trees. *IEEE Trans Visual Comput Graph* 2018; 24: 174–183.

48. Ren D, Amershi S, Lee B, et al. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Trans Visual Comput Graph* 2017; 23: 61–70.
49. Wongsuphasawat K, Moritz D, Anand A, et al. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Trans Visual Comput Graph* 2016; 22: 649–658.
50. Vartak M, Huang S, Siddiqui T, et al. Towards visualization recommendation systems. *ACM SIGMOD Rec* 2017; 45: 34–39.
51. Tableau. Tableau prep. URL https://www.tableau.com/products/prep (2020).
52. Trifacta. Trifacta data wrangling tools & software. URL https://www.trifacta.com/ (2020)
53. Tableau. Tableau. URL http://www.tableau.com/ (2020)
54. Shearer C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehous* 2000; 5: 13–22.
55. Turkay C, Pezzotti N, Binnig C, et al. Progressive data science: Potential and challenges. *arXiv preprint* 2018; 1812.08032: 1–10.
56. Stolper C, Perer A and Gotz D. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Trans Visual Comput Graph* 2014; 20: 1653–1662.
57. Elmqvist N and Fekete JD. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Trans Visual Comput Graph* 2010; 16: 439–454.
58. Heer J and Kandel S. Interactive analysis of big data. *Big Data* 2012; 19: 50–54.
59. Dimara E and Perin C. What is interaction for data visualization? *IEEE Trans Visual Comput Graph* 2020; 26(1): 119–129.
60. Crisan A and Munzner T. Uncovering data landscapes through data reconnaissance and task wrangling. In: *2019 IEEE visualization conference (VIS)*, Vancouver, BC, Canada, 2019, pp.46–50.
61. Munzner T. A nested model for visualization design and validation. *IEEE Trans Visual Comput Graph* 2009; 15(6): 921–928.
62. Fisher R. The use of multiple measurements in taxonomic problems. *Ann Eugenics* 1936; 7: 179–188.
63. Gleicher M, Barve A, Yu X, et al. Boxer: Interactive comparison of classifier results. *Comput Graph Forum* 2020; 39(3): 181–193.
64. Dua D and Graff C. The UCI machine learning repository - mammographic mass data set. URL https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass (2020).
65. Dua D and Graff C. The UCI machine learning repository - cervical cancer (risk factors) data set. https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29 (2020).
66. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
67. Elter M, Schulz-Wendtland R and andWittenberg T. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *Med Phys* 2007; 34(11): 4164–4172.
68. Shobha K and Savarimuthu N. Clustering based imputation algorithm using unsupervised neural network for enhancing the quality of healthcare data. *J Amb Intel Hum Comput* 2020; 12(2): 1771–17811.
69. Azam M and Bouguila N. Multivariate bounded support laplace mixture model. *Soft Comput* 2020; 24(7): 13239–13268.
70. Ahishakiye E, Wario R, Mwangi W, et al. Prediction of cervical cancer basing on risk factors using ensemble learning. In: *2020 IST-Africa conference (IST-Africa)*, Kampala, Uganda, 2020, pp.1–12. IEEE.
71. Ahmed M, Kabir M, Kabir M, et al. Identification of the risk factors of cervical cancer applying feature selection approaches. In: *International conference on electrical, computer and telecommunication engineering*, Rajshahi, Bangladesh, 2019, pp.1–5.
72. Ijaz MF, Attique M and Son Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors* 2020; 20(10): 2809.
73. D'Ignazio C. Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Inf Des J* 2017; 23: 6–18.
74. Gray J, Bounegru L, Milan S, et al. Ways of seeing data: Toward a critical literacy for data visualizations as research objects and research devices. In: Kubitschko, Sebastian and Kaun, Anne (Eds) *Innovative methods in media and communication research*. Springer, 2016, pp.227–251. Cham, Switzerland. https://link.springer.com/chapter/10.1007/978-3-319-40700-5_12
75. Correa CD, Chan YH and Ma KL. A framework for uncertainty-aware visual analytics. In: *2009 IEEE symposium on visual analytics science and technology*, 2009, pp. 51–58.
76. Sacha D, Senaratne H, Kwon BC, et al. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Trans Visual Comput Graph* 2016; 22: 240–249.
77. Munzner T. *Visualization analysis and design*. CRC press, 2014. Boca Raton, FL 33487-2742. https://www.routledge.com/Visualization-Analysis-and-Design/Munzner/p/book/9781466508910#
78. Galesic M and Garcia-Retamero R. Graph literacy: A crosscultural comparison. *Med Decis Mak* 2011; 31: 444–457.
79. Zhang L, Stoffel A, Behrisch M, et al. Visual analytics for the big data era — a comparative review of state-of-the-art commercial systems. In: *2012 IEEE conference on visual analytics science and technology (VAST)*, Seattle, WA, USA, 2012, pp.173–182.
80. Behrisch M, Streeb D, Stoffel F, et al. Commercial visual analytics systems–advances in the big data analytics field. *IEEE Trans Visual Comput Graph* 2019; 25(10): 3011–3031.
81. Liu S, Maljovec D, Wang B, et al. Visualizing high-dimensional data: Advances in the past decade. *IEEE Trans Visual Comput Graph* 2017; 23: 1249–1268.
82. Angelini M, May T, Santucci G, et al. On quality indicators for progressive visual analytics. In: *EuroVis workshop on visual analytics (EuroVA)*, Porto, Portugal, 2019, pp.25–29.