Info Vis

Article

Information Visualization
2020, Vol. 19(4) 273–287
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1473871619896101
journals.sagepub.com/home/ivi
SAGE

Visualization in the preprocessing phase: Getting insights from enterprise professionals

Alessandra Maciel Paz Milani[1], **Fernando V. Paulovich**[2] **and Isabel Harb Manssour**[1]

Abstract

The current information age has increasingly required organizations to become data-driven. However, analyzing and managing raw data is still a challenging part of the data mining process. Even though we can find interview studies proposing design implications or recommendations for future visualization solutions in the data mining scope, they cover the entire workflow and do not fully focus on the challenges during the preprocessing phase and on how visualization can support it. Moreover, they do not organize a final list of insights consolidating the findings of other related studies. Hence, to better understand the current practice of enterprise professionals in data mining workflows, in particular, during the preprocessing phase, and how visualization supports this process, we conducted semi-structured interviews with 13 data analysts. The discussion about the challenges and opportunities based on the responses of the interviewees resulted in a list of 10 insights. This list was compared with the closest related works, improving the reliability of our findings and providing background, as a consolidated set of requirements, for future visualization research articles applied to visual data exploration in data mining. Furthermore, we provide greater details on the profile of the data analysts, the main challenges they face, and the opportunities that arise while they are engaged in data mining projects in diverse organizational areas.

Keywords

Visualization, preprocessing, visual data exploration, data mining, interviews

Introduction

The data-driven society in which we live led us to accumulate massive volumes of data in the most variety of domains. The process of data analysis for knowledge extraction is still a very challenging, laborious activity. During the process of data exploration, data analysts spend most of their time on data preparation activities,[1] that is, the preprocessing phase, when we consider data mining[2] workflows, such as Knowledge Discovery in Databases (KDD)[3] or Cross-Industry Process for Data Mining (CRISP-DM).[4] As examples of the demanding activities that are part of the preprocessing phase, we can list completeness and conformity of data quality, since there is not a single technique or tool to solve all data issues automatically.[5,6] Therefore, intense interaction between raw data and data analysts is required to perform the decisions on how to proceed with the data management.[1,7]

Consequently, the preprocessing purpose of transforming “the raw input data into an appropriate format for subsequent analysis”[8] may often not be carried out impartially, which means new issues may arise due to

[1] Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil
[2] Dalhousie University, Halifax, NS, Canada

Corresponding author:
Alessandra Maciel Paz Milani, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS, 90619-900, Brazil.
Email: paz.alessandra@gmail.com

the data analysts. For instance, they can update missing values with the mean calculated based on other instances in their dataset instead of the median to avoid outliers or they can even ignore data, for example, deleting instances due to missing values in a specific attribute, which was supposed to be fixed before proceeding with the data analysis. Thus, no matter how robust the algorithm created for data mining is, if bad data from a source are used or a data manipulation strategy is wrongly selected, it may lead to the identification of wrong patterns and misunderstanding in the final results.[1]

Under these circumstances, visualization techniques and visual data exploration could play an important role in data analysis while providing meaningful insights.[7,9,10] However, most of the visualization studies are concerned with the end of the process when sharing the final results of the analysis. Likewise, we can find interview studies with enterprise professionals proposing design implications[11,12] or recommendations[13] for future visualization solutions in the data mining scope, but they cover the entire workflow and do not focus fully on the challenges during the preprocessing phase and on how visualization can support it. Moreover, they do not organize a final list of insights consolidating the findings of other related studies.

In this article, we aim to gather requirements of how visualization can be used as a powerful tool to be incorporated into the toolkit of the data analysts during the preprocessing phase to foster visual data exploration. We conducted an interview study with 13 enterprise professionals to investigate their working practices. As a result, we present a consolidated list of 10 insights as to how visualization can support the preprocessing activities based on the data analysts' perspective on data exploration. Furthermore, when analyzing the responses of the interviewees, we provide greater details on the profile of the data analysts, the main challenges they face, and the opportunities that arise while they are engaged in data mining projects in diverse organizational areas, for example, e-commerce and finance.

It is important to highlight that the summarization of practical items, such as 10 rules of thumb, provides an overview of the requirements in the preprocessing phase for new visualization efforts, speeding up newcomers' progress. We also hope it serves as a background for future studies on visualization research applied to data mining, contributing to create awareness of the current gaps and to increase the adoption of visualization techniques as part of the daily practice of data analysts, mainly earlier in their workflow.

The remainder of this article is structured as follows: section "Related work" describes the literature review methodology and the interview studies focusing on capturing the experience of data analysts while evaluating design implications in the data mining scope. Subsequently, section "Interview study" outlines the procedure developed to perform the interviews, the profile of the participants, and the results and analysis of the interviews. Section "Insights for new visualizations" presents the list of 10 insights resulting from our study and details the comparative analysis with the related work. Section "Limitations" summarizes the opportunities for improvements in our study. Finally, Section "Conclusion and future work" presents our conclusions and plans for future work.

Related work

We conducted a state-of-the-art literature review to explore interview studies capturing the experience of data analysts while visualizing data during the data mining process. More specifically, we were interested in studies presenting visualization guidelines, challenges, opportunities, or gaps in the preprocessing phase. The selected studies presented a discussion on data analysis from the perspective of enterprise professionals and used interviews with semi-structured questionnaires as a data collection instrument. They are referenced in this work as RW1 for Batch and Elmqvist,[11] RW2 for Kandel et al.,[12] and RW3 for Alspaugh et al.[13]

RW1 developed a variant of contextual inquiry to observe eight data analysts in their work environment. All the participants worked for the US Government in Washington, DC. Their experience in data mining ranged from 4 to 20 years. The interview analysis was very detailed; however, the main limitation of the study is the lack of representation of professionals from different sectors. On the contrary, RW2 interviewed 35 enterprise analysts who were working in 25 organizations across a variety of industries. Although most of the participants were located in Northern California in the United States, this scenario brought good coverage of heterogeneous experiences and responses to be analyzed. However, the activities for the preprocessing phase were not fully explored since the study aimed to characterize the space of analytic workflows as a whole.

Even though RW3 did not aim primarily to explore visualization options, its results, based on interviews with 30 data analysts located in the San Francisco Bay Area in the United States, were still relevant to us, in particular, because they presented an extensive discussion on data exploration practices, which included visualization as a tool.

To summarize, these three studies proposed design implications (RW1 and RW2) or recommendations

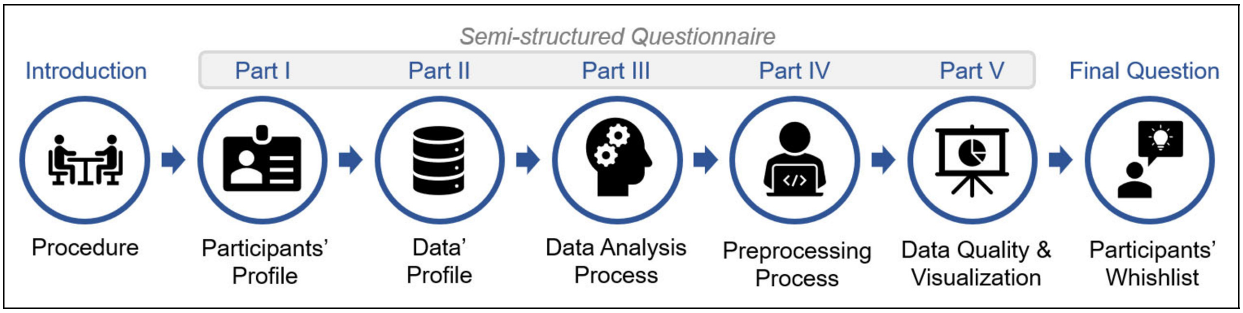


Figure 1. Overview of the interview process followed during our study.

(RW3) for future tools in data exploration or visual analytics research. Their investigation contributed to identifying challenges, opportunities, and barriers to adopt visualization during exploratory data analyses. Hence, they were used to ratify most of the items included in our final list of insights for new visualizations.

Nevertheless, we can still highlight relevant differences when comparing them with the proposal of our study. First, in our research, we explore aspects to broaden the understanding of how the preprocessing phase is performed in data mining workflows and we instigate the discussion on how visualization could contribute to that process. Moreover, we go into greater detail concerning the profile of the data analysts, including a description of their work process, details on data type and source, tools and technologies, and strategies for data mining or machine learning (ML) in use. Finally, we compiled a more straightforward list of requirements for future visualization solutions in this research area, considering the inputs received by enterprise professionals combined with the review of these three related works.

Interview study

As a qualitative data collection instrument, we developed a semi-structured questionnaire to guide the interviews with the data analysts. Most of the questions were open-ended to capture as much information as possible during the interviews. Some questions covered the participant's profile with a few demographic items. Others were intended to encourage the participants to describe their working practices to provide an overview of their data exploration processes. In addition, some questions were phrased specifically to address the visualization strategies as part of the preprocessing activities. Furthermore, few related works[11,12,14] were used as reference points during the development of the procedure and the definition of the questions. The interview process is summarized in Figure 1.

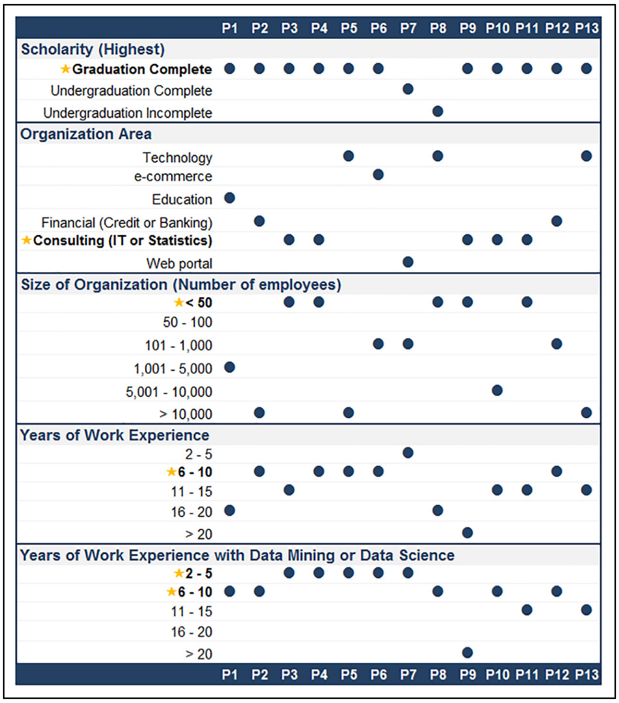
Participants

We set as a goal to interview between 10–15 data analysts considering the research methods in human–computer interaction.[15] The participants were recruited based on their engagement with the practice of data mining. We used online platforms, such as LinkedIn and Meetup, and our professional network to identify potential participants. We interviewed a total of 13 professionals, 12 male and 1 female, with ages ranging from 26 to 42 years. They were located in three different cities from Brazil: Porto Alegre, São Paulo, and Rio de Janeiro.

Our participants worked in different areas, such as technology consulting and services, education, finances, web portals, statistical consulting, and e-commerce. However, 12 of them worked in the private sector, and only 1 participant had a governmental job. There were three cases where they held positions at the industry and the academy at the same time. The range of their company size was significantly wide, from 3 to close to 100,000 collaborators. Their organizational roles varied from director or manager (31%) to researcher (23%), but most of them were officially data scientists or data analysts (46%).

The majority of participants (85%) had received master's degrees in computer science, engineering, statistics, or business administration. One of them completed a PhD program, and three were PhD candidates. Their background during their undergraduate studies included different areas, such as physics, statistics, engineering, and business administration. However, computer science–related areas were still predominant among this group.

The length of experience of the participants in the technology field ranged from 6 to 15 years and, with regards to data exploration more specifically, the range was reduced to 2–10 years. That happened because



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
Scholarity (Highest)													
★**Graduation Complete**	●	●	●	●	●	●			●	●	●	●	●
Undergraduation Complete							●						
Undergraduation Incomplete								●					
Organization Area													
Technology					●			●					●
e-commerce						●							
Education	●												
Financial (Credit or Banking)		●										●	
★**Consulting (IT or Statistics)**			●	●					●	●	●		
Web portal							●						
Size of Organization (Number of employees)													
★**< 50**			●	●				●	●		●		
50 - 100													
101 - 1,000						●	●					●	
1,001 - 5,000	●												
5,001 - 10,000										●			
> 10,000		●			●								●
Years of Work Experience													
2 - 5							●						
★**6 - 10**		●		●	●	●						●	
11 - 15			●							●	●		●
16 - 20	●							●					
> 20									●				
Years of Work Experience with Data Mining or Data Science													
★**2 - 5**			●	●	●	●	●						
★**6 - 10**	●	●						●		●		●	
11 - 15											●		●
16 - 20													
> 20									●				
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13

Figure 2. Profile information for the 13 participants of our interview study, which includes level of education, work organization area and size, and years of work experience.

62% of the participants started working in positions outside data mining. Further details on the participants' profile is shown in Figure 2.

Procedure

Each participant was interviewed continually, and the sessions lasted from 30 to 60 min. The same environment configuration was used for all participants, face-to-face or online conversations, that is, calls or video conferences. First, we introduced the procedure and presented the consent form, in compliance with our Research Ethics Committee (REC). Subsequently, we briefly introduced our study and we provided participants with the opportunity to ask any questions regarding the explained items.

The interview was guided by a semi-structured questionnaire consisting of five parts and a total of 25 questions, which is available in Appendix 1. A copy of the questionnaire was shared with the participants during the interview. In addition, we asked participants to consider their most recent data analysis projects while answering the questions.

A pilot interview was run to confirm the clarity of the questions and the approximate duration required for the activity. Since it occurred as planned, the content of the pilot interview was regarded as part of this study, as participant number 1. The interviews were performed in May, June, and July 2018, by the same interviewer. During each session, the interviewer took extensive notes of the answers. Parts of the sessions were recorded, with the consent of participants, and the audio was used to review the notes.

We developed the analysis code of the responses primarily following the same structure used for the questionnaire, divided into five parts. Afterward, the questions related to each part worked as a second level of coding. We tabulated the collected data following these two levels, which resulted in 325 entries, that is, each entry is the transcript for the open responses provided by each of the 13 participants. In more details: Part I, participant profile, resulted in 117 entries since there were nine questions; Part II, data profile, resulted in 52 entries since there were four questions; Part III, data analysis process, resulted in 52 entries since there were four questions; Part IV, preprocessing activities, resulted in 52 entries since there were four questions; and Part V, visualization techniques, resulted in 52 entries since there were four questions. Later, the content of each question was analyzed, comparing the responses of all participants. During that step, the third level of code was created to group similar responses. In the next subsection, we describe the recurring patterns and the significant elements observed during this analysis. As a rule, we considered the items reported by more than two participants. However, those items emphasized as important, even if only by one participant, were discussed as well.

Analysis of the interviews and results

The results and discussion based on the analysis of the responses were grouped into four items: data profile, data analysis process, preprocessing activities, and visualization of data quality issues. The most relevant aspects are described in the following paragraphs. In relation to the numerical computation in this analysis, it is important to note we are only counting explicit responses. Therefore, for some situations, we cannot assume the other participants agree or disagree with a particular point since their answers were not counted.

Data profile. The information captured about the source, format, and type of data is summarized as part of Figure 3. Regarding the volume of the datasets in use, it ranges from a small number of data records, that is, which can be processed in simple spreadsheet, to Big Data[16] infrastructures, with billions of records and more than 100,000 features.

Data analysis process. Participants described their work process similarly to KDD, ML, or CRISP-DM

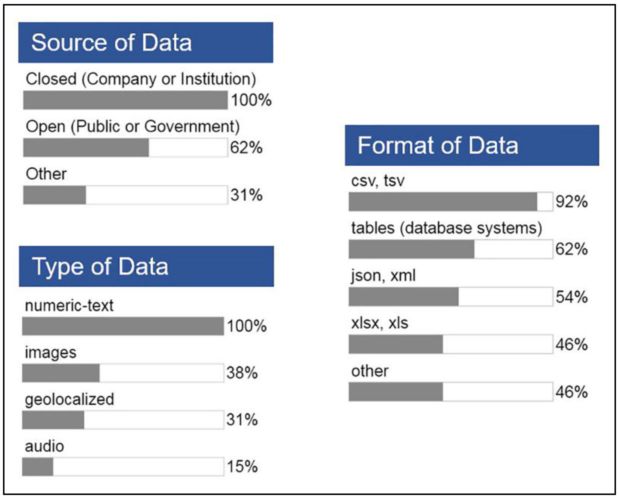


Figure 3. Data profile: details on source, type, and format of data. Based on the responses of the 13 participants of our interview study.

workflows, see Figure 4 for details. Moreover, the participants mentioned that the steps may vary according to the scope and type of project. For some cases, these workflow tasks were mixed; for instance, (1) *Business understanding* and (2) *Data understanding* from CRISP-DM were added as pre-steps in the KDD and ML workflows. One participant added a new step (0) *Research* to represent the literature review in the domain under analysis, including model evaluations, prior to starting any other regular step.

When asked about the activities that usually require the most investment of time or that cause the most difficulties during execution, the reference to the preprocessing phase was almost unanimous. As reasons for that, they mentioned bad quality of the data, lack of data standardization, infrastructure limitation, and mainly the efforts to understand the raw data prior to deciding on any transformations, for instance, data cleaning or the creation of new features. However, for three participants, the preprocessing stage was not highly demanding.

The first works with deep learning with images and their cycle started directly on (3) *Select ML algorithm* and (4) *Train model*, in reference to the ML workflow. The second considered (1) *Business understanding* and (2) *Data understanding*, in reference to CRISP-DM, more demanding. That occurred because they were developing a new solution and were not following the same structure of on-demand projects as most of the other participants. The third worked in a new organization that provides financial services; the company invested in its system architecture since the conception, leading to few data issues and no need to integrate with legacy systems.

Business understanding was the second task indicated as highly demanding because it requires domain expertise and, in some cases, the clients do not know what to ask or look for in their own data. Other items were also mentioned, such as data collection in the case of heterogeneous and complex systems and model deployment in the production environment.

Regarding their data mining strategies, the most indicated were clustering, association, classification, and regression analysis. In addition, many participants mentioned the dimensionality reduction strategy used as part of preprocessing. One participant said that this was not a good strategy for their context and explained that if there are 300 attributes reduced to 10 dimensions, it will be necessary to guarantee all the 300 attributes arrive with quality in the production environment. Then, keeping the model working as planned after deployment adds more complexity to the process. Thus, they preferred to invest in a strategy that only selects the really important attributes. Furthermore, principal component analysis (PCA)

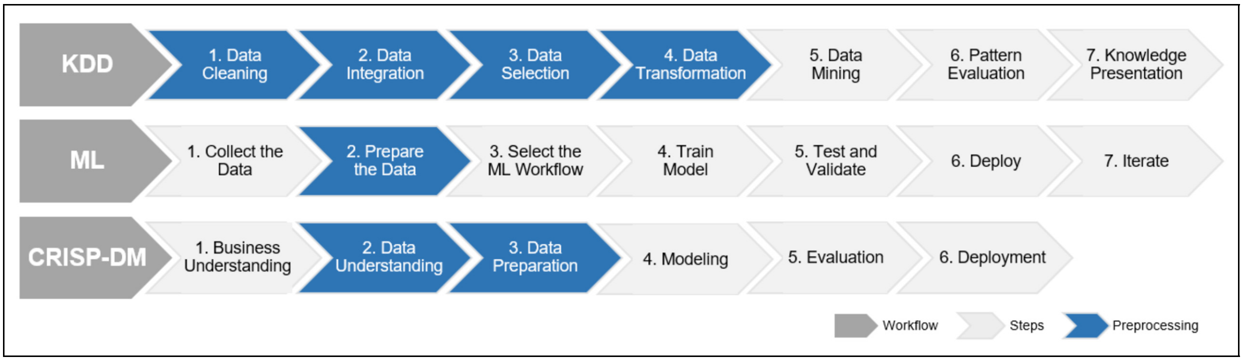


Figure 4. Three examples of workflows used during data analysis: Knowledge Discovery in Databases (KDD),[2] Machine Learning (ML),[17] and Cross-Industry Process for Data Mining (CRISP-DM).[4] The steps highlighted in blue are considered in the scope of the preprocessing phase.

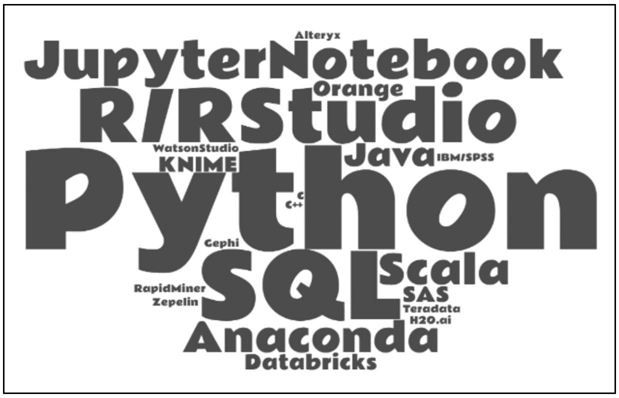


Figure 5. List of the tools and technologies indicated by the 13 participants of our interview study. The larger the text font, the more referenced the item.

was indicated as still useful, but only with the purpose of understanding which attributes are interesting and should be kept, and not with the intention of working with dimensionality reduction in later stages.

The technological basis of the participants was predominantly composed of Python (100%), SQL (69%), and R (54%). These and other technological artifacts are represented in a Word Cloud shown in Figure 5.

Preprocessing activities. Nine participants reported preprocessing activities as laborious since they require a lot of manual intervention. Therefore, they were indicated as highly dependent on professional experience and domain expertise. Although they had already created a particular toolbox of strategies and scripts to make this process easier, the majority of the situations still requires the development of customized scripts to be aligned to the reality of their projects. In this context, Python[18] and R[19] play an important role. Four participants mentioned using tools, such as Databricks,[20] KNIME,[21,22] Gephi,[23,24] and Orange,[25,26] in some moments to support this process. Only one participant said that most of the preprocessing activities were performed directly on Spark.[27,28]

When asked to share further details about the preprocessing tasks, the participant most described, or even emphasized, the following three activities. It is important to notice that the order of each activity is not the same for all participants and may vary according to their project engagement.

1. *Analysis.* Some participants considered a period of time to conduct an assessment of the business area to understand the problem and the data, especially when a domain expert was not involved. They described performing an exploratory analysis of raw data using statistical methods to generate data summaries. Subsequently, behaviors and distributions of these data were evaluated and the next activities were decided based on that. The understanding of how the variables are related was also considered within this exploratory analysis. Another item mentioned was the strategic plan to clean and standardize the data.
2. *Cleaning and standardization of data.* Most participants described performing the general cleaning of the data, trying to ensure the variables are from the same type and other standardizations, for example, data transformation to match the syntax rules defined by the database where newly arrived data are being appended. In addition, few participants reported investing more time in the treatment of missing values, since there is the need to understand, for example, if they are system errors or forms where people do not need to fill in that information or even if they result from an incorrect cross-over during data collection. One participant classified this activity as data enrichment, which could be considered a part of the data quality process.
3. *Feature selection.* They reported evaluating the variables that may be interesting for the model and, from those, deciding the new variables to be created. In addition, some participants indicated they spent considerable time in this activity of categorical variable definition. One participant cited as an example that the cardinality of the variables could be a problem. Since sometimes, the feature binarization is required as a transformation strategy for the ML algorithm, for example, a nominal variable can be encoded using binary attributes by creating a new variable for each of the *n* categories. Then soon there would be a lot of new variables that require tracking, leading to extra complexity. Thus, they indicated the need to be careful to understand which technique is going to be selected for each type of the variable being treated.

Additional challenges and frequent problems were indicated while describing their preprocessing efforts. The next items summarize them.

1. *Data volume and high dimensionality.* Opposite realities were reported: first, a group with a large volume of data and several attributes, for example, 500,000 columns in a table, where such high dimensionality becomes a challenge. On the other side, there were participants who noticed insufficient data, for example, not a minimum number of records to conduct the analysis safely.

2. *Processing time.* Three participants reported some issues with their technical resources, which eventually became the bottleneck for some projects due to waiting time to process their data.
3. *Access to the data.* Another point mentioned was the difficulty to access the data, due to data confidentiality restrictions, owing to particularities of the businesses, such as financial services and healthcare.
4. *Data quality.* Eight participants considered data quality a frequent point of concern. Regarding the most frequent issues, the number one, mentioned by 92% of participants, was missing values (null/empty), followed by missing records (69%), inconsistent–ambiguous data (62%), and incorrect issues, such as duplicates (54%) and outliers/nonstandard (54%). In addition, two participants indicated that the raw data always have problems, such as missing data and outliers. Hence, their starting point is looking for these issues. When they are not present, they then continue the investigation drilling down the specific variable to better understand its behavior. They emphasized this process as very dependent on the knowledge of the analyst performing the activity. Conversely, three participants recognized that they ignore some errors, such as incorrect–duplicated and inconsistent–ambiguous data, depending on the scope of the project and the volume of data.

Visualization of data quality. The beginning of the final part of the questionnaire related to the previous question on data quality issues but focused on how the participants notice these issues. The idea was to acquire further information on the visual identification of data issues, which could be used as a guideline during the development of new visualization techniques. However, when working with the text–numeric type of data, all participants reported the use of scripts to perform the data analysis, for example, generation of the total count of Null per column. Hence, most of them relied primarily on the validation of the absolute numbers, based on their script outputs, rather than on visual exploration or use of any visualization techniques in the process. For unstructured data, for example, audio and images, the participants mentioned the need for a manual inspection.

When using visualization to support their analysis, they mentioned generating graphics, such as barplot, lines, radar plot, boxplot, scatterplot, and histogram, which are available in visualization libraries for Python, for example, Matplotlib[29] and Seaborn,[30] and R, for example, ggplot2.[31] To identify outliers, four participants indicated that boxplot could help to visualize the distribution. Other five participants mentioned the use of additional resources, such as the visualizations available on Hadoop,[32] Orange, Gephi, Databricks, and KNIME.

Five participants emphasized that missing data was the most common problem related to data quality. In addition, they mentioned that tools like SAS[33] can help with the identification of the missing data and even perform transformations automatically. Nevertheless, the solution to this problem cannot be seen so simply, and the validation of these transformations still requires manual inspection. In these cases, one participant said that first they used VIM,[34,35] a graphical user interface available as an R package, to build visualizations to help understand the patterns of these missing values or *NA*s, which stands for Not Applicable, Not Available, or Not Announced.

So we could ask ourselves, what is the reason for them not to use, or use very little, visualization techniques during the process? Three participants argued that it occurs because they were dealing with a very large volume of data, which results in difficulties to visualize the data. In addition, after the solution deployment, the preprocessing must be automatized and cannot be dependent on any manual intervention in the production environment. Then, a visualization could be used only during the initial problem analysis and for model changes. Other three participants mentioned that the choice related to the capacity of the current tools to handle data processing. Free tools, for example, Orange, cannot process huge volumes, being valid only for proof-of-concept purposes. One participant observed that even tools that promise to handle Big Data, for example, Gephi, did not do that in their experience. Moreover, one participant highlighted that even for the most robust tools, which could handle graphic rendering, it was still hard to capture any meaningful information from a crowded visualization if there was too much data.

In addition, five participants stated that generating the visualization was time-consuming. Thus, due to the timeline of the projects, they preferred to invest their time in other activities and then only generate the final visualization that would be shared with the business team and/or clients. One participant also said their current scripting approach, which allowed to look directly at the numbers, was enough, which means there was no need to add any visualization technique during their analysis. Another participant mentioned that they did not know how to use visualization to support preprocessing activities, demonstrating a lack of communication between the visualization research community and the professionals of the enterprise.

In conclusion, the participants were encouraged to mention any visualization techniques or additional features to their current tools that could support their preprocessing activities. Their *wishlist* was considered to build the 10 insights introduced in the next section.

Insights for new visualizations

During our interviews, only one participant mentioned visualization was not a differential for the activities they were performing during preprocessing. Two other participants expressed that they felt confident with their set of tools. However, the 10 remaining participants demonstrated an interest in different ways to explore their data with visualization techniques. Based on these feedbacks and complementary to the discussion started in the previous sections, we present a list of 10 insights for visualization in data exploration in this section.

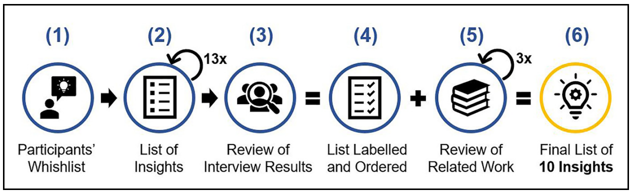


Figure 6. Process to derive the list of 10 Insights.

We compiled the final list of insights following an iterative, incremental coding method, which we explain in the next six steps, also illustrated in Figure 6. (1) The list started based on the inputs received from Participant 1 while explaining his *wishlist*. (2) Every input from a new participant was considered to review the latest version of the list, checking for similarities and complementing the background of the existing items or adding new items to the list. (3) After the completion of the interviews, all the records of the responses were reviewed, including all prior entries, to evaluate if any other item could be added based on the most common inputs, primarily related to challenges and improvement opportunities while describing any particular activity. (4) The items were labeled and ordered from the most to the least frequent. The items that were not mentioned by at least two participants were not included in the final list. (5) We merged the list of recommendations for tool development or design implications available in the related works with the list obtained in Step 4, which resulted in one additional insight. (6) Finally, we ordered the list considering Step 4 for the insights in common with the related work, that is, from Insight 1 to 6, then the insights that were only identified in our study, that is, from Insight 7 to 9, and finally the additional insight not captured by our interviews, that is, Insight 10. In Figure 7, we added details on the list of insights and the correlation of each source that mentioned them.

To simplify the description of the comparison with the related works, we will continue using the following code: RW1 for Batch and Elmqvist,[11] RW2 for Kandel et al.,[12] and RW3 for Alspaugh et al.[13]

1. Keep it simple

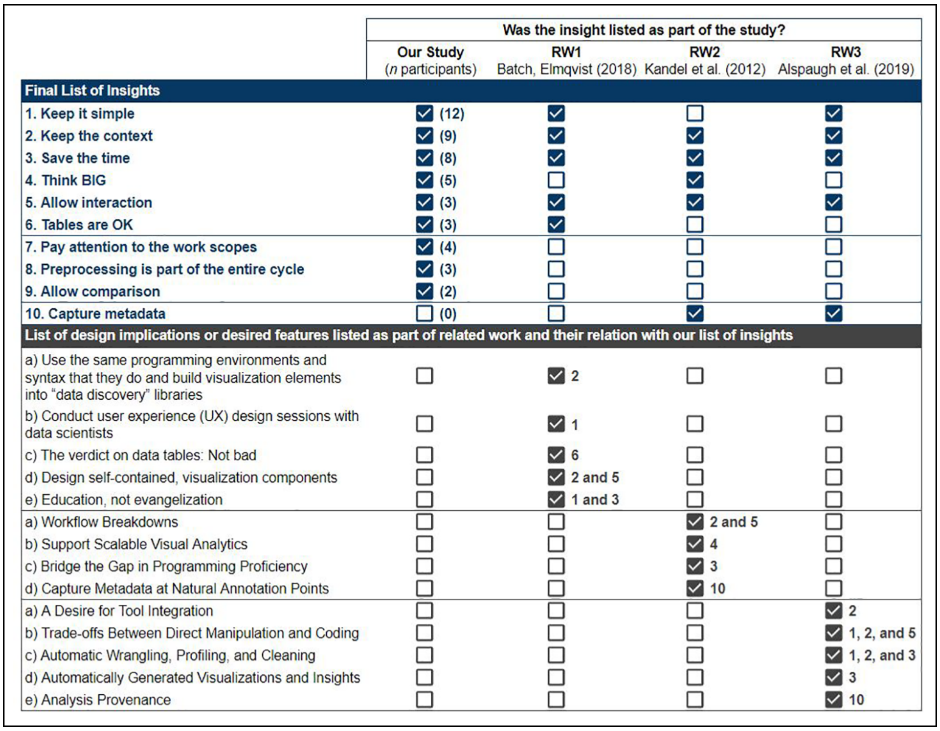
For the majority of the cases, the existing visualizations or more traditional charts should fulfill the demand, without the need for novel visualization techniques, but rather focusing on reusable artifacts and recommendation features according to the type of data and what is intended to be presented. Moreover, even though Python's and R's current visualization packages and libraries are easy to use, they still require some level of programming. Hence, a more ready-to-play alternative, such as Tableau[36] and Qlik,[37] but easier to use, could encourage the use during the preprocessing phase instead of just at the end of the process.

The perception that traditional charts are considered good was only stated by RW1. Moreover, RW1 noticed a lack of usability attention for visualization solutions applied to data mining. Therefore, user experience (UX) design sessions were indicated, and this can support to keep the solution simple for real scenarios use. However, only RW3 objectively mentioned the need for easier tools as desired by data analysts.

2. Keep the context

Any new solution should remain compatible with the most used tools for data mining, currently Python and R, to build an uninterrupted work environment, preventing data analysts from losing the context under investigation while alternating among several different tools. Complementary, RW1 stated that it is important to keep the same syntax of the programming environments used by data analysts. In addition, it indicated the relevance of considering the integration with command line interfaces and of building "visualization elements into data discovery libraries." Although RW2 did not objectively mention it as part of the programming environment, this article referred to the need for visualization tools to avoid the breakdown of the workflows, hence, directly promoting connections to the existing environments. The same was indicated by RW3, which is not focused on the visualization features but was considered important for data exploration tools as a whole.

Furthermore, new tools should allow the evaluation of multiple rows and attributes on the same view, without losing the context under investigation. Thus, there is a need to plan the use of interaction techniques,



	Was the insight listed as part of the study?			
	Our Study (*n* participants)	RW1 Batch, Elmqvist (2018)	RW2 Kandel et al. (2012)	RW3 Alspaugh et al. (2019)
Final List of Insights				
1. Keep it simple	☑ (12)	☑	☐	☑
2. Keep the context	☑ (9)	☑	☑	☑
3. Save the time	☑ (8)	☑	☑	☑
4. Think BIG	☑ (5)	☐	☑	☐
5. Allow interaction	☑ (3)	☑	☑	☑
6. Tables are OK	☑ (3)	☑	☐	☐
7. Pay attention to the work scopes	☑ (4)	☐	☐	☐
8. Preprocessing is part of the entire cycle	☑ (3)	☐	☐	☐
9. Allow comparison	☑ (2)	☐	☐	☐
10. Capture metadata	☐ (0)	☐	☑	☑
List of design implications or desired features listed as part of related work and their relation with our list of insights				
a) Use the same programming environments and syntax that they do and build visualization elements into "data discovery" libraries	☐	☑ 2	☐	☐
b) Conduct user experience (UX) design sessions with data scientists	☐	☑ 1	☐	☐
c) The verdict on data tables: Not bad	☐	☑ 6	☐	☐
d) Design self-contained, visualization components	☐	☑ 2 and 5	☐	☐
e) Education, not evangelization	☐	☑ 1 and 3	☐	☐
a) Workflow Breakdowns	☐	☐	☑ 2 and 5	☐
b) Support Scalable Visual Analytics	☐	☐	☑ 4	☐
c) Bridge the Gap in Programming Proficiency	☐	☐	☑ 3	☐
d) Capture Metadata at Natural Annotation Points	☐	☐	☑ 10	☐
a) A Desire for Tool Integration	☐	☐	☐	☑ 2
b) Trade-offs Between Direct Manipulation and Coding	☐	☐	☐	☑ 1, 2, and 5
c) Automatic Wrangling, Profiling, and Cleaning	☐	☐	☐	☑ 1, 2, and 3
d) Automatically Generated Visualizations and Insights	☐	☐	☐	☑ 3
e) Analysis Provenance	☐	☐	☐	☑ 10

Figure 7. Complete list of the insights. (Top of figure, dark blue box) We present the final list of insights, their frequency in our study, that is, how many participants mentioned it and their connection with other studies. (Bottom of figure, gray box) We present the list of design implications or desired features we could identify in the three related works and their relation to our final list of insights, indicated by the number of the insights.

such as *focus* + *context*, where "a selected subset of the structure (focus) is presented in detail while the rest of the structure is shown in low detail to help the viewer maintain context,"[10] therefore avoiding the *change blindness* effect related to the difficulty to notice changes made during an eye movement.[38]

3. Save the time

Complementing the previous point, the new visualization tools should consider intuitive features and little need for configuration and/or coding, aiming to keep the agility in the working process. Data analysts also regarded the visualization as "too time-consuming to be worth their efforts" during the discussion in RW1. The same was observed in RW3, where the data analysts expressed difficulties around visualizations, such as choosing the right type of chart. Similarly, RW2 discussed this idea as required to "bridge the gap in programming proficiency," since most of the professionals without "hacker" skills, per their study classification, faced difficulties to manipulate data from diverse sources and especially during the wrangling tasks.

Thus, a solution that is embedded into the toolkit of the data analysts and automatically generates some examples or basic templates to support its use and provides recommendations of visualization techniques based on the type of data could be very useful. As a consequence, this approach should avoid some unsuitable uses, such as the use of barplot for time series or line plots for ranking, when they are better in the opposite relation.

4. Think BIG

New visualizations should support scalable solutions, considering Big Data needs. Even though not all participants mentioned this item as critical in their scope (5 of 13, see Figure 7), it is a growing demand, and the development of techniques that can handle this scenario is urged. It was indicated that when dealing with large volumes of data, the data rendering can be complicated even to plot simple visualizations. In that case, different alternatives should be planned, for example, using density or aggregation plotting. Consequently, it should require the evaluation of new strategies, such as data reduction by selecting a sample and server-side preprocessing. The same was discussed in RW2 under the statement "scaling visualization requires addressing both perceptual and computational limitations." RW2 was published in 2012, and this subject remains a critical challenge.

Another alternative is to consider the progressive paradigm, which enables the data analyst to inspect partial results as they become available and interact with the algorithm to prioritize items of interest instead of waiting for full data processing, as explained by Stopler et al.[39] while introducing the progressive visual analytics (PVA).

5. Allow interaction

It is important to provide more than static reports. Moreover, allowing the data analyst to perform flexible data manipulation within visualization tools is fundamental. RW1 indicated the visualization components should enable full-fledged interaction, such as zooming and panning, filtering, and details on demand.[40] It is aligned with the techniques suggested by us in Insight 2, *Keep the context*. As an example, one participant mentioned that a solution similar to Orange UI's proposal, but in a more robust and online version, could contribute to filling this gap, while for RW3 "embedding interactive visualizations within notebook-style" is a better approach considering the emerging trends.

Two good examples of interactive visualization studies in the scope of visual data exploration are VizAssist[41] and VisExemplar.[42] They also planned some assistant features to support with visualization recommendation based on the data analysis needs, which is also related to Insight 1 *Keep it simple*. Concerning preprocessing activities particularities, Heer et al.[43] propose the predictive interaction framework for interactive systems that covers general design considerations for data transformations.

6. Tables are OK

As we could observe during the interviews, most of the participants are still using tabular data during their analysis (see Figure 3). Therefore, aligned with the Insight 1 *Keep it simple*, the tabular format is considered a suitable choice for visual representation. The same was noticed in RW1. Files to store tabular data and structured database tables are widely used. However, there are still opportunities to be explored for table views, such as combining different interaction options and visualization techniques, such as table lens[44] or pixel-oriented.[45]

7. Pay attention to the work scopes

During our interviews, two work scopes were indicated as lacking attention by current visualizations solutions, which remains an opportunity for future works. One concerns the creation of new variables, features, which usually requires a lot of analysis time during preprocessing activities. Thus, the new studies should continue exploring the combination of feature selection techniques[46] with visualization techniques to generate proposals, such as t-Distributed Stochastic Neighbor Embedding (t-SNE).[47]

The other is related to the deep learning scope for visual interpretation of why each decision was made, which is under the scope of studies to support the interpretability of ML.[48,49] In addition, aligned with Insight 5 *Allow interaction*, more interactive visualizations to support the parameterization options are needed, such as Deep playground[50,51] an interactive visualization of neural networks.

8. Preprocessing is part of the entire cycle

For many data mining workflow processes, such as Visual Analytics[52] and KDD,[2] preprocessing is represented as part of a flow in a one-way direction, similarly to a waterfall approach. However, we could notice during the interviews that for most cases multiple interactions were required among preprocessing activities and all the other stages during the same cycle. Except for confirmatory analysis, where most of the process was already automated and little interaction was needed, for other cases, especially for initial data exploration, multiple back and forwards in the raw data occurred.

9. Allow comparison

Considering adding features that allow the comparison of data prior to and after its transformation is important to support the preprocessing decision. It could follow a similar approach as proposed by Kindlmann and Scheidegger,[53] which discussed the importance of knowing whether data transformations respected the original data. Furthermore, one participant mentioned that despite preprocessing activities being very fundamental and at some level performed by all data analysts, few people are truly proficient at them. Hence, this visual support could contribute for more data analysts to adopt visualization as part of their daily strategies, since most of them complained about the difficulties during data cleaning or wrangling activities.

In addition, for the scenarios of ML, support the contrast between the test and train data, and the validation of the model based on different preprocessing strategies. However, during the model testing, "the integration level must be shallow to prevent overfitting and conflation of testing and training data," as observed by Lu et al.[54]

10. Capture metadata

Besides the two previous insights, if automatic exploratory tasks or data transformations are needed, it is important to present the logic underneath them because, as identified by RW2 and RW3, data analysts desired to continue working with control and visibility of what the tool was doing. Thus, the creation of metadata for the dataset under analysis and data preparation are fundamental to this process.

Moreover, this metadata can be added to the data mining project documentation, helping to build the principle of transparency on activities performed, which is aligned to initiatives, such as the European Union General Data Protection Regulation.[55]

11. Discussion

The last insight presented in our list, 10 *Capture Metadata*, was the only one seen in the related works that was not captured during our interviews. However, the Insights 7 *Pay attention to the work scopes*, 8 *Preprocessing is part of the entire cycle*, and 9 *Allow comparison* in our list were not mentioned by any of the indicated related works, which brings new topics for discussion. Moreover, none of the other insights appeared together in the final list of recommendations or implications for design, as shown in Figure 7.

Although RW1 was very well organized, introducing relevant points to this discussion, an important item related to the need for scalable solutions, Insight 4 *Think BIG*, was not listed in its final implications for design. Similarly, despite RW2 being one of the first studies addressing this subject and reporting important perceptions from enterprise data analysis, it still did not cover our entire list, nor did it present its design implications in an approach that is as straightforward as ours. Besides, it was not concerned with the particular needs of data mining workflows. While RW3 also contributed with this discussion, their primary focus was neither visualization nor preprocessing activities in data mining. Thus, many of its recommendations covered data exploration at a higher level of the process than ours.

In terms of the evaluation of the usability, von Zernichow and Roman[56] explored approaches of visual data profiling in tabular data cleaning and transformation processes. While validating their software prototype, they identified usability issues and suggestions for further research that also can be related to our list of insights, as, for example, visual-recommend system approaches to suggest relevant and domain-specific charts to the user (Insight 1 *Keep it simple* and 3 *Save the time*) and explore direct table manipulation (Insight 2 *Keep the context* and 6 *Tables are OK*).

In addition to the discussion of the 10 insights, two preprocessing activities can be mentioned as relevant to be considered in the scope of future studies: data dependency and data normalization. A couple of references can be mentioned covering data transformation aiming data normalization and using visual interaction, such as Profiler[57] and Wrangler,[5] which later resulted in a commercial solution named Trifacta.[58] Likewise, Tableau Prep[59] provides a visual and direct way to combine, shape, and clean data. Tableau Prep is part of Tableau ecosystem, and it is comprised two products: “Builder” for building data flows and “Conductor” for scheduling, monitoring, and managing flows. Even though these are relevant references, there are still opportunities to discuss, for instance, how to integrate these proposals under the most used tools for data analysis? How to explore the comparison of data transformations decisions with the impact in the data mining model building? Which visualization techniques can be used to support data quality?

As summarized in Figure 8, we hope to contribute with a straight and easy-to-understand list of items that require attention when planning new visualization solutions as part of the alternatives to lower adoption barriers. Moreover, despite our focus on the preprocessing phase for many of our questions, we consider these insights are also applicable to other phases of the data mining workflow, which includes the final visualizations used to report the analysis and findings.

Limitations

With respect to opportunities for improving our study, we can list two main items. First regarding the procedure, the number of questions was designed to guarantee that each interview session would take no longer than 1 h, in an attempt to capture a higher number of positive returns to our participation invitation. However, a more open strategy for data collection such as an experiment where participants are instructed to perform a list of tasks and it is possible to observe how they deal with them to solve certain problems could contribute to acquire further details about daily practices. Likewise, that approach would require an additional number of hours, at least 2 h for each participant session based on RW1 study and possibly reducing the list of participants available to join the activity.

The second opportunity is regarding the participant’s profile. Most of our interviewees were working in the IT Industry. Additional participants from different organization structures could contribute to a different perspective. Also, we notice lack of female representation but that seems to be a bigger issue in the Science, Technology, Engineering, and Mathematics (STEM) areas. Therefore, despite our efforts to recruit a variety of participants, the data collected and its analysis cannot be considered a representation of all data analysts.

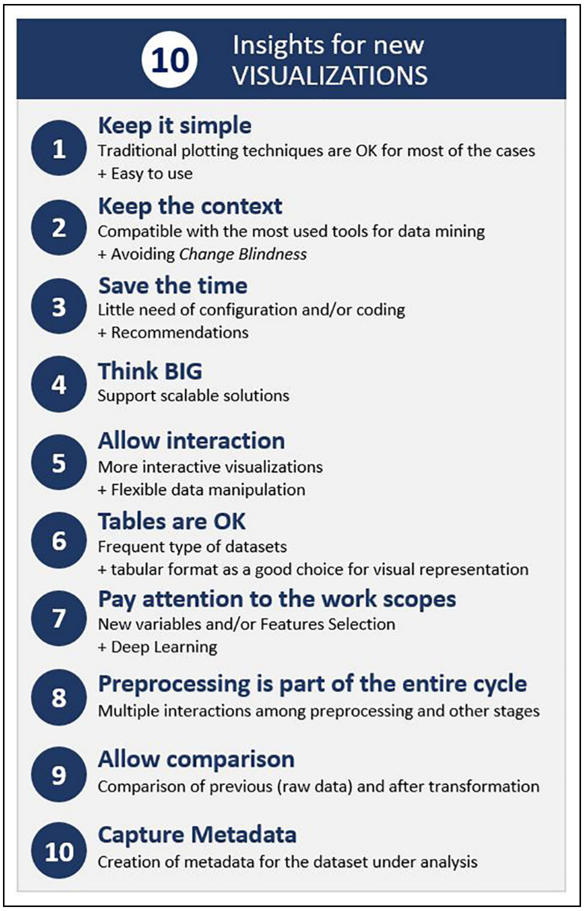
Conclusion and future work

We interviewed 13 enterprise professionals to understand their data analysis practices in data mining, how they use visualization during the preprocessing phase and which features could support them during this process. In addition, we presented the methodology used for data collection in this interview study and the results obtained from the interviews.

Our main contribution was the organization of the challenges and opportunities identified during our analysis of the interviews, which resulted in a list of 10 insights. This list of insights was then compared with the closest related works, improving the reliability of our findings, and, at the same time, encouraging the discussion about uncovered considerations.

Even though some insights appeared in previous studies, an in-depth analysis of the related works was necessary to identify and relate their findings to our final list of insights. Through our study, we also summarized practical items to be considered during the planning and development stages of new visualization solutions, aiming to lower the barriers to adopt visualization as part of any data mining workflow. Ultimately, this study contributes as a source of requirements to fill the visualization gap during the data understanding, exploration, and preparation in early phases.

As future work, regarding the interview study procedure, we can list two main items: (1) in-depth interviews or user-centered experiments to further investigate visualization alternatives; (2) extend participants' profile by considering professionals who are doing data analysis and do not have a solid foundation in computer science or statistics. Based on their different perspectives of data analysis, evaluate if new insights should be considered as part of the requirements. Also, while contemplating the requirements elicited by our study, several future work opportunities arise. One possibility is to develop preliminary prototype systems considering the list of insights and then evaluate the prototypes while conducting in-depth interviews or user-centered experiments with the participation of domain experts.



10 Insights for new VISUALIZATIONS

1 Keep it simple
Traditional plotting techniques are OK for most of the cases
+ Easy to use

2 Keep the context
Compatible with the most used tools for data mining
+ Avoiding *Change Blindness*

3 Save the time
Little need of configuration and/or coding
+ Recommendations

4 Think BIG
Support scalable solutions

5 Allow interaction
More interactive visualizations
+ Flexible data manipulation

6 Tables are OK
Frequent type of datasets
+ tabular format as a good choice for visual representation

7 Pay attention to the work scopes
New variables and/or Features Selection
+ Deep Learning

8 Preprocessing is part of the entire cycle
Multiple interactions among preprocessing and other stages

9 Allow comparison
Comparison of previous (raw data) and after transformation

10 Capture Metadata
Creation of metadata for the dataset under analysis

Figure 8. Consolidated list of insights for new visualizations solutions.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We want to acknowledge that this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Also, this study was achieved in cooperation with Hewlett Packard Brasil LTDA. and HP Brasil Indústria e Comércio de Equipamentos Eletrônicos LTDA. using incentives of Brazilian Informatics Law (Law n° 8.2.48 of 1991). Lastly, the Emerging Leaders in the Americas Program (ELAP) with the support of the Government of Canada.

ORCID iDs

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

References

1. Dasu T and Johnson T. *Exploratory data mining and data cleaning*. 1st ed. New York: John Wiley & Sons, 2003.
2. Han J, Kamber M and Pei J. *Data mining: concepts and techniques*. 3rd ed. San Francisco, CA: Morgan Kaufmann Publishers, 2011.
3. Piateski G and Frawley W. *Knowledge discovery in databases*. Cambridge, MA: MIT Press, 1991.
4. Shearer C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehous* 2000; 5(4): 13–22.
5. Kandel S, Paepcke A, Hellerstein J, et al. Wrangler: interactive visual specification of data transformation scripts. In: *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '11*, Vancouver, BC, Canada, 7–12 May 2011, pp. 3363–3372. New York: ACM.
6. Hellerstein JM. Quantitative data cleaning for large databases, 2008, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6419
7. Jugulum R. *Competing with high quality data: concepts, tools, and techniques for building a successful approach to data quality*. New York: John Wiley & Sons, 2014.
8. Tan PN, Steinbach M and Kumar V. *Introduction to Data mining*. 1st ed. Boston, MA: Addison-Wesley, 2005.
9. Ferreira de, Oliveira MC and Levkowitz H. From visual data exploration to visual data mining: a survey. *IEEE T Vis Comput Gr* 2003; 9(3): 378–394.
10. Ward MO, Grinstein G and Keim D. *Interactive data visualization: foundations, techniques, and applications, second edition - 360 degree business*. 2nd ed. Natick, MA: A. K. Peters, 2015.
11. Batch A and Elmqvist N. The interactive visualization gap in initial exploratory data analysis. *IEEE T Vis Comput Gr* 2018; 24(1): 278–287.
12. Kandel S, Paepcke A, Hellerstein JM, et al. Enterprise data analysis and visualization: an interview study. *IEEE T Vis Comput Gr* 2012; 18(12): 2917–2926.
13. Alspaugh S, Zokaei N, Liu A, et al. Futzing and moseying: interviews with professional data analysts on exploration practices. *IEEE T Vis Comput Gr* 2019; 25: 22–31.
14. Lam H, Bertini E, Isenberg P, et al. Empirical studies in information visualization: seven scenarios. *IEEE T Vis Comput Gr* 2012; 18(9): 1520–1536.
15. Lazar J, Feng JH and Hochheiser H. *Research Methods in Human-Computer Interaction*. 2nd ed. Amsterdam: Elsevier, 2017.
16. De Mauro A, Greco M and Grimaldi M. What is big data? a consensual definition and a review of key research topics. *AIP Conf Proc* 2015; 1644: 97–104.
17. Altexsoft. Machine learning: bridging between business and data science, 2019, https://www.altexsoft.com/whitepapers/machine-learning-bridging-between-business-and-data-science/
18. Python. Python, 2018, https://www.python.org/
19. R. The R project for statistical computing, 2018, https://www.r-project.org/
20. Databricks. Databricks: making big data simple, 2018, https://databricks.com/
21. Berthold MR, Cebron N, Dill F, et al. KNIME—the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor Newsl* 2009; 11(1): 26–31.
22. KNIME. KNIME: open for innovation, 2018, https://www.knime.com/
23. Bastian M, Heymann S and Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: *Proceedings of the Third international AAAI conference on weblogs and social media*, San Jose, CA, 17–20 May 2009, pp. 361–362. Menlo Park, CA: AAAI.
24. Gephi. The Open Graph Viz Platform, 2018, https://gephi.org/
25. Demšar J, Curk T, Erjavec A, et al. Orange: data mining toolbox in python. *J Mach Learn Res* 2013; 14: 2349–2353.
26. Orange. Orange: data mining fruitful and fun, 2018, https://orange.biolab.si/
27. Zaharia M, Chowdhury M, Franklin MJ, et al. Spark: cluster computing with working sets. In: *Proceedings of the 2nd USENIX conference on hot topics in cloud computing*, Boston, MA, 22–25 June 2010, p. 10. Berkeley, CA: USENIX.
28. Spark A. Apache spark: unified analytics engine for big data, 2018, https://spark.apache.org/
29. Matplotlib. Matplotlib: Python plotting—Matplotlib 3.0.2 documentation, 2018, https://matplotlib.org/
30. Seaborn. seaborn: statistical data visualization - 0.9.0 documentation, 2018, https://seaborn.pydata.org/
31. ggplot2. ggplot2—tidyverse, 2018, https://ggplot2.tidyverse.org/
32. Hadoop A. Apache Hadoop, 2018, https://hadoop.apache.org/
33. SAS. SAS analytics, 2018, https://www.sas.com/
34. Kowarik A and Templ M. Imputation with the R package VIM. *J Stat Softw* 2016; 74(7): 1–16.
35. Templ M, Alfons A, Kowarik A, et al. Vim: visualization and imputation of missing values, 2018, https://cran.r-project.org/web/packages/VIM/index.html
36. Tableau. Tableau, 2018, http://www.tableau.com/
37. Qlik. Qlik: data analytics for modern business intelligence, 2018, https://www.qlik.com
38. Rensink RA. Seeing, sensing, and scrutinizing. *Vision Res* 2000; 40(10): 1469–1487.
39. Stolper CD, Perer A and Gotz D. Progressive visual analytics: user-driven visual exploration of in-progress analytics. *IEEE T Vis Comput Gr* 2014; 20(12): 1653–1662.
40. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE symposium on visual languages. VL '96*, Boulder, CO, 3–6 September 1996, pp. 336–343. Washington, DC: IEEE.
41. Bouali F, Guettala A and Venturini G. Vizassist: an interactive user assistant for visual data mining. *Vis Comput* 2016; 32(11): 1447–1463.
42. Saket B, Kim H, Brown ET, et al. Visualization by demonstration: an interaction paradigm for visual data exploration. *IEEE T Vis Comput Gr* 2017; 23(1): 331–340.

43. Heer J, Hellerstein JM and Kandel S. Predictive interaction for data transformation, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.692.1613
44. Rao R and Card SK. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In: *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '94*, Boston, MA, 24–28 April 1994, pp. 318–322. New York: ACM.
45. Keim D. Designing pixel-oriented visualization techniques: theory and applications. *IEEE T Vis Comput Gr* 2000; 6(1): 59–78.
46. Guyon I and Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003; 3: 1157–1182.
47. Maaten L and Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; 9: 2579–2605.
48. Bratko I. Machine learning: between accuracy and interpretability. In: Riccia GD, Lenz HJ and Kruse R (eds) *Learning, networks and statistics*. New York: Springer, 1997, pp. 163–177.
49. Molnar C. Interpretable machine learning, 2019, https://christophm.github.io/interpretable-ml-book/
50. TensorFlow. A neural network playground, 2019, https://playground.tensorflow.org
51. Smilkov D, Carter S, Sculley D, et al. Direct-manipulation visualization of deep networks, 2017, arXiv:1708.03788
52. Keim D, Kohlhammer J and Ellis G. *Mastering the information age: solving problems with visual analytics*. Goslar: Eurographics Association, 2010.
53. Kindlmann G and Scheidegger C. An algebraic process for visualization design. *IEEE T Vis Comput Gr* 2014; 20(12): 2181–2190.
54. Lu Y, Garcia R, Hansen B, et al. The state-of-the-art in predictive visual analytics. *Comput Graph Forum* 2017; 36(3): 539–562.
55. Commission E. Eu general data protection regulation, 2019, https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en
56. von Zernichow BM and Roman D. Usability of visual data profiling in data cleaning and transformation. In: *Proceedings of the OTM Confederated International Conferences "On the Move to Meaningful Internet Systems,"* Rhodes, 23–28 October 2017, pp. 480–496. New York: Springer.
57. Kandel S, Parikh R, Paepcke A, et al. Profiler: Integrated statistical analysis and visualization for data quality assessment. In: *Proceedings of the international working conference on advanced visual interfaces, AVI '12*, Capri Island, 21–25 May 2012, pp. 547–554. New York: ACM.
58. Trifacta. Trifacta data wrangling tools & software, 2018, https://www.trifacta.com/
59. Tableau. Tableau prep, 2019, https://www.tableau.com/products/prep

Appendix 1

Data collection instrument—Questionnaire developed for semi-structured interview.

Part I—Questions to map the participant profile.

1. What is your work location (country/city)?
2. What is your gender/sex?
3. What is your age?
4. What is your education? Which area?
5. What is the place of work?
6. Which area/department?
7. What is your official title/role in this organization?
8. How much time of experience in the area of technology?
9. How much experience with preparation and/or preprocessing of data?

Given your most recent data analysis, please answer the following questions.

Part II—Questions to identify the data profile.

10. What are the sources of this data?
11. What is the format of this data?
12. What types of data were used?
13. What is the volume?

Part III—Questions to identify the process involved in data analysis.

14. What are the main activities/tasks performed in the data analysis process?

For this question, three workflow examples were introduced as described in Figure 4.

15. Which of these activities (mentioned in question 14) do you consider need to invest more time and/or have more difficulties to achieve? Why?
16. What strategies/techniques have been used for data mining and/or machine learning?

For this question, five examples were introduced: Anomaly Detection, Clustering or Association Analysis, Classification, Regression, Dimensionality Reduction, and others-please list what else.

17. Development environment/technology/platform.

For this question, 18 examples were introduced: Java, Python, R, Scala, SQL, Weka, Orange, Jupyter

Notebook, KNIME, Databricks, Dataiku, IBM/SPSS, SAS, RapidMiner, Alteryx, Anaconda, H2O.ai, Teradata, and others.

Part IV—Questions to identify data preparation and/or preprocessing activities.

18. How did you prepare and/or preprocess this data before transforming it or running any ML algorithms?
19. Do you use any tools to assist you in the preparation and/or preprocessing of data?

[YES] Which ones? What is the purpose of each? Why were they chosen?

[NO] Have you used any? [YES] Why did you stop? [NO] Why do not you use it?

20. What are the biggest challenges (or recurrent problems) faced during the data preparation process?
21. What are the key data quality issues faced during the preparation process?

For this question, six examples were introduced: missing-missing record, missing-missing value (null/empty), inconsistent-measurement units, inconsistent-ambiguous data, inconsistent-misspelling, incorrect-duplicated, incorrect-outliers (non-standard data), and others-please list what else.

Part V—Questions related to how they visualize the data quality issues and to identify visualization techniques used.

22. Considering the following problems (listed be same as in question 21), what is important to understand to identify the problem? How do you visualize/perceive if they are present?
23. Does the tool you have use during the preparation or preprocessing of the data provide some visualization technique to support the interpretation of the data?

[YES] What would they be? Which ones do you use? Why?

24. In your opinion, what types of analysis should the visualization tool support in data preparation activities?
25. Is there any additional visualization technique that you think might support this process?

As a wrap-up question, the participants were instigated to answer which are the features they would consider as part of their *wishlist*.