

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

ALESSANDRA MACIEL PAZ MILANI

PREPROCESSING PROFILING MODEL FOR VISUAL ANALYTICS

Porto Alegre

2019

PÓS-GRADUAÇÃO - *STRICTO SENSU*



Pontifícia Universidade Católica
do Rio Grande do Sul

**PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM**

**PREPROCESSING PROFILING
MODEL FOR VISUAL ANALYTICS**

ALESSANDRA MACIEL PAZ MILANI

Dissertation submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. PhD. Isabel Harb Manssour
Co-Advisor: Prof. PhD. Fernando Vieira Paulovich

**Porto Alegre
2019**

Ficha Catalográfica

M637p Milani, Alessandra Maciel Paz

Preprocessing Profiling Model for Visual Analytics / Alessandra Maciel Paz Milani . – 2019.

119 p.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profª. Dra. Isabel Harb Manssour.

Co-orientador: Prof. Dr. Fernando Vieira Paulovich.

1. Visual Analytics. 2. Visualization Techniques. 3. Data Mining. 4. Preprocessing. I. Manssour, Isabel Harb. II. Paulovich, Fernando Vieira. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Salete Maria Sartori CRB-10/1363

Alessandra Maciel Paz Milani

Preprocessing Profiling Model for Visual Analytics

This Dissertation has been submitted in partial fulfillment of the requirements for the degree of Master of Computer Science, of the Graduate Program in Computer Science, School of Technology of the Pontifícia Universidade Católica do Rio Grande do Sul.

Sanctioned on August 29, 2019.

COMMITTEE MEMBERS:

Prof. Dra. Maria Cristina Ferreira de Oliveira
(ICMC/USP)

Prof. Dr. Duncan Dubugras Alcoba Ruiz
(PPGCC/PUCRS)

Prof. Dr. Fernando Vieira Paulovich
(Dalhousie University – Co-advisor)

Prof. Dra. Isabel Harb Manssour
(PPGCC/PUCRS - Advisor)

To my beloved Leonardo.

“The greatest value of a picture is when it forces us to notice what we never expected to see.”
(John Tukey)

ACKNOWLEDGMENTS

Many people contributed to the completion of this master's dissertation, and at this point, I want to express my sincere appreciation to all of them.

To begin, I want to express my special thanks to Prof. Isabel as my supervisor. She always provided me with valuable feedback, learning opportunities, and guided me through all phases of my master's course, which introduced me to new possibilities in the academic research. In the same way, to Prof. Fernando, my co-supervisor, for essential constructive feedback and insightful comments that helped me to shape and fine-tune this dissertation.

To the staff, students and teachers of PUCRS for the experiences shared throughout the master's course. Especially to colleagues Olimar, Henrique, and Juliana, for support and feedback, and to Prof. Sabrina for all the knowledge shared and encouragement. Likewise, I must express my gratitude to the staff at Dalhousie University, and to the friends I met there who helped me go through the harsh Canadian winter in the warmest way.

I also want to thank Lucas for the support with the development of our prototype tool, and to all participants of the interview study for their valuable inputs. Both essential contributions to build this dissertation.

To family and friends who understood my absence during these 2.5 years of dedication. Especially to my mother Mirna and my grandmother Guta, examples of strong and wise women, who have inspired me to remain curious and pursue my studies as the best strategy for achieving my goals.

To Leonardo for the unconditional partnership, patience, and love. Without him, my master's course would not have been possible.

To conclude, I 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 nº 8.248 of 1991). Lastly, with the support of the Government of Canada during the period of my academic exchange with the Emerging Leaders in the Americas Program (ELAP).

PREPROCESSING PROFILING MODEL FOR VISUAL ANALYTICS

ABSTRACT

In the information age, we have evolved the ability to collect and store data, create sophisticated data mining methods, and generate rich visualizations to share the information resulting from the data analysis process. However, analyzing and managing raw data is still a challenging part of this process, mainly with regards to data preprocessing, which aims to transform this raw data into an appropriate format for subsequent analysis. Although we can find studies proposing design implications or recommendations for future visualization solutions in the data analysis scope, they do not focus on the challenges during the Preprocessing phase and on how visualization can support it. Likewise, the current Visual Analytics Models are not considering preprocessing an equally important phase in their process, such as Data, Models, Visualization, and Knowledge. Thus, with this study, we aim to contribute to the discussion of how we can use and combine methods of visualization and data mining to assist data analysts during the preprocessing activities. To achieve that, we are introducing the Preprocessing Profiling Model for Visual Analytics, which contemplates a set of features to inspire the implementation of new solutions. In turn, these features were designed considering a list of insights we obtained during an interview study with thirteen data analysts. The main contributions in our study are three: (a) the Preprocessing Profiling Model for Visual Analytics as a solution to assist during Preprocessing phase. (b) The list of ten insights, as a consolidated set of requirements for future visualization research studies applied to preprocessing and data mining. (c) The 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: Visual Analytics, Visualization Techniques, Data Mining, Preprocessing.

MODELO DE PRÉ-PROCESSAMENTO PARA ANÁLISE VISUAL

RESUMO

Na era da informação, desenvolvemos a capacidade de coletar e armazenar dados, criar métodos sofisticados de mineração de dados e gerar visualizações ricas para compartilhar as informações resultantes do processo de análise de dados. No entanto, analisar e gerenciar dados brutos ainda é uma parte desafiadora desse processo, principalmente no que diz respeito ao pré-processamento de dados, que visa transformar esses dados brutos em um formato apropriado para análises subsequentes. Embora possamos encontrar estudos propondo implicações ou recomendações para futuras soluções de visualização no escopo da análise de dados, eles não se concentram nos desafios da fase de pré-processamento, nem em como a visualização pode suportá-la. Da mesma forma, os modelos atuais de análise visual não consideram o pré-processamento como uma fase igualmente importante em seus processos. Assim, com este estudo, pretendemos contribuir para a discussão de como podemos usar e combinar métodos de visualização e mineração de dados para auxiliar os analistas de dados durante as atividades de pré-processamento. Para isso, apresentamos um modelo de pré-processamento com análise visual, que contempla um conjunto de recursos para inspirar a implementação de novas soluções. Por sua vez, esses recursos foram projetados considerando uma lista de ideias (*Insights*) que obtivemos durante um estudo de entrevista com treze analistas de dados. As principais contribuições de nosso estudo são três: (a) o modelo de análise visual para auxiliar durante a fase de pré-processamento. (b) A lista de dez *Insights*, como um conjunto consolidado de requisitos para futuros estudos de pesquisa de visualização aplicados ao pré-processamento e à mineração de dados. (c) Os detalhes sobre o perfil dos analistas de dados, os principais desafios que eles enfrentam e as oportunidades que surgem enquanto eles estão envolvidos em projetos de mineração de dados em diversas áreas da organização.

Palavras-Chave: Análise Visual, Técnicas de Visualização, Mineração de Dados, Pré-Processamento.

LIST OF FIGURES

Figure 2.1 – Overview of the main concepts and terms presented in Background.	29
Figure 2.2 – The Analytical Problems can be solved using Automatic Analysis, Visualization, or the combination of both, Visual Analytics [KMT10].	33
Figure 2.3 – Three examples of workflows used during data analysis: Knowledge Discovery in Databases (KDD) [HKP11], Machine Learning (ML) [Altb], and Cross Industry Process for Data Mining (CRISP-DM) [She00]. The steps highlighted in blue are considered in the scope of the preprocessing phase.	35
Figure 2.4 – Example of taxonomy of data quality issues combined with methods for detecting each issue and visualizations for assessing their output [KPP+12].	35
Figure 3.1 – Research Design: list of the main activities.	39
Figure 4.1 – The inclusion criteria for each analyzed study was progressive until Step 4. For Steps 5 and 6, we changed two of the prior inclusion criteria. First, for Step 5, while searching for new references on the list featured in RW1 [BE18], the year of publication was unlimited, which allowed the selection of RW2 [KPHH12], from 2012. Second, for Step 6, while searching for citations, studies which contributed with the perceptions of professional data analysts on the data exploration process were selected even if the study did not focus on visualization. Hence, a third study was selected, RW3 [AZL+19]. Among the venues for crucial prior studies, three digital libraries, i.e., ACM, IEEE, and Springer, were selected aiming to cover the most relevant journals and conferences in our research scope, in addition to studies that went through a rigorous review process.	42
Figure 4.2 – Overview of the Interview Process followed during our study.	43
Figure 4.3 – Additional information on the profile of the thirteen participants.	45
Figure 4.4 – Process to derive the list of 10 Insights.	52
Figure 4.5 – Complete list of the insights. (Top of figure, dark blue box) We present the final list of insights, their frequency in our study, i.e., 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 insight.	52
Figure 4.6 – Consolidated list of insights for new visualizations solutions.	58
Figure 5.1 – The Visual Analytics process proposed by Keim et al. [KKE10]	60

Figure 5.2 – Overview of the Preprocessing Profiling Model for Visual Analytics process. We are extending the VA process proposed by Keim et al. [KKE10]. Each node (represented through rounded rectangle) corresponds to a different phase, and their transitions are represented through arrows. 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.	67
Figure 5.3 – The Preprocessing Profiling Model Architecture. List of components to be considered while developing a new solution. The Preprocessing Profiling features are indicated in the blue box on the right, and the list of insights is indicated on the bottom of the image.	69
Figure 5.4 – The sequence of steps in the Visual Analytics process.	70
Figure 5.5 – Example of Python programming code to run the report.	71
Figure 5.6 – High-level comparison between the (a) original Pandas Profiling Report and (b) Data Profiling Report. The orange boxes indicate new items implemented. In both examples, the Iris dataset is used as input.	72
Figure 5.7 – Data Profiling Report - Variables. Information of a Numeric Variable, with (a) the missing values frequency bar. Also, details for each tab: (b) Quantile and Descriptive Statistics; (c) Histogram; (d) Frequency of the most common values; (e) Extreme Values, five Minimum and five Maximum are listed; (f) Boxplot.	73
Figure 5.8 – Data Profiling Report - Missing Values. Four visualizations are presented, one per tab: (a) Matrix, (b) Barplot, (c) Heatmap, and (d) Dendrogram. Output generated based on Cervical Cancer dataset.	74
Figure 5.9 – Data Profiling Report - Correlations. Output generated based on Cervical Cancer dataset. (a) Pearson result. (b) Table with complete information.	75
Figure 5.10 – Prototype A - Data Profiling. Fifth Section, Sample of the first lines of the dataset.	75
Figure 5.11 – Data Profiling Report - Relation Matrix. Pairplot to visualize the similarities and differences between the species.	76
Figure 5.12 – Example of Facets Dive [Goo], using their pre-loaded dataset example.	76
Figure 5.13 – Preprocessing Profiling Report. Menu of content and the Details for the Cervical cancer dataset with missing values from the original file.	78
Figure 5.14 – Preprocessing Profiling Report - Overview of Classification Results for Cervical Cancer dataset.	79

Figure 5.15 – Preprocessing Profiling Report - Results of Classification of each Imputation Strategy based on Iris dataset. Six visualizations are presented: (a) Classification Report, (b) Confusion Matrix, (c) Error Distribution, (d) Precision Recall Curves, (e) Precision Recall Curves (Individually), and (f) Flow of Classes.	80
Figure 5.16 – Preprocessing Profiling Report - Flow of Classes per round of pre-processing strategy.	81
Figure 5.17 – Preprocessing Profiling Report - Matrix of Nullity combined with Class Prediction Error	81
Figure 5.18 – A high-level illustration of the prototype coverage in comparison to the Preprocessing Profiling Model Architecture. For each corresponding component indicated in the Architecture (see Figure 5.3 for details), we are adding a box with a list of items implemented in the scope of our prototypes. The “(…)” indicates that more items could be listed. Also, there is the indication of the features that are covered (indicated by the green sign) or not (the red sign). For the last feature in the list, (I) Interaction, we are considering it implemented partially, as our understanding is that more user interaction options should be implemented to have a full advantage of the interactivity benefits.	83
Figure 6.1 – Tim’s steps for Preprocessing Profiling phase and an illustrative sample of the correspondent visualizations for each step.	86
Figure 6.2 – Data Profiling Report - Overview section for Iris dataset. (a) Warning information regarding petal_width column high correlation with petal_lenght. That is the reason why one variable appears in Rejected status on the Variable Types breakdown.	86
Figure 6.3 – Relation Matrix: pairplot to visualize the similarities and differences between the species.	87
Figure 6.4 – Missing Values Matrix with the dirty dataset for Iris.	88
Figure 6.5 – Relation Matrix: pairplot to visualize the similarities and differences between the species.	88
Figure 6.6 – Data Profiling Report - Variable section with detailed information for sepal_length variable.	89
Figure 6.7 – Preprocessing Profiling Report - Baseline results for the classification of Iris original dataset.	90
Figure 6.8 – Preprocessing Profiling Report - Classification Results for different Missing Values Imputation for Iris dataset with dirty data. The classes are identified as Set (blue) for Iris Setosa, as Ver (orange) for Iris Versicolor, and Vir (green) for Iris Virginica.	91

Figure 6.9 – Preprocessing Profiling Report - Flow of Classes Visualization. Classification results for different missing values imputation strategies for Iris dataset with dirty data.	92
Figure 6.10 – Data Profiling - Section Overview. Detail of the Mammographic Masses dataset. (a) Dataset information with columns, rows, and the size of the dataset. (b) Variable types distribution. (c) Missing values distribution and breakdown of types identified in the dataset. (d) Warnings list.	93
Figure 6.11 – Data Profiling - Section Variables. Detail on the first three columns of the Mammographic Masses dataset.	93
Figure 6.12 – Data Profiling - Section Variables. Detail of the first column of the Mammographic Masses dataset. (a) Statistics for the variable. (b) Common Values in details highlighting the value 55.0 with one occurrence. (c) Boxplot in details for the same variable.	94
Figure 6.13 – Data Profiling - Section Missing Values. Detail of nullity matrix of the Mammographic Masses dataset.	94
Figure 6.14 – Data Profiling - Section Correlations. Details of Spearman for Mammographic Masses dataset.	95
Figure 6.15 – Preprocessing Profiling - Classification Results for different Missing Values Imputation for Mammographic Masses dataset.	96
Figure 6.16 – Preprocessing Profiling - Comparison of Classification Results for different Missing Values Imputation for Mammographic Masses dataset. (a) Baseline, missing values removed. (b) Mean Imputation.	97
Figure A.1 – Example of visualization techniques used in studies on visual data exploration: (a) Boxplot [Ora]; (b) Matrix with Dense Pixel [KKA95]; (c) Radial Graph combined with Histogram [AHH ⁺ 14]; (d) Heatmap combined with Line chart [WFW ⁺ 17]; (e) Scatterplot combined with Glyphs [KPB14]; (f) Sankey [ACF ⁺ 16]; (g) Scatterplot [WMA ⁺ 16]; (h) Parallel Coordinates [BGV16]; (i) Treemap [KPS16].	111
Figure A.2 – Example of a decision tree to select the most appropriate visualization technique for numeric data [HH].	112
Figure A.3 – Example of a decision tree to select the most appropriate visualization technique for categorical data [HH].	113
Figure C.1 – Consent form used on the interview study - page 1 of 2.	117
Figure C.2 – Consent form used on the interview study - page 2 of 2	118

LIST OF TABLES

Table 5.1 – Is the study presenting details on the following items? (1) Process or Model or Workflow or Pipeline; (2) Preprocessing is considered an explicit phase on the workflow; (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 understanding.	64
Table 5.2 – List of the nine features and respective descriptions.	66
Table 6.1 – List of attributes and respective descriptions for the Mammographic Masses dataset.	95
Table D.1 – List of the main technologies used for the prototype development.	119

LIST OF ACRONYMS

CRISP-DM – Cross-Industry Standard Process for Data Mining

DM – Data Mining

DMBOK – Data Management Body of Knowledge

IV – Information Visualization

KDD – Knowledge Discovery in Databases

MAR – Missing at random

MCAR – Missing completely at random

ML – Machine Learning

NMAR – Not missing at random

PVA – Predictive Visual Analytics

REC – Research Ethics Committee

SRQ – Secondary research questions

STEM – Science, Technology, Engineering, and Mathematics

VA – Visual Analytics

CONTENTS

1	INTRODUCTION	25
2	BACKGROUND	29
2.1	DATA	29
2.2	DATA MINING	30
2.3	INFORMATION VISUALIZATION	31
2.4	VISUAL ANALYTICS	32
2.5	DATA MANAGEMENT	33
2.6	DATA QUALITY	33
2.7	PREPROCESSING	34
3	METHODOLOGY	37
3.1	RESEARCH QUESTIONS	37
3.2	OBJECTIVES	37
3.3	EXPECTED CONTRIBUTIONS	38
3.4	RESEARCH DESIGN	38
4	INSIGHTS FOR NEW VISUALIZATION	41
4.1	RELATED WORK	41
4.2	INTERVIEW STUDY	43
4.2.1	PARTICIPANTS	44
4.2.2	PROCEDURE	44
4.2.3	ANALYSIS OF THE INTERVIEWS AND RESULTS	46
4.3	INSIGHTS FOR NEW VISUALIZATIONS	51
4.3.1	KEEP IT SIMPLE	53
4.3.2	KEEP THE CONTEXT	53
4.3.3	SAVE THE TIME	54
4.3.4	THINK BIG	54
4.3.5	ALLOW INTERACTION	55
4.3.6	TABLES ARE OK	55
4.3.7	PAY ATTENTION TO THE WORK SCOPES	55
4.3.8	PREPROCESSING IS PART OF THE ENTIRE CYCLE	56
4.3.9	ALLOW COMPARISON	56

4.3.10	CAPTURE METADATA	56
4.4	DISCUSSION AND LIMITATIONS	57
5	PREPROCESSING PROFILING MODEL FOR VA	59
5.1	RELATED WORK	59
5.1.1	VISUAL ANALYTICS MODELS	59
5.1.2	VISUALIZATION DURING PREPROCESSING	60
5.1.3	VISUALIZATION OF DATA QUALITY ISSUES	61
5.1.4	TOOLS AND SYSTEMS	62
5.1.5	DISCUSSION	64
5.2	THE PREPROCESSING PROFILING MODEL	66
5.3	ARCHITECTURE	69
5.4	PROTOTYPE DESIGN	70
5.4.1	OUTLINE	70
5.4.2	DATA PROFILING	71
5.4.3	PREPROCESSING PROFILING	77
5.5	DISCUSSION AND LIMITATIONS	82
6	MODEL VALIDATION	85
6.1	USAGE SCENARIO: UNDERSTANDING A DATASET AND ITS PREPROCESSING IMPACTS	85
6.1.1	UNDERSTANDING THE DATA	86
6.1.2	UNDERSTANDING THE IMPACTS OF PREPROCESSING	89
6.2	USE CASE: EVALUATING A HEALTHCARE DATASET	92
6.3	DISCUSSION AND LIMITATIONS	97
7	CONCLUSION	99
	REFERENCES	101
	APPENDIX A – Visualizaton Techniques	111
	APPENDIX B – Interview Process: Questionnaire	115
	APPENDIX C – Interview Process: Consent Form	117
	APPENDIX D – Prototype Technologies	119

1. INTRODUCTION

The volume of data and its complexity is increasing over the years, with that follows growing demand for the analysis of this data. However, analyzing and managing raw data is still a challenging part of this process, mainly with regards to data preparation, or preprocessing, which aims to transform this “raw data into an appropriate format for subsequent analysis” [TSK06]. Some authors even mention these activities as the most laborious and time-consuming in data analysis workflows [DJ03, TSK06, KPHH11, TPB⁺18].

We have evolved in the ability to collect and store data, create sophisticated mechanisms to build data mining methods, and generate visual representations to present the information resulting from this process. Therefore, no matter how robust the algorithm created for data mining is, if dirty 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 [DJ03, KCH⁺03].

Besides, even though automated processes are essential and likely to happen, we still need human inputs as a critical piece behind this mechanism. Especially when we are trying to achieve data quality since for many situations, the domain expert is required to decide how to proceed with data cleaning [KCH⁺03]. In such times where data can come from everywhere, e.g., smart sensors and online social networks, we will hardly find datasets without data quality issues. Then, detecting data quality issues and evaluating which are the best strategies to correct them in preparation for the next steps are crucial. In this case, as observed by Lu et al. [LCM⁺17] “one of the major strengths of visualization is enabling users to identify erroneous data quickly”.

Hence, the use of information visualization techniques can play an essential role in data analysis while providing meaningful insights [dOL03, Jug14, WGK15]. However, most of the visualization studies are concerned with the end of the process, when sharing the final results of the analysis. Furthermore, early studies proposing Visual Analytics (VA) Models sought to solve the requirements of the VA process, but they were not considering preprocessing as an equal phase in the process [KKE10]. Also, the more recent studies in VA [SSS⁺14, RF16, FWR⁺17] are focused on Knowledge Generation Models, then remaining the opportunity to continue the discussion on how preprocessing can be supported by this combination of visualization techniques and data mining methods.

In addition, although we can find studies proposing visualization methods to assist with preprocessing activities [KPHH11, KPP⁺12], most of them are focused on data transformation and data cleaning activities. Thus, we can still observe opportunities to be discussed, such as: (a) alternative visualizations to cover the same data quality issue by different perspectives; (b) visualizations to support the evaluation of the preprocessing impacts in further

phases; and (c) list of guidelines and features to support novel visualizations in the context of preprocessing.

Likewise, we can find interview studies with data analysts, enterprise professionals, proposing design implications [BE18, KPHH12] or recommendations [AZL⁺19] for future visualization solutions in the data mining scope, but they cover the entire workflow and do not focus fully in 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 scenario, we built our study efforts to answer the research question “How can we assist the preprocessing activities with visualization techniques during a visual analytics workflow?”. Thus, our main objective is to explore visualization techniques to support the activities part of preprocessing phase, mainly, the data understanding and the evaluation of the preprocessing strategies and its impacts to further phases of the data mining workflows. To achieve that, first, we conducted an interview study with thirteen data analysts to investigate their working practices. The discussion about the challenges and opportunities based on the responses of the interviewees resulted in a list of ten insights. This list was compared with the closest related works, improving the reliability of our findings and providing background.

Next, we proposed the Preprocessing Profiling Model for VA, which is an extension of the VA Model [KKE10]. In it, we created a new phase called Preprocessing Profiling to highlight the need to consider preprocessing activities as an essential part of the workflow. As part of our Model presentation, we introduced a set of features that should be considered when we design new solutions within this scope. These features were compiled based on the state-of-the-art literature review and the list of insights obtained from our interview study. Additionally, to validate the Preprocessing Profiling Model, we performed a qualitative analysis in two different use case scenarios. For this activity, we developed two prototypes: one focused on Data Profiling, and another on Preprocessing Profiling. As a result, we built a discussion about the possibilities of using our Model and how it could be used to assist data analysts in performing preprocessing activities.

As the main contributions of our study, we can list.

- The Preprocessing Profiling Model, as a proposal to fill the gaps within preprocessing activities in Visual Analytics, along with its prototype design and usage scenarios discussion.
- The list of ten insights, as a consolidated set of requirements for future visualization research studies applied to preprocessing and data mining.
- Furthermore, we provide details on the profile of the data analysts, the main challenges they face, and the opportunities that arise while they are engaged in data mining projects in diverse organizational areas.

The remainder of our study is structured as follows: Chapter 2 presents some basic concepts, terminology, and a brief overview of the literature on subjects related to data analysis. Subsequently, Chapter 3 describes the details of the methodology used in our study. Chapter 4 outlines the procedure developed to perform the interviews, the profile of the participants, and the results and analysis of the interviews. Chapter 5 presents the Pre-processing Profiling Model, its architecture, and details on the prototype design. Chapter 6 describes our Model validation discussion. Finally, Chapter 7 presents our conclusions and plans for future work.

2. BACKGROUND

John Tukey, a remarkable statistician, still in 1960s defined data analysis as: “procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of statistics which apply to analyzing data” [Tuk62]. Since then, we have seen **data analysis** referenced and applied in different contexts. Thus, in this chapter, we introduce some basic concepts, terminology, and a brief overview of the following subjects related to data analysis: Data (2.1), Data Mining (2.2), Information Visualization (2.3), Visual Analytics (2.4), Data Management (2.5), Data Quality (2.6), and Preprocessing (2.7). Figure 2.1 shows an overview of the main items covered.

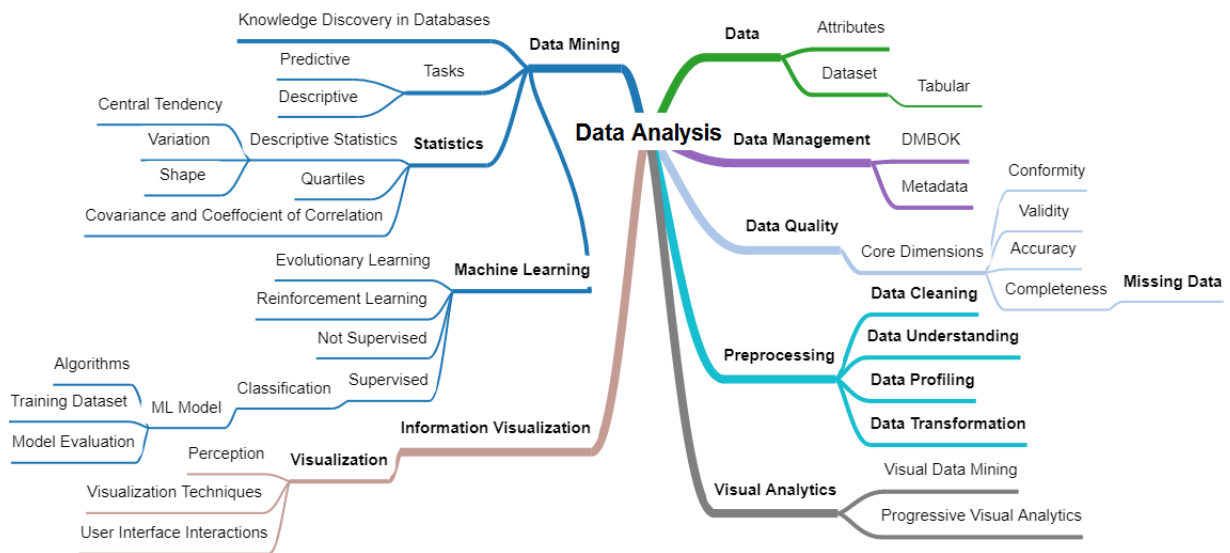


Figure 2.1 – Overview of the main concepts and terms presented in Background.

2.1 Data

According to Tan et al. [TSK06], a **data object** can be named in different ways, e.g., record, case, sample, and observation. These data objects can be described by several attributes that capture the characteristics of an object. In turn, the attributes may also appear referenced with different names, e.g., variable, field, feature, and dimension. Additionally, each attribute can be classified in different types, such as categorical (qualitative), e.g., nominal and ordinal, or numeric (quantitative), e.g., interval and ratio.

The collection of data objects can be defined as **dataset**, and even though there are many types of datasets, they can be often seen as a tabular file [TSK06]. In that case, the rows correspond to the records and the columns to the variables. The data sources

may vary, from corporate systems, such as large stores handling hundreds of millions of transactions per week, to social media, producing digital pictures and videos [HKP11].

As stated by Han et al. [HKP11] (p. 5) “the fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools”. In this scenario, the demand for turning this extensive collection of data into knowledge motivated different area of studies to emerge to attend this need [TSK06, HKP11]. As part of that, we can list data mining, information visualization, and visual analytics. They are described in the next sections.

2.2 Data Mining

Data Mining (DM) is the process of discovering interesting patterns, or useful information, from large amounts of data [TSK06, HKP11]. What differs the definition in Tan et al. [TSK06], when compared to Han et al.[HKP11], is the reference that the discovering process should be automated. In any case, both state DM as a **Knowledge Discovery in Databases** (KDD) process. In turn, KDD [PF91] appears referenced in literature with different names, e.g., Knowledge Discovery from Data, Data Mining, and Knowledge Discovery. Also, the definition and steps varies. Han et al. [HKP11] list seven steps as part of the KDD process: (1) data cleaning, to remove noise and inconsistent data; (2) data integration, to combine data sources; (3) data selection, to retrieve from the database the relevant data; (4) data transformation, to transform and consolidate the data; (5) data mining, to apply methods to extract data patterns; (6) pattern evaluation, to identify unusual patterns; and (7) knowledge presentation, to present mined knowledge to users.

As indicated by Tan et al. [TSK06], the four core DM tasks, or functionalities, are Cluster Analysis, Association Analysis, Anomaly Detection, and Predictive Modeling. These tasks can be divided into two main groups: **Predictive** and **Descriptive**. They are used to specify which types of patterns are being searched during DM process. Then, according to Han et al. [HKP11] (p. 15) Descriptive mining tasks “characterize properties of the data in a target data set” while Predictive mining tasks “perform induction on the current data in order to make predictions”.

Being DM an interdisciplinary area of study, it adopts techniques from many domains [TSK06, HKP11]. Among the areas that play an essential role, we can list Statistics, Artificial Intelligence, and Information Visualization. Information Visualization will be described in the next section, while Statistics and Artificial Intelligence are summarized in the next paragraphs.

With reference to the statistics key terms, we need to highlight the **descriptive statistics**. Berenson et al. [BLSK12] (p. 4) define this as “the methods that help collect, summarize, present, and analyze a set of data”. Considering the example of a numeri-

cal data, it can be characterized by three properties: (a) Central Tendency, e.g., statistical measures as Mean, Median, and Mode; (b) Variation, e.g., statistical measures as Range, Variance, and Standard Deviation; and (c) Shape, e.g., statistical measures as Skewness and Kurtosis. Moreover, additional methods to describe the data could be used as part of the exploratory data analysis, just to mention a few, the computation of Quartiles, which split a set of data into four equal parts, and the Covariance and the Coefficient of Correlation to examine the relationship between two numerical variables [BLSK12, BB09].

In relation to Artificial Intelligence, it is a vast area of research, and we are considering a subset related to **Machine Learning** (ML). One of the first definitions of ML was stated by Mitchell [Mit97] as an automatic process to improve the performance in the execution of some task through the experience. In a more recent definition, Marsland [Mar14] (p. 4) states that “ML is about making computers modify or adapt their actions (whether these actions are making predictions, or controlling a robot) so that these actions get more accurate, where accuracy is measured by how well the chosen actions reflect the correct ones”.

Faceli et al. [FLGC11] explain the ML methods, or algorithms, can be organized according to their learning task type as Supervised Learning (predictive goal) or Unsupervised Learning (descriptive goal), similarly to previous explained for DM tasks. Additionally, Marsland [Mar14] describes as four types of learning, the same Supervised Learning and Unsupervised Learning, and additional two, Reinforcement Learning and Evolutionary Learning. He states that the most common type is the Supervised Learning. Also, for the problems that can be solved using Supervised Learning, one famous example is the **Classification**, which considers a set of data to be used as training data. This training data consists of the input data, or exemplars of each class, and the target data, or labels of each class. Then the process of ML training refers to “the use of computational resources in order to build a model of data in order to predict the outputs on new data” [Mar14] (p. 11). It follows another critical concept, the model evaluation, which involves the selection of appropriate metrics to test the trained algorithm and evaluate the accuracy of the produced results. In turn, different measurements can be used to interpret the performance of a classifier, e.g., accuracy, precision, and recall. As well as different algorithms can be used to build the model, e.g., Decision Tree, Nearest Neighbour, and Support Vector Machines. In our study, we are using the term DM model to refer to this process just described.

2.3 Information Visualization

As indicated by Ware [War04] (p. 2) “we acquire more information through vision than through all of the other senses combined”, hence, “visualizations have a small but crucial and expanding role in cognitive systems”. In relation to **Visualization**, it is defined by Ward et al. [WGK15] (p. 1) as “the communication of information using graphical rep-

representations”. These graphical representations might include different types of data and visualization techniques. Also, they can be applied daily in a variety of areas, such as traffic heatmap, weather charts, or a graph of stock market activities [WGK15].

Additionally, Ware [War04] explains that if the visualization is presented well, it allows rapid interpretation of a huge amount of data, which makes this one of the greatest benefits in its application. Also, he lists other advantages of visualization, such as often enabling problems with the data itself to become apparent and it facilitates hypothesis formation. For that reason, the success of the **Information Visualization (IV)** systems depends on the capacity to create visualizations that transform data into a perceptually efficient visual format.

To assist in this process, multiple **visualization techniques** and **user interface interaction** strategies can be implemented. The book “Interactive Data Visualization” [WGK15] can be used as a reference for the most used visualizations and their frequent applications. Nevertheless, in Appendix A we exemplify a couple of visualizations used during visual data analysis. Moreover, we exemplify two decision trees to select the most appropriate visualization technique based on numeric data and categorical data.

2.4 Visual Analytics

Tukey [Tuk77] introduces the Exploratory Data Analysis, which contributed to the movement from the Confirmatory Data Analysis, using visual representations merely to present results, to a new approach of interacting with data and results. Next, advances in graphical user interfaces supported the development of IV research area, and a new concept of **Visual Data Mining** was proposed by merging IV and DM techniques [Won99, Kei01, dOL03].

From that, a multidisciplinary research area emerged, the **Visual Analytics (VA)**. Wong and Thomas [WT04] (p. 20) were one of the first to use the term, and they explained VA as “an outgrowth of the fields of scientific and information visualization but includes technologies from many other fields, including knowledge management, statistical analysis, cognitive science, decision science”. Moreover, the combination of automatic analytical methods and interactive visual interfaces is essential for any VA solution [TC05, KKE10, KMT10]. This idea is illustrated in Figure 2.2.

In the context of very large and complex datasets, an alternative paradigm named *Progressive* has gained increasing attention over the past years: the **Progressive Visual Analytics** [SPG14] and the **Progressive Analytics** [FP16]. This paradigm 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 to process a whole dataset at once [TPB⁺18].

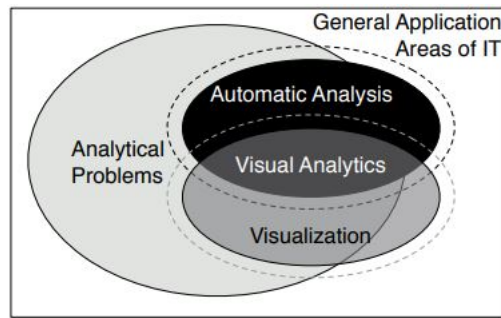


Figure 2.2 – The Analytical Problems can be solved using Automatic Analysis, Visualization, or the combination of both, Visual Analytics [KMT10].

2.5 Data Management

Data Management can be defined as the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their life cycles [Int10]. It is also known by many other terms, including, for instance, Information Management, Enterprise Data Management, Data Resource Management, and Information Asset Management.

Keim et al. [KAF⁺08] emphasize that efficient data management is a critical component for VA since its input is the data to be analyzed. A list of guiding principles for data management are listed on Data Management Body of Knowledge (DMBOK) [Int10]. Among them, two relevant principles from this study can be listed: (1) Managing data means managing the **quality of data**, the data quality is seen as a primary goal. (2) It takes **metadata** to manage data, and a data cannot be manipulated directly, rather it must be understood and defined in the form of metadata, i.e., data about the data.

The term metadata can be ambiguous depending on the domain and interests of users [SSS⁺14]. According to Jugulum [Jug14], it can be defined as a set of data that describes and gives information about other data. It can cover the description of the origin, structure, or characteristics of these data. In our study, when referencing metadata, we are considering more than the description of the dataset, e.g., the label of columns and the type of variables. Beyond that, we are considering the summary of information that can support as input for further analysis, for instance, during the data understanding or data profiling.

2.6 Data Quality

Data quality leads to quality of information. Therefore, it plays a key role in the success of data management [Int10]. Jugulum [Jug14] reviews the concepts, tools, and techniques for building a successful approach to data quality. He states four core data quality

dimensions: (1) conformity, i.e., measure of a data element's adherence to required formats as specified in metadata documentation; (2) validity, i.e., data correspond to reference tables or lists in metadata; (3) accuracy, i.e., measure of whether the value of a given data element is correct and reflects the real-world case; and (4) completeness, which is related to the presence of core source data elements that must be present in order to complete a given business process. In our study, non-completeness will be referenced as missing data and complementary definition is explained as follows.

Statistical studies on **missing data**, or missing values, are not something new, and there are different literature focusing on this subject since the 1970s [LR02]. According to Little and Rubin [LR02], the missing data can be classified in 3 types: (1) Missing completely at random (MCAR), does not mean the pattern itself is random, but rather the reason for the miss is independent of any other data collected. (2) Missing at random (MAR), the reason for the miss does not depend on the value that is missing in itself but on another variable that is being collected. (3) Not missing at random (NMAR), the reason for the missing value is related to the value itself.

2.7 Preprocessing

Essential concepts that help to contextualize **data preparation** or **preprocessing** have already been presented in previous sections. For example, one first definition for preprocessing is presented during the definition of KDD steps (Section 2.2 and Figure 2.3), when the steps from 1 to 4 are indicated as different forms of data preprocessing [HKP11]. This description for preprocessing activities is similar to VA perspective of a typical set including “data cleaning, normalization, grouping, or integration of heterogeneous data sources” [KKE10]. In other words, for Tan et al. [TSK06] (p. 3), the steps of preprocessing include “fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting records and features that are relevant to the data mining task at hand”.

We can observe variations in which tasks are considered part of the preprocessing, due to different areas of research and their particularities for preparing the data under analysis. An example is shown in Figure 2.3 with one possible workflow of ML, KDD, and CRISP-DM [She00].

In general, a broad definition for the preprocessing purpose can be explained as “transforming the raw input data into an appropriate format for subsequent analysis” [TSK06], which should comprehend a number of different strategies, methods, and techniques for **data understanding**, e.g., similarity and dissimilarity between data objects. Also, for **data transformations**, e.g., aggregations and normalization or standardization of variables.

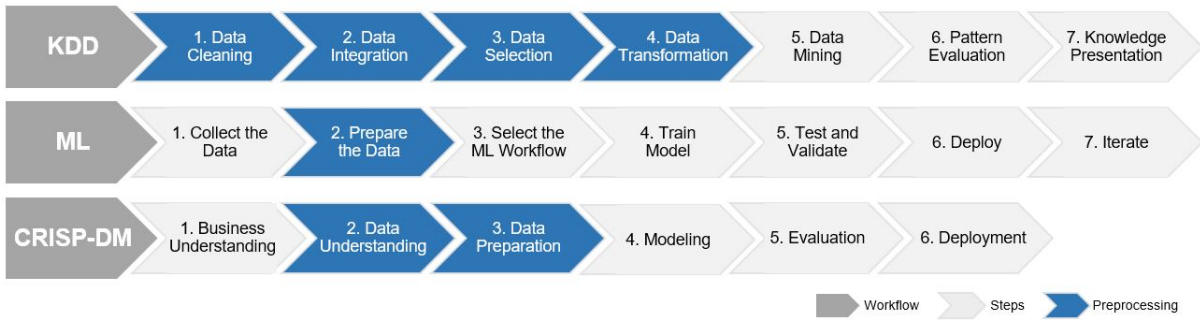


Figure 2.3 – Three examples of workflows used during data analysis: Knowledge Discovery in Databases (KDD) [HKP11], Machine Learning (ML) [Altb], and Cross Industry Process for Data Mining (CRISP-DM) [She00]. The steps highlighted in blue are considered in the scope of the preprocessing phase.

The possible strategies to be used during the preprocessing phase should be based on the type of data, the data quality, and the objectives of the ongoing DM, IV, or VA activity. Hence, the resulting combinations bring complexity to the process, mainly because there is not a single technique or tool to solve all data issues automatically [KPHH11, Hel08]. For that reason, data preprocessing is mentioned as one of the most laborious and time-consuming steps in the overall workflow [DJ03, TSK06, KPHH11, TPB+18].

Kandel et al. [KPP+12] presented an example of the combination of data quality issues, detection methods, and visualization techniques that can be used to analyze their output. It is shown in Figure 2.4. Additionally, Kim et al. [KCH+03] can be used as a reference for a comprehensive review of the **taxonomy of dirty data** and the possible strategies for cleaning transformation methods. This process is known as data wrangling by some authors [KPHH11], and is also called as data cleansing and scrubbing by other references [RD00].

Type	Issue	Detection Method(s)	Visualization
Missing	Missing record	Outlier Detection Residuals then Moving Average w/ Hampel X84	Histogram, Area Chart
		Frequency Outlier Detection Hampel X84	Histogram, Area Chart
Inconsistent	Missing value	Find NULL/empty values	Quality Bar
	Measurement units	Clustering Euclidean Distance	Histogram, Scatter Plot
Incorrect	Erroneous entry	Outlier Detection z-score, Hampel X84	Histogram, Scatter Plot
		Clustering Levenshtein Distance	Grouped Bar Chart
		Clustering Atomic Strings	Grouped Bar Chart
		Clustering Structure Extraction	Grouped Bar Chart
		Clustering Structure Extraction	Grouped Bar Chart
		Type Verification Function	Quality Bar
Extreme	Numeric outliers	Outlier Detection z-score, Hampel X84, Mahalanobis distance	Histogram, Scatter Plot
		Outlier Detection Residuals vs. Moving Average then Hampel X84	Area Chart
		Frequency Outlier Detection Unique Value Ratio	Bar Chart
Schema	Primary key violation	Frequency Outlier Detection Unique Value Ratio	Bar Chart

Figure 2.4 – Example of taxonomy of data quality issues combined with methods for detecting each issue and visualizations for assessing their output [KPP+12].

Turning to the **missing data** issue, complementary to table presented by Kandel et al., after identifying the missing values presence, some possible cleaning strategies can be performed to replace them by a constant value, e.g., zero; or use the resulting calculation of mean, median, or the most frequent number of the attribute in question. Still, according to Kim et al. [KCH⁺03], a variety of options can be performed, and for most of the dirty data types, it should require intervention by a domain expert to decide how to proceed. Thus, information resulting from data profiling, part of the data understanding, should be valuable input to support on this decision process.

Data profiling can be defined as the activity of creating informative summaries of a database [Joh09]. This information can range from simple to complex statistics, such as the total number of missing records in a table, to structural properties as functional dependencies of crucial records. DMBOK [Int10] refers to data profiling tools as critical to supporting on preprocessing activities. Moreover, these tools are defined as a set of algorithms with two main purposes: (1) statistical analysis and evaluation of the quality of the data values in the dataset, and (2) exploration of the relationships between sets of datasets. Therefore, the statistical analysis provides valuable guidance and useful insight in data preprocessing. Also, it should promote the unbiasedness of data analysts during the data transformation choices [DJ03].

In addition, during the discussion on how the presence of dirty data can impact the reliability of the DM model, Kim et al. [KCH⁺03] (p. 96) mention “if the dataset is not to be properly cleansed before being used for training and testing a model, at least a larger dataset should be used to reduce the impact of dirty data”. However, as also mentioned by the authors, that may not always be possible due to data sources constraints, or in other situations when the primary interest is a small number of outlying data, e.g., fraud detection in bank systems, then even a low proportion of data issues should contribute to misleading results.

To conclude, as observed in the previous example, the preprocessing decisions made often have significant implications for the following phases [TPB⁺18, FLGC11, TSK06], which endorses the importance of these activities.

3. METHODOLOGY

In this chapter, we present the details of the methodology used in our study. It covers the research questions, objectives, and expected contributions. Finally, we describe the research design.

3.1 Research Questions

In order to guide this study, the main research question defined is “How can we assist the preprocessing activities with visualization techniques during a visual analytics workflow?”. Additionally, we formulated secondary research questions (SRQ), which are described in the next paragraphs.

SRQ1. Which data visualization techniques are used in the scope of preprocessing?

SRQ2. What kind of problems arise in practical experience during preprocessing?

SRQ3. How data mining methods can support during preprocessing?

SRQ4. How can the resulting information of dataset exploration be presented in a way to support data analysts decision in the next phases?

In response to SRQ1, SRQ2 and SRQ3 we reviewed the related works, as reported in Sections 4.1 and 5.1. Nevertheless, we built a questionnaire and interviewed data analysts to capture additional evidence (Section 4.2) which originated a set of Insights (Section 4.3). In response to SRQ4, we are proposing the Preprocessing Profiling Model, described in Chapter 5.

3.2 Objectives

Guided by the research questions, our main objective is to explore visualization techniques to support the activities which are part of preprocessing phase, mainly, the data understanding and the evaluation of the preprocessing strategies and its impacts to further stages of data mining workflows. Furthermore, five specific objectives were defined to focus our study to answer our research questions.

1. To investigate the current practice of data analysts in data mining workflows.
2. To identify the main problems faced during the preprocessing phase and how visualization supports this process.

3. To identify data preparation strategies that are relevant during data profiling and the identification of data quality problems under analysis.
4. To develop a Preprocessing Profiling Model that can be extended to different use case scenarios.
5. To develop a prototype solution that can be used for Preprocessing Profiling Model evaluation for at least one data analysis issue identified in the second and third aforementioned specific objectives.

3.3 Expected Contributions

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

1. Compile and report the inputs of data analysts on their practice, tools mostly used in data mining, and their perceptions of how visualization can support preprocessing activities.
2. Consolidate a set of requirements for future visualization research based on the discussion of challenges and opportunities obtained on the responses of the data analysts.
3. Develop a Model which increases the capacity of VA in the Preprocessing phase. Hence, supporting data analysts to improve their data understanding and evaluation of preprocessing impacts. As a consequence, promoting the quality of data and supporting decision making on data preparation strategies.

3.4 Research Design

According to Lam et al. [LBI⁺12], the assessment in the scope of IV is complicated since for a complete understanding of a solution, it involves not only the evaluation of the visualizations themselves but also the complex processes that the solution must support. Thus, it is not trivial to develop hypotheses or a set of variables to examine and measure information numerically during experiments. Therefore, following the definitions of the research approach described by Creswell [Cre14], our research fits as an experimental and qualitative study. The thirteen main activities of the research design and their order of execution are presented in Figure 3.1.

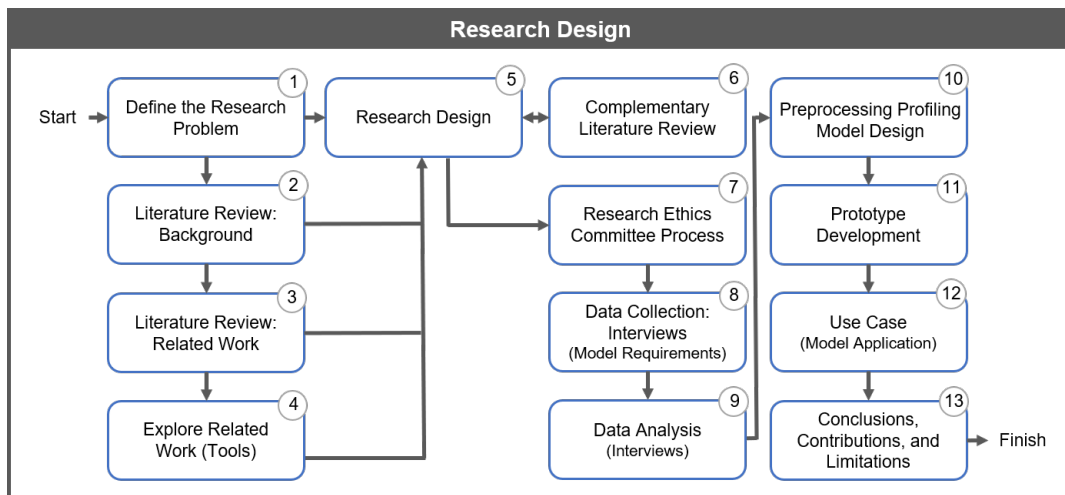


Figure 3.1 – Research Design: list of the main activities.

After the definition of the research problem (Figure 3.1-1), the next first activities (2, 3, and 4) comprised the search and review of the state-of-the-art and related works. Based on that, the research design was revised (5). Later, before start working on the Preprocessing Profiling Model design (10), it was decided to interview data analysts to capture relevant information about their processes and needs (8). Thus, the analysis resulting from this activity (9) contributed as requirements to build a Model as close as possible to attend real situations. However, to proceed with the interviews, an additional activity was added (7), we followed the Research Ethics Committee (REC) protocol to validate our study and get their approval. The complete documentation is available on Plataforma Brasil [DATb] (CAAE number 89239418.0.0000.5336). In parallel to the REC process, we performed a complementary literature review on related works (6).

Although we planned two activities as part of the Preprocessing Profiling Model validation (12), we performed one to explore the Model usage through two use case scenarios. The other, to run a new interview study with data analysts using the developed prototype in different use case scenarios, is now considered as part of our future work opportunities.

In the next chapters, we present detailed analysis and discussion resulted from these activities 8 and 9. In Chapter 5, the Preprocessing Profiling Model proposal and the discussion resulted from activities 10 e 11. To conclude the main body of this study, the results of activity 12 are reported in Chapter 6. Our final considerations (13) are compiled in Chapter 7.

4. INSIGHTS FOR NEW VISUALIZATION

Although 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 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. This is explained in Section 4.1.

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 thirteen data analysts. The information about the participants, the procedure, and the analysis of the interviews and responses are presented in Section 4.2.

The discussion about the challenges and opportunities based on the responses of the interviewees resulted in a list of ten insights, which are explained in Section 4.3. This list was compared with the closest related works, improving the reliability of our findings and providing background, as a consolidated set of requirements. Finally, in Section 4.4, we present the discussion and limitations in our study.

4.1 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. However, since during the exploratory search for the related work we could not find studies focusing on the preprocessing phase, we then decided to also include studies related to an upper level, e.g., data mining, data analysis, or data science, since their workflows contemplate preprocessing activities.

In brief, Figure 4.1 shows the literature review procedure. Initially, four steps were planned following a systematic literature review process. However, we decided to add two new steps, since up to Step 4 only one study met all the inclusion criteria, presented in Figure 4.1. Thus, Steps 5 and 6 followed the snowballing search methodology [Woh14], in an attempt to select additional research, which resulted in a final list of three studies. All these 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 [BE18], RW2 for Kandel et al. [KPHH12], and RW3 for Alspaugh et al. [AZL⁺19].

State-of-the-art Literature Review – Procedure

Literature review started with an exploratory search using several online sources without limitation of publication year. However, only one study was identified as similar work. Hence, we decided to move to a more formal search of state-of-the-art research. The next steps summarize the protocol used. For the first steps, from 1 to 4, we followed a similar approach to a Systematic Literature Review, and for the latest steps, from 5 to 6, Snowballing.

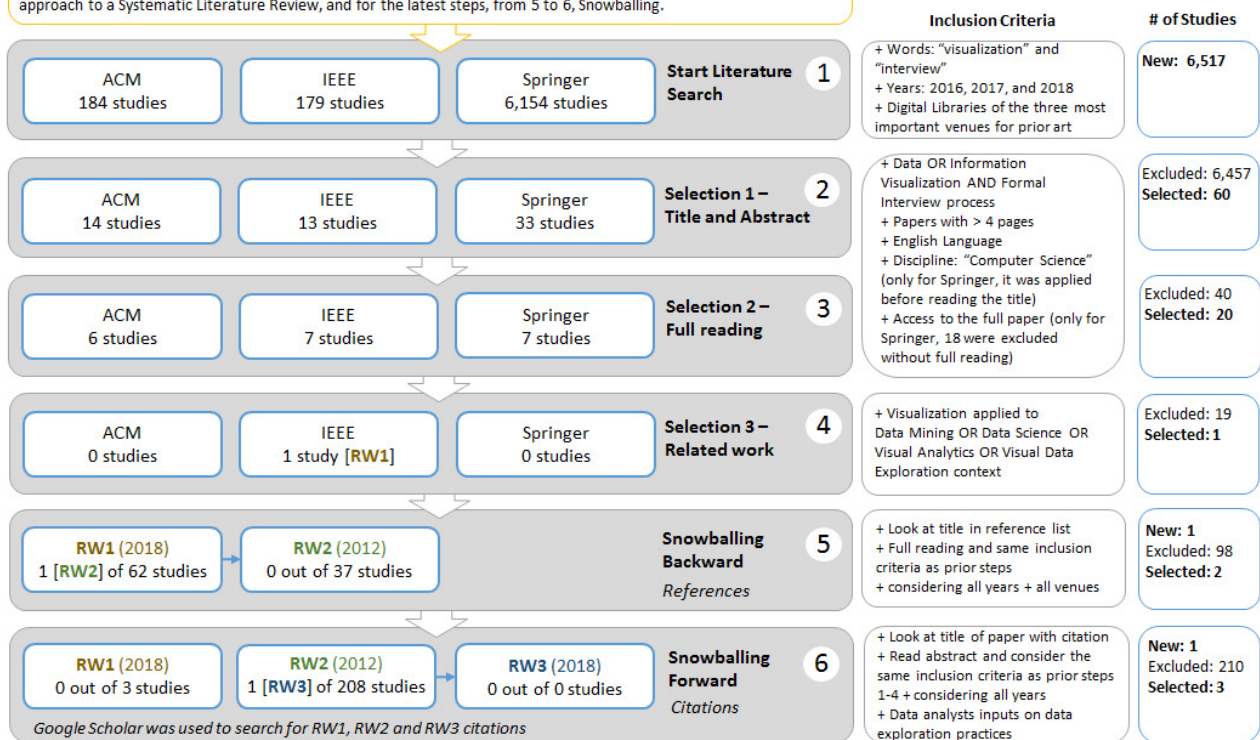


Figure 4.1 – The inclusion criteria for each analyzed study was progressive until Step 4. For Steps 5 and 6, we changed two of the prior inclusion criteria. First, for Step 5, while searching for new references on the list featured in RW1 [BE18], the year of publication was unlimited, which allowed the selection of RW2 [KPHH12], from 2012. Second, for Step 6, while searching for citations, studies which contributed with the perceptions of professional data analysts on the data exploration process were selected even if the study did not focus on visualization. Hence, a third study was selected, RW3 [AZL⁺19]. Among the venues for crucial prior studies, three digital libraries, i.e., ACM, IEEE, and Springer, were selected aiming to cover the most relevant journals and conferences in our research scope, in addition to studies that went through a rigorous review process.

RW1 developed a variant of contextual inquiry to observe eight data analysts in their work environment. All the participants worked for the U.S. Government in Washington, D.C.. Their experience in data science ranged from four to twenty 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 U.S., 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 thirty data analysts located in the San Francisco Bay Area, in the U.S., were still relevant to us, in particular because they presented an extensive discussion on data exploration practices, which included visualization as a tool.

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

4.2 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 in order 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 [KPHH12, BE18, LBI⁺12] were used as reference points during the development of the procedure and the definition of the questions. The interview process is summarised in Figure 4.2.

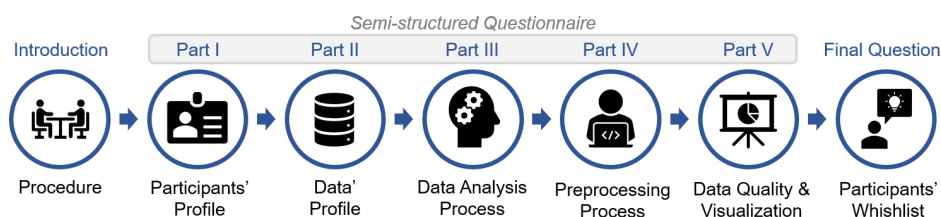


Figure 4.2 – Overview of the Interview Process followed during our study.

4.2.1 Participants

We set as a goal to interview between 10 to 15 data analysts considering the research methods in Human-Computer Interaction [LFH17]. 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 thirteen professionals, twelve male and one female, with ages ranging from 26 to 42. 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. Twelve of them worked in the private sector, and only one 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 three to close to a hundred thousand 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. One of them completed a Ph.D. program, and three were Ph.D. candidates. Their background during their undergraduate studies included different areas such as Physics, Statistics, Engineering, and Business. 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 to 10 years. That happened because 62% of the participants started working in positions outside data mining. Further details on the participants' profile is shown in Figure 4.3.

4.2.2 Procedure

Each participant was interviewed continually, and the sessions lasted from 30 to 60 minutes. The same environment configuration was used for all participants, face-to-face or online conversations, i.e., calls or video conferences. First, we introduced the procedure and presented the consent form, which is available in Appendix C, in compliance with 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 B. A copy of the ques-

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
Scholarity (Highest)													
★ Graduation Complete	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Undergraduation Complete							<input checked="" type="checkbox"/>						
Undergraduation Incomplete								<input checked="" type="checkbox"/>					
Organization Area													
Technology					<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>					<input checked="" type="checkbox"/>
e-commerce						<input checked="" type="checkbox"/>							
Education	<input checked="" type="checkbox"/>												
Financial (Credit and Banking)		<input checked="" type="checkbox"/>											<input checked="" type="checkbox"/>
★ Consulting (IT and Statistics)			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		
Web portal							<input checked="" type="checkbox"/>						
Size of Organization (Number of employees)													
★ < 50			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		
50 - 100													
101 - 1,000						<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						<input checked="" type="checkbox"/>
1,001 - 5,000	<input checked="" type="checkbox"/>												
5,001 - 10,000										<input checked="" type="checkbox"/>			
> 10,000		<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>								<input checked="" type="checkbox"/>
Years of Work Experience													
2 - 5								<input checked="" type="checkbox"/>					
★ 6 - 10		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
11 - 15			<input checked="" type="checkbox"/>							<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
16 - 20	<input checked="" type="checkbox"/>							<input checked="" type="checkbox"/>					
> 20									<input checked="" type="checkbox"/>				
Years of Work Experience with Data Mining or Data Science													
★ 2 - 5			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						
★ 6 - 10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
11 - 15											<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
16 - 20													
> 20									<input checked="" type="checkbox"/>				
Source of Data													
★ Closed (Company)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Open (Public or Gov)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
Other	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>										<input checked="" type="checkbox"/>
Format of Data													
★ csv, tsv	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
xlsx, xls		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>							<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
json, xml		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
tables (database systems)	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
other	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>					<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Type of Data													
★ numeric-text	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
geolocalized		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>							<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	
images	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>				
audio	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>										
Tools and Technology													
Alteryx		<input checked="" type="checkbox"/>											
Anaconda			<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>							<input checked="" type="checkbox"/>
C		<input checked="" type="checkbox"/>											
C++		<input checked="" type="checkbox"/>											
Databricks			<input checked="" type="checkbox"/>										<input checked="" type="checkbox"/>
Gephi													<input checked="" type="checkbox"/>
H2O.ai							<input checked="" type="checkbox"/>						
Hadoop		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>									
HIVE		<input checked="" type="checkbox"/>											
IBM/SPSS												<input checked="" type="checkbox"/>	
Java	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>							<input checked="" type="checkbox"/>				
Jupyter Notebook	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>	
KNIME		<input checked="" type="checkbox"/>								<input checked="" type="checkbox"/>			
Orange	<input checked="" type="checkbox"/>							<input checked="" type="checkbox"/>					
PHP								<input checked="" type="checkbox"/>					
Power BI		<input checked="" type="checkbox"/>											
★ Python	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
★ R and RStudio	<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
RapidMiner			<input checked="" type="checkbox"/>										
SAS		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>									<input checked="" type="checkbox"/>	
Scala/Spark	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>					<input checked="" type="checkbox"/>						<input checked="" type="checkbox"/>
★ SQL	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
Tableau		<input checked="" type="checkbox"/>											
Teradata				<input checked="" type="checkbox"/>									
Watson Studio													<input checked="" type="checkbox"/>
Zepelin			<input checked="" type="checkbox"/>										
Strategies for Data Mining and/or Machine Learning													
Anomaly Detection	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			
Clustering; Association	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Classification	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
★ Regression	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Dimensionality	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Others	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Figure 4.3 – Additional information on the profile of the thirteen participants.

tionnaire was shared with the participants during the interview. Additionally, 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. Afterwards, 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, i.e., each entry is the transcript for the open responses provided by each of the thirteen participants. In more details: Part 1, Participant Profile, resulted in 117 entries since there were nine questions; Part 2, Data Profile, resulted in 52 entries since there were four questions; Part 3, Data Analysis Process, resulted in 52 entries since there were four questions; Part 4, Preprocessing Activities, resulted in 52 entries since there were four questions; Part 5, 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.

4.2.3 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 4.3. Regarding the volume of the datasets in use, it ranges from a small number of data records, i.e., which can be processed in simple spreadsheet, to

Big Data [dMGG15] infrastructures, with billions of records and more than 100 thousand features.

Data Analysis Process

Participants described their work process similarly to KDD, Machine Learning (ML), or CRISP-DM workflows (see Figure 2.3 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*, in order to represent the literature review in the domain under analysis, including DM model evaluations, prior to starting any 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, preprocessing was not highly demanding.

One 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 DM model deployment in the production system environment.

Regarding their data mining strategies, the most indicated were Clustering, Association, Classification, and Regression Analysis. Additionally, many participants mentioned the dimensionality reduction strategy used as part of preprocessing. One participant said that for their context this was not a good strategy, 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 DM 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) 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.

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 [Pyt] and R [Fou] play an important role. Four participants mentioned using tools such as Databricks [Data], KNIME [BCD⁺09, KNI], Gephi [BHJ09, Gep], and Orange [DCE⁺13, Ora] in some moments to support this process. Only one participant said that most of the preprocessing activities were performed directly on Spark [Apa].

When asked to share further details about the preprocessing tasks, 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, e.g., data transformation to match the syntax rules defined by the database where newly arrived data are being appended. Additionally, 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 DM 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 vari-

able 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 strategy for the ML model, e.g., 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 variable being treated.

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

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

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.

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.

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%), Inconsistency-Ambiguous data (62%), and Incorrect Issues, such as Duplicates (54%) and Outliers/Non-Standard (54%). Additionally, two participants indicated that the raw data always has 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, e.g., 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, e.g., 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, line, radar plot, boxplot, scatterplot, and histograms, which are available in visualization libraries for Python, e.g., matplotlib [Hun07] and seaborn [Was], and R, e.g., ggplot2 [Wic10]. In order 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 [Had], Orange, Gephi, and Databricks.

Five participants emphasized that missing data was the most common problem related to data quality. In addition, they mentioned that tools like SAS [SAS] 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 [KT16, TAKP], a graphical user interface available as an R package, to build visualizations to help understand the patterns of these missing values or *NAs*, 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. Additionally, 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 DM model changes. Other three participants mentioned that the choice related to the capacity of the current tools to handle data processing. Free tools, e.g., 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, e.g., 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.

Additionally, 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.

Finally, 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 ten insights introduced in the next section.

4.3 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 they felt confident with their set of tools. However, the ten remaining participants demonstrated an interest in different ways to explore their data with visualization techniques. Based on these feedbacks and complementary to the discussion started in the previous sections, in this section, we present a list of ten insights for visualization in data exploration.

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 4.4:

1. The list started based on the inputs received from participant one 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 labelled 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. We ordered the list considering Step 4 for the insights in common with the related work, i.e., from Insight 1 to 6, then the insights that were only identified in our study, i.e., from Insight 7 to 9, and lastly the additional insight not covered by our interviews, i.e., Insight 10.

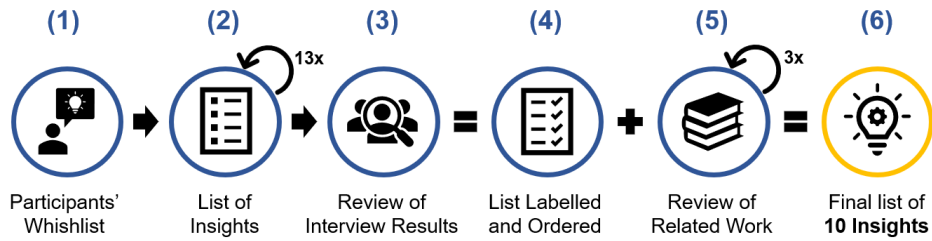


Figure 4.4 – Process to derive the list of 10 Insights.

In Figure 4.5, we added details on the list of insights and the correlation of each source that mentioned them. Also, we complement the explanation for each item in the next subsections. To simplify the description of the comparison with the related works, we will continue using the following code: RW1 for Batch and Elmqvist [BE18], RW2 for Kandel et al. [KPHH12], and RW3 for Alspaugh et al. [AZL+19].

	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. (2018)
Final List of Insights				
1. Keep it simple	<input checked="" type="checkbox"/> (12)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
2. Keep the context	<input checked="" type="checkbox"/> (9)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
3. Save the time	<input checked="" type="checkbox"/> (8)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
4. Think BIG	<input checked="" type="checkbox"/> (5)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
5. Allow interaction	<input checked="" type="checkbox"/> (3)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
6. Tables are OK	<input checked="" type="checkbox"/> (3)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Pay attention to the work scopes	<input checked="" type="checkbox"/> (4)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Preprocessing is part of the entire cycle	<input checked="" type="checkbox"/> (3)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. Allow comparison	<input checked="" type="checkbox"/> (2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Capture metadata	<input type="checkbox"/> (0)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
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	<input type="checkbox"/>	<input checked="" type="checkbox"/> 2	<input type="checkbox"/>	<input type="checkbox"/>
b) Conduct user experience (UX) design sessions with data scientists	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1	<input type="checkbox"/>	<input type="checkbox"/>
c) The verdict on data tables: Not bad	<input type="checkbox"/>	<input checked="" type="checkbox"/> 6	<input type="checkbox"/>	<input type="checkbox"/>
d) Design self-contained, visualization components	<input type="checkbox"/>	<input checked="" type="checkbox"/> 2 and 5	<input type="checkbox"/>	<input type="checkbox"/>
e) Education, not evangelization	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1 and 3	<input type="checkbox"/>	<input type="checkbox"/>
a) Workflow Breakdowns	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 2 and 5	<input type="checkbox"/>
b) Support Scalable Visual Analytics	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 4	<input type="checkbox"/>
c) Bridge the Gap in Programming Proficiency	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 3	<input type="checkbox"/>
d) Capture Metadata at Natural Annotation Points	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 10	<input type="checkbox"/>
a) A Desire for Tool Integration	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 2
b) Trade-offs Between Direct Manipulation and Coding	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1, 2, and 5
c) Automatic Wrangling, Profiling, and Cleaning	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1, 2, and 3
d) Automatically Generated Visualizations and Insights	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 3
e) Analysis Provenance	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 10

Figure 4.5 – Complete list of the insights. (Top of figure, dark blue box) We present the final list of insights, their frequency in our study, i.e., 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 insight.

4.3.1 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 [Taba] and Qlik [Qli], 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.

4.3.2 Keep the Context

Any new solution should remain compatible with the most used tools for data mining, currently Python and R, in order to build an uninterrupted work environment, preventing data analysts from losing the context under investigation while alternating among several different tools. Complementary, RW1 stated it is important to keep the same syntax of the programming environments used by data analysts. Additionally, 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 study 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 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" [WGK15], therefore avoiding the *change blindness* effect, related to the difficulty to notice changes made during an eye movement [Ren00].

4.3.3 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 such as the use of barplot for time series, or line plots for ranking, when they are better in the opposite.

4.3.4 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 4.5), 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. [SPG14] while introducing the Progressive Visual Analytics (PVA).

4.3.5 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 [Shn96]. 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 [BGV16] and VisExemplar [SKBE17]. 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. [HHK15] propose the Predictive Interaction framework for interactive systems that covers general design considerations for data transformations.

4.3.6 Tables Are OK

As we could observe during the interviews, most of the participants are still using tabular data during their analysis (see Figure 4.3). Therefore, aligned with the Insight 1 *Keep it simple*, the tabular format is considered a good 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 like Table Lens [RC94] or Pixel-oriented [Kei00].

4.3.7 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 [GE03] with visualization techniques to generate proposals such as t-Distributed Stochastic Neighbor Embedding (t-SNE) [MH08].

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 [Bra97, Mol]. In addition, aligned with Insight 5 *Allow interaction*, more interactive visualizations to support the parameterization options are needed, such as Deep playground [Ten, SCS⁺17] an interactive visualization of neural networks.

4.3.8 Preprocessing Is Part of the Entire Cycle

For many data mining workflow processes, such as VA [KKE10] and KDD [HKP11], 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.

4.3.9 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 [KS14], 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.

Additionally, 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. [LGH⁺17].

4.3.10 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 iden-

tified 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 [Com].

4.4 Discussion and Limitations

With respect to opportunities for improving our study, we can list two main items: first regarding to the procedure. The number of questions was designed to guarantee that each interview session would take no longer than one hour, 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 two hours 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, such as government, could contribute to a different perspective. Also, we notice lack of female representation, but that seems to be a bigger issue in the STEM (science, technology, engineering, and mathematics) 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.

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 4.5.

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. 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 [vZR17] 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*).

As summarized in Figure 4.6, 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.

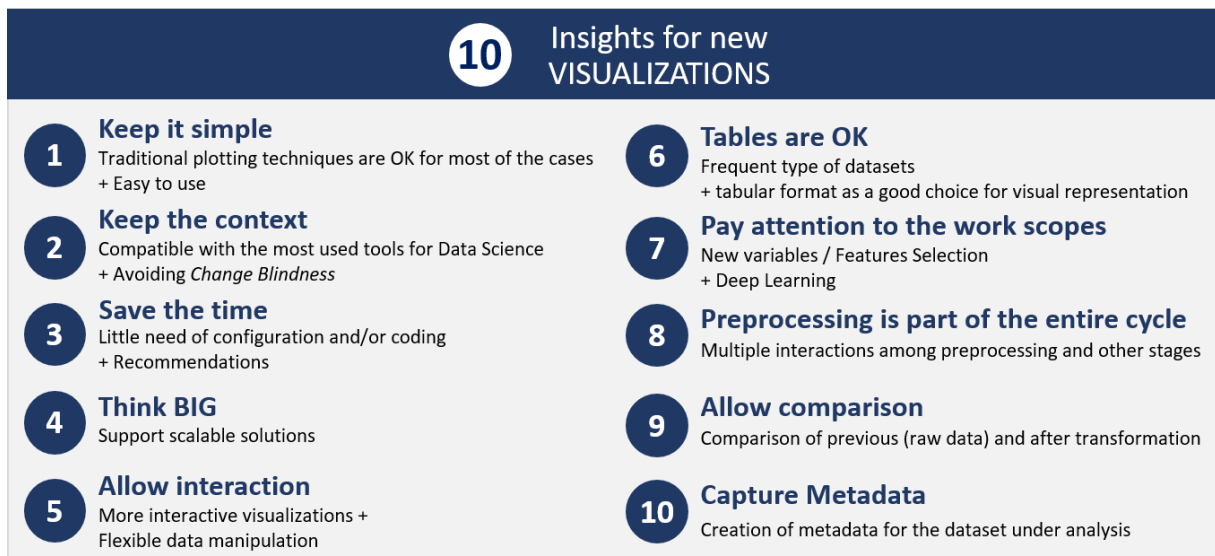


Figure 4.6 – Consolidated list of insights for new visualizations solutions.

5. PREPROCESSING PROFILING MODEL FOR VA

The activities of data preprocessing and their context as part of data analysis are described in the Background (Chapter 2). On top of that and similarly to the concept of Data Profiling, we are using the term of Preprocessing Profiling to indicate the activity of creating informative summaries while performing the data preprocessing activities.

We identified the opportunity to explore the preprocessing activities as part of the VA Model. Motivated by the insights obtained from the data analysts such as 8 - Preprocessing is part of the entire cycle, explained in Chapter 4, preprocessing should not be seen as part of a batch or waterfall approach, but as an activity being constructed during the whole cycle. Moreover, the preprocessing decisions at this phase may have significant impacts on the next steps of the process. Also, we have observed that preprocessing activities are frequently overlooked by data analysts, even though they assume to spend a lot of time involved in these activities during their data mining workflow.

In this section, we present a review of related works that are relevant for our proposed Model definition (5.1). Next, we introduce the Preprocessing Profiling Model for VA (5.2) and the details on its architecture for implementation (5.3). Later, we describe a prototype tool planned as proof of concept of our Model (5.4). Finally, we present our final discussion and limitations (5.5).

5.1 Related Work

This section covers essential related works that influenced the Preprocessing Profiling Model presented in our study. These works are grouped in four areas according to their focus on Visual Analytics Models (5.1.1), Visualization during preprocessing (5.1.2), Visualization of data quality issues (5.1.3), and Tools and Systems (5.1.4).

5.1.1 Visual Analytics Models

As part of the Visual Analytics (VA) discussion, Keim et al. [KKE10] contribute with an overview of the different phases in the VA process, which is illustrated in Figure 5.1. Their process combines automatic and visual analysis methods with human interaction to gain insights and promote knowledge generation. Despite their notorious relevance to VA area, their process still requires more detail when covering the preprocessing activities and the generation of knowledge. The representation of their process as a waterfall flow also deserves further discussion.

To cover the gap in knowledge generation, Sacha et al. [SSS⁺14] describe a novel model for Knowledge Generation in VA. Other works emerged inspired by these previous works, such as Ribarsky and Fisher [RF16] addressing the human-machine interaction loop complementary to [SSS⁺14]; and Federico, Wagner et al. [FWR⁺17] explaining the role of explicit knowledge in the analytical reasoning process when proposing a conceptual model for knowledge-assisted visualizations, grounded in a more theoretical representation as proposed by van Wijk [Wij05]. These three references share the focus on the “Human” side, i.e., cognitive science and knowledge generation aspects. Thus, despite [SSS⁺14] being one of the works that most describes the “Computer” side, during its explanation of the “Exploration loop”, the opportunities to continue the discussion about the data profiling and preprocessing challenges are still existent.

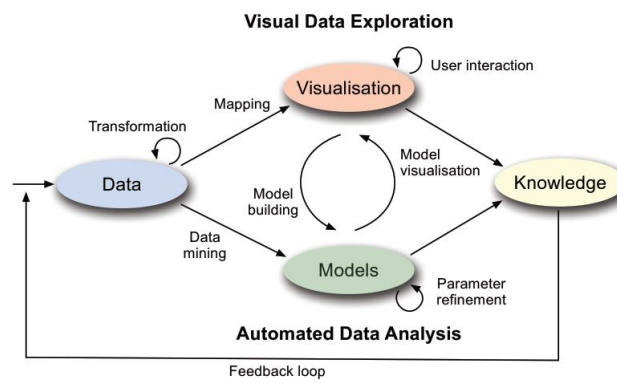


Figure 5.1 – The Visual Analytics process proposed by Keim et al. [KKE10]

Another new group of work in VA are those discussing the Predictive Visual Analytics (PVA), which links predictive analytics methods with interactive visualizations [LGH⁺17, LCM⁺17]. Lu et al. [LGH⁺17] define five steps for the PVA: (1) Data Preprocessing, (2) Feature Engineering, (3) Modeling, (4) Result Exploration and Model Selection, and (5) Validation. Authors highlight two specific aspects of Predictive systems: (a) VA can be integrated to the entire workflow, not following a specific order of steps, and (b) data mining model testing can be applied during the validation step. According to the authors, even though preprocessing appears as the first step of the PVA pipeline, it is one of the most neglected steps since most of the studies in the area tend to focus on modeling and result exploration.

5.1.2 Visualization During Preprocessing

In Lu et al. [LGH⁺17] state-of-the-art review of PVA, the authors noted that the preprocessing step is commonly removed from the main analytic workflow. Although the authors could identify interesting works where preprocessing was integrated, e.g., CO-QUITO [KPS16], a visual interface to assist with the definition of cohorts with temporal constraints, these works are still not entirely dedicated to cover preprocessing problems.

In this context, few relevant works can be cited. One of them is the Predictive Interaction framework for interactive systems [HHK15] that covers general design considerations for data transformations. As the main discussion, Heer et al. [HHK15] propose that the data analyst can decide the next steps of data transformation by highlighting features 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 [KPHH11] is introduced as a system for interactive data transformations. The authors propose an interface language to support data transformation with a mixed interface of suggestions and user interaction while providing visual resources. Both papers provide novel techniques in the scope of preprocessing, but they are limited to the data transformations.

One of the most comprehensive proposals about preprocessing is Profiler [KPP⁺12], an integrated statistical analysis, and visualization tool for assessing quality issues. It uses data mining methods to support anomaly detection. As part of the system architecture discussion, relevant details are presented for combinations of data quality issues, detection methods, and visualizations used by their planned procedures. However, there is still the opportunity to explore different ways to view frequent data issues, e.g., missing values.

In conclusion, for all these papers, two opportunities remain. First, how to integrate them under the most used tools for data analysis. Second, how to explore the comparison of data transformation decisions with its impact on data mining model building, processing, and resulting output.

5.1.3 Visualization of Data Quality Issues

There are comprehensive literature available on how to handle data errors strategies, e.g., [RD00, KCH⁺03]. Among the different types of data quality issues, the missing data are one of the most frequent referenced. Thus, the following works are focused on how to use visualization in support of missing values understanding.

On Templ et al. [TAF12] work, they criticize that no matter how good the classification mechanism for missing data have been planned, they still have limitations such as the right identification of missing cause while working with multivariate data with missing values in several variables. Subsequently, [TAF12] (p. 32) argue to the importance of visualization to solve the related questions on incomplete data: “Visualization of missing values provides a fast way to distinguish between MCAR and MAR situations, as well as to gain insight into the quality and various other aspects of the underlying data at the same time”. Hence, they introduce Visualization and Imputation of Missing Values (VIM) [TAF12, KT16], implemented as an R package with the same name, which provides functions for graphical presentation

of missing data. VIM explores different visualization techniques such as histogram, barplot, scatterplots, parallel coordinate, and boxplot.

Another reference is Missingno [Bil18], implemented as a Python [Pyt] package. It is restricted on visualization options but includes exciting ideas such as a nullity matrix with data-dense display (limited to 50 labelled variables), a barplot for simple valid and invalid data counts, and a heatmap for correlations between the variables with missing values.

During an empirical study to evaluate the best design for graph interpretation with missing data, Eaton et al. [EPD05] observe that data interpretation is negatively impacted when there was a poor indication of the missing values. However, for other cases, even when missing data were indicated clearly, the users continued with their analysis and tried to find trends on the partial data. This corroborates with Sjobergh and Tanaka [ST17] discussion on the importance of developing different ways of visualizing missing values as an attempt to avoid misleading interpretations impacted by how the visualization procedure was developed. Further, they propose a coordinated multiple view framework for visualizing the missing values.

In common, all these works are limited in user interactions and customization options to the provided visualizations, and in their capacity to handle high volumes of data.

5.1.4 Tools and Systems

During the search for available tools and systems supporting data preprocessing activities, five solutions were selected to exemplify some possibilities: Trifacta Wrangler, a paid solution to support preprocessing activities; OpenRefine, as free tool to assist with data cleaning; Tableau, as a robust paid solution for data analysis; Facets, as an online and free solution for the initial data exploration; and Orange, as an end to end free solution for data mining. They are succinctly presented in this subsection.

Trifacta Wrangler [Tri] is a data preparation tool, including raw data evaluation, cleaning and running transformations. It works with the concept of *Flow* where the elements part of the project flow are visually organized. It was developed by the same authors of other works referenced throughout our study [KPHH12, KPP⁺12, KPHH11]. Hence, there are similarities to what was already described (Subsection 5.1.2), howsoever, three additional items can be highlighted: (1) regarding its limitation as a paid tool, the free edition only runs on the local machine, and it does not allow collaboration and sharing features. (2) Even though we need to download the tool, it is still a web-based solution running in the backend, thus with compromised performance even to run simple operations for a small volume of data. (3) Its predictive transformations are a differential with smart options; however, it is all attached to the tool. The same occurs for the visualizations plots, i.e., they cannot be integrated with other solutions in R and Python environment.

With a similar and simpler proposal than Trifacta Wrangler, we can refer to OpenRefine [Ope] as an open-source alternative to support datasets exploration. It is also a useful tool to make the data preprocessing more accessible to any user interested in exploring the data with a minimal level of visual support. However, it is not a visual data exploration tool, and so it is limited to visualization techniques, i.e., it has only text tables and barplot with no user interface interaction.

Tableau [Taba] is an interactive data visualization system focused on business intelligence. In a primary analysis, the most reliable features are the user-friendly design with drag and drop properties, and the vast visualization technique options, e.g., bubble chart, gantt, maps, and boxplot. Additionally, it provides a couple of automatically functions to support the users, such as mark types and scales transformation based on the data under analysis. As limitations, despite the student's free license, it is a paid tool which restricts public usage. Moreover, it requires some previous understanding of the visualization techniques to select the best approach. Although it is a complete solution, it has not an option to run data summaries on raw data to support the identification of data issues before start working on the final visualizations. For that, Tableau Prep [Tabb] was released to provide a visual and direct way to combine, shape, and clean data. It is comprised of two products: "Builder" for building data flows, similarly to Trifacta proposal, and "Conductor" for scheduling, monitoring and managing flows.

Facets [Goo] is an open-source visualization tool to assist in understanding and analyzing machine learning datasets. There are two main views: (1) Facets Overview summarizes statistics for the dataset under analysis. It provides an understanding of the distribution of values across the variables on the dataset. Moreover, it allows uncovering issues like unexpected and missing values. The number of variables under analysis is limited to six numeric and nine categorical. Also, there are some usability issues as it creates panels and sub-panels with scroll bars hindering the data analysis. (2) Facets Dive is an interactive interface for exploring the relationship between data points across all the different variables of the dataset. They implemented a pixel-oriented visualization technique that allows interaction with the data. However, there are still challenges related to how to identify any data patterns, once it was not implemented any hint or suggestion in how to visualize the data, and so similar to Tableau the data analyst should know what they are looking for in advance.

Orange [Ora] is an open-source machine learning and data visualization tool. It allows interactive data analysis workflow and a visual programming approach. Another plus for Orange is the user interface created with help features supporting novice users while working on the tool. It includes many standard visualizations, e.g., barplot and scatterplot, as well as some advanced visualizations, e.g., mosaic display and silhouette plot. As the main limitation for Orange is the capacity to handle dataset with high volume of data. Additionally, even though it has a visual programming approach, the user must know in advance what is necessary to do to evaluate their data and proceed with any data mining step. There are

other paid solutions covering the entire workflow for data mining that can be classified in a similar approach as presented by Orange, such as KNIME [KNI] and Alteryx [Alta]. They promise more robust solutions for commercial purposes, but they were not evaluated as part of our current review. In any case, they should require some training prior to use them.

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? Which visualization techniques can be used to support data quality?

5.1.5 Discussion

We evaluated six items to guide our comparison between the most relevant Related Works (RWs) and the scope of our study. The results are summarized in Table 5.1, and each item is explained in the next paragraphs.

Table 5.1 – Is the study presenting details on the following items? (1) Process or Model or Workflow or Pipeline; (2) Preprocessing is considered an explicit phase on the workflow; (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 understanding.

Subsection	Related Work	Reference	(1)	(2)	(3)	(4)	(5)	(6)
5.1.1	Keim et al. (2010)	[KKE10]	✓					
	Sacha et al. (2014)	[SSS ⁺ 14]	✓					
	Ribarsky and Fisher (2016)	[RF16]	✓					
	Federico, Wagner et al. (2017)	[FWR ⁺ 17]	✓					
	Lu et al. (2016)	[LCM ⁺ 17]	✓	✓	✓			
	Lu et al. (2017)	[LGH ⁺ 17]	✓	✓	✓			
5.1.2	Kandel et al. (2011)	[KPHH11]			✓			✓
	Kandel et al. (2012)	[KPP ⁺ 12]			✓		✓	✓
	Heer et al. (2015)	[HHK15]					✓	
	Krause et al. (2016)	[KPS16]						
5.1.3	Eaton et al. (2005)	[EPD05]						✓
	Templ et al. (2012)	[TAF12]			✓			✓
	Sjobergh and Tanaka (2017)	[ST17]						✓
	Bilogur (2018)	[Bil18]						✓

Items (1) *Process or Model or Workflow or Pipeline* and (2) *Preprocessing is considered an explicit phase on the workflow* evaluated if the RWs address our central problem regarding the importance of Preprocessing to be considered as an equally important phase in the process. Most of the works in PVA are guided by the KDD workflow [HKP11]. Consequently, we can observe that only these RWs [LGH⁺17, LCM⁺17] present preprocessing for-

mally as a phase during their pipeline presentation and discussion. However, they address data mining problems in the scope of Predictive tasks, which does not cover the Descriptive tasks as in the initial VA processes referenced [KKE10, SSS⁺14, RF16, FWR⁺17].

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 were not expecting a complete taxonomy under discussion. On the contrary, we were considering if the RWs were at least bringing into consideration the existence of the complexity in selecting different strategies. In spite of that, not all RWs considered it. The PVA RWs [LGH⁺17, LCM⁺17] contributed with a high-level discussion on the topic. Additionally, the two most relevant RWs in visualization during Preprocessing [KPHH11, KPP⁺12] covered that aspect. Especially [KPP⁺12], which presents a good overview of its related works review and system architecture explanation. Finally, [TAF12] also mentioned the complexity of handling missing values, even if focused on only one data issue.

We consider Item (4) *Preprocessing impacts in the next phases* an essential topic for discussion. It should be addressed during Item 3, but focusing on the impact that the decisions made during the preprocessing may cause to further phases. For instance, the data format of the variables used as parameters during the data mining model building, or the imputation strategies for missing values, may impact the possible methods that can be used, or the processing time, or even the final patterns that can be observed in the results. However, this is not a topic of discussion for any RW, at least not clearly or explicitly.

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 [HHK15] has the goal of proposing a framework, and then, it addressed the item. The majority of the other RWs that could contribute to this item was designed as Systems, therefore, only [KPP⁺12] was added to the list because it provides valuable contributions while presenting its system architecture.

Multiple RWs [EPD05, KPHH11, KPP⁺12, TAF12, ST17, Bil18] can be listed as part of Item (6) *Visualizations for data quality issue understanding*. Nevertheless, all of them still need to invest in visualizations with more capabilities for user interaction, and strategies to handle Big Data volumes. Due to our selection criteria of RWs, i.e., concerned specifically with missing values, we understand that a complementary investigation is required to cover different data quality issues, but we are confident these RWs should support us with this initial discussion.

It is interesting to observe that [LCM⁺17] (p. 194) state as “one of the major strengths of visualization is enabling users to quickly identify erroneous data” when contextualizing the use of visualization during preprocessing in PVA. However, contradictorily to the opportunity of a research area to be explored, in a more recent state-of-the-art in PVA, [LGH⁺17] indicates preprocessing activities are still receiving little attention as part of this research area.

In conclusion, although we can find RWs proposing VA Models or visualization methods to assist with preprocessing activities, we can still observe opportunities to be discussed. From this list, the following items receive less attention than the others: (a) Preprocessing as an equal phase in VA process; (b) Alternative visualizations to cover the same data quality issue by different perspectives; (c) Visualizations to support the evaluation of the preprocessing impacts in further phases; (d) List of guidelines and features to support novel visualizations in the context of this study. To continue this discussion and support filling these gaps, we are proposing the Preprocessing Profiling Model for VA, which is presented in the next section.

5.2 The Preprocessing Profiling Model

In this section, we describe the Preprocessing Profiling Model for Visual Analytics. First, we outline the nine features that we identified as important to be observed while planning new solutions in compliance with our proposed approach. Later, we explain our Model process and its relation to these features.

We matured the proposed features based on the review of related works (Section 5.1) and, especially, on the list of insights obtained in our interview study (Section 4.3). In Table 5.2, we provide a summary description of the meaning and motivation to each of the nine features.

Table 5.2 – List of the nine features and respective descriptions.

Feature	Meaning	Motivation
F1 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.
F2 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 still being processed. Hence, data analysts can visualize huge volumes of data in a continuous flow.
F3 Metadata	Ability to generate informative summaries of the preprocessing activities.	The data computation of other features, e.g., F4 and F5 , should be the source of this feature, 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.
F4 Data Mining	Use of data mining methods to support preprocessing activities.	Data quality assessment can benefit from the use of data mining algorithms, e.g., the identification of data errors and recommendations on data transformation. Additionally, supporting the validation of the preprocessing strategies and data mining model testing.
F5 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 to making decisions on data transformation demands, not only to fix data issues but to better integrate with the planned data mining model. Later, it should be combined with visualization techniques.
F6 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 F4 , this feature can support the evaluation of the data mining model based on different preprocessing strategies.
F7 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.
F8 Template	Ability to generate automatically initial visualizations or basic templates.	It refers to a solution that generates initial visualizations or template options based on the data under analysis. This feature, when combined with F7 , should avoid some inappropriate uses, e.g., the use of Barplot expecting to see trends or time series.
F9 Interaction	Use of visualization interaction techniques to support flexible data exploration.	Although this feature seems obvious for visualization practitioners, it is still an important point of concern to allow the data analysts to perform flexible data manipulation instead of static reports.

In addition to the list of features, we are considering the Visual Analytics process proposed by Keim et al. [KKE10] to devise the Preprocessing Profiling Model. As a result, our Model is formalized as an extension of this process [KKE10], in which we include a new phase named Preprocessing Profiling and new possible transitions among the phases. An overview of the Preprocessing Profiling Model is shown in Figure 5.2.

The term *Preprocessing Profiling* was coined to indicate the activity of creating informative summaries while performing the data preprocessing activities. It is inspired by the concept of Data Profiling, i.e., the activity of creating informative summaries of a database [Joh09].

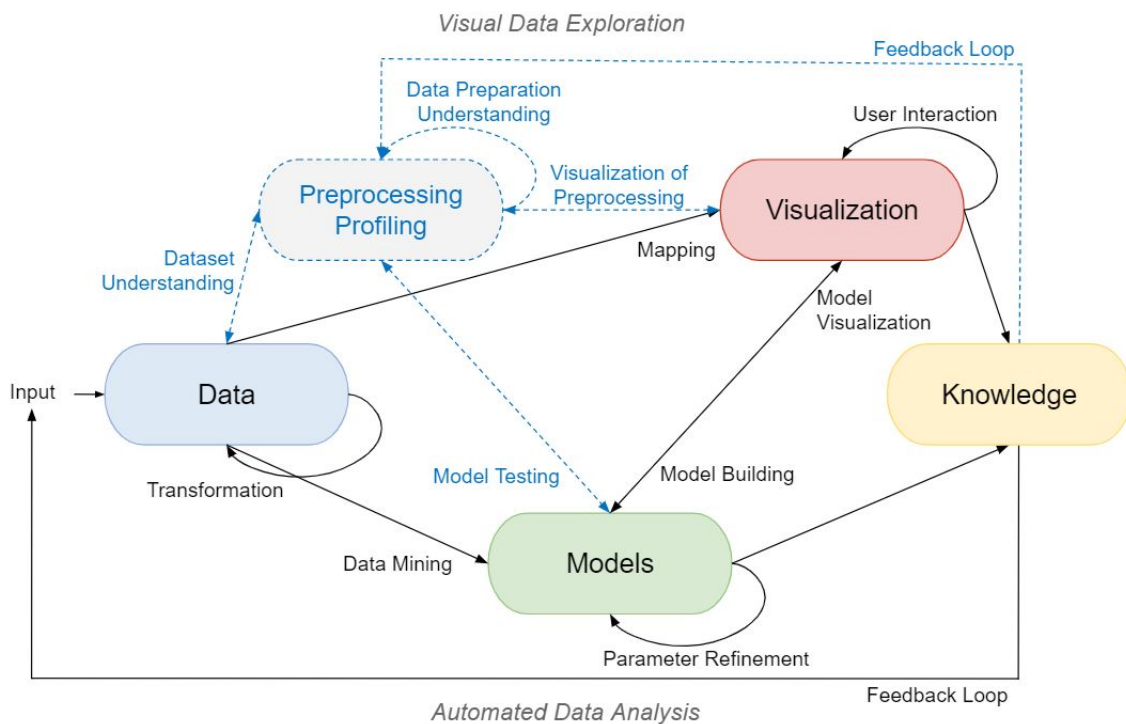


Figure 5.2 – Overview of the Preprocessing Profiling Model for Visual Analytics process. We are extending the VA process proposed by Keim et al. [KKE10]. Each node (represented through rounded rectangle) corresponds to a different phase, and their transitions are represented through arrows. We added 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.

Despite the description of some preprocessing activities in the original Data phase, e.g., data cleaning, normalization, and other tasks as part of the Transformation transition, by adding Preprocessing Profiling as a phase, we put activities such as the data profiling and the evaluation of preprocessing strategies prior to Model Building in the critical path, i.e., as an equally important phase. However, the activities planned as part of Transformation (Data \leftrightarrow Data) can still occur, since, for example, the dataset input may require data standardization to integrate with different sources of data before proceeding with any analysis. Also, the other four original phases and their transitions remain the same.

Thus, next, we focus on explaining only the new transitions. Furthermore, we are providing some examples of how the features presented in Figure 5.3, and here identified by their codes (**F1** to **F9**), can be associated with this process.

The new transition of **Dataset Understanding** (Data ↔ Preprocessing Profiling) intends to explore the dataset, its data types, values distribution, and other descriptive statistics **F5** that will be important to create the data profiling, i.e., metadata **F3**, and then support the data analyst decisions while 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 **F3**. 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 **F7**, or even initial visualizations as templates can be presented to support this activity **F8**.

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 **F4**, 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 **F6**.

All the transitions leaving the Preprocessing Profiling phase have a way back on the same connection (i.e., the arrows in Figure 5.2). Different from the original VA process (Figure 5.1), which can be read as one-way direction, such as a waterfall approach, the PrAVA considers the possibility of multiple interactions between two phases during the same cycle. For that reason, we also added a new **Feedback Loop** (Knowledge → Preprocessing Profiling). Nonetheless, the model proposed by Sacha et al. [SSS⁺14] better describes the different loops in this scope of knowledge generation and should be used as a reference for the subject.

The feature **F2** is related to Big Data scenarios. In these cases, during a flow such as Data → Preprocessing Profiling, we can consider the progressive paradigm [SPG14] as an alternative to producing partial results while the entire dataset is still being processed. Also, for a flow such as Preprocessing Profiling → Visualization, aggregation techniques [EF10] could be used to support generating visual representations more efficiently.

In reference to **F1** and **F9**, 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 [HK12] (p. 53) “interactive tools for data analysis should make technically proficient users more productive while also empowering users with limited programming skills”.

The current VA process presented in Figure 5.1 can continue as-is since it covers cases of confirmatory analysis. However, our approach includes cases in which data adjust-

ments are identified in different phases of the process, and is not limited to the first time data are selected and transformed. Moreover, we advocate about the advantage of visualization techniques during the preprocessing activities, and not only to generate the final visualizations. Hence, our proposed approach considers the Preprocessing Profiling as a prominent phase. Based on that, the Preprocessing Profiling Model for Visual Analytics should enable the data analysts to increase their ownership of the data under analysis, master the impacts of preprocessing activities to the data mining model building, and contribute to the final phase of knowledge generation.

5.3 Architecture

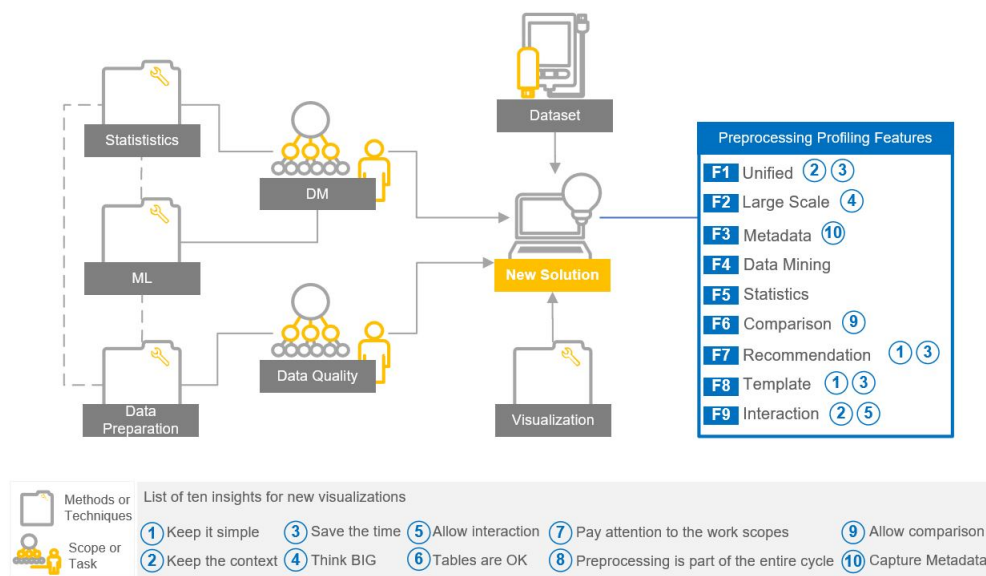


Figure 5.3 – The Preprocessing Profiling Model Architecture. List of components to be considered while developing a new solution. The Preprocessing Profiling features are indicated in the blue box on the right, and the list of insights is indicated on the bottom of the image.

While planning new solutions in compliance to the Preprocessing Profiling Model, despite of considering the list of nine features, we should also consider some inputs, for instance, what is the DM scope? Based on that, which ML or Statistical methods can be used to solve the problem under question? Moreover, which data quality issues are intended to be addressed? Which leads to another question, which data preparation strategies can be used? Finally, which visualization techniques can be used to support on this context?

There are several different possible answers to these questions. For that reason, our conceptual Model is planned to work as an extensible system architecture. This implies that independent of the prior answers, the nine features listed for the Preprocessing Profiling can still be considered as part of the new solution in development. The components of this explained architecture are illustrated in Figure 5.3.

Ideally, all the features should be implemented as part of the new solution. However, the feature list should be viewed as a set of practices to be covered to maximize the performance of the activities performed during the Preprocessing Profiling phase. The more they are considered, the more effective the solution will be. In the next section, we describe the design for a new solution considering these components.

5.4 Prototype Design

As a proof of concept for the Preprocessing Profiling Model, we developed a prototype solution, which generates as output two dynamic reports. One is named *Data Profiling* and it is related to the sequence of Data ↔ Preprocessing Profiling ↔ Visualization (Figure 5.4-1, yellow). The other is named *Preprocessing Profiling* and it is related to the sequence of Data ↔ Preprocessing Profiling ↔ Models ↔ Visualization (Figure 5.4-2, blue). In the next subsections, we outline their design and we present some examples of the implementation for each report.

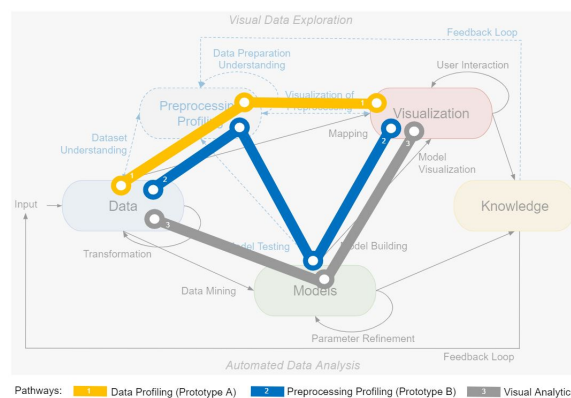


Figure 5.4 – The sequence of steps in the Visual Analytics process.

5.4.1 Outline

As primary requirements, the prototype solution should present: a user-friendly design, with widely used visualization techniques, that can be saved and used as meta-data. Also, it should be integrated with the most popular development environments for data analysis. Furthermore, it should allow user interactions. Hence, the prototype solution was developed as a dynamic report. It was written in Python, Javascript, HTML5, and CSS3, which requires a browser to open the output file for the report. Considering an example of integration with Jupyter Notebook, a few lines of code are required to generate the report in the output cell.

We used the implementation of the Pandas Profiling package [Pan] as the starting point of development. This package (version 1.4.1) originally covers some of the planned items for the *Data Profiling* scope. However, beyond fixing minor bugs and adjusting the layout for the standardization, several new items were implemented as explained in Subsection 5.4.2.

The prototype is written in Python, Javascript, HTML5, and CSS3, which requires a browser to open the output file for the report. The list of technologies used for the prototype development is available in D. Moreover, Javascript implementations allow some traditional web user interaction. For example, show additional details while mouse is over an object, create links to different sections, and allow to expand or hide information.

Considering an example of integration with Jupyter Notebook, few lines of code are required to generate the interactive report, as indicated in Figure 5.5.

```
In [ ]: #Import Libraries
import pandas as pd
import pandas_profiling
#Read dataset and create Pandas Dataframe Object
df = pd.read_csv('E:\\workspace\\dataset\\risk_factors_cervical_cancer.csv', encoding='UTF-8')
#Call Data Profiling Report informing the dataframe
pandas_profiling.ProfileReport(df)
```

Figure 5.5 – Example of Python programming code to run the report.

5.4.2 Data Profiling

The *Data Profiling* report is planned to support data analysts during the data understanding required as part of the preprocessing activities. The first version of the report is divided into five sections: Overview, Variables, Missing Values, Correlations, and Sample. Two additional sections are planned, but they are not fully integrated into the current version of the prototype report.

A high-level comparison between the original version of Pandas-Profiling and our prototype is presented in Figure 5.6. Beyond the fix of minor bugs, the main items changed are the following:

- New procedure to verify the presence of missing values in the dataset to be analyzed. For instance, datasets available on the UCI Machine Learning Repository [DGa] use the character “?” to represent nulls, and in the original Pandas Profiling implementation it was not computed as part of missing values, but as a valid value. Other characters were considered as part of the validation such as “na”, “n/a”, and “null”.
- New section for Missing Values, considering four new visualizations.

- Report layout adjustments for the standardization, e.g., text font type and size, and the use of colours on visual representations. Also the disposition of information were revised in an attempt to improve the readability of the report, such as the inclusion of tabs in Correlations.
- New visualizations: horizontal barplot for missing values, boxplot, and histogram.

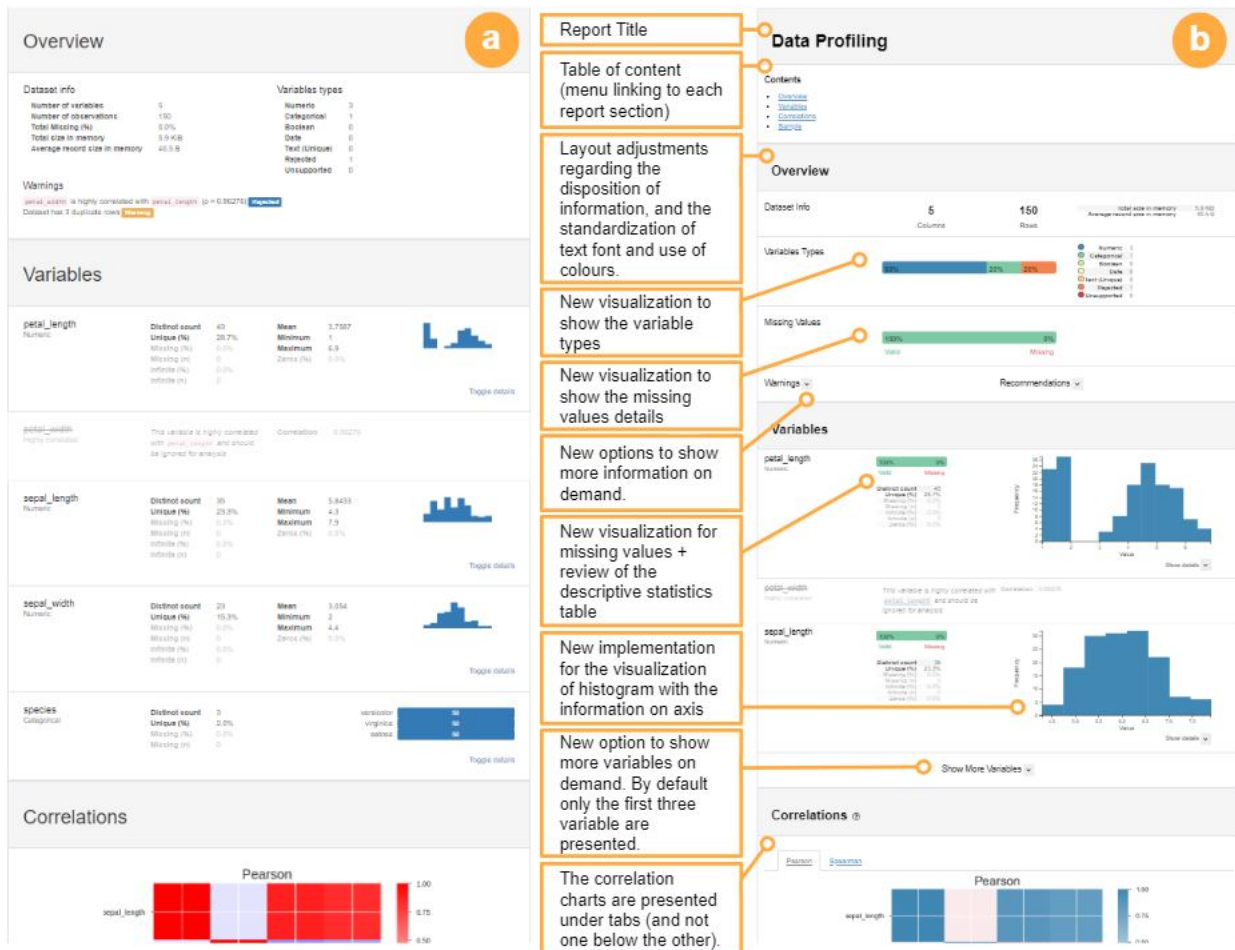


Figure 5.6 – High-level comparison between the (a) original Pandas Profiling Report and (b) Data Profiling Report. The orange boxes indicate new items implemented. In both examples, the Iris dataset is used as input.

The dataset for Cervical Cancer (Risk Factors) from the UCI Machine Learning Repository [DGB] was used as data input. After running the Python code mentioned before (Figure 5.5), the *Data Profiling* report is generated in the output cell of the Jupyter Notebook. Each section is explained in the next paragraphs.

Overview

In the first section of the *Data Profiling* report, details about the dataset are presented, such as the total number of rows and columns, distribution of variable types, percentage of missing values, Warnings, and Recommendations. Regarding the Warnings,

they may be generated indicating relevant aspects to be observed in the dataset under analysis, such as if a specific variable has a higher number of missing values or zeros, or if two variables are highly correlated. There is an option to expand/hide this information. The same is planned for Recommendations, which lists suggestions of how to proceed with the data wrangling or which are appropriated visualization techniques based on the data under analysis. All this information should support the data analyst not only to get an overall view, but also can be used as metadata.

Variables

Information about each variable, i.e., attribute or column, of the dataset is presented in this section. The initial view option is illustrated in Figure 5.6, and it shows a summary table with statistical data and a visualization, Histogram if the variable type is numeric or Horizontal Barplot to indicate the frequency of values. Further details are available when “Show Details” is selected (Figure 5.7). In addition to the data understanding that supports data transformation decisions, the visualizations presented under this section contribute to the identification of data issue patterns, e.g., Boxplot (Figure 5.7-f) and Histogram (Figure 5.7-c) for outliers, and Horizontal Barplot (Figure 5.7-a) to indicate the frequency of missing values.



Figure 5.7 – Data Profiling Report - Variables. Information of a Numeric Variable, with (a) the missing values frequency bar. Also, details for each tab: (b) Quantile and Descriptive Statistics; (c) Histogram; (d) Frequency of the most common values; (d) Extreme Values, five Minimum and five Maximum are listed; (f) Boxplot.

Missing Values

This section is dedicated to support the understanding of missing values in the dataset. Four visualizations are presented, and they were implemented based on Missingo [Bil18].

The Nullity Matrix (Figure 5.8-a) is a data-dense display that supports the identification of patterns for the missing values. The records are shown in different colours, i.e., dark grey for valid records and white for the missing values. Considering this particular case illustrated in Figure 5.8, even without any prior information regarding the dataset used for input, it is possible to observe three patterns quickly. First, there are two columns with high nullity (Figure 5.8-e). Second, the first and the last eight columns seem to be fully informed as no occurrence of white spaces is observed (Figure 5.8-f). Third, many columns seem to be nullity correlated, i.e., when one column has a missing value for a particular row, there is a high chance of the other columns in this group having missing values as well (Figure 5.8-g).

Moreover, the first and second statements mentioned before can be confirmed when looking at the Barplot (Figure 5.8-b). The total count of valid values is informed, and it is possible to see the proportion of missing values per column. Furthermore, the third statement can be confirmed when looking at the Heatmap (Figure 5.8-c) since this visualization shows the relationships within pairs of variables having missing values. Complementarily, the Dendrogram chart (Figure 5.8-d) uses a hierarchical clustering algorithm to support revealing trends for a group of variables.

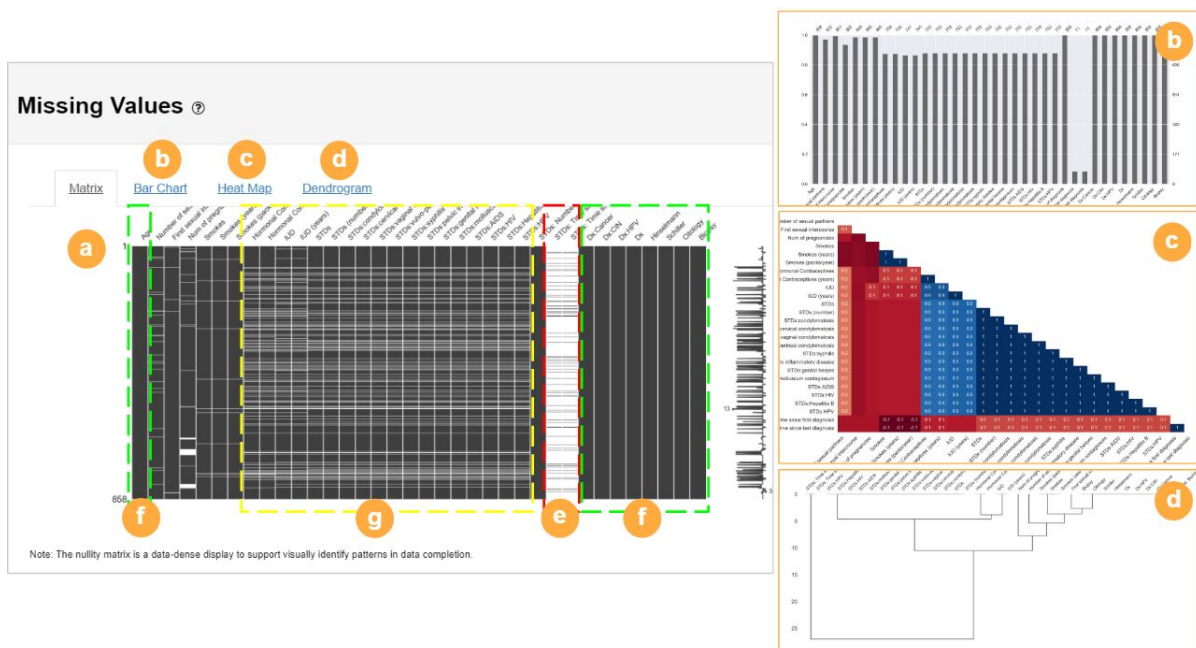


Figure 5.8 – Data Profiling Report - Missing Values. Four visualizations are presented, one per tab: (a) Matrix, (b) Barplot, (c) Heatmap, and (d) Dendrogram. Output generated based on Cervical Cancer dataset.

Correlations

This section is dedicated to the Correlation Coefficient (Figure 5.9). Two measures are used, e.g., Pearson and Spearman, but others can be incorporated. This visualization supports a faster identification of variables with higher or lower relation, through the use of a Heatmap to show the scale ranging from -1 (red) to 1 (blue). Moreover, when clicking on “Show Details”, the numeric coefficients of all pairs of variables are displayed on the bottom of the chart.

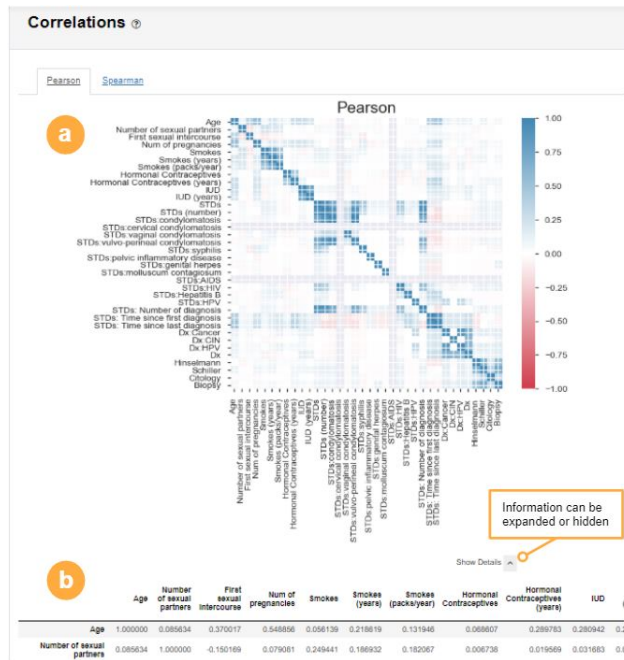


Figure 5.9 – Data Profiling Report - Correlations. Output generated based on Cervical Cancer dataset. (a) Pearson result. (b) Table with complete information.

Sample

This final section shows the first five rows for all columns of the dataset as a sample of the available data. Figure 5.10 illustrates that. Despite seeming trivial, this brings agility to the analysis process by keeping all the relevant information about the dataset in an integrated view.

Sample

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives (years)
0	18	4.0	15.0	1.0	0.0	0.0	0.0	
1	15	1.0	14.0	1.0	0.0	0.0	0.0	
2	34	1.0	NaN	1.0	0.0	0.0	0.0	
3	52	8.0	16.0	4.0	1.0	37.0	37.0	
4	46	3.0	21.0	4.0	0.0	0.0	0.0	

Figure 5.10 – Prototype A - Data Profiling. Fifth Section, Sample of the first lines of the dataset.

Additional Sections

Other sections could be incorporated to this prototype. For instance, considering the needs for the dataset evaluation while working in a classification problem, one possibility is the Relation Matrix of Classes. It is a pairplot to visualize the similarities and differences between the classes, and we can combine different visualization techniques, e.g., Histogram and Scatterplot. Also, we can take advantage of using colours and markers to increase the clarity of the classes distribution. Figure 5.11 shows an example for Iris dataset [DGc] implemented based on Seaborn [Was] visualization package.

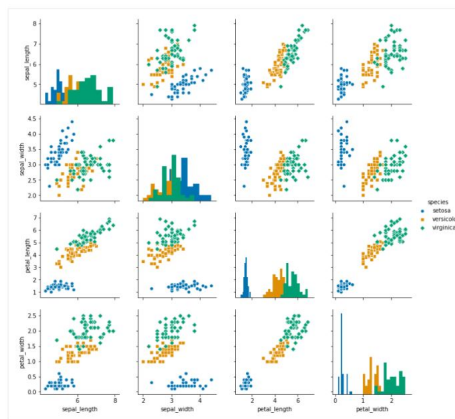


Figure 5.11 – Data Profiling Report - Relation Matrix. Pairplot to visualize the similarities and differences between the species.

Another section planned to be incorporated to the prototype is based on Facets Dive [Goo], mentioned in Subsection 5.1.4. This visualization should enable the data analysts to explore and play with the raw data through an interactive interface. Figure 5.12 shows an example of Facets Dive pre-loaded dataset and three variables selected for evaluation: Country (X-Axis), Education (Y-Axis), and Occupation (Display Colour).

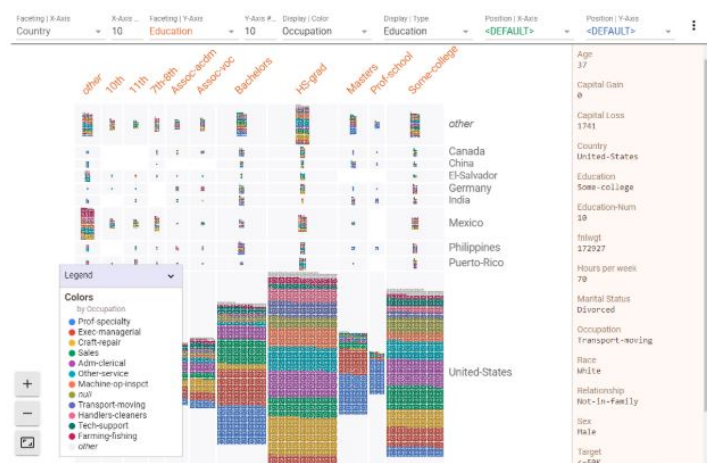


Figure 5.12 – Example of Facets Dive [Goo], using their pre-loaded dataset example.

5.4.3 Preprocessing Profiling

The *Preprocessing Profiling* report focuses on preprocessing activities related to data cleaning, transformations, and the evaluation of the impacts on the data mining 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). Four sections were planned: Dataset Details, Overview of Classification Results, Classification Results in Details per Imputation Strategy, and Diving into the classification results.

The basic operation of the *Preprocessing Profiling* prototype occurs as follows:

1. Receives the dataset (Pandas dataframe object);
2. Performs a validation if the received dataset has missing values, otherwise randomly adds null records to the dataset;
3. Performs the data transformations based on different strategies to handle the missing values;
4. Divides the data into training (70%) and testing (30%);
5. Trains a model using the classification algorithm (Decision Tree) and uses it to predict the classes of the test data;
6. Generates the visualization for the dataset overview, preprocessing metadata, and the classification results.

In relation to item 2, this procedure was implemented as part of the exercise in evaluating the imputation strategies for missing values even when the dataset does not have them by default. Thus, the procedure adds nulls randomly in 75% of the lines, except in the column correspondent to the class labels. This implementation is not only resourceful for our tests during the proof of concept, but as well as for possible real scenarios in need of performing anticipated evaluations of some preprocessing strategies based on different scenarios simulations.

Turning to item 3, five different strategies of data imputation are performed: one removes all the rows that have at least one missing value, and it is named *Baseline (no missing)*. Other replaces all missing values by zero, named *Constant(=zero)*. A third and fourth replace missing values by mean and median values computed based on all records on the same column; they are named *Mean* and *Median*. The last one replaces missing values by the most frequent value in the corresponding column, named *Most Frequent*. When any attribute in the dataset is recognized as boolean, even if it is still a numeric type, the strategies of *Mean* and *Median* are not executed. That occurs because the current implementation applies the same data transformation to all attributes with missing values.

The same set of training and testing data is used for each imputation strategy, except the first, *Baseline*. The *Baseline* runs the classification after removing all the rows with missing data, or the original dataset when it does not have any missing values. That is the reason for different numbers for the Support when *Baseline* is compared with the other imputation strategies.

Dataset Details

This is the first section after the table of contents, and it shows a Sample table of the dataset under analysis. Figure 5.13 illustrates it. Also, it includes the Nullity Matrix, that is important to keep the major picture of the missing values distribution of the dataset under analysis at a particular round of testing. Both visualizations follow the same implementation as the *Data Profiling* report.

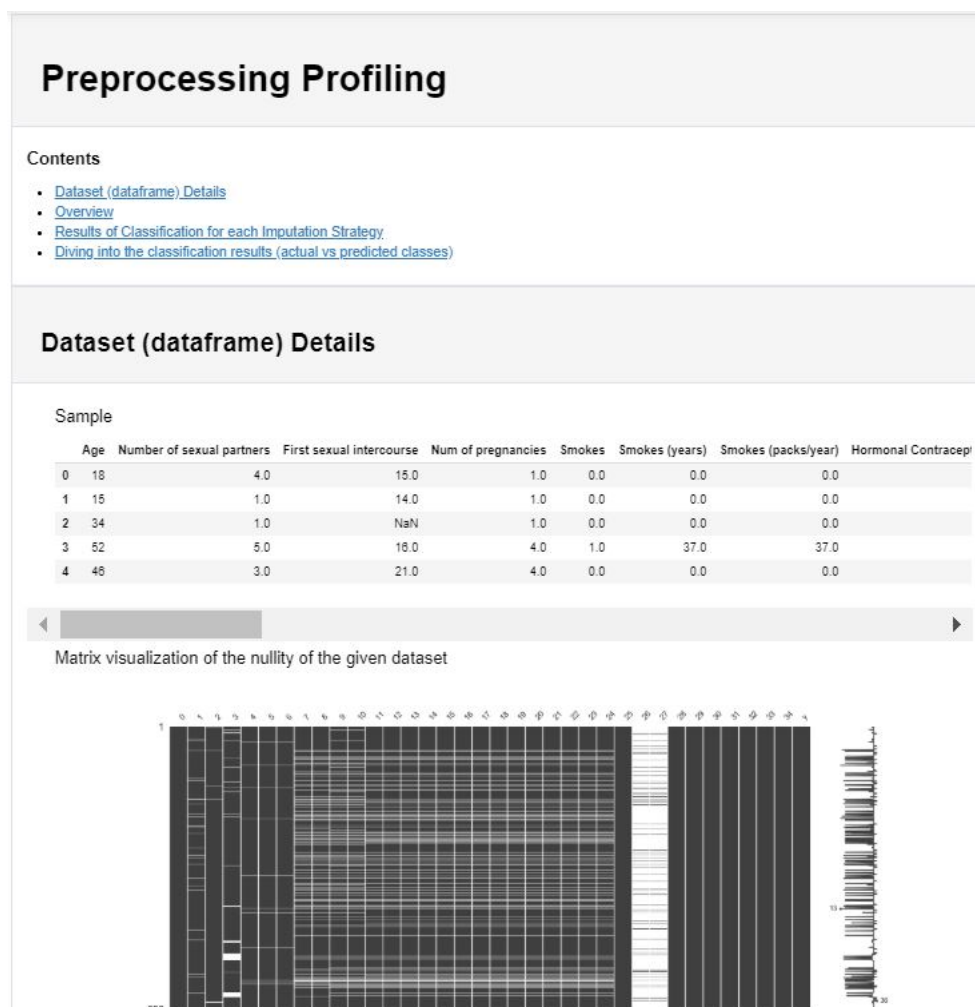


Figure 5.13 – Preprocessing Profiling Report. Menu of content and the Details for the Cervical cancer dataset with missing values from the original file.

Overview of Classification Results

The second section shows a summary table presenting in each line the different imputation strategies for missing values. In each column the metric or report used for classification model evaluation is informed, e.g., Classification Report, Confusion Matrix, Class Prediction Error Distribution, and Accuracy. This consolidated view facilitates the comparison of the results. The results for Cervical Cancer dataset are presented in Figure 5.14.



Figure 5.14 – Preprocessing Profiling Report - Overview of Classification Results for Cervical Cancer dataset.

Results of Classification for Each Imputation Strategy

The third section covers additional details of the classification results. The first three visualizations (Figure 5.15-a, b, and c) are the same available in the Overview of Classification Results (previous explained), and there are two new ones for Precision Recall Curves (Figure 5.15-d and e). These five visualizations were implemented using Yellowbrick [BB19].

The last visualization, Flow of Classes (Figure 5.15-f), aims to provide a new perspective for the comparison of actual versus predicted classification. Also, it should support the analysis of the unbalanced distribution of classes when compared to the Error Distribution (Figure 5.15-c). This visualization was implemented based on Sankey Diagram from D3 [Bos].



Figure 5.15 – Preprocessing Profiling Report - Results of Classification of each Imputation Strategy based on Iris dataset. Six visualizations are presented: (a) Classification Report, (b) Confusion Matrix, (c) Error Distribution, (d) Precision Recall Curves, (e) Precision Recall Curves (Individually), and (f) Flow of Classes.

Diving Into the Classification Results

This final section consists of two new visualizations: Flow of Classes and Matrix of Nullity + Class Prediction Error. They appear in individual tabs and should support the comparison of actual versus predicted results for each class.

In contrast to the previous Flow of Classes (Figure 5.15-f) shown for each imputation strategy, in this new visualization (Figure 5.16), we are combining all the results. The main idea remains in showing the volume of correct (blue link) and wrong (yellow link) predicted classes by showing the percentage of total instances in each condition, instead of only the absolute numbers. Moreover, some interaction options are planned, for instance, the possibility to drill down on records (rows) that are part of each group when selecting (mouse click) a specific link bar. Also, by selecting the preprocessing strategy, the procedure of the data transformation should appear. This is built on top of the metadata for preprocessing that should be automatically computed in parallel to each processing, and can be updated by the data analyst with their annotations for future reference.

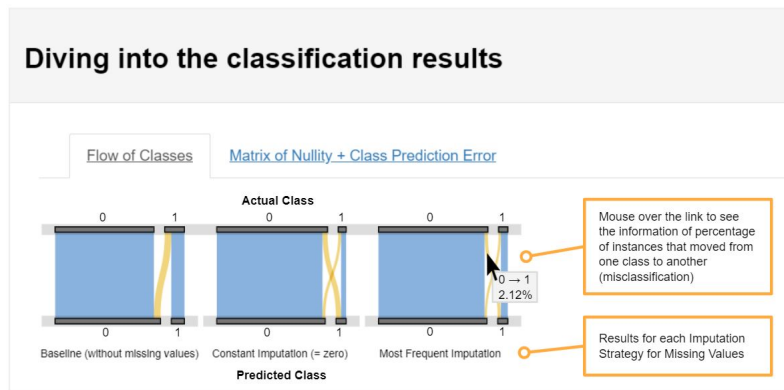


Figure 5.16 – Preprocessing Profiling Report - Flow of Classes per round of preprocessing strategy.

The second visualization aims to support pattern identification on misclassification that may be caused by missing values. Figure 5.17 shows an example of the Matrix of Nullity combined with the results of a Class Prediction Error for a binary problem. It is built based on the idea of Matrix Dense Pixel [Kei00], similar to our previous implementation of the Nullity Matrix. The colour palette remains the same, dark grey means valid values, and white means missing values. Besides, new colours represent the records with classification problems. In our example, if the record (row) was misclassified, it appears in yellow if it is actually Class 1 and was wrongly predicted as Class 2. Likewise, it appears in blue if it is actually Class 2 and was wrongly predicted as Class 1.

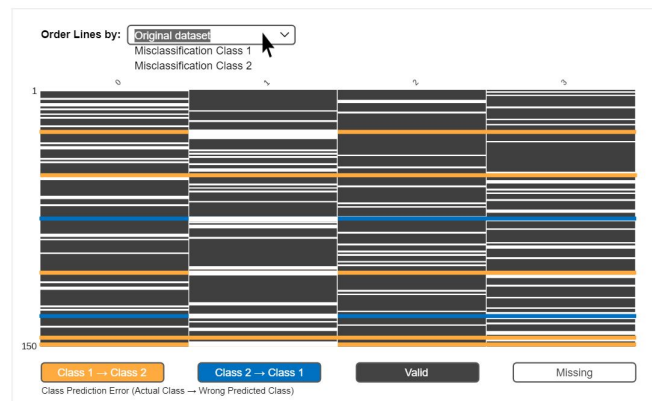


Figure 5.17 – Preprocessing Profiling Report - Matrix of Nullity combined with Class Prediction Error

As part of the component planned for this visualization, we can also list interaction options such as the action to reorder the records by the type of error, e.g., ordering the rows to show first the misclassifications of Class 1, and in that case, potentially putting in evidence any pattern that may be hidden on the original dataset order. In the same way of the previous visualization of Flow of Classes, the data analyst can select a record to see the values corresponding to all the attributes of the specific row. Moreover, a table listing each group of misclassified records should be presented on the bottom of the matrix.

As a final remark, further investigation of aggregation strategies [EF10] is still planned to allow this visual metaphor to scale while analyzing big datasets. This needs to be carefully evaluated; otherwise, a wrong design decision may introduce issues on data distribution that may impair the visual identification of any pattern.

5.5 Discussion and Limitations

The early related works presenting the VA Models sought to solve the requirements of the VA process and are used as inspiration to our work. However, they are not considering preprocessing as an equal phase in the process. In addition, the more recent VA Models are focused on Knowledge, the Human side, then remaining the opportunity to continue the discussion on the Computer side, particularly regarding the activities involved on data preprocessing.

Furthermore, although Kandel et al. works are not proposing a new Model focused on preprocessing, their discussion in data transformation, data cleaning, and the assessment of data anomalies brings outstanding contributions that are used as a reference during the design of the features and the ideas for the prototype implementation. In the same way, VIM, Missingno, Tableau, Trifacta, and Facets are used as inspiration. However, as prior mentioned, there are remaining opportunities for discussion, and aiming to support filling these gaps we presented the Preprocessing Profiling Model for VA.

During the conception of our Model, we take into consideration the requirements obtained during the interview process in combination with related work review. We designed our Model and its architecture to be generalist, with the ambition to attend the most varied use cases. It should be extensible as new demands arise. However, our Model still needs a rigorous assessment with domain experts and may be revisited after this process. In any case, to mitigate potential design issues, we have developed two prototypes: one focused on Data Profiling and another in Preprocessing Profiling activities.

The prototypes presented are an initial version to support our Model use case explanation. In reference to the Architecture of Preprocessing Profiling Model, we present in Figure 5.18 the main tasks and techniques covered in the prototypes. They still have limitations regarding the user interface interactions, data mining methods to uncover data issues, and support to Big Data. However, the current limitations are related to the prototype implementation, and not to the Model design. Despite these limitations, the prototypes are operational, and we can proceed with the Preprocessing Profiling Model discussion to explore how they can assist the data analysts in preprocessing activities. The Model validation is described in the next chapter.

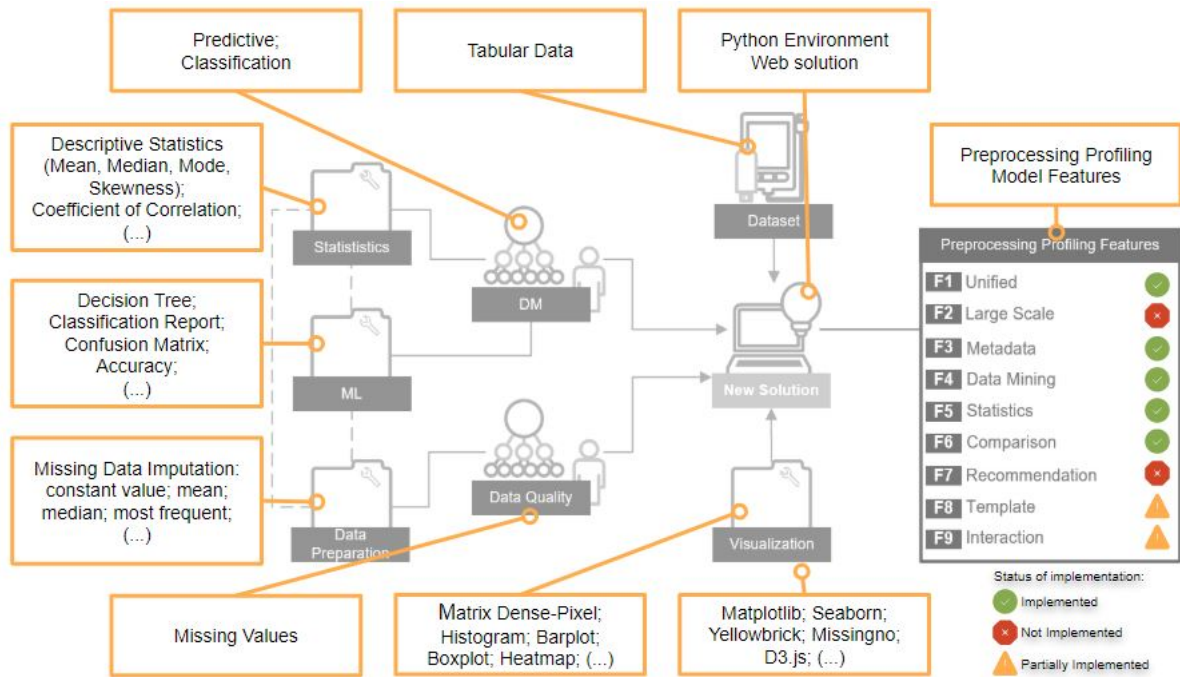


Figure 5.18 – A high-level illustration of the prototype coverage in comparison to the Preprocessing Profiling Model Architecture. For each corresponding component indicated in the Architecture (see Figure 5.3 for details), we are adding a box with a list of items implemented in the scope of our prototypes. The “(…)” indicates that more items could be listed. Also, there is the indication of the features that are covered (indicated by the green sign) or not (the red sign). For the last feature in the list, (I) Interaction, we are considering it implemented partially, as our understanding is that more user interaction options should be implemented to have a full advantage of the interactivity benefits.

6. MODEL VALIDATION

In this chapter, we present a qualitative validation to the Preprocessing Profiling Model for VA. We exemplify the capabilities of our proposal Model by presenting its utilization with two different scenarios. First, we describe a hypothetical scenario and persona to show how the Preprocessing Profiling Model supports to understand the raw data under analysis, and the impacts of the preprocessing strategies (Section 6.1). The second use case is considering our initial analysis of an original dataset selected from the UCI Machine Learning Repository (Section 6.2). Lastly, we compile a final discussion on our Model validation (Section 6.3).

6.1 Usage Scenario: Understanding a Dataset and Its Preprocessing Impacts

Tim, a biology student, is searching for strategies on how to solve the taxonomic problems of his current research. He collected data about a new group of flowers, and he is interested in identifying species of flowers by the attributes measured from a morphologic variation of the flowers.

Tim read about ML algorithms that can be used for classification problems. Also, he is aware that not only the ML algorithm is enough for the final solution of his problem, but as well as the preprocessing strategies he selects. Considering that, before using his data, he decides to explore different approaches using a similar known problem of Anderson's Iris dataset [Fis36]. He downloads the dataset from the UCI Machine Learning Repository website [DGc]. The original dataset contains 50 samples from each of three species of Iris, i.e., Iris Setosa, Iris Virginica, and Iris Versicolor. For each sample, four attributes were measured in centimetres: length of sepal, length of petal, width of sepal, and width of petal. Additionally, a fifth column informs the correspondent class of each sample.

Tim is familiar with Python programming development environment and its popular libraries to support with DM problems. He uses the VA workflow to guide his analysis, and the Preprocessing Profiling Model along with its developed prototypes to perform his activities. Under this scenario, his steps related to the Preprocessing Profiling phase are indicated in Figure 6.1. Branch A is related to understanding the data, and it is explained in Subsection 6.1.1, while branch B is related to understanding the impacts of preprocessing, and it is explained in Subsection 6.1.2.

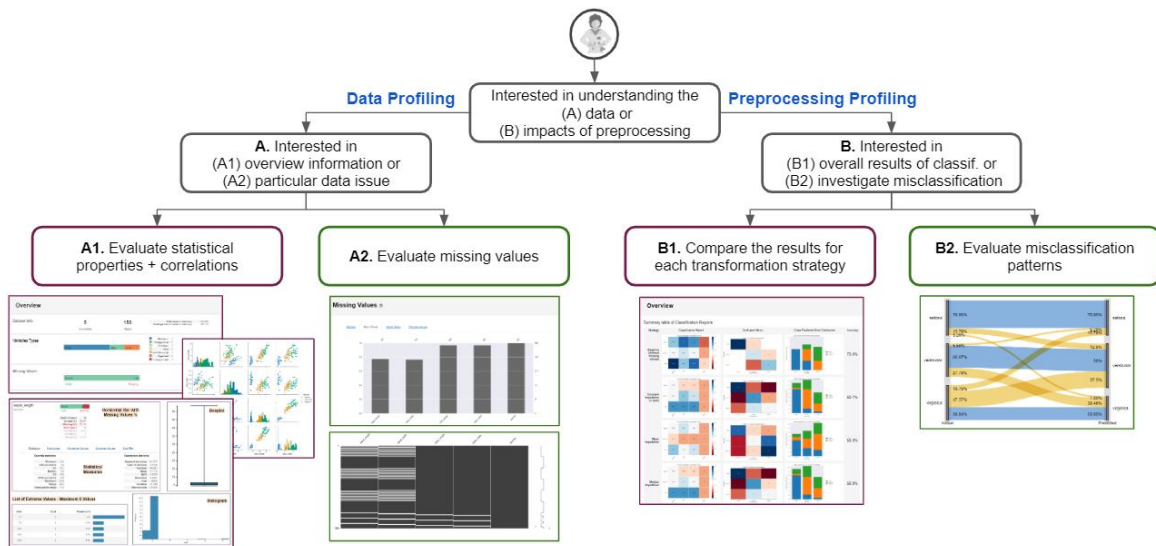


Figure 6.1 – Tim’s steps for Preprocessing Profiling phase and an illustrative sample of the correspondent visualizations for each step.

6.1.1 Understanding the Data

Tim explores the original Iris dataset to get a better understanding of the available raw data. He starts by running descriptive statistics using his programming skills in Python. However, many lines of code and outputs with plain text would be required to generate all the information he wanted. Then, he decides to use the Data Profiling Prototype that is integrated to his Python environment to generate the first report for his data analysis of Iris dataset. While analyzing the first section of the report, he can see some information regarding the number of records (rows) and variables (columns), the dataset size, and variable types distribution, as shown in Figure 6.2.

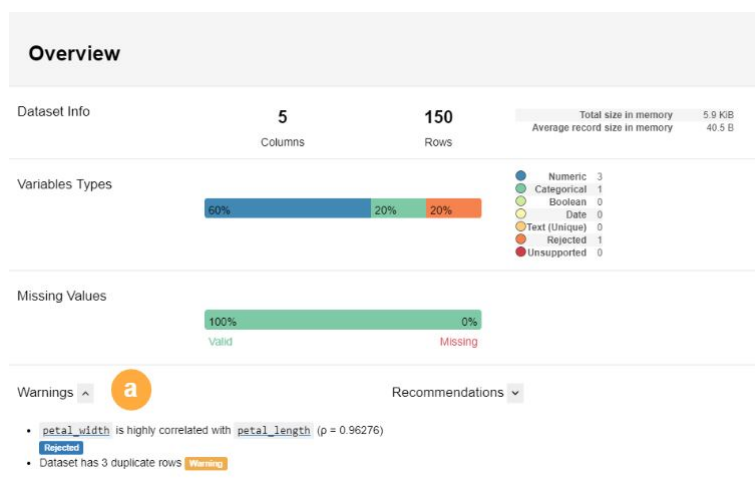


Figure 6.2 – Data Profiling Report - Overview section for Iris dataset. (a) Warning information regarding petal_width column high correlation with petal_length. That is the reason why one variable appears in Rejected status on the Variable Types breakdown.

Tim realizes the petal columns are highly correlated with each other, as indicated in Figure 6.2-a. Even though Tim has previously generated the covariance and correlation matrix, when he was executing his initial set of Python code, he still considers challenging to observe the relation between two variables just by looking at the output with plain text. After all, he confirms this information also while seeing the visualization for Correlation Coefficient.

Next, by evaluating the Relation Matrix visualization (Figure 6.3), two new thoughts arise regarding the group of Iris Setosa. First, there is a point close to zero that differs from most of the other values on the same species, which may indicate an outlier. Second, flowers from Setosa are well separated from the other two classes, while Versicolor and Virginica are overlapping, which may cause some uncertainty during the classification process.

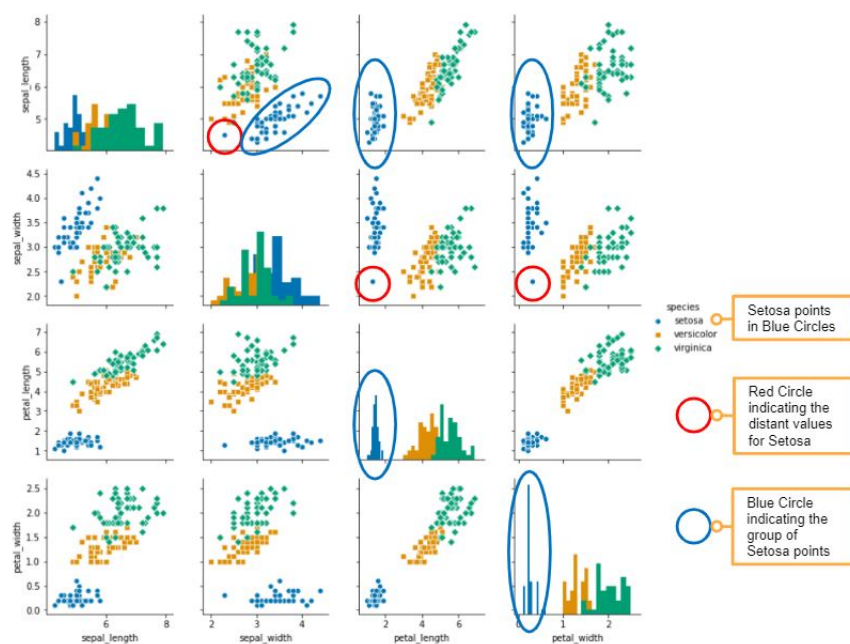


Figure 6.3 – Relation Matrix: pairplot to visualize the similarities and differences between the species.

After completing this initial data exploration and get familiarized with the original Iris dataset details, Tim is now interested in evaluating the impacts of different preprocessing strategies. Mainly, he wants to investigate the possibility of data transformation for missing values. For that, he creates a new dirty dataset for Iris considering the following rules:

- a Duplicate the first 12 instances from each class (total number of rows increased to 186).
- b For this set of duplicated instances, two values were replaced by blank for each variable, i.e., a total of 24 missing values were introduced (8 instances per class). It represents an addition of 16% missing values when considering each class, or 2.6% of missing values when considering the total number of records (186 rows * 5 columns = 930 records).

- c For the remaining duplicated instances (4 per class), 1 value of each variable is replaced by a value eight times higher than the current one, i.e., where it is 3.5 should be changed to 28. It represents an addition of 8% of outliers.
- d Only one value by instance (row) is replaced at a time for the changes (b) and (c). Also, the changes followed a sequential order, starting with the first new duplicated instance.
- e Additional missing values were added to the columns `sepal_length` and `sepal_width` in the first original rows. In the end, the total amount of missing values in the entire dataset is 10.6%.

Considering this new dirty dataset, Tim performs the same sequence of analysis described for the original Iris Dataset. Additionally, he explores the Missing Values section, which was not observed earlier since the original dataset was clean of data issues. The nullity matrix for the new dirty dataset is presented in Figure 6.4. Also, in this new round, he notices how even a small percentage of outliers could cause a significant impact on the overall visual understanding of the variables relationships. For example, he considers challenging to see the cluster of flowers species in the new dirty dataset, as shown in Figure 6.5.

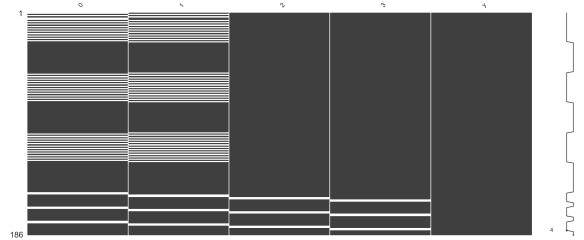


Figure 6.4 – Missing Values Matrix with the dirty dataset for Iris.

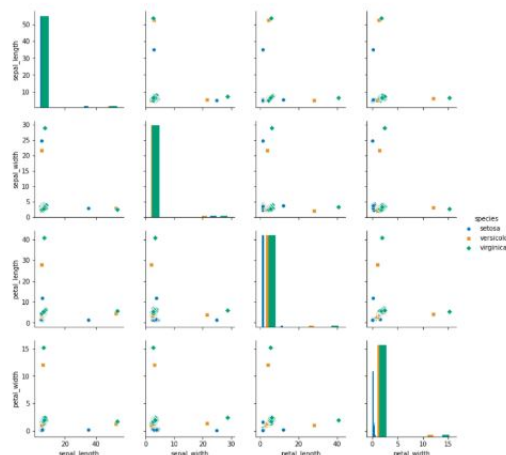


Figure 6.5 – Relation Matrix: pairplot to visualize the similarities and differences between the species.

In that case, Tim examines the details under the Variable section of the Data Profiling Report. For instance, in Figure 6.6 is presented the information for `sepal_length` in different ways, such as the percentage of missing values, histogram and boxplot to see the distribution of values and identify the outliers. As a result, he can confirm the new distribution of missing and extreme values of each variable.

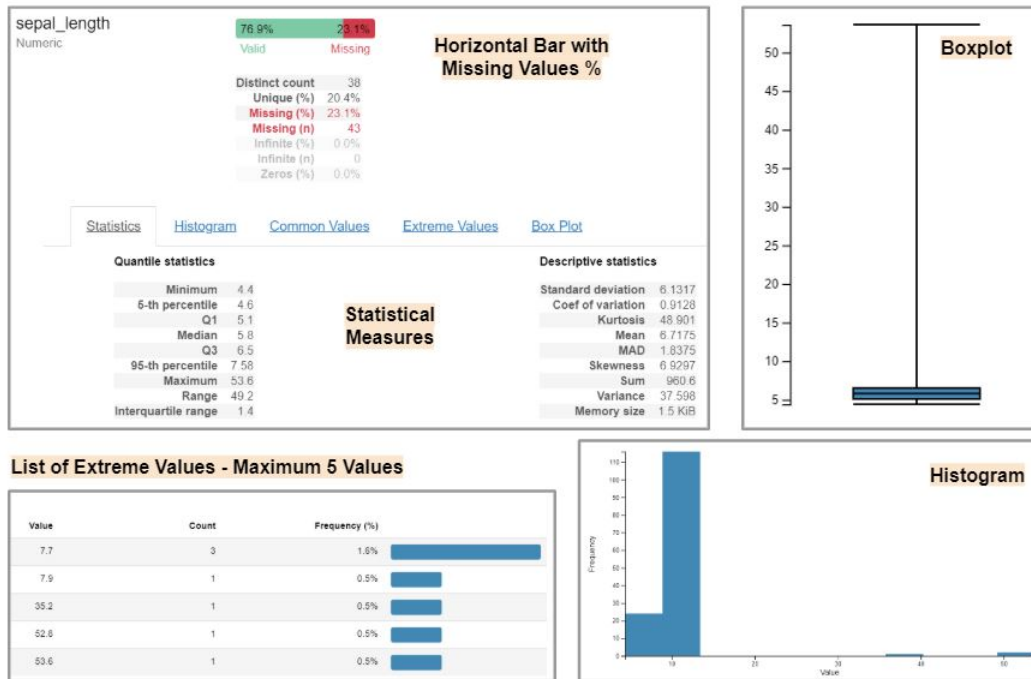


Figure 6.6 – Data Profiling Report - Variable section with detailed information for `sepal_length` variable.

With the completion of his activities in understanding the data, Tim moves to the second branch of his exercise. He is now interested in understanding the impacts of preprocessing strategies for his classification problem, which is the subject of the next subsection.

6.1.2 Understanding the Impacts of Preprocessing

Tim uses the Preprocessing Profiling prototype to build a classification model for the Iris problem and evaluate the quality of the results. He starts by creating a baseline of the flowers classification using the original Iris dataset. Before proceeding with the dirty Iris dataset evaluation, Tim is curious to evaluate four different combinations of variables input in the classification model training. First, he considers all variables. Second, beyond the variable with the classifier (species) only two others are considered: `sepal_length` and `sepal_width`. Third, three variables of the dataset: `petal_width`, `sepal_length`, and `sepal_width`. Finally, he considers another combination of three variables: `petal_length`, `sepal_length`, and `sepal_width`.

In the Preprocessing Profiling prototype, the training and testing data are generated with a constant percentage of distribution, but the rows (instances) are randomly selected every execution. Then, considering the possible variation on the results, Tim runs three times each combination and keeps the results for the round that presented a higher percentage on Accuracy metric. During this comparison, he confirms the previous observation that `petal_length` and `petal_width` columns are highly correlated; consequently, there is no significant impact on the classification results if he uses only one of these columns to train his model. The classification results for these four rounds are shown in Figure 6.7.

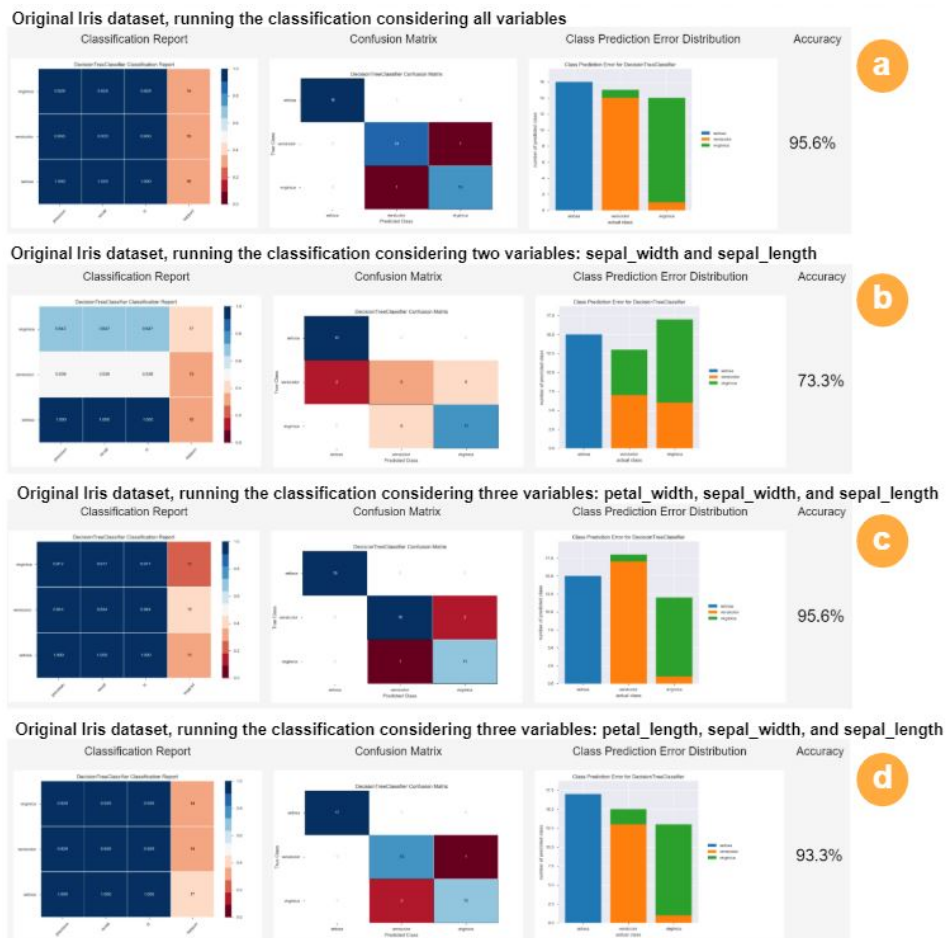


Figure 6.7 – Preprocessing Profiling Report - Baseline results for the classification of Iris original dataset.

Now, using the dirty Iris dataset, he moves to the evaluation of the possible impacts caused by data transformations to handle the missing values. Five different strategies of data imputation are performed. One removes all the rows that have at least one missing value, and it is named as Baseline (no missing). Other replaces all missing values by zero, named as Constant(=zero). A third and fourth replace missing values by mean and median values computed based on all records on the same column, they are named as Mean and Median. The last one replaces missing by the most frequent value in the corresponding column.

Since Tim needs to inform only the dataset as input, running a couple of lines of Python code, he performs multiple rounds using the different combinations of variables as input. Also, he can save the HTML report files generated as the output of each round for further reference. Figure 6.8 shows an overview of the classification results for the round considering only the two variables related to sepal attributes.



Figure 6.8 – Preprocessing Profiling Report - Classification Results for different Missing Values Imputation for Iris dataset with dirty data. The classes are identified as Set (blue) for Iris Setosa, as Ver (orange) for Iris Versicolor, and Vir (green) for Iris Virginica.

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. However, with the presence of data issues and the need to perform imputation strategies, the classification results are negatively impacted. Also, in the example of Figure 6.8, Tim observes a significant variation on the accuracy metric for the Mean imputation strategy when compared to the others. With that, it is clear to him the need of identifying outliers and removing them before continuing, or, for this particular case, he could use the Median to avoid data with high magnitude to dominate results.

Furthermore, while comparing the Flow of Classes visualization for different rounds, he is able to observe two new situations that are not possible with the prior perspectives. First, 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 of test using the four variables (Figure 6.9-a), four imputation strategies present the same accuracy percentage as a result (91.1%). However, he can notice a new 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 6.9-b), he can notice even more variations among the possible combination flows. Second, he considers essential to have different views for the same classification results, mainly when using a dataset with data quality issues. Analyze the same information from different perspectives can support a broader understanding.

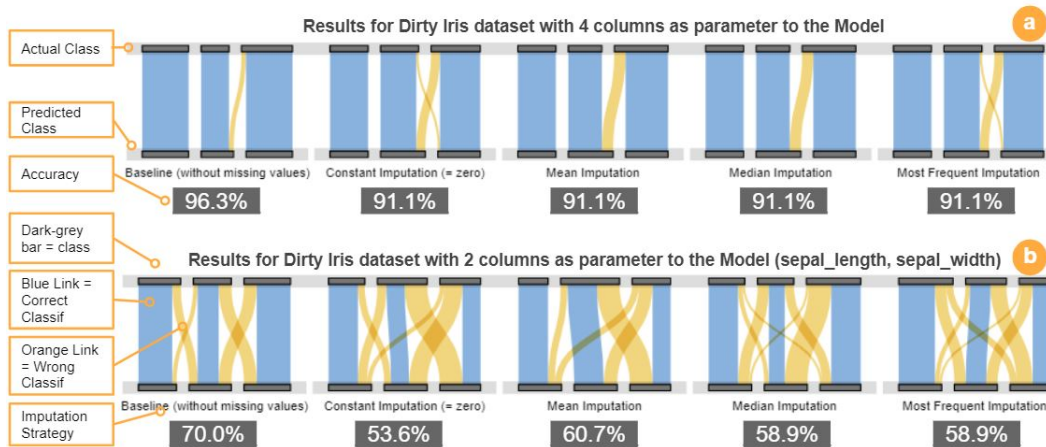


Figure 6.9 – Preprocessing Profiling Report - Flow of Classes Visualization. Classification results for different missing values imputation strategies for Iris dataset with dirty data.

In conclusion, Tim considers these insights a reinforcement of the importance in exploring data transformation strategies, especially when dealing with data issues, before moving to further phases in the VA, or any DM workflow. Beyond that, for any new data analysis engagement, he now considers to combine it with the DM model evaluation during the Preprocessing.

6.2 Use Case: Evaluating a Healthcare Dataset

To proceed with this use case analysis, we looked into online repositories for datasets that could be used in the scope of classification methods. Therefore, we selected a dataset from the UCI Machine Learning Repository related to breast cancer screening method. The dataset contains the discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age [DGd]. Without any previous knowledge on the dataset, we decided to start by running the Data Profiling Prototype to collect information about the dataset and start the data understanding. In the next paragraphs, we are sharing some details on this activity.

First, while reading the information available in the Overview section of the output report, we could confirm the number of columns and rows (Figure 6.10-a), as well as the distribution of variable types (Figure 6.10-b), predominantly numeric. Also, we could observe the presence of missing values and the information of which character was used in the original dataset to represent the not informed values (Figure 6.10-c). Additionally, in the Warnings list (Figure 6.10-d), we could confirm which were the columns with missing values, and a new highlight regarding the highly skewed distribution for one column. It is important to notice the original dataset downloaded did not contain headers, and then the columns appear named as numbers in this report.

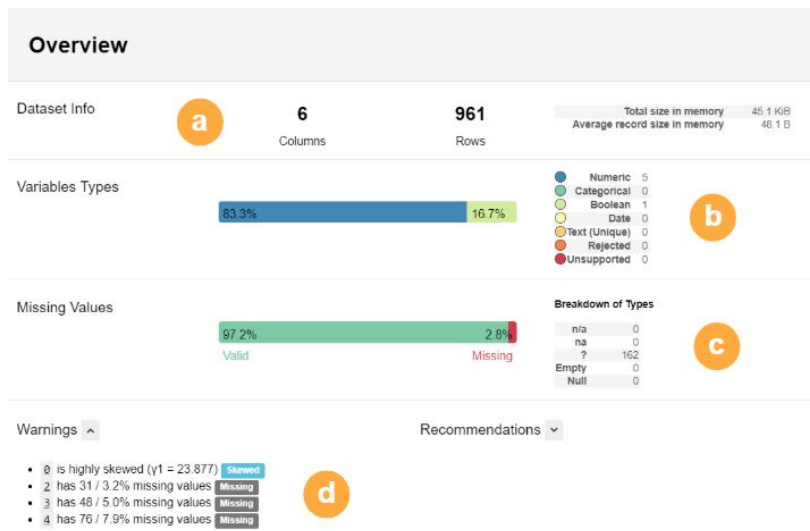


Figure 6.10 – Data Profiling - Section Overview. Detail of the Mammographic Masses dataset. (a) Dataset information with columns, rows, and the size of the dataset. (b) Variable types distribution. (c) Missing values distribution and breakdown of types identified in the dataset. (d) Warnings list.

After, we explored the Variables section. Figure 6.11 shows the information for the first three variables in the dataset composed of 6 columns. The first variable, column 0, presented high positive Skewness. Then, we opened the details for this variable and checked complementary information, seen in Figure 6.12. Subsequently, we were able to notice a possible outlier value (55.0) as highlighted in Figure 6.12-b and c.

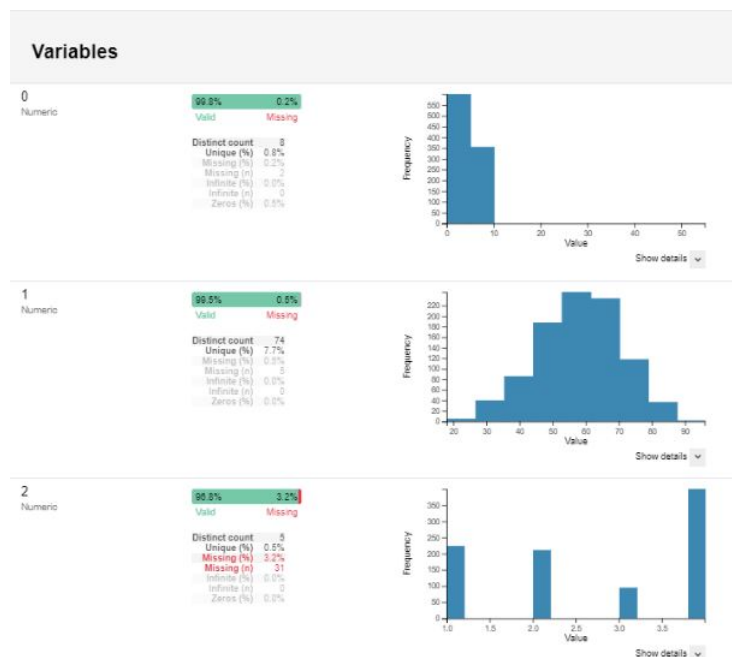


Figure 6.11 – Data Profiling - Section Variables. Detail on the first three columns of the Mammographic Masses dataset.

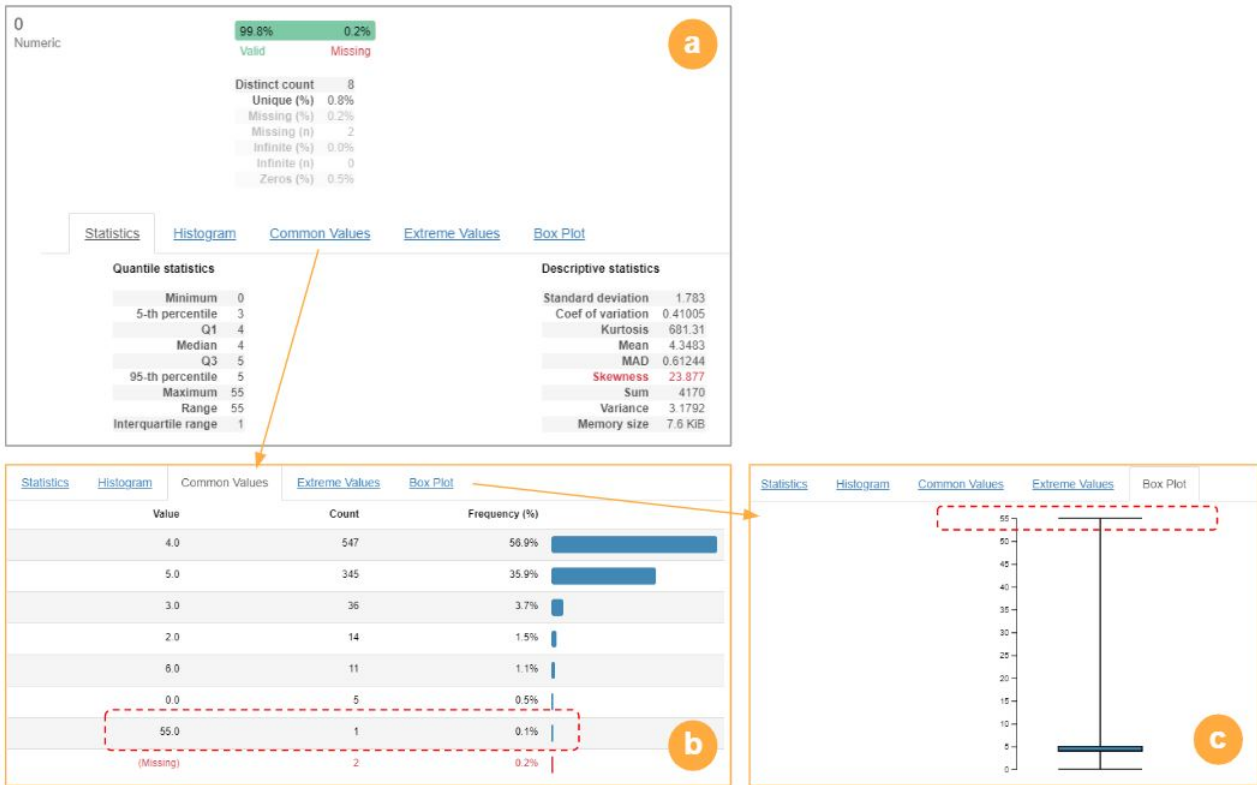


Figure 6.12 – Data Profiling - Section Variables. Detail of the first column of the Mammographic Masses dataset. (a) Statistics for the variable. (b) Common Values in details highlighting the value 55.0 with one occurrence. (c) Boxplot in details for the same variable.

We continued the dataset understanding by evaluating Missing Values section (Figure 6.13). We could observe the higher percentage of missing values in column 4 (7.9%) as initially listed in the Warnings list (Figure 6.10-d). However, non-significant correlation with these missing values and any pattern could be noted.

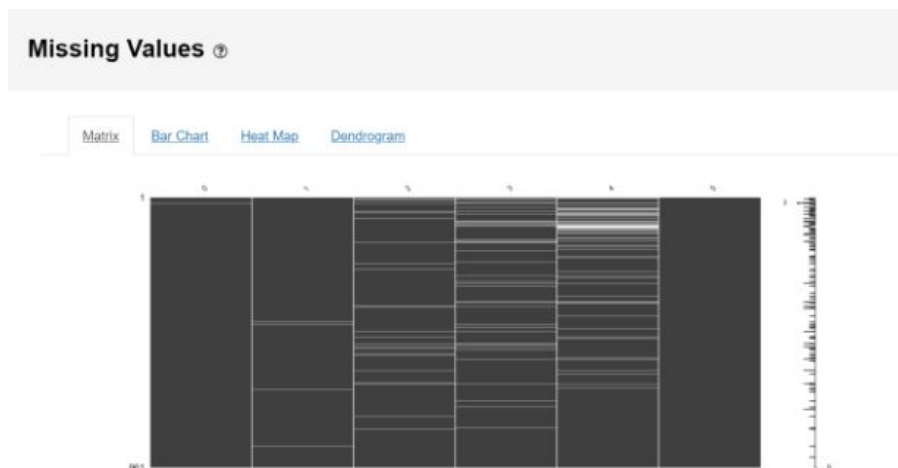


Figure 6.13 – Data Profiling - Section Missing Values. Detail of nullity matrix of the Mammographic Masses dataset.

Additionally, we observed the relationship between each pair of variables using the Spearman's rank correlation coefficient, as illustrated in Figure 6.14. With this visualization, we saw a strong connection between column 2 and 3. This information could be useful in case we need to remove columns to avoid potential bias in the classification.

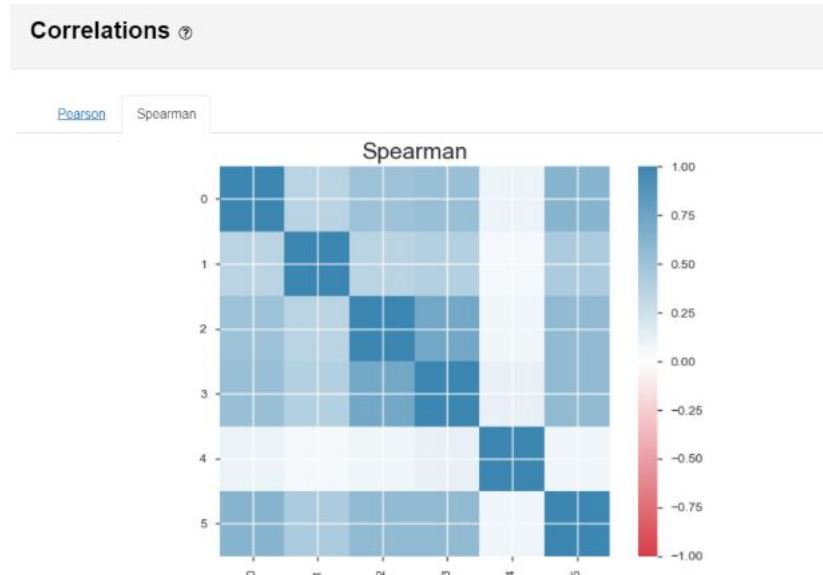


Figure 6.14 – Data Profiling - Section Correlations. Details of Spearman for Mammographic Masses dataset.

As a final step, to confirm the previous observations, we consulted the documentation available for the Mammographic Masses dataset on the UCI Machine Learning Repository website. In Table 6.1, we list the relation of the column in our report with their meaning and the expected values for each variable.

Table 6.1 – List of attributes and respective descriptions for the Mammographic Masses dataset.

Column	Variable Name	Description and Expected Values
0	BI-RADS	1 to 5 (ordinal, non-predictive).
1	Age	patient's age in years (integer).
2	Shape	mass shape: round=1; oval=2; lobular=3; irregular=4 (nominal).
3	Margin	mass margin: circumscribed=1; microlobulated=2; obscured=3; ill-defined=4; spiculated=5 (nominal).
4	Density	mass density: high=1; iso=2; low=3; fat-containing=4 (ordinal).
5	Severity	benign=0 or malignant=1 (binominal, goal field).

Based on that, we were able to conclude: (a) For column 0, BI-RADS assessment, the value "55.0" identified as a potential outlier, in fact, could be considered bad data since the expected values were ranging from 1 to 5. According to the official dataset documentation, the values 0 and 6 were not expected, then we were initially considering them as noise. However, later we confirmed BI-RADS assessment categories in the American College of Radiology [ACR], and we confirmed that 0 means "Incomplete – Need Additional Imaging

Evaluation and/or Prior Mammograms for Comparison” and 6 means “ Known Biopsy-Proven Malignancy”. In any case, this variable should not be used as part of the classification model. (b) Column 2, Shape, and 3, Margin, seemed to be highly correlated. However, there was no indication in the official documentation for that. So, we continued considering both. (c) Column 5, the only variable without missing values, corresponds to Severity, i.e., it contains the class of each instance.

At this point, we completed the initial understanding of the dataset, and we decided to move to the evaluation of the missing values imputation strategies. We used the entire original dataset, except the column 0, BI-RADS.

We ran multiple comparison rounds using the Preprocessing Profile prototype. For all rounds performed, we could observe some variation in the classification results. The maximum variation in accuracy noted was 6.4% between Baseline and Mean imputation strategies, as shown in Figure 6.15. Rather than evaluating the better imputation strategy for this problem, our concerns remained in observe if the visual resources developed as part of the Preprocessing Profile prototype helped to evaluate any possible impacts on the different strategies.



Figure 6.15 – Preprocessing Profiling - Classification Results for different Missing Values Imputation for Mammographic Masses dataset.

In addition to the classification results reports such as Confusion Matrix and Error Distribution, we also used the visualization Flow of Classes in an attempt to find any pattern for misclassification. An example of that is presented in Figure 6.16. Nevertheless, we could not observe any significant insight for the dataset and the imputation strategies in use. The other visualization planned for misclassification analysis, the Matrix of Nullity combined with the information of the Class Prediction Error, could support us in this activity.

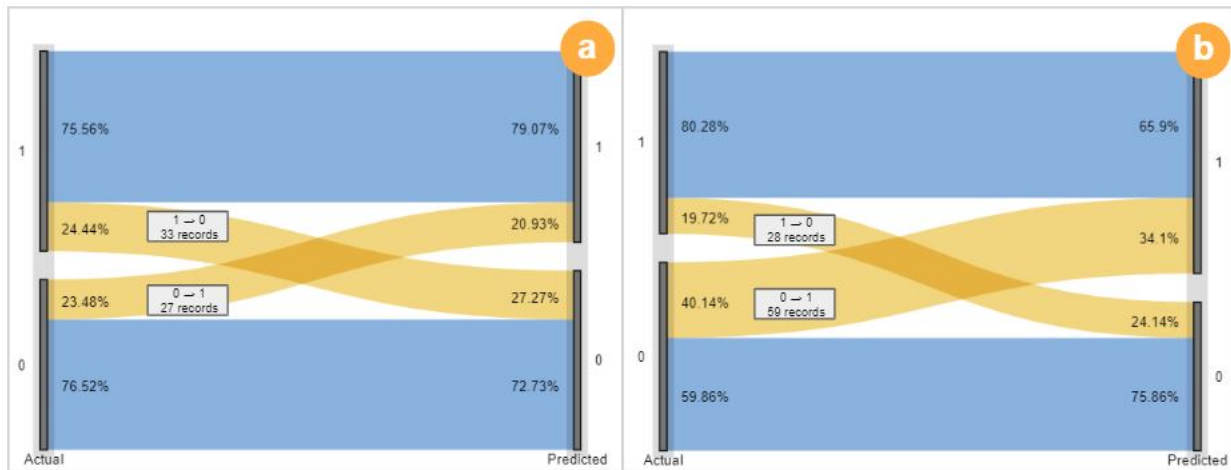


Figure 6.16 – Preprocessing Profiling - Comparison of Classification Results for different Missing Values Imputation for Mammographic Masses dataset. (a) Baseline, missing values removed. (b) Mean Imputation.

6.3 Discussion and Limitations

During the Preprocessing Profiling Model validation we observed that our Model can provide important resources to data analysts while performing the preprocessing activities, which answers our research question “How can we assist the preprocessing activities with visualization techniques during a visual analytics workflow?”. An example within the scope of data profiling is the understanding of the presence of outlier values while analyzing the variables. As shown in Figure 6.6 and 6.12-c, by looking at the boxplot this understanding is facilitated. Turning to the preprocessing strategies evaluation, we could compare the classification results for multiple data transformation strategies at a glance. For example, when showing the Classification Report and Confusion Matrix data combined with heatmap, despite simple, it should improve the perception of the results.

Although most of the visualizations are simple, they still demonstrate more benefits to understand the data when compared to viewing the plain text. Also, the simplicity should favour the understanding, since it does not require a prior explanation as most of them are already part of the culture of the data analysts. Even the visualizations that are bringing a new perspective are designed to be as user-friendly as possible. For instance, the matrix of nullity when using a scheme of colours, which white, or blank spaces, represents missing and dark grey represents valid values, should not require significant efforts to be understood.

Additionally, through practicing on developed prototypes, three main advantages can be mentioned. First, considering we have the dataset loaded in 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 by running a couple of lines of code. Consequently, we are contributing to simplify the working proce-

dures during the Preprocessing phase, that is a big concern since it is frequently reported as the most laborious tasks. 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. Finally, this detailed information about the dataset and data preparation can be used as metadata for preprocessing. Thus, it can be added to the data mining project documentation, helping to build the principle of transparency on activities performed as part of the Preprocessing phase, which is aligned to initiatives such as the European Union General Data Protection Regulation [Com].

Regarding the limitations of the Preprocessing Profiling Model validation, we can list three critical subjects. To begin, a variety of additional experiments could be added as part of our Model validation. Even using the Iris dataset, that is good for the usage scenario purposes, we could have generated different dirty versions as part of our comparison. For example, one alternative could be to automatically introduce multiple patterns for missing values, and later, without knowing which one was generated, trying to recognize the correspondent pattern visually.

Furthermore, not all the planned features were implemented. Both prototypes still require development enhancements to be thoroughly used as part of our Model validation. Particularly to add more user interaction features, instead of presenting static reports and visualizations. As well as the scalability capacity to handle Big Data volumes needs attention, in the same way as described during the Model's Architecture presentation (Section 5.3). Also, the coverage of additional DM problems and data quality issues. After that, we should be able to evaluate the Preprocessing Profiling Model against more complex use case scenarios.

To conclude, although we built the Preprocessing Profiling Model on top of the requirements obtained on the interviews with the data analysts, we did not validate it with them. Thus, as remaining work, we intend to evaluate the Model using the prototypes developed while conducting in-depth interviews or user-centred experiments with the participation of domain experts in VA area and enterprise professionals in DM.

7. CONCLUSION

In our study, we presented the results obtained from the interview process with thirteen data analysts 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. Also, we described in details the methodology used for data collection and the process to derive the ten insights, the most significant outcome of this process.

Based on the list of insights and the review of the related works, we proposed the Preprocessing Profiling Model for Visual Analytics. Next, we explained the Architecture of our Model with a list of features to be contemplated during the implementation of new solutions in this scope. Moreover, we presented the design of a prototype solution that was used as proof of concept during the Model validation in two different usage scenarios.

The main contributions as part of our study can be summarized as following:

- The introduction of the Preprocessing Profiling Model as an alternative to support the data analysts during the Preprocessing phase. By enabling better methods for data understanding and evaluation of preprocessing impacts, it promotes the quality of the data and the decision making on data preparation strategies.
- The organization of the challenges and opportunities identified during our analysis of the interviews, which resulted in a list of ten insights. This list of insights was 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.
- The summarization of practical items to be considered during the planning and development phases 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 a future work, we plan three main items. To begin, (a) append to the list of ten insights a detailed indication of the visualization techniques that can be associated with each insight. Next, (b) continue the development of the Preprocessing Profiling Model prototypes. For Data Profiling, the visualizations concerned with missing values, and for Preprocessing Profiling, the visualizations related to the understanding of misclassification patterns resulting from the different data transformation strategies. To conclude, (c) evaluate the Preprocessing Profiling Model using the prototypes developed while conducting in-depth interviews, or user-centred experiments, with the participation of domain experts.

REFERENCES

- [ACF⁺16] Angelini, M.; Corriero, R.; Franceschi, F.; Geymonat, M.; Mirabelli, M.; Remondino, C.; Santucci, G.; Stabellini, B. “A visual analytics system for mobile telecommunication marketing analysis”. In: *Proceedings of the EuroVis Workshop on Visual Analytics, 2016*, pp. 7–11.
- [ACR] ACR. “American college of radiology BI-RADS atlas 5th edition”. Source: <https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads>, May 2019.
- [AHH⁺14] Alsallakh, B.; Hanbury, A.; Hauser, H.; Miksch, S.; Rauber, A. “Visual methods for analyzing probabilistic classification data”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, Dec 2014, pp. 1703–1712.
- [Alta] Alteryx. “Alteryx”. Source: <https://www.alteryx.com>, May 2019.
- [Altb] Altexsoft. “Machine learning: Bridging between business and data science”. Source: <https://www.altexsoft.com/whitepapers/machine-learning-bridging-between-business-and-data-science/>, May 2019.
- [Apa] Apache. “Apache Spark: Unified analytics engine for big data”. Source: <https://spark.apache.org/>, Dec 2018.
- [AZL⁺19] Alspaugh, S.; Zokaei, N.; Liu, A.; Jin, C.; Hearst, M. A. “Futzing and moseying: Interviews with professional data analysts on exploration practices”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, Jan 2019, pp. 22–31.
- [BB09] Brase, C.; Brase, C. “Understandable statistics: Concepts and methods”. Houghton Mifflin Company, 2009, 816p.
- [BB19] Bengfort, B.; Bilbro, R. “Yellowbrick: Visualizing the Scikit-Learn Model Selection Process”, *Journal of Open Source Software*, vol. 4, Mar 2019, pp. 1–5.
- [BCD⁺09] Berthold, M.; Cebron, N.; Dill, F.; Gabriel, T.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. “KNIME - the Konstanz information miner: Version 2.0 and beyond”, *ACM SIGKDD Explorations Newsletter*, vol. 11, Nov 2009, pp. 26–31.
- [BE18] Batch, A.; Elmqvist, N. “The interactive visualization gap in initial exploratory data analysis”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, Jan 2018, pp. 278–287.

- [BGV16] Bouali, F.; Guettala, A.; Venturini, G. “Vizassist: an interactive user assistant for visual data mining”, *The Visual Computer*, vol. 32, Nov 2016, pp. 1447–1463.
- [BHJ09] Bastian, M.; Heymann, S.; Jacomy, M. “Gephi: an open source software for exploring and manipulating networks”. In: *Proceedings of the AAAI Conference on Weblogs and Social Media*, 2009, pp. 361–362.
- [Bil18] Bilogur, A. “Missingno: a missing data visualization suite”, *Journal of Open Source Software*, vol. 3, Feb 2018, pp. 1–4.
- [BLSK12] Berenson, M.; Levine, D.; Szabat, K.; Krehbiel, T. “Basic business statistics: Concepts and applications”. Pearson Education, 2012, 859p.
- [Bos] Bostock, M. “Data-driven documents”. Source: <https://d3js.org/>, May 2019.
- [Bra97] Bratko, I. “Machine learning: Between accuracy and interpretability”. In: *Proceedings of Learning, Networks and Statistics*, 1997, pp. 163–177.
- [Com] Commission, E. “EU general data protection regulation”. Source: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en, May 2019.
- [Cre14] Creswell, J. “Research Design: Qualitative, Quantitative, and Mixed Methods Approaches”. SAGE Publications, 2014, 273p.
- [Data] Databricks. “Databricks: Making big data simple”. Source: <https://databricks.com/>, Dec 2018.
- [DATb] DATASUS. “Plataforma Brasil”. Source: <http://plataformabrasil.saude.gov.br/login.jsf>, Dec 2018.
- [DCE⁺13] Demšar, J.; Curk, T.; Erjavec, A.; Gorup, V.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; Štajdohar, M.; Umek, L.; Žagar, L.; Žbontar, J.; Žitnik, M.; Zupan, B. “Orange: Data mining toolbox in python”, *Journal of Machine Learning Research*, vol. 14, Jan 2013, pp. 2349–2353.
- [DGa] Dua, D.; Graff, C. “The UCI machine learning repository”. Source: <http://archive.ics.uci.edu/ml>, May 2019.
- [DGb] Dua, D.; Graff, C. “The UCI machine learning repository - cervical cancer (risk factors) data set”. Source: <https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29>, May 2019.
- [DGC] Dua, D.; Graff, C. “The UCI machine learning repository - iris data set”. Source: <https://archive.ics.uci.edu/ml/datasets/Iris>, May 2019.

- [DGd] Dua, D.; Graff, C. “The UCI machine learning repository - mammographic mass data set”. Source: <https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>, May 2019.
- [DJ03] Dasu, T.; Johnson, T. “Exploratory Data Mining and Data Cleaning”. John Wiley & Sons, 2003, 203p.
- [dMGG15] de Mauro, A.; Greco, M.; Grimaldi, M. “What is big data? a consensual definition and a review of key research topics”. In: Proceedings of the AIP Conference, 2015, pp. 97–104.
- [dOL03] de Oliveira, M. C.; Levkowitz, H. “From visual data exploration to visual data mining: A survey”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, Jul 2003, pp. 378–394.
- [EF10] Elmqvist, N.; Fekete, J.-D. “Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, May 2010, pp. 439–454.
- [EPD05] Eaton, C.; Plaisant, C.; Drisd, T. “Visualizing missing data: Classification and empirical study”. In: Proceedings of the Conference on Human-Computer Interaction, 2005, pp. 861–872.
- [Fis36] Fisher, R. “The use of multiple measurements in taxonomic problems”, *Annals of Eugenics*, vol. 7, Sep 1936, pp. 179–188.
- [FLGC11] Faceli, K.; Lorena, A.; Gama, J.; Carvalho, A. “Inteligência artificial: uma abordagem de aprendizado de máquina”. LTC, 2011, 396p.
- [Fou] Foundation, R. “The R project for statistical computing”. Source: <https://www.r-project.org/>, Dec 2018.
- [FP16] Fekete, J.-D.; Primet, R. “Progressive analytics: A computation paradigm for exploratory data analysis”, *arXiv preprint*, vol. 1607.05162, Jul 2016, pp. 1–10.
- [FWR+17] Federico, P.; Wagner, M.; Rind, A.; Amor-Amorós, A.; Miksch, S.; Aigner, W. “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, 2017, pp. 92–103.
- [GE03] Guyon, I.; Elisseeff, A. “An introduction to variable and feature selection”, *Journal of Machine Learning Research*, vol. 3, Mar 2003, pp. 1157–1182.
- [Gep] Gephi. “The open graph viz platform”. Source: <https://gephi.org/>, Dec 2018.

- [Goo] Google. “Facets: Visualizations for ML datasets - PAIR”. Source: <https://pair-code.github.io/facets/>, Dec 2018.
- [Had] Hadoop, A. “Apache hadoop”. Source: <https://hadoop.apache.org/>, Dec 2018.
- [Hel08] Hellerstein, J. “Quantitative data cleaning for large databases”, White paper, United Nations Economic Commission for Europe, 2008, 42p.
- [HH] Holtz, Y.; Healy, C. “from data to viz”. Source: <https://www.data-to-viz.com/>, May 2019.
- [HHK15] Heer, J.; Hellerstein, J.; Kandel, S. “Predictive interaction for data transformation”. In: Proceedings of the Biennial Conference on Innovative Data Systems Research, 2015, pp. 1–7.
- [HK12] Heer, J.; Kandel, S. “Interactive analysis of big data”, *XRDS: Crossroads, The ACM Magazine for Students - Big Data*, vol. 19, Sep 2012, pp. 50–54.
- [HKP11] Han, J.; Kamber, M.; Pei, J. “Data Mining: Concepts and Techniques”. Morgan Kaufmann Publishers, 2011, 744p.
- [Hun07] Hunter, J. “Matplotlib: A 2D graphics environment”, *Computing in Science & Engineering*, vol. 9, May-Jun 2007, pp. 90–95.
- [Int10] International, D. “The DAMA Guide to the Data Management Body of Knowledge”. Technics Publications, 2010, 406p.
- [Joh09] Johnson, T. “Data profiling”, *Encyclopedia of Database Systems*, 2009, pp. 604–608.
- [Jug14] Jugulum, R. “Competing with High Quality Data: Concepts, Tools, and Techniques for Building a Successful Approach to Data Quality”. John Wiley & Sons, 2014, 304p.
- [KAF+08] Keim, D.; Andrienko, G.; Fekete, J.-D.; Görg, C.; Kohlhammer, J.; Melançon, G. “Visual analytics: Definition, process, and challenges”. In: *Information visualization*, Springer, 2008, pp. 154–175.
- [KCH+03] Kim, W.; Choi, B.; Hong, E.; Kim, S.; Lee, D. “A taxonomy of dirty data”, *Data Mining and Knowledge Discovery*, vol. 7, Jan 2003, pp. 81–99.
- [Kei00] Keim, D. “Designing pixel-oriented visualization techniques: theory and applications”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, Jan 2000, pp. 59–78.
- [Kei01] Keim, D. “Visual exploration of large data sets”, *Communications of the ACM*, vol. 44, Aug 2001, pp. 38–44.

- [KKA95] Keim, D.; Kriegel, H.; Ankerst, M. “Recursive pattern: a technique for visualizing very large amounts of data”. In: Proceedings of the Conference on Visualization, 1995, pp. 279–286.
- [KKE10] Keim, D.; Kohlhammer, J.; Ellis, G. “Mastering the Information Age: Solving Problems with Visual Analytics”. Eurographics Association, 2010, 168p.
- [KMT10] Keim, D.; Mansmann, F.; Thomas, J. “Visual analytics: How much visualization and how much analytics?”, *ACM SIGKDD Explorations Newsletter*, vol. 11, May 2010, pp. 5–8.
- [KNI] KNIME. “KNIME: Open for innovation”. Source: <https://www.knime.com/>, Dec 2018.
- [KPB14] Krause, J.; Perer, A.; Bertini, E. “INFUSE: Interactive feature selection for predictive modeling of high dimensional data”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, Dec 2014, pp. 1614–1623.
- [KPHH11] Kandel, S.; Paepcke, A.; Hellerstein, J.; Heer, J. “Wrangler: Interactive visual specification of data transformation scripts”. In: Proceedings of the Conference on Human Factors in Computing Systems, 2011, pp. 3363–3372.
- [KPHH12] Kandel, S.; Paepcke, A.; Hellerstein, J.; Heer, J. “Enterprise data analysis and visualization: An interview study”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, Dec 2012, pp. 2917–2926.
- [KPP+12] Kandel, S.; Parikh, R.; Paepcke, A.; Hellerstein, J.; Heer, J. “Profiler: Integrated statistical analysis and visualization for data quality assessment”. In: Proceedings of the Conference on Advanced Visual Interfaces, 2012, pp. 547–554.
- [KPS16] Krause, J.; Perer, A.; Stavropoulos, H. “Supporting iterative cohort construction with visual temporal queries”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, Jan 2016, pp. 91–100.
- [KS14] Kindlmann, G.; Scheidegger, C. “An algebraic process for visualization design”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, Dec 2014, pp. 2181–2190.
- [KT16] Kowarik, A.; Templ, M. “Imputation with the R package VIM”, *Journal of Statistical Software*, vol. 74, Oct 2016, pp. 1–16.
- [LBI+12] Lam, H.; Bertini, E.; Isenberg, P.; Plaisant, C.; Carpendale, S. “Empirical studies in information visualization: Seven scenarios”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, Sep 2012, pp. 1520–1536.

- [LCM⁺17] Lu, J.; Chen, W.; Ma, Y.; Ke, J.; Li, Z.; Zhang, F.; Maciejewski, R. “Recent progress and trends in predictive visual analytics”, *Frontiers of Computer Science*, vol. 11, Apr 2017, pp. 192–207.
- [LFH17] Lazar, J.; Feng, J.; Hochheiser, H. “Research Methods in Human-Computer Interaction”. Elsevier Science, 2017, 560p.
- [LGH⁺17] Lu, Y.; Garcia, R.; Hansen, B.; Gleicher, M.; Maciejewski, R. “The state-of-the-art in predictive visual analytics”, *Computer Graphics Forum*, vol. 36, Jun 2017, pp. 539–562.
- [LR02] Little, R.; Rubin, D. “Statistical Analysis with Missing Data”. John Wiley & Sons, 2002, 408p.
- [Mar14] Marsland, S. “Machine learning: an algorithmic perspective”. Chapman and Hall/CRC, 2014, 457p.
- [MH08] Maaten, L.; Hinton, G. “Visualizing data using t-SNE”, *Journal of Machine Learning Research*, vol. 9, Nov 2008, pp. 2579–2605.
- [Mit97] Mitchell, T. “Machine Learning”. McGraw-Hill, 1997, 421p.
- [Mol] Molnar, C. “Interpretable machine learning: A guide for making black box models explainable”. Source: <https://christophm.github.io/interpretable-ml-book/>, May 2019.
- [Ope] OpenRefine. “Openrefine”. Source: <http://openrefine.org/>, Dec 2018.
- [Ora] Orange. “Orange: Data mining fruitful and fun”. Source: <https://orange.biolab.si/>, Dec 2018.
- [Pan] Pandas-profiling. “Create HTML profiling reports from pandas dataframe objects”. Source: <https://github.com/pandas-profiling/pandas-profiling>, May 2019.
- [PF91] Piatetski, G.; Frawley, W. “Knowledge Discovery in Databases”. MIT Press, 1991, 540p.
- [Pyt] Python. “Python”. Source: <https://www.python.org/>, Dec 2018.
- [Qli] Qlik. “Qlik: Data analytics for modern business intelligence”. Source: <https://www.qlik.com>, Dec 2018.
- [RC94] Rao, R.; Card, S. “The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information”. In: *Proceedings of the Conference on Human Factors in Computing Systems*, 1994, pp. 318–322.

- [RD00] Rahm, E.; Do, H. "Data cleaning: Problems and current approaches", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 23, Dec 2000, pp. 3–13.
- [Ren00] Rensink, R. "Seeing, sensing, and scrutinizing", *Vision Research*, vol. 40, Jun 2000, pp. 1469–1487.
- [RF16] Ribarsky, W.; Fisher, B. "The human-computer system: Towards an operational model for problem solving". In: *Proceedings of the Hawaii International Conference on System Sciences*, 2016, pp. 1446–1455.
- [SAS] SAS. "Sas analytics". Source: <https://www.sas.com/>, Dec 2018.
- [SCS⁺17] Smilkov, D.; Carter, S.; Sculley, D.; Viégas, F.; Wattenberg, M. "Direct-manipulation visualization of deep networks", *arXiv preprint*, vol. 1708.03788, Aug 2017, pp. 1–5.
- [She00] Shearer, C. "The CRISP-DM model: the new blueprint for data mining", *Journal of Data Warehousing*, vol. 5, Fall 2000, pp. 13–22.
- [Shn96] Shneiderman, B. "The eyes have it: A task by data type taxonomy for information visualizations". In: *Proceedings of the IEEE Symposium on Visual Languages*, 1996, pp. 336–343.
- [SKBE17] Saket, B.; Kim, H.; Brown, E.; Endert, A. "Visualization by demonstration: An interaction paradigm for visual data exploration", *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, Jan 2017, pp. 331–340.
- [SPG14] Stolper, C.; Perer, A.; Gotz, D. "Progressive visual analytics: User-driven visual exploration of in-progress analytics", *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, Dec 2014, pp. 1653–1662.
- [SSS⁺14] Sacha, D.; Stoffel, A.; Stoffel, F.; Kwon, B. C.; Ellis, G.; Keim, D. A. "Knowledge generation model for visual analytics", *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, Dec 2014, pp. 1604–1613.
- [ST17] Sjöbergh, J.; Tanaka, Y. "Visualizing missing values". In: *Proceedings of the Conference Information Visualisation*, 2017, pp. 242–249.
- [Taba] Tableau. "Tableau". Source: <http://www.tableau.com/>, Dec 2018.
- [Tabb] Tableau. "Tableau Prep". Source: <https://www.tableau.com/products/prep>, May 2019.

- [TAF12] Templ, M.; Alfons, A.; Filzmoser, P. “Exploring incomplete data using visualization techniques”, *Advances in Data Analysis and Classification*, vol. 6, Apr 2012, pp. 29–47.
- [TAKP] Templ, M.; Alfons, A.; Kowarik, A.; Prantner, B. “VIM: Visualization and imputation of missing values”. Source: <https://cran.r-project.org/web/packages/VIM/index.html>, Dec 2018.
- [TC05] Thomas, J.; Cook, K. “Illuminating the path: the research and development agenda for visual analytics”. IEEE Computer Society, 2005, 186p.
- [Ten] TensorFlow. “A neural network playground”. Source: <https://playground.tensorflow.org>, May 2019.
- [TPB+18] Turkay, C.; Pezzotti, N.; Binnig, C.; Strobel, H.; Hammer, B.; Keim, D.; Fekete, J.-D.; Palpanas, T.; Wang, Y.; Rusu, F. “Progressive data science: Potential and challenges”, *arXiv preprint*, vol. 1812.08032, Dec 2018, pp. 1–10.
- [Tri] Trifacta. “Trifacta data wrangling tools & software”. Source: <https://www.trifacta.com/>, Dec 2018.
- [TSK06] Tan, P.; Steinbach, M.; Kumar, V. “Introduction to Data Mining”. Pearson Education, 2006, 769p.
- [Tuk62] Tukey, J. “The future of data analysis”, *The Annals of Mathematical Statistics*, vol. 33, Mar 1962, pp. 1–67.
- [Tuk77] Tukey, J. “Exploratory Data Analysis”. Addison-Wesley Publishing Company, 1977, 688p.
- [vZR17] von Zernichow, B.; Roman, D. “Usability of visual data profiling in data cleaning and transformation”. In: *Proceedings of the On the Move to Meaningful Internet Systems*, 2017, pp. 480–496.
- [War04] Ware, C. “Information Visualization: Perception for Design”. Elsevier Science, 2004, 512p.
- [Was] Waskom, M. “seaborn: statistical data visualization”. Source: <https://seaborn.pydata.org/>, Dec 2018.
- [WFW+17] Wang, Z.; Ferreira, N.; Wei, Y.; Bhaskar, A.; Scheidegger, C. “Gaussian cubes: Real-time modeling for visual exploration of large multidimensional datasets”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, Jan 2017, pp. 681–690.

- [WGK15] Ward, M.; Grinstein, G.; Keim, D. “Interactive Data Visualization: Foundations, Techniques, and Applications”. AK Peters/CRC Press, 2015, 548p.
- [Wic10] Wickham, H. “ggplot2: Elegant Graphics for Data Analysis”. Springer New York, 2010, 213p.
- [Wij05] van Wijk, J. “The value of visualization”. In: Proceedings of the IEEE Visualization, 2005, pp. 79–86.
- [WMA⁺16] Wongsuphasawat, K.; Moritz, D.; Anand, A.; Mackinlay, J.; Howe, B.; Heer, J. “Voyager: Exploratory analysis via faceted browsing of visualization recommendations”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, Jan 2016, pp. 649–658.
- [Woh14] Wohlin, C. “Guidelines for snowballing in systematic literature studies and a replication in software engineering”. In: Proceedings of the Conference on Evaluation and Assessment in Software Engineering, 2014, pp. 1–10.
- [Won99] Wong, P. “Guest editor’s introduction: Visual data mining”, *IEEE Computer Graphics and Applications*, vol. 19, Sep 1999, pp. 20–21.
- [WT04] Wong, P.; Thomas, J. “Guest editor’s introduction: Visual analytics”, *IEEE Computer Graphics and Applications*, vol. 24, Sep 2004, pp. 20–21.

APPENDIX A – VISUALIZATION TECHNIQUES

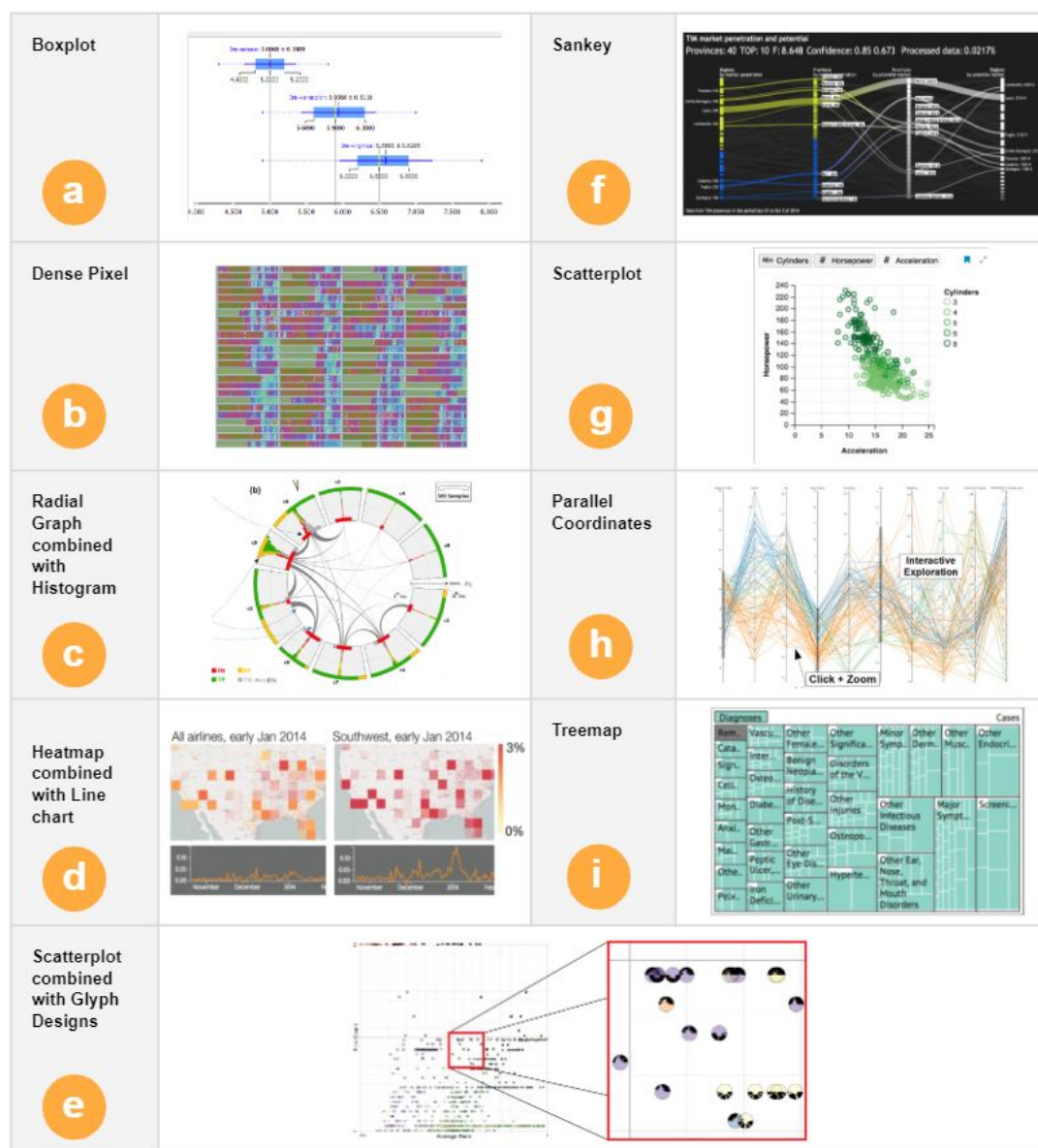


Figure A.1 – Example of visualization techniques used in studies on visual data exploration: (a) Boxplot [Ora]; (b) Matrix with Dense Pixel [KKA95]; (c) Radial Graph combined with Histogram [AHH⁺14]; (d) Heatmap combined with Line chart [WFW⁺17]; (e) Scatterplot combined with Glyphs [KPB14]; (f) Sankey [ACF⁺16]; (g) Scatterplot [WMA⁺16]; (h) Parallel Coordinates [BGV16]; (i) Treemap [KPS16].

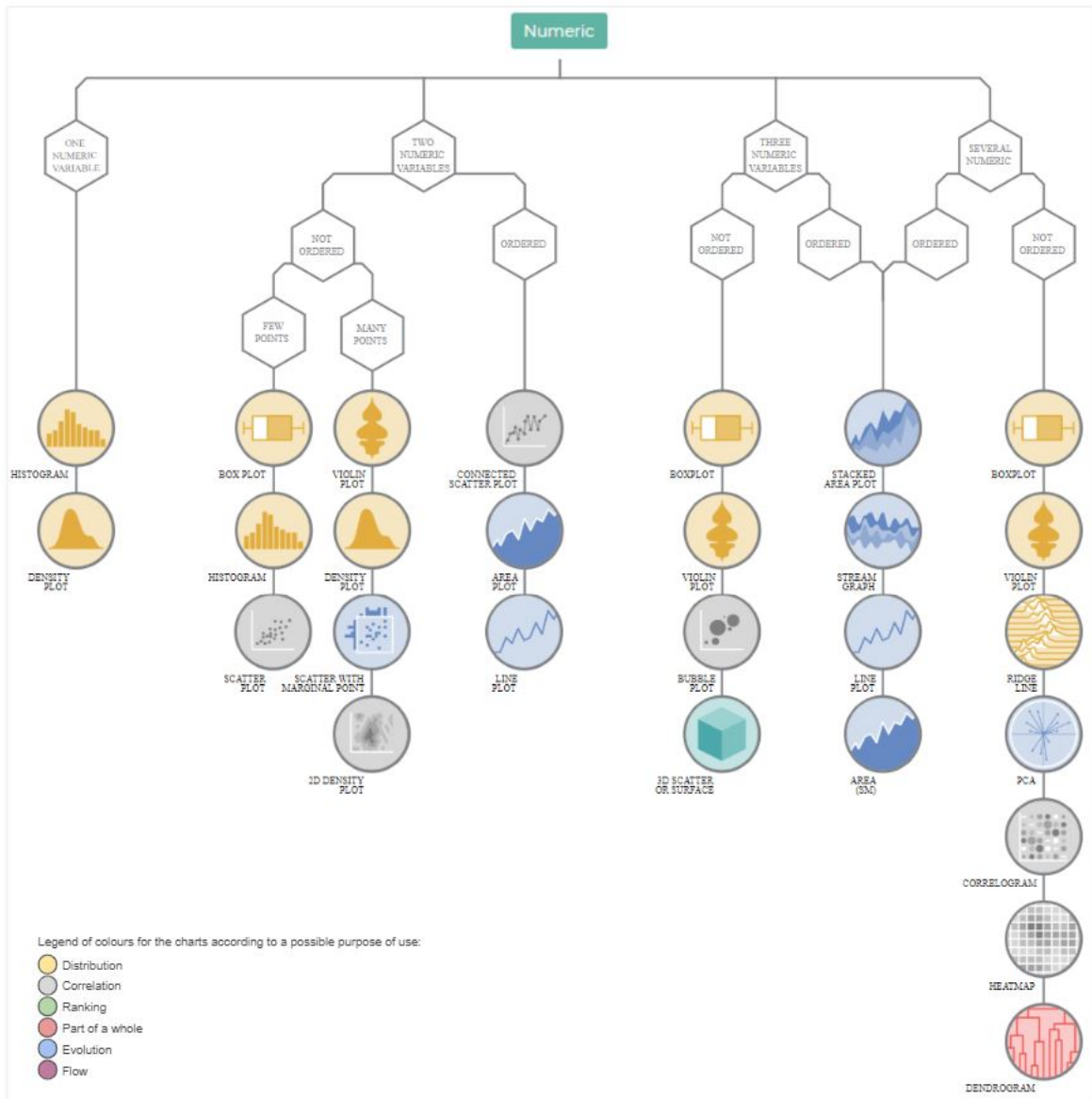


Figure A.2 – Example of a decision tree to select the most appropriate visualization technique for numeric data [HH].

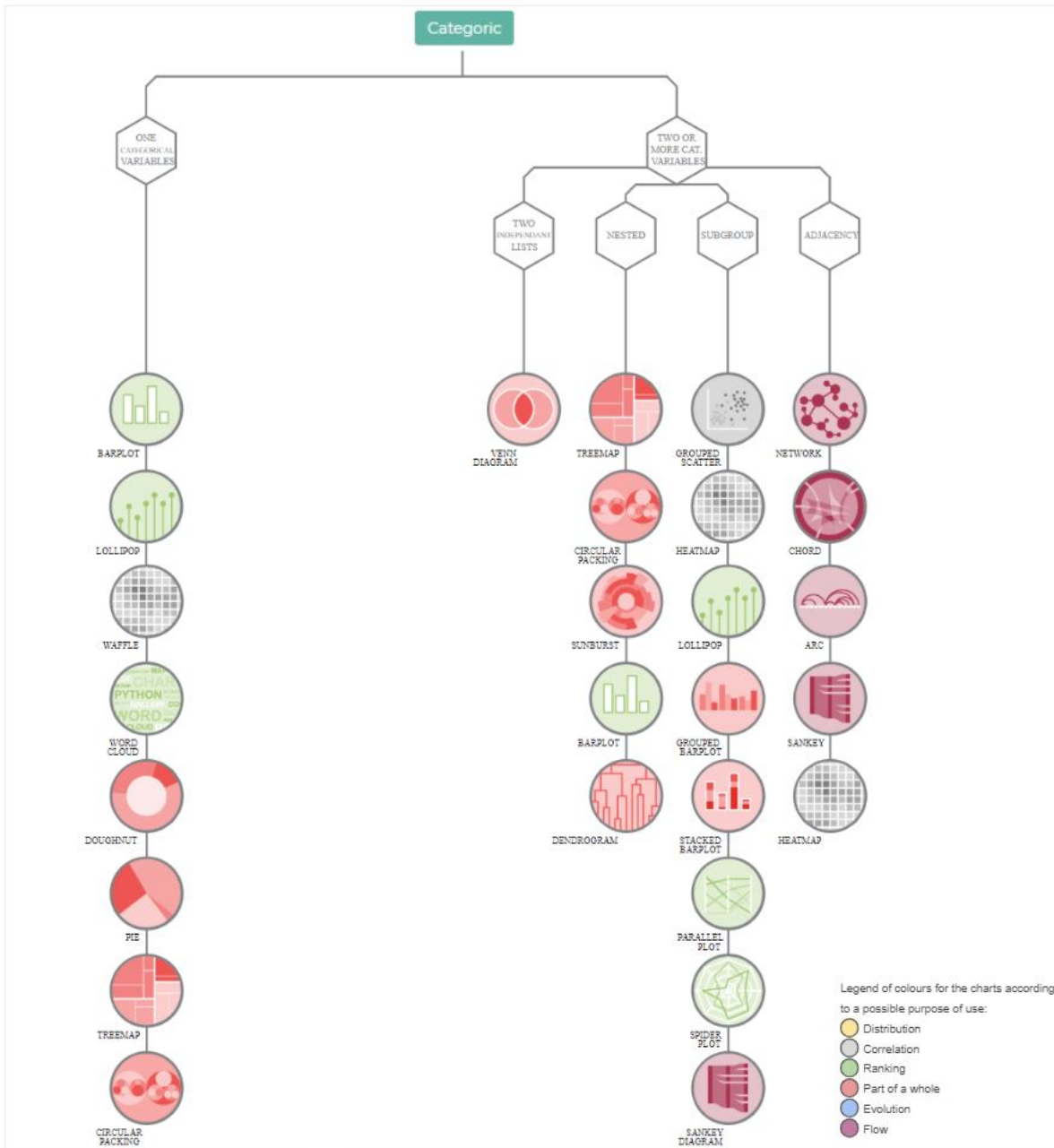


Figure A.3 – Example of a decision tree to select the most appropriate visualization technique for categorical data [HH].

APPENDIX B – INTERVIEW PROCESS: QUESTIONNAIRE

Data collection instrument developed to guide semi-structured interview.

Part 1 - 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. 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 2 - 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 3 - 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 2.3.

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.

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, Rapid-Miner, Alteryx, Anaconda, H2O.ai, Teradata, and others.

Part 4 - 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 6 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 5 - 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*.

APPENDIX C – INTERVIEW PROCESS: CONSENT FORM

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Nós, Alessandra Maciel Paz Milani (aluna de mestrado) e Isabel Harb Manssour (professora orientadora), responsáveis pela pesquisa **técnicas de visualização para entender e analisar conjuntos de dados**, estamos fazendo um convite para você participar como voluntário nesse estudo.

Esta pesquisa pretende desenvolver um modelo de visualização de dados para auxiliar na exploração de dados durante a fase de pré-processamento da descoberta do conhecimento em banco de dados. Para isso precisamos mapear atividades e recursos de visualização de dados utilizados durante esse processo e a avaliação do modelo proposto. Não há benefícios a curto prazo para os participantes dessa pesquisa, contudo, ao término desse estudo são esperadas duas contribuições principais: otimização das tarefas de preparação dos dados e diminuição de problemas com a qualidade dos dados durante sua preparação e transformação.

Para a coleta dos dados poderão ser utilizadas diferentes técnicas, tais como: entrevista seguindo roteiro semiestruturado e observação do desenvolvimento das atividades pré-estabelecidas no sistema sob análise. Entendemos que há riscos mínimos durante essas atividades como: divulgação de dados confidenciais (quebra de sigilo) e desconforto ou constrangimento durante gravações de áudio e/ou vídeo. Lembrando que o objetivo deste estudo não é avaliar o participante, mas, sim, avaliar os processos de trabalho e o sistema computacional que o participante estará usando durante o teste. O uso que se faz dos registros efetuados durante o teste é estritamente limitado a atividades acadêmicas e buscaremos garantir seu anonimato e confidencialidade.

Outras informações importantes:

- As informações desta pesquisa serão confidenciais, e serão divulgadas apenas em eventos ou publicações científicas, não havendo identificação dos participantes, a não ser entre os responsáveis pelo estudo, sendo assegurado o sigilo sobre sua participação.
- Você tem garantido o seu direito de não aceitar participar ou de retirar sua permissão, a qualquer momento, sem nenhum tipo de prejuízo ou retaliação, pela sua decisão.
- Você tem direito ao ressarcimento das despesas diretamente decorrentes de sua participação na pesquisa, como custo de transporte para deslocamento e/ou lanche.
- Ao assinar este termo de consentimento, você não abre mão de nenhum direito legal que teria de outra forma.
- Somente assine este termo de consentimento a menos que tenha tido a oportunidade de fazer perguntas e tenha recebido respostas satisfatórias para suas dúvidas.
- Se você concordar em participar deste estudo, você rubricará todas as páginas e assinará e datará duas vias originais deste termo de consentimento. Você receberá uma das vias para seus registros e a outra será arquivada pelo responsável pelo estudo.
- Durante todo o período da pesquisa você tem o direito de esclarecer qualquer dúvida ou pedir qualquer outro esclarecimento, bastando para isso entrar em contato com:
 - Alessandra – telefone (51) 98415-1686; e-mail alessandra.paz@acad.pucrs.br
 - Isabel – telefone (51) 99955-4948; e-mail isabel.massour@pucrs.br
- Caso você tenha qualquer dúvida quanto aos seus direitos como participante de pesquisa, entre em contato com Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Rio Grande do Sul (CEP-PUCRS) em (51) 33203345, Av. Ipiranga, 6681/prédio 50, sala 703, Porto Alegre – RS, e-mail: cep@pucrs.br, de segunda a sexta-feira das 8h às 12h e das 13h30 às 17h. O CEP é um órgão independente constituído de profissionais das diferentes áreas do conhecimento e membros da comunidade. Sua responsabilidade é garantir a proteção dos direitos, a segurança e o bem-estar dos participantes por meio da revisão e da aprovação do estudo, entre outras ações.

Rubrica do participante

Rubrica dos pesquisadores responsáveis

Figure C.1 – Consent form used on the interview study - page 1 of 2.

Eu, _____, após a leitura deste documento e de ter tido a oportunidade de conversar com o pesquisador responsável, para esclarecer todas as minhas dúvidas, acredito estar suficientemente informado, ficando claro para mim que minha participação é voluntária e que posso retirar este consentimento a qualquer momento sem penalidades ou perda de qualquer benefício. Estou ciente também dos objetivos da pesquisa, dos procedimentos aos quais serei submetido, dos possíveis danos ou riscos deles provenientes e da garantia de confidencialidade e esclarecimentos sempre que desejar.

Diante do exposto expresse minha concordância de espontânea vontade em participar deste estudo.


Assinatura do participante da pesquisa

Assinatura de uma testemunha

DECLARAÇÃO DO PROFISSIONAL QUE OBTVEU O CONSENTIMENTO

Expliquei integralmente este estudo ao participante. Na minha opinião e na opinião do participante, houve acesso suficiente às informações, incluindo riscos e benefícios, para que uma decisão consciente seja tomada.

Data: _____



Alessandra Maciel Paz Milani
Aluna de Mestrado em Ciência da Computação
PPGCC – Escola Politécnica



Isabel Harb Manssour
Professora do PPGCC – Escola Politécnica

Figure C.2 – Consent form used on the interview study - page 2 of 2

APPENDIX D – PROTOTYPE TECHNOLOGIES

Table D.1 – List of the main technologies used for the prototype development.

Name	Scope	Available on	Data Profiling	Preprocessing Profiling
python	Programming Environment	https://www.python.org/	3.7.4	3.7.4
pandas	Python - Dataset manipulation	https://pandas.pydata.org/	0.23.4	0.23.4
numpy	Python - Scientific Computing	https://numpy.org/	1.15.4	1.15.4
matplotlib	Python - Visualization - multiple charts	https://matplotlib.org/2.1.2/index.html	3.0.2	3.0.2
missingno	Python - Visualization - Missing values	https://github.com/ResidentMario/missingno	0.4.2	Custom version
jinja2	Python - Template management	http://jinja.pocoo.org/docs/2.10/	2.10.0	2.10.0
sklearn	Python - Machine learning algorithm	https://scikit-learn.org/stable/	-	0.21.2
yellowbrick	Python - Visualization - ML (sklearn)	https://www.scikit-yb.org/en/latest/	-	0.9.1
six	Python - Compatibility	https://pypi.org/project/six/	1.12.0	-
d3	Javascript - Visualization - Multiple charts	https://d3js.org/	5.9.7	5.9.7
d3 array	Javascript - Visualization	https://github.com/d3/d3-array	-	1.2.4
d3 path	Javascript - Visualization	https://github.com/d3/d3-path	-	1.0.7
d3 shape	Javascript - Visualization	https://github.com/d3/d3-shape	-	1.3.5
d3 sankey	Javascript - Visualization - Sankey chart	https://github.com/d3/d3-sankey	-	0.12.1
jquery	Web - Front end tool	https://jquery.com/	3.4.1	3.4.1
bootstrap	Web - Front end tool	https://getbootstrap.com/	3.3.6	3.3.6



Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br