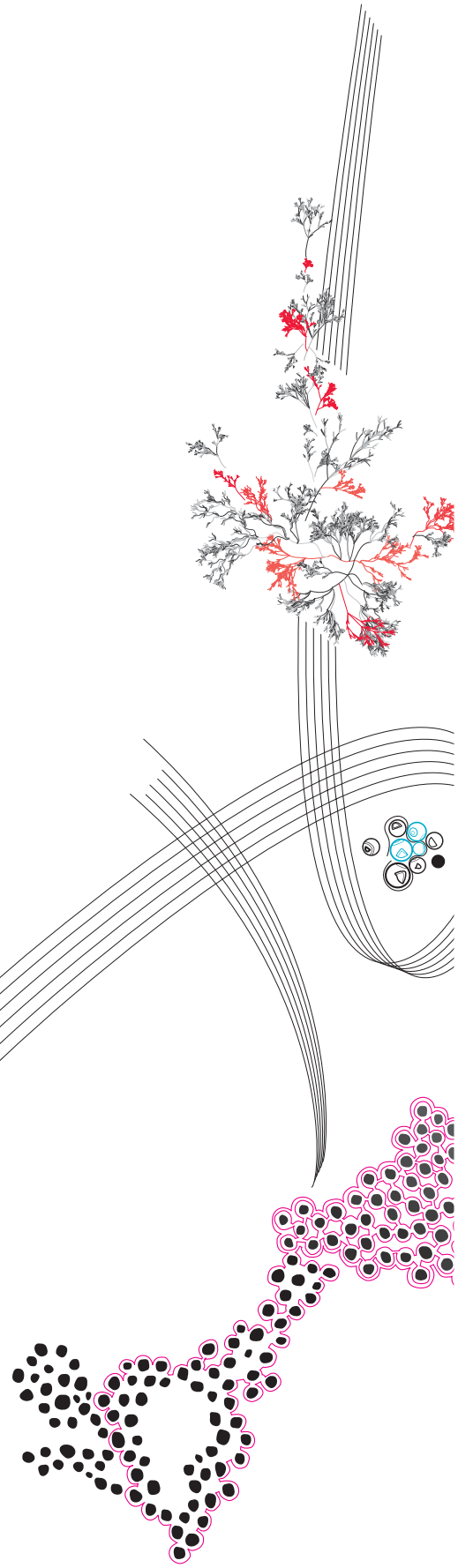

Msc Business Information Technology
Thesis

# Optimizing Data Quality with a Scoring-Based Cleansing Framework


Nabila Pindya

University Supervisor:
Dr. Maya Daneva
Dr. Lucas Meertens
Dr. JeewanieJayasinghe Arachchige

Company Supervisor:
Ivana Mishikj


27 August 2024


Department of Business Information Technology
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente


UNIVERSITY OF TWENTE.

Tembo

# Acknowledgement

I wish to extend my deepest gratitude to the individuals and organizations who have contributed to the successful completion of my thesis:

First and foremost, with the utmost sincerity, I express my gratitude to my one and only God, Allah *subhanahu wata'ala*, the source of all blessings. His guidance and the numerous opportunities He has bestowed upon me, including the chance to pursue my Master's degree—a long-held aspiration—have been crucial. His divine grace has supported me through the challenges and triumphs of my academic journey. It is by His blessings that I stand here today, humbled and deeply thankful.

I am also profoundly grateful to the University of Twente for granting me the opportunity to pursue a Master's degree. My appreciation extends to the Indonesian Ministry of Communication and Informatics for their financial support, which was essential for completing my studies. I would like to express my heartfelt thanks to my thesis supervisors, Maya, Jeewanie, Lucas, and Ivana. Their support and guidance were instrumental in completing this thesis.

To my beloved family, especially my late mother in heaven, thank you for bringing me into this world and nurturing me into who I am today. I am deeply grateful to my father, who has always supported my decisions and provided invaluable advice, and to my younger brother, Farhan, my constant companion. Words cannot fully convey my deep gratitude. Your unwavering support and prayers have been a source of strength and inspiration throughout this challenging journey. Your love and encouragement have carried me through the most difficult times, and for that, I am profoundly thankful. I would like to extend special thanks to my friend, Tifani, who greatly assisted me during my studies, providing essential knowledge and information. To Dita and Gina, my confidants for the past 24 years, thank you for your constant support and encouragement.

My heartfelt appreciation also goes to Nisa, Karisma, and Mba Luluk, who always shared their food with me, and to my friends Ally, Kimia, Camile, Pratyush, and everyone else who has been part of my life here. Your contributions have been invaluable. I also thank Ossie, Rina and all my friends in Indonesia, who are too numerous to mention individually.

Lastly, I want to express my deepest thanks to my partner, Sebastian, for always listening to my stories, frustrations, and selfishness. Thank you for standing by me for nearly two years. Hopefully, we can grow and progress together in the future.

Enschede, 27 August 2024

Nabila Pindya

# Abstract

The effectiveness of business decisions heavily relies on the quality of the data used in the decision-making process. On the other hand, poor-quality data misleads business decisions and leads to financial loss for a company and its reputation. Therefore, the quality of data is of utmost importance in any domain. However, maintaining high-quality data is challenging, especially in the manufacturing domain, due to its inherent complexity and ever-changing nature. Thus, this research focuses on minimizing errors, enhancing efficiency, and guaranteeing high-quality data for decision-making. The research strives to develop a framework to enhance data quality to achieve this. The research examined the past literature to reveal which attributes are significant to the data quality. Further, it explored existing data quality-enhancing frameworks. Design Science Engineering Cycle, as proposed by Wieringa in 2014 , is chosen as the methodology for this research to develop a data quality framework. Besides that, a workshop is held to gather information from a manufacturing company. The proposed framework primarily focuses on four dimensions of data quality: accuracy, completeness, consistency, and timeliness. The proposed data quality enhancing framework comprises four pivotal stages: Assess, Action, Enhancement, and Quality Scoring. One significant component of this framework is the scoring process, which encompasses pre-scoring and post-scoring. During this stage, the data undergoes evaluation to determine its overall quality. The conclusion of this research culminated in the successful development of a validated framework using the designed prototype. The validation results affirm that this framework can yield high-quality data as anticipated. From the validation results, accuracy was 95.75% before the cleaning process was carried out, Consistency was 99.98%, and scores for these two dimensions reached 100% after the cleaning process was completed. The data is being used for the validation is limited to only structure data. Consequently, this research significantly contributes by offering an effective method for upholding and enhancing data quality in the manufacturing sector, facilitating more accurate and reliable decision-making.

*Keywords*: Data quality, Data quality dimension, Data cleaning, Data cleansing, Data preprocessing, Manufacturing domain

# Contents

# List of Figures

# List of Tables

# Abbreviation

| | |
|---|---|
| **DMF** | Data Management Framework |
| **DSEC** | Design Science Engineering Cycle |
| **DQD** | Data Quality Dimension |
| **DQCF** | Data Quality Cleansing Framework |
| **KNN** | K-Nearest Neighbors |
| **EDA** | Exploratory Data Analysis |
| **NIDP** | Number In-Range Data Point |
| **NDV** | Number of Duplicate Value |
| **NNMV** | Number of Non-Missing Value |

# Chapter 1

# Introduction

## 1.1 Research Background

In today's world, data has become a critical element across various sectors, particularly in business. Many successful companies rely on data to drive profits [15]. Data serves multiple purposes, including providing a foundational basis for decision-making [35], which is crucial for the survival of an organization or company.

As the sheer volume of data continues to grow exponentially, ensuring the quality of data has become vitally important to support reliable business decision-making. Therefore, organizations must prioritize maintaining data quality, as data is considered a valuable asset [11]. Thus it is essential for any company to ensure the quality of data at each of the stages of the data life cycle (from the creation, storing, and using) as data quality is a major worry for several organizations [14]. Besides that, according to research conducted by [7], integrating data from various sources is a challenge due to the complexity of the data structure and types. This complexity makes it difficult to combine data. In addition, there is the problem of unclear data quality standards, as mentioned by [4], which creates difficulties in ensuring data quality during implementation.

Poor quality data can seriously impact a business, resulting in inaccurate insights, faulty analysis, and increased operating costs [18]. This should be a significant concern for every organization or company [20]. According to [30], poor data quality can cause an average annual loss of $15 million and in 2014, [14] stated that businesses incur annual costs of $13.3 million due to bad data and that the US economy suffers an annual loss of up to $3 trillion as a result. This poor quality data's adverse impact disrupts daily operations and can damage a company's reputation and hinder business growth. Inaccurate data can lead to wrong business decisions, resulting in missed opportunities and reduced profits. Additionally, the additional costs of correcting insufficient data and dealing with the consequences of bad decisions can burden a company's resources.

Therefore, data quality assurance is essential for evaluating, measuring, and improving data quality. This process ensures that the data meets the requirements and standards set within the [22] organization. Data quality assurance covers a wide range of critical activities, from data validation and cleansing to ongoing monitoring to maintain data integrity. By implementing a rigorous data quality assurance process, organizations can ensure that the data they use in analysis and decision-making is accurate, complete, and reliable. This not only improves operational efficiency but also provides a competitive advantage by enabling companies to make more informed and strategic decisions.

Data cleaning and data cleansing are often used interchangeably, but they refer to different processes. According to [12], data cleaning involves identifying and rectifying

errors or inconsistencies in a dataset, while data cleansing encompasses a more thorough approach. This involves not only detecting errors, but also making modifications, replacing, or removing inaccurate data to create a reliable and consistent dataset.

Previous research has developed various techniques such as data mining and machine learning. For example, in the paper written by [16], preventive and predictive algorithms are used for the data cleansing process, which helps identify and correct data errors before and while the data is used. In addition, in research conducted by [17], various outlier detection methods were applied to identify and analyze the distribution of outliers in the dataset. This approach allows researchers to understand unusual data patterns and take appropriate actions to manage them. Although these techniques have shown improvements in data quality, they are still unable to fully address data complexity, especially in manufacturing. Data in the manufacturing sector is often highly heterogeneous and dynamic, which adds to the challenge of ensuring data quality and consistency. Therefore, despite significant improvements in data quality through these techniques, a more holistic and sophisticated approach is still needed to overcome the various challenges in data management in the manufacturing sector. This new approach needs to handle data complexity more effectively, ensuring that the data used is reliable to support accurate analysis and informed decision-making.

Despite significant theoretical advances in organizational management, many companies still struggle to apply these effectively. For instance, in this research the use case is a manufacturing company that collects extensive data from multiple sensors across its operations has overlooked the crucial task of managing the quality of this data. Recently, the company came to realize that poor data quality could pose a threat to its operational efficiency and overall business continuity. This realization has prompted the company to initiate a comprehensive effort to implement a more sophisticated data management system, incorporating advanced technology for data analysis and monitoring. The primary aim of this endeavor is to enhance data accuracy and reliability, minimize machine downtime, and optimize production output, ultimately leading to a notable increase in company profits.

## 1.2 Research Scope

The scope of this research covers essential aspects related to quality of data especially in the manufacturing domain. This research identifies data quality dimensions (DQD) and improves data quality as the primary focus. For the DQD, in this research it will be focused only on four dimensions which are accuracy, completeness, consistency, and timeliness. The inspection will focus on improving data quality through data cleansing processes in the manufacturing domain. This project aims to develop and design a practical framework for improving data quality. The proposed framework will be validated through prototypes applied to real case studies in the manufacturing industry. Through this approach, the framework developed can not only improve data quality but can also be implemented practically to solve data quality problems faced by companies in the manufacturing sector. Validation using prototypes and case studies will provide empirical evidence regarding the effectiveness of the proposed framework, ensuring that the resulting solutions are widely applicable in real industrial environments.

## 1.3 Problem Context

Based on the facts presented in section 1.1, it is essential to establish a data quality framework in the manufacturing industry, given the limited research in this area. A thorough review in Chapter 2 revealed that only 11 studies specifically addressed data quality in the manufacturing context, highlighting a significant gap compared to other sectors such as health services, education, and government. This gap presents important opportunities for further research. Developing a robust framework in this area will not only enhance data quality but also improve operational efficiency and decision-making capabilities in the manufacturing industry.

Secondly, manufacturing companies gather a substantial volume of data from various operational machines. However, these companies currently lack an effective framework for ensuring the quality of this data. The absence of such a framework means that the produced data may not be of high quality, potentially impacting the accuracy of any subsequent analysis. Unverified data has the potential to lead to errors in decision-making, thus adversely affecting the company's operational performance and efficiency. It is, therefore, essential to develop and implement a robust framework for verifying data quality to ensure the reliability of the resulting data for further analysis.

## 1.4 Research Goal and Objectives

Based on the research gaps identified in the section 1.1, it is clear that there is a need to improve data quality in the manufacturing industry. The main goal of this research is to develop effective strategies for improving data quality in order to minimize errors, enhance efficiency, and guarantee the use of truly high-quality data for decision-making purposes. To achieve this, the research started by thoroughly investigating different dimensions of data quality that can be utilized as guidelines or requirements to ensure data quality. The research also aims to scrutinize various frameworks that can effectively measure, evaluate, and enhance data quality and also to achieve high-quality data that can be utilized for decision-making or any other essential actions within the organization. Through a critical analysis of the existing research on data quality management, this research aims to explore the techniques that can optimize the data quality assurance process in the context of manufacturing data. By reviewing numerous academic articles and research papers, the research seeks to identify gaps and opportunities in data quality. It also aims to provide comprehensive insight and motivation for further research into optimal approaches to data quality.

In pursuit of this primary goal, the research focused on two specific objectives. First, to develop a comprehensive and practical framework for improving data quality. This framework is meticulously crafted to provide users with comprehensive guidance, ensuring that resulting data meets high quality standards. By utilizing this framework, users can methodically execute necessary processes, ranging from data checking and cleansing to validation and analysis, with consistent and precise standards. Second, to create a prototype that will validate the framework. Through this approach, the research not only contributes to improving data quality in the manufacturing industry but also provides a strong foundation for practical implementation and continued research in this area. Hence, this research can offer significant and impactful solutions for the manufacturing industry in managing and improving the quality of their data.

## 1.5 Research Question

To effectively pursue the research goal, a main research question and a series of sub-research questions have been developed. These questions form the basis for exploring and analyzing the relevant existing literature. The sub-research questions are divided into three categories: knowledge questions, which aim to gather information or understanding about the topic; design questions, which aim to develop or propose a solution, model, or system for a particular problem or need; and validation questions, which aim to assess the effectiveness, accuracy, or reliability of the solution, model, or system. The research questions are articulated below:

**Main Research Question:**
How can data quality be improved by considering most crucial data quality factors in manufacturing domain?

**Sub-Research Questions**

- **Sub-RQ1 (knowledge question)**: What are the motivations for conducting research on data quality framework?
  *Motivation*: The objective of this question is to gain a comprehensive understanding of the underlying motivations and driving factors behind the research. The research aims to establish the significance of a data quality framework in data processing, providing contextualization to the research.

- **Sub-RQ2 (knowledge question)**: What are the fundamental data quality dimensions found in current literature?
  *Motivation*: The primary objective of this question is to discern the prevailing trends in data quality dimensions, with the purpose of informing the selection of the most appropriate data quality measures for the ensuing research.

- **Sub-RQ3 (knowledge question)**: What are available frameworks/methodologies to enhance the data quality in the current literature?
  *Motivation*: The objective of the sub-research question is to elucidate the framework for the development of a data quality framework in the context of the manufacturing domain, with the ultimate goal of achieving high-quality data.

- **Sub-RQ4 (knowledge question)**: What is the current state of data cleansing (automatic) for improving the quality of data?
  *Motivation*: The objective of this sub-research question is to acquire an in-depth comprehension of the current state-of-the-art in data cleansing, with the aim of designing a data cleansing prototype.

- **Sub-RQ5 (knowledge question)**: What are the challenges faced when implementing frameworks/procedures of data quality in manufacturing domain?
  *Motivation*: This question intends to identify the difficulties related to existing methods, which is essential for developing solutions that address these important issues. Evaluating the limitations enables the creation of new and innovative strategies to overcome challenges and achieve successful implementation of good data quality in the manufacturing industry.

- **Sub-RQ6 (design question)**: How can a robust data quality framework be designed to enhance data quality and operational efficiency specifically for the manu-

facturing domain?
*Motivation*: This question aims to design a good data quality framework that can produce high-quality data that will boost the company's analysis performance.

- **Sub-RQ7 (validation question)**: To what extent does the implementation of a developed framework effectively enhance data quality?
*Motivation*: This research question seeks to examine how the framework can effectively enhance the accuracy, consistency, and completeness of the data. This is crucial as improved data quality directly influences decision-making, operational efficiency, and ultimately, the overall success of the organization.

## 1.6 Research Structure

This thesis is structured into eight chapters, each with a specific focus. Chapter 1 serves as an introduction, outlining the research background, defining research scope and problem context, formulating research questions, highlighting the research goal, and introduces the research structure. In Chapter 2, a Systematic Literature Review (SLR) is conducted to address sub-research questions 1-5, synthesizing theoretical knowledge on relevant topics and identifying research gaps. Chapter 3 is explaining about the methodology that used in this research.

Chapter 4 is discussing data understanding to data quality dimension (DQD) and the result of the workshop. The design of the data quality framework and its explanation are presented in Chapter 5, and Chapter 6 provides insight into the validation using the prototype that has been develop. Chapter 7, the findings are discussed and finally, in Chapter 8 , along with contribution, limitation and recommendations for future research. Figure 1.1 show the summary of the overall structure of the research.

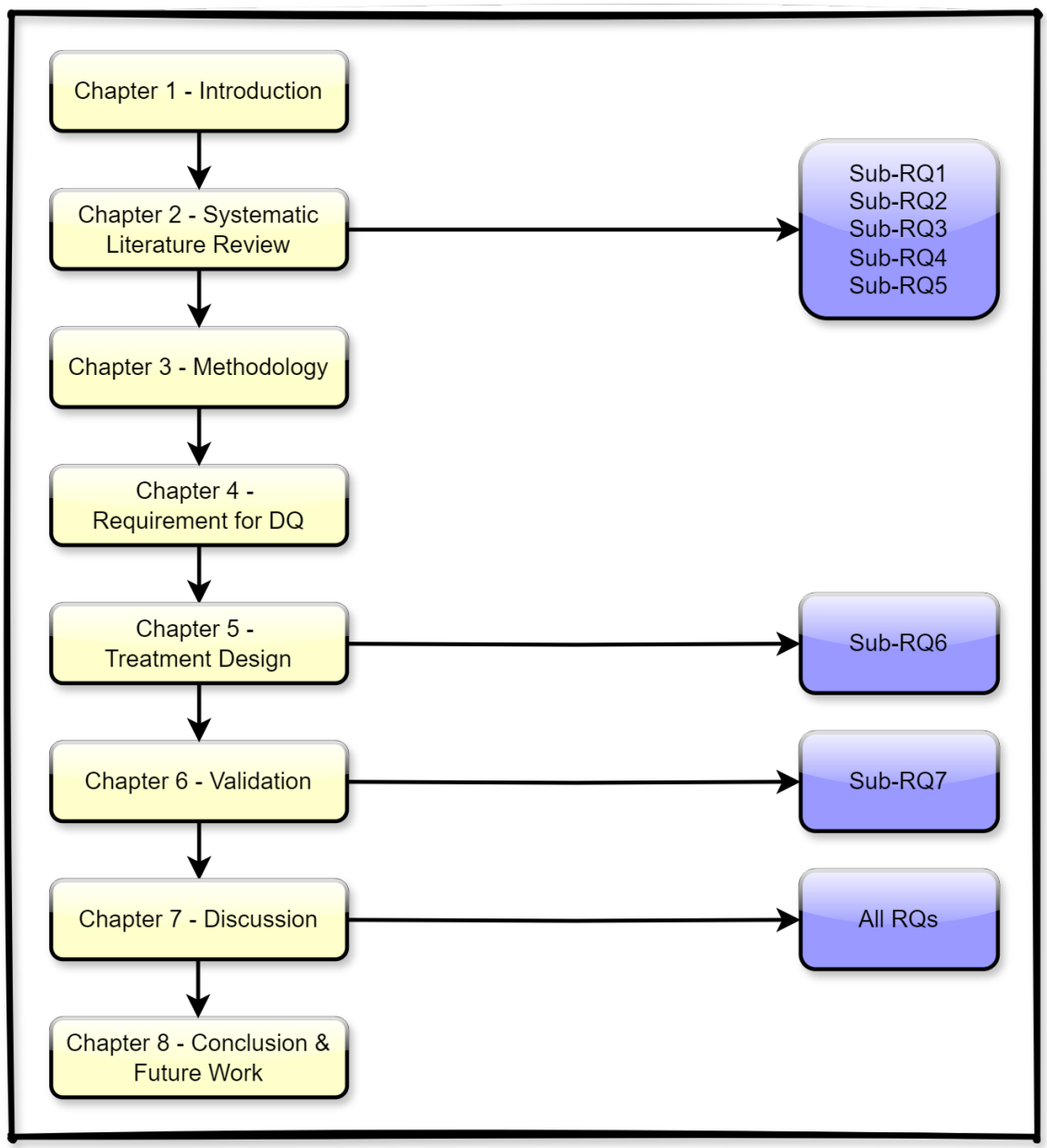


FIGURE 1.1: Research Structure

## 1.7 Chapter Summary

This chapter is dedicated to discussing the motivation behind this research, encompassing the research background, research scope, problem context, research goal, research question, and research structure.

The first section offers an general overview of nature of data in the today's world and significance of data quality in the manufacturing domain. It delves into the dearth of comprehensive research on data quality in manufacturing and the challenges faced by many companies in managing the quality of their data. This underscores the pressing need for a robust framework to ensure optimal data quality in the manufacturing industry.

Following this, we delve into the main research question and sub-research questions, along with the motivation behind each question. These questions have been formulated to

guide the research in identifying and addressing key challenges in ensuring data quality in the manufacturing sector.

Lastly, the chapter outlines the research goals, which include the development and validation of a data quality framework that can be effectively applied in manufacturing companies. These goals aim to offer practical and innovative solutions to the data quality issues faced by the manufacturing industry today and the chapter presents the research structure, offering a clear road map of how the research will unfold through various stages.

# Chapter 2

# Systematic Literature Review

This chapter explains the methodology used for the SLR in this research, following the guidelines proposed by Kitchenham & Charters [24]. The literature review is a vital component of academic research, and its efficacy is dependent on a rigorous and structured approach. A systematic methodology ensures that the review process is organized and comprehensive, and it begins with the planning phase. This phase involves formulating research questions, selecting appropriate scientific databases, developing search strategies, and establishing inclusion and exclusion criteria. The review process entails selecting relevant articles and extracting pertinent data from them. These steps are crucial for identifying and incorporating studies pertinent to the research questions and objectives of the study, as well as for extracting pertinent information from the selected articles. Therefore, it is essential to adhere to a structured approach to ensure that the literature review process is thorough and effective.

## 2.1 Planning the Review

When starting a systematic literature review, it is important to have a well-organized approach to guarantee a thorough and rigorous analysis of the research available. The planning of this systematic literature review is based on research conducted by Kitchenham and Charters.

### 2.1.1 Scientific Database

In this systematic literature review, three scientific databases were selected to obtain relevant academic publications for answering the research questions. The scientific databases selected for this systematic literature review are as follows:

- Scopus (https://www.scopus.com)
- Web of Science (https://webofscience.com)
- IEEE (https://ieeexplore.ieee.org)

The selected scientific databases were chosen for their extensive coverage of academic literature pertinent to the proposed topic. Furthermore, these databases are recognized as being among the top five most reputable academic sources.

### 2.1.2 Search Query Formulation

A keyword selection process was planned to formulate proficient search queries for scientific databases used in this systematic literature review. The chosen keywords were extracted directly from the objective of the research. The keywords are "data quality", "manufacturing", "dimensions", "improvement", and "data pre-processing". A comprehensive list of keywords associated with the main query will be executed on selected scientific databases to retrieve relevant literature for the review can be see in Table 2.1.

TABLE 2.1: Query Search Keywords

| Data Quality | Improvement | Artefact | Manufacturing |
|---|---|---|---|
| Data Quality | Improving | Framework | Manufacturing |
| | Data Pre-processing | Architecture | Production Machine |
| | Data Cleansing | Model | Production System |
| | Data Cleaning | | |
| | Dimension | | |

After categorizing keywords presented in Table 2.1, search queries were defined using "OR" and "AND" logical operators in each scientific database. The queries are listed below:

- Scopus (advance search):
  TITLE-ABS-KEY(
  ("Data Quality")
  AND
  (Improving OR "Data Pre-processing" OR "Data Cleansing" OR "Data Cleaning" OR "Dimension")
  AND
  (Framework OR Architecture OR Model)
  AND
  (Manufacturing OR "Production Machine" Or "Production System"))

- Web of Science (advance search):
  TS=(
  ("Data Quality") AND (Improving OR "Data Pre-processing" OR "Data Cleansing" OR "Data Cleaning" OR "Dimension") AND (Framework OR Architecture OR Model) AND (Manufacturing OR "Production Machine" Or "Production System"))
  OR
  TI=(
  ("Data Quality") AND (Improving OR "Data Pre-processing" OR "Data Cleansing" OR "Data Cleaning" OR "Dimension") AND (Framework OR Architecture OR Model) AND (Manufacturing OR "Production Machine" Or "Production System"))
  OR
  AB=(
  ("Data Quality") AND (Improving OR "Data Pre-processing" OR "Data Cleansing" OR "Data Cleaning" OR "Dimension") AND (Framework OR Architecture OR Model) AND (Manufacturing OR "Production Machine" Or "Production System"))

- IEEE (https://ieeexplore.ieee.org)
  ("Data Quality")
  AND
  (Improving OR "Data Pre-processing" OR "Data Cleansing" OR "Data Cleaning" OR "Dimension")
  AND
  (Framework OR Architecture OR Model)
  AND
  (Manufacturing OR "Production Machine" Or "Production System")

### 2.1.3 Inclusion and Exclusion Criteria

Kitchenham and Charters [24] have highlighted the importance of establishing clear inclusion and exclusion criteria while conducting a systematic literature review. These criteria act as guidelines to ensure that only relevant studies are included in the review and irrelevant ones are excluded. Such criteria are crucial to maintain the validity and integrity of the review and to prevent any biases that may arise due to the inclusion of irrelevant studies. In this systematic literature review, the inclusion and exclusion criteria are outlined in Table 2.2.

TABLE 2.2: Inclusion and Exclusion Criteria

| **Inclusion Criteria** | **Exclusion Criteria** |
|---|---|
| Literature conducted in English | Duplicate Literature |
| Literature published in last 10 years | Incomplete or unavailable literature |
| Literature was a journal or conference proceedings in Computer Science, Engineering, Decision Sciences, Mathematics, Social Sciences and Business, Management and Accounting | Irrelevant literature based on its abstract to this study's defined research questions |

For this systematic literature review, inclusion criteria have been set for publications written in English and published within the last 10 years. The reason behind this is to ensure that the publications are accessible and easily understood and provide a contemporary perspective on the subject matter. To ensure academic integrity, the research is limited to Computer Science, Engineering, Decision Sciences, Mathematics, Social Sciences and Business, Management and Accounting. Peer-reviewed journals and conference proceedings are the preferred sources of literature for scholarly discourse within these disciplinary boundaries.

Steps were taken to eliminate any duplicate or inaccessible literature to ensure an efficient and accurate review process. Any sources that were not available in full were excluded, as comprehensive analysis and interpretation require complete access. Additionally, a manual assessment was conducted to identify and exclude irrelevant literature based on its abstract. The objective was to refine the selection process and only retain publications that directly contribute to the research objectives.

## 2.2 Selection Process

The process of selecting literature began by running a search query in chosen scientific databases. However, since the search results may contain irrelevant studies, applying the inclusion and exclusion criteria established earlier is essential. This step ensures that the research collection is of high quality and that the data extraction process is streamlined by focusing on relevant or highly relevant studies.The selection process is illustrated in Figure 2.1.

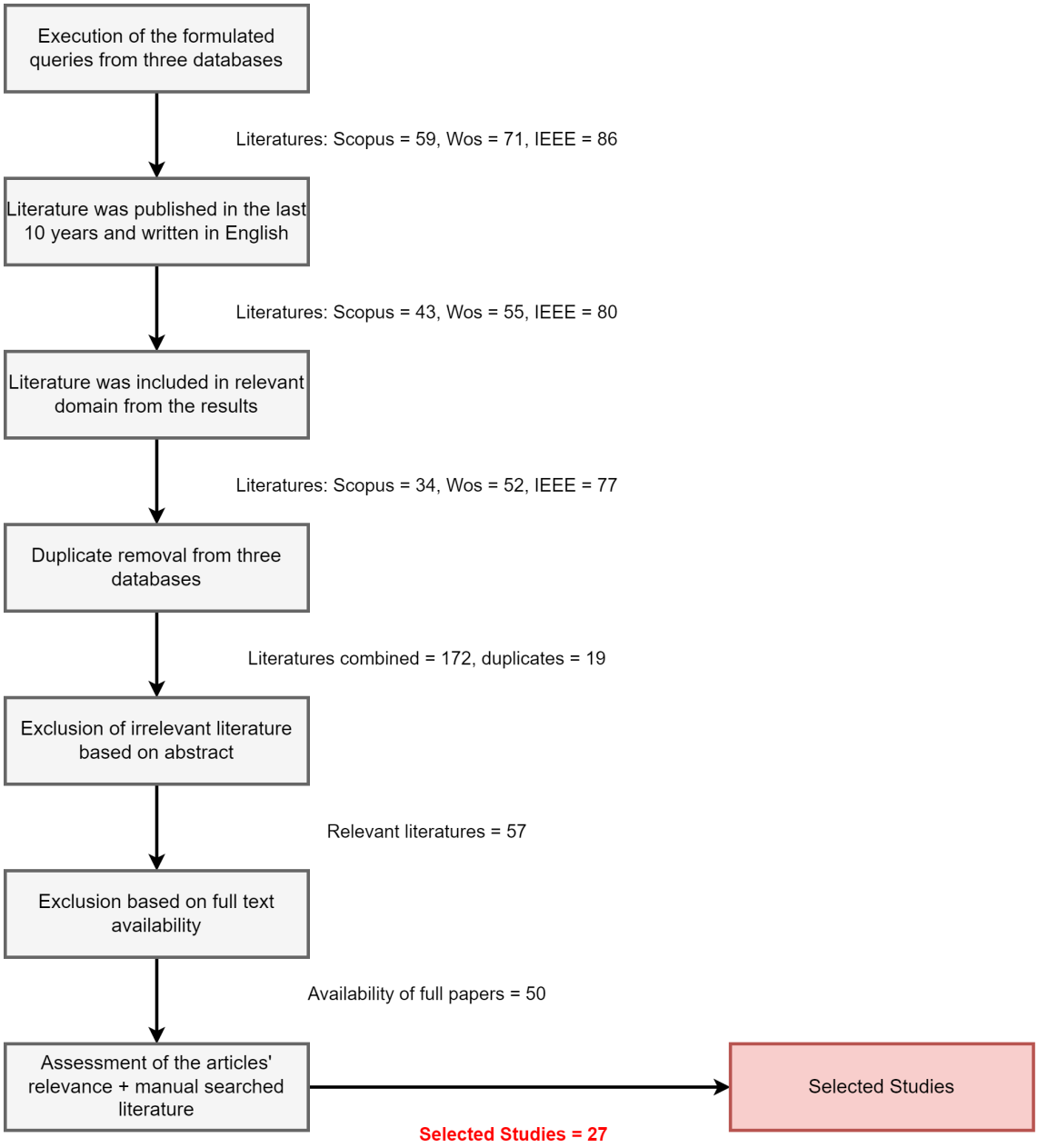


FIGURE 2.1: Literature Selection Phases

Several stages were involved in the process of selecting literature, which was directed by the inclusion and exclusion criteria specified in Table 2.2. After going through these stages, a total of 27 pertinent articles were selected by applying the criteria and necessitating manual review.

## 2.3 Data Extraction

Upon completing the application of inclusion and exclusion criteria, the relevant information obtained from the literature review will be elaborated on in the subsequent chapter. The process of extracting data involves a meticulous combination of both quantitative and qualitative methods, which have been carefully designed to provide a comprehensive analysis of the current state of existing literature. By integrating these two approaches, this study aims to create a more nuanced and detailed overview of the research landscape in this area, which will enable us to gain a deeper understanding of the topic at hand.

## 2.4 Quantitative Analysis

A high-level quantitative analysis is conducted to determine the objective of the literature, the research techniques employed, and the context related to this Systematic Literature Review (SLR). The outcome of this quantitative analysis is classified into four in order to analysis purposes: data quality dimension (D), data cleaning (DC), whether the research used manufacturing data (M), and whether the research proposed framework or methodology (FM). The results are depicted in Table 2.3.

Table 2.3: Quantitative Analysis Based on Target

| NO | Reference | Categories | | | |
|---|---|---|---|---|---|
| | | D | DC | M | FM |
| P1 | Ali, T.Z., Maatuk, A.M., Abdelaziz, T.M., Elakeili, S.M. (2020) [3] | | x | | x |
| P2 | Al-Masri, E., Bai, Y. (2019) [2] | x | x | | x |
| P3 | Burggraf, P., Dannapfel, M. (2018) [5] | | | x | x |
| P4 | Burkhardt, A., Berryman, S., Brio, A., Ferkau, S., Hubner, G., Lynch, K., Mittman, S., Sonderer,K. (2017) [6] | | | x | x |
| P5 | Chen, Q., Liu, y., Huo, S., Duan, F., Cai, Z. (2021) [8] | x | | x | |
| P6 | Chernov, Y. (2016) [9] | x | | | x |
| P7 | Ding, N.B., Mit, E. (2023) [16] | x | x | | x |
| P8 | Günther, C.L., Colangelo, E., Wiendahl, H.H., Bauer. C. (2019) [17] | x | | x | x |
| P9 | Ji, C. (2023) | x | x | x | |
| P10 | Juneja, A., Das, N.N. (2019) [21] | x | x | | x |
| P11 | Kirchen, I., Schutz, D., Folmer, J., Vogel-Heuuser, B. (2017) [23] | x | | x | x |
| P12 | Kong, W., Qiao, F., Wu, Q. (2020) [25] | x | | x | |
| P13 | Li, C., Zhu, Y. (2015) [26] | x | x | | x |
| P14 | Li, X. (2022) | | x | | x |

| | | | | | |
|---|---|---|---|---|---|
| P15 | Munawar (2021) [27] | x | x | | x |
| P16 | Poon, L., Farshidi, S., Li, N., Zhao, Z. (2021) [28] | | x | | x |
| P17 | Schlegel, P., Buschman, D., Ellerich, M.m Schmitt, R. H. (2020) [29] | x | | x | x |
| P18 | Sitawati, H.D., Ruldeviyani, Y., Hidayanto, A.N., Amanda, R.S., Nugroho, A.G. (2021) [30] | x | | | x |
| P19 | Soto, P.C., Ramzy, N., Ocker, F., Vogel-Heuser, B. (2021) [31] | | x | x | x |
| P20 | Sreenivas, P., Srikrishna, C.V. (2013) [32] | | x | | x |
| P21 | Taleb, I., Dssouli, R., Serhani, M. A. (2015) [33] | x | x | | x |
| P22 | Tsai, W.L., Chan, Y.C. (2019) [34] | x | | | x |
| P23 | Wahyudi, T., Isa, M.S. (2023) [35] | x | | | x |
| P24 | Widad, E., Saida, E., Gahi, Y. (2023) [37] | x | x | | x |
| P25 | Xu, D., Zhang, Z., Shi, J. (2022) [39] | x | | x | x |
| P26 | Yuan, T., Adjallah, K.H., Sava, A., Wang, H., Liu, L. (2021) [40] | x | | x | |
| P27 | Zhang, J., Gao, R.X (2021) [41] | | x | | |

## 2.5 Qualitative Analysis

The systematic literature review employs qualitative analysis methods to thoroughly explore the literature to discover recurring themes, patterns, and trends. In this particular systematic literature review, the qualitative analysis is presented in the form of a table based on the target set, as depicted in Table 2.3. The outcome of the qualitative analysis is displayed in the Table 5.1 in appendix.

## 2.6 Visualization of the Findings

### 2.6.1 Year-based Trend

The following section displays the graphical representations of the outcomes from the systematic literature review. The initial representation illustrates a trend based on years to obtain valuable inferences into the changing trends and structures within the academic domain by scrutinizing the number of literature categorized by years. The academic perspective on the distribution and growth of research output over the years is presented in Figure 2.2 as a result of this analysis.

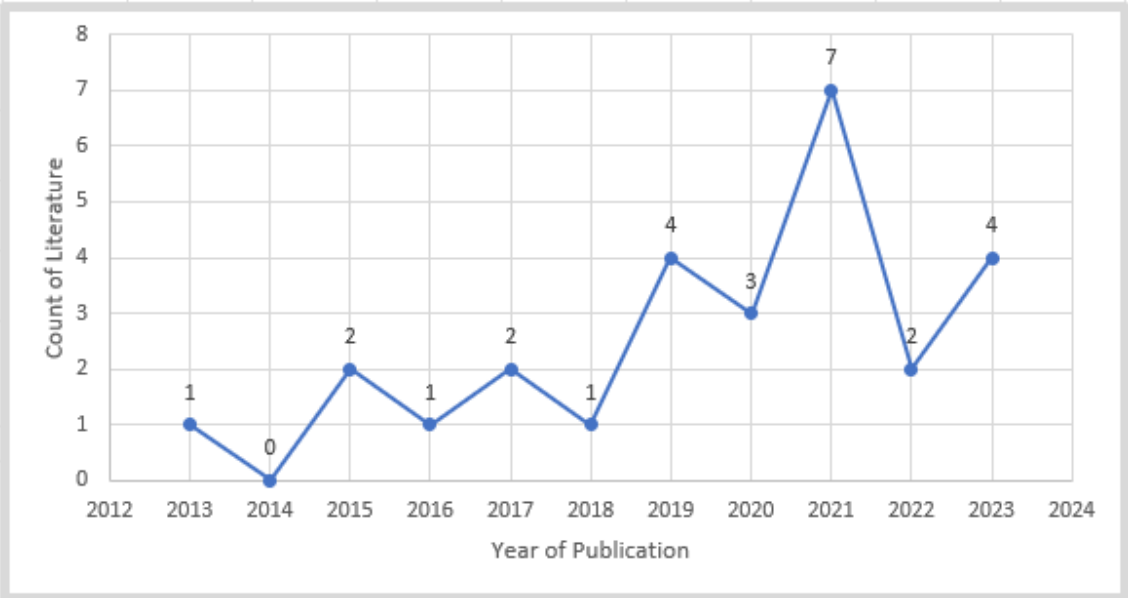


FIGURE 2.2: Year-based Trends of the Reviewed Literature

The information depicted in Figure 2.2 shows a significant growth in the quantity of published research papers from 2013 to 2023. This suggests an increasing interest in the specific research topic. The trend began with a small number of published papers, ranging from 1 to 2 between 2013 and 2018. However, there has been a gradual rise in the number of publications with a few intermittent variations. The subsequent years have shown a mixture of fluctuations with varying publication counts. Nevertheless, it is worth noting that there was an immense surge in the number of published papers in 2019, which indicates a probable turning point or increased scholarly attention during that period on the search query formulation of this study.

The graph illustrates an increasing trend in the number of publications from 2019 to 2023. The research output has been consistently high each year. However, the most significant surge in research output is observed in 2021, when it reaches its peak with seven publications. This notable increase indicates that the research field has attracted more attention and academic engagement, possibly due to the interest of related parties in data quality, particularly in the manufacturing sector.

### 2.6.2 Distribution of Literature in Different Sectors

Figure 2.3 reveals the distribution of literature reviewed in this study classified by sector areas. The analysis of trends based on these sector areas has identified some intriguing patterns and potential research gaps that can be explored in this study.

After analyzing the literature obtained through the search query, it was found that manufacturing is the most frequently discussed sector with a total of 11 publications. The second most discussed sector is big data, followed by warehouse, finance, IoT, and other data. Interestingly, even though the search query was defined as "manufacturing," there were other sectors discussed in the findings. Upon deeper examination, it was discovered that some of the literature that focused on other manufacturing sectors was actually discussing data preparation or data cleaning.

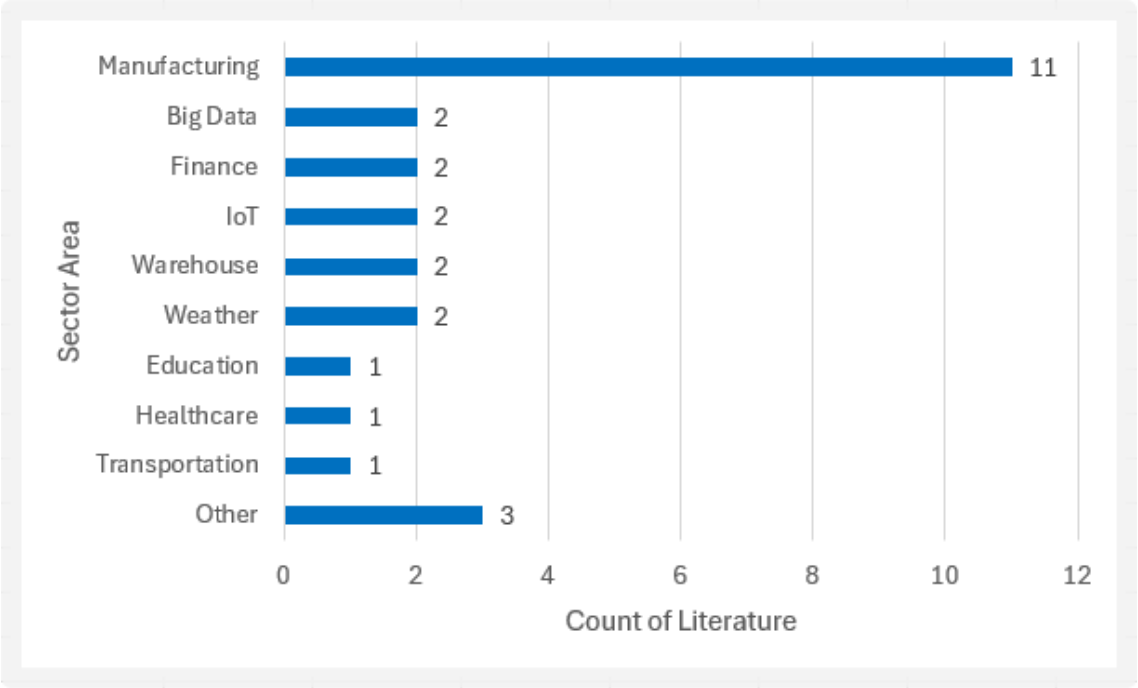


FIGURE 2.3: Distribution of Literature in Subject Area of the Reviewed Literature

## 2.7 Answer to Sub-Research Question

### 2.7.1 Sub-RQ1: What are the motivations for conducting research on data quality framework?

The first sub-research question aims to investigate the motivations behind researching data quality frameworks. Several factors contribute to this motivation.

According to [30], data is highly valuable to organizations and businesses in the realm of business. When data is collected and processed correctly, it can be transformed into knowledge and meaningful information that can be leveraged to enhance decision-making and streamline operations. The importance of data quality in the business world is well-understood, as it plays a crucial role in ensuring that businesses operate efficiently and make informed decisions. This is further emphasized by research conducted by [35], which highlights data quality's significance in modern business practices.

Another factor, in a recent conversation with representatives from the manufaturing company, it was brought to light that they are grappling with a major issue - the lack of certainty around the quality of their data. The absence of a well-defined framework for data management has made it difficult for them to analyze and draw insights from the available information accurately. Furthermore, the absence of established standards or rules for data collection, storage, and usage has only added to the situation's complexity. It is imperative for the organization to address these challenges and put in place a robust data quality framework to ensure that they have access to reliable, high-quality data that can support their business objectives.

### 2.7.2 Sub-RQ2: What are the fundamental data quality dimensions found in current literature?

Data quality assessment is greatly influenced by its dimension, making it an essential factor. When selecting data quality dimensions, discussing them with the organization or company is important. To determine the appropriate dimensions for measuring data quality levels, it is necessary to consider the dimensions relevant to specific business processes. Many other factors may come into play when identifying data quality dimensions, such as data source conditions, data needs within an organization, and internal data policies.

Various data quality dimensions have been employed in this study, as shown in Figure 2.4; it can be seen that the most important dimension being used is completeness [2, 9, 16, 17, 21, 23, 29, 30, 33, 34, 35, 37], followed by accuracy [2, 3, 16, 17, 21, 23, 30, 33, 34, 35, 37], consistency [8, 9, 16, 17, 21, 23, 33, 34, 35, 37], timeliness [2, 8, 9, 17, 21, 27, 30, 33, 34, 35, 40], accessibility [2, 8, 23, 29], relevance [17, 23, 29, 39], conformity [9, 21, 37], Appropriate Amount [23, 29, 39], Integrity [8, 9, 21], Traceability [23, 27, 40], correctness [9, 27], currency [30, 27], conciseness [27, 40], ease of manipulation [39, 40], interpretability [29, 40], security [27, 40], uniqueness [21, 37], variance [23, 40], applicability [27], availability [23], calculation [19], compliance [23], credibility [23], data comprehension [9], free of error [39], imbalance [39], maintainability [27], plausibility [39], precision [2], readability [37], reliability [19], speed [27], synchronicity [23], validity [9], understandability [40], value-added [40], and reputation [40]. According to [35], the most commonly used dimensions are completeness, timeliness, accuracy, and consistency, and this statement is valid based on the findings of this study in Figure 4.1.

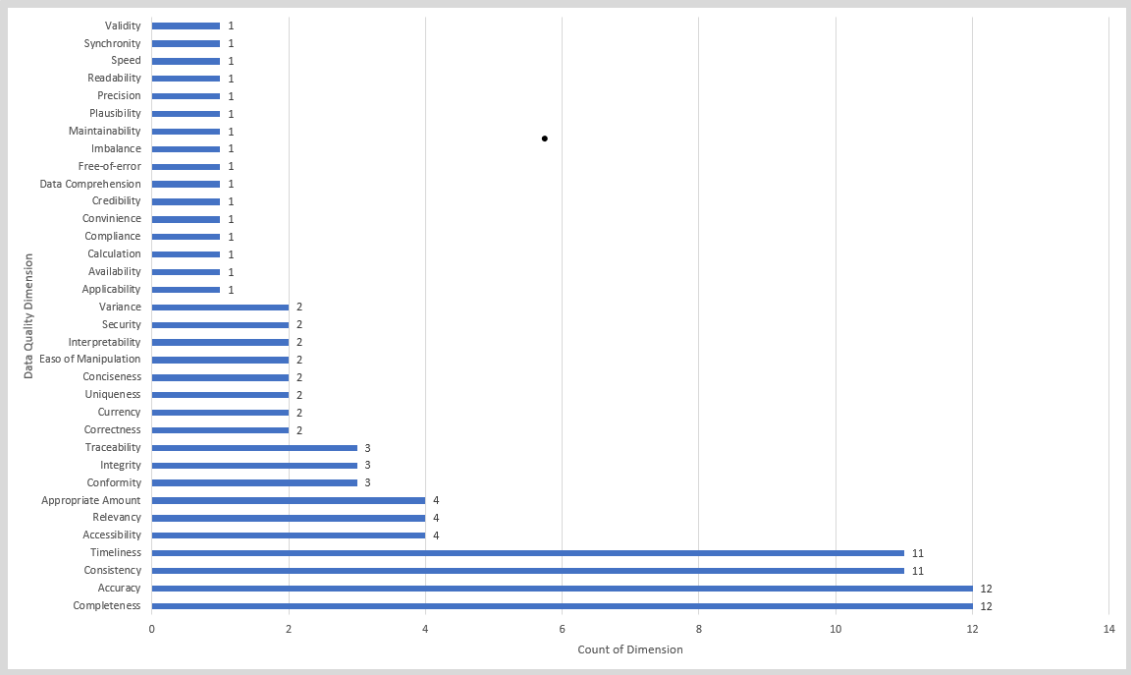


FIGURE 2.4: Data Quality Dimension in the Reviewed Literature

### 2.7.3 Sub-RQ3:What are available frameworks/methodologies to enhance the data quality in the current literature?

Based on the results of this study, a number of frameworks are available. However, not all of the literature pertains specifically to data quality frameworks; some are related to data

cleaning (preprocessing) frameworks, which are also relevant to data quality. The identified frameworks will be categorized into two groups: data quality frameworks and preprocessing or data cleaning frameworks. This section will focus on explaining the findings of the data quality framework, and the preprocessing or data cleaning framework will be explained in the next section.

For data quality framework, there are several frameworks available. One such framework, developed by [35], focuses on data quality assessment and uses the Total Data Quality Management (TDQM) method, which has three phases: Define, measure, and analyze.

Another framework, developed by [5], is called Data Quality-based Process Enabling (DQPE). This framework is based on the available quality of data and aims to improve data quality and optimize process design. It contains two processes: Continuous Data Quality Improvement Management (CDQIM) for continuous improvement of data quality processes, and Data Quality-dependent Process Design (DQPD) for alternative design processes based on available data quality.

[39]'s framework focuses on improving data quality in the early stages before it is used for further processing or analysis. This framework aims to resolve quality issues for large data sets and consists of several processes: scope project, collect data, raw data, data quality assessment, data quality improvement, data quality control, train model, and deployment in production. Other framework is focus on data quality improvement process [6].

### 2.7.4 Sub-RQ4: What is the current state of data cleansing (automatic) for improving the quality of data?

As mentioned earlier, several frameworks are available for data cleaning or preprocessing with different methods, as shown in table 2.4. In this context, the paper by [26] primarily deals with cleaning indoor positioning data. The paper's authors have utilized several techniques, such as statistical and probabilistic theory, graph theory, machine learning, numerical optimization, and computing-domain theory. They have employed computing domain theory as their algorithm for designing an efficient and effective data-cleansing process. Additionally, they have used various methods to cleanse the data, including uncertainty modeling, fault correction, exploiting heterogeneity, and imputing missing values. These techniques are commonly found in the literature on data cleansing.

In another literature, [31] have contributed to the field of data cleaning by also focusing on missing values imputation. To handle this problem, the author utilized queries and a reasoner to identify and eliminate missing values. Additionally, another common technique in data cleaning, outlier detection and correction, was implemented. This problem is also discussed by [28]. Statistical traditional methods such as z-score and IQR and unspervised clustering are being used for anomaly detection.

In their practical work, [31] utilized Python and the Pandas library for their advanced capabilities in managing, analyzing, and exploring data. These tools play a crucial role in simplifying the data cleaning and preparation process for further analysis. Similarly, [28] applied these techniques, specifically using Python's scikit-learn and PYoD packages, for implementing an unsupervised clustering approach, highlighting the effectiveness of these technologies in addressing data quality issues.

In their paper [41], the authors provide a summary of several important techniques used in data cleansing. They mention that missing values and outliers are two common data problems. Common methods to deal with missing values include linear interpolation and regressive modeling. The paper also mentions several algorithms like recurrent neural networks (RNN) for time series data and convolutional neural networks (CNN) for image

TABLE 2.4: Methods used in Data Quality Improvement

| Reference | Methods/Algorithm | Tools |
|---|---|---|
| Ali, T.Z., Maatuk, A.M., Abdelaziz, T.M., Elakeili, S.M.(2020) | Single-source data and Multiple-source data | N/A |
| Kong, W., Qiao, F., Wu, Q.(2020) | Data value density | N/A |
| Poon, L., Farshidi, S., Li, N., Zhao, Z.(2021) | Statistical Traditional Method and unsupervised clustering | Python's scikit-learn and PYoD packages |
| Soto, P.C., Ramzy, N., Ocker, F., Vogel-Heuser, B.(2021) | Utilized queries and a reasoner | Python's Pandas library |
| Sreenivas, P., Srikrishna, C.V.(2013) | Moving averages, mean filters, and median filters | N/A |

imputation that can be used for this purpose. On the other hand, statistical methods such as Gaussian, distance-based, density-based, and cluster-based methods are commonly used for detecting outliers.

According to a study by [32], various cleaning techniques are necessary for different types of data. The study delves into diverse data preprocessing methods, such as the application of filters like moving averages, mean filters, and median filters to minimize noise in time series data. Additionally, data normalization, data sampling, and data segmentation are crucial steps in the data cleaning process. The paper by [25] explains the use of the rate of change of data value density, and clean single-source data and multiple-source data can be found in [3].

### 2.7.5 Sub-RQ5: What are the challenges faced when implementing frameworks/procedures of data quality in manufacturing domain?

The implementation of data quality is a complex and challenging process that confronts several obstacles. One of the most significant hurdles is the high cost and time required for the process. According to [17], data quality implementation demands a considerable amount of time and money. Moreover, data preparation, which is a crucial stage for ensuring data quality, consumes around 80 percent of the total work, as pointed out by [31].

Detecting anomalies in data quality is another major challenge as quality anomalies are usually hidden and may not be readily visible in the data. Furthermore, the diversity of data sources and data types contributes to the complexity of the data structure, making data integration difficult. In a study conducted by [7], it was observed that the variety of

data sources and types can significantly impair the data structure, making it challenging to integrate and further complicating the process of ensuring data quality.

## 2.8 Chapter Summary

This chapter provides a comprehensive guide to the procedures for planning and executing a Systematic Literature Review, underscoring the significance of selecting appropriate scientific databases, devising effective search queries, and establishing stringent inclusion and exclusion criteria. Through this meticulous selection process, 27 pertinent pieces of literature pertaining to data quality were identified. The chapter also examines trends in literature over time and across subject areas, revealing a notable dearth of attention to data quality research within the manufacturing industry. This underscores the emergence of fresh and consequential research prospects in the sector, poised to significantly advance both practice and theory in manufacturing. Besides that it also add explain the answer to the sub-research question 1-5 where from this SLR it was found that the 4 dimensions that are most often used are the existing framework and the current state of data cleansing in previous research.

# Chapter 3

# Methodology

This chapter will discuss the methodology used in this research to answer the research question and research methodology.

## 3.1 Research Methodology

The research methodology used in this research is the Design Science Engineering Cycle (DSEC) as proposed by [38]. The choice of this methodology was based on the reason that the concept of this methodology makes more sense because this methodology functions as a systematic guide for data collection and analysis, which aims to ensure the reliability and integrity of the findings. It emphasizes the development of conceptual models based on theory and their evaluation against requirements. DSEC is especially relevant in practical fields where the main goal is to develop real-world solutions. Although commonly used in information systems, software engineering, and design research, DSEC principles can be adapted to a variety of research domains.

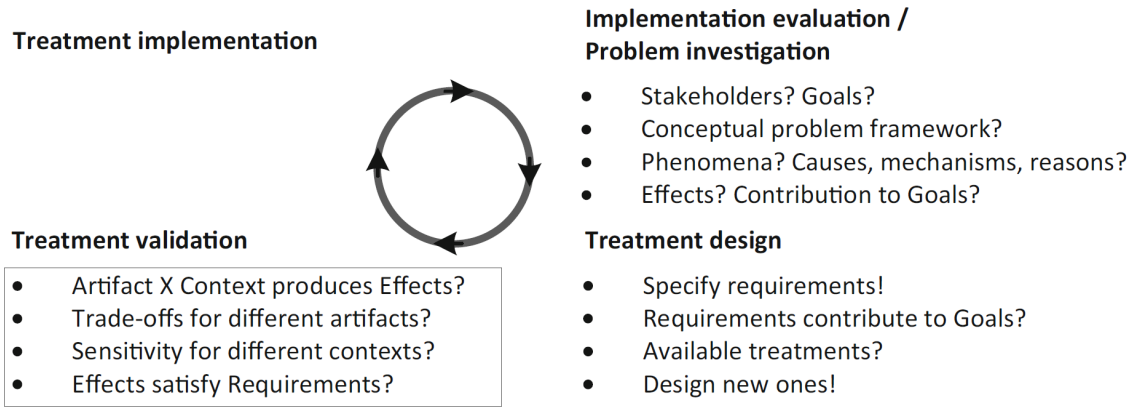


FIGURE 3.1: Design Science Engineering Cycle [38]

The design procedure encompasses four primary phases: problem investigation, treatment design, treatment validation, and treatment implementation, which collectively form the design cycle. The design cycle represents a component of the broader engineering cycle,

which encompasses the implementation and assessment of validated treatments in practical settings. Consequently, the scope of this thesis can be delineated as follows:

### 3.1.1 Problem Investigation

The first step in DSRM is problem investigation. This stage aims to find out more deeply about the roots of the existing problems. In this research, researchers conducted observations on matters related to data quality, especially in the manufacturing domain. Then also identify the goals of this research. Some of the methods used in the research are as follows:

1. **Systematic Literature Review**:
   In this approach, researchers thoroughly examine various academic papers, articles, and publications using the guidelines outlined by Kitchenham (2007) [24] to gain a foundational understanding of data quality. The aim of this literature review is to compile comprehensive information about the concepts, dimensions, and best practices in data quality management. This process consists of systematic steps to ensure the identification, critical analysis, and synthesis of all relevant sources. By adhering to [24] guidelines, researchers can guarantee that the literature review is carried out with a rigorous and transparent methodology, ensuring that the resulting findings are dependable and valuable for further research.

   The findings of the literature review is discussed in Chapter 2, where it outlines the fundamental concepts and key discoveries related to data quality in detail. This section will also offer an understanding of the crucial aspects of data quality and their application in different scenarios.

2. **Workshop**:
   As previously mentioned in the section 1.1, this research focuses on a use case involving data quality challenges within a manufacturing company. To gain a comprehensive understanding of the company and its data quality issues, a workshop was conducted with the company's data team.

   The workshop aimed to uncover detailed information about the company's data quality challenges and to gather direct insights from experts in the field. Discussions during the workshop encompassed topics such as data understanding, user stories, and the specific data quality dimensions required by the company.

   The outcomes of this workshop are crucial for informing the development of practical and effective solutions. The information gathered will be utilized to ensure that the proposed solution is not only theoretically sound but also aligned with the company's actual needs. A detailed account of the workshop's findings and insights will be presented in section 4.3.

### 3.1.2 Treatment Design

The next step in the process involves treatment design, aiming to tackle the identified problem by devising creative and feasible solutions. This process will be thoroughly explained in Chapter 5. Treatment design refers to a conceptual model crafted to offer effective problem-solving strategies.

During this phase, the design process doesn't just prioritize creativity but also draws on various frameworks and insights from past research. This ensures that the proposed solution is not only innovative but also thoroughly tested and reliable. These frameworks

offer valuable guidance for developing models that are not just theoretical but also practical and applicable in real-world situations.

### 3.1.3 Treatment Validation

The third step in the process involves treatment validation, which aims to rigorously assess the effectiveness of the design solutions proposed in the treatment design stage. This validation ensures that the solution not only achieves the desired objectives but also meets all the established requirements.

Various techniques are employed in the validation process, one of which is prototyping, which is used for this research. In the context of this research, a prototype is currently being developed and will be thoroughly described in Chapter 6. This prototype serves as a tool to validate the proposed solution, particularly the data quality framework, using the use cases outlined in the 1.1 section.

For the treatment implementation and implementation evaluation, it's important to note that due to time constraints, are not within the scope of this research.

## 3.2 Chapter Summary

This chapter offers a thorough explanation of the utilization of DSEC as proposed by Wieringa in this research and outlines the rationale behind selecting this methodology. The explanation encompasses each stage of DSEC, starting from problem investigation and treatment design to treatment validation.

# Chapter 4

# Requirements for Data Quality

This chapter discusses the meaning of data quality along with several data quality dimensions, a brief explanation of the use case and insights obtained from SLR (chapter 2) and workshops that have been conducted.

## 4.1 Data Quality

Data quality is a very important and widely discussed topic today. However, what exactly is meant by data quality? Accroding to [36], data quality is defined as data that meets user needs, emphasizing accuracy, completeness, and relevance. [33] asserts that data quality is a well-established concept in the database community and has been a primary focus of database research for many years. Data quality encompasses a range of activities, including planning, implementation, and control. These activities utilize various data quality management techniques to ensure that the resulting data is prepared for consumption and can meet the needs of data users [30]. Data quality pertains to the degree to which existing data aligns with the specific requirements of its users, ensuring precision, accuracy, consistency, and reliability.

### 4.1.1 Data Quality Dimension

Upon understanding the concept of data quality, it's important to consider how to measure whether existing data meets quality standards. This measurement can be achieved through DQD, which are elements used to assess data quality [33]. Wang has identified 15 dimensions of data quality grouped into four main categories: Intrinsic data quality, Contextual data quality, Representational data quality, and Accessibility data quality (Figure:4.1). For the purpose of this research, the focus is specifically placed on four dimensions of data quality: accuracy, consistency, completeness, and timeliness. The selection of these dimensions is based on the results of the SLR which can be seen in section 2.7.2 and the workshop, which will be further discussed in section 4.3. The definition of selected DQD as follow:

1. **Accuracy**
   Accuracy plays a crucial role in evaluating data quality as it measures the extent to which data reflects correct and reliable information. According to [33] & [21], accuracy is to see whether the data has been recorded accurately and represents realistic values. This evaluation involves comparing the amount of correct data to the total amount of data in a dataset. Not only does it consider the correctness of the facts, but it also ensures that the data is free from errors that could potentially

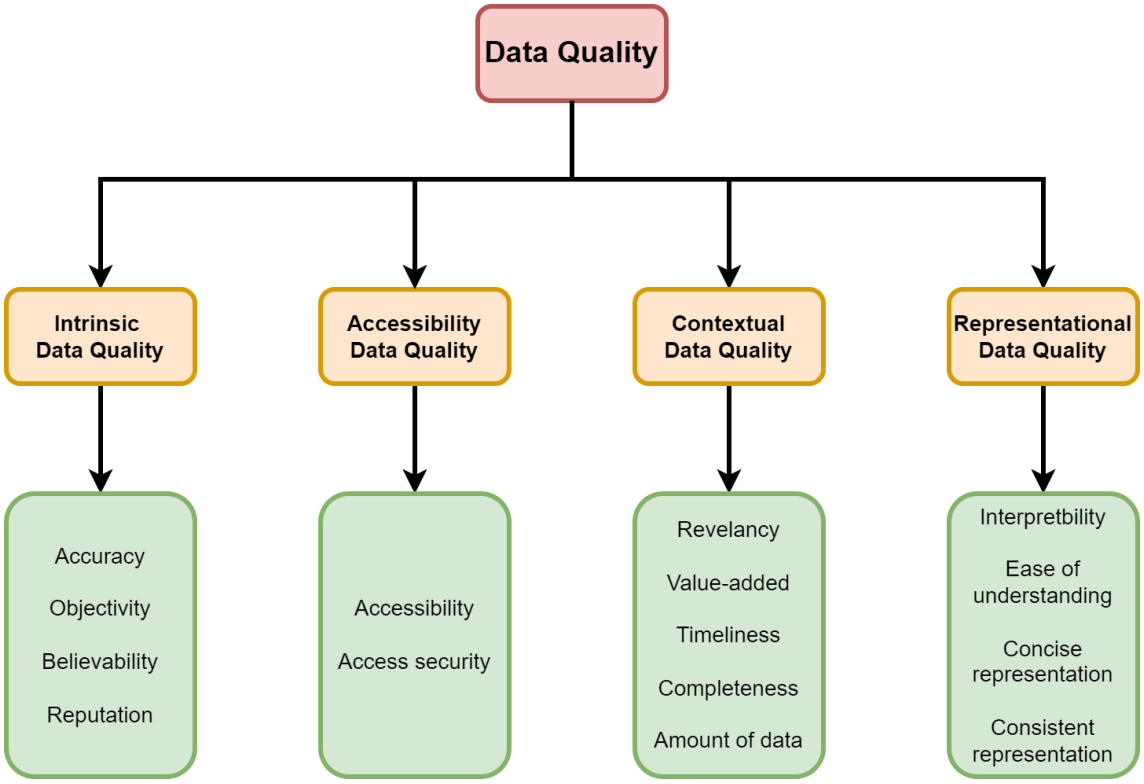


FIGURE 4.1: Data Quality Dimension [36]

impact business decisions and operations. The significance of accuracy is especially pronounced in industries heavily reliant on precise and efficient decision making based on data.

2. **Consistency**
   Consistency refers to the degree of conformity of data to its designated format and structure [33], ensuring that each data element aligns accurately with the specified format specifications. This process is essential for minimizing errors in data processing and promoting effective integration between systems.

   According to [17], maintaining consistency in data is crucial to presenting information in a uniform format and adhering to established semantic rules. This alignment ensures compatibility with existing entries and prevents format discrepancies that can lead to misinterpretation and inaccurate analysis.

3. **Completeness**
   Completeness within an information system denotes the system's capability to accurately mirror the full spectrum of real-life scenarios [30]. Within a database or application, all attributes should contain comprehensive values, ensuring that all crucial entries are thoroughly filled in which means no missing values [36]. Inadequate data can result in erroneous interpretations or suboptimal decisions. Therefore, it is vital to guarantee the collection and accurate representation of all required information within the system. This is paramount in fortifying the system's reliability to support effective decision-making and operations.

4. **Timeliness**
   The extent to which data can accurately represent the true value at the time of need

[36] is a critical consideration for the accessibility and availability of information. Data timeliness not only assesses the accuracy of data in reflecting real-time events, but also encompasses the duration of the data's relevance throughout its life cycle [8]. This underscores the importance of implementing a system capable of updating and presenting data promptly to support efficient and accurate decision-making.

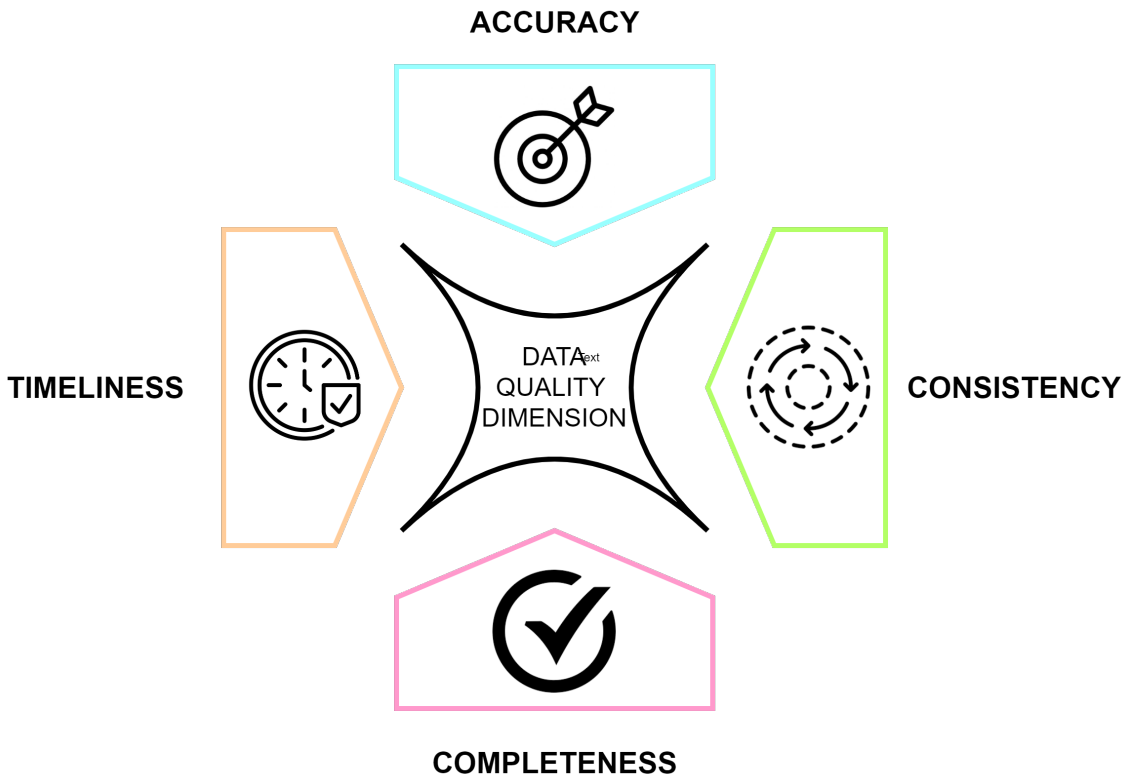


FIGURE 4.2: Selected Data Quality Dimension

## 4.2 Use Case

As mentioned in the section 1.1 , the research will focus on addressing the data quality issues in a manufacturing company. The aim is to develop a data quality framework that can help companies assess and enhance the quality of their data. To validate the framework, a prototype will be created and tested using data from the company's detergent making machine. This will provide valuable insights into the effectiveness and practicality of the framework in real-world settings. For more details about the dataset used and the validation process, refer to the chapter 6.

## 4.3 Workshop

In order to gain more understanding about the data, the workshop is being held with Data Team. Over the course of 90 minutes, the attendees were involved in lively conversations and practical exercises, including the creation of a user story that outlines how a project can provide value to the end user. The goal of the workshop was to define the data quality

dimension and rules, and the participants' diverse insights and knowledge greatly contributed to the discussion, leading to the development of innovative solutions and concrete strategies. From the workshop there are several things gained from the workshop which will be discussed further below.

### 4.3.1 Data Understanding

The dataset utilized for this research has been sourced from a sophisticated machine that comprises a vast collection of information obtained from multiple sensors. However, owing to temporal limitations, only six sensors will be considered in this research. The data is readily available in a parquet file. More information about the data as follow:

**Data Collection Methodology**

- **Real-Time Collection**: Data is collected in real-time directly from machines on the production shop floor located in Serbia.
- **Raw Data Capture**: The data is unfiltered. It is raw as captured from the Data Monitoring System (DMS).
- **Storage and Transfer**:
  - Local Storage: Initially stored locally before being pushed to Azure Cloud.
  - Transfer Delay: Minimal delays occur during data transfer, maintaining near real-time availability.

When working with data, there are certain quality issues that are commonly encountered. The first issue is related to data anomalies, such as persistent zero values, which indicate that there is a problem. Another issue is empty files, where the data files are either empty or contain only zeros. The second issue is related to data loss concerns, such as missing data and static values where data values do not change as expected. These issues might mask underlying operational problems and should be addressed.

### 4.3.2 User Story

The user story is a crucial work unit in product development that seeks to align product features with customer requests. From the user's perspective, these elements are presented in simple language, making it easier to understand and increase customer satisfaction. During the workshop, a number of important user points were identified and divided into two main category requirements: Operational requirement and Data Quality requirement. In the context of this research, the main focus will be given to the Data Quality category. This category is considered very important because it directly relates to the accuracy, consistency, completeness, and timeliness of data, all of which are key factors in ensuring that the data used for analysis and decision making is of high quality. Through this approach, research aims to optimize how data quality can be improved and maintained, with the hope that this will bring significant benefits to both end users and the product development process as a whole.

### 4.3.3 Operational Requirements

- Tracing back problems to get an overview of machine history.

- Notifications on wrong signal values to correct and utilize data for operations.
- Storing outliers to relate them to machine performance variations.
- Fast data delivery for capturing quick changes in machine operation (e.g., wrinkling).
- Data aggregation per shift to match the operational scope and improve response to machine speed changes.
- Assurance that data corresponds to logging settings for reliability.
- Tracking data volume to confirm the reliability of the DMS tool.

### 4.3.4 Data Quality Requirement

- Automated data quality checking system to minimize manual effort and ensure data reliability.
- Detection of out-of-bound data points to understand machine anomalies.
- Overview of faulty data points to improve the system.
- Requirement of no missing data for accurate process statistics.
- Ability to adjust data quality thresholds (e.g., watchdog settings) to manage data storage efficiently (filtering noise, outliers).
- Desire for both high-granularity and aggregated data for in-depth analysis and broader trends.
- Monitoring data quality to ensure data meets expectations before it enters the pipeline.
- Ensuring data consistency to maintain expected standards (e.g., temperature data within a specified range).
- Need for valuable data that is relevant to the production or machine running status.

### 4.3.5 Selected Data Quality Dimension

During the workshop, there was an extensive discussion regarding the critical aspects that needed to be considered for the research. The results showed a consensus among the participants that there were four primary dimensions of data quality that would be used in this research project (Figure 4.2). These dimensions were identified as Accuracy, Completeness, Consistency, and Timeliness. Based on the workshop result, the participants agreed that these dimensions would form the foundation for the data quality assessment framework for the project. The definition of each dimension can be seen in table 4.1

TABLE 4.1: Data quality dimension definition by users

| Dimension | Definition |
|---|---|
| Accuracy | Ensure the correct value in accordance with existing requirements or regulations |
| Completeness | Maintaining a balance between data volume and system capacity while collecting adequate data is crucial to avoid losing critical information while handling high-frequency logging |
| Consistency | Assess uniformity within data sets and between data sets related by examining duplicate data |
| Timeliness | Operational data must be actionable for real-time system monitoring and alarms that require immediate attention to prevent operational failures |

## 4.4 Chapter Summary

This chapter elucidates the significance of data quality and its dimensions, laying down a crucial knowledge base for this research. A profound grasp of data quality and its dimensions not only facilitates the identification and evaluation of data quality but also establishes a robust framework for implementing effective data quality management techniques. Additionally, the report of the workshop that has been held for in-depth understanding of data is also explained. The report covers various aspects of data, such as the understanding of data, how the data was collected, the issues with the data, the data user story, and selected data quality dimensions.

# Chapter 5

# Treatment Design

This chapter explains briefly the current framework from the company (DMF) and outlines the development of the artifact design which is the Data Quality Cleansing Framework (DQCF). The present discourse provides a comprehensive and scholarly overview of the meticulous process involved in crafting the target framework, starting from the existing framework. The discourse, thus, offers a detailed exposition of the process involved in the development of a target framework.

## 5.1 Data Management Framework (DMF)

A framework for data management is presently available in the company, which encompasses various processes, namely acquisition, collection, transfer, storage, transformation, and reporting. These processes are visually represented in the accompanying figure 5.1.

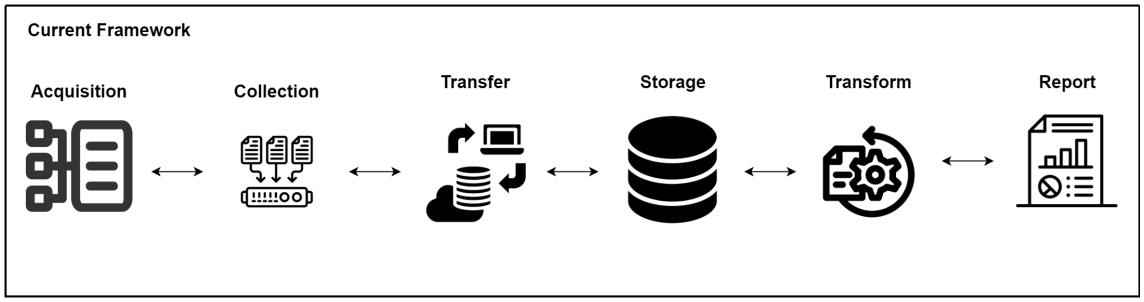


FIGURE 5.1: Data Management Framework (DMF)

The first stage in the DMF pertains to data acquisition. Data acquisition refers to the process of capturing raw data, whether structured or unstructured, from various resources, in this case, the machine and its sensors, in real-time. The data typically conforms to the format directly provided by the sensor. After the completion of the initial stage, the subsequent step involves data collection, during which all previously captured raw data is gathered and consolidated. These two phase, data acquisition and data collection, are similar processes involving obtaining, organizing, and analyzing data.

After the completion of the meticulous data collection process, the subsequent critical stage involves the transfer of the amassed raw data. This pivotal phase entails the seamless transfer of all the raw data to the designated storage repository, which constitutes the fourth stage of the process. Once the data is securely stored, the subsequent step involves the transformation of the data as per the specific requirements. This transformation process

is instrumental in ensuring that the data is tailored to meet the analytical and operational needs. Subsequently, this leads to the final process where a comprehensive report detailing the insights and findings derived from the data is meticulously generated.

Based on the DMF, several deficiencies persist in its implementation and functionality. One of the most salient inadequacies pertains to data quality, a critical factor in effective data processing and analysis. The absence of guaranteed data quality is disconcerting as it could yield inaccurate and misleading results, potentially impacting users in decision-making and other business processes reliant on this data.

Furthermore, the deficiencies in data quality aspects within the DMF underscore the necessity for additional mechanisms to ensure data accuracy, completeness, consistency, and timeliness throughout the processing process. Addressing data quality is imperative due to its direct influence on the reliability and validity of the output. Therefore, there is an urgent need to methodically incorporate data quality enhancement and monitoring features into the framework.

## 5.2 Data Quality Cleansing Framework (DQCF)

The primary consideration of the proposed framework is data quality aspects, encompassing the implementation of data checking, data cleansing, and data normalization and de-duplication techniques, all of which can significantly enhance data quality. This development would not only enhance the framework's usability but also bolster user confidence in the accuracy of the processed data. It is evident that the DMF inadequately acknowledges the importance of data quality, despite its essential role in achieving optimal and reliable results in data processing.

With a view to the future, a Data Quality Cleansing Framework (DQCF) was developed to fulfill the requirements for data processing. This framework is constructed upon the DMF. The DQCF encompasses multiple stages: Assess, Action, Enhancement, and Quality Scoring, and it entails two types of activities, differentiated by color: orange for data processing and green for the scoring process. Data processing encompasses all necessary operations for data cleansing, while the scoring process involves evaluating data quality, as depicted in Figure 5.2. The integration of this framework with the existing one is aimed at not only enhancing the overall effectiveness of the DMF but also paving the way for future advancements.

The integration of the DMF with DQCF takes place during the collection stage within the DMF. Upon completion of the data collection process, the subsequent step is not the transfer stage, but rather the Assess stage. The assessment process serves as the primary stage within the DQCF then later when the data is cleaned, it will go to transfer stage in DMF. Further details about each stage encompassed by the DQCF are outlined in the following subsection.

### 5.2.1 Assess

The first stage in the DQCF is the Assess stage, which involves several key processes. Data partitioning, data profiling, and pre-scoring are carried out during this stage.

1. **Data Partitioning**
   In this process, raw data is sorted by sensor name to streamline analysis. This is necessary as the dataset originates from a machine with diverse sensors, each possessing unique properties. Organizing the data by sensor enables more targeted and efficient analysis. This approach ensures that each sensor is meticulously examined

and analyzed based on its specific characteristics and contribution to the machine. By doing so, we can enhance the accuracy of the analysis and pinpoint and address issues that may only be evident at the individual sensor level. This guarantees that the data utilized in the analysis is of superior quality and relevance.

2. **Data Profiling**
   The data profiling process is a series of activities and processes aimed at identifying and investigating [1] data, includes statistical checks of data types, detection of missing and null values, validation of duplicate data, and identification of outliers. This multi-stage process provides a comprehensive understanding of the data being analyzed.

   Several methods are executed during data profiling to further analyze and understand the data.

   (a) **Check and change data type**: The aim of this process is to verify that the existing data types comply with applicable regulations. If differences or discrepancies are identified, the incompatible data type will be replaced with the correct one. This step is crucial for producing high-quality data ready for further use, such as data analysis. When the data type does not match the desired format, it can lead to irrelevant or misleading analysis results. Therefore, ensuring the appropriateness of data types is a critical step in maintaining data integrity and validity, ultimately supporting more accurate and data-driven decision-making.

   (b) **Check missing value**: An prevalent data quality challenge involves missing values, where their absence may signify incomplete data [10]. It is typical for datasets to have missing values, and it is crucial to thoroughly investigate and rectify these omissions to prevent statistical bias and maintain the integrity and coherence of the analysis.

   (c) **Check outliers**: In addressing data quality, identifying outliers is critical as these observations significantly deviate from the majority of data points [13]. Within DQCF, two principal methodologies are employed to detect such outliers effectively. The first method involves implementing predefined rules based on minimum and maximum values. This rule-based approach allows for the straightforward identification of data points that fall outside established thresholds, providing a quick filter for potential outliers. The second method incorporates the K-Nearest Neighbors (KNN) algorithm. This more sophisticated technique identifies outliers based on the distance and similarity of a point to its nearest neighbors in the data space. This algorithmic approach offers a dynamic and context-sensitive analysis, enhancing the robustness of the outlier detection process in complex datasets.
   The use of two different methods for checking outliers allows users to compare the results of each method and determine whether there are significant differences in identifying outliers in the dataset. Employing two different approaches provides users with a more comprehensive understanding of outlier patterns in their data and ensures that outlier detection is not reliant on one method alone. This is important for increasing the reliability and validity of the data cleansing process, as well as ensuring that the resulting data accurately reflects the actual situation without any bias or undetected errors associated with using only one method.

(d) **Check duplicate**: When working with data, it's crucial to prioritize data quality by actively seeking to avoid duplicate entries. Duplicate data can significantly impact the accuracy of analytical results, leading to errors and bias [37]. By eliminating duplicates, analysts can enhance the reliability and integrity of their findings for more informed decision-making.

3. **Pre-scoring**
   Upon completing data profiling, a pre-scoring process is initiated. This stage stands out as one of the most pivotal within the framework, offering a significant advantage. At this juncture, a scoring system is deployed to assess the accuracy, consistency, and completeness of the raw data. The primary aim is to discern any substantial disparities between the data's condition before the cleansing process.

   This pre-scoring stage provides crucial insights into how the raw data's quality can influence the ultimate results. By conducting pre-scoring, we can pinpoint areas necessitating special attention and ensure the efficacy of the data cleansing process. For a more comprehensive understanding of data quality assessment methods and the resultant findings, please refer to section 5.2.4.

By conducting these processes during the Assess stage, DQCF ensures a thorough initial evaluation of the data used in the analysis. This early identification of data quality issues enables effective remedial measures to be implemented in later stages of the framework.

### 5.2.2 Action

Once the Assess stage is complete, the next phase in the DQCF process is known as Action stage with a primary focus on data cleansing. During this stage, two sub-processes are carried out to ensure optimal data quality. The data cleansing process is meticulously designed to address existing issues and enhance the integrity of the dataset.

The initial sub-process involves removing missing data or NA values, which is crucial to overcoming imperfections that can significantly impact data analysis. Following the identification and deletion of missing data, the process proceeds to eliminating duplicate data. The goal is to ensure that each entry in the dataset is unique, thus avoiding redundancy that could disrupt subsequent analysis.

Within this cleansing stage, two fundamental procedures are systematically employed to safeguard the integrity and usability of the data:

1. **Remove missing data/NA**
   The removal of missing values or NAs is pivotal, as unaddressed issues can distort analysis and lead to inaccurate conclusions which will make a problem to decision making for the business.

2. **Remove duplicate** The elimination of duplicates is essential to prevent redundancy, which can distort data analysis outcomes by giving undue weight to repeated entries.

These cleansing processes are conducted iteratively to continually refine the dataset. Through the rigorous and iterative application of these cleansing techniques, datasets become more reliable and better aligned with the desired quality criteria. Ultimately, this ensures that the data supports accurate and effective decision-making.

Iterative cleansing not only improves data quality but also fortifies the foundation of the dataset, making it a robust resource for subsequent analytical tasks and facilitating the generation of more meaningful insights. Consequently, this cleansing process represents a

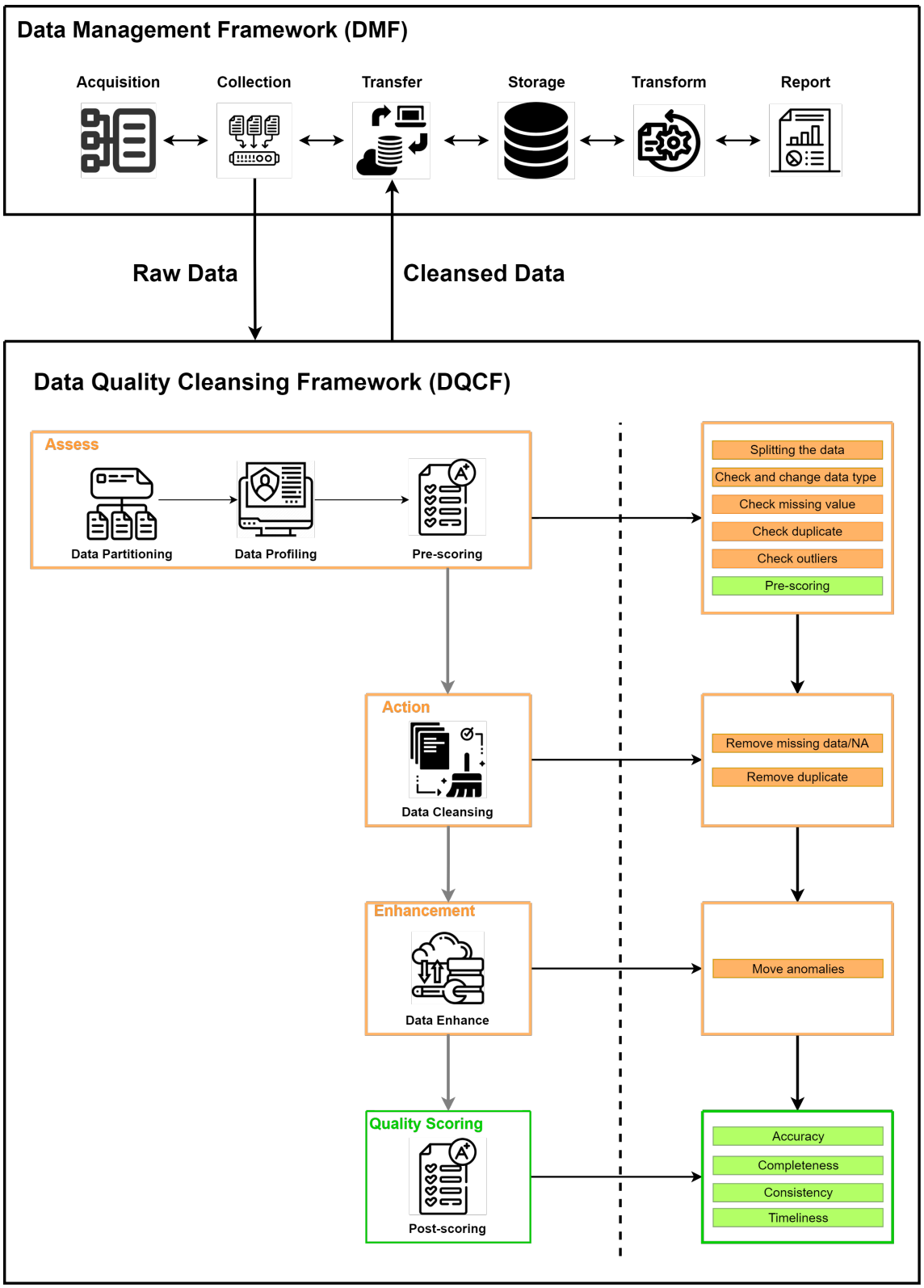


FIGURE 5.2: Data Quality Cleansing Framework

pivotal step in the data management framework, guaranteeing that the dataset used in analysis is free from imperfections that could influence the final research outcomes.

### 5.2.3 Enhancement

The third stage of DQCF is referred to as "enhancement," marking a critical phase following the comprehensive data cleansing process. During this stage, the focus shifts to a more nuanced handling of outliers identified during the assessment phase. Instead of eliminating these outliers, they are strategically relocated to a separate file designated for this purpose. This file aggregates outliers collected from various sensors, ensuring their preservation for potential future analyses.

Deliberate preservation of outliers is crucial, as it allows for their availability without compromising the integrity of the main dataset. By segregating these values, analysts can maintain a clean primary dataset for standard operations while retaining the ability to access the outlier data for detailed investigations or specialized analytical tasks. This approach acknowledges the value outliers can hold in providing insights into abnormal conditions or identifying errors in data collection methodologies. Furthermore, storing outliers separately facilitates a dual analysis strategy, where routine analytics are performed on the cleaned dataset and exploratory analyses on the outlier-rich dataset, enhancing the overall robustness and depth of the analytical framework.

This strategy ensures the cleanliness and usability of the main dataset while optimizing the information utility of data that deviates from normal patterns. This methodological refinement in handling outliers exemplifies a balanced approach to data management, ensuring thoroughness in data quality enhancement while safeguarding against potential loss of critical information.

### 5.2.4 Quality Scoring

The final phase of the DQCF encompasses the "Quality Scoring," a pivotal stage intended to evaluate the quality of data following the antecedent processes of data profiling, data cleansing, and data enhancement. This stage assumes significance as it quantitatively assesses the readiness of the dataset for decision-making and operational utilization. Quality evaluation at this juncture is executed through a systematic scoring mechanism, which entails the assignment of a score reflecting the overall data quality. This scoring process yields a percentage value indicative of the achieved level of data quality.

Within the data scoring process, the assessment is structured around four fundamental dimensions: Accuracy, Consistency, Completeness, and Timeliness. These dimensions were chosen based on organizational requisites and the outcomes of a Systematic Literature Review (SLR), which highlighted these attributes as most frequently discussed and crucial in prior research, as depicted in Figure 2.4 and also from the result of the workshop in section 4.3. Each dimension addresses a distinct aspect of data quality:

1. **Accuracy**
   It is essential to ensure that the data accurately represents real-world values in this research. The primary focus is to ensure that the data falls within an acceptable range, matching the expected values and remaining relevant to the context. To achieve a high level of accuracy, we are utilizing data that falls within a predetermined minimum and maximum value range and then applying the following formula to determine the number of correct values.

   $$\frac{NIDP}{N} * 100 \tag{5.1}$$

2. **Completeness**
   Completeness is a fundamental concept used to assess the presence of all necessary

data in a dataset, typically by identifying any missing values. In this research, we conducted a completeness assessment by scrutinizing the dataset for any such gaps. Missing data can signify potential issues in the data collection process and may have an impact on subsequent analysis and findings. The process of ensuring data completeness includes identifying and quantifying missing values in each dataset variable. Techniques such as analyzing missing value patterns and filling in missing data (imputation) or even remove it can be employed to address this issue. Additionally, further analysis may be conducted to understand the cause of missing data, whether it is random or systematic.

$$\frac{NNMV}{N} * 100 \tag{5.2}$$

3. **Consistency**
   Ensuring consistency in data evaluation is crucial for assessing uniformity within a dataset and across related datasets. In this research, consistency evaluation is conducted by examining the data for any signs of duplication. Duplicate data could indicate issues in data collection or processing, potentially impacting the overall analysis results. The entire dataset meticulously reviewed for repeated entries and ensured that unique data was accurately represented. Upon identifying duplicate data, we took necessary actions such as merging or deletion to eliminate redundancy and enhance the integrity of the dataset.

$$(1 - (\frac{NDV}{N})) * 100 \tag{5.3}$$

4. **Timeliness**
   Timeliness is a critical element of data quality that aids in assessing the currency of the data. When gauging timeliness, two primary components should be considered: Currency and Volatility. Currency pertains to the time lapse between data generation and its use. Data with a brief time gap between creation and utilization is generally more pertinent and accurate, as it reflects the most recent conditions or situations. The second component, Volatility, denotes the duration of data relevance, signifying how long the data remains valid and useful. Data with low volatility remains relevant and accurate for an extended period, while data with high volatility may swiftly become outdated and less applicable.

$$MAX[(1 - \frac{Currency}{Volatility}), 0] \tag{5.4}$$

Table 4.1 of the DQCF documentation furnishes definitions and comprehensive descriptions of these dimensions. Furthermore, to facilitate a structured and replicable assessment, DQCF integrates specific formulas for evaluating each dimension, as delineated above. These formulas establish a standardized approach for calculating the quality scores, ensuring that the assessment is both objective and aligned with best practices in data quality management. The aims of this quality scoring is to guarantees that data meets the operational and analytical requirements of the organization and adheres to high-quality standards, thereby bolstering effective and reliable business decisions.

## 5.3 Chapter Summary

This chapter contains a comprehensive explanation of the DMF and DQCF, including effective methods for integrating the two framework. It provides detailed insights into the functions of each framework and explains every step of the process. Additionally, the chapter outlines the methods used to scoring various aspects of data quality, such as accuracy, consistency, completeness, and timeliness, giving readers a thorough understanding of how the framework can improve overall data quality and how to calculate the score.

# Chapter 6

# Validation

This section intends to scrutinize a compact prototype of a data cleansing process developed to validate DQCF. The primary objective of this prototype is to assess the efficacy of the designed framework and recognize its favorable influence on data quality. By implementing this prototype, the aim is to ensure that the steps in DQCF are correctly executed to enhance the accuracy, consistency, completeness, and timeliness of data, thereby rendering the resulting data more reliable for analysis and decision-making.

The validation process harnesses data gathered from detergent making machine in a manufacturing company. This dataset encompasses readings from numerous sensors, resulting in millions of data rows. However, this research focuses on the analysis of only six sensors. The collected data is in the form of time series data recorded daily. This selection of time series data enables the observation of changes and patterns over time, which is crucial for data quality analysis. Through the utilization of this subset of data, the aim is to effectively carry out the validation of the DQCF, consequently demonstrating the tangible impact of implementing this framework for enhancing the quality of the data produced. This chapter is organized into sections that delve into critical facets of prototype development and implementation.

Primarily, a discussion on the tools and technology employed in crafting this prototype will be undertaken. The selection of appropriate tools and technology assumes paramount importance to ensure the effective and efficient operation of the prototype. Subsequently, the user interface of the prototype will be depicted, offering a visual representation of user interactions with the system. An intuitive and user-friendly interface is essential to facilitate the user's seamless execution of the steps within the DQCF.

Ultimately, the chapter will present the findings of the data quality assessment, serving as the ultimate objective of this research endeavor. These findings will elucidate the effectiveness of the DQCF in enhancing the quality of the data generated. Through these subsections, readers will garner a comprehensive understanding of the prototype development process, the resultant achievements, and how this prototype substantiates DQCF validation. Consequently, it is aspired that this research will impart a seminal contribution to the realm of data quality in the manufacturing domain.

## 6.1 Tools

To create a data cleansing prototype, the primary tool utilized is Python. Python was chosen due to its widespread popularity as a programming language, supported by a diverse set of features that are highly suitable for data processing. Python boasts a rich ecosystem with various libraries that facilitate different stages of data processing, ranging from cleansing, analysis to visualization. In developing this prototype, several key libraries were employed. The utilization of these libraries enables the development of efficient and effective prototypes, with the capacity to handle and analyze large volumes of data optimally. The list of the library that has been used can be seen below.

Table 6.1: Tools

| Name | Function |
|---|---|
| Pandas | This library is crucial for tasks such as data manipulation, analysis, and processing. It really helps to work with the data |
| Numpy | This library helps to do easy scientific calculation in the prototype. |
| Matplotlib | Visualisation in the prototype is supported by this library. |
| Ipywidgets | library that shows interactive widgets. |
| Ydata profiling | Pandas profiling is a library that provide a detailed EDA to see the background of the data such as the number of variable exist in the data set, duplicate data, missing value, number of observation, etc. |
| Ipython | helps to display media in Jupyter notebook in this case is to support the show the profiling report by Ydata profiling |

## 6.2 User-Interface

The user interface display in this prototype uses Jupyter Notebook with a very simple design, supported by the use of the Ipywidgets and IPython libraries. This prototype is equipped with several functions that are used to display important information about the available dataset, making it easier for users to analyze the data before the data is processed.

This prototype is divided into seven tabs which can be seen in Figure 6.1. One tab contains information about the entire raw dataset, including data from several sensors. The other six tabs each display specific information about the six predefined sensors. With this division, users can easily access and analyze data from each sensor separately or as a whole.

The Raw Data tab features a table displaying data and information on the data type of each column. Additionally, it includes a Clear Data button to delete data and a Show Data button to redisplay the data. This functionality enables users to conveniently control data visibility based on their analysis requirements.

In the other tabs, the initial view consists of one button and four accordion widgets:

- Show Data (button)
- Profiling Report
- Outliers
- Pre-scoring

- Post-scoring

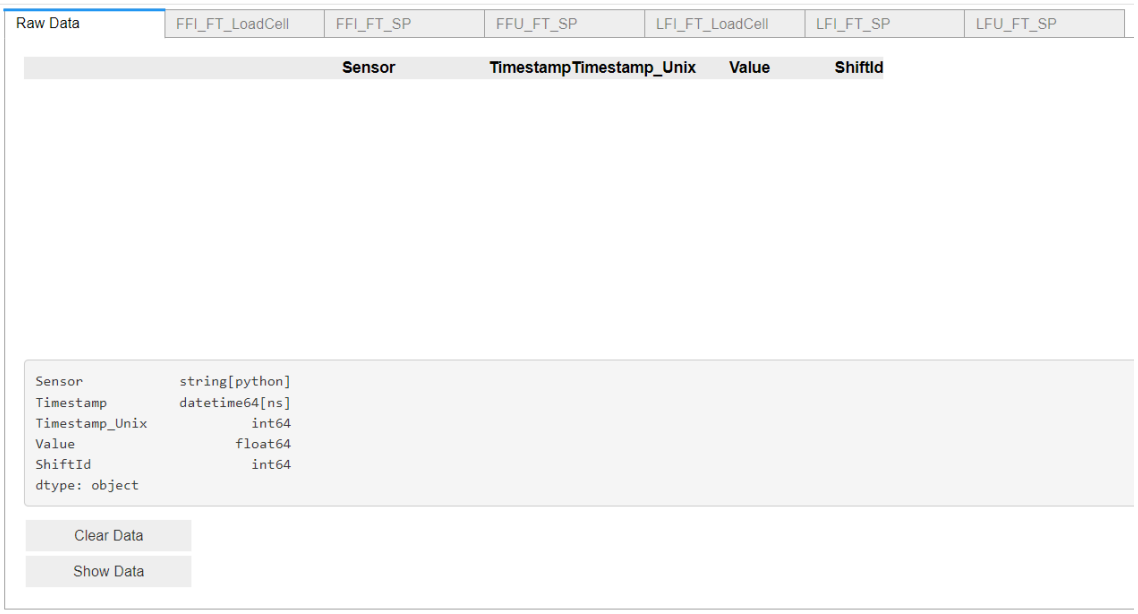


FIGURE 6.1: Main Interface of Raw Data

Within this tab, the displayed data is restricted to the first 5 rows, offering a preliminary overview of the dataset's contents without overwhelming the display with excessive information which can be seen in Figure 6.2. This layout allows users to navigate and analyze data from various perspectives, ensuring an effective and organized analysis workflow.

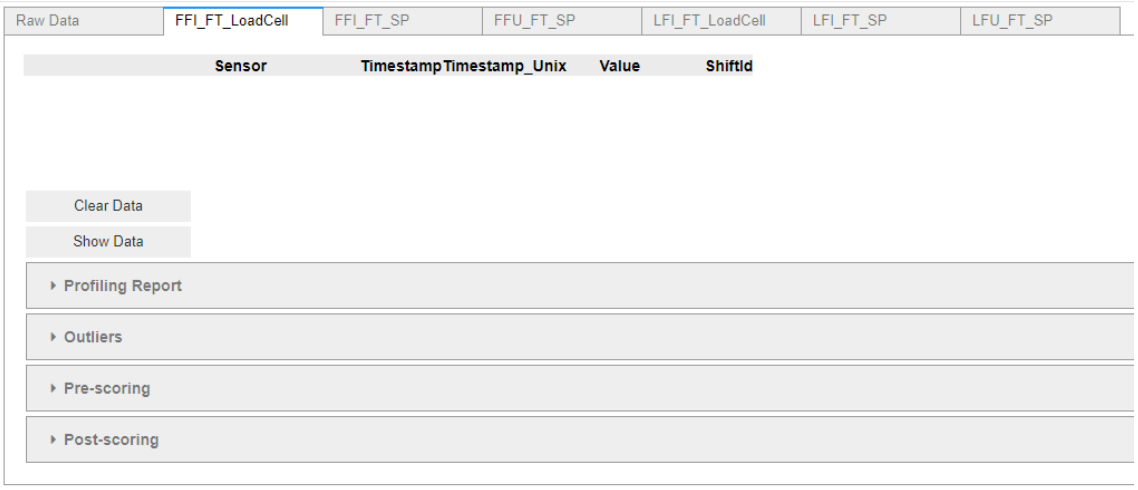


FIGURE 6.2: Main Interface of Sensor

### 6.2.1 Profiling Report

The profiling report is a key feature of the accordion widgets available in this prototype. It is designed to offer comprehensive information on the existing sensor data. This widget displays important statistics such as the number of variables, observations, and missing

cells, as shown in Figure 6.3. Additionally, it provides a list of any duplicate data in the dataset, details on variable correlations, and other pertinent information. The profiling report enables users to gain a thorough understanding of the condition and quality of the sensor data being utilized. Armed with this detailed insight, users can more effectively conduct data analysis and pinpoint potential issues in the dataset. As a valuable tool in the data validation and cleansing process, the profiling report helps pave the way for a more in-depth analysis stage.

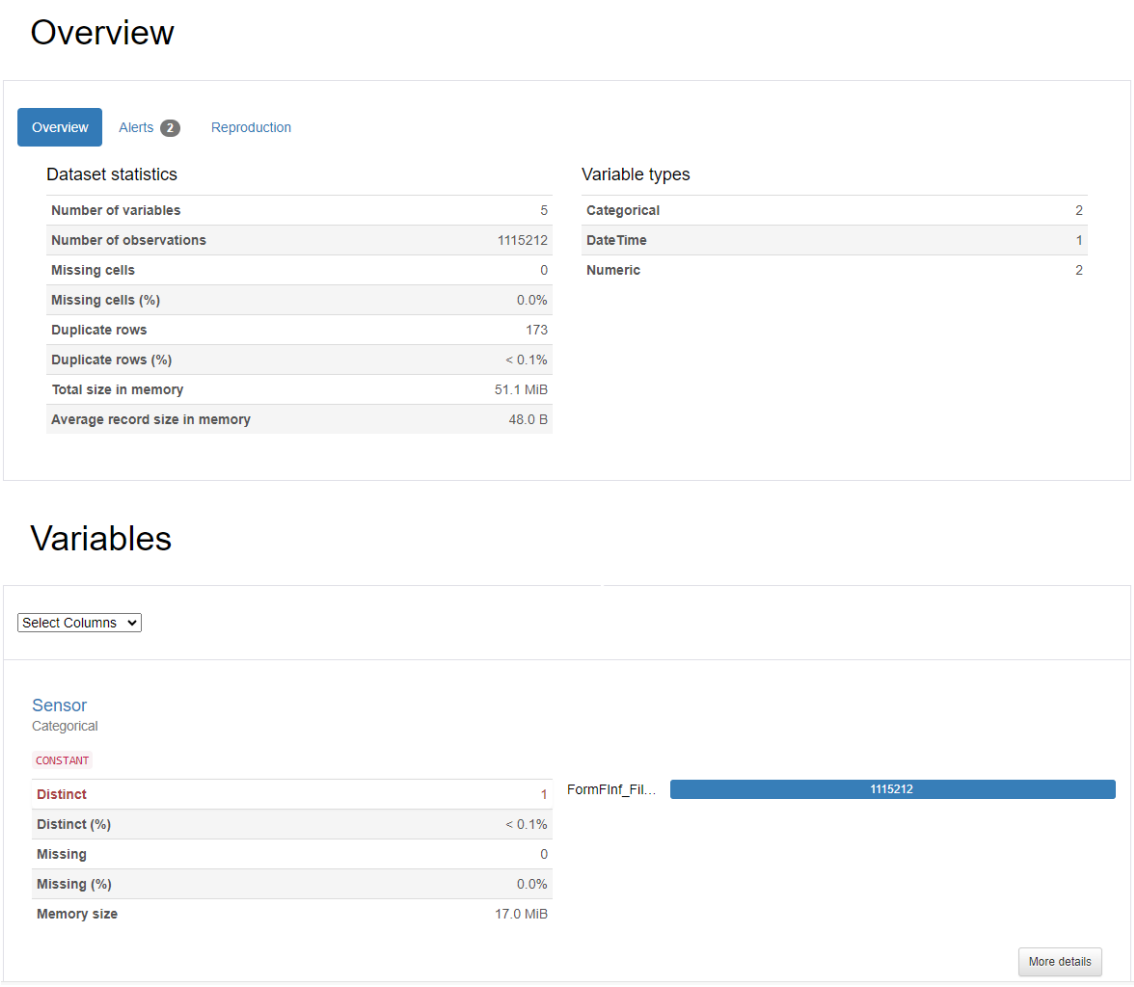


FIGURE 6.3: Profiling Report

### 6.2.2 Outliers

Contained within this accordion widget are two graphs providing insights into outliers. The first graph illustrates the results of outlier detection based on predetermined minimum and maximum values Figure 6.5. The second graph showcases outlier detection outcomes using the K-Nearest Neighbors (KNN) algorithm Figure 6.6. These graphs offer a comprehensive view of data distribution, facilitating the identification of outliers and enabling a more thorough and precise analysis of the existing data. This detailed analysis enhances the effective utilization of the data for various purposes. The detail information about this will be explain in section 6.3.1

### 6.2.3 Pre-scoring and Post-scoring

These two accordion widgets share a similar appearance but serve distinct purposes. The pre-scoring widget presents the results of the initial scoring of the raw sensor data to provide an overview of data quality and characteristics prior to the cleansing stage. In contrast, the Post-scoring widget showcases scoring results (Figure 6.7) after the data has undergone the cleansing process outlined in DQCF. There is only one difference in this accordion, where in the pre-scoring, the data displayed is only accuracy, consistency and completeness, whereas in the post-scoring accordion there is one additional dimension, namely timeliness. An explanation of this can be seen in section 6.3.2.

This setup enables users to compare data quality before and after cleansing and comprehend the impact of each step within the framework. The information within these widgets is essential for ensuring that the data used in the final analysis is of high-quality and free from anomalies or errors that could undermine the accuracy of the results. This clear and informative structure facilitates effective data access and evaluation, thereby supporting more precise decision-making based on validated data.

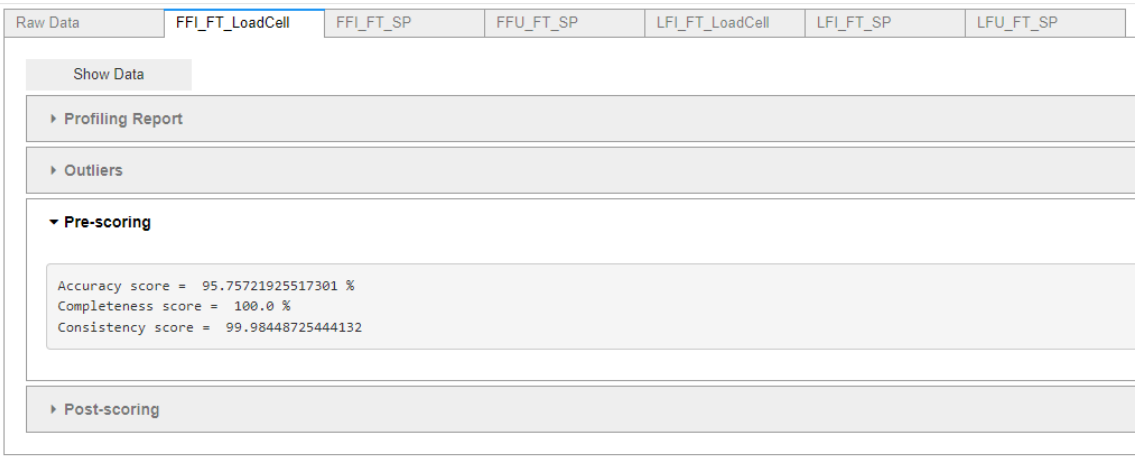


FIGURE 6.4: Pre-scoring

## 6.3 Findings of the Data Quality Score

The purpose of this prototype was to serve as a tool for validating DQCF. In this section, a detailed explanation of the results obtained from the development of this prototype will be provided, covering aspects such as outlier detection, pre-scoring scoring, and post-scoring.

### 6.3.1 Outliers

In this prototype, two methodologies is utilized to detect outliers. The first methodology involves setting predetermined minimum and maximum values. For Sensor A, the minimum value is 11 and the maximum value is 21. An analysis in Figure 6.5 reveals 47,316 outliers, with outlier values ranging from 1.5335 to 35.3700, indicating significant anomalies in the data.

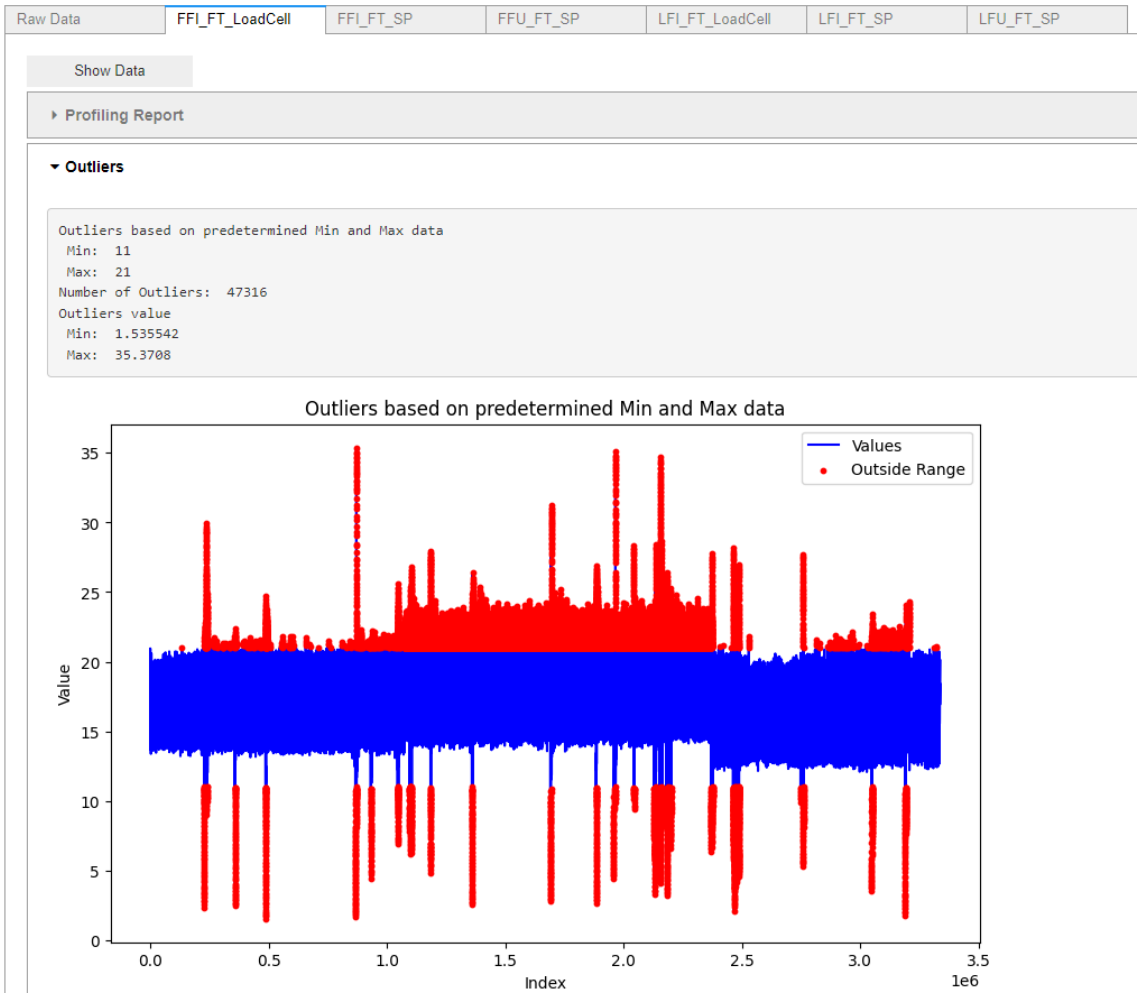


FIGURE 6.5: Outliers based on Minimum and Maximum Values

The second approach relies on the utilization of the K-Nearest Neighbors (KNN) algorithm for the purpose of outlier detection (Figure 6.6). Upon conducting an analysis, it was observed that the outliers identified through the application of KNN exhibited no discernible disparities in their distribution when compared to those detected through the utilization of the minimum and maximum value approach. This finding indicates that the two methodologies possess nearly identical capacities in identifying outliers. Consequently,

the use of KNN as an alternative technique for outlier detection can be deemed credible, particularly in scenarios where consideration is given to algorithmic complexity and computational requisites.

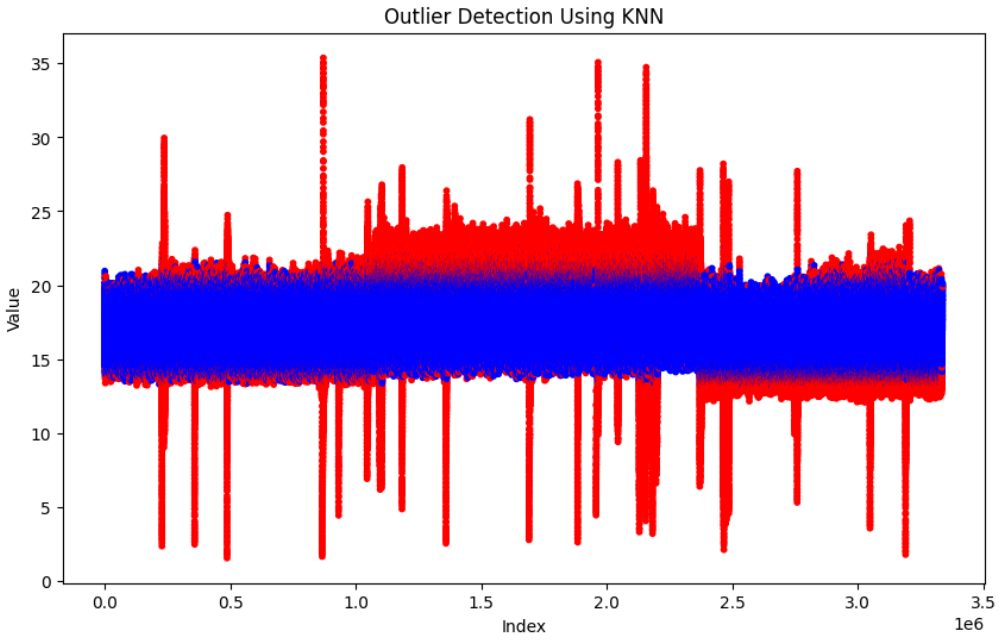


FIGURE 6.6: Outliers Detection Using KNN

Following outlier detection, those identified based on minimum and maximum values are aggregated into a dedicated source specifically created for the same sensor. This represents the improvement stage, aiming to ensure all outliers are appropriately identified and managed. This step is crucial for enhancing data quality before further analysis, ensuring clean and accurate data for research purposes.

### 6.3.2 Pre-scoring and Post-scoring

In the pre-scoring process, the data is initially assessed before undergoing the data cleansing process. The formula used in this process is identical to that used in the post-scoring, detailed in the section 6.7. The primary disparity between these processes is that the post-scoring involves an additional Timeliness score, unlike the pre-scoring, where the Timeliness score is not factored in. For this process, the validation is done using two different datasets. The first dataset is manufacturing data that is collected from manufacturing company and another dataset is sales data from kaggle. The purpose of this is to see whether the framework is working only for manufacturing data or it is can be used generally for all data.

#### Manufacturing Data

During the pre-scoring, an evaluation is conducted to pinpoint areas that necessitate improvement before data cleansing. This encompasses an assessment of the completeness, consistency, and accuracy of the data. The intention is to obtain an initial overview of the raw data's quality and to delineate the steps required to enhance the data quality.

In the post-scoring stage, subsequent to the data being cleaned in line with the framework's procedures, a re-scoring is conducted, incorporating the Timeliness component. Timeliness gauges the data's currency and its relevance within the current analysis context. By integrating the Timeliness score, the post-scoring delivers a more comprehensive assessment of the cleaned data's quality.

1. **Accuracy**
   **Pre-scoring**
   In the initial assessment, the accuracy score was 95.75%, influenced by the presence of 47,316 outliers out of a total of 1,115,212 data rows. This high number of outliers had a noticeable impact on the accuracy score. The accuracy score can be calculated using the formula in 5.1. By plugging in the relevant values into this formula:

$$\frac{1067896}{1115212} * 100 = 95.75\%$$

   **post-scoring**
   Following the data cleansing process, where the outliers were moved to a special source for outliers, the accuracy score improved to 100% which can be seen in Figure 6.7 with calculation like this after the cleansing process done:

$$\frac{1067731}{1067731} * 100 = 100\%$$

   This process involved removing all identified outliers, ensuring that the main dataset is now free from anomalies and errors. These results demonstrate a substantial enhancement in data accuracy, from 95.75% to 100%, indicating that the cleaned dataset meets the expected quality standards.

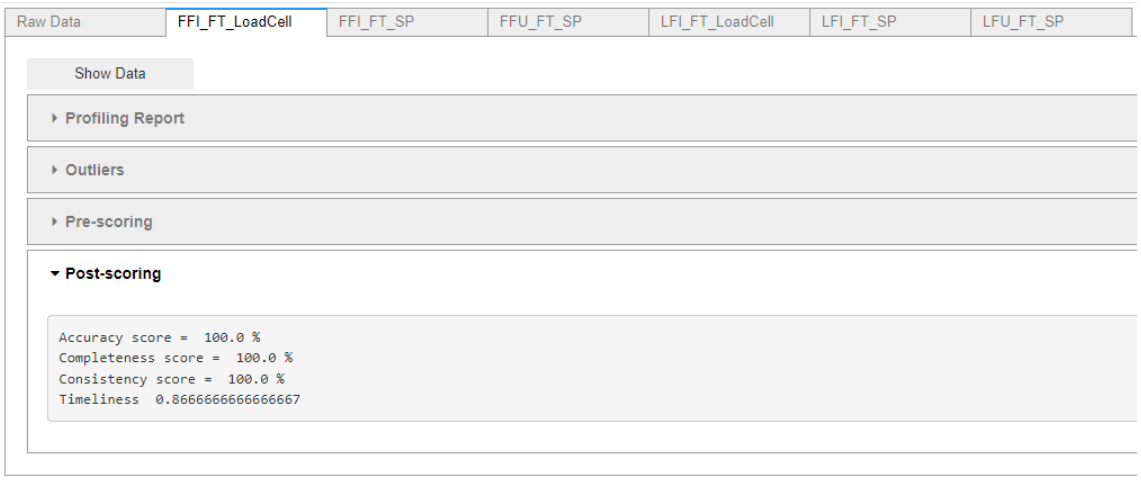


FIGURE 6.7: Post-scoring

2. **Completeness**
   **Pre-scoring and Post-scoring**
   The pre-scoring results revealed a perfect 100% completeness score for the Sensor A dataset, attributed to the absence of any missing values (NA) upon inspection.

Given the lack of missing values, there was no variance between the pre-scoring and post-scoring for the completeness dimension.

The completeness score is derived by inputting the count of non-missing values into the formula 5.2.

$$\frac{1115212}{1115212} * 100 = 100\%$$

A 100% completeness score indicates that the Sensor A dataset is entirely complete, with no missing values. This is crucial for ensuring data accuracy and thorough analysis. High data completeness guarantees that no missing value could impact analysis results and data interpretation.

Hence, both the pre-scoring and post-scoring phases yield flawless results for the completeness dimension, signifying that the dataset meets the key data quality criterion of completeness. This underscores the importance of verifying and upholding data completeness within the data quality assurance process in this research. More comprehensive insights into the pre-scoring outcomes and data cleansing procedures will be provided in the subsequent section of this research.

3. **Consistency**
   **Pre-scoring**
   In the pre-scoring process for the consistency dimension of Sensor A, a remarkable score of 99.98% was achieved, revealing the high consistency of the dataset used. This score is determined by the number of non-duplicate values in the Sensor A dataset.

   The consistency score is calculated using the following formula with values obtained before data cleansing:

$$(1 - (\frac{1115212}{1115212})) * 100 = 99.98\%$$

   **Post-scoring**
   And after data cleansing, by removing duplicate values, the consistency score becomes:

$$(1 - (\frac{0}{1067731})) * 100 = 100\%$$

   Following the prescribed data cleansing stages in the framework, a highly satisfactory perfect consistency score of 100% was attained. Through this process, identification and removal of duplicate data ensured that each entry in the dataset was unique and consistent.

   The initial score of 99.98% already indicated a very high level of consistency within the dataset, even before data cleansing. However, after thorough data cleansing, all duplicate values were removed, leading to a perfect consistency score of 100%. Achieving a perfect consistency score after data cleansing assures that the dataset is devoid of duplication and highly consistent. This serves to uphold data integrity and reliability in subsequent analysis.

4. **Timeliness** As elucidated previously, the scoring process for the timeliness dimension is specifically conducted during the post-scoring phase owing to its distinct calculation methodology compared to other dimensions. While other dimensions utilize the dataset's content as a basis for scoring, the timeliness dimension relies on the data's production and usage times (referred to as currency) and its relevance over time (referred to as volatility) as reference points. Consequently, in the scoring for

timeliness, only the variable denoting the data's production time is derived from the dataset, with the other two variables manually configured.

Scoring for timeliness yields disparate values compared to other dimensions. Whereas scoring in other dimensions is expressed in percentages, timeliness scoring results manifest as binary values: 0 denotes untimely data, while 1 indicates timely data. The formula for calculating the timeliness score (5.4) is provided below:

$$MAX[(1 - \frac{4}{30}), 0] = 0.86$$

In this research, the timeliness score is 0.86, closely approaching 1, signifying the timeliness of the dataset. This result was achieved by setting the data usage time on October 2 and the data production time on September 28, resulting in a currency of 4 days, with volatility set at 30 days. A higher currency value yields a lower timeliness score.

Consequently, the obtained timeliness results denote the relevance and currency of the data for analytical purposes. Timely data is crucial as it ensures the accuracy of analyses and subsequent decisions by aligning with the latest conditions. This explanation underscores the importance of considering currency and volatility in evaluating timeliness and their influence on the final timeliness score.

**Sales Data**

Following the validation using data obtained from manufacturing companies, the next validation is done using sales data and no significant difficulties were encountered. The pre-scoring results demonstrated that the dataset has excellent quality with an Accuracy score of 98.79%, Consistency of 100%, and Completeness of 100%. This indicates that the dataset used is in very good condition.

In the post-scoring stage, the scores for Accuracy, Consistency, and Completeness all increased to 100%, indicating that the data cleansing and management process successfully improved data quality. However, the Timeliness dimension scored 0.33 due to the currency data value is 20 days and the volatility being set for 30 days. Despite the lower Timeliness score, this reflects the challenges of maintaining up-to-date data in a dynamic operational cycle.

These validation results confirm that DQCF effectively enhances data quality in various domains. With high scores on Accuracy, Consistency, and Completeness, and a clear Timeliness evaluation, this research successfully demonstrates that the implemented framework can generate high-quality datasets. This dataset is now ready for further analysis and improved decision-making in manufacturing and sales environments.

## 6.4 Chapter Summary

The chapter delves into the prototype developed to validate DQCF. It encompasses various aspects, including the tools that is used, an explanation of the prototype's appearance,and the findings from applying the framework. Each element is elaborated upon to offer a comprehensive understanding of the prototype's function and effectiveness.

The method used to detect outliers is thoroughly explained, encompassing the two main approaches applied: detection based on predetermined minimum and maximum values, and the utilization of the K-Nearest Neighbors (KNN) algorithm. The results of these two approaches were compared to ensure consistency and accuracy in identifying outliers.

Additionally, this section discusses the scoring results from the pre-scoring and post-scoring processes using two different datasets, manufacturing data and sales data. The pre-scoring process involves assessing data before cleansing, resulting in an initial score reflecting the quality of the raw data. Following the data cleansing process, a post-scoring is conducted, encompassing all dimensions of data quality, including timeliness. The final results exhibit a marked improvement in data quality after implementing DQCF.

These findings authenticate the DQCF's effectiveness in assessing and enhancing data quality. By delivering satisfactory results across various dimensions of data quality, the prototype underscores its reliability as a potent validation tool. A comprehensive explanation of the outlier detection method, the scoring process, and the attained improvements in data quality corroborate the successful achievement of the research objectives. Further analysis of the results and the implications of these findings will be addressed in the subsequent section of this research.

# Chapter 7

# Discussion

In this chapter, the elucidation of the answers to the sub-research inquiries will be undertaken, in addition by a thorough comparative analysis between the newly proposed framework and pre-existing frameworks. The primary objective of this comparative delving is to discern and delineate the congruities and disparities between the Data Quality Control Framework (DQCF) under consideration and its antecedent frameworks. The aim is to elucidate the innovative characteristics and the overarching superiority of the proposed framework. This effort will equip readers with a broad understanding of how this new framework is poised to offer more efficacious and rapid solutions in improving data quality.

## 7.1 Summarize of Answer to Sub-research Question 1-5

- According to [30], data that is collected and processed correctly is very valuable for business organizations because it can improve decision making and simplify operations. [35] research also highlights the importance of data quality in modern business practices. Conversations with manufacturing company representatives reveal that they face major challenges regarding data quality, underscoring the urgent need for a robust data quality framework. The most commonly used dimensions of data quality are completeness, timeliness, accuracy, and consistency, and these statements are valid based on the findings of this research.

- Based on the results of this research, the identified frameworks can be categorized into two groups: data quality frameworks and data pre-processing or cleaning frameworks. Data quality frameworks such as the one developed by [35] use the TDQM method, while other frameworks such as DQPE by [5] and the [39] framework focus on improving data quality through various processes. This section describes the findings regarding the data quality framework, while the data preprocessing or cleaning framework will be explained in the next section.

- As mentioned earlier, several frameworks are available for data cleaning or preprocessing with different methods. For example, [26] uses statistical, probabilistic, and computational domain theory for cleaning indoor positioning data, as well as methods such as uncertainty modeling and error correction. [31] focuses on imputation of missing values using queries and reasoning, while [28] uses statistical methods and unsupervised clustering for anomaly detection. [41] summarizes important techniques in data cleaning such as linear interpolation and regressive modeling for missing values, as well as Gaussian-based and cluster-based methods for outlier detection. [32]

highlights various cleaning techniques for various types of data, including filters to reduce noise in time series data, as well as data normalization and segmentation.

- Implementing data quality is a complex process that faces several obstacles, including high costs and time required, as pointed out by [17] and [31]. Detecting anomalies in data quality is also a big challenge because anomalies are often hidden and difficult to spot. Additionally, the diversity of data sources and types adds complexity, making data integration difficult and complicating the process of ensuring data quality, as [7] found.

## 7.2 Answer to Sub-Research Question 6-7

### 7.2.1 Sub-RQ6: How can a robust data quality framework be designed to enhance data quality and operational efficiency specifically for the manufacturing domain?

The primary aim of this research is to construct a framework tailored to ensuring the production of high-quality data that can be effectively utilized for analysis and decision-making, with a particular focus on offering substantial benefits to stakeholders, especially companies operating within the manufacturing sector.

To achieve this goal, the framework was conceptualized by drawing upon existing frameworks from prior research, subsequently refining it to align with the specific requirements and operational context of the company. This method was employed to ensure that the resulting framework is well-suited to the organization's distinctive needs.

The framework is structured around four key stages: Asses, Action, Enhancement, and Quality Scoring. Each stage encompasses a set of interconnected processes, collectively aimed at establishing an effective framework for advancing data quality.

1. **Assess**
   The Assessment stage encompasses activities such as data partitioning, creation of data profiles, data type validation and transformation, identification of missing values, duplicates, and outliers, as well as pre-scoring. The primary objective at this stage is to identify existing data quality issues and evaluate the initial quality of the data, in order to formulate appropriate measures for subsequent stages.

2. **Action**
   The Action stage is dedicated to data cleansing, encompassing the removal of missing and duplicate data to enhance the integrity of the dataset, thereby ensuring the availability of clean and reliable data for analysis.

3. **Enhancement**
   The Improvement stage involves the implementation of additional measures to enhance and enrich the data based on prior evaluation findings. For instance, in this research, measures were taken to relocate outliers to an alternate source, thus ensuring that the data subjected to analysis is not skewed by extreme values and represents a true reflection of the underlying phenomena.

4. **Quality Scoring**
   Performs a post-scoring on processed data to ensure that it meets established quality standards. This stage is a crucial step for final validation of data quality before it is used in analysis and decision making because at this stage, users can directly see

the real value of the data quality itself after all the processes in this framework have been completed.

An integral aspect of this framework is the scoring process, which encompasses pre-scoring and post-scoring. During this stage, the data undergoes evaluation to determine its overall quality. Each data point is assigned a numerical value as a percentage, allowing users to directly gauge the quality of the data. This scoring system provides clear direction on areas needing improvement and ensures that each phase of the framework significantly contributes to enhancing data quality, setting this framework apart from others.

This comprehensive framework not only ensures the availability of high-quality data for analysis and decision-making but also equips stakeholders with valuable tools to continuously assess and enhance the quality of their data. It is anticipated that the implementation of this framework will yield substantial advancements in the efficiency and effectiveness of data processing for companies, particularly those within the manufacturing sector, and could potentially serve as a benchmark for other organizations grappling with similar challenges pertaining to data quality.

### 7.2.2 Sub-RQ7: To what extent does the implementation of a developed framework effectively enhance data quality?

The assessment of the proposed framework's effectiveness involved the design and testing of a prototype using a dataset obtained from a manufacturing company. The validation results revealed substantial improvements in data quality following the implementation of all the processes outlined in the framework. Notably, the pre-scoring and post-scoring comparison, as detailed in Subsection 6.3.2, demonstrated significant changes in data quality dimensions, including accuracy, completeness, and consistency. Furthermore, the dimension of timeliness exhibited results only after the data had undergone processing and enhancement within the framework.

Beyond the evaluation with the manufacturing dataset, the prototype underwent testing using a sales dataset. The results of this test similarly exhibited a noteworthy shift in data quality scores, both before and after the data was subjected to the cleansing process stipulated by the framework. This outcome underscores the capacity of the proposed framework to elevate data quality across diverse types of datasets, extending beyond the confines of manufacturing data.

The implementation of this comprehensive framework yielded considerable enhancements in data quality, rendering the data deployed for analysis and decision-making more dependable. This not only amplified operational efficiency but also conferred a competitive edge upon companies, underscoring its particular relevance within the manufacturing sector. With high-quality data at their disposal, organizations can make more informed and strategic decisions, heighten productivity, and mitigate the risk of errors.

In summary, the proposed framework has demonstrated efficacy and successfully achieved its intended objectives. Furthermore, validation has indicated that the framework can be applied not only to manufacturing datasets but also to other varieties, such as sales datasets. Consequently, the framework holds promise for widespread application across diverse domains, offering substantial benefits to organizations leveraging it to uphold high data quality standards.

## 7.3 Comparison DQCF with Existing Frameworks

In the prior section 7.2.1, the origins of the framework are delineated, citing the influence of various frameworks in antecedent research. Section 2.7.3 expounds upon several frameworks documented in previous literature. This section aims to furnish a more comprehensive overview of the factors that differentiate the proposed framework from antecedent frameworks.

The framework proposed by [33] comprises of four principal components: BIG DATA Sources, Data Quality Class Selection, Data Quality Pre-processing Evaluation, and Data Storage. Each section encompasses several processes resembling those in DQCF, such as data cleansing, data enrichment, and data quality profiling.

Another scholarly work conducted by [2] introduced the QoDID Framework, which is comprised of five distinct stages: acquire, assess, process, improve, and integrate. The proposed framework, DQCF, shares several stages with QoDID, particularly in terms of assessment, process, and improvement. Additionally, a comparable data pre-processing stage was also identified in the research by [37].

TABLE 7.1: Analysis of Previous Frameworks

| No | Reference | Steps | | | | |
|---|---|---|---|---|---|---|
| | | **PC** | **P** | **C** | **E** | **S** |
| 1 | Ali, T.Z., Maatuk, A.M., Abdelaziz, T.M., Elakeili, S.M. (2020) [3] | | x | x | x | |
| 2 | Al-Masri, E., Bai, Y. (2019) [2] [2] | | x | x | x | |
| 3 | Burkhardt, A., Berryman, S., Brio, A., Ferkau, S., Hubner, G., Lynch, K., Mittman, S., Sonderer,K. (2017) [6] | | x | | | x |
| 4 | Ding, N.B., Mit, E. (2023) [16] | | x | x | | |
| 5 | Juneja, A., Das, N.N. (2019) [21] | | x | x | x | |
| 6 | Soto, P.C., Ramzy, N., Ocker, F., VogelHeuser, B. (2021) [31] | | | x | | |
| 7 | Sreenivas, P., Srikrishna, C.V. (2013) [32] | | | x | | |
| 8 | Taleb, I., Dssouli, R., Serhani, M. A. (2015) [33] | | x | x | x | |
| 9 | Tsai, W.L., Chan, Y.C. (2019) [34] | | | | | x |
| 10 | Wahyudi, T., Isa, M.S. (2023) [35] | | x | | | x |
| 11 | Xu, D., Zhang, Z., Shi, J. (2022) [39] | | | | x | x |

Note:

- PC: Pre-quality Check
- P: Profilling
- C: Cleansing
- E: Enhancement
- S: Scoring

Despite these similarities, it is important to emphasize that DQCF has differences from previous research. Previous studies have largely focused on improving data quality without paying careful attention to comprehensive data quality assessment. Although

certain studies include a scoring process, precise details such as assessing data quality dimensions are often lacking, there is no data profiling, no data processing and the most significant difference with DQCF is that in previous studies, the scoring process was only carried out once (Table 7.1. whereas in DQCF, the scoring process is carried out twice, namely on raw data and data that has gone through a cleansing process.

The DQCF places great emphasis on careful and specific data quality assessment. The scoring process is carefully designed to measure overall data quality, providing actual percentage scores for each dimension of data quality, including accuracy, consistency, completeness, and timeliness. This rigorous approach ensures that improvements in data quality can be measured objectively and transparently, providing a clear assessment of the effectiveness of each step in the framework.

Additionally, this framework not only identifies and corrects data quality issues, but also provides a structured mechanism for ongoing data quality assessment and reporting. Therefore, this framework presents a more holistic and comprehensive approach to ensuring high-quality data.

By integrating a comprehensive assessment process, this framework not only improves data quality, but also equips users with the tools to measure and monitor the quality of their data in real time. This provides the added benefit of ensuring that the data produced is truly reliable and ready for effective analysis and decision making.

In conclusion, the DCQF offers a more comprehensive and effective solution compared to previous frameworks. An emphasis on detailed and continuous data quality assessment ensures that each step in the data processing process contributes to real, measurable quality improvements. This provides great added value for companies, especially in the manufacturing sector, which depend on high-quality data for their operations and strategic decision-making.

# Chapter 8

# Conclusion and Future Work

This chapter concludes the research while highlighting contributions, limitations and recommendations for future research.

## 8.1 Conclusion

In conclusion, this research has culminated in the development of a framework designed to ensure high-quality data. The main result of this research is a framework that is inspired by various existing frameworks from previous research and then developed by adapting to the specific needs of the manufacturing domain. Subsequently, this framework, DQCF, is integrated into the existing framework of the company.

The DQCF comprises four pivotal stages: Assess, Action, Enhancement, and Quality Scoring. Each stage encompasses several interrelated processes which ensure the high quality of data.

To ascertain the relevance and efficacy of the developed framework, a prototype was utilized as a testing tool. This prototype facilitates real-time testing of the framework, aiding in the identification of strengths and areas for improvement. Validation results, employing manufacturing datasets, evince that the proposed framework yields highly satisfactory results, manifesting substantial enhancements in data quality. During the validation process, the post-scoring for all data quality dimensions attained 100%, except for timeliness, which is binary (0 or 1), underscoring the efficacy of the framework in ensuring high-quality data. So, it is not just a framework that provides some kind of steps or guidelines to data analysts or anyone analyzing data, but it also provides some kind of measure of data quality. In addition, validation was also conducted using sales datasets, yielding similarly satisfactory results, suggesting that the DQCF can be universally employed.

This research contributes significantly to the literature and practice in the realm of data quality, particularly within the manufacturing sector. Validation through prototypes corroborates the preparedness of the framework for real-world implementation, bolstering decision-making processes and augmenting operational efficiency.

Ultimately, this research substantiates that the proposed framework can yield substantial improvements in data quality, rendering it an invaluable tool for more dependable analysis and decision-making. With appropriate implementation, companies can accrue long-term benefits from high-quality data, including a competitive advantage and enhanced operational performance.

## 8.2 Contribution

a. **Scientific Contribution**
The academic significance of this research lies in its innovative and simple approach in generating high-quality data. By establishing the proposed framework, this research contributes significantly to the theoretical framework regarding data quality, especially in the manufacturing context, which to date has rarely been discussed in the literature.

The approach used in this research combines several existing frameworks, then modifies them to meet specific needs in the manufacturing industry. This makes an important novel contribution to understanding and managing data quality in this sector. This research not only offers a practical solution to the data quality problems faced by manufacturing companies, but also enriches the academic literature by providing a strong theoretical foundation and relevant practical applications.

Thus, this research helps fill a gap in the research of data quality in manufacturing and provides useful guidance for researchers and practitioners in their efforts to improve data quality. The proposed framework can serve as a reference for further research and practical implementation in industry, ultimately contributing to increased operational efficiency and better decision making based on accurate and reliable data.

b. **Practical Contribution**
The primary objective of this research is to address data quality issues within the manufacturing sector. By directly gathering data from manufacturing companies, valuable insights have been obtained. The research has led to the development of a practical framework with specific steps to enhance data accuracy, consistency, completeness, and timeliness.

The subsequent phase involves disseminating the research findings to manufacturing companies for integration and execution. Therefore, this research not only adds value in the academic sphere but also offers practical benefits by assisting companies in managing and enhancing their data quality. Ultimately, this contributes to better decision-making and improved operational efficiency within these organizations.

## 8.3 Limitation

Although this research has made significant progress in enhancing the quality of data, several limitations need to be noted. The following is a list of limitations encountered in this research:

### Handling Complex Data

The validation of the framework is restricted to structured data prevalent in manufacturing companies, thereby excluding an assessment of its efficacy in processing unstructured data. Subsequently, the robustness of the framework in handling more complex and diverse data formats remains unexamined.

### Prototype Development Time:

Temporal constraints within this research, particularly in prototype development, have impeded the potential for creating optimal prototypes that would serve to validate the proposed framework in a comprehensive fashion. Augmenting the available timeframe would

enable enhancements in both the quality and functionality of the prototype, thereby facilitating more thorough and exhaustive appraisals. Such a measure would prove instrumental in meticulously identifying the strengths and weaknesses of the framework.

**Real Implementation:**

This research aligns with the design cycle expounded by Wieringa (2014), which encompasses four primary stages: Problem Statement, Treatment Design, Treatment Validation, and Implementation and Implementation Evaluation. However, owing to time restrictions, only the initial three stages could be executed. The inability to carry out the implementation and implementation evaluation stages due to these limitations hinders the capacity to evaluate the framework's performance within authentic operational environments over the long haul.

**Handling Outliers:**

The methodology adopted for handling outliers revolves around transferring them to a distinct file sans comprehensive identification or analysis of their causative factors or characteristics. As such, this approach falls short of providing exhaustive insights and holistic solutions.

## 8.4 Future Research Recommendation

It is important to take into account the following suggestions for prospective studies that aim to enhance and broaden the application of the established framework:

**Data Organization**

In forthcoming research, it would be valuable to explore the utilization of this framework for processing unstructured data in order to broaden its sphere of application and adaptability. Consequently, the framework would be able to handle a diverse array of data types, not confined solely to structured data. This extension would empower the framework to manage various data formats and sources, thereby offering heightened flexibility across a spectrum of industrial contexts.

**In-depth Examination of Outliers**

The existing method for addressing outliers in this research involves isolating them into a separate dataset without carrying out further analysis. In future inquiries, it is suggested to conduct a more comprehensive analysis to elevate the efficacy of outlier management. Subsequent research should consider employing a wider array of techniques for handling outliers, incorporating an analysis of their origins, characteristics, and impact on overall data integrity. Such an approach would yield profound insights and a more comprehensive resolution to the outlier conundrum.

### Real-world Application

It is advisable to directly implement this framework on an operational system to evaluate its efficacy and suitability in an authentic environment. Subsequent research endeavors are imperative to complete outstanding phases, such as full-scale implementation and its subsequent assessment, to ascertain the effectiveness and sustainability of the framework. By means of real-world implementation, researchers can identify practical challenges and fine-tune the framework to better align with operational requisites.

### Cost-based Implementation Strategy

Future investigations should encompass an examination of the expenses associated with implementing this framework in practical systems. This necessitates conducting a cost-benefit analysis to ensure that the benefits garnered align with the necessary costs. Such an approach will aid organizations in making well-informed decisions regarding the investments required to enhance data quality through the implementation of this framework.

By addressing these suggestions, future research stands to fortify the developed framework and ensure its broad and effective application across diverse operational landscapes. Continual exploration will facilitate overcoming current constraints and will expand the potential utilization of the framework across a spectrum of data types and industrial contexts, thereby delivering widespread advantages to adopting organizations.

# Bibliography


[1] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Data profiling. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 1432–1435, 2016. doi:10.1109/ICDE.2016.7498363.

[2] E. Al-Masri and Y. Bai. Invited paper: A service-oriented approach for assessing the quality of data for the internet of things. In *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)*, pages 9–97, 2019. doi:10.1109/SOSE.2019.00013.

[3] T.M. Ali, T.Z.and Abdelaziz, A.M. Maatuk, and S.M. Elakeili. A framework for improving data quality in data warehouse: A case study. In *2020 21st International Arab Conference on Information Technology (ACIT)*, pages 1–8, 2020. doi:10.1109/ACIT50332.2020.9300119.

[4] S.V.D. Berghe and K.V. Gaeveren. Data quality assessment and improvement: A vrije universiteit brussel case study. volume 106, pages 32–38. Elsevier B.V., 2017. doi:10.1016/j.procs.2017.03.006.

[5] P. Burggräf, M. Dannapfel, R. Förstmann, T. Adlon, and C. Fölling. Data quality-based process enabling: Application to logistics supply processes in low-volume ramp-up context. In *2018 International Conference on Information Management and Processing (ICIMP)*, pages 36–41, 2018. doi:10.1109/ICIMP1.2018.8325838.

[6] A. Burkhardt, S. Berryman, A. Brio, S. Ferkau, G. Hubner, K. Lynch, S. Mittman, and K. Sonderer. Measuring manufacturing test data analysis quality. In *2018 IEEE AUTOTESTCON*, pages 1–6, 2018. doi:10.1109/AUTEST.2018.8532518.

[7] L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. volume 14. Committee on Data for Science and Technology, 2015. doi:10.5334/dsj-2015-002.

[8] Q. Chen, Y. Liu, S. Hou, F. Duan, and Z. Cai. Data-driven methodology for state detection of gearbox in phm context. In *2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing)*, pages 1–6, 2021. doi:10.1109/PHM-Nanjing52125.2021.9612946.

[9] Y. Chernov. Test-data quality as a success factor for end-to-end testing - an approach to formalisation and evaluation. pages 95–101, 01 2016. doi:10.5220/0005971700950101.

[10] Po Chan Chiu, Ali Selamat, Ondrej Krejcar, King Kuok Kuok, Siti Dianah Abdul Bujang, and Hamido Fujita. Missing value imputation designs and methods of nature-inspired metaheuristic techniques: A systematic review. *IEEE Access*, 10:61544–61566, 2022. doi:10.1109/ACCESS.2022.3172319.

[11] C. Cichy and S. Rass. An overview of data quality frameworks. *IEEE Access*, 7:24634–24648, 2019. doi:10.1109/ACCESS.2019.2899751.

[12] Jonas Cleveland. Demystifying data cleansing vs data cleaning: A comparative analysis. https://jonascleveland.com/data-cleansing-vs-data-cleaning/. Accessed: 2024-08-07.

[13] Divya D. and Suvanam Sasidhar Babu. Methods to detect different types of outliers. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 23–28, 2016. doi:10.1109/SAPIENCE.2016.7684114.

[14] Dasu Dasari and P.Suresh Varma. Employing various data cleaning techniques to achieve better data quality using python. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 1379–1383, 2022. doi:10.1109/ICECA55336.2022.10009079.

[15] Martin De Saulles. *The Business of Data: Commercial Opportunities and Social Challenges in a World Fuelled by Data*. 06 2020. doi:10.4324/9780429427022.

[16] N.B. Ding and E. Mit. A framework of data quality assurance using machine learning. In *2023 13th International Conference on Information Technology in Asia (CITA)*, pages 88–93, 2023. doi:10.1109/CITA58204.2023.10262802.

[17] C.L Günther, E. Colangelo, H.H Wiendahl, and C. Bauer. Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. *Procedia Manufacturing*, 29:583–591, 2019. "18th International Conference on Sheet Metal, SHEMET 2019"x"New Trends and Developments in Sheet Metal Processing". URL: https://www.sciencedirect.com/science/article/pii/S2351978919301477, doi:10.1016/j.promfg.2019.02.114.

[18] Anders Haug, Frederik Zachariassen, and Dennis Liempd. The costs of poor data quality. *Journal of Industrial Engineering and Management*, 4, 07 2011. doi:10.3926/jiem..v4n2.p168-193.

[19] C. Ji. Reliability evaluation and prediction of mechanical system based on machine learning technology. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, pages 01–05, 2023. doi:10.1109/AIKIIE60097.2023.10390487.

[20] L. Jiang, D. Barone, A. Borgida, and J. Mylopoulos. Measuring and comparing effectiveness of data quality techniques. In Pascal van Eck, Jaap Gordijn, and Roel Wieringa, editors, *Advanced Information Systems Engineering*, pages 171–185, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[21] A. Juneja and N.N. Das. Big data quality framework: Pre-processing data in weather monitoring application. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 559–563, 2019. doi:10.1109/COMITCon.2019.8862267.

[22] M. Kindling and D. Strecker. Data quality assurance at research data repositories. *Data Science Journal*, 21, 2022. doi:10.5334/dsj-2022-018.

[23] I. Kirchen, D. Schütz, J. Folmer, and B. Vogel-Heuser. Metrics for the evaluation of data quality of signal data in industrial processes. In *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, pages 819–826, 2017. doi:10.1109/INDIN.2017.8104878.

[24] B. Kitchenham. Guidelines for performing systematic literature reviews in software engineering, 2007. URL: https://www.researchgate.net/publication/302924724.

[25] W. Kong, F. Qiao, and Q. Qidi. Real-manufacturing-oriented big data analysis and data value evaluation with domain knowledge. volume 35, pages 1–24, 06 2020. doi:10.1007/s00180-019-00919-6.

[26] X. Li. Cleansing and analytics of indoor positioning data. In *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, pages 334–336, 2022. doi:10.1109/MDM55031.2022.00075.

[27] Munawar. Extract transform loading (etl) based data quality for data warehouse development. In *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, volume 1, pages 373–378, 2021. doi:10.1109/ICCSAI53272.2021.9609770.

[28] L. Poon, S. Farshidi, N. Li, and Z. Zhao. Unsupervised anomaly detection in data quality control. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2327–2336, 2021. doi:10.1109/BigData52589.2021.9671672.

[29] P. Schlegel, D. Buschmann, M. Ellerich, and R.H. Schmitt. Methodological assessment of data suitability for defect prediction. *Quality Innovation Prosperity*, 24:170–185, 2020. doi:10.12776/QIP.V24I2.1443.

[30] D.H. Sitawati, Y. Ruldeviyani, A.N. Hidayanto, R.S Amanda, and A.G. Nugroho. Data quality improvement: Case study financial regulatory authority reporting. In *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, pages 272–277, 2022. doi:10.1109/ISMODE53584.2022.9743087.

[31] P.C. Soto, N Ramzy, F Ocker, and B Vogel-Heuser. An ontology-based approach for preprocessing in machine learning. In *2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES)*, pages 000133–000138, 2021. doi:10.1109/INES52918.2021.9512899.

[32] P. Sreenivas and C.V. Srikrishna. An analytical approach for data preprocessing. In *2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA)*, pages 1–12, 2013. doi:10.1109/C2SPCA.2013.6749435.

[33] I. Taleb, R. Dssouli, and M.A. Serhani. Big data pre-processing: A quality framework. In *2015 IEEE International Congress on Big Data*, pages 191–198, 2015. doi:10.1109/BigDataCongress.2015.35.

[34] W.L. Tsai and Y.C. Chan. Designing a framework for data quality validation of meteorological data system. *IEICE TRANSACTIONS on Information and Systems*, 29:583–591, 2019. doi:10.1016/j.promfg.2019.02.114.

[35] T. Wahyudi and S.M. Isa. Data quality assessment using tdqm framework: A case study of pt aid. *Journal of Theoretical and Applied Information Technology*, 15, 2023. URL: www.jatit.org.

[36] Richard Y. Wang and Diane M. Strong. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, mar 1996. doi:10.1080/07421222.1996.11518099.

[37] E. Widad, E. Saida, and Y. Gahi. Quality anomaly detection using predictive techniques: An extensive big data quality framework for reliable data analysis. *IEEE Access*, 11:103306–103318, 2023. doi:10.1109/ACCESS.2023.3317354.

[38] Roelf J. Wieringa. *Design science methodology for information systems and software engineering*. Springer, Germany, 2014. 10.1007/978-3-662-43839-8. doi:10.1007/978-3-662-43839-8.

[39] D. Xu, Z. Zhang, and J. Shi. A data quality assessment and control method in multiple products manufacturing process. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pages 1–5, 2022. doi:10.1109/DSIT55514.2022.9943883.

[40] T. Yuan, K.H. Adjallah, A. Sava, H. Wang, and L. Liu. Issues of intelligent data acquisition and quality for manufacturing decision-support in an industry 4.0 context. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 2, pages 1200–1205, 2021. doi:10.1109/IDAACS53288.2021.9660957.

[41] J. Zhang and R. Gao. Deep learning-driven data curation and model interpretation for smart manufacturing. *Chinese Journal of Mechanical Engineering*, 34, 12 2021. doi:10.1186/s10033-021-00587-y.

Appendix

TABLE 8.1: Qualitative Analysis of Literature

| Literature | Main Purpose | Proposed Future Research |
|---|---|---|
| E. Al-Masri and Y. Bai (2020) [2] | "In this paper, we introduce the Quality of Data for IoT Devices (QoDID) framework that attempts to address key challenges associated with the collection and processing of raw sensor data used in IoT systems or applications. Throughout the paper, we discuss the overall architecture, use cases and implementation details of QoDID. We further use our QoDID framework to provide insights on improving the data acquisition processes involving devices and sensors in IoT systems" | "For future work, we plan to extend our QoDID framework to include a ranking mechanism for enabling existing IoT applications to integrate QoD through a service- oriented approach." |
| T.M. Ali, T.Z.and Abdelaziz, A.M. Maatuk, and S.M. Elakeili (2019) [3] | "This paper presents a general framework for the implementation of data cleaning according to the scientific principles followed in the data warehouse field, where the framework offers guidelines that define and facilitate the implementation of the data cleaning process to the enterprises interested in the data warehouse field." | N/A |

| | | |
|---|---|---|
| P. Burggräf, M. Dannapfel, R. Förstmann, T. Adlon, and C. Fölling (2017) [5] | "Within the scope of this paper, a methodological framework has been developed, which comprises continuous data quality improvement management (CDQIM) interlinked with data quality-related process design (DQPD)." | "Regarding future developments, the theoretically developed and partly implemented DQPE framework should be further put into practice. Two aspects must be considered. On the one hand, the proposed DQPE framework and especially the CDQIM framework should be further implemented in different industrial contexts. Existing methods and techniques in organizations, which are aligned with DQPE requirements, must be promoted and extended to further processes and to the entire organization. Aspects of the methodology, which have not yet been implemented to a satisfactory extent, must be introduced. With regard to highly networked companies, the application of CDQIM elements should not only be enhanced locally, but rather disseminated throughout affiliated plants." |

| | | |
|---|---|---|
| A. Burkhardt, S. Berryman, A. Brio, S. Ferkau, G. Hubner, K. Lynch, S. Mittman, and K. Sonderer [6] | "This paper presents a fully automated test data quality measurement developed by the authors to facilitate analysis of manufacturing test operations, resulting in a single number used to compare manufacturing test data quality across programs and factories, and focusing effort cost-effectively. The automation enables program and factory users to see, understand, and improve their test data quality directly. Immediate improvements in test data quality speed manufacturing test operation analysis, reducing elapsed time and overall spend in test operations." | N/A |
| L. Cai and Y. Zhu (2015) [7] | "First, this paper summarizes reviews of data quality research. Second, this paper analyzes the data characteristics of the big data environment, presents quality challenges faced by big data, and formulates a hierarchical data quality framework from the perspective of data users." | N/A |
| Q. Chen, Y. Liu, S. Hou, F. Duan, and Z. Cai (2021) [8] | "The main work of this paper is as followes. (1) The data quality issues are discussed in the context of PHM. (2) The PHM framework is proposed for improving the reliability of equipment. (3) Several machine learning algorithms are introduced for state detection. (4) The proposed technology is applied to real cases, and the results are analyzed and visualized in detail" | N/A |

| | | |
|---|---|---|
| Y. Chernov (2016) [9] | "Therefore, our intention is to find a reasonable approach, suitable to the evaluation and helpful in finding the ways to improve the test-data. And the major question remains: how we can formally define the test-data quality." | N/A |
| N.B. Ding and E. Mit (2023) [16] | "This research aims to solve the challenges mentioned above through the framework designed. The framework is designed to integrate two algorithms into the data management lifecycle. Preventive algorithm meant to be applied in the earlier stage of the data management lifecycle while predictive algorithm is applied for the purpose of data cleansing. It is foresee that this framework is able to improve the data quality. Future work will be the implementation of the framework." | "Future work for this research is to develop the rules of the machine learning in order to execute the framework designed." |

| | | |
|---|---|---|
| C.L Günther, E. Colangelo, H.H Wiendahl, and C. Bauer (2019) [17] | "In this paper, we propose a methodology that simplifies the execution of DQ evaluations and improves the understandability of its results. One of its main concerns is to make DQ assessment usable to small and medium- sized enterprises (SME). The approach takes selected, context related structured or semi-structured data as input and uses a set of generic test criteria applicable to different tasks and domains. It combines data and domain driven aspects and can be partly executed automated and without context specific domain knowledge. The results of the assessment can be summarized into quality dimensions and used for benchmarking. The methodology is validated using data from the enterprise resource planning (ERP) and manufacturing execution system (MES) of a sheet metal manufacturer covering a year of time. The particular application aims at calculating logistic key performance indicators. Based on these conditions, data requirements are defined and the available data is evaluated considering domain specific characteristics" | "Methods to improve the traceability of DQ dimensions, e.g. by visualizations of results, should be developed. Future work might also address a guideline for the preparation and improvement phase and the application to different contexts and domains." |

| | | |
|---|---|---|
| C. Ji (2023) [19] | "This paper reviews the traditional mechanical reliability evaluation methods and points out their limitations in terms of data quality, data quantity and accuracy of analysis results. This paper proposes a mechanical system reliability assessment and prediction method based on machine learning technology, focusing on key links such as data preprocessing, feature engineering, model selection and model evaluation." | N/A |
| A. Juneja and N.N. Das (2019) [21] | "We propose addressing various aspects of the raw data to improve its quality in the pre-processing stage, as the raw data may not usable as-is. We are exploring process like Cleansing to fix as much data as feasible, Noise filters to remove bad data, as well sub-processes for Integration and Filtering along with Data Transformation/Normalization. We evaluate and profile the Big Data during acquisition stage, which is adapted to expectations to avoid cost overheads later while also improving and leading to accurate data analysis." | N/A |

| | | |
|---|---|---|
| I. Kirchen, D. Schütz, J. Folmer, and B. Vogel-Heuse (2017) [23] | "The main contribution of this paper is the development of a generic model for the objective evaluation of data quality of industrial signal data. This enables the supply of a basis for decision making in data mining processes concerning effective and efficient use of the available data." | "In future research the further introduced numerical indicator needs to be considered. Measurements for these indicators are required and an aggregation rule to combine all indicators is demanded. The resulting key indicator is qualified to evaluate data quality objectively in all required dimensions. Further- more, the validation of the DQM should be performed by real industrial signal data solely. Only this can prove the practical applicability unequivocally." |
| W. Kong, F. Qiao, and Q. Qidi (2020) [25] | "The main motivation of this paper is to explore methods for analyzing and evaluating big data with domain knowledge. For this purpose, real production data from a semiconductor manufacturing workshop are adopted as the data object. First, a series of data analysis techniques with domain knowledge are developed for diagnosing the imperfections. Then, corresponding data processing techniques with domain knowledge are proposed for solving those data quality problems according to specific flaws in the data. Furthermore, this paper proposes quantitative calculation methods of data value density to determine the extent to which data quality can be improved by the proposed data processing techniques." | "Moreover, the work in this paper has the potential to be further extended and applied to other big data applications beyond the manufacturing industry. Nevertheless, our work has several limitations: the data analysis and processing techniques with domain knowledge that we proposed are in their infancy and must be systematically organized to better suit other big data applications. In addition, the calculation methods of the rate of change of the data value density must be further discussed and modified for a more precise evaluation of data quality and data value." |

| | | |
|---|---|---|
| X. Li (2022) [26] | "In this paper, we first introduce the data quality issues consisting in indoor positioning data and propose a cleansing framework to handle such issues. Subsequently, we formulate four specific research questions in order to settle related quality issues. In addition, we present promising methodologies and comprehensive evaluation criteria to resolve our proposed research questions." | N/A |
| Munawar (2021) [27] | "In this paper, an ETL framework is proposed which incorporates data quality to improve information processes in data warehouse development through 'the story' of process whilst others framework more to technical approach. In order to be useful, the proposed framework compared with other framework in case of advantages and disadvantages for future improvement." | "Therefore, our future work is synthesizing our proposed ETL DQ-based framework with our previous DQ-based works in requirements analysis and conceptual design [35], logical design [36] and physical design [37] to be a single framework for the DW development as described in [38,39]" |

| | | |
|---|---|---|
| L. Poon, S. Farshidi, N. Li, and Z. Zhao (2021) [28] | "This study introduces an unsupervised anomaly detection approach based on models comparison, consensus learning, and a combination of rules of thumb with iterative hyper-parameter tuning to increase data quality." | "As the next course of action in this research, we plan to employ different anomaly detection models and evaluate their performance against each other. Automatic hyper-parameter tuning on other models could then be explored. In addition, determining outliers can be selected via <true, false> labels of the selected models. Some anomaly detection models can also output an anomaly score to determine the chance of a data point being an outlier. Finally, it is interesting to explore unsupervised ensemble learning further. It shows promising results with the consensus approach, and it can be an addition for anomaly detection on unlabelled data. A more balanced result via a "consensus" can give the data better assurance of labeling a data point as an outlier." |
| Schlegel, D. Buschmann, M. Ellerich, and R.H. Schmitt (2020) [29] [29] | This paper provides a domain specific concept to assess data suitability of various data sources along the production chain for defect prediction." | N/A |
| D.H. Sitawati, Y. Ruldeviyani, A.N. Hidayanto, R.S Amanda, and A.G. Nugroho (2021) [30] | "The objective of this study is to analyze the quality of critical data from the Integrated Report for the monthly period that is reported to the financial regulatory authorities based on the dimensions used to measure and propose recommendations to assist Financial Regulatory Authority in conducting assessments." | "Future work can also be continued by measuring the data quality management maturity periodically." |

| | | |
|---|---|---|
| P.C. Soto, N Ramzy, F Ocker, and B Vogel-Heuser (2021) [31] | "This paper presents a framework for semantic preprocessing, which is evaluated at the example of an industrial use case from the semiconductor industry" | "Based on the this research, we suggest three directions for future work. First, SPARQL is limited regarding arithmetic op- erations, which reflects in the identification of single construct outliers. However, the current approach can extract data which could then be processed using advanced arithmetic tools (e.g., numpy) to further improve the results. Second, various ML algorithms could be used in the comparison of the ontology- based approach and the ”baseline” one. Finally, the effect of data cleaning on highly optimized ML algorithms, i.e., with hyperparameter optimization, should also be investigated." |
| P. Sreenivas and C.V. Srikrishna (2013) [32] | "This paper revisits the preprocessing technique of Data Mining. A sequential flow diagram is proposed for different databases and data sources which are addressed through analysed framework. Through a case study and calculating cyclomatic complexity of different sequences of preprocessing the appropriateness and efficiency of proposed method is evaluated. It has been observed that right selection of an appropriate sequence in cleaning improves the data mining process by saving time taken for each step." | N/A |

| | | |
|---|---|---|
| I. Taleb, R. Dssouli, and M.A. Serhani (2015) [33] | "This paper addresses the QBD at the pre-processing phase, which includes sub-processes like cleansing, integration, filtering, and normalization. We propose a QBD model incorporating processes to support Data quality profile selection and adaptation." | "Further implementations will tackle the data pre-processing evaluation and dynamic assessment of data quality rules. Quality adaptation will be possible through the DQP adapter taking the appropriate actions to adjust and reevaluate the quality profile. Yet, big data sampling is very important to maximize data quality, and minimize the processing time." |
| W.L. Tsai and Y.C. Chan (2019) [34] | "This paper proposes the Total Meteorological Data Quality (TMDQ) framework based on the Total Quality Management (TQM) perspective, especially considering the systematic nature of data warehousing and process focus needs. In practical applications, this paper uses the proposed framework as the basis for the development of a system to help meteorological observers improve and maintain the quality of meteorological data in a timely and efficient manner." | "Therefore, in the future, the four quality dimensions of the TMDQ framework will be further developed for application with different meteorological elements, thereby enabling comprehensive meteorological data validation." |
| T. Wahyudi and S.M. Isa (2023) [35] | "The approach taken in this study to evaluate the quality of the data are Total Data Quality Management (TDQM) and the six dimensions from the DAMA white paper. The results of this evaluation procedure can be used to examine the company's existing data quality and to provide recommendations for changes that need be made internally." | N/A |

| | | |
|---|---|---|
| E. Widad, E. Saida, and Y. Gahi (2023) [37] | "To achieve this, we suggest a novel approach that allows a comprehensive detection of Big Data quality anomalies related to six quality dimensions: Accuracy, Consistency, Completeness, Conformity, Uniqueness, and Readability. Moreover, the framework allows for sophisticated detection of generic data quality anomalies through the implementation of an intelligent anomaly detection model without any correlation to a specific field. Furthermore, we introduce and measure a new metric called "Quality Anomaly Score," which refers to the degree of anomalousness of the quality anomalies of each quality dimension and the entire dataset" | "In future work, we aim to extend the suggested framework by addressing the detected anomalies and automatically correcting them to the appropriate data value rather than simply removing them, which will improve the dataset's quality. Automatically correcting anomalies will improve the dataset's overall quality and guarantee reliable and comprehensive information for subsequent analyses and decision-making. Moreover, it will save valuable time and effort that would otherwise be required for manual inspection and correction of anomalies." |
| D. Xu, Z. Zhang, and J. Shi (2022) [39] | "n this study, a data assessment matrix for imbalanced multivariate time series data from complex manufacturing process is designed to measure the data quality quantitatively." | "here are many potential future directions of this topic such as define more precise data quality assessment dimensions for single or multiple signals in manufacturing, define standard data quality assessment metrics for data collected from one production line, design data quality improvement cycle." |
| T. Yuan, K.H. Adjallah, A. Sava, H. Wang, and L. Liu (2021) [40] | "In this paper, we will present the key issues of intelligent data acquisition and data quality for manufacturing decision-support in industry 4.0." | "Further investigation results will be provided in a more detailed work considering data uncertainty and error propagation to decision-making risks assessment." |

| | | |
|---|---|---|
| J. Zhang and R. Gao (2021) [41] | "his paper summarizes several key techniques in data curation where breakthroughs in data denoising, outlier detection, impu- tation, balancing, and semantic annotation have demonstrated the effectiveness in information extraction from noisy, incomplete, insufficient, and/or unannotated data. Also highlighted are model interpretation methods that address the “black-box” nature of DL towards model transparency." | "As research on DL-enabled manufacturing continues to accelerate, several topics that closely relate to data curation and model interpretation are summarized here, as recommendations for future study: Uncertainty quantification. Physics-informed learning. Mitigating false discovery." |