Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2022, 9(3):172-178

Review Article ISSN: 2394-658X

A Review and Analsis on Data Cleaning and Quality Assurance

Pranay Mungara

ABSTRACT

The exponential growth of research, technology, and engineering over the last decade has resulted in massive amounts of data being produced across many different industries. The act of doing medical research results in the generation of data on a continuous basis, and the accumulation of substantial amounts of data from the real world becomes a "data disaster." The availability of data and the quality of the data are the foundations upon which effective data mining and analysis are built. Cleaning the data is an important step in achieving high data quality. You must meet this requirement. Finding and fixing "dirty data" is what data cleaning is all about. This procedure serves as the foundation for data management and analysis. Furthermore, data cleaning is a type of technology that is frequently used to improve the quality of data. There is, however, a lack of guidance in the existing literature on real-world research regarding how to set up and carry out data cleansing in a manner that is both efficient and ethical. To tackle this problem, we put up a data cleansing strategy for actual research. Duplicate, missing, and outlier data are the three main types of dirty data that the methodology aims to address. Furthermore, we suggested a routine data cleaning technique that can be used as a guide for future research that incorporates similar technology.

Key words: Data cleaning, Quality Assurance, Data management

INTRODUCTION

The measurement of data quality (DQ) is a crucial part of determining the relevance of data-driven decisions. Like the machine-based decisions employed by ranking algorithms, industrial robots, and self-driving cars—all instances of the expanding domain of artificial intelligence such decisions are present in our daily lives. Machine learning (ML) models' mistake rates are significantly impacted by insufficient data, as shown by [1]. Furthermore, high-quality data is essential for human-based decisions. For example, sales statistics are usually considered while deciding whether to promote or stop making a given product.

In the US, 84% of CEOs worry about the quality of their decisions, and "organisations believe poor data quality to be responsible for an average of fifteen million dollars per year in losses" [2]. Despite the obvious correlation between data and sound decision-making, this remains the case. In light of this shift, data quality is now "no longer a question of 'hygiene,' but rather has become critical for operational excellence," and addressing this issue poses the greatest challenge to corporate data management professionals.

To ensure the dependability of decision-making, the author put out five criteria for DQ metrics in [3]. These requirements are as follows: "the existence of minimum and maximum metric values (R1), the interval scaling of the metric values (R2), the quality of the configuration parameters and the determination of the metric values (R3), the sound aggregation of the metric values (R4), and the economic efficiency of the metric (R5)." On the other hand, there are academics that assert "that a more general approach is required" in order to evaluate the benefits and reliability of a DQ metric. This section will provide an overview of four main dimensions of DQ, along with popular metrics that are used to calculate them. Though the list of measures is not full, it should provide some insight into the research that has been carried out in this field. This is because our DQ tool review reveals the presence of metrics that are comparable to those that are included here. When it comes to academic and business settings, the significance of the function that vast amounts of real-world data play is growing [4]. In order to learn more about the worth of this data, we need data mines and analyses.

Real-world data from medical practice, on the other hand, are typically generated without the stringent procedural controls that are utilised in clinical laboratory studies. Some examples of quality difficulties include duplicate, missing, and outlier data. Integrating data from several sources, such as several hospitals or hospital systems, could necessitate different rules.

This can lead to problems with the original data. Everywhere you look in RWS, you'll find "dirty data." One of these is that the database's efficiency is negatively impacted by the high quantity of store space that is required for duplicate data. Another is that a lot of potentially useful information can be lost due to incorrect processing of missing data. In addition, data analytics and critical computations might be severely affected by inconsistent or inaccurate outlier data.

These outcomes might even lead to inappropriate paths for later academic study, which can result in a total loss of time, effort, and financial resources [5].

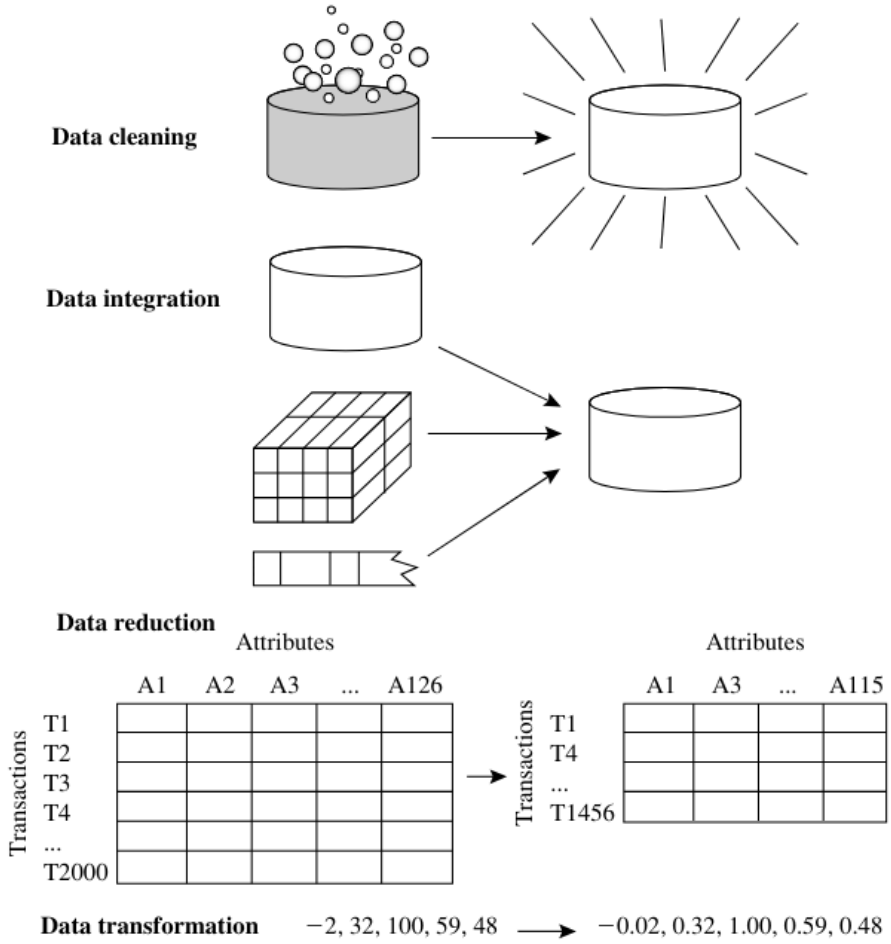

Figure 1: Forms of data preprocessing

The methods of data pretreatment that are covered in this study are summarised in Figure 1. It is important to note that the previous classification should not be considered mutually exclusive. An example of data cleaning and data reduction would be the elimination of superfluous data. This might be considered a sort of data cleaning. In a nutshell, data collected from the real world are typically unclean, lacking in consistency, and incomplete. The accuracy and efficiency of the subsequent mining process can be improved by the use of data pretreatment techniques, which can assist improve the quality of the data. Preprocessing data is a crucial part of knowledge discovery because it ensures that high-quality decisions are based on accurate information.

The detection of data abnormalities, the early correction of such anomalies, and the reduction of the amount of data that needs to be analysed can all lead to significant payoffs for decision making.

LITERATURE REVIEW

A. Categorizing Issues with Data Quality

The degree to which the data in terms of its accuracy, completeness, consistency, and timeliness are able to fulfil the requirements that are anticipated by certain users is referred to as the data's quality. The pattern-layer and instance-layer difficulties are two ways to categorise issues with data quality, depending on the discovery level. Problems can also be categorised as having one or many sources of data, depending on the data that is being used. Because of this, issues with data quality are typically grouped into four types: single-source pattern layer, multi-source pattern layer, single-source instance layer, and multi-source instance layer (Figure 2).

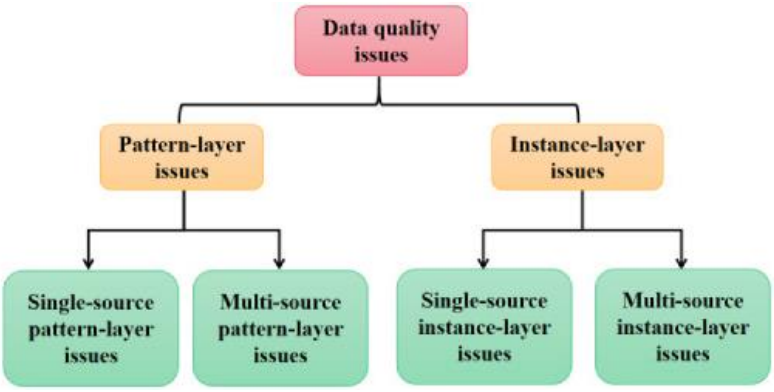

Figure 2: Data quality issues classification

B. Causes of Data Quality Issues

System design errors are the root cause of pattern layer problems. Lack of integrity requirements and simplistic architectural design are pattern-layer issues with respect to single data sources. Conflicts in structure and nomenclature across sources are another possible pattern layer difficulty when multiple sources of data are involved. Data governance for RWS primarily focuses on issues in the instance layer that are caused by unhandled faults in the pattern layer, rather than the pattern layer itself [6].

One of RWS's main goals in data governance is reducing the prevalence of human mistake at the instance layer, where the vast majority of data problems originate. At the single-source instance layer, data record exceptions often occur due to missing values, similar or duplicate records, or input errors. Case recording is a common time for input mistakes in data sources that depend heavily on human intervention, such as patient health records and information retrieved from mobile health monitoring devices. There is a high incidence of input errors in these databases. People often make mistakes when entering data, which can lead to duplicate or identical records.

On the other hand, they may also occur when two cases of the same patient are stored over the same time period, each of which has a different level of completeness. This second instance happens frequently when exporting data for multiple time periods at once, for example, from January to June and then from June to December consecutively. Inaccuracies in the recording method or the patient's deliberate concealment (such as a refusal to reveal relevant information) could lead to missing values.

There is also the possibility that the absence of values is due to errors in data storage or error clearance that are the consequence of problems with the equipment. Data that is extremely sensitive may also be difficult to access in certain circumstances (for example, payment information for medical insurance) [7].

At the instance layer, problems might arise not only for single sources but also for multisource configurations, on top of all the problems that can arise for single sources. Data aggregation and uneven data time are two of these issues. Records that are identical to one another or duplicates that are the result of recognising the same material as distinct objects (that is, using different expressions) are the most commonly encountered issues among these.

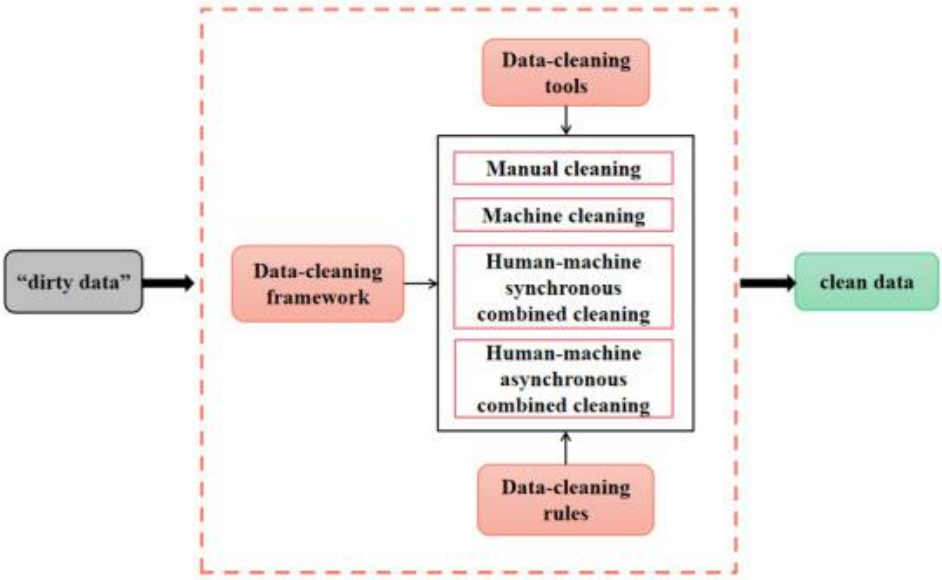

Figure 3: Basic process of data cleaning

Even though it's quite accurate, manual cleaning is only suitable for smaller data sets because of how long it takes. Machine cleaning, on the other hand, is better appropriate for processing larger data sets due to the fact that the procedure is entirely automated. The cleaning plan and programme, on the other hand, must still be prepared in advance, which makes maintenance in later stages problematic. Figure 3 illustrates the fundamental process of data cleaning. In the synchronous human-machine technique, problems that the machine is unable to solve are manually addressed through an appropriate interface through the use of human intervention. This approach is useful because it lessens the amount of effort that is required for manual cleaning while simultaneously lowering the level of difficulty involved in developing a strategy for machine cleaning. Asynchronous human-machine strategies are conceptually similar to their synchronous counterparts. On the other hand, issues that the machine is unable to resolve are not treated immediately.

This is contrary to the synchronous method. Instead, an issue report is generated, and the cleaning process moves on to the subsequent stage. Therefore, manual processing takes place after cleaning, and it is the way that the majority of cleaning software employs [8].

ENHANCING DATA QUALITY FOR OPTIMAL DATA ANALYTICS PERFORMANCE

Many scholars have spent a lot of time studying data quality, specifically how it affects things like timeliness, reliability, accuracy, completeness, and consistency. The data quality can be assessed using these dimensions as benchmarks. These studies provide the framework for future study on the different ways to improve data quality management by highlighting the significance of addressing each factor to ensure the data's dependability and integrity.

According to [9], the approaches of data cleansing and profiling play a crucial role in locating and correcting problems that are present in several datasets. The importance of automated cleansing algorithms that make use of machine learning to identify abnormalities and inconsistencies is brought to light. With the use of sophisticated profiling tools, businesses are able to gain an understanding of the structure and quality of their data, which lays the framework for changes in data quality that are specifically targeted. According to [10], machine learning algorithms, and more specifically predictive analytics models, play a significant part in the process of forecasting data flaws and inconsistencies. Both the data quality and the ability to make proactive judgements are enhanced by these predictive models.

Real-time data processing has been made easier by the introduction of Big Data technologies, which are depicted in figure 4. As a result, concurrent real-time data quality management solutions have become necessary. Scholars place a strong emphasis on the creation of scalable frameworks that are capable of maintaining data quality in real-time. This enables organisations to adapt quickly to emerging data difficulties and to preserve the quality of streaming data.

Figure 4: Big data analytics

The nature of the complex link that exists between data analytics and data quality is investigated in this study. The purpose of this study is to research the methodology, tools, and best practices that are utilised in the field of data analytics in order to improve the quality of the data. The purpose of this study is to provide insights into how organisations and institutions may leverage the power of data analytics to secure the integrity and reliability of their data assets. This will be accomplished by evaluating real-world case studies and industry applications.

Figure 5: Data quality dimensions

As stated in [11] here, we highlight various techniques, including data profiling, data cleansing, and standardisation, and explain their roles in discovering and correcting data inconsistencies. The critical relationship between data quality and the efficacy of data analytics methodologies is illuminated by real-world case studies. These examples show how data quality projects can pay off in the real world. Better decision-making, higher customer satisfaction, and simplified operational processes are some of these advantages. The consequences of inaccurate data, such as flawed corporate plans and inaccurate predictive models, are also explored in the study. The abstract places an emphasis on the financial and reputational problems that are connected with mediocre data. It argues for a proactive strategy, in which organisations engage in solid data governance structures, innovative tools, and qualified individuals in order to guarantee that the quality of their data remains consistent. It is possible for organisations to harness the actual power of data analytics by doing so, which will ultimately lead to the achievement of sustainable growth, the promotion of innovation, and the encouragement of competitive advantage. In addition, the dimensions of the data quality were displayed in figure 5.

OVERVIEW ON DATA CLEANING ON REAL WORLD DATA

A database's "data cleaning" is the process of correcting or eliminating inaccurate, incomplete, or misleading information. One example of a data-intensive profession is a document that has been generated or reproduced inaccurately. Other examples include a corporation that operates in the banking, insurance, and commerce industries. Within the realms of transportation and telecommunications, a data set might be utilised. Using a cleaning algorithm, you may perform a methodical audit of your data to ensure that it is free of errors. Data cleaners that can detect and fix a broad range of specific mistakes are a prominent component of data cleaning tools. An algorithm, a set of rules, and a table search are used to do this. Duplicate records seem to be present. You can save both time and money by using an algorithm. Database administration demands a lot of energy and time, but the reward is substantial.

Compared to manually fixing errors, this method is less cost-effective. Cleaning the data is an important operation that must be completed by professionals working in data warehouses, database managers, and developers alike. Data warehouse component populating, recent data integration, and operational system maintenance of real-time dedupe processes are just a few examples of the many possible applications of deduplication, substantiation, and householding. The goal is to achieve the highest possible level of precision and dependability in the data, which will ultimately result in enhanced customer service, reduced expenses, and increased peace of mind [12]

A. **Using Measure of central tendency for each class:** This data cleansing procedure is an additional strategy that is comparable to the one that was discussed earlier inside this paragraph. On the other hand, it is somewhat different from the first one since in the second one, we ensured that the same amount of data was filled in for each class. When compared to the central tendency method, this one allows us more

leeway to guess the data (mode, median, or mean) that goes with each category [13]. This is the difference between the two methods.

```
X_train = train.drop(columns = 'SalePrice')
Y_train = train['SalePrice']
print("shape of X_train df = ",X_train.shape)
print("shape of y_train df = ",Y_train.shape)
```

```
shape of X_train df =  (1460, 80)
shape of y_train df =  (1460,)
```

Figure 6: Assigning dependent and independent variables

Also, keep in mind that this method requires an in-depth familiarity with the dataset's domain before it can be applied. If you want absolutely pinpoint accuracy, this is a must. For instance, thinking about the dataset that came before it. We follow the same process as previously when it comes to this plan. After loading the dataset and incorporating the libraries, we need to identify which columns have a null value percentage more than 20% and exclude them.

Then, from the remaining columns, check to see if any of them include numeric values, and then check once more to see what proportion of those columns contain null values; the results of this check are displayed in figure 6. Working with this library is a little less difficult than working with Python's programming language. Here the variables that are dependent and variables that are independent are characteristic of a data frame. With this approach, we separate the two components. In contrast, one variable is used for the independent characteristics and another for the dependent feature.

CONCLUSION

The goals of the data cleaning technology, a few key features of the technology, and approaches to developing effective solutions were covered in this review study. For instance, data cleaning technologies in common business applications aim to keep data quality and consistency at the same level as what is required by the data warehouse. This includes databases that store information about customers and products. With enough time and effort put into research and analysis, one can draw many important conclusions. Improving the data's quality is more than just a process; it's a strategic necessity. Systematic detection of data problems, full cleansing, integration, and transformation are at the most fundamental level. The basis of dependable and high-quality data is comprised of these tactics, which are backed by modern analytics. The ability to make judgements based on accurate information is facilitated by improved data quality. The route that their businesses will take is determined by decision-makers, who rely on data that is accurate and consistent. Businesses are now able to trust the data that underpins their crucial decisions, which directly leads to more reliable strategies and outcomes. This is made possible by effective data analytics.

REFERENCES

[1]. Ehrlinger L., Haunschmid V., Palazzini D., Lettner C. (2019). A DaQL to monitor the quality of machine data, in Proceedings of the International Conference on Database and Expert Systems Applications (DEXA), volume 11706 of Lecture Notes in Computer Science. (Cham: Springer;), 227–237.

[2]. Moore S. (2018). How to Create a Business Case for Data Quality Improvement. Available Online at: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement (January 2022).

[3]. Schalk E, Hentrich M. Real-world data. Dtsch Arztebl Int. 2022;119(8):134. doi: 10.3238/arztebl.m2022.0035. https://europepmc.org/abstract/MED/35506295 .arztebl.m2022.0035

[4]. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. JAMA. 2018;320(9):867–868. doi: 10.1001/jama.2018.10136.2697359

[5]. Bogani R, Theodorou A, Arnaboldi L, Wortham RH. Garbage in, toxic data out: a proposal for ethical artificial intelligence sustainability impact statements. AI Ethics. 2022:1–8. doi: 10.1007/s43681-022-00221-0. https://europepmc.org/abstract/MED/36281314 .221

[6]. Jahan M, Hasan M. A robust fuzzy approach for gene expression data clustering. Soft Comput. 2021;25(23):14583–14596. doi: 10.1007/s00500-021-06397-7

[7]. Berger B, Waterman MS, Yu YW. Levenshtein distance, sequence comparison and biological database search. IEEE Trans Inf Theory. 2021;67(6):3287–3294. doi: 10.1109/tit.2020.2996543.
[8]. Smith, J., Johnson, A., & Brown, K. (2018). Data Quality Dimensions: A Review of the State of the Art, 2018.
[9]. Chen, S., Liu, Y., & Wang, H. (2020). Enhancing Data Quality through Predictive Analytics: A Machine Learning Approach, 2020.
[10]. Li, X., Zhou, H., & Ma, L. (2021). Real-time Data Quality Management in the Era of Big Data: Challenges and Solutions, 2021.
[11]. Ilyas, I. F., & Chu, X. (2019). Data cleaning. Morgan & Claypool.
[12]. Jordanov, Ivan, Nedyalko Petrov, and Alessio Petrozziello. "Classifiers accuracy improvement based on missing data imputation." Journal of Artificial Intelligence and Soft Computing Research 8 (2018).
[13]. Zhang, Robert F., and Ryan J. Urbanowicz. "A scikitlearn compatible learning classifier system." In Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion, pp. 1816-1823. 2020.