Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/GC03.20250536
Published: February 2, 2026

RawtoReady: A Web-Based Automated Tool for Data Cleaning and Preparation

Gio Miguel R. Bihasa, Pam Cassedy T. Dumangeng, Kim Caryl H. Esperanza, Louella Josephine A. Ng and Richard N. Monreal
School of Information Technology
Mapua University Makati, Philippines
gmrbihasa@mymail.mapua.edu.ph, pctdumangeng@mymail.mapua.edu.ph,
kchesperanza@mymail.mapua.edu.ph, ljang@mymail.mapua.edu.ph, rmonreal@mapua.edu.ph

Abstract

Data cleaning is often a time-consuming and complex task that requires technical skills or expensive software, creating barriers for students, researchers, and professionals who need reliable datasets for analysis. This study developed *RawtoReady*, a web-based automated data cleaning system that provides one-click processing features such as handling missing values, removing duplicates, standardizing text and dates, validating emails, applying fuzzy standardization, and detecting anomalies through a guided four-step workflow of uploading data, selecting cleaning options, running the process, and downloading the cleaned output with a transparent summary. The system was evaluated using a user survey designed based on Edward de Bono's Six Thinking Hats methodology, with responses collected from 56 participants to assess usability, effectiveness, and user satisfaction. Results showed that users found RawtoReady efficient, accessible, and easy to use, significantly reducing the time and effort required for manual data cleaning while improving confidence, organization, and control in data preparation. These findings indicate that RawtoReady is a practical and effective solution for democratizing data cleaning, supporting data-driven learning, research, and innovation in alignment with Sustainable Development Goals 4 (Quality Education) and 9 (Industry, Innovation, and Infrastructure).

Keywords

Data cleaning, automated data preprocessing, web-based system, data quality, dataset preparation

1. Introduction

In the data-driven world of today, datasets are crucial for researchers, corporations, and students to make decisions, create models, and provide insights. Raw data, however, is rarely available for use right away. Analysis is severely hampered by common errors such as outliers, duplicates, missing information, and inconsistent formatting. If these "dirty data" issues are not resolved, they may impair accuracy, make data less interpretable, and result in findings that are not trustworthy (Rahm and Do, 2000). As a result, one of the most important and time-consuming steps in the data preparation pipeline is data cleaning. Research continuously shows that between 40 and 60 percent of analysts' time is spent on cleaning, frequently at the expense of more in-depth analysis and interpretation (Chiriyankandath et al., 2020; Martins et al., 2025).

Even while data quality is crucial, many of the solutions available are either too expensive for regular users or require highly skilled technical knowledge. While enterprise technologies like Alteryx and Trifacta offer automation but are too costly for students and small businesses, well-known libraries like Pandas offer flexibility but require programming knowledge. OpenRefine still requires manual setting and has trouble with really large datasets, despite being open-source and helpful for exploratory cleaning. Because of this, novice users, students, and practitioners who

are not familiar with coding encounter many difficulties when preparing their data. This leaves a need for automated, lightweight solutions that combine dependable preprocessing with user-friendliness.

This project, Raw to Ready, directly responds to these challenges by providing a simple, accessible web-based system for one-click data cleaning. Users can upload CSV files, choose from automated cleaning options such as handling missing values, removing duplicates, and normalizing text, and then download a cleaned dataset along with a transparent summary of changes. The system follows a guided, four-step process: uploading the dataset, selecting cleaning options, running the automated cleaning process, and downloading the cleaned output (Figure 1).

Figure 1. Four-step Process of RawtoReady

RawtoReady features several cleaning capabilities, including missing value handling, duplicate removal, column name standardization, text normalization, email validation, fuzzy standardization of categorical values, and anomaly detection. It also provides a real-time data preview feature, allowing users to compare raw and cleaned data side by side. The interface is designed to be user-friendly through drag-and-drop uploads, simple check boxes, dropdown menus, and one-click automation.

1.1 Research Objectives

The main objective of the project is to develop a data-cleaning system that enables users to efficiently process and prepare raw datasets for analysis by providing automated cleaning options, anomaly detection, and easy data export. Specifically, it intends to do the following: (1) Provide essential and advanced data-cleaning functions that improve dataset quality and consistency; (2) Design a user-friendly interface that allows users to easily upload and interact with their datasets; (3) Generate outputs that help users understand the cleaning proves; (4) Enable users to download the cleaned dataset in a format ready for analysis; and (5) Ensure the system operates efficiently, accurately, and reliably while maintaining the secure handling of user data.

2. Literature Review

2.1 The Data Cleaning Problem

The foundational premise of all data analysis is that data quality directly and significantly impacts the quality of results. However, a vast body of literature identifies a critical bottleneck in this process: real-world data is inherently "messy," "dirty," and complex (Buddiga 2020; Chiriyankandath 2020) A systematic literature review by Côté et al. (2023) confirms that the field of data cleaning is itself vast and fragmented, highlighting the difficulty of establishing a single, coherent solution. Hosseinzandeh et al. (2021) likewise report that detecting and repairing “dirty” data remains one of the most time-consuming challenges in data management. Incomplete, inaccurate, and inconsistent records continue to affect the reliability of analytics, showing that the data quality problem persists even with modern tools. The problem is not trivial; it involves a "normal workflow" of multiple, distinct stages, including handling duplicates, missing values, and outliers (Guo et al. 2023).

This multi-stage process is complicated by its deep, historical dependency on manual human intervention. Ridzuan and Zainon (2019), in their review of big data methods, note that the data cleansing process is inherently "complex

and time-consuming." They, along with Guo et al. (2023), emphasize the "importance of domain expert" for validation, which creates a severe bottleneck, as expert time is both expensive and scarce. Valêncio et al. (2020) state this more forcefully, noting that most methods "demand heavy user interaction," a model that is simply "infeasible" for modern large-scale databases.

2.2 Traditional and Manual Data Cleaning Process

Traditional approaches to data cleaning, which are either fully manual or rely on simple, rigid rule-based systems, are demonstrably insufficient for the scale and complexity of modern data. Panwar (2024), in a comparative analysis, states that these "traditional data cleansing methods... often fall short," struggling to adapt to the volume and variety of contemporary data environments.

The limitations of these methods are not just theoretical but have been clearly quantified. A usability study by Hameed et al. (2025) on their Tasheeh system provides a stark comparison: manual data cleaning took participants an average of one hour and seven minutes to achieve only 83% accuracy. Supporting this, Abdelaal et al. (2023) evaluated a broad range of traditional and modern data cleaning methods within their REIN benchmark framework and found that rule- and constraint-based systems typically achieved an F1 accuracy of only around 60-70%, whereas learning-based and ensemble approaches often exceeded 85-99% accuracy. They further observed that traditional methods rely heavily on human-defined rules and constraints that make them rigid and unable to efficiently handle new or evolving data errors. This highlights that manual methods are not only slow but also "error-prone," a sentiment echoed by Oladipupo et al. (2023). This "time-consuming and error-prone manual preprocessing" establishes the clear and urgent need for automation to achieve gains in both efficiency and reliability.

2.3 Automated and ML-Based Data Cleaning Approaches

In response to the failures of manual methods, the research community has developed a wide spectrum of automated solutions. These approaches can be grouped by their increasing level of intelligence and autonomy.

The most foundational level of automation involves automating the core, known workflow. Oladipupo et al. (2023) exemplify this by designing an automated Python script to handle standard tasks like imputation via the Interquartile Range (IQR) method and feature normalization. Similarly, Yasin and Khorsheed (2025) present an automated method for large databases that also focuses on imputation, outlier management, and normalization. Chiriyankandath et al. (2020) developed a more comprehensive Automated Data Cleaning (ADC) framework to holistically manage profiling, duplicate detection, and various imputation methods, successfully demonstrating that a well-designed framework can achieve "nearly the same quality as those cleaned manually" with a "significant reduction in preprocessing time."

A more advanced category of automation involves domain-specific and heuristic-based systems. These tools move beyond generic statistics to incorporate specialized knowledge. For example, Shi et al. (2021), working with complex Electronic Health Records (EHRs), integrated a "Clinical Knowledge Database (CKD)" to automate expert decisions—such as using fuzzy search on medical units or validating values against biological plausibility—a task that simple statistical outlier detection would fail. Likewise, Hameed et al. (2025) developed Tasheeh to specifically automate the repair of structural CSV errors, while Valêncio et al. (2020) built a configurable environment that uses "physical-semantic data similarity" and multilingual support to achieve 90% cleaning efficiency.

The current "paradigm shift," as described by Panwar (2024), is the use of AI and Machine Learning (ML) to drive the cleaning process. This approach uses AI models to "learn complex patterns and anomalies beyond predefined rules," resulting in superior accuracy and efficiency. Kilani et al. (2025) concur, noting that ML methods provide adaptability and Deep Learning (DL) models can achieve the highest precision, such as 94.7%. This is demonstrated in practice by Jäger and Biessmann (2024), whose Conformal Data Cleaning (CDC) system uses ML to detect incorrect cells that violate learned patterns, not just fill in blank ones. However, this power comes with a critical caveat. A landmark study by Pradhan et al. (2024) provides a stark warning that "Automated Data Cleaning can Hurt Fairness," as naive algorithms may disproportionately remove data from minority groups, thereby embedding bias into the final model.

Finally, the most advanced research aims for fully autonomous pipeline generation. These systems automate the planning of the cleaning process itself. Krishnan and Wu (2019) developed AlphaClean, which treats cleaning as a "search problem" to find the optimal sequence of cleaning operations. This concept is shared by Nargesian et al. (2021) with their Auto-Clean framework, which programmatically selects the best operators for a given dataset. The latest evolution of this concept is AutoDCWorkflow by Li et al. (2024), which leverages Large Language Models (LLMs) to inspect a dataset and auto-generate a full cleaning workflow from scratch.

2.4 Research Gap and Motivation for the Proposed System

This comprehensive review demonstrates that automated data cleaning solutions are demonstrably faster (Panwar 2024; Hameed et al. 2025) more accurate , and more powerful (Krishnan and Wu 2019) than traditional manual methods. However, it also reveals a significant and persistent "accessibility gap" that motivates the RawtoReady project.

The current landscape of tools, as benchmarked by Martins et al. (2025), is highly fragmented. Users are forced to choose between different tools (OpenRefine, Dedupe, etc.) that all have specific strengths, weaknesses, and usability trade-offs. This is reinforced by Kilani et al. (2025), who note that user-friendly tools are often less scalable, while scalable tools are not user-friendly.

This creates a clear divide:

- Code-Based Tools: Many powerful solutions require technical expertise in Python (Oladipupo et al. 2023) or are complex, research-oriented frameworks (Nargesian et al. 2021; Li et al. 2024)
- Complex Configuration: Advanced systems can be computationally expensive (Kilani et al. 2025) or require careful configuration by an expert to avoid unintended consequences, such as the fairness violations described by Pradhan et al. (2024).

This leaves students, researchers, and non-technical professionals—the target users of this project—without a viable option. The existing literature lacks a tool that is simultaneously simple, powerful, and accessible. The motivation for RawtoReady is to democratize data preparation by synthesizing the proven, core automation techniques from the literature (Guo et al. 2023; Chiriyankandath et al. 2020) into a "one-click" web application. This project fills the identified gap by prioritizing accessibility and efficiency, providing a practical tool that empowers non-expert users to quickly and reliably clean their data.

3. Methods

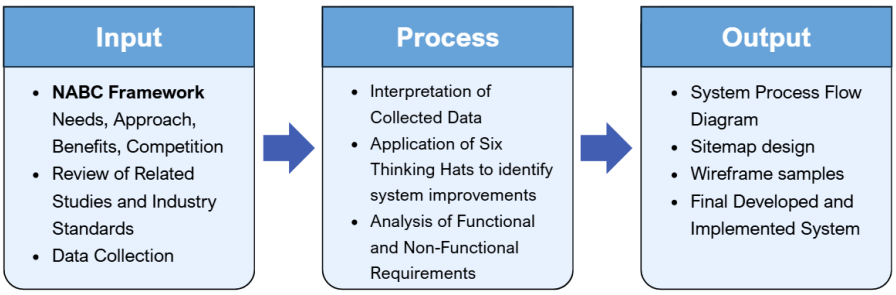

Figure 2. IPO Chart

Figure 2 illustrates the methodology of RawtoReady using the IPO model. In the input stage, the NABC framework is applied to identify the system's needs, define its approach, determine its benefits, and analyze existing competition. This stage also includes the review of related studies and industry standards to ensure alignment with best practices, as well as data collection through user surveys to gather requirements and insights for system development.

In the process stage, the collected data are interpreted and evaluated. The Six Thinking Hats method is used to examine the system from different perspectives, allowing the identification of strengths, weaknesses, risks, and possible

improvements. At the same time, functional and non-functional requirements are analyzed to clearly define the system's features and performance expectations.

In the output stage, the results of the analysis are transformed into concrete system outputs, including the system process flow diagram, sitemap design, and wireframe samples, which serve as the basis for developing and implementing the final RawtoReady system.

4. Data Collection

To gather comprehensive insights regarding RawtoReady, the group applied Edward de Bono's Six Thinking Hats framework in creating the survey questionnaire. As De Bono (1985) discussed, each hat represents a different thinking perspective, and hence guided the formation of the survey questions. Specifically, the survey questions involve a combination of Likert scale and open-ended questions per hat.

The White Hat approach involves neutrality and objectivity, focusing on information, data, and facts. The Red Hat approach focuses on emotions, aiming to capture the respondents' feelings about RawtoReady. The Black Hat approach identifies weaknesses, threats, and cautions, highlighting potential limitations and disadvantages of RawtoReady. The Yellow Hat approach covers strengths and benefits, identifying positive perception and feedback of RawtoReady. The Green Hat approach recognizes opportunities and recommendations, emphasizing ideas for improvement. Lastly, the Blue Hat approach centers on overall reflection towards RawtoReady. Notably, the use of Six Thinking Hats fosters critical thinking by encouraging the exploration of diverse perspectives and the comprehensive analysis of situations and solutions, thereby improving problem-solving skills (Wongsa and Cojorn 2024).

5. Results and Discussion

5.1 Respondents

A total of 56 participants responded to the survey using convenience sampling, a method in which respondents are selected based on their easy accessibility and availability to the researchers (Rahi 2017). Among them, 39 (69.6%) were students or beginners, 13 (23.2%) were researchers, and 4 (7.1%) were data professionals. This distribution reflects a mix of primarily academic users and a smaller group of more experienced data professionals, which aligns with the target users of the proposed system (Figure 3)

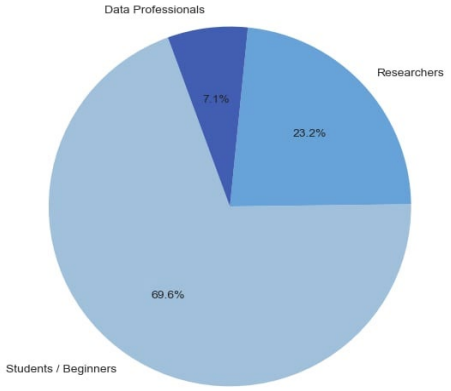

Figure 3. Respondents Distribution

5.2 Survey Results

5.2.1 Information and Data (White Hat)

Table 1. Descriptive Statistics on Information and Data

Questions	Mean	Verbal Description
RawtoReady provided enough information about the changes made during cleaning.	4.45	Strongly Agree
I could easily track what steps were applied to my dataset.	4.25	Agree
The cleaning summary was clear and easy to understand.	4.48	Strongly Agree

Respondents strongly agreed that RawtoReady provided sufficient information regarding the changes made during the data cleaning process, with a weighted mean of 4.45. Most participants also strongly agreed that they could easily track the steps applied to their dataset during cleaning (M=4.25) (Table 1), and that the cleaning summary was clear and easy to understand (M=4.28). Open-ended responses revealed that the majority of users did not require additional information after cleaning, frequently replying "None" or "N/A," indicating high satisfaction with the tool's current reporting. However, some participants suggested improvements to enhance functionality and user experience, such as providing clearer explanations of detected anomalies, offering previews for specific flagged anomalies, including visual summaries, creating more comprehensive cleaning reports, and improving the handling of error codes like #N/A and #DIV/0!. These suggestions point to potential areas for strengthening transparency in future versions of the system.

5.2.2 User Experience and Feelings (Red Hat)

Table 2. Descriptive Statistics on User Experience and Feelings

Questions	Mean	Verbal Description
I felt confident while using RawtoReady.	4.54	Strongly Agree
The interface felt user-friendly and not intimidating.	4.59	Agree
I felt that my data was being cleaned properly.	4.48	Strongly Agree
I would personally use RawtoReady again in future projects.	4.54	Strongly Agree

Respondents strongly agreed that they felt confident while using RawtoReady, indicating a positive overall experience. They also strongly agreed that the interface was user-friendly and not intimidating, making it easy to navigate even for users with limited data cleaning experience (Table 2). A majority of participants strongly agreed that they would use RawtoReady again in future projects. Open-ended responses further supported these findings, with many users describing their experience as highly positive, expressing satisfaction, confidence, and relief. Participants noted that the tool was intuitive, efficient, and easy to use, saving time and effort compared to manual data cleaning. The user interface was frequently praised for being clear, friendly, and beginner-friendly, which made the process less intimidating. Some users reported brief moments of confusion when working with messier datasets, but still recognized the tool's functionality and workflow. Overall, the system left users feeling confident, satisfied, and willing to use it for future projects.

5.2.3 Risks and Limitations (Black Hat)

Table 3. Descriptive Statistics on Risks and Limitations

Questions	Mean	Verbal Description
Some parts of RawtoReady felt confusing or hard to use.	2.41	Disagree
I worried that some important data might be lost during cleaning.	2.93	Neutral
I feel that RawtoReady still has limitations compared to more advanced tools.	3.00	Neutral

Respondents generally disagreed that any part of RawtoReady felt confusing or difficult to use (Table 3), and they expressed a neutral stance regarding potential data loss during the cleaning process as well as the tool's limitations compared to more advanced alternatives. Open-ended responses indicated that most participants did not encounter major challenges or frustrations, reflecting overall satisfaction with the tool's reliability and simplicity. However, some minor concerns were mentioned, including the accidental loss of important data during cleaning, longer processing times with larger datasets, and limited visibility of modified or removed rows. Additional suggestions included clearer explanations and improvements for features such as anomaly detection and sidebar navigation, as well as an optional review step before finalizing changes to ensure that no critical data is lost.

5.2.4 Strengths and Benefits (Yellow Hat)

Table 4. Descriptive Statistics on Strengths and Benefits

Questions	Mean	Verbal Description
RawtoReady saved me time compared to manual data cleaning.	4.57	Strongly Agree
The tool improved the overall quality of my dataset.	4.41	Strongly Agree
I found RawtoReady easier to use than other cleaning tools I know.	4.43	Strongly Agree
I would recommend RawtoReady to classmates, colleagues, or coworkers.	4.61	Strongly Agree

Respondents strongly agreed that RawtoReady significantly saved them time compared to manual data cleaning (Table 4), improved the overall quality of their datasets, was easier to use than other cleaning tools, and that they would recommend it to classmates, colleagues, or coworkers. Open-ended responses revealed that most users found the automatic data cleaning feature to be the most helpful, as it reduced manual effort and enhanced data quality. Many appreciated the tool's ability to detect and fix missing values, duplicates, and inconsistencies with just a few clicks. Others highlighted the cleaning summary and anomaly detection as particularly useful for providing transparency and understanding the changes made to datasets. Additionally, some participants noted that the advanced options and organized interface made the tool easier to navigate and customize according to individual needs.

5.2.5 Suggestions and Improvements (Green Hat)

Table 5. Descriptive Statistics on Suggestions and Improvements

Questions	Mean	Verbal Description
I can think of many ways RawtoReady could be improved further.	3.68	Agree
I would like the tool to support more file formats beyond CSV.	4.48	Strongly Agree
I see potential for RawtoReady to be used in more industries or fields.	4.54	Strongly Agree
RawtoReady could include more advanced features for professional use	4.43	Strongly Agree

Respondents agreed that there are still several ways RawtoReady could be improved (Table 5). They strongly supported expanding file format compatibility beyond CSV, recognized the tool's potential for use across more industries, and suggested adding advanced features for professional use. Open-ended responses indicated that while most users were satisfied with the current features, they recommended improvements such as supporting Excel or JSON files, adding data visualization or exploratory data analysis (EDA) tools, and enhancing usability through colors, tooltips, and short tutorials. A few participants also suggested advanced options like custom null replacements, manual data type selection, and validation rules. Overall, these responses indicate that users value the tool's simplicity but see opportunities for it to become more versatile and functional.

5.2.6 Process and Organization (Blue Hat)

Table 6. Descriptive Statistics on Process and Organization

Questions	Mean	Verbal Description
Using RawtoReady helped me think more clearly about my data preparation process.	4.43	Strongly Agree
The tool made the cleaning process feel more organized and structured.	4.59	Strongly Agree
I felt more in control of the overall data preparation when using RawtoReady.	4.38	Strongly Agree

Respondents strongly agreed that using RawtoReady helped them think more clearly about their data preparation process (Table 6), made the cleaning tasks feel more organized and structured, and allowed them to feel more in control of overall data handling. Open-ended feedback emphasized that the tool is a valuable part of their workflow, saving time by automating repetitive tasks such as handling duplicates, imputing missing values, and standardizing formats. Participants noted that this time-saving feature allows them to focus more on analysis, visualization, and model building. Users described RawtoReady as flexible, easy to use, and well-integrated into their workflow. Beginners appreciated its straightforward approach, which does not require advanced technical skills, while advanced users valued its efficiency in reducing errors and streamlining repetitive tasks. Overall, RawtoReady is regarded as a practical, reliable, and effective tool that strengthens the foundation of data preparation and promotes more organized workflows.

5.3 Design of Proposed System

The system provides a guided, four-step process for dataset cleaning:

- Step 1: Upload Dataset
  The user begins by logging in and uploading a dataset through the sidebar. Once uploaded, the dataset appears in the Raw Data Preview tab, along with dataset details such as its data type and descriptive statistics.

- Step 2: Choose Cleaning Options
  Users select how they want to handle issues in the dataset. The sidebar provides two options:
  - Missing Values: Fill with N/A, mean, median, most common, or drop rows
  - Advanced Options: Remove duplicates, standardize column names, normalize text, fix date formats, validate emails, fuzzy standardize values, and detect anomalies.
- Step 3: Run Cleaning
  After selecting the desired options, the user clicks Run Cleaning. The system then processes the dataset based on the chosen configurations.
- Step 4: Download Cleaned Dataset
  Results are presented across three main tabs:
  - Raw Data Preview - shows the original dataset, with a Summary of Cleaning and a download option
  - Cleaned Data Preview - displays the processed or cleaned dataset along with the summary and download option.
  - Anomalies Detected - list rows flagged as anomalies and provide recommendations, also with the summary and download option (Figure 4).

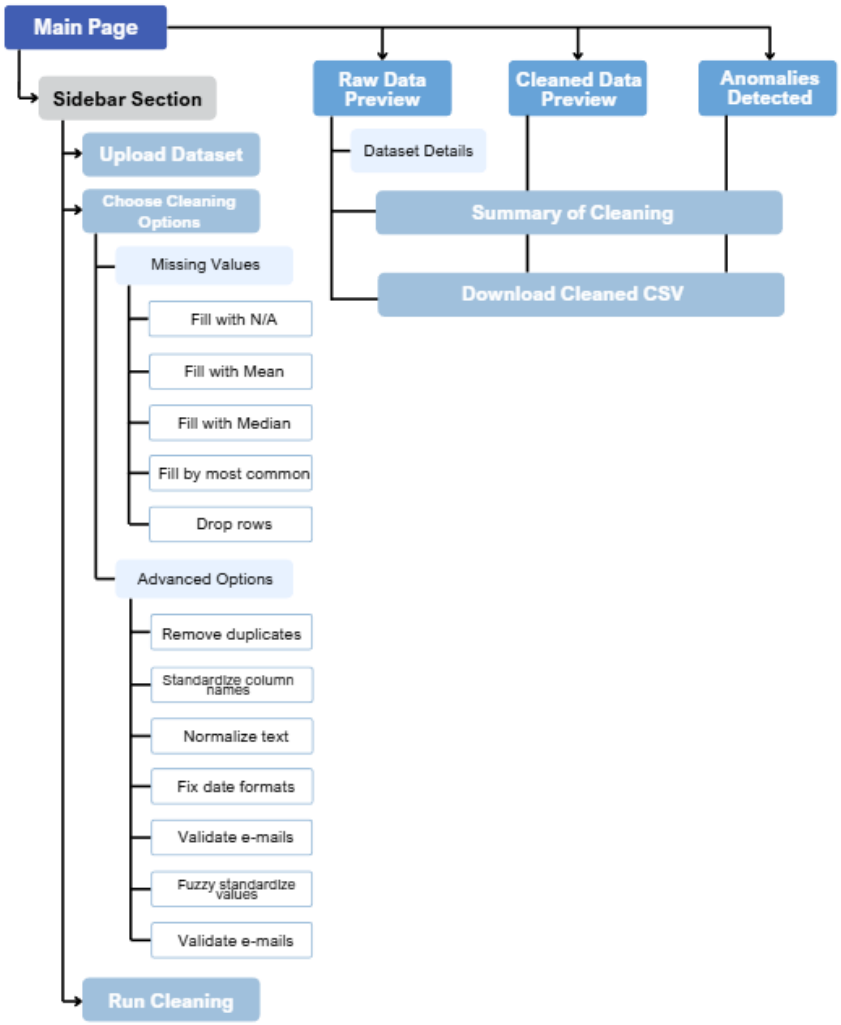

Figure 4. Sitemap of RawtoReady

5.4 Functional and Non-functional Requirements

5.4.1 Functional Requirements

1. Upload Dataset

This function enables users to upload a raw CSV dataset into the system. The system shall:

- Accept only CSV files up to 200MB.
- Provide options for file browsing or drag-and-drop.
- Validate the uploaded file format and readability before processing.
- Display a preview of the raw dataset after a successful upload.

2. Choose Cleaning Options

This function allows the user to select different cleaning operations. The system shall:

- Provide options for handling missing values.
- Allows standardization of column names for consistency.
- Allow normalization of text values.
- Fix inconsistent date formats to ensure uniformity.
- Validate emails to ensure correctness.
- Apply fuzzy standardization to group similar text values together.
- Enable anomaly detection to flag unusual or irregular rows in the dataset.

3. Apply Cleaning

This function allows the user to apply the selected cleaning operations. The system shall:

- Automatically run selected cleaning operations when the user clicks the "Run Cleaning" button.
- Generate a summary of the cleaning process.
- Display a preview of the cleaned dataset.
- Display a preview of rows flagged as anomalies.

4. Save and Download

This function allows the user to save and export the cleaned dataset. The system shall:

- Allow the user to download the cleaned dataset in CSV format.

5.4.2 Non-Functional Requirements

1. Usability
- Users can navigate between uploading, cleaning, and downloading datasets with minimal clicks.
- The interface shall provide tooltips or short guides to guide users throughout the cleaning process.

2. Reliability
- The system shall handle errors gracefully, providing clear messages for invalid file formats, failed uploads, or processing issues.
- The system shall ensure data integrity, preventing loss or corruption of uploaded datasets.

3. Performance
- The system shall process up to 200MB efficiently, completing cleaning operations within a reasonable time.
- File upload, preview, and download operations shall respond without noticeable delay.

4. Supportability
- The system codebase shall be modular and documented to allow maintenance and the addition of new features.

5.5 Modules

The system comprises several functional modules that work together to automate data cleaning. The Data Upload Module handles the import and validation of datasets. The Data Cleaning Module performs the necessary cleaning operations on the dataset chosen by the user. The Data Preview Module generates reports such as Raw Data Preview and Cleaned Data Preview, allowing users to examine the dataset before and after cleaning. The Anomaly Detection Module identifies potential issues in the dataset for further review. The Download Module enables users to export the cleaned dataset for further use. Lastly, the Summary Module provides an overview of the cleaning result. Together, these modules enable the system to process datasets efficiently and provide users with actionable insights.

5.6 Screen Layout

The **Home Page** of RawtoReady serves as the main entry point for users. It provides an overview of the system and guides users through the initial steps of the data cleaning process. The left sidebar contains navigation options and an upload panel where users can drag and drop CSV files or browse from their device. The main display area welcomes the user and presents tabs for data previews that ensures a clear workflow from upload to analysis.

Once a .csv file/dataset has been uploaded, the **Raw Data Preview Page** displays the first few rows of the raw data for initial inspection. It allows users to review the dataset's structure, column names, and sample values before performing any cleaning. The sidebar on the left presents the available cleaning options, including handling missing values and accessing advanced cleaning features (Figure 5).

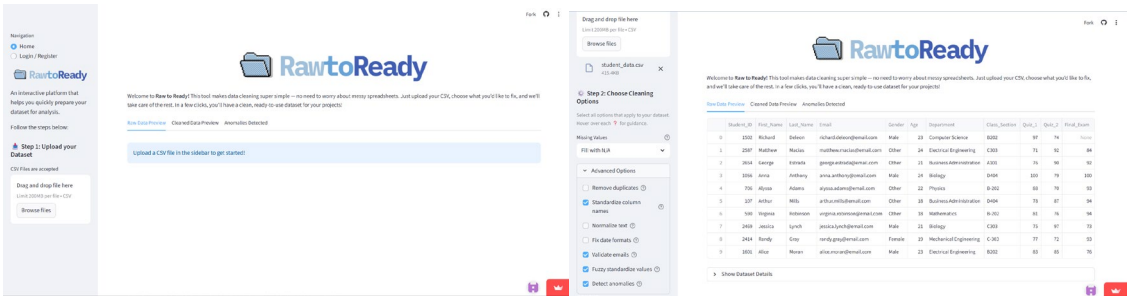

Figure 5. Home Page & Raw Data Preview Page (Left-to-Right)

After running the cleaning process, the **Cleaned Data Preview Page** displays the processed dataset along with a summary of the cleaning results. The summary indicates that the total number of rows remained at 5,000, with 250 missing values successfully fixed, and no anomalies detected. This page allows users to review the cleaned data before exporting it, with a button available to save or download the cleaned CSV file for further use.

The **Anomalies Detected tab** is also shown below. Similar to the Raw Data and Cleaned Data Preview pages, this section displays a dataframe showing the rows identified as anomalies. Since anomaly handling usually depends on the domain of the dataset, the system provides a recommendation panel to guide users in deciding how to address these irregularities (Figure 6)

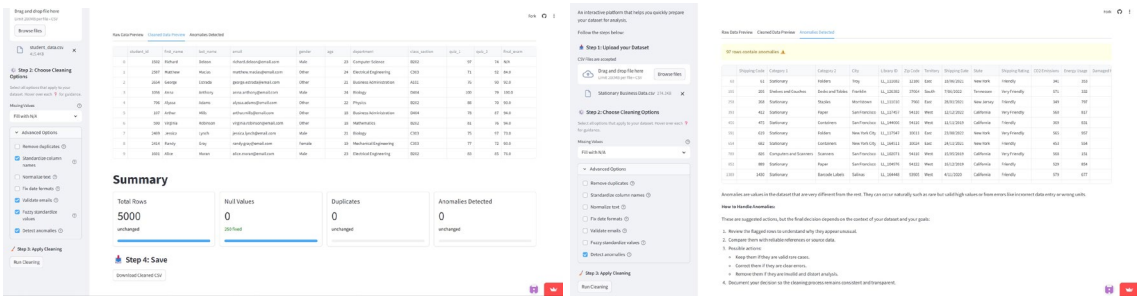

Figure 6. Cleaned Data Preview Tab & Anomalies Detected Tab (Left-to-Right)

6. Conclusion

In this study, the researchers developed RawtoReady, a web-based automatic data cleaning system designed to simplify and speed up dataset preparation. The system provides users with one-click solutions for handling missing values, removing duplicates, standardizing text and dates, validating emails, and detecting anomalies. Its goal is to make data cleaning accessible to students, researchers, and professionals without requiring coding skills or costly tools. Through its user-friendly interface and clear cleaning summaries, RawtoReady enables users to efficiently transform raw data into ready-to-analyze datasets.

The 6 Thinking Hats methodology guided the data collection process by framing the user survey across different perspectives, including usability, effectiveness, risks, and potential improvements. With responses from 56 participants, the survey findings revealed that users found RawtoReady easy to use, efficient, and helpful in reducing the effort and time required for manual data cleaning. The SWOT analysis further identified the system's strengths, including its user-friendly interface, time-saving features, and clear data summaries, as well as its weaknesses, such as limited file format support, limited visibility of data modifications, and processing time for larger datasets. These insights were valuable in determining the current performance of the system and user acceptance.

Acknowledgements

The authors would like to express their sincere gratitude to Mapúa University, particularly the School of Information Technology, for providing academic guidance and institutional support throughout this research. Special appreciation is extended to Sir Richard Monreal for his valuable mentorship, guidance, and continuous support during the development of this study. The authors also wish to thank all individuals who willingly participated and cooperated in the testing and evaluation of RawtoReady, whose feedback and involvement greatly contributed to the improvement and validation of the system.

References

Abdelaal, M., Hammacher, C. and Schoening, H., Rein: A comprehensive benchmark framework for data cleaning methods in ML pipelines, arXiv preprint, 2023. Available: https://arxiv.org/abs/2302.04702

Buddiga, S.K.P., The importance of data cleaning in machine learning: Best practices and techniques, European Journal of Advances in Engineering and Technology, vol. 7, no. 10, pp. 70–73, 2020.

Chiriyankandath, J.J., Kamath, A., Khatib, M. and Nawal, P., Automated data cleaning, University of Mumbai, 2020.

Côté, P.O., Nikanjam, A., Ahmed, N., Humeniuk, D. and Khomh, F., Data cleaning and machine learning: A systematic literature review, arXiv preprint, 2023. Available: https://arxiv.org/abs/2310.01765

De Bono, E., Six thinking hats, Penguin Books, London, 1985.

Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Cui, Y., Jiang, X., Sheng, S., Li, Q. and Gao, R., Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint, JMIR Medical Informatics, vol. 11, e48911, 2023.

Hameed, M., Vitagliano, G., Panse, F. and Naumann, F., Repairing raw data files with TASHEEH, ACM SIGMOD Record, vol. 54, no. 1, pp. 90–99, 2025.

HosseinZadeh, M., Azhir, E., Ahmed, O.H., Ghafour, M.Y., Ahmed, S.H., Rahmani, A.M. and Vo, B., Data cleansing mechanisms and approaches for big data analytics: A systematic study, Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 1, pp. 99–111, 2021.

Jäger, S. and Biessmann, F., From data imputation to data cleaning — Automated cleaning of tabular data improves downstream predictive performance, Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS 2024), Proceedings of Machine Learning Research, vol. 238, pp. 3394–3402, 2024.

Kilani, S.O., Amao, I.K., Ojo, N.A. and Samson, P.A., AI and machine learning-powered automated data cleaning methods: Improving data quality, International Journal of Advance Research Publication and Reviews, vol. 2, no. 4, pp. 35–47, 2025.

Krishnan, S. and Wu, E., AlphaClean: Automatic generation of data cleaning pipelines, arXiv preprint, 2019. Available: https://doi.org/10.48550/arXiv.1904.11827

Li, L., Fang, L., Ludäscher, B. and Torvik, V.I., AutoDCWorkflow: LLM-based data cleaning workflow auto-generation and benchmark, arXiv preprint, 2024. Available: https://doi.org/10.48550/arXiv.2412.06724

Martins, P., Cardoso, F., Váz, P., Silva, J. and Abbasi, M., Performance and scalability of data cleaning and preprocessing tools: A benchmark on large real-world datasets, Data, vol. 10, no. 5, p. 68, 2025.

Nargesian, F., Zhu, E., Pu, K.Q. and Miller, R.J., Auto-clean: A programmatic data cleaning framework, IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3057–3070, 2021.

Oladipupo, M.A., Obuzor, P.C., Bamgbade, B.J., Adeniyi, A.E., Olagunju, K.M. and Ajagbe, S.A., An automated Python script for data cleaning and labeling using machine learning technique, Informatica, vol. 47, no. 6, 2023.

Panwar, V., AI-powered data cleansing: Innovative approaches for ensuring database integrity and accuracy, International Journal of Computer Trends and Technology, vol. 72, no. 4, pp. 116–122, 2024.

Pradhan, R., Zhu, J., Glavic, B. and Salimi, B., Automated data cleaning can hurt fairness in machine learning-based decision making, IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 12, pp. 7368–7379, 2024.

Rahi, S., Research design and methods: A systematic review of research paradigms, sampling issues and instruments development, International Journal of Economics & Management Sciences, vol. 6, no. 2, pp. 1–5, 2017.

Rahm, E. and Do, H.H., Data cleaning: Problems and current approaches, IEEE Data Engineering Bulletin, vol. 23, no. 4, pp. 3–13, 2000.

Ridzuan, F. and Zainon, W.M.N., A review on data cleansing methods for big data, Procedia Computer Science, vol. 161, pp. 731–738, 2019.

Shi, X., Prins, C., Van Pottelbergh, G., Mamouris, P., Vaes, B. and De Moor, B., An automated data cleaning method for electronic health records by incorporating clinical knowledge, BMC Medical Informatics and Decision Making, vol. 21, no. 1, p. 258, 2021.

Valêncio, C.R., Jardini, T., Martins, V.H.P., Colombini, A.C. and Fortes, M.Z., A system proposal for automated data cleaning environment, ITEGAM Journal of Engineering and Technology for Industrial Applications, vol. 6, no. 25, pp. 4–15, 2020.

Wongsa, S. and Cojorn, K., Enhancing the students' problem-solving ability through situation-based learning with the six thinking hats technique, Journal of Advanced Sciences and Mathematics Education, 2024.

Yasin, H.M. and Khorsheed, A.K., Automated data cleaning in large databases using machine learning methods, Asian Journal of Research in Computer Science, vol. 18, no. 5, pp. 364–386, 2025.

Biographies

Gio Miguel R. Bihasa is a third-year Bachelor of Science in Data Science student at Mapúa University Makati.

Pam Cassedy T. Dumangeng is a third-year Bachelor of Science in Data Science student at Mapúa University Makati.

Kim Caryl H. Esperanza is a third-year Bachelor of Science in Data Science student at Mapúa University Makati.

Louella Josephine A. Ng is a third-year Bachelor of Science in Data Science student at Mapúa University Makati.

Richard N. Monreal is a faculty member of Mapúa University, where he teaches Data Science, Information Systems, Entertainment and Multimedia Computing, Computer Science, and Information Technology courses under the School of Information Technology. He holds a Doctorate degree in Information Technology from the University of the Cordilleras.