International Journal of Multidisciplinary Allied Research Review and Practices (2455-1570)www.ijmarrp.in

Smart Data Cleaner: An Intelligent Framework for Automated Data Preprocessing, Interactive Visualization, and Conversational Analytics

Siddesh.U.Hallale, Aneek Chowdhury, Sayyam Choudhary, Satyam Kumar
Department of Computer Engineering, MIT ADT University, Pune, India

Abstract

The rapid growth of data across enterprises, educational institutions, healthcare, finance, and digital platforms has increased the need for efficient data preprocessing techniques. Real-world datasets often contain missing values, duplicate records, inconsistent formats, incorrect data types, and other quality issues that reduce the accuracy of analysis and decision-making. Existing data-cleaning methods are largely rule-based, require significant manual effort, and often demand technical expertise. Although several tools provide data preprocessing or visualization capabilities, few offer an integrated platform that combines automated data cleaning, interactive visualization, conversational analytics, and intelligent pattern discovery.

This research proposes **Smart Data Cleaner**, a unified framework designed to simplify data preprocessing and exploratory data analysis for both technical and non-technical users. The system enables users to upload CSV or Excel datasets, automatically clean and preprocess data, generate descriptive statistics, create interactive visualizations, and identify hidden patterns, correlations, trends, and anomalies within the dataset. Rather than performing predictive analytics, the framework focuses on descriptive insights that help users better understand their data and support informed decision-making. A conversational chatbot powered by the **Google Gemini API** allows users to interact with datasets using natural language, making analytical tasks more accessible without requiring programming knowledge.

The proposed framework will be developed using **Python**, **Fast API**, **React.js**, **TypeScript**, **PostgreSQL**, **SQL Alchemy**, **JWT Authentication**, **Pandas**, **NumPy**, **Scikit-learn**, **Plotly**, **Matplotlib**, **Git**, **GitHub**, and **Docker** to provide a secure, scalable, and user-friendly platform. The proposed system aims to reduce manual effort, improve data quality, simplify exploratory analysis, and establish a foundation for future research in Explainable Artificial Intelligence (XAI), intelligent data-quality assessment, and machine learning-assisted data analytics.

Keywords:

Data Cleaning, Data Preprocessing, Data Visualization, Conversational Analytics, Pattern Discovery, Google Gemini API, Machine Learning, Fast API, React.js.

1. Introduction

1.1 Background

The rapid growth of digital technologies has led to an exponential increase in the amount of data generated by businesses, educational institutions, healthcare organizations, financial sectors, and online platforms. High-quality data is essential for accurate analysis, machine learning, and informed decision-making. However, real-world datasets often contain:

- Missing values
- Duplicate records
- Incorrect data types
- Inconsistent data formats
- Outliers and anomalies

These issues reduce data quality and can lead to inaccurate analytical results.

Studies indicate that **data scientists spend nearly 80% of their time cleaning and preparing data**, while only a small portion of their time is spent on actual analysis. This highlights the need for efficient and intelligent data preprocessing techniques.

1.2 Problem Statement

Traditional data-cleaning methods mainly rely on manual inspection, predefined rules, and domain expertise. Although several tools support individual preprocessing tasks, they often:

- Require technical knowledge.
- Focus only on specific cleaning operations.
- Lack integrated visualization and analytical capabilities.
- Do not provide conversational interaction for non-technical users.
- Require users to switch between multiple applications.

As data continues to grow in size and complexity, these limitations make existing approaches less efficient and more time-consuming.

1.3 Research Motivation

The increasing dependence on data-driven decision-making has highlighted the need for efficient and accessible data preprocessing tools. Existing solutions often focus on individual tasks such as data cleaning or visualization and generally require programming knowledge or technical expertise. This creates challenges for students, educators, researchers, and business professionals who need to analyze data but may not have a strong background in data science.

The motivation behind this research is to develop a unified platform that automates common data preprocessing tasks, generates interactive visualizations, discovers meaningful patterns within datasets, and enables users to interact with data through natural language. By integrating these capabilities into a single application, the proposed framework aims to reduce manual effort, improve data quality, and make data analysis more accessible and efficient for users across different domains.

1.4 Objectives

This research aims to:

- Develop an intelligent framework for automated data preprocessing.
- Improve data quality by handling missing values, duplicates, and inconsistent data.
- Generate interactive visualizations for better data understanding.
- Discover hidden patterns, trends, and anomalies through descriptive analytics.
- Integrate a chatbot powered by Google Gemini API for natural language interaction.
- Provide an easy-to-use platform for students, researchers, educators, and business professionals.

2. Literature Review / Background Study

2.1 Data Quality Dimensions

Data quality is one of the most important factors in data analytics and machine learning. Poor-quality data can produce inaccurate results and reduce the effectiveness of decision-making. According to existing research, data quality is generally evaluated using the following dimensions:

- **Accuracy** – Correctness of data values.
- **Completeness** – Presence of all required information.
- **Consistency** – Uniformity of data across different sources.
- **Validity** – Compliance with predefined rules and formats.
- **Uniqueness** – Elimination of duplicate records.

- **Timeliness** – Availability of up-to-date information.

Maintaining these dimensions ensures that datasets are reliable for analysis and decision-making.

2.2 Traditional Data Cleaning Techniques

Traditional data-cleaning methods mainly depend on manual inspection, predefined rules, and statistical techniques to improve data quality. Common approaches include:

- Manual data inspection and correction
- Rule-based validation
- Duplicate record detection
- Missing value replacement using mean, median, or mode
- Statistical outlier detection

Although these techniques are effective for small datasets, they require significant manual effort and technical expertise. They also become less efficient when working with large and continuously changing datasets.

2.3 Intelligent Data Cleaning and Conversational Analytics

Recent research has introduced intelligent approaches that combine automation, visualization, and conversational interfaces to simplify data analysis. Machine learning techniques can identify anomalies, detect duplicate records, discover hidden patterns, and improve preprocessing with minimal human intervention. Modern conversational systems further enable users to interact with datasets using natural language instead of writing complex queries.

These developments make data analysis more accessible for beginners while reducing the time required for preprocessing and exploration.

2.4 Data Visualization and Pattern Discovery

Data visualization plays an important role in understanding complex datasets. Interactive charts and dashboards help users identify trends, correlations, distributions, and anomalies that may not be visible in raw tables. Recent studies also emphasize descriptive pattern discovery, where analytical systems summarize recurring behaviors and meaningful relationships within the data without performing predictive analysis.

Combining visualization with automated preprocessing enables users to obtain reliable insights more efficiently.

2.5 Research Gap

A review of existing literature shows that most available solutions focus on individual aspects of data preprocessing such as duplicate detection, missing value handling, visualization, or conversational analytics. Very few studies propose a unified framework that integrates all these capabilities into a single platform.

Furthermore, many existing systems require programming knowledge or technical expertise, making them difficult for students and non-technical users. There is also limited research on combining automated data cleaning with descriptive pattern discovery and natural language interaction in one user-friendly application.

The proposed **Smart Data Cleaner** addresses this gap by integrating automated preprocessing, interactive visualization, conversational analytics, and intelligent pattern discovery within a single platform, providing an efficient and accessible solution for data analysis.

3. Methodology

3.1 Research Approach

This research follows a **design and development** approach to develop an intelligent data preprocessing framework. The proposed system is designed after studying existing research on data cleaning, visualization, and conversational analytics. The development process includes the following stages:

- Literature review of existing data-cleaning techniques.
- Analysis of common data quality issues.
- Design of an integrated preprocessing framework.
- Development of a web-based application.
- Performance evaluation using sample datasets.

The objective is to provide an automated and user-friendly platform that simplifies data preprocessing and exploratory data analysis.

3.2 Proposed Smart Data Cleaner Framework

The proposed framework consists of the following modules:

1. Data Upload Module

- Accepts CSV and Excel datasets.
- Validates file format before processing.

2. Data Preprocessing Module

- Detects and removes duplicate records.
- Identifies and handles missing values.
- Corrects inconsistent data types.
- Performs basic data validation.

3. Data Analysis Module

- Generates descriptive statistics.
- Identifies trends, correlations, and anomalies.
- Discovers hidden patterns within the dataset.

4. Visualization Module

- Creates interactive charts and graphs.
- Supports Bar Charts, Pie Charts, Line Graphs, Histograms, and Scatter Plots.
- Provides visual summaries for better understanding.

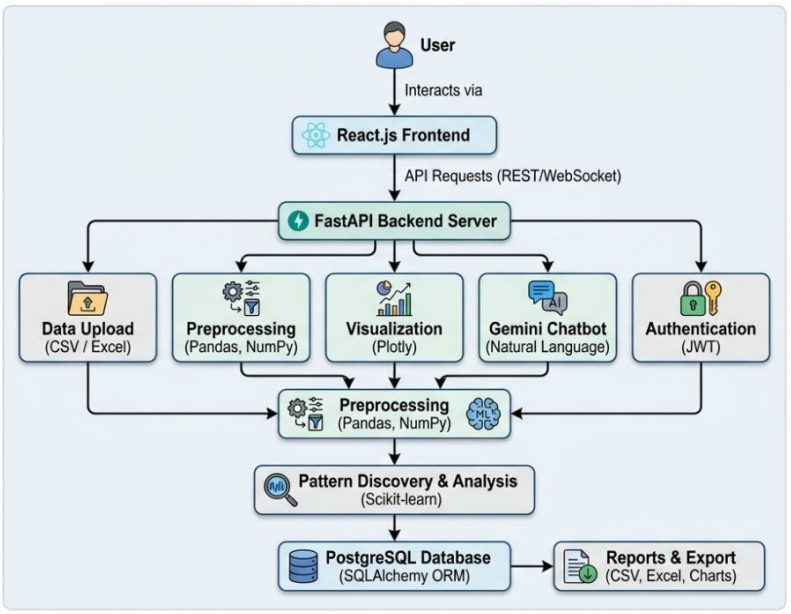
5. Conversational Analytics Module

- Uses the Google Gemini API for natural language interaction.
- Allows users to search, filter, sort, group, and analyze datasets through chatbot queries.
- Explains analytical results in simple language.

6. Export Module

- Downloads the cleaned dataset.
- Exports generated reports and visualizations for future use.

Figure 1: Smart Data Cleaner Framework Architecture



3.3 Evaluation Metrics

The performance of the proposed framework will be evaluated using the following metrics:

- **Data Cleaning Accuracy** – Percentage of correctly identified and cleaned records.
- **Duplicate Detection Rate** – Effectiveness in identifying duplicate entries.
- **Missing Value Handling** – Accuracy of missing value treatment.
- **Processing Time** – Time required to clean and analyse datasets.
- **Pattern Discovery Effectiveness** – Ability to identify meaningful trends and relationships.
- **User Experience** – Ease of use and accessibility of the chatbot and visualization features.
- **Visualization Quality** – Clarity and usefulness of generated charts and dashboards.

The proposed framework will be tested using datasets from different domains to evaluate its efficiency, usability, and overall performance.

4. Discussion / Analysis

4.1 Importance of Automated Data Preprocessing

Data preprocessing is one of the most time-consuming stages in the data analysis pipeline. Existing studies highlight that poor-quality data directly affects analytical accuracy and decision-making. Automating common preprocessing tasks such as duplicate removal, missing value handling, and data validation helps reduce manual effort and improves the overall reliability of datasets.

4.2 Pattern Discovery and Data Analysis

Modern analytical systems are expected to do more than clean data. They should also help users understand the information hidden within datasets. The proposed framework focuses on descriptive pattern discovery by identifying:

- Recurring trends
- Correlations between variables
- Unusual or abnormal records
- Frequently occurring patterns

These insights assist users in understanding their data without performing predictive analysis.

4.3 Interactive Visualization

Visualization transforms processed data into an easily understandable format. Interactive charts allow users to explore distributions, trends, and comparisons more effectively than raw tables. Graphical representation also helps identify inconsistencies and hidden relationships that may otherwise remain unnoticed.

The framework supports multiple visualization techniques, including:

- Bar Charts
- Pie Charts
- Line Graphs
- Histograms
- Scatter Plots

4.4 Conversational Analytics

Recent research has demonstrated the growing importance of conversational interfaces in data analysis. Instead of writing complex queries, users can interact with datasets using natural language.

The chatbot integrated into the proposed framework allows users to:

- Search datasets
- Filter records
- Sort data
- Group information

- Request statistical summaries
- Understand discovered patterns through simple explanations

This improves accessibility for users with limited programming knowledge.

4.5 Challenges in Existing Systems

Despite the availability of various data preprocessing and analysis tools, existing systems have several limitations in providing a simple and integrated data-analysis workflow. The major challenges include:

- **Manual preprocessing:** Identification and handling of missing values, duplicate records, inconsistent data, and other data-quality issues often require manual effort.
- **Fragmented workflow:** Data cleaning, visualization, analysis, and pattern identification are often performed through separate tools or stages.
- **Limited accessibility:** Existing solutions may require technical knowledge to perform common data-processing and exploratory analysis tasks.
- **Limited natural-language interaction:** Users generally interact with data through predefined options, commands, or technical procedures rather than simple natural-language queries.
- **Difficulty in data exploration:** Extracting useful trends, relationships, and patterns from a dataset can require multiple analysis steps and manual interpretation.
- **Lack of an integrated solution:** A single platform that combines automated preprocessing, visualization, analysis, pattern discovery, and conversational data exploration is not commonly available in a simple user-oriented workflow.

These challenges demonstrate the need for an integrated system that can simplify data preprocessing and exploratory analysis while reducing the effort required from users.

4.6 Comparison with Existing Approaches

Feature	Traditional Tools	Smart Data Cleaner
Automated Data Cleaning	Partial	✓
Interactive Visualization	Limited	✓
Pattern Discovery	Limited	✓
Conversational Analytics	✗	✓
CSV & Excel Support	✓	✓
Beginner Friendly	Limited	✓
Unified Platform	✗	✓

4.7 Summary of Findings

Based on the literature reviewed and the proposed framework, the following findings can be summarized:

- Automated data preprocessing significantly reduces the time and effort required to prepare datasets for analysis.
- Interactive visualizations improve data interpretation by presenting trends, distributions, and relationships in a clear and understandable manner.
- Pattern discovery helps identify hidden correlations, recurring trends, and anomalies, enabling users to gain meaningful insights from data.
- Conversational analytics enhances accessibility by allowing users to interact with datasets using natural language instead of programming queries.
- An integrated framework that combines preprocessing, visualization, pattern discovery, and chatbot-based interaction provides a more efficient and user-friendly solution than using multiple standalone tools.

5.Findings / Expected Results

The proposed **Smart Data Cleaner** framework integrates automated data preprocessing, visualization, analytical functions, pattern discovery, and conversational interaction into a single platform. The expected outcomes of the proposed system are:

- Automated preprocessing is expected to reduce the manual effort involved in identifying and handling common data-quality issues.
- Interactive visualizations are expected to improve understanding of data trends, distributions, and relationships.
- Integrated analytical features are expected to make dataset exploration faster and more convenient.
- The conversational chatbot is expected to simplify data exploration by allowing users to interact with datasets using natural-language queries.
- Pattern discovery is expected to assist in identifying useful trends, relationships, and potential anomalies in the data.
- The integrated workflow is expected to improve accessibility for both technical and non-technical users by bringing major data-analysis functions into one application.

Overall, the proposed framework is expected to provide a **simpler, faster, and more accessible approach to data preprocessing and exploratory data analysis**.

5.1 Performance Evaluation

Evaluation Parameter	Expected Outcome
Data Cleaning	Accurate removal of duplicates and handling of missing values
Data Visualization	Clear and interactive graphical representation
Pattern Discovery	Identification of meaningful trends and correlations
Chatbot Interaction	Easy natural language querying of datasets
User Accessibility	Suitable for both technical and non-technical users
Overall Efficiency	Reduced preprocessing time and improved workflow

5.2 Expected Applications

The proposed framework can be applied in various domains, including:

- Educational institutions for student performance analysis.
- Business organizations for sales and customer data analysis.
- Healthcare systems for medical record preprocessing.
- Financial institutions for transaction and market trend analysis.
- Research organizations for preparing datasets before machine learning.
- Government organizations for analyzing large public datasets.

Conclusion

This research helped us understand the importance of data preprocessing and the challenges involved in working with real-world datasets. From the literature reviewed, it was observed that although many tools support data cleaning, visualization, and data analysis, most of them focus on specific tasks and often require technical knowledge to use effectively. This highlighted the need for a simple and integrated solution that can make data preprocessing easier and more accessible.

Based on these findings, we proposed **Smart Data Cleaner**, a framework that combines automated data cleaning, interactive visualization, conversational analytics, and pattern discovery within a single platform. The proposed system is designed to help user clean datasets, explore meaningful patterns, and interact with data using natural language, making the overall analysis process more efficient and user-friendly.

Since the project is currently in the design and development phase, the proposed framework has not yet been fully implemented or evaluated. The next stage of this work will focus on developing the application, testing it with real-world datasets, and evaluating its performance. We also plan to explore advanced features such as Explainable Artificial Intelligence (XAI), intelligent data-quality assessment, and machine learning-assisted analytics to further improve the capabilities of the system.

References

[1] P.-O. Côté, A. Nikanjam, N. Ahmed, D. Humeniuk, and F. Khomh, "Data Cleaning and Machine Learning: A Systematic Literature Review," *Automated Software Engineering*, vol. 31, no. 2, p. 54, 2024, doi: 10.1007/s10515-024-00453-w.
https://doi.org/10.1007/s10515-024-00453-w
[Published: June 2024]

[2] F. Neutatz, B. Chen, Y. Alkhatib, J. Ye, and Z. Abedjan, "Data Cleaning and AutoML: Would an Optimizer Choose to Clean?," *Datenbank-Spektrum*, vol. 22, no. 2, pp. 121–130, 2022, doi: 10.1007/s13222-022-00413-2.
https://doi.org/10.1007/s13222-022-00413-2
[Published: May 13, 2022]

[3] P. Maddigan and T. Susnjak, "Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models," *IEEE Access*, vol. 11, pp. 45181–45193, 2023, doi: 10.1109/ACCESS.2023.3274199.
https://doi.org/10.1109/ACCESS.2023.3274199
[Published: May 2023]

[4] L. Zha, J. Zhou, L. Li, R. Wang, Q. Huang, S. Yang, J. Yuan, C. Su, X. Li, A. Su, T. Zhang, C. Zhou, K. Shou, M. Wang, W. Zhu, G. Lu, C. Ye, Y. Ye, W. Ye, Y. Zhang, X. Deng, J. Xu, H. Wang, G. Chen, and J. Zhao, "TableGPT: Towards Unifying Tables, Nature Language and Commands into One GPT," arXiv preprint arXiv:2307.08674, 2023.
https://arxiv.org/abs/2307.08674
[Published: July 17, 2023 — arXiv preprint]

[5] W. Zhang, Y. Wang, Y. Song, V. J. Wei, Y. Tian, Y. Qi, J. H. Chan, R. C.-W. Wong, and H. Yang, "Natural Language Interfaces for Tabular Data Querying and Visualization: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 6699–6718, 2024, doi: 10.1109/TKDE.2024.3400824.
https://doi.org/10.1109/TKDE.2024.3400824
[Published: November 2024]

[6] Y. Wu, Y. Wan, H. Zhang, Y. Sui, W. Wei, W. Zhao, G. Xu, and H. Jin, "Automated Data Visualization from Natural Language via Large Language Models: An Exploratory Study," *Proceedings of the ACM on Management of Data*, vol. 2, no. 3, Art. no. 115, 2024, doi: 10.1145/3654992.
https://doi.org/10.1145/3654992
[Published: 2024]

[7] S. Zhang, Z. Huang, and E. Wu, "Data Cleaning Using Large Language Models," arXiv preprint arXiv:2410.15547, 2024.
https://arxiv.org/abs/2410.15547
[Published: October 21, 2024 — arXiv preprint]

[8] L. Li, L. Fang, B. Ludäscher, and V. I. Torvik, "AutoDCWorkflow: LLM-based Data Cleaning

Workflow Auto-Generation and Benchmark," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7766–7780, 2025, doi: 10.18653/v1/2025.findings-emnlp.410.
https://doi.org/10.18653/v1/2025.findings-emnlp.410
[Published: November 2025]

[9] G. Ouyang, J. Chen, Z. Nie, Y. Gui, Y. Wan, H. Zhang, and D. Chen, "nvAgent: Automated Data Visualization from Natural Language via Collaborative Agent Workflow," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pp. 19534–19567, 2025, doi: 10.18653/v1/2025.acl-long.960.
https://doi.org/10.18653/v1/2025.acl-long.960
[Published: July 2025]

[10] A. Akella, A. Kaul, K. Narayanam, and S. Mehta, "Quality Assessment of Tabular Data Using Large Language Models and Code Generation," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2713–2748, 2025, doi: 10.18653/v1/2025.emnlp-industry.183.
https://doi.org/10.18653/v1/2025.emnlp-industry.183
[Published: November 2025]