International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 46-54

An AI Assisted Web Tool for Survey Data Cleaning and Statistical Reporting

Dr. J. Yogapriya [1],Abinithi S [2],Dhivya V [3],Harshavarthini V [4]
[1] Professor Department of Computer Science and Engineering Kongunadu college of Engineering and Technology Tamilnadu, India
[2,3,4] Bachelor of Engineering Department of Computer Science and Engineering Kongunadu college of Engineering and Technology Tamilnadu, India

Abstract. Survey-based research often faces data quality issues such as missing values, duplicate entries, inconsistent responses, and outliers, which can affect the accuracy of statistical analysis and decision-making. This paper proposes an AI- assisted web-based system for automated survey data cleaning and statistical reporting. The system integrates rule-based techniques and Large Language Model (LLM) capabilities to detect missing values, remove duplicates, identify outliers, and standardize survey responses with minimal manual intervention. After preprocessing, the platform automatically generates descriptive statistical reports that help researchers quickly analyze cleaned datasets. The proposed system improves the reliability and efficiency of survey data analysis while reducing manual preprocessing effort. It supports applications in healthcare surveys, educational research, and workforce studies, contributing to data-driven decision-making aligned with Sustainable Development Goals (SDGs) including Good Health and Well-Being, Quality Education, Decent Work and Economic Growth, Indutry Innovation and Infrastructure, and Partnerships for the Goals.

Keywords — Artificial Intelligence, Survey Data Cleaning, Large Language Models, Data Preprocessing, Statistical Reporting, Survey Analytics, Data Quality Management.

I. Introduction

Survey-based research is widely used in domains such as healthcare, education, social sciences, and market research to collect structured information from large populations. Surveys enable organizations and researchers to understand public opinion, evaluate policies, analyze behavioral patterns, and support evidence-based decision-making. However, the quality of survey analysis largely depends on the accuracy and consistency of the collected data. In many real-world scenarios, survey datasets contain issues such as missing responses, duplicate entries, inconsistent formats, and extreme values. These data quality problems often arise due to manual data entry errors, incomplete responses, or poorly structured survey forms, which can significantly affect the reliability of statistical analysis.

Traditional data cleaning methods often require extensive manual effort and domain expertise to identify and correct such issues. Researchers typically perform preprocessing tasks such as removing duplicate records, handling missing values, and standardizing response formats before conducting analysis. These manual processes are

time-consuming and may introduce human errors, particularly when working with large-scale datasets.

Recent advances in artificial intelligence and natural language processing have created opportunities to automate many aspects of data preprocessing and validation. In particular, Large Language Models (LLMs) provide powerful capabilities for interpreting text-based responses, detecting inconsistencies, and assisting in data standardization tasks. Integrating AI-driven techniques with traditional rule-based data cleaning methods can significantly improve the efficiency and accuracy of survey data preparation.

This research proposes an AI-assisted web-based platform for automated survey data cleaning and statistical reporting. The system combines rule-based preprocessing with LLM- assisted validation to detect missing values, remove duplicate entries, identify outliers, and standardize survey responses. By automating these processes and generating statistical summaries of cleaned datasets, the proposed system aims to improve the reliability, efficiency, and usability of survey-based research data and generates statistical insights from processed datasets.

II. Related Works

A. Survey Data Quality and Preprocessing

Survey datasets often suffer from quality issues such as missing values, inconsistent responses, duplicate records, and data entry errors. Previous studies have highlighted that improper data preprocessing can significantly affect the reliability of statistical analysis and predictive modeling. Researchers have therefore emphasized the importance of structured preprocessing techniques including data validation, normalization, and outlier detection before conducting survey analysis.

B. Traditional Data Cleaning Techniques

Traditional data cleaning methods rely on rule-based approaches and statistical techniques to detect inconsistencies in datasets. These methods typically involve manual procedures such as identifying missing values, removing duplicate entries, and applying normalization rules to standardize responses. Although effective in small datasets, manual data cleaning becomes increasingly difficult and time- consuming when working with large-scale survey data.

C. Automated Data Cleaning Systems

Several automated data cleaning frameworks have been proposed to reduce the complexity of preprocessing tasks. These systems utilize algorithms to detect anomalies, remove duplicates, and apply transformation rules to improve data quality. Automated approaches improve efficiency but often struggle with unstructured responses and textual survey inputs where contextual interpretation is required.

D. Artificial Intelligence in Data Processing

Artificial Intelligence techniques have been widely applied in data processing and analytics to improve data management workflows. Machine learning models can identify patterns and inconsistencies within datasets, enabling automated error detection and correction. AI-based systems also support scalable data preprocessing pipelines capable of handling large volumes of survey data.

E. Large Language Models for Data Validations

Recent advancements in Large Language Models (LLMs) have significantly enhanced automated text processing capabilities. LLMs can understand contextual information in natural language responses and assist in identifying inconsistencies or ambiguous entries within datasets. Their ability to interpret unstructured text makes them particularly useful in survey data validation and preprocessing tasks.

F. Natural Language Processing in Survey Analysis

Natural Language Processing (NLP) techniques have been used to analyze open-ended survey responses and textual feedback. NLP methods can categorize responses, detect sentiment, and standardize textual data for further statistical analysis. These approaches enable better interpretation of qualitative survey responses compared to traditional rule-based preprocessing methods.

G. Research Gap

Despite the progress in automated data cleaning and AI-driven analytics, many existing systems focus only on isolated preprocessing tasks and lack integrated solutions for survey data analysis. There is a need for a unified platform that combines rule-based data cleaning with AI-assisted validation and automated statistical reporting. The proposed system addresses this gap by providing an AI-assisted web-based platform that streamlines survey data cleaning.

Proposed Methodology

The proposed system presents an AI- assisted framework designed to automate survey data cleaning and statistical reporting. The methodology combines rule-based preprocessing techniques with Large Language Model (LLM)– assisted validation to improve the quality and reliability of survey datasets. Initially, survey data collected from online questionnaires or digital forms is uploaded through a web-based interface. The system performs preliminary validation to ensure that the dataset structure and formatting meet the required standards for processing.

Once the dataset is accepted, the preprocessing module automatically identifies missing values, duplicate records, and inconsistent responses. Rule-based techniques are applied to handle missing entries and remove redundant data, while statistical methods are used to detect outliers within numeric responses. In addition, the LLM-based validation module analyzes textual survey responses to detect irregularities and assist in response standardization. After completing the data cleaning process, the system generates descriptive statistical summaries that help researchers interpret the cleaned dataset efficiently. This automated workflow reduces manual preprocessing effort and improves the accuracy and consistency of survey data analysis.

III. System Architecture

A. Overall Architectural Design

The proposed system follows a layered architecture that integrates client interaction, backend processing, AI-based decision logic, and data storage. It enables seamless data flow from survey input to response analysis while supporting dynamic question flow

using branching rules. The design ensures scalability, efficiency, and reliable survey data processing.

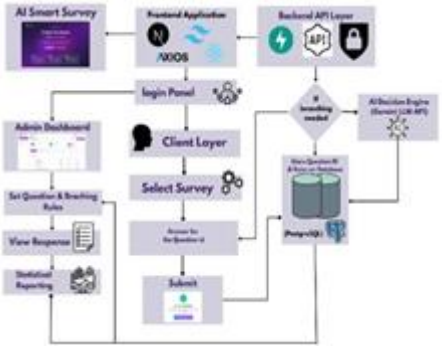


Figure 1: High-level architecture Diagram

A . Client Layer
The client layer provides the user interface for interacting with the survey system. It is built using a modern frontend framework and enables users to select surveys, submit responses, and view outputs. This layer ensures smooth user interaction and efficient communication with the backend system.

B. User Interface Layer
The user interface is developed using Next.js and Tailwind CSS to create a responsive and dynamic frontend. It manages API communication using Axios and ensures seamless data exchange between the client and backend layers. This layer enhances usability and accessibility for end users

C. Application Layer
The application layer handles core system functionalities including API processing, authentication, and validation. It is implemented using FastAPI and REST APIs to manage survey flow and user interactions. This layer acts as the central processing unit of the system.

D. AI Decision Engine
The AI decision engine integrates a Large Language Model to process user responses and apply branching logic. It dynamically determines the next question based on predefined rules and user input. This ensures adaptive survey flow and improves response relevance.

E. Database Layer
The database layer stores survey questions, branching rules, and user responses in a structured format. It uses PostgreSQL to manage and retrieve data efficiently. This layer ensures data consistency, reliability, and secure storage.

F. Admin Dashboard and Reporting
The admin dashboard enables survey management, rule configuration, and response monitoring. It provides statistical reporting and data visualization for analysis. This

module supports decision-making by presenting processed data in an organized and interpretable format.

The process begins when the user accesses the survey platform through the web interface. The user selects a survey and initiates the interaction, where the first question is displayed based on the predefined survey structure.

The user provides responses through the interface, which are immediately sent to the backend system for processing. The system performs validation checks to ensure that the input is complete, consistent, and in the correct format.

Once validated, the response is passed to the AI decision engine, which evaluates the input based on predefined branching rules. The system dynamically determines the next question, ensuring an adaptive and relevant survey flow.

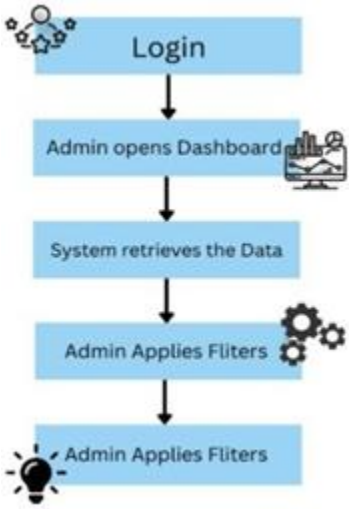


Figure 2: Data Flow Diagram Level 0

The responses are continuously stored in the database in a structured format, maintaining the association between questions and user inputs. This ensures data consistency and allows real-time tracking of survey progress.

After completing the survey, the collected data is processed and made available for analysis through the admin dashboard. The system generates summarized outputs and statistical reports, enabling efficient interpretation and decision-making.

IV. Implementation Details

A. System Development Environment

The proposed system is implemented as a web-based application that enables automated survey data cleaning and statistical reporting. The development environment includes modern web technologies for building an interactive user interface and efficient

backend processing. The frontend interface allows users to upload survey datasets, view processed results, and access statistical reports through a centralized dashboard.

B. Data Upload and Processing Module

The data upload module allows researchers to submit survey datasets collected from online questionnaires or digital forms. Once uploaded, the system performs initial validation to ensure that the dataset structure is correct and compatible with the processing pipeline. The uploaded data is then forwarded to the preprocessing module where automated data cleaning operations are performed.

C. Data Cleaning and Validation Module

The data cleaning module applies rule- based algorithms to detect and correct common data quality issues such as duplicate records, missing values, and inconsistent responses. Statistical methods are also used to identify outliers in numeric survey data. In addition, the system incorporates AI-assisted validation to analyze textual responses and detect inconsistencies within open-ended survey inputs.

D. Data Standardization and Storage

After cleaning and validation, the dataset undergoes standardization to ensure consistent formatting and structure. Text responses are normalized, categorical responses are unified, and numerical values are formatted appropriately. The processed dataset is then stored in the system database for further analysis and report generation. The system also maintains structured metadata and indexed records to enable efficient data retrieval, querying, and integration with the automated statistical reporting module.

E. Statistical Reporting Module

The statistical reporting module generates automated descriptive statistics from the cleaned dataset. This includes summary statistics, frequency distributions, and basic analytical insights that help researchers understand key patterns in survey responses. The generated reports allow users to quickly interpret data without performing manual statistical calculations.

F. Visualization and User Dashboard

The final implementation stage provides a web-based dashboard that displays processed survey data and generated reports. The dashboard presents analytical results using visual elements such as tables and statistical summaries, allowing researchers to easily evaluate the quality and outcomes of survey datasets. This interactive interface improves accessibility and supports efficient data-driven analysis.

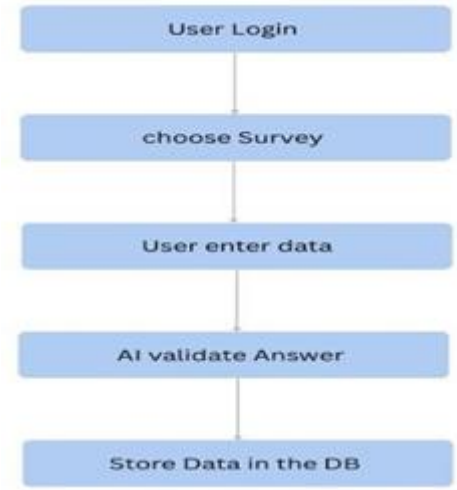


Figure 3: Data Flow Diagram Level 1

Overall, the implementation integrates user-friendly frontend interaction, automated backend processing, AI-assisted data cleaning, and statistical reporting into a unified survey analysis platform. The seamless flow of data from survey dataset upload to automated preprocessing and statistical report generation enables efficient and reliable research workflows. By combining rule-based validation, Large Language Model– assisted text analysis, and web-based dashboards, the system ensures scalability, accuracy, and accessibility for researchers and analysts. This integrated implementation significantly reduces manual effort in survey data preparation and improves the reliability of analytical insights derived from survey datasets.

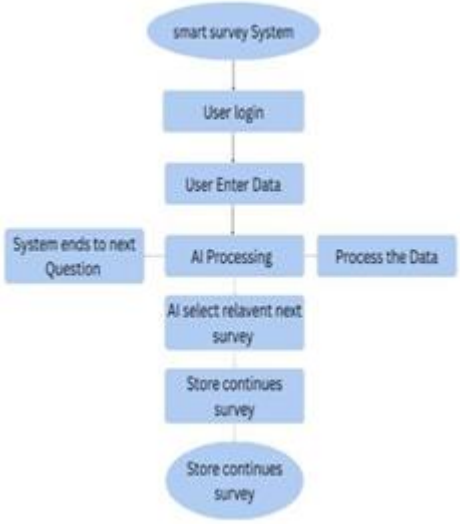


Figure 4: Data Flow Diagram Level 2

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 46-62

Figure 5: User Page

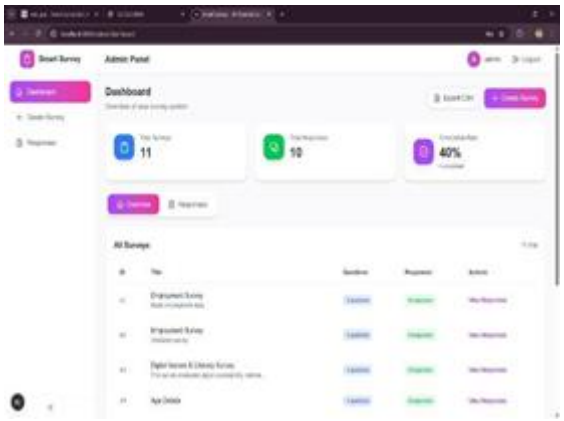
V. Output

A. User Page Output

The user interface displays the survey in a structured and interactive format, allowing users to respond to questions dynamically. Based on user input, the system adapts the flow of questions using branching logic to ensure relevance. The page also provides real-time validation and a smooth submission process for completing the survey.

B. Admin Page Output

The admin dashboard presents collected survey data in a structured and organized format. It allows administrators to view responses, manage surveys, and analyze results through tabular and statistical representations. The dashboard supports efficient monitoring and decision-making by providing clear insights into user responses.



Figurec: Admin Page

VI. Conclusion and Future Enhancement

The AI-assisted Survey Data Cleaning and Statistical Reporting Platform demonstrates how intelligent digital technologies can improve the reliability and efficiency of survey-

based research. By integrating rule-based preprocessing, Large Language Model–assisted text analysis, and automated statistical reporting, the system simplifies the preparation and analysis of survey datasets.

The proposed approach enables automatic detection of missing values, duplicate entries, inconsistent responses, and outliers, thereby improving data quality and reducing manual effort. The web-based interface and automated reporting features allow researchers to quickly interpret cleaned datasets and obtain meaningful insights for decision-making.

While the current system effectively supports automated data cleaning and statistical reporting, there is significant scope for future enhancement. Future improvements may include advanced analytics, improved natural language processing for open-ended responses, and integration with large-scale survey platforms to expand analytical capabilities.

Reference :-

1. Smith, J., & Anderson, R. (2022). AI-Based Adaptive Survey Systems for Dynamic Data Collection. International Journal of Artificial Intelligence Applications.
2. Kumar, S., & Patel, R. (2021). Cloud-Based Survey Management Platforms Using Modern Web Technologies. Journal of Cloud Computing Systems.
3. Ahmed, N., & George, P. (2020). Role-Based Authentication in Web Applications Using Secure Access Control Models. International Conference on Smart Data Security.
4. Li, W., & Chen, Y. (2021). Real-Time Data Synchronization in Distributed Web Applications. Journal of Web Engineering and Systems.
5. Sharma, V., & Singh, A. (2022). Implementation of RESTful APIs Using FastAPI for Scalable Backend Systems. International Journal of Software Engineering.
6. Brown, T., & Wilson, K. (2021). PostgreSQL for Structured Data Storage in Enterprise Applications. Database Management Review.
7. Lopez, M., & Williams, J. (2020). Cloud Computing Architectures for Large-Scale Data Collection Systems. Journal of Cloud Technologies.
8. Rahman, A., & Hussain, M. (2021). AI-Based Decision Engines for Intelligent Workflow Automation. IEEE Conference on Intelligent Systems.
9. Chen, L., & Wu, Y. (2022). Dynamic Form Generation and Validation in Modern Web Applications. Journal of Interactive Computing.
10. Thomas, L., & Raj, P. (2022). Integrating Machine Learning for Survey Data Classification and Prediction. Journal of AI Research and Applications.
11. Zhang, H., & Kim, S. (2023). Secure Data Storage and Encryption Models in Cloud-Based Applications. Journal of Cybersecurity and Cloud Systems.
12. Mehta, S., & Nair, D. (2021). Offline Data Synchronization Techniques for Web and Mobile Platforms. Journal of Distributed Computing Systems.