Automated Data Quality Checking & Cleaning Workflow

1. Introduction

1.1 Background

In data analytics workflows, incoming datasets often require repetitive and manual validation before analysis can begin. Analysts routinely check for missing values, duplicates, schema inconsistencies, and basic data validity issues using spreadsheets or ad-hoc scripts. This manual approach is time-consuming, inconsistent across analysts, and prone to human error, especially when datasets are received frequently from different sources.

As data volume and frequency increase, the lack of a standardized and automated data quality process leads to unreliable inputs flowing into dashboards, reports, and models, ultimately impacting decision-making quality.

1.2 Objective

The objective of this project is to design and implement an automated, analyst-friendly workflow that:

- Accepts raw CSV datasets through a simple form-based interface
- Automatically performs standardized data quality checks
- Safely cleans common and repetitive data issues
- Produces a cleaned dataset ready for analysis
- Generates a clear data quality summary for decision-making

2. Personas & Stakeholders

2.1 Primary Persona - Data Analyst

Responsibilities:

- Preparing datasets for analysis and reporting
- Validating incoming data quality
- Cleaning and standardizing data

Pain Points:

- Repeating the same quality checks for every dataset
- Manually identifying and fixing duplicates and missing values
- Lack of a consistent, documented process for data validation

2.2 Secondary Persona - Analytics Manager / Reviewer

Responsibilities:

- Ensuring data reliability across reports and dashboards
- Reviewing analyst outputs for correctness
- Maintaining data standards across teams

Pain Points:

- Limited visibility into how datasets were cleaned
- Inconsistent quality checks across analysts
- Difficulty auditing data preparation steps

3. Current State (AS-IS)

3.1 Workflow Overview

1. Analyst receives a CSV dataset
2. Opens the file manually in Excel / scripts
3. Checks for missing values and duplicates
4. Applies ad-hoc cleaning rules
5. Saves a new version of the dataset
6. Repeats the process for every new dataset

3.2 Pain Points by Persona

Data Analyst:

- High time investment in repetitive tasks
- Cognitive fatigue from repeated validations
- Lack of a reusable and standardized process

Analytics Manager:

- No standardized cleaning process
- Difficult to verify if data quality checks were performed
- Inconsistent outputs across analysts

4. Problem Statement

There is no standardized, automated mechanism to validate and clean incoming datasets before analysis. As a result, analysts repeatedly perform manual data quality checks, leading to inefficiencies, inconsistent outcomes, and a higher likelihood of errors propagating into analytical outputs and business decisions.

5. Future State Goals (TO-BE)

The future-state workflow aims to:

- Standardize data quality validation across datasets
- Automatically resolve safe and repetitive data issues
- Clearly separate cleaned data from quality metadata
- Maintain transparency and human review for risky cases
- Improve analyst productivity and confidence in data inputs

6. TO-BE Workflow

1. Analyst uploads a CSV file via a form interface
2. The system automatically parses the dataset
3. Data quality checks are executed
4. Safe issues are automatically cleaned
5. A data quality score and decision are generated
6. A cleaned dataset is produced for download
7. A summary is displayed to the user

7. Outputs

The workflow produces two clearly separated outputs to ensure clarity and usability. The first output is a cleaned dataset in CSV format that contains only validated and standardized rows. This dataset is immediately ready for downstream analysis, reporting, or dashboarding without requiring additional manual preprocessing.

The second output is a data quality report that summarizes the validation process. This report includes the number of rows before and after cleaning, the count of duplicate records removed, a list of identified data quality issues, and an overall quality score with a final approval or review decision. Keeping the cleaned dataset and the quality report separate avoids confusion between analytical data and metadata, and improves transparency and auditability.

8. Tooling & Implementation

The workflow is implemented using n8n as the primary automation and orchestration platform. n8n handles the end-to-end process including CSV parsing, data quality validation, safe auto-cleaning, quality scoring, and final output generation. The n8n Form Trigger is used to provide a simple and structured interface for uploading datasets and initiating the workflow, enabling a smooth analyst experience without requiring external tools or APIs.

Lovable is used as a prototyping tool to design and visualize the analyst-facing user interface. It helps demonstrate how users would interact with the system, while the actual execution logic and data processing remain fully implemented within n8n.

9. Role of Lovable (Prototype)

Lovable was used to design a lightweight, analyst-facing interface that illustrates the intended user flow of the automated data quality system. The prototype focuses on three core interactions: uploading datasets, viewing data quality results, and downloading the cleaned dataset. It serves as a visual and conceptual layer that communicates product intent and usability, while all data validation, cleaning, and scoring logic is executed entirely within the n8n workflow.

10. Expected Impact

By automating repetitive data quality checks and safe cleaning steps, the workflow significantly reduces the time analysts spend on manual data preparation. It establishes consistent and auditable data quality standards across datasets, improving trust in analytical outputs. As a result, analysts can move faster from data intake to insight generation, enabling quicker and more reliable decision-making.

11. Assumptions & Constraints

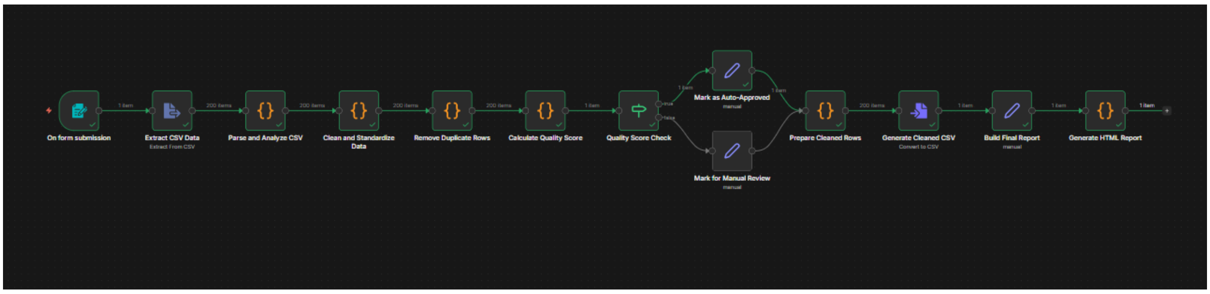
- Input datasets are provided in CSV format
- Cleaning rules are generic and non-domain-specific
- Safe issues can be auto-cleaned, while risky issues require human judgment
- No external databases or storage systems are integrated
- The workflow is designed for internal analyst usage and decision support

How AI Helped Structure the Thinking

AI tools were used to help structure the workflow, identify logical decision points, and ensure clarity in separating data quality checks, cleaning logic, and final outputs. Core data cleaning logic remains rule-based and explainable.

N8n

https://github.com/Adityabeecha/automated-data-quality-checker


On form submission
1 item
Extract CSV Data
Extract From CSV
200 items
Parse and Analyze CSV
200 items
Clean and Standardize Data
200 items
Remove Duplicate Rows
200 items
Calculate Quality Score
1 item
Quality Score Check
true
false
1 item
Mark as Auto-Approved
manual
Mark for Manual Review
manual
1 item
Prepare Cleaned Rows
200 items
Generate Cleaned CSV
Convert to CSV
1 item
Build Final Report
manual
1 item
Generate HTML Report
1 item