Analytics Validation & Safe Insight Generation System

Yugaash Sridhar

Department of Computer Science and Engineering
IIITDM Kanchipuram, India

Abstract

In data-driven environments, analytics pipelines frequently operate directly on raw datasets without sufficient validation of data quality and structure. This practice often leads to misleading insights and unreliable decision-making. This paper presents a data-quality–first analytics validation system that evaluates the structural reliability of structured datasets before generating insights. The proposed approach introduces layered validation, interpretable health scoring, and safe, rule-based insight generation to prevent incorrect analytics. The system is domain-agnostic, scalable across diverse datasets, and designed using conservative validation logic.

Index Terms

Data Quality, Analytics Validation, Data Profiling, Health Score, Safe Analytics

I. Introduction

Analytics systems play a crucial role in modern decision-making across industries. However, most analytics tools assume that input datasets are analytically valid and directly generate summaries and visualizations. In real-world scenarios, datasets often contain identifiers, ambiguous numeric fields, unconverted date attributes, missing values, and constant columns. These structural issues can significantly distort analytical outcomes.

This work proposes a validation-first analytics framework that prioritizes correctness over aggressive insight generation.

II. Proposed Methodology

A. Data Ingestion

The system accepts structured CSV datasets without enforcing domain-specific assumptions.

B. Data Profiling

Profiling computes dataset structure, schema, and missing values.

C. Data Quality Validation

Rule-based checks detect identifiers, ambiguous identifiers, date-like columns, and constant attributes.

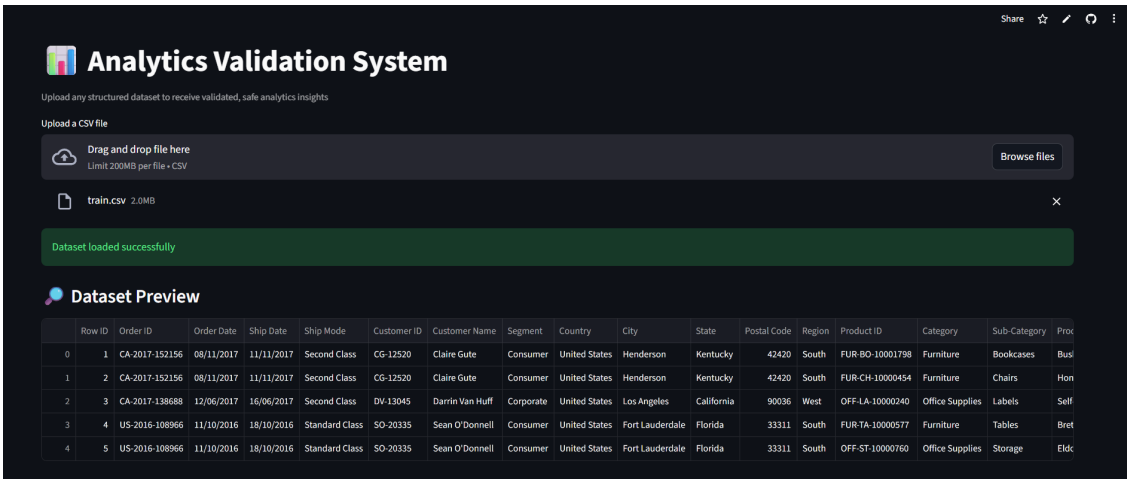

Fig. 1. Analytics Validation System – CSV upload interface

Fig. 2. Dataset preview and automatically detected schema for train.csv

D. Data Health Scoring

A Data Health Score (0–100) quantifies dataset reliability.

E. Safe Insight Generation

Insights are generated only from validated attributes.

III. Visualization Strategy

Only safe insights are visualized.

IV. Results and Discussion

Trend analysis is intentionally blocked when date columns are not validated.

Fig. 3. Data profiling summary showing rows, columns, and missing values

Fig. 4. Rule-based data quality checks identifying unsafe analytical columns

Fig. 5. Computed Data Health Score (77/100) with detected risk factors

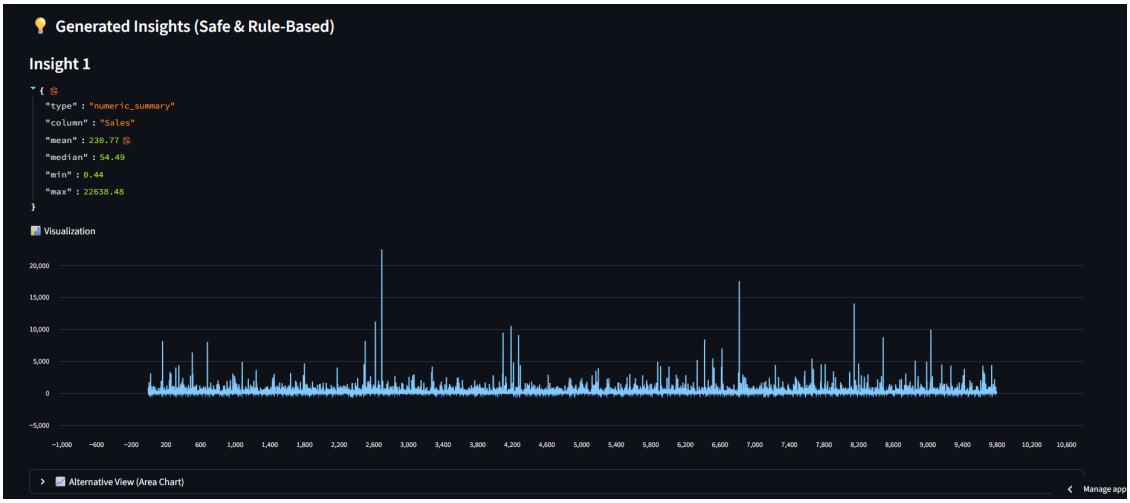

Fig. 6. Rule-based safe insights generated from validated columns

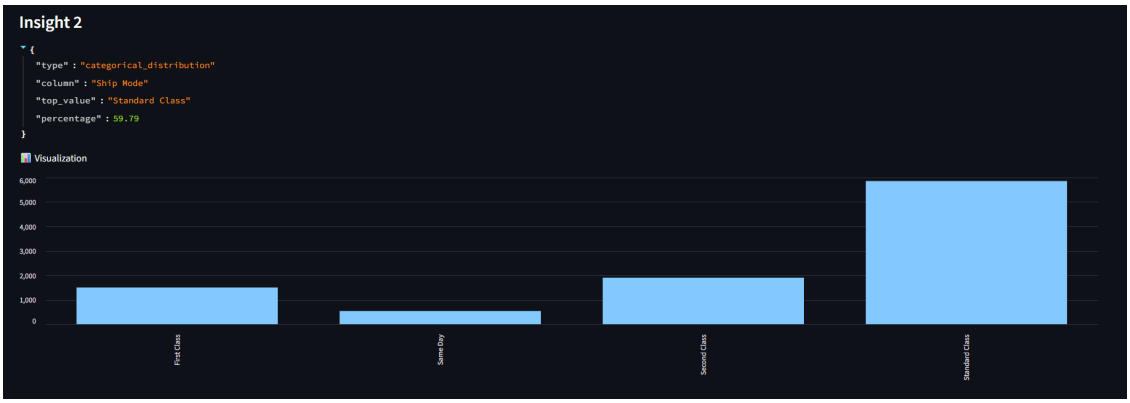

Fig. 7. Numeric sales insight visualized using line and area charts

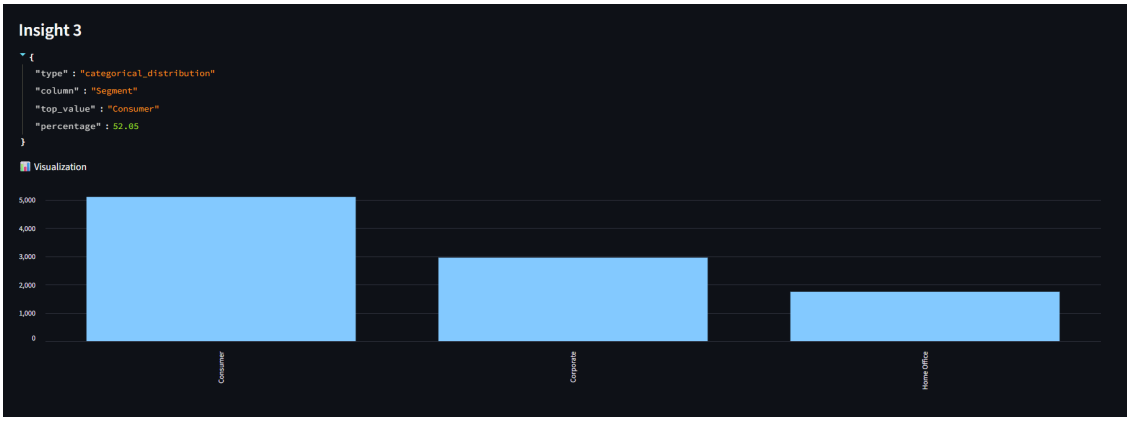

Fig. 8. Categorical distribution visualized using bar charts

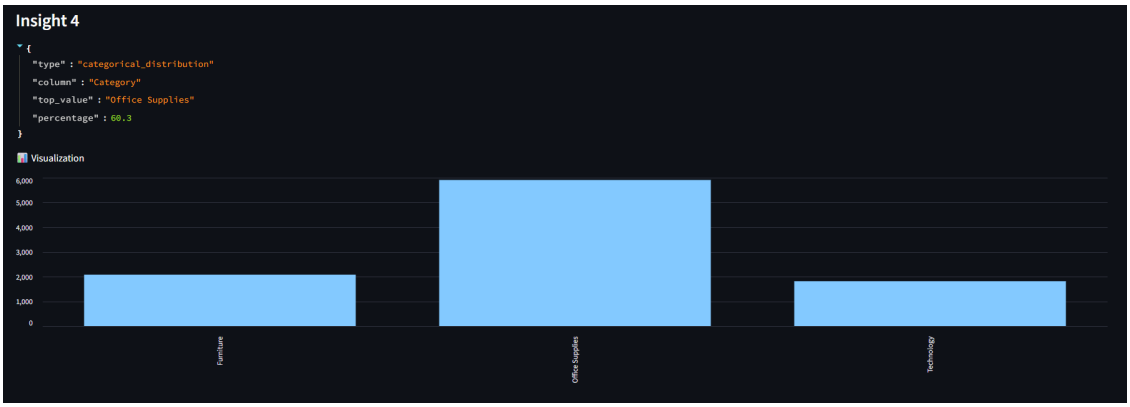

Fig. 9. Group comparison visualized using bar and radar charts

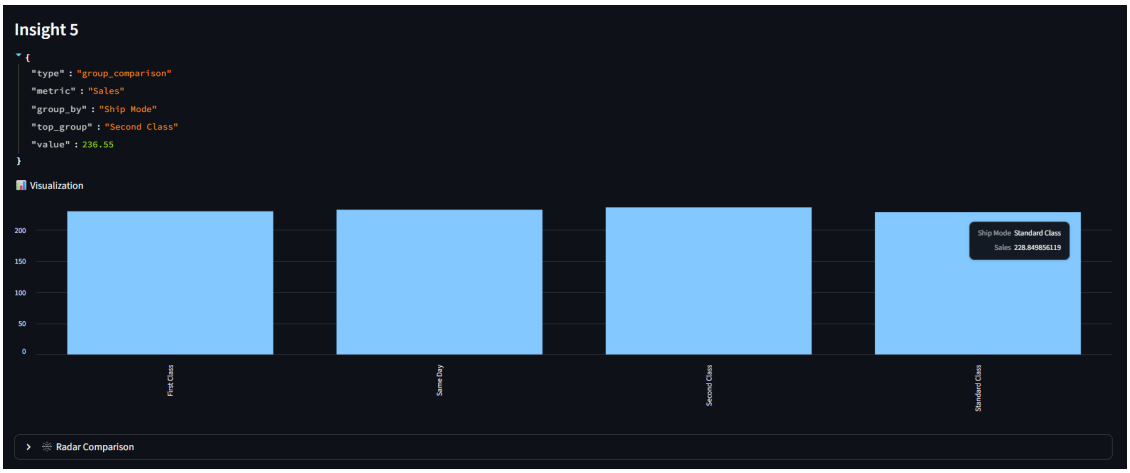

Fig. 10. Trend readiness insight indicating detected but unconverted date fields

The validation-first approach significantly improves trust in analytics outputs.

V. Conclusion

This paper presented a data-quality–first analytics validation system that prevents misleading analytics by validating datasets before insight generation. The approach improves correctness, transparency, and trustworthiness in data-driven decision-making.

References

[1] T. Redman, *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2001.
[2] ISO/IEC 25012:2008, *Software Engineering—Software Product Quality Requirements and Evaluation (SQuaRE)*.
[3] A. Abedjan et al., "Detecting Data Errors: Where Are We and What Needs to Be Done?" *VLDB Endowment*, 2016.