

InsightForge – Automated Statistical Analysis & Report Generator

Aastha Thakur¹, Rohan. Mehta²

Department of Computer Science and Engineering

Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India

Abstract—Exploratory statistical analysis of tabular datasets is a routine but time-consuming task for students, researchers, and analysts, typically requiring manual selection of appropriate statistical tests, repeated switching between analysis and documentation tools, and hand-crafted report writing to communicate findings. This paper presents InsightForge, an automated statistical analysis and report generation tool that ingests tabular datasets (CSV/XLSX), performs automated exploratory data analysis and appropriate statistical test selection based on variable types and distributions, and generates a structured, human-readable analysis report combining summary statistics, visualizations, and plain-language interpretation of results. The system was implemented using a Python-based analysis pipeline for descriptive statistics, normality and hypothesis testing, and correlation analysis, coupled with an automated report-generation module producing formatted PDF/HTML output. InsightForge was evaluated on a set of benchmark datasets of varying size and variable composition, comparing automated test selection and output against manual analysis performed by a trained analyst, and measuring report-generation time and user-rated report clarity. Results show that automated test selection matched the manual analyst's chosen statistical test in 91.3% of evaluated cases, average end-to-end report generation time was under 12 seconds for datasets up to 50,000 rows, and sampled users rated generated report clarity favorably with an average score of 4.3 out of 5. These findings indicate that automated statistical analysis and report generation can substantially reduce the time and expertise burden of routine exploratory data analysis while maintaining analytical soundness comparable to manual practice.

Index Terms—Automated Statistical Analysis, Report Generation, Exploratory Data Analysis, Hypothesis Test Selection, Data Visualization, Natural Language Summarization, Python, Data Science Tools.

I. INTRODUCTION

Exploratory statistical analysis, a foundational step in data-driven research and business analytics, typically requires an analyst to compute descriptive statistics, select and apply appropriate hypothesis tests based on variable types and data distributions, generate supporting visualizations, and finally compile these findings into a coherent written report. This workflow, while conceptually well established, is time-consuming and requires statistical expertise that is not uniformly available to all students and early-career analysts.

Automating the analysis-selection and report-generation stages of this workflow can substantially reduce the time and expertise burden of routine exploratory analysis, while offering a consistent starting point that reduces the likelihood of applying an inappropriate statistical test. Prior automated statistical tools have generally focused narrowly on either analysis execution or visualization, with fewer tools additionally automating the synthesis of results into a structured, readable report.

This paper presents InsightForge, a tool that combines automated exploratory data analysis,

B. Automated Statistical Test Selection

Rule-based approaches to statistical test selection, generally based on variable type (categorical/continuous), number of groups being compared, and distributional assumptions such as normality, have been explored in decision-support literature as a means of guiding non-expert users toward statistically appropriate test choices.

III. SYSTEM DESIGN

A. Pipeline Architecture

InsightForge is implemented as a four-stage pipeline: (1) data ingestion and cleaning, (2) automated exploratory analysis and variable-type classification, (3) statistical test selection and execution, and (4) report synthesis and formatted output generation, as summarized in Table I.

B. Automated Test Selection Logic

The test-selection module applies a rule-based decision tree considering variable type, number of comparison groups, and a preliminary Shapiro-Wilk normality check, selecting among common tests such as

statistical test selection, and structured report generation into a single pipeline, and evaluates its analytical accuracy, processing speed, and report clarity against manual analysis practice.

II. LITERATURE REVIEW

A. Automated Exploratory Data Analysis Tools

Existing automated EDA tools, including open-source libraries that generate descriptive-statistics and visualization reports from tabular data, have demonstrated the feasibility of automating the summary-statistics and visualization stages of analysis, though most such tools stop short of automated hypothesis-test selection or narrative interpretation.

t-test, ANOVA, chi-square, and their non-parametric counterparts (Mann-Whitney U, Kruskal-Wallis), along with Pearson or Spearman correlation as appropriate.

C. Report Generation Module

The report-generation module combines computed statistics, auto-generated visualizations (histograms, boxplots, correlation heatmaps), and templated plain-language interpretation text into a structured PDF/HTML report, with statistical results embedded alongside brief natural-language explanations of what each result indicates.

TABLE I — InsightForge Pipeline Stages

Stage	Function
Data Ingestion & Cleaning	Import CSV/XLSX, handle missing values, detect variable types
Exploratory Analysis	Compute descriptive statistics and generate summary visualizations
Test Selection & Execution	Select and run appropriate statistical test(s) based on data characteristics
Report Synthesis	Generate structured PDF/HTML report with results and plain-language interpretation

IV. IMPLEMENTATION AND EVALUATION METHODOLOGY

The analysis pipeline was implemented in Python using standard statistical and data-manipulation libraries, with the report-generation module built to output formatted PDF and HTML reports from a templated document structure populated with computed results and visualizations.

Evaluation was conducted on a set of ten benchmark datasets ranging from 200 to 50,000 rows and varying in categorical/continuous variable composition. Automated test selection was compared against test choices made independently by a trained statistical analyst reviewing the same datasets, and report-generation time was measured across dataset sizes. A sample of student and analyst users rated the clarity of generated reports on a 5-point scale.

V. RESULTS AND DISCUSSION

Automated test selection matched the manual analyst's chosen test in 91.3% of evaluated cases across the ten benchmark datasets, with most disagreements arising

in borderline normality cases where the automated normality threshold and analyst judgment diverged. Report-generation time scaled sublinearly with dataset size, remaining under 12 seconds for datasets up to 50,000 rows, as summarized in Table II. Sampled users rated generated report clarity favorably at an average of 4.3 out of 5, with particularly positive feedback on the plain-language interpretation accompanying each statistical result, summarized in Table III.

VI. CONCLUSION

This paper presented InsightForge, an automated statistical analysis and report-generation tool combining exploratory analysis, rule-based test selection, and structured narrative reporting. Evaluation showed strong agreement with manual analyst test selection (91.3%), fast report generation across dataset sizes, and favorable user-rated report clarity, supporting the practical value of automated analysis tools in reducing the time and expertise burden of routine exploratory statistical work.

TABLE II — Report Generation Time by Dataset Size

Dataset Size (rows)	Avg. Report Generation Time (s)
---------------------	---------------------------------

Dataset Size (rows)	Avg. Report Generation Time (s)
200	1.8
2,000	3.4
20,000	7.6
50,000	11.7

TABLE III — Summary of Evaluation Results

Metric	Result
Agreement with manual analyst test selection	91.3%
Max report generation time (50,000 rows)	11.7 s
Average user-rated report clarity (out of 5)	4.3

VII. REFERENCES

- [1] McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.
- [2] Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications.
- [3] Shapiro, S.S., and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3-4), 591-611.
- [4] Wickham, H., and Grolemund, G. (2017). R for Data Science. O'Reilly Media.
- [5] Virtanen, P., Gommers, R., Oliphant, T.E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261-272.
- [6] Montgomery, D.C., and Runger, G.C. (2018). Applied Statistics and Probability for Engineers (7th ed.). John Wiley & Sons.
- [7] Few, S. (2009). Now You See It: Simple Visualization Techniques for Quantitative Analysis. Analytics Press.