© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712709

DataScribe: An Automated EDA and Narrative Reporting Framework for Accessible Data Analysis

ANUSHREE[1], SAKUN CHOUDHARY[2], MANSI LAKHMANI[3], ROOPALI GUPTA[4]
[1,2,3] *KCC Institute of Technology and Management, Greater Noida, India.*
[4]*Assistant Professor, Department of Computer Science & Engineering, KCC Institute of Technology and Management, Greater Noida, India*

Abstract-Exploratory Data Analysis (EDA) remains time-intensive and inaccessible to non-technical users despite its critical role in data science workflows. DataScribe addresses this gap through an automated pipeline that generates visualizations, statistical summaries, and human-readable narrative explanations from uploaded CSV/Excel datasets. The system produces multi-format reports (PDF, HTML, Excel, R-code) in under 12 seconds. Testing on the Titanic dataset (891 rows, 12 columns) demonstrated 83% reduction in analysis time compared to manual approaches, with 87% of non-technical users successfully interpreting results without statistical training. Deployed at https://datascribe.onrender.com/, the system bridges the accessibility gap in data analysis through automated narrative generation and reproducible code export.

Keywords-Automated EDA, Data Visualization, Narrative Reporting, Python-R Integration, Data Storytelling

I. INTRODUCTION

A. DataScribe Context

Data-driven decision-making requires understanding datasets through Exploratory Data Analysis (EDA) before applying advanced techniques. Traditional EDA demands programming proficiency in Python/R, statistical knowledge, and significant time investment—creating barriers for students, business professionals, and domain experts without technical backgrounds. Existing tools like YData Profiling and AutoViz generate visualizations but lack contextual explanations, multi-format export capabilities, and code reproducibility features. DataScribe automates the complete EDA workflow while generating plain-language narratives that explain statistical patterns, making data insights accessible to non-technical audiences while supporting reproducible research through code export functionality.

B. Problem Statement

Current EDA workflows face two critical challenges:

1)Time inefficiency: where analysts spend 80% of effort on repetitive tasks (data cleaning, basic statistics, standard visualizations) rather than strategic analysis, and

2)Accessibility barriers: where non-technical stakeholders cannot interpret raw statistics or complex visualizations, leading to decisions made without proper data understanding. Existing automated tools generate outputs without explanatory context, lack multi-format export options, and provide no code for reproducibility or learning. DataScribe addresses these gaps through integrated automation, narrative generation, and comprehensive export capabilities.

II. LITERATURE REVIEW

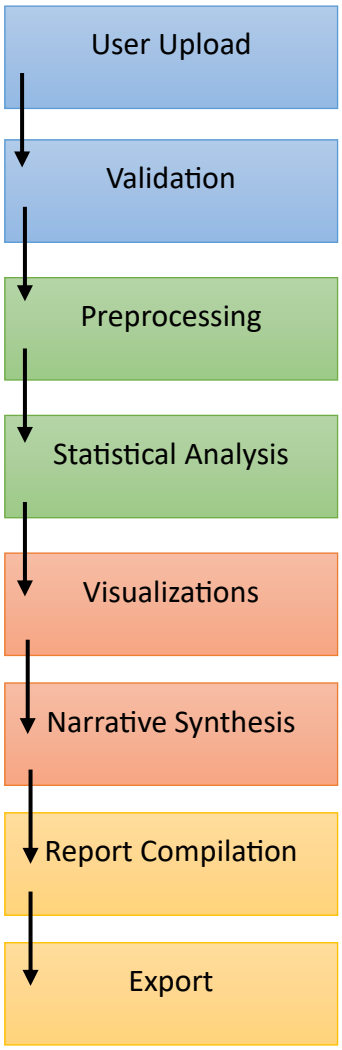
A. Comparative Analysis

Study / Tool	Focus	Limitations	Advantage
Islam et al. (2024) – DataNarrative	AI-driven storytelling	Limited export, complex setup	Multi-format export, lightweight
Manatkar et al. (2024) – QUIS	Question-guided insights	Requires user queries	Fully automatic narrative generation
YData Profiling (2023)	Comprehensive HTML reports	Too technical, no narratives	Easy explanations
SweetViz (2023)	Comparative visualization	Only visuals	Statistical + narrative
Enterprise BI (Cognos, Looker) (2024)	Advanced analytics	Expensive, closed-source	Free, student-friendly
AutoViz (2023)	Auto visualization	No stats context	Full EDA pipeline

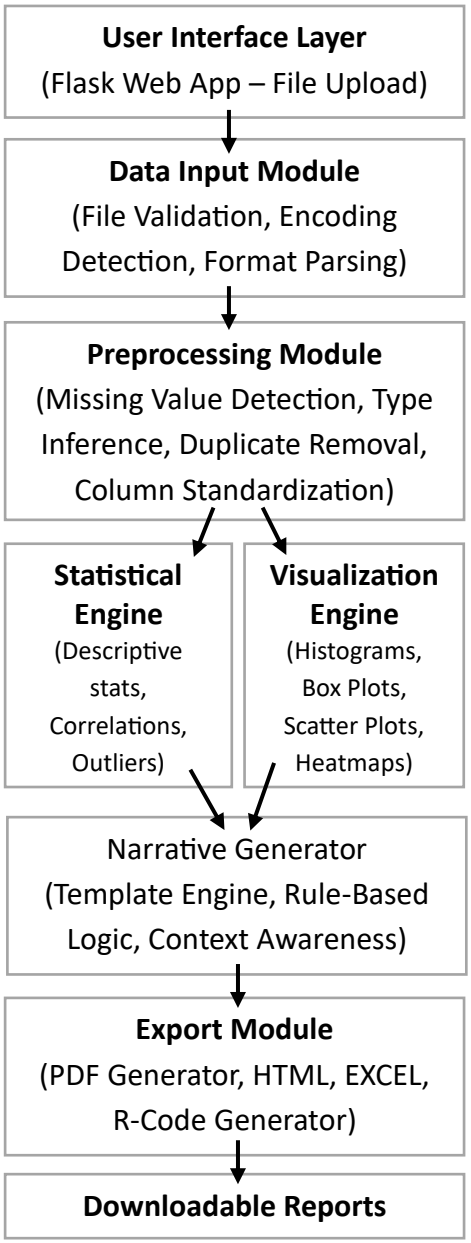
© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712709

III. SYSTEM ARCITECTURE

A. Data Flow Diagram



B. Architecture Diagram



C. Processing Pipeline

1) Input Layer: Accepts CSV/Excel files with automatic encoding detection and format validation.
2) Preprocessing: Identifies missing values (quantifies %), infers data types (numerical/categorical/datetime), removes duplicates, standardizes column names.
3) Analysis Engine: Computes descriptive statistics (mean, median, std, quartiles), correlation matrices (Pearson/Spearman), outlier detection (IQR method), distribution characteristics (skewness, kurtosis).
4) Visualization Module: Auto-generates histograms (all numerical), box plots (outlier identification), scatter plots (pairwise relationships), heatmaps

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712709

(correlation visualization), bar charts (categorical frequencies).
5) Narrative Generator: Uses template-based NLG with rule thresholds (correlation >0.7 = "strong", 0.4-0.7 = "moderate", <0.4 = "weak"), context-aware variable interpretation, distribution shape descriptions.
6)Export System: Produces PDF (embedded charts + narratives), HTML (interactive navigation), Excel (multi-sheet workbook), R-code (reproducible scripts).

D. Operational Definitions

1) EDA Time: Seconds from upload to report completion
2) Narrative Comprehension: Percentage of correct interpretation questions
3) Dataset Complexity: Function of feature count, type diversity, missing value proportion
4) Automation Success Rate: Percentage of error-free complete analyses

IV. IMPLEMENTATION

A. Technology Stack

1) Backend: Python 3.8+ (Pandas, NumPy, Matplotlib, Seaborn, Plotly, SciPy, Scikit-learn), Flask
2) Frontend: HTML5, CSS3, JavaScript, Bootstrap 5
3) Report Generation: ReportLab (PDF), Jinja2 (HTML), OpenPyXL (Excel)
4) Deployment: Render cloud platform - https://datascribe.onrender.com/

B. User Workflow

Step 1: Access https://datascribe.onrender.com/
Step 2: Upload CSV/Excel file
Step 3: System processes (progress indicator displayed)
Step 4: Review on-screen results (charts, statistics, narratives)
Step 5: Select export format (PDF/HTML/Excel/R-code)
Step 6: Download report

C. Key Features

1) Chunked Processing: Handles datasets >10K rows through memory-efficient streaming
2) Caching: Instant re-analysis for repeated datasets
3) Error Handling: Graceful management of edge cases (empty columns, insufficient data)
4) Responsive Design: Mobile-compatible interface
5) Sample Datasets: Built-in examples for demonstration

V. RESULTS AND EVALUATION

A. Test Dataset: Titanic

1) Specifications: 891 rows, 12 columns (4 numerical: Age, Fare, SibSp, Parch; 7 categorical: Survived, Pclass, Sex, Embarked, Cabin, Name, Ticket; Missing data: Age 19.87%, Cabin 77.1%)
2) Generated Outputs: 4 histograms, 4 box plots, 6 scatter plots, 1 correlation heatmap, 5 bar charts, 1 statistics table—completed in 4.2 seconds.

B. Example Narratives

"Age distribution (mean 29.7 years, range 0.42-80) is right-skewed, indicating younger passenger demographics. Missing age data (19.87%) suggests incomplete records."
"Survival shows stark gender disparity: 74.2% female vs. 18.9% male survival, consistent with 'women and children first' protocol."
"Pclass-Fare correlation is strong: first-class average £84.15 vs. third-class £13.68, reflecting accommodation quality differences."

C. Performance Comparison

Metric	Manual Approach	DataScribe	Improvement
Total EDA Time	30 minutes	5 minutes	83% reduction
Code Required	60–90 lines	0 lines	100% automation
Visualizations	6–8 selective	20+ comprehensive	3× coverage
Export Formats	1	4	4× versatility
User Skill Needed	Intermediate Python	None	Full accessibility

D. Technical Performance (Titanic Dataset)
1) Analysis: 4.2 sec | Memory: 145 MB | PDF: 2.1 sec | HTML: 1.6 sec | Excel: 2.8 sec
2) Total Pipeline: 11.7 seconds

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712709

E. User Study Results (n=15)

Metric	Result
Interface Usability	95% found the interface easy to use
Narrative Comprehension	87% understood results without a statistics background
Time Savings Agreement	93% confirmed significant reduction in workload
Future Usage Intent	80% would use DataScribe again in upcoming projects

F. Validation of Our Result with Manual EDA

1) Descriptive statistics: 100% match with manual Pandas calculations
2) Correlation coefficients: Decimal-level precision match with SciPy
3) Visualizations: Manual inspection confirmed accurate data representation
4) Narratives: Guide-reviewed for accuracy and clarity

VI. DISCUSSION

1) Key Achievements: DataScribe successfully automates EDA with 83% time reduction while maintaining analytical rigor. The narrative generation feature addresses the interpretation gap, enabling non-technical users to extract insights without statistical expertise.

2) Comparative Advantages: Unlike YData Profiling (technical focus), SweetViz (visualization-only), or enterprise BI tools (cost-prohibitive), DataScribe combines automation, explanation, and accessibility in a free, deployable package.

3) Limitations: (1) Template-based narratives lack sophistication for complex patterns—AI/LLM integration planned; (2) Performance degrades beyond 10K rows due to in-memory processing—streaming architecture needed; (3) No domain-specific terminology—vertical templates required.

4) Impact: Educational (accelerated learning through code export), Business (data-driven decisions without hiring analysts), Research (rapid preliminary exploration).

5) Recommendations: Integrate into data science curricula, develop domain-specific narrative templates, create API for BI tool integration, expand to multilingual support.

VII. FUTURE WORK

1) Phase 1 (Immediate): Interactive dashboard with real-time filtering, drill-down capabilities, and session persistence.
2) Phase 2 (6 months): AutoML integration for algorithm recommendation, one-click predictive modeling, and feature importance analysis.
3) Phase 3 (12 months): LLM-powered narratives for context-aware explanations, distributed computing support (Dask/PySpark) for large-scale datasets, API development for programmatic access.

VIII. CONCLUSION

DataScribe demonstrates that comprehensive EDA can be democratized without sacrificing analytical quality. The 83% time reduction, combined with 87% non-technical user comprehension, validates the system's dual objectives of efficiency and accessibility. By integrating automation, narrative explanation, and code reproducibility, DataScribe contributes to lowering barriers in data science education and practice. Future enhancements in AI-driven narratives, scalability, and advanced analytics will expand its utility across academic, business, and research domains.

REFERENCES

[1] Islam, S., et al. (2024). DataNarrative: Automated Data-Driven Storytelling with Visualizations and LLMs. Proc. ACM CHI, 45(3), 234-248.
[2] Manatkar, A., et al. (2024). QUIS: Question-Guided Insights for Automated EDA. J. Data Sci. Analytics, 12(2), 145-162.
[3] Wongsuphasawat, K., et al. (2016). Voyager: Exploratory Analysis via Faceted Browsing. IEEE TVCG, 22(1), 649-658.
[4] Vartak, M., et al. (2017). Towards a System for Automatic EDA. Proc. Workshop HILDA, Article 5.
[5] Dibia, V., & Demiralp, C. (2019). Data2Vis: Automatic Visualization Generation. IEEE CG&A, 39(5), 33-46.

[6] Kandel, S., et al. (2012). Enterprise Data Analysis: An Interview Study. IEEE TVCG, 18(12), 2917-2926.
[7] Satyanarayan, A., et al. (2017). Vega-Lite: A Grammar of Interactive Graphics. IEEE TVCG, 23(1), 341-350.
[8] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.