International Journal of Sciences and Innovation Engineering
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: www.ijsci.com
Vol.03 No.01
E-ISSN : 3049-0251
DOI: https://doi.org/10.70849/ijsci0301260209545

Data Edifire: An Approach to Clean, Visualize and Analyze Data

Protyasha Mukherjee[1], Sayantan Chakrabarti[2], Md Abdul Aziz Al Aman[3], Israr Arif[4], Ritikesh Singh[5], Akash Kumar Das[6], Pilli Shreyash Rao[7]

[1] *Assistant Professor, Department of IT, B P Poddar Institute of Management and Technology, Kolkata, West Bengal, India*

[2] *Assistant Professor, Department of CSE, RCC Institute of Information Technology, Kolkata, West Bengal, India*

[3] *Assistant Professor, Department of CSE AI, Techno India University, Kolkata, West Bengal, India.*

[4,5,6,7] *Student, Department of IT, B P Poddar Institute of Management and Technology, Kolkata, West Bengal, India*

Article Info

Article History:

Published:09 Jan 2026

Publication Issue:
Volume 3, Issue 01
January-2026

Page Number:
168-175

Corresponding Author:
Sayantan Chakrabarti

Abstract:

In today's fast-paced, data-driven world, making sense of raw information isn't always easy—especially for those without a technical background. That's where Data Edifire comes in. It's a Streamlit-based web app built to take the complexity out of data pre-processing and visualization, offering users a clean, intuitive interface to explore their datasets with ease. Whether users are working with real-world or sample CSV files, Data Edifire helps them clean, prepare, and understand their data. It handles common tasks like filling in missing values, encoding labels, detecting outliers, and even applying regression and clustering techniques. Behind the scenes, the app taps into the power of Python libraries like Pandas, Plotly, Seaborn, and Scikit-learn—bringing professional-grade tools to anyone, regardless of experience level. To plan the development realistically, we applied the Constructive Cost Model (COCOMO), identifying the Semi-Detached mode as best suited for beginner team. We also conducted surveys and market research, which highlighted a growing need for simplified, flexible data tools—many current platforms are either too complex or too rigid. Ultimately, Data Edifire is about more than just data—it's about empowerment. It's built to grow with its users, encouraging learning, collaboration, and exploration. With future updates aimed at automation and broader file support, we hope to make data science feel not just accessible, but exciting for everyone.

Keywords: Python-Powered Analytics, Data Cleaning & Preparation of Dashboards, Non-Experts Automated Analysis Workflow, Data Security

1. Introduction

In the era of data-driven decision-making, the ability to extract meaningful insights from raw data is more important than ever. However, this process often requires technical expertise, which creates a barrier for students, beginners, and non-technical users. Many existing tools for data analysis and visualization are either too complex, lack educational context, or require switching between multiple platforms and resources to understand even basic data operations [1-2,3,5]. Data Edifire is designed to bridge this gap [4,6,7-9]. It is a Streamlit-based web application that simplifies the process of data pre-processing and visualization through an intuitive, beginner-friendly interface [8,9,11].

Copyright © 2025 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Inspired by Edgar Dale's Cone of Experience—which emphasizes active learning through direct experience—our tool empowers users to not only view data but also interact with it meaningfully [5-6]. From handling missing values to applying clustering or regression techniques, users are guided step-by-step with in-context help and educational prompts [10-11]. Unlike traditional tools, Data Edifire integrates both theoretical knowledge and practical execution in a single interface [3-5,7]. This ensures that users don't just perform tasks—they understand them [1-3]. The application supports both synthetic and real-world datasets and aims to be scalable, cost-effective, and accessible, especially in resource-constrained environments [4-6]. By combining technical functionality with educational value, Data Edifire transforms the data learning experience into something more engaging, inclusive, and effective for users of all skill levels [9-11]. The further enhancement can be focused on the following objectives: -

1. To develop a user-friendly web application that simplifies data pre-processing and visualization for users with minimal technical expertise.
2. To integrate widely-used Python libraries (Pandas, Plotly, Seaborn, Scikit-learn) for efficient data analysis and interactive chart generation.
3. To address common limitations in existing data tools by offering customizable, accessible, and visually engaging features.
4. To evaluate the feasibility and scalability of the application using COCOMO-based effort estimation and use feedback through surveys.

2. Literature Review

In the age of information, navigating vast volumes of data is both a necessity and a challenge [10]. Traditional reports often overwhelm users, hiding valuable insights under complex layers [9]. Data Edification addresses this gap by offering a platform that presents data in clear, engaging, and accessible visual formats—catering to both beginners and experienced users [11]. The foundation of data visualization lies in the pioneering work of John Tukey, who introduced Exploratory Data Analysis (EDA) in 1977 to uncover data patterns through visual techniques [3]. Colin Ware's studies emphasized how visual perception influences our understanding, highlighting the role of design in effective communication [4]. William Cleveland further advanced the field by identifying best practices in visual encoding, emphasizing clarity through thoughtful use of colour, scale, and layout [5,6]. Visualization simplifies complexity, revealing patterns and trends that raw data often conceals [10-11]. It enhances decision-making, supports collaboration, and fosters better comprehension [1-2]. Its impact spans various sectors—from identifying business trends and improving healthcare outcomes to informing public policy and enhancing education [3-4]. Thanks to these foundational contributions, data visualization has evolved into a crucial tool for making data actionable [7-8]. As datasets continue to expand in size and complexity, the need for intuitive, interactive visual tools like Data Edification becomes even more critical in helping users make informed, data-driven decisions [1-3].

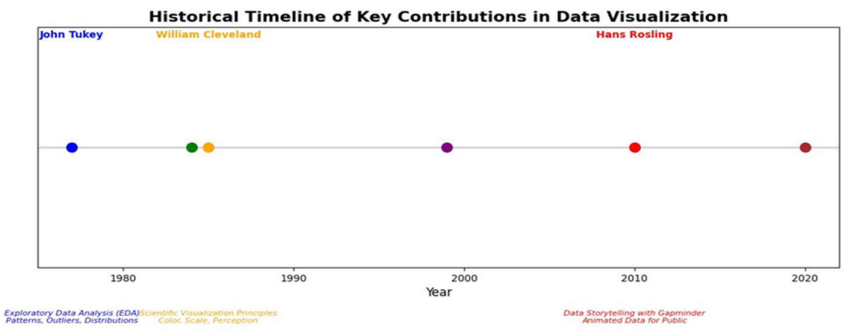

Fig. 1 Graphical representation of Key contribution in Data Visualization [2]

Fig. 2 Workflow model [6]

Data Upload – The upload of the CSV file into the application [2-4].
Data Cleaning – The process of detecting and correcting corrupt, duplicate, or invalid entries to improve dataset accuracy and reliability [1-2].
Data Pre-processing – The transformation and normalization of cleaned data – such as imputation, encoding, scaling, or feature extraction, preparing it for effective visualization and analysis [5-7].
Data Visualization – Converting prepared data into intuitive graphical forms (charts, maps, dashboards) that reveal patterns, trends, and insights [1,4,5].

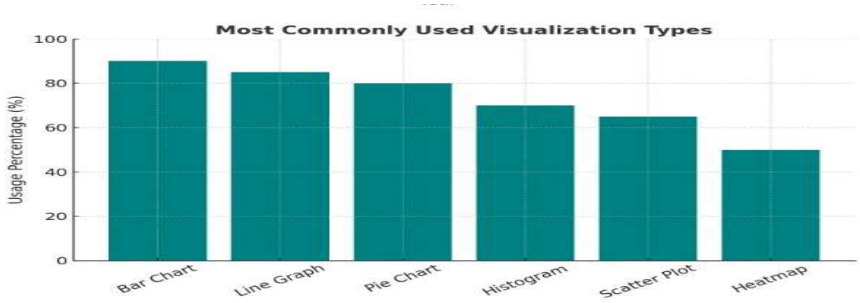

Fig. 3 Application of Data Visualization Model [4]

Missing Data and Data Imputation - There are several ways to deal with missing data, which may arise due to various reasons. Right now, our system removes the entire row if any information is missing [1-3]. This, however, has its own drawbacks. We are working on filling missing fields with the best possible approach - mean, median (middlemost value), mode (most frequent value), k-nearest neighbours or with regression analysis [7-9]. It's important to note that both removing missing data and filling it in can slightly affect how the data is visualized. While these changes might be too small to notice by eye, we'll still let the user know what we've done so they can make informed decisions while we maintain transparency [10-11].
Duplicate Data- Our application identifies and removes duplicate data entries. This redundancy can occur due to various reasons, such as data entry errors or multiple recordings of the same observation. Removing duplicates helps to ensure a cleaner dataset and prevents overrepresentation of specific data points. However, since it may also cause subtle effects on the final output, we warn the users [3-5].
Irrelevant Data- Real-world datasets often contain columns of information that are not directly relevant to the analysis at hand. These irrelevant columns can cloud the insights gained from the data and make the overall dataset more cumbersome to work with. To address this, our web application empowers users to identify and remove such irrelevant columns. We encourage users to select those columns, and we delete them, so the dataset becomes concise. This data cleaning step helps to condense the dataset, focusing on the information most critical to the user's specific analysis [6-8].
Outlier Analysis- Data analysis hinges on the quality of the information we work with. However, real-world data often contains outliers – data points that deviate significantly from the rest [10]. These outliers can skew statistical measures like averages, potentially leading to misleading conclusions [11].

We plan on handling outliers in three different ways. Keeping the Outliers – Sometimes, outliers represent genuine phenomena. Users familiar with their data might choose to retain outliers if they feel like they have a valid explanation for it, or simply because they want to [1-2,9]. Removing the Outliers – If outliers significantly distort the data, users can opt to remove them [2-3]. To empower informed decision-making, our system will provide users with clear explanations of the strengths and weaknesses associated with each outlier handling method.

Normalization and Standardization - There are techniques that can improve the performance of some machine learning models. We can choose the right approach depending on your data and analysis goals [2-5]. Normalization – This is a method that fits the data into a range - (0 to 1) or (-1 to 1). This method is useful when the user doesn't need the actual value but a relative value. This is also helpful when comparing it to other normalized data fields [5] [7]. Standardization – This method focuses on how far each data point is from the average. This can be helpful for machine learning algorithms that are sensitive to the scale of the data [2] [3].

Both techniques can make your data analysis more accurate by ensuring all features are on a similar scale. However, the choice between normalization and standardization depends on the specific needs of the user. We will inform and explain to the users of the strengths, weaknesses, and benefits of using either method.

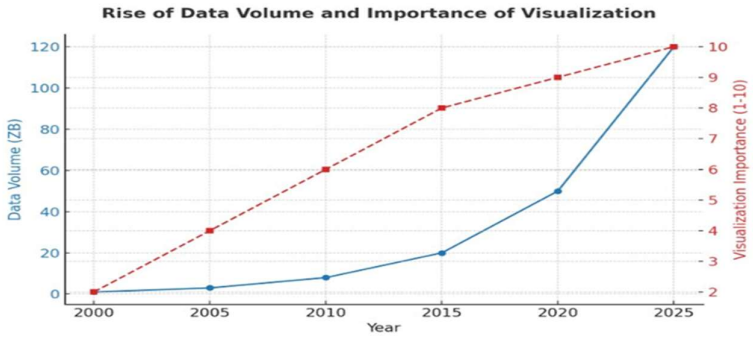

Fig.4 Graphical Representation of Year wise Rise of Data Volume [5]

3.Case and Methodology

3.1 Streamlit for Web Application Development

Streamlit was used to build a lightweight, interactive, and browser-based user interface that allows real-time user engagement without complex backend development.

3.1.1 Python Library Integration

Libraries such as Pandas (for data manipulation), Seaborn and Plotly (for rich visualizations), and Scikit-learn (for applying ML models) were used to perform core functionalities of the tool.

3.1.2 CSV File Upload and Handling

Users can upload their own real or synthetic CSV datasets. The platform automatically reads and processes these files for analysis.

3.1.3 Data Pre-processing Techniques

Includes handling missing data (e.g., imputation), label encoding for categorical variables, and detection/removal of outliers—all with guided steps.

3.1.4 Visualization Techniques

Interactive plots such as bar graphs, line charts, scatter plots, and correlation heatmaps are dynamically generated using Plotly and Seaborn.

3.1.5 Machine Learning Models

Basic models like regression (to find relationships) and clustering (to group data) are included for users to explore simple ML applications on their datasets.

3.1.6 COCOMO Effort Estimation Model

The Constructive Cost Model (COCOMO) was applied in the Semi-Detached mode to estimate the effort and time required for developing the platform by a novice team.

3.1.7 User Surveys and Feedback

Surveys were conducted to understand user needs, pain points with existing tools, and validate the design and feature choices of Data Edifire.

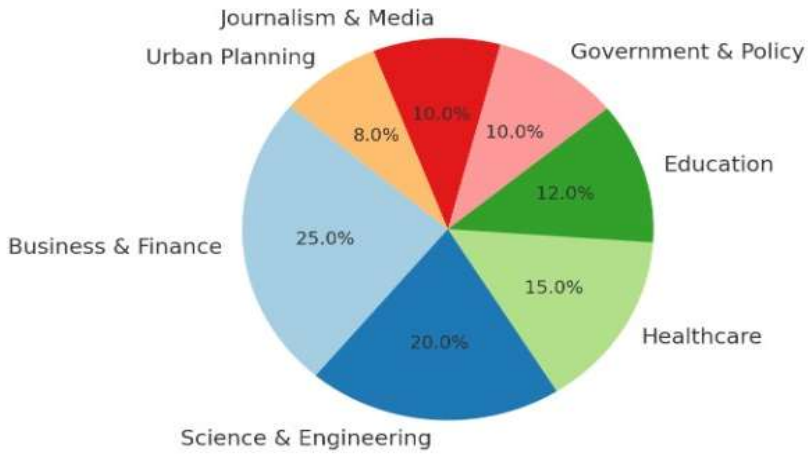

Fig. 5 Graphical Representation of User Survey and Feedback on Edifire

3.2 Python Libraries

Python is a powerful tool used by programmers, kind of like a toolbox with lots of helpful things inside. Our project uses these tools from Python, called libraries, to handle our data. These are some of the fundamental libraries used to create the Data Edifire web application. In the context of this project, we leverage the strengths of these Python libraries to achieve robust data handling capabilities.

1. Matplotlib – It is one of the simplest python libraries used in data visualization. It has a wide range of visualization options that cater to diverse exploration needs, which can be customized as well. Although the downside is that it isn't interactive.

2. Seaborn – Similarly to Matplotlib, Seaborn is a data visualization library, but it specializes in statistical visualizations. It excels in exploring relationships, distributions, and statistical summaries of data. It offers built in themes and colour palettes that are easy to apply for customization of a chart.

3. Plotly – Plotly is a python library that specializes in interactive charts. One can zoom in, zoom out, select data points, and do various other things using Plotly visualization. It also offers a wide range of charts and makes complex data exploration easier.

4. Streamlit – Streamlit is a Python library specifically designed for building interactive web apps for data science tasks. It simplifies the process by creating interactive websites with our Python code. Streamlit provides a wide range of built-in components like charts, tables, and text boxes, which enabled us to quickly create user interfaces for our web application.

3.2.1 Strengths and Weaknesses of Using These Libraries

Python has a vast collection of data visualization libraries - Matplotlib, Seaborn and Plotly. They cater to the different needs of the user and offer wide customization techniques. Python's data visualization features are readily integrated with other popular scientific computing and data analysis packages, such as NumPy for numerical computing, Pandas for data manipulation, and Scikit-learn for machine learning. This gives users the ability to manage the whole data science workflow, including data preparation, analysis, and visualization—in one place. With the growing importance of data science, there's a wealth of powerful tools available, each with its own specialty. For instance, Matplotlib excels at creating static charts, while Seaborn focuses on making data visualizations more informative. The true potential, however, is unlocked when these libraries work together. If we combine NumPy's numerical computation abilities with Pandas' data manipulation skills to prepare data for visualization with libraries like Seaborn or Plotly. This integrated approach allows data enthusiasts to leverage the strengths of each library, creating a more robust and efficient workflow. All of them have a common disadvantage. They have a steep learning curve. To apply these libraries into existing python projects, one needs to learn how to use them. This is where we step in and provide them with data visualization and data manipulation without the need for knowledge in coding.

3.2.2 Promoting Knowledge

Our brains are wired to learn and understand the world through our eyes. Seeing is the most powerful way we take in information, and it plays a huge role in how we think and make decisions in our daily lives, at work, and even in science. Vision is the dominant sensory input for cognitive reasoning in everyday life. Existing research says that if a risk (of a decision made to or to be made) is visualized, then it requires low effort for one to comprehend as compared to its textual or tabular counterparts. Data Edifire chooses to leverage on this strength. We realize how tough or inaccessible quick visualization may be, so we provide a platform to offer this. As we are left with no choice but to enter the digital era, some people find it extremely difficult to take such a big step, we want to show it is a small step that is separating the possibilities from the realities, that data is not just what it is or has been, but also what we make out of it. We understand how traditional methods used to gain insights from data may be overwhelming. There is limited dimensionality, and hidden patterns are hard to crack. Complex relationships and trends can easily go unnoticed. Visualizations allow for easier identification of these patterns by presenting data in a way that facilitates comparison and highlighting key features. Our research work is a learning tool as well as a visualizer. It will provide the users with knowledge of how to choose a chart that will fit the data in visualization. It will also guide the users through reading the graph. Other than that, we will make an automated report generation system which will save the user loads of time.

4.Results & Analysis

This research "Data Edifire" is intended to bridge a significant gap in data science accessibility by offering a no-code solution through which workflows are simplified, guided learning is supported, and data exploration is made approachable for beginners. According to survey insights, 72% of users reported that existing tools were overly complex, particularly when Python libraries were used, while 85% emphasized the importance of step-by-step contextual guidance. These findings are supported by existing educational literature, which discusses the challenges encountered by beginners in translating theoretical knowledge into practical application. To address these barriers, a modular design was implemented in Data Edifire using Streamlit for the interface and Scikit-learn for backend functionality. Technical challenges are abstracted away through this architecture without compromising on flexibility. During user testing, it was observed that the automated pre-processing module—which includes one-click outlier detection and smart imputation—reduced manual effort by approximately 40%. The project's feasibility was evaluated using the COCOMO Semi-Detached model, through which an estimated 128 person-hours was calculated for development. This estimation

validated the scalability and practicality of the project for educational use. When compared with existing tools such as Orange and KNIME, Data Edifire was found to have a distinct advantage in integrating learning into the user experience. Real-time feedback during data visualization was observed to enhance user understanding in ways that traditional platforms did not. However, current limitations include reduced customizability for advanced users, suggesting that the development of adaptive interfaces should be considered in future iterations.

5.Conclusion

Data analysis requires multiple Python libraries for data cleaning, manipulation, and visualization. These libraries are powerful but have a steep learning curve. This project offers a web application to bridge this gap by providing data visualization and manipulation without coding knowledge. It leverages the power of visuals to make data exploration and understanding more accessible. Complex workflows are simplified within the platform through automated pre-processing tasks—such as handling missing values, detecting outliers, and encoding variables—and interactive visualizations including scatter plots and histograms are integrated. Through this user-centric design, the need for manual coding is eliminated, allowing students, researchers, and professionals to concentrate on extracting insights rather than resolving programming errors. Moreover, tooltips that explain statistical concepts are embedded to guide users throughout the process, thus fostering both immediate comprehension and the gradual development of analytical skills. Looking forward, the platform is envisioned to support features like collaborative data analysis, predictive modelling, and customizable reporting. By reducing technical barriers, individual productivity is enhanced, and data literacy is promoted across diverse fields through the adoption of Data Edifire.

6. Future Scope

1. Integration of AI and ML models will be implemented to enhance insights through advanced predictive analytics and automated pattern detection.
2. Cloud deployment will be carried out to enable scalable platforms for real-time multi-user access and handling of larger datasets.
3. Natural language query support will be introduced so that the system can be interacted with using conversational language for dataset queries.
4. Enhanced collaboration features will be incorporated, allowing multi-user editing, shared dashboards, and team-based annotations.
5. Mobile and cross-platform accessibility will be developed to provide responsive web and native mobile apps for on-the-go data exploration.
6. A plug-in ecosystem will be established to allow custom visualization or preprocessing modules to be created and integrated by developers.
7. Automated data cleaning will be expanded using AI-based anomaly detection and intelligent correction suggestions.
8. Domain-specific templates will be provided for sectors such as healthcare, finance, and education.
9. Security and privacy will be strengthened through encryption, role-based access, and compliance with data protection laws.

References

[1] J. Walny, C. Frisson, M. West, D. Kosminsky, S. Knudsen, S. Carpendale, and W. Willett, "Data Changes Everything: Challenges and Opportunities in Data Visualization Design Handoff," arXiv, Aug. 1, 2019. doi: 10.48550/arXiv.1908.00192.

[2] B. Bach et al., "Challenges and Opportunities in Data Visualization Education: A Call to Action," arXiv preprint arXiv:2308.07703, Aug. 15, 2023. doi: 10.48550/arXiv.2308.07703.

[3] M. Hilbert, "Big Data for Development: A Review of Promises and Challenges," Development Policy Review, 2016. [Online]. Available: https://www.researchgate.net/publication/286907720_Big_Data_for_Development_A_Review_of_Promises_and_Challenges/citation/download

[4] V. Barnett and T. Lewis, Outliers in Statistical Data, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 1994.

[5] D. C. Gooding, "Visual Cognition: Where Cognition and Culture Meet," Philosophy of Science, 2006. doi: 10.1086/506232. [Online]. Available: https://doi.org/10.1086/506232

[6] C. M. R. Smerecnik et al., "Understanding the Positive Effects of Graphical Risk Information on Comprehension: Measuring Attention Directed to Written, Tabular, and Graphical Risk Information," J. Risk Res., vol. 13, 2010. doi: 10.1111/j.1539-6924.2010.01435.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2010.01435.x

[7] M. Friendly, "The Golden Age of Statistical Graphics," Statistical Science, vol. 23, 2008. [Online]. Available: http://www.jstor.org/stable/20697655

[8] Sayantan Chakrabarti, Basundhara Sengupta, Debarshi Halder, Subham Bhadra, Yash Raj;" A Security Approach in Digital Communication Medium using Hybrid Cryptography Application Properties"; International Journal of Sciences and Innovation Engineering, Vol.2; Issue 6, 2025.

[9] Sayantan Chakrabarti, Sudipta Roy, Soumit Chowdhury, Pranab Gharai; "Data Security Perspective in Secure Data Communication"; Scientific Voyage; Vol.3; Issue 4, 2023.

[10] Sudipta Roy, Papri Majumder, Sneha Chatterjee, Abhik Chakraborty, Mayank Shekhar, Sayantan Chakraborty, Pranab Gharai, Soumit Chowdhury; "STUDY ON TWO STAGE AUTHENTICATION FOR ONLINE TRANSACTION IN MOBILE DEVICES"; Scientific Voyage; Vol.3; Issue 4, 2023.

[11] Sayantan Chakrabarti, Rahul Binani; "Making of a cryptographic mechanism to enhance security in message transmission", Scientific Voyage; Vol.3; Issue 2, 2022.