DataAnalyzer: An AI-Driven Frameworkfor Automated Data Cleaning and Intelligent Business Analytics

Affan Khan
Department of AIML
Universal College of Engineering
Mumbai, India
affik7194@gmail.com

Ayaan Khan
Department of AIML
Universal College of Engineering
Mumbai, India
mayaanikhan7@gmail.com

Mohammed Bagdadi
Department of AIML
Universal College of Engineering
Mumbai, India
mohammed163835@gmail.com

Armaan Khilji
Department of AIML
Universal College of Engineering
Mumbai, India
armaankhilji987@gmail.com

Anas Dange
Department of AIML
Universal College of Engineering
Mumbai, India
anas.dange@universal.edu.in

***Abstract*— Effective data analytics and machine learning depend heavily on data quality, but real-world datasets frequently have missing values, outliers, duplicate records, and structural irregularities that lower analytical reliability. An AI-based interactive data cleaning and analysis platform that automates the data preparation lifecycle and facilitates effective data exploration is presented in this work. Python and the Streamlit framework are used in the development of the suggested system, which incorporates automated data cleaning methods such as KNN-based imputation for managing missing values and Isolation Forest for anomaly detection. The platform offers an interactive Exploratory Data Analysis (EDA) module that facilitates statistical summaries and visual data exploration in addition to preprocessing. The system includes an analytics interface based on the Large Language Model (LLM) to improve usability. The integration of this method bridges the gap between automated data cleaning and intuitive data interpretation. The experimental use of the system demonstrates that this proposed approach reduces manual preprocessing effort and also improves dataset readiness for downstream analytics and various machine learning tasks.**

Keywords*— *Data Cleaning, Artificial Intelligence, Data Preprocessing, Outlier Detection, Missing Value Imputation, Exploratory Data Analysis, Interactive Data Analytics.

I. INTRODUCTION

Effectively overseeing vast datasets and conducting exploratory data analysis are foundational for extracting meaningful business insights in modern data-driven industries [1]. Historically, manual data cleansing techniques and disjointed analytical platforms have proven to be labor-intensive, error-prone, and increasingly inadequate for handling the volume and velocity of contemporary information.[2] As a result, data professionals frequently encounter significant hurdles, as real-world data is often noisy; without rapid, automated preprocessing, the accuracy of subsequent modeling is severely compromised [3]. While managing data quality is vital for interpreting complex structures, relying solely on traditional methods often exposes severe generalization gaps and inefficiencies [4]. To address these industry demands, our project, **DataAnalyzer AI Pro**, unites automated data preparation with Large Language Model (LLM)-powered semantic capabilities to deliver a comprehensive, end-to-end data management solution. [5]

A. Problem Statement

Despite the rapid evolution of digital analytics, numerous enterprises still struggle to efficiently govern and rectify their raw data. A primary challenge is the absence of a unified platform that allows users to seamlessly execute advanced preprocessing tasks such as multi-strategy imputation and outlier detection across diverse file formats like JSON, Excel, and CSV, which otherwise forces analysts to rely on writing repetitive code instead of plain English [6]. Moreover, organizations suffer from disjointed workflows that lack adaptable systems to intelligently detect and contextualize errors without heavy manual effort. [7] Existing tools also demonstrate a notable deficiency in integrating advanced unsupervised anomaly detection models, such as Isolation Forests, leaving critical gaps in evaluating high-dimensional data quality [8]. This disconnect between automated cleaning utilities and natural language query interfaces fundamentally restricts an organization's capacity to access real-time business intelligence and execute rapid, automated decision-making [9]. Ultimately, current systems fail to provide fully integrated features like automated quality scoring within the machine learning pipeline, highlighting an urgent need for a cohesive, automated approach to data preparation and exploration [10].

B. *Solution*

To resolve such operational bottlenecks, **DataAnalyzer AI Pro** serves as a unifying bridge and offers an accessible framework that seamlessly links the

ingestion of raw data to a secure and interactive analytical environment. Through a Streamlit-driven web interface, it enables customers to consistently clean, tidy, and examine their data as a centralized data administration solution [6][7]. This architecture includes a backend that allows for smooth file uploads and the implementation of complex automated preprocessing procedures designed to meet strict requirements for data quality [4].

The solution is twofold which includes the following:

1. **An Advanced Data Cleansing Engine:** This program, which was created for data professionals, automates crucial preprocessing operations such text casing changes to guarantee structural regularity, Z-score standardization, and K-Nearest Neighbours (KNN) imputation. It keeps data quality consistently good across a range of organizational needs by streamlining the process of ingesting and updating files while cleverly handling missing entries and structural anomalies. Additionally, customers have the option to manually modify particular parameters, guaranteeing that any changes made to critical data are accurately and strictly regulated [3][9][10].
2. **An Interactive EDA and LLM Interface:** Second module provides a secure environment for Exploratory Data Analysis (EDA). Analysts can visually assess distribution plots, review correlation matrices, and retrieve actionable business intelligence by querying the data in plain English via an integrated Large Language Model (Llama-3.1-8b-instant). By incorporating charting libraries like Plotly and Seaborn, the application generates sharp, interactive visualizations that allow users to rapidly identify hidden patterns specific to their unique analytical use cases [2][6].

Under the hood, a highly scalable Python architecture seamlessly links these components together. It utilizes Streamlit's session state to synchronize operations in real time, making it highly capable of managing massive datasets without performance drops [2]. To round out the workflow, the platform features automated reporting that instantly updates users on data quality scores and maintains a transparent log of all applied cleaning operations.

C. *Objective*

Building a single, web-based data analysis platform that safely handles dataset transformations and keeps an exhaustive record of all performed cleaning operations is the primary driving force behind this project. Providing customers with real-time data quality indicators, such as originality and completeness scores, immediately on an interactive dashboard is one of the main goals [8][9]. The program includes web-based tools for producing statistical summaries and categorical distributions to further improve analytical accessibility, giving big datasets instant clarity [10].

This solution connects its front-end interface with a potent processing engine powered by high-performance Python packages like Scikit-learn and Pandas to provide seamless data handling. Reliable, current, and consistent access to the examined data is guaranteed by this configuration. The program also enables users to download their modified datasets in a variety of file formats, greatly increasing its usefulness for further stages of modeling. By carrying out automated procedures intended to protect data accuracy and integrity, the system complies with contemporary data science standards throughout the procedure. Specifically, the project seeks to achieve the following:

1. **Real-Time Quality Monitoring**: The platform includes a dynamic dashboard devoted to reporting data quality in order to provide customers with immediate analytical feedback. The application incorporates a Larells to bridge the gap between raw data processing and executive decision-making. It computes and displays important metrics, such as Uniqueness and Completeness Scores, in addition to an overall Quality Score derived from the frequency of duplicates and empty cells.
2. **Enhanced Exploratory Analysis**: By offering a comprehensive set of exploratory visualization tools, the project fosters a deep comprehension of dataset structures. Automatically created distribution histograms, category heatmaps, and correlation matrices that update in real time as the cleaning processes are conducted allow users to dynamically track data relationships.
3. **AI-Driven Business Intelligence**: The program incorporates a Large Language Model (LLM) to bridge the gap between executive decision-making and raw data processing. Using a Pandas dataframe agent, customers may ask natural language inquiries to extract sophisticated business intelligence, completely avoiding the need to manually develop Python scripts or SQL queries during the first exploration phase.
4. **Scalable and Flexible Data Export**: It is crucial to make sure the freshly produced data can be readily fed into external reporting applications or downstream machine learning pipelines. Users may easily download their clean data as JSON, Excel, or CSV files thanks to the platform's flexible exporting features, which also provide an extensive Markdown report that details the full analytical lifecycle [1][3][7][8].

II. METHODOLOGY

Building this project shows the need of systematic architecture that focuses on creating and integrating an online application that allows users to upload and automatically analyze their datasets. This platform is specifically designed to combine advanced preprocessing tools, exploratory visualization, and intelligent data cleansing into a single, coherent, interactive environment. The methodology that follows describes the algorithmic processing methods and underlying technical infrastructure that power the system. In the end, it provides an approachable, user-focused solution designed to get around the enduring challenges of handling noisy and inconsistent raw records, Consequently, the conventional challenges of extracting significant and practical insights from large and intricate datasets are eliminated.

A. *System Architecture and Design*

This framework, which runs on a simplified client-server architecture, combines a web interface with a potent processing engine dedicated to automate data analysis. Streamlit is used by the application's core to manage file uploads, carry out analytical tasks, and produce interactive graphics in a single setting [4]. Using commonly used formats like JSON, Excel, and CSV, users can upload raw datasets, which the system evaluates to generate dynamic reporting dashboards. The software enables analysts to

watch statistical summaries, find hidden data patterns, and automatically fix missing items by integrating visualization tools directly with the purification algorithms. It completely supports Exploratory Data Analysis (EDA) by generating box plots, correlation heatmaps, category charts, and distribution histograms. Outlining data distributions, highlighting variable linkages, and spotting systemic abnormalities without the need for manual programming all depend on the application of these visual techniques.

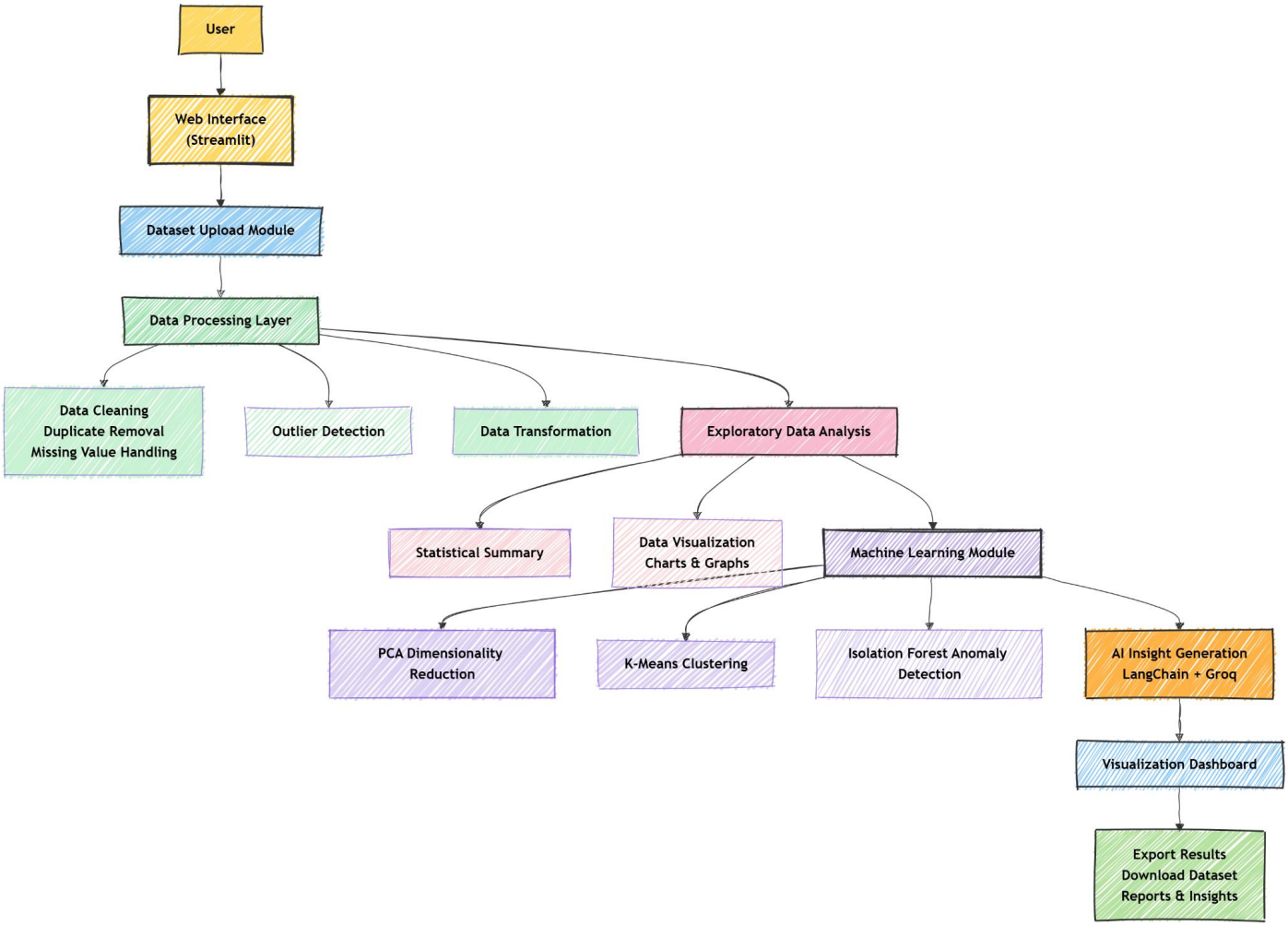

Fig 1. System Architecture of the Data Analysis Platform

B. Data Upload and Input Processing

To ensure smooth data ingestion, the platform incorporates a dedicated input module that lets users upload datasets directly through a web interface, completely bypassing the need for technical setup [5][9]. Because compatibility with standard sources is essential, the application accepts multiple file formats—namely CSV, Excel, and JSON—enabling analysts to effortlessly handle structured records gathered across different environments

After the file is uploaded, the backend in this system uses the Pandas library to parse it and create a structured dataframe from the raw data. In order to manipulate data and perform further analytical processing in a highly effective manner, this transformation is an essential step [2][7]. Additionally, the architecture is designed to securely archive an identical copy of the original dataset prior to implementing any preprocessing alterations in order to ensure dependability and guard against unintentional data loss. Users are free to return to the baseline data anytime needed thanks to this safeguard, which maintains stringent data integrity standards [1].

Immediately after loading the files, the system triggers an automated preliminary inspection. It extracts some vital dataset metrics, including the overall dimensional shape (total row and column counts), individual attribute data types, and the exact distribution of any missing values. Then by projecting this initial summary onto the dashboard, the platform allows users to quickly grasp the underlying structure and quality of their data, properly preparing them for the more intensive cleaning and exploratory phases that follow.

C. Data Cleaning and Preprocessing

Preparing and cleansing raw data are indispensable phases in any analytical workflow, given that real-world information frequently suffers from structural inconsistencies, absent values, and redundant records [7]. To rectify these defects and ensure high data reliability prior to modeling, the platform incorporates automated preprocessing routines. This is to ensure the accuracy in the conclusion of our dataset [10].

The system also features versatile mechanisms for resolving missing information, which adapt depending on the underlying data types. When handling numerical fields, the platform leverages statistical trends to estimate and replace blanks, providing users with options like K-Nearest

Neighbors (KNN) imputation, interpolation, or basic substitutions using the mean, median, or mode. Structural consistency is preserved by applying forward filling, mode replacement, or inserting a default label such as "Unknown" to fill empty cells.

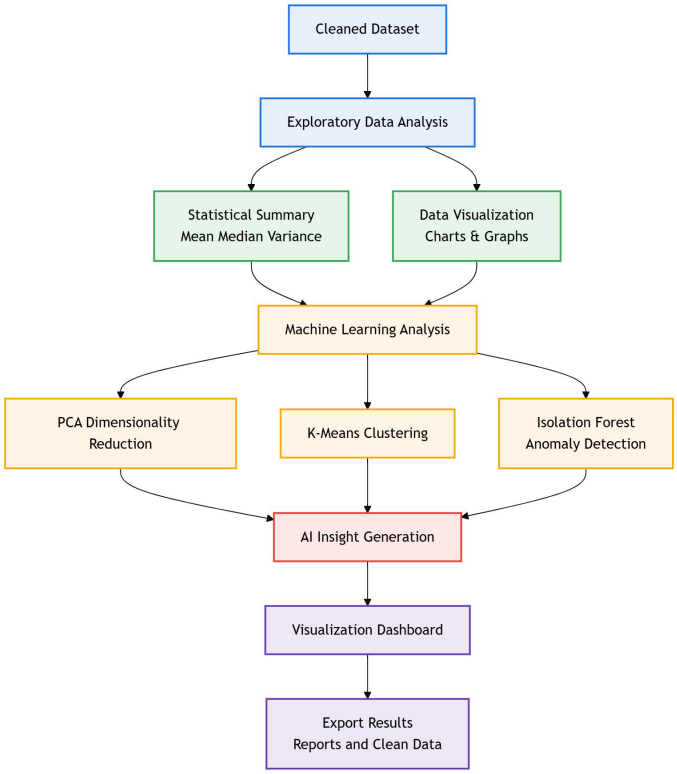

Fig 2. Analytical Workflow Diagram

The software does targeted text standardization in addition to handling missing numbers and categories. This entails eliminating superfluous whitespace, removing special characters, and requiring title-case formatting in all textual columns. These particular modifications guarantee that qualitative data stays highly consistent and ordered, greatly simplifying the ensuing visualization and analysis procedures [4][8].

D. Outlier Detection and Treatment

Anomalous data points differ from the typical observations in a dataset and are often caused by measurement errors, data entry errors, or irregular events. If they are left unchecked, these outliers impair statistical analysis's accuracy and have a major effect on subsequent machine learning results [1][10]. The platform uses a hybrid strategy that combines sophisticated machine learning anomaly detection techniques with traditional statistical methods to ensure analytical dependability.

Statistically, the framework evaluates data spread using the Interquartile Range (IQR) to flag values falling significantly outside the central 50% boundary, alongside Z-score calculations that identify outliers based on their standard deviation from the dataset's mean [9].

Beyond classical statistics, the system integrates the Isolation Forest algorithm via the Scikit-learn library [8]. This unsupervised machine learning model effectively isolates rare observations by randomly selecting features and partitioning values, making it highly capable of detecting anomalies within expansive and high-dimensional datasets

Once such anomalies are isolated, the architecture provides options to either filter them out or mitigate their effects through mathematical transformations. Supported techniques in this architecture includes Min-Max scaling, logarithmic transformations, and Z-score standardization, which collectively normalize these extreme values while preserving the natural and accurate structure of the data for accurate modeling. [5]

E. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is executed to grant users a clear comprehension of their dataset's underlying architecture and properties. By automatically evaluating the uploaded records, the framework yields statistical and visual summaries that expose hidden relationships, overarching patterns, and data irregularities. To examine numerical distributions, the platform dynamically renders box plots and histograms, enabling analysts to visualize the spread of continuous values and visually corroborate the presence of potential outliers. [1][9]

Furthermore, the system conducts correlation analysis by generating interactive heatmap matrices. These visuals directly map the associative strength between varying numerical features, immediately highlighting both weak and prominent dependencies

For categorical variables, the interface supplies donut charts, pie charts, and bar graphs to distinctly illustrate the frequency and distribution of qualitative data classes

Complementing these visual tools in the system, the architecture leverages the Pandas library to compute comprehensive descriptive statistics. Metrics such as kurtosis, skewness, variance, median, and mean are automatically tabulated, furnishing a robust numerical snapshot of the information before analysts proceed to further analytical stages

.

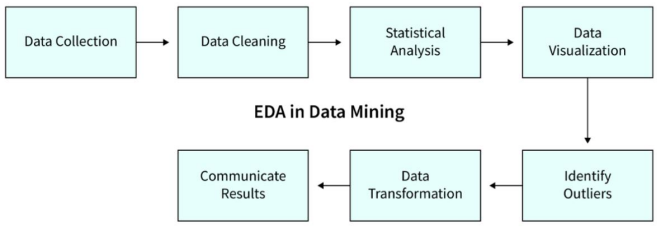

Fig 3 . EDA in Data Mining

F. Machine Learning Based Analysis

To extract more profound insights from the ingested data, the architecture incorporates foundational machine learning methodologies. These approaches are critical for uncovering latent patterns, isolating atypical observations, and mitigating dataset complexity to facilitate superior analysis [8].

Specifically, the platform utilizes Principal Component Analysis (PCA) to perform dimensionality reduction. By compressing the number of variables while retaining the most critical structural information, PCA significantly simplifies massive datasets, rendering them much more accessible for downstream evaluation and visual representation [3].

Furthermore, the framework deploys K-Means clustering, a widely adopted unsupervised algorithm designed to aggregate analogous data points into distinct clusters. This technique is highly effective for unearthing natural classifications and structural groupings hidden within the raw information.

Fig 4. Simplified Functional Block Diagram of the proposed system

Finally, to evaluate data quality and identify systemic irregularities, the system integrates the Isolation Forest These machine learning techniques allow the system to uncover hidden structures and provide more meaningful insights from the data. isolates rare and anomalous data points from normal observations.

G. AI-Powered Insight Generation

The framework has a specialized AI-driven insight creation module to help users understand difficult analytical results. The platform easily processes queries to provide insightful, text-based explanations straight from the uploaded datasets by integrating a Large Language Model using the Groq API with LangChain's dataframe agents. Asking questions in simple, natural language allows users to communicate with the system. After analyzing these inputs and scanning the pertinent data structures, the model automatically creates succinct summaries that emphasize important patterns and important findings [1][3][4][5].

In the past, obtaining this kind of strategic knowledge required a great deal of manual research and specific programming abilities.

The technology completely removes that technological barrier by utilizing this automated natural language method, enabling people to find hidden patterns without having to write tedious SQL or Python programs during the discovery phase.

In the end, this feature is particularly helpful for non-technical stakeholders since it converts inflexible analytical metrics into an easily understood manner for quick, data-driven decision-making.ve and highly accessible format for rapid, data-driven decision-making [3][8].

H. Result Visualization and Export

To present analytical outcomes in a highly structured and accessible manner, the platform incorporates dynamic visualization dashboards. By automatically generating diverse graphical representations, the system empowers analysts to efficiently decode underlying patterns, associative trends, and complex relationships within the data, which is an essential requirement for driving accurate and reliable data-driven decision-making.

algorithm via the Scikit-learn library [4][7].

In In addition to its exploratory visual skills, the architecture has strong extraction features to guarantee that the processed data becomes as useful and as accurate as possible. Users can safely download their fully cleaned datasets for use in subsequent applications after crucial preparation procedures, such as the elimination of duplicate entries and the clever resolution of missing values, are completed [6]. Additionally, the solution enables the direct export of textual insights produced by AI, statistical summaries, and visual measurements.

All enhanced records and synthesized business knowledge can be easily stored, shared with stakeholders, and incorporated into official organizational paperwork or presentations thanks to this extensive export flexibility [4].

III . DISCUSSION

By providing a completely integrated environment for preprocessing, visual exploration, and advanced analytics, the proposed Data Analysis and Visualization framework radically changes how practitioners interact with raw data.

The architecture simplifies the extraction of usable intelligence from otherwise complex datasets by combining automated data purification, exploratory evaluations, machine learning algorithms, and artificial intelligence-driven insight development.
In the end, this unified approach removes the conventional dependence on disparate analytical tools, enabling analysts to carry out whole data operations in a single, web-accessible setting.

In terms of user experience, the platform provides an easy-to-use, interactive interface where datasets can be loaded and assessed with ease. Through automatic preprocessing capabilities that explicitly target missing entry resolution, duplicate record removal, and structural outlier separation, the system ensures that raw data is rigorously prepared for downstream modelling. Exploratory Data Analysis Exploratory Data Analysis (EDA) visualizations, such as correlation heatmaps, box plots, and histograms, allow users to graphically map variable associations, evaluate distributions, and quickly identify systemic abnormalities in order to further improve data interpretation.

The platform's incorporation of fundamental machine learning techniques, such as dimensionality reduction, clustering, and anomaly detection models, is a key component of its analytical capabilities. By using these techniques, the system can automatically reveal underlying structural groups in the data, providing far more in-depth analytical insights. Additionally, by adding an AI
-powered insight generation module, people may use simple natural language queries to interact with their datasets. This feature effectively bridges the gap between raw metrics and strategic understanding by automatically condensing complex analytical outputs into easily comprehensible textual explanations, making advanced data science highly accessible to both technical practitioners and non-technical stakeholders [3][2][8][7].

IV . CONCLUSION

Combining data exploration, cleaning, and insight generating, the recently created data analysis and visualization system provides a useful web-based solution. This integrated platform enables users to upload raw files, carry out preprocessing procedures, and instantly provide graphical summaries rather than requiring them to juggle several disparate tools [1][2].

The application uses automated purification techniques to ensure the dataset is extremely dependable before delving into in-depth research. By eliminating duplicate rows, imputing missing items, and identifying structural outliers, these built-in methods address typical data problems. After the data has been cleaned, users may rapidly determine how their variables connect to each other and what the overall data distributions look like by using interactive visual tools like box plots, histograms, and correlation heatmaps [3][4].

Additionally, the platform uses machine learning techniques to go beyond simple charts. The system may automatically uncover hidden structures and uncommon observations that might not be apparent during manual inspection by using methods like clustering, dimensionality reduction, and anomaly detection [5][6].

The AI-driven insight creation module is one of this setup's main benefits. Instead of writing complicated code, users may query their datasets using everyday language thanks to its Large Language Model. Also, the system responds with summaries of the analytical findings in plain English. This particular feature removes technical barriers so that people with very little training in data science can easily analyse their data and also make data-driven and well-informed judgments [7][8].

In the end, the entire analytical workflow is significantly accelerated by combining automated data preparation, interactive graphics, and AI-powered querying into a single, easily accessible workspace. By adding more sophisticated predictive models to the pipeline and extending the backend to handle large or real-time data streams, future updates could further enhance these capabilities [9][10].

V. REFERENCES

[1] **Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani (2019)** . Exploratory Data Analysis using Python., International Journal of Innovative Technology and Exploring Engineering (IJITEE)

[2] **Vijay Panwar (2024)** . AI-Powered Data Cleansing: Innovative Approaches for Ensuring Database Integrity and Accuracy., International Journal of Computer Trends and Technology (IJCTT)

[3] **Rahul Cherekar (2024).** Automated Data Cleaning: AI Methods for Enhancing Data Quality and Consistency, International Journal of Emerging Trends in Computer Science and Information Technology (IJETCSIT)

[4] **Jingyu Zhu, Xintong Zhao, Yu Sun, Shaoxu Song, Xiaojie Yuan (2024)**. Relational Data Cleaning Meets Artificial Intelligence: A Survey, Data Science and Engineering

[5] **Shuo Zhang, Zezhou Huang, Eugene Wu (2024)**. Data Cleaning Using Large Language Models, arXiv:2410.15547 [cs.DB]

[6] **Alberto Sánchez Pérez, Paolo Papotti, Alaa Boukhary, Luis Castejón Lozano, Adam Elwood (2025)**. An LLM-Based Approach for Insight Generation in Data Analysis, North American Chapter of the Association for Computational Linguistics (NAACL 2025)

[7] **Anjali Kapoor (2025)**. AI-Driven Data Cleaning: Intelligent Detection and Correction of Data Errors, International Journal of Computer Technology and Electronics Communication (IJCTEC)

[8] **Zuleaizal Sidek, Sharifah Sakinah Syed Ahmad, Noor Hasimah Ibrahim Teo (2025)**. Unsupervised outlier detection in high-dimensional text data: a comparative analysis, Bulletin of Electrical Engineering and Informatics

[9] **Sanjeet Singh, Geetika Madaan, HR Swapna, Amrinder Singh, Binay Kumar Pandey, A. Shaji George, Digvijay Pandey (2025)**. Unleashing the Power of AI and Data Analysis: Transforming Insights into Action, Interdisciplinary Approaches to AI, Internet of Everything, and Machine Learning (IGI Global)

[10] **Alhassan Mumuni, Fuseini Mumuni (2025)**. Automated data processing and feature engineering for deep learning and big data applications: A survey, Journal of Information and Intelligence