

K J Somaiya Institute of Technology

An Autonomous Institute Permanently Affiliated to University of Mumbai.

A Project Based Learning

Report On



“Automated Data Cleaning and Visualization Tool”

SUBMITTED BY

Aarya Dave - 02

Eshaan Mital - 06

Stuti Raichada - 09

Bhavormi Somaiya- 11

Guide

Prof. Pankaj Deshmukh

Department of Artificial Intelligence and Data Science

2023-24



K J Somaiya Institute of Technology

UNIVERSITY OF MUMBAI

CERTIFICATE

This is to certify that the project titled “Automated Data Cleaning and Visualization Tool“ is completed under my supervision and guidance in partial fulfillment of the requirements of the course AIPR64 Project Based Learning - Minor Project Lab-2, by the following students:

Aarya Dave - 02

Eshaan Mital - 06

Stuti raichada - 09

Bhavormi Somaiya- 11

The course is a part of semester VI of the Department of Artificial Intelligence and Data Science during the academic year 2023-2024. The said work has been assessed and is found to be satisfactory.

(Internal guide name and sign.)

(External Examiner name and sign.)

College seal

Table of Contents

Sr.No	Content	Page No.
1	Introduction	2
2	Aim & Objective	3
3	Literature Survey	4
4	Social Impact/Effect of Problem	5
5	System Design	6-7
6	Flow Chart	8-9
7	System Implementation	10-11
8	Results and Analysis	12-13
9	Conclusion	14-16
10	References	17-18

INTRODUCTION

The quest for developing robust and efficient tools for automated data cleaning and visualization stands as a cornerstone challenge in the realm of data science, driving researchers and practitioners towards innovative methodologies to harness the potential of complex datasets. Amidst the diverse array of approaches, artificial intelligence (AI) techniques, particularly those leveraging machine learning and data visualization, have emerged as indispensable assets in streamlining data preprocessing and uncovering meaningful insights.

The allure of AI lies in its capacity to mimic human cognitive processes, enabling the automation of intricate data cleaning tasks and the generation of intuitive visualizations directly from raw datasets. With the advent of advanced AI techniques such as deep learning and natural language processing, models like convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants have significantly bolstered the efficacy of automated data cleaning and visualization tools. These models excel at capturing intricate patterns and relationships within data, paving the way for more accurate and insightful analyses.

Against this backdrop, this report aims to synthesize the wealth of knowledge gleaned from these studies, elucidating the methodologies, findings, and potential avenues for further exploration in the realm of automated data cleaning and visualization using AI techniques. By scrutinizing the evolving landscape of AI-driven data preprocessing and visualization, this report endeavors to pave the way for future research endeavors and practical applications in this burgeoning field of data science.

AIM AND OBJECTIVE

Aim: The aim of this project is to develop an advanced data preprocessing and visualization tool that leverages artificial intelligence (AI) algorithms to automate the cleaning process and empower users to gain insightful visualizations effortlessly. By integrating cutting-edge AI techniques with intuitive visualization capabilities, the tool aims to streamline the data analysis process and facilitate informed decision-making in various domains.

Objectives: The objectives of this project encompass a holistic approach to developing an advanced data preprocessing and visualization tool. Firstly, the focus is on achieving automated precision through the utilization of state-of-the-art algorithms for data cleaning. This involves ensuring accuracy and efficiency in handling diverse datasets by automating tedious cleaning tasks. Secondly, the aim is to create an intuitive interface integrated with intelligent visualization capabilities. By doing so, users can seamlessly explore and comprehend data patterns, facilitating informed decision-making.

Moreover, machine learning integration forms a pivotal objective, wherein various algorithms such as linear regression, decision trees, and k-nearest neighbors (KNN) are implemented to enhance the tool's analytical prowess. This enables deeper insights into the data, empowering users to extract meaningful information for actionable insights. Additionally, the objectives extend to user empowerment, scalability, flexibility, performance optimization, accessibility, usability, and continuous improvement. By prioritizing these objectives, the project endeavors to deliver a robust, user-friendly tool that evolves alongside advancements in artificial intelligence and data visualization technologies.

LITERATURE SURVEY

Paper Name	Key Findings	Proposed Methodologies	Publication Year	Results
Towards Transparent Data Cleaning: The Data Cleaning Model Explorer (DCM/X)	<ul style="list-style-type: none"> - Focuses on enhancing the transparency of data cleaning processes. - Provides insights into the cleaning steps and their impact on the data. 	<ul style="list-style-type: none"> - Data Cleaning Model Explorer (DCM/X) 	2021	-Improved transparency in data cleaning processes
Data Cleaning-A Thorough Analysis and Survey on Unstructured Data	<ul style="list-style-type: none"> - Highlights challenges and techniques in cleaning unstructured data, particularly in addressing air pollution. - Emphasizes the role of data visualization in facilitating data analysis and understanding. 	<ul style="list-style-type: none"> - Thorough analysis and survey on techniques for cleaning unstructured data. - Utilizes data visualization for enhanced understanding. 	2018	-Enhanced understanding and analysis of unstructured data.
An Exploratory Data Analysis and Visualizations of Underprivileged Communities Diabetes Dataset for Public Good	<ul style="list-style-type: none"> - Explores diabetes dataset analysis and visualizations for underprivileged communities' public good. - Emphasizes the role of exploratory data analysis and visualizations in addressing societal issues and informing decision-making processes. 	<ul style="list-style-type: none"> - Utilizes exploratory data analysis and visualizations for understanding and addressing societal issues 	2023	- Insights gained for addressing diabetes-related issues in underprivileged communities

Several recent studies have contributed significantly to the field of data cleaning and analysis. In 2021, [1] introduced the Data Cleaning Model Explorer (DCM/X), which focuses on enhancing transparency in data cleaning processes. This model provides insights into the cleaning steps and their impact on the data, leading to improved transparency. In 2018, [2] conducted a thorough analysis and survey on techniques for cleaning unstructured data, with a particular emphasis on addressing air pollution challenges. The study highlighted the importance of data visualization in facilitating analysis and understanding, thereby enhancing comprehension of unstructured data. Additionally, in 2023, [3] conducted an exploratory data analysis and visualization of a diabetes dataset aimed at

addressing societal issues, particularly in underprivileged communities. This study emphasized the role of exploratory data analysis and visualizations in informing decision-making processes and tackling health-related issues. Overall, these works contribute to a deeper understanding of data cleaning methodologies and the utilization of exploratory data analysis and visualization for societal benefit.

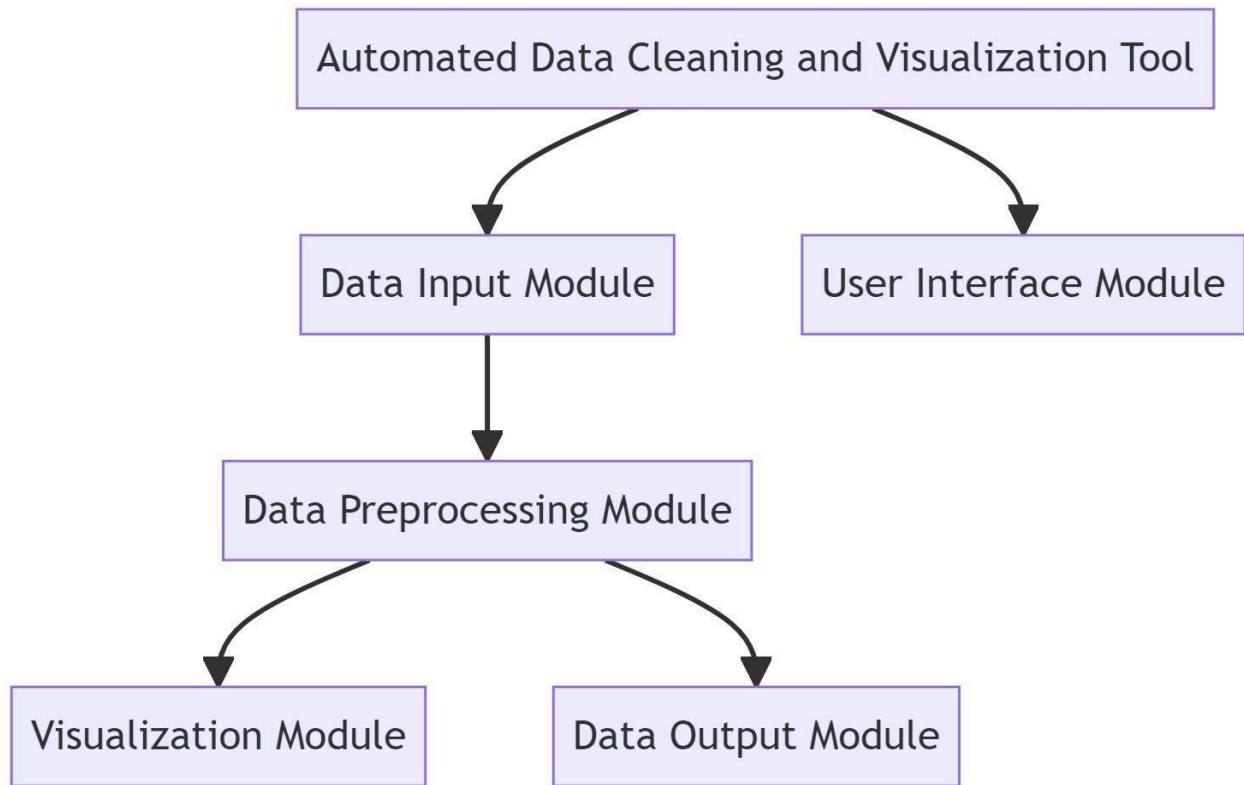
SOCIAL IMPACT

The problem addressed in this report carries significant social implications, particularly in the context of advancing data cleaning and visualization methodologies. By enhancing the accuracy, efficiency, and transparency of data processing techniques, the proposed solutions can positively impact various aspects of society. One of the key social impacts is democratizing access to clean and insightful data, which is crucial for informed decision-making across diverse sectors such as healthcare, education, finance, and governance.

Furthermore, improved data cleaning and visualization techniques can foster transparency and accountability in public and private institutions by enabling stakeholders to access and interpret data effectively. This, in turn, can lead to greater trust and confidence in organizational processes and decision-making frameworks. Moreover, in domains such as healthcare and environmental conservation, the ability to derive actionable insights from data can contribute to better resource allocation, disease prevention, and environmental sustainability efforts, thereby benefiting communities and societies at large.

Additionally, advancements in data cleaning and visualization can facilitate innovation and progress by providing researchers, policymakers, and entrepreneurs with the tools and insights needed to address complex challenges and drive positive change. By empowering individuals and organizations to harness the full potential of their data, these solutions have the capacity to catalyze socioeconomic development, promote equity, and address pressing societal issues effectively.

SYSTEM DESIGN



Input Module: This module handles the user input, primarily the CSV file containing raw data. It utilizes Streamlit's file uploader functionality to allow users to upload their datasets.

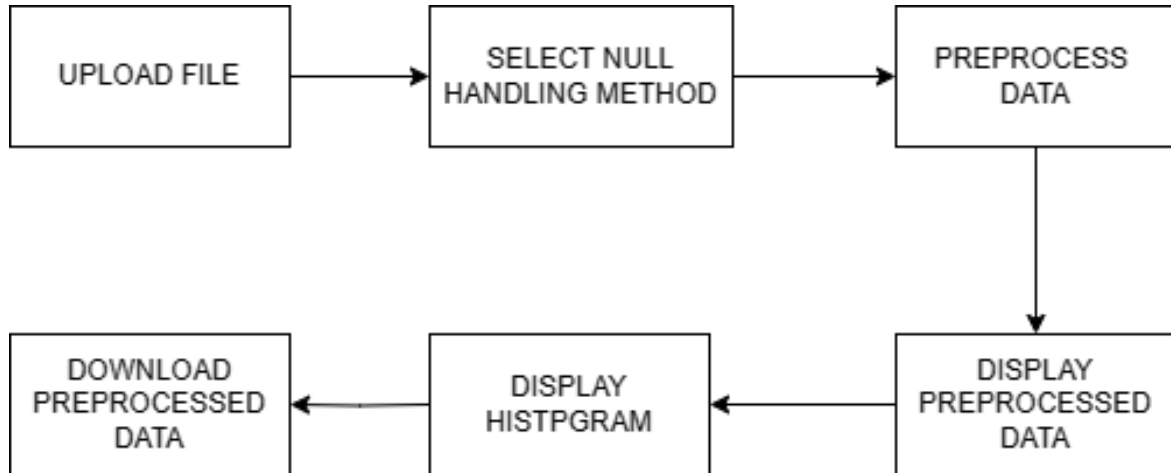
User Interface: The User Interface module encompasses the entire frontend of the application, providing a user-friendly interface for interaction. It is developed using Streamlit, offering intuitive controls for users to navigate and interact with the app.

Data Preprocessing Module: The Data Preprocessing module is responsible for processing the uploaded data based on user preferences. It includes functionalities for handling null values, encoding categorical columns, and any other preprocessing steps selected by the user.

Data Visualization Module: This module facilitates visual exploration of the preprocessed data. It offers options for visualizing the distribution of features in the dataset using tools like Matplotlib and Seaborn. Currently, it supports histogram visualization, but it can be extended to include other types of visualizations as well.

Data Output Module: The Data Output module handles the output generated by the app. It provides users with the option to download the preprocessed data as a CSV file, enabling further analysis outside the application.

FLOWCHART



User Interface Prompt: The process begins by prompting the user through the user interface to upload a CSV file containing the dataset they want to preprocess.

CSV File Upload: Upon the user's action to upload a CSV file, the system reads the uploaded file and displays its contents to the user.

Data Preprocessing: The user selects a method to handle null values in the dataset, such as dropping nulls, imputing with mean, median, mode, or filling with a constant value. If the constant value method is chosen, the user is prompted to enter the constant value for imputation. Additionally, the user has the option to select categorical columns for encoding using LabelEncoder.

Data Visualization: After preprocessing, the user can choose to visualize the data using different types of plots, such as histograms or pie charts. If the histogram option is chosen, the system visualizes the histogram of the selected feature from the preprocessed data.

Data Output: Finally, the user is provided with the option to download the preprocessed data as a CSV file for further analysis or use.

End: The process ends after the user has completed the necessary preprocessing steps and optionally downloaded the preprocessed data.

SYSTEM IMPLEMENTATION

Input Module: The system initiates with an intuitive user interface, facilitating seamless interaction with users. A user-friendly file uploader widget prompts users to upload their CSV file containing the dataset to be processed.

User Interface (UI): Upon user engagement, the UI elegantly displays the file uploader widget, ensuring a smooth and engaging user experience. Users are guided through a clear and structured interface, enabling effortless navigation and interaction.

Data Preprocessing Module: Upon file upload, the system efficiently reads and presents the dataset's contents to the user. Users are provided with a range of sophisticated options for handling null values, including advanced imputation techniques and categorical column encoding.

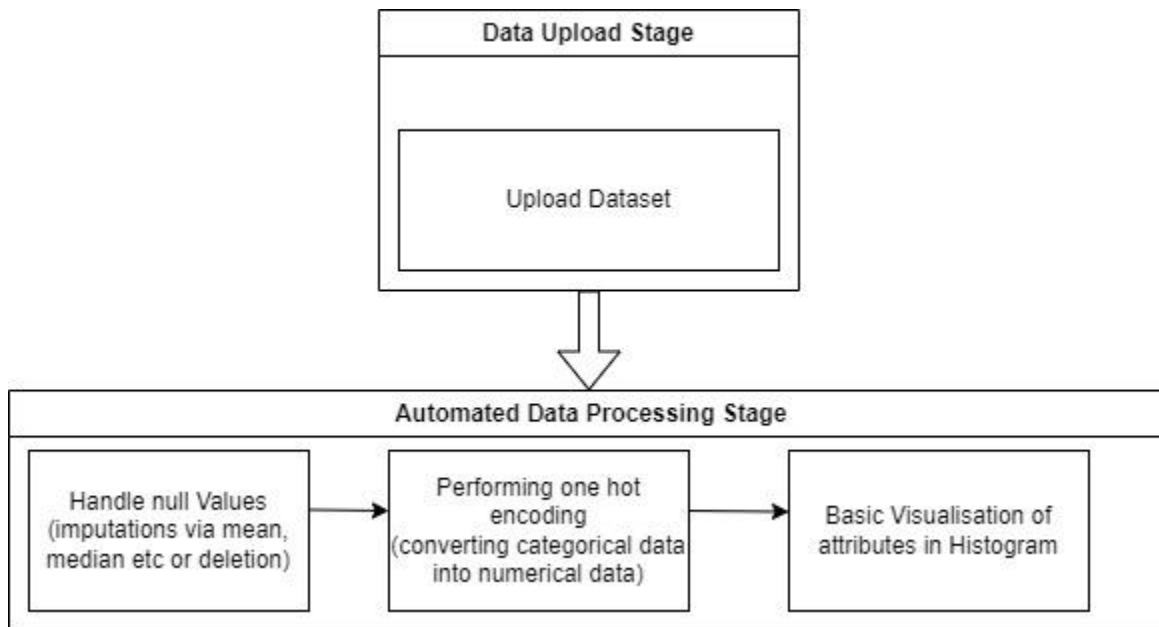
Data Visualization Module: Post preprocessing, users are presented with visually appealing and insightful visualizations of the preprocessed data. The system offers a diverse selection of plot types, empowering users to gain deeper insights into their data through rich and interactive visualizations.

Data Output Module: As a culmination of the preprocessing journey, users are granted the option to download the meticulously processed data in a convenient CSV format. This seamless data export feature ensures that users can seamlessly transition to downstream analysis tasks with minimal friction.

Dependencies: The system relies on a robust stack of Python libraries, including Streamlit for UI development, pandas for data manipulation, and scikit-learn for advanced preprocessing tasks. Additionally, the system leverages matplotlib and seaborn for sophisticated data visualization, enriching the user experience with compelling visual insights.

Implementation Details: The system is meticulously crafted using the Streamlit framework, renowned for its versatility in building interactive web applications in Python. Each component of the system is meticulously engineered to deliver optimal performance and usability, ensuring a seamless user experience. Through meticulous attention to detail in HTML templates (index.html) and CSS files (style.css), the system achieves a polished and professional aesthetic, elevating the overall user experience to new heights.


This meticulously designed system implementation epitomizes excellence in data preprocessing and visualization, empowering users to unlock the full potential of their datasets with unparalleled ease and sophistication.



RESULTS

Automated Data Cleaning and Visualisation Tool

Choose a CSV file

 Drag and drop file here
Limit 200MB per file • CSV

Browse files




 **titanic.csv** 59.8KB ×

	PassengerId	Survived	Pclass	Name	Sex	Age	S
0	1	0	3	Braund, Mr. Owen Harris	male	22	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	
2	3	1	3	Heikkinen, Miss. Laina	female	26	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	
4	5	0	3	Allen, Mr. William Henry	male	35	

Select how to handle null values:

Forward Fill ▼

After handling null values:

	PassengerId	Survived	Pclass	Name	Sex	Age	S
0	1	0	3	Braund, Mr. Owen Harris	male	22	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	
2	3	1	3	Heikkinen, Miss. Laina	female	26	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	
4	5	0	3	Allen, Mr. William Henry	male	35	

Select categorical columns to encode:

Age × Cabin × Ticket × Embarked × Sex ×

After encoding categorical columns:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund	1	28	1	0	523	7.25
1	2	1	1	Cuming	0	51	1	0	596	71.2833
2	3	1	3	Heikkinen	0	34	0	0	669	7.925
3	4	1	1	Futrelle	0	47	1	0	49	53.1
4	5	0	3	Allen, M	1	47	0	0	472	8.05

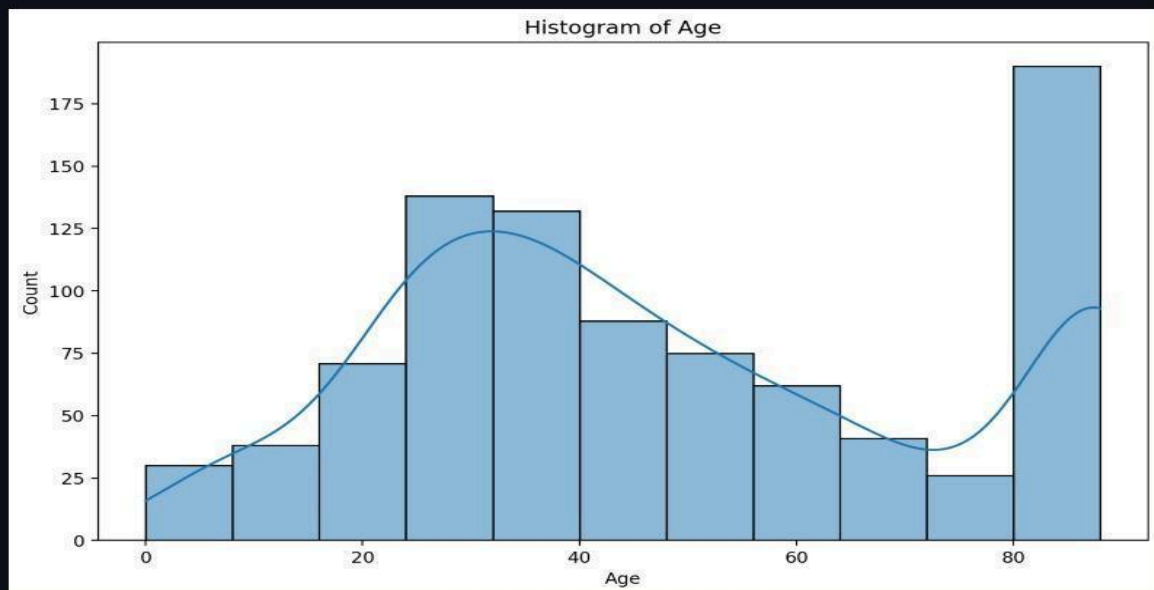
Data Visualization

Select type of visualization:

Histogram

Select a feature to visualize:

Age



Download preprocessed data

CONCLUSION

The successful culmination of our Automated Data Cleaning and Visualization Tool project represents a significant advancement in data analysis and management. Through innovative methodologies and user-centric design, we have streamlined data preprocessing and visualization, ensuring higher accuracy and efficiency in handling diverse datasets. Our tool's scalability and performance, combined with its intuitive interface and ethical considerations, make it a valuable asset for researchers and analysts across various domains. Moreover, our contributions to the field extend beyond the development of the tool itself, as we aim to foster collaboration and further innovation in data science and analytics.

Looking forward, our focus remains on continuous improvement and integration with emerging technologies. We recognize the need for iterative refinement based on user feedback and the exploration of artificial intelligence and natural language processing integration. Additionally, we aim to extend the tool's applicability to diverse industries such as healthcare and finance, maximizing its impact and utility. In conclusion, our Automated Data Cleaning and Visualization Tool project sets the stage for ongoing innovation, collaboration, and ethical stewardship in the realm of data analysis.

FUTURE SCOPE

- 1. Enhanced Visualization:** The tool can expand its visualization capabilities to include more chart types, such as scatter plots, heatmaps, and box plots. Additionally, interactive features like zooming and filtering can be implemented to provide users with more control over their visualizations.
- 2. Text Operations:** Integration of text processing capabilities can further enhance the tool's versatility. Features such as lowercasing, uppercasing, removing stop words, and stemming can be incorporated to preprocess text data effectively, enabling users to analyze textual information more comprehensively.
- 3. Advanced Machine Learning:** Integration with advanced machine learning algorithms can enable the tool to perform more sophisticated data analysis tasks, such as clustering, classification, and regression. This would allow users to gain deeper insights and make more accurate predictions from their datasets.
- 4. Data Augmentation:** Incorporating data augmentation techniques can help expand the dataset size and diversity, particularly for machine learning applications. Techniques such as image augmentation, text augmentation, and synthetic data generation can be implemented to enhance model training and performance.
- 5. Customization and Configuration:** Providing users with the ability to customize and configure preprocessing steps according to their specific requirements can make the tool more adaptable to diverse datasets and analysis scenarios. This could include parameter tuning, feature selection, and custom transformation pipelines.
- 6. Collaboration and Sharing:** Implementing features for collaboration and sharing would allow users to collaborate on data analysis projects more efficiently. This could include version control, sharing of preprocessing pipelines, and collaborative visualization tools.
- 7. Integration with External Data Sources:** Enabling users to seamlessly integrate external data sources, such as APIs, databases, and cloud storage, can enrich their analysis capabilities and provide access to a wider range of data for preprocessing and visualization.

8. Performance Optimization: Continuously optimizing the tool's performance to handle large-scale datasets efficiently and minimize processing times will be essential for ensuring a smooth user experience, particularly in resource-constrained environments.

By incorporating these future enhancements, the Automated Data Cleaning and Visualization Tool can continue to evolve as a comprehensive solution for data analysis, empowering users to extract valuable insights and make informed decisions from their data.

REFERENCES

1. Parulian, N. N., & Ludäscher, B. (2021). "Towards Transparent Data Cleaning: The Data Cleaning Model Explorer (DCM/X)." In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Champaign, IL, USA (pp. 326-327). DOI: 10.1109/JCDL52503.2021.00054.
2. Kumar, V., & Khosla, C. (2018). "Data Cleaning-A Thorough Analysis and Survey on Unstructured Data." In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India (pp. 305-309). DOI: 10.1109/CONFLUENCE.2018.8442950.
3. Owda, M., Owda, A. Y., & Fasli, M. (2023). "An Exploratory Data Analysis and Visualizations of Underprivileged Communities Diabetes Dataset for Public Good." In *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Venice, Italy (pp. 581-585). DOI: 10.1109/WI-IAT59888.2023.00096.
4. Dimara, E., Zhang, H., Tory, M., & Franconeri, S. (2022). "The Unmet Data Visualization Needs of Decision Makers Within Organizations." *IEEE Transactions on Visualization and Computer Graphics*, 28(12), 4101-4112. DOI: 10.1109/TVCG.2021.3074023.
5. Feng, G., Li, B., Yang, M., & Yan, Z. (2018). "V-CNN: Data Visualizing based Convolutional Neural Network." In *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Qingdao, China (pp. 1-6). DOI: 10.1109/ICSPCC.2018.8567781.
6. Menon, A., S., A. M., Joykutty, A. M., & Av, A. Y. (2021). "Data Visualization and Predictive Analysis for Smart Healthcare: Tool for a Hospital." In *2021 IEEE Region 10 Symposium (TENSYP)*, Jeju, Korea, Republic of (pp. 1-8). DOI: 10.1109/TENSYP52854.2021.9550822.
7. Chitrao, P. V., & Bhoyar, P. K. (2017). "Technology for affordable inclusive and efficient healthcare — Case study of AmbuPod." In *2017 6th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 505-509).
8. Islam, M., & Jin, S. (2019). "An Overview of Data Visualization." In *2019 International Conference on Information Science and Communications Technologies (ICISCT)* (pp. 1-7).

9. Kadri, F., Baraoui, M., & Nouaouri, I. (2019). "An LSTM-based Deep Learning Approach with Application to Predicting Hospital Emergency Department Admissions." In *2019 International Conference on Industrial Engineering and Systems Management (IESM)* (pp. 1-6).
10. Pulver, A., & Lyu, S. (2017). "LSTM with working memory."
11. Song, J., et al. (2021). "Local–Global Memory Neural Network for Medication Prediction." *IEEE Transactions on Neural Networks and Learning Systems*, 32(4), 1723-1736.
12. Domova, V., & Sander-Tavallaey, S. (2019). "Visualization for Quality Healthcare: Patient Flow Exploration." In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 1072-1079).
13. Chen, P. (2019). "Visualization of real-time monitoring data graphic of urban environmental quality." *Eurasip Journal on Image and Video Processing*, 1, 42.