IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DESIGN AND DEVELOPMENT OF HANDLING UNSTRUCTURED DATA MAINTENANCE AND INSIGHTS EXTRACTION

[1]Dr. M. Saraswathi, [2]Madhuvarsu Subhash Chandra Manvith, [3]Dhanokonda Balatripurasundari

[1] Assistant Professor, [2] UG Scholar, [3]UG Scholar

[123]Department of Computer Science and Engineering,

[123]SCSVMV University, Kanchipuram, Tamilnadu, India.

Abstract: Data Insight Pro is a comprehensive Tkinter-based graphical user interface (GUI) application designed to facilitate unstructured data management and enable users to extract meaningful insights from their datasets. It provides a user-friendly platform for uploading Excel or CSV files, performing data preprocessing tasks, conducting statistical analyses, and generating insightful visualizations.

One of the key features of Data Insight Pro is its ability to preprocess data by handling missing values, removing duplicates, and detecting and handling outliers. This data cleaning process is crucial for improving data quality and enhancing the accuracy of subsequent analyses and visualizations. Additionally, the application incorporates a simple linear regression model to assess the impact of data preprocessing on model performance, calculating the accuracy before and after data cleaning.

Users can download the cleaned dataset as a CSV file and the generated visualizations as PDF or PNG files, with the option to include log information for better documentation and reproducibility. Data Insight Pro aims to streamline the data exploration, analysis, and visualization process, empowering users to make informed decisions based on high-quality data and insightful representations.

Index Terms **- U**nstructured Data, Data Visualization, Data Preprocessing, Tkinter GUI, Unstructured Data Management, Linear Regression, Insights Extraction, Data Analysis.

Introduction

The primary objectives of the Data Insight Pro project are to facilitate efficient data management, preprocessing, visualization, and analysis for users dealing with unstructured datasets. Through a user-friendly interface, the application aims to empower users to upload Excel (.xlsx) or CSV (.csv) files seamlessly, ensuring robust file handling mechanisms to guarantee smooth dataset importation. Core functionalities include essential data preprocessing features like data exploration, missing value handling, and duplicate removal, ensuring data cleanliness and accuracy. Moreover, the application offers a diverse array of visualization options such as histograms, bar plots, pie charts, scatter plots, line plots, heatmaps, and box plots, enabling users to gain insights and identify patterns effectively. Downloadable outputs in various formats (CSV, PDF, PNG) further enhance usability, allowing users to share findings or conduct further analysis. Robust error handling mechanisms and enhancements in usability, scalability, and flexibility ensure that the Data Insight Pro application caters to the diverse analytical needs of users, facilitating informed decision-making processes.

I. LITERATURE SURVEY

Journal	Description	Drawbacks
Data Visualization with Python: A Project-Based Introduction by Gatto, A., & Boyarsky, A. (2022)	Offers a project-based introduction to data visualization in Python, guiding readers through hands-on projects to create effective visualizations.	Limited coverage of advanced visualization techniques and more complexity.
Seaborn: statistical data visualization by Waskom, M. L. (2021)	Discusses Seaborn, a Python library for statistical data visualization, highlighting its features and usage for creating appealing visualizations.	May have fewer customization options compared to Matplotlib.
Introducing Python: Modern Computing in Simple Packages by Lubanovic, B. (2019)	Provides an introduction to Python programming covering various aspects of modern computing with simple packages.	May lack in-depth coverage of advanced Python topics.
Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython by McKinney, W. (2017)	Focuses on data manipulation techniques using Pandas and NumPy libraries in Python, along with interactive computing with IPython.	Limited coverage of advanced data analysis and machine learning topics.
Matplotlib: A 2D graphics environment by Hunter, J. D. (2007)	Describes Matplotlib, a 2D graphics environment in Python, and its capabilities for creating visualizations.	Steeper learning curve compared to higher-level plotting libraries.
Scikit-learn: Machine learning in Python by Pedregosa, F., et al. (2011)	Introduces Scikit-learn, a machine learning library in Python, covering various machine learning algorithms and tools.	Limited support for deep learning and advanced neural network architectures.
Data Science from Scratch: First Principles with Python by Grus, J. (2019)	Offers a foundational understanding of data science principles using Python, covering topics from basic data manipulation to machine learning algorithms.	Less comprehensive compared to specialized data science textbooks.
Data Visualization with Python and JavaScript: Scrape, Clean, Explore & Transform Your Way to Better Understanding of Data by Finzer, W. (2019)	Explores data visualization techniques using Python and JavaScript, focusing on practical methods for improving data understanding.	Requires familiarity with JavaScript for full utilization of techniques.

Introduction to Machine Learning with Python: A Guide for Data Scientists by Müller, A. C., & Guido, S. (2016)	Provides a guide to machine learning using Python, covering essential concepts, algorithms, and best practices for data scientists.	Limited coverage of advanced machine learning topics and deep learning techniques.
Python Data Analytics: Data Analysis and Science using Pandas, Matplotlib and the Python Programming Language by Nelli, F. (2015)	Focuses on data analysis and science using Python libraries such as Pandas and Matplotlib, demonstrating practical applications in Python.	May lack in-depth coverage of advanced data analysis techniques.
Python Data Science Handbook: Essential Tools for Working with Data by VanderPlas, J. (2016)	Serves as a comprehensive guide to data science tools in Python, covering essential tools and techniques for working with data effectively.	May require prior familiarity with Python programming and data manipulation.
Interactive Data Visualization in Python by Chung, D. (2016)	Explores interactive data visualization techniques in Python, demonstrating methods for creating engaging and dynamic visualizations.	May have a learning curve for implementing complex interactive visualizations.
Python Data Visualization Cookbook: Over 120 Recipes to Analyze and Visualize Data by Saha, B. (2022)	Provides a collection of recipes for analyzing and visualizing data in Python, offering practical solutions to common data visualization challenges.	Recipes may be tailored to specific use cases and may require adaptation for different scenarios.
The Grammar of Graphics by Wilkinson, L. (2005)	Explores the principles of graphical representation and data visualization using the grammar of graphics approach, providing a theoretical foundation.	May require a deeper understanding of statistical graphics principles for full comprehension.
ggplot2: Elegant Graphics for Data Analysis by Wickham, H. (2016)	Describes ggplot2, a data visualization package for the R programming language, focusing on its capabilities for creating elegant graphics.	Limited coverage of Python implementation and integration with other data analysis libraries.
Storytelling with Data: A Data Visualization Guide for Business Professionals by Knaflic, C. N. (2015)	Guides business professionals in creating effective data visualizations that tell compelling stories and communicate insights effectively.	May focus more on storytelling aspects rather than technical aspects of data visualization.
Show Me the Numbers: Designing Tables and Graphs to Enlighten by Few, S. (2012)	Focuses on designing tables and graphs to effectively communicate data insights, emphasizing clarity and simplicity in data presentation.	May lack coverage of advanced data visualization techniques and tools.
Data Points: Visualization That Means Something by Yau, N. (2013)	Discusses the principles of effective data visualization and provides examples of visualizations that convey meaningful insights and stories.	May require prior knowledge of statistical concepts and data visualization techniques.
Web Application Development with R Using Shiny by Beeley, C. (2013)	Introduces Shiny, a web application framework in R, and demonstrates its capabilities for developing interactive data visualization applications.	Focuses on R programming language and may require familiarity with R for full utilization.
Taxed and Untaxed Data: A Guide for Researchers by Krause, E. F. (2016)	Provides guidance for researchers in managing and analyzing taxed and	May focus more on data management and legal

	untaxed data, offering practical advice and best practices for data handling.	considerations rather than technical aspects.

II. PROBLEM STATEMENT

The problem addressed by the Data Insight Pro system is the need for a comprehensive, integrated, and user-friendly solution for unstructured data management, analysis, and visualization. Specifically, this paper aims to solve the following problems:

- Lack of a unified platform
- Inaccessibility for non-technical users
- Ineffective assessment of data quality
- Inadequate data preprocessing capabilities
- Limited visualization options

III. MODULES AND PROJECT DESCRIPTION

Module 1: GUI Module (Tkinter):

- Provides the graphical user interface (GUI) for the application.
- Enables user interaction through buttons, dropdowns, and text fields.
- Facilitates file upload, data preprocessing, visualization selection, and download functionalities.

Module 2: Data Processing Module (Pandas):

- Handles data manipulation and preprocessing tasks.
- Reads dataset files (Excel or CSV) and displays file information.
- Analyzes data for null values, provides statistical analysis, and handles missing values.
- Splits data into training and testing sets, applies linear regression for accuracy calculation, and removes outliers.

Module 3: Visualization Module (Matplotlib and Seaborn):

- Generates various types of visualizations based on user selection.
- Supports histograms, bar plots, pie charts, scatter plots, line plots, heat maps, and box plots.
- Displays visualizations within the GUI canvas area for easy interpretation.

Module 4: File Management Module (OS and PdfPages):

- Manages file operations such as saving cleaned datasets and visualizations.
- Supports saving data in CSV format and visualizations in PDF or PNG format.
- Provides options for users to download the processed data and generated visualizations for further analysis.

Module 5: Logging and Error Handling Module:

- Logs important messages, including file uploads, preprocessing results, and visualization generation.
- Handles exceptions gracefully and provides error messages to guide users in resolving issues.

Module 6: Iterative Improvement and Feedback Module:

- Allows users to provide feedback or report issues with the application.
- Supports iterative development by incorporating user feedback and suggestions for future updates and enhancements.

ADVANTAGES:

- User-Friendly Interface
- Comprehensive Data Analysis
- Flexible Visualization Options
- Improved Accuracy Assessment
- Documentation and Reporting

V. RESULTS AND DISCUSSION

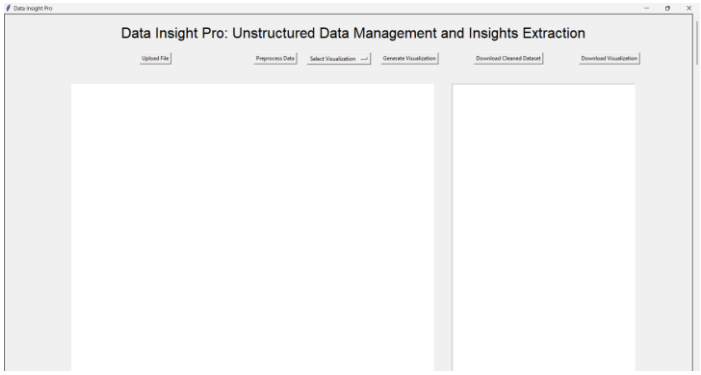

Fig 1: Data insight pro Home page

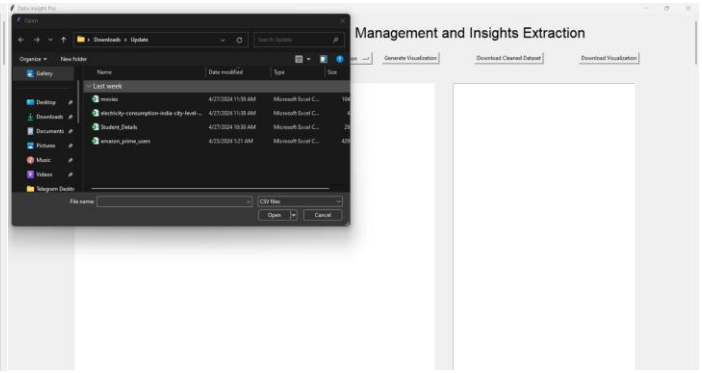

Fig 2: Uploading File

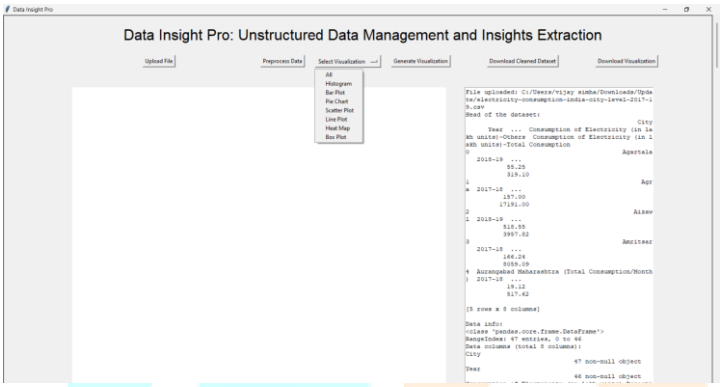

Fig 3: Select Visualization Type

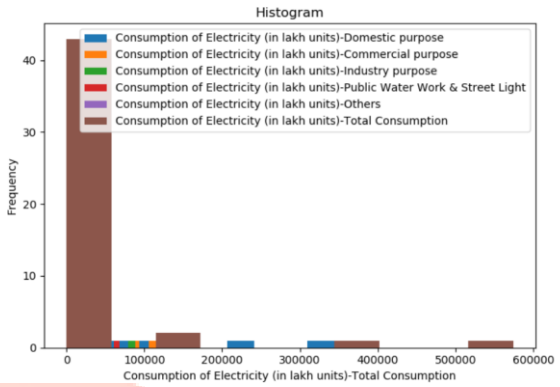

Fig 4: Generate Selected Visualization

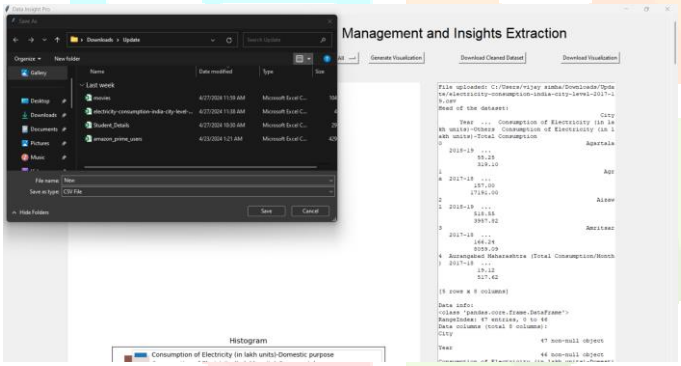

Fig 6: Save Cleaned Dataset

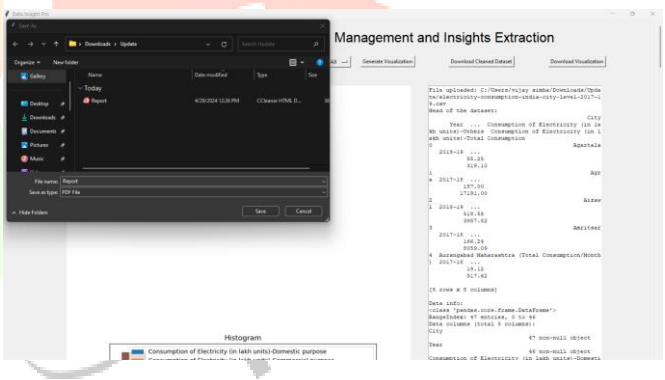

Fig 7: Save Generated Visualizations in Image (.png) or Document (.pdf)

Fig 8: Generated Visualizations in Document (.pdf)

```
File uploaded: C:/Users/vijay simha/Downloads/Update/electricity-consumption-india-city-level-2017-19.csv
Head of the dataset:
City     Year  ...  Consumption of Electricity (in lakh units)-Others  Consumption of Electricity (in lakh units)-Total Consumption
0                          Agartala  2018-19  ...                 55.25                 319.10
1                              Agra  2017-18  ...                157.00               17191.00
2                            Aizawl  2018-19  ...                518.55                3957.82
3                          Amritsar  2017-18  ...                166.24                8059.09
4  Aurangabad Maharashtra (Total Consumption/Month)  2017-18  ...            19.12            517.62

[5 rows x 8 columns]

Data info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 47 entries, 0 to 46
Data columns (total 8 columns):
City                                                                       47 non-null object
Year                                                                       46 non-null object
Consumption of Electricity (in lakh units)-Domestic purpose                46 non-null float64
Consumption of Electricity (in lakh units)-Commercial purpose              46 non-null float64
Consumption of Electricity (in lakh units)-Industry purpose                46 non-null float64
Consumption of Electricity (in lakh units)-Public Water Work & Street Light  46 non-null float64
Consumption of Electricity (in lakh units)-Others                          44 non-null float64
Consumption of Electricity (in lakh units)-Total Consumption               47 non-null float64
dtypes: float64(6), object(2)
memory usage: 3.1+ KB


Null values:
City                                                                       0
Year                                                                       1
Consumption of Electricity (in lakh units)-Domestic purpose                1
Consumption of Electricity (in lakh units)-Commercial purpose              1
Consumption of Electricity (in lakh units)-Industry purpose                1
Consumption of Electricity (in lakh units)-Public Water Work & Street Light    1
Consumption of Electricity (in lakh units)-Others                          3
Consumption of Electricity (in lakh units)-Total Consumption               0
dtype: int64

Statistical analysis:
       Consumption of Electricity (in lakh units)-Domestic purpose  ...  Consumption of Electricity (in lakh units)-Total Consumption
count        46.000000     ...        47.000000
mean      19821.499271     ...     30522.137615
std       62583.287221     ...    100845.760188
min          13.400000     ...        23.370000
25%         238.680000     ...       415.143000
50%         749.770000     ...      1946.860000
75%        4266.173413     ...      9301.958660
max      344276.000000     ...    574293.000000

[8 rows x 6 columns]

Accuracy before data cleaning: 0.78

Accuracy after data cleaning: 0.90

Outliers removed from the dataset.
Visualization generated.
```

Fig 9: Log Info saved along with the visualizations

VI. ACKNOWLEDGMENT

In conclusion, the "Data Insight Pro" project presents a comprehensive solution for data management and insights extraction, leveraging Python libraries such as Tkinter, Pandas, Matplotlib, and Seaborn. The project aims to empower users to upload, preprocess, visualize, and analyze datasets efficiently, facilitating data-driven

decision-making processes. By providing a user-friendly graphical interface, the application enables users to interact with their data seamlessly, gaining valuable insights through various visualization techniques.

Throughout the implementation process, key functionalities such as file upload, data preprocessing, visualization generation, and file management have been meticulously developed and integrated into the application. The project follows established methodologies for testing and validation, ensuring the reliability, accuracy, and usability of the final product. Through unit testing, integration testing, user acceptance testing, and performance testing, the application's robustness and effectiveness have been thoroughly evaluated.

REFERENCES

- Lubanovic, B. (2019). Introducing Python: Modern Computing in Simple Packages. O'Reilly Media.
- McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
- Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
- Grus, J. (2019). Data Science from Scratch: First Principles with Python. O'Reilly Media.
- Finzer, W. (2019). Data Visualization with Python and JavaScript: Scrape, Clean, Explore & Transform Your Way to Better Understanding of Data. Packt Publishing Ltd.
- Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.
- Nelli, F. (2015). Python Data Analytics: Data Analysis and Science using Pandas, Matplotlib and the Python Programming Language. Apress.
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
- Chung, D. (2016). Interactive Data Visualization in Python. Packt Publishing Ltd.
- Gatto, A., & Boyarsky, A. (2022). Data Visualization with Python: A Project-Based Introduction. Packt Publishing Ltd.
- Saha, B. (2022). Python Data Visualization Cookbook: Over 120 Recipes to Analyze and Visualize Data. Packt Publishing Ltd.
- Wilkinson, L. (2005). The Grammar of Graphics. Springer Science & Business Media.
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer International Publishing.
- Knaflic, C. N. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. John Wiley & Sons.
- Few, S. (2012). Show Me the Numbers: Designing Tables and Graphs to Enlighten. Analytics Press.
- Yau, N. (2013). Data Points: Visualization That Means Something. John Wiley & Sons.
- Beeley, C. (2013). Web Application Development with R Using Shiny. Packt Publishing Ltd.
- Krause, E. F. (2016). Taxed and Untaxed Data: A Guide for Researchers. Research Data Alliance.