Data Insights Explorer: Windows Application for Statistical Analysis and Integrated LLM Inquiry

- **A Project by VISWA R**

Introduction

This project aims to develop a Windows application that provides a comprehensive suite of tools for data analysis, visualization, and exploration. The application will enable users to upload datasets, perform statistical analysis, generate insightful plots, and interact with the data through a Q&A interface.

Features

1. **Home Screen:**
   ○ Upload Button: Allows users to select and upload their datasets in various formats (e.g., CSV,).
2. **Statistical Analysis:**
   ○ Descriptive Analysis: Calculates and displays essential statistics such as mean, median, mode, count of null values, and percentage of outliers.
   ○ Correlation Matrix: Computes and visualizes the correlation matrix between variables in the dataset.
   ○ Best Correlated Pair: Identifies and presents the pair of variables with the highest correlation, along with their values.
3. **Plotting:**
   ○ Plot Selection: Offers a variety of plot types (e.g., scatter plot, bar chart, histogram, line plot, box plot, pie chart, heatmap) for users to choose from.
   ○ Best Correlated Pair Plots: Generates plots of the best-correlated pair of variables using the selected plot type.
4. **Q&A:**
   ○ Chat Interface: Provides a conversational interface for users to ask questions about the dataset.
   ○ Intelligent Responses: Utilizes natural language processing to understand user queries and generate relevant answers based on the dataset.
   ○ Out-of-Scope Handling: Gracefully handles questions that are unrelated to the dataset or beyond the scope of the application.

Implementation

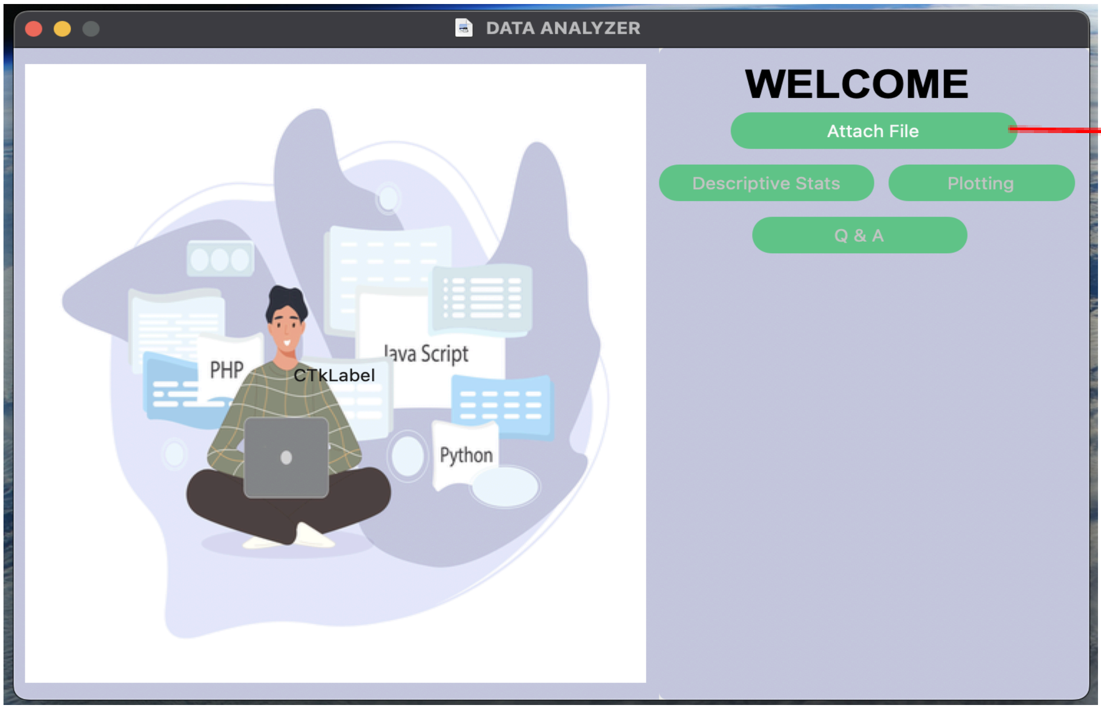
The application's interface was developed using a customized Tkinter library, providing a user-friendly graphical interface for Windows desktop environments. The statistical analysis and plotting functionalities were implemented using libraries such as NumPy, pandas, and Matplotlib, ensuring robust and efficient data processing and visualization.

For the Q&A feature, the application leverages the power of prompt engineering techniques in conjunction with Gemini AI, a sophisticated language model. This combination enables the application to understand and respond to user queries effectively, providing accurate and relevant information based on the dataset.

Benefits

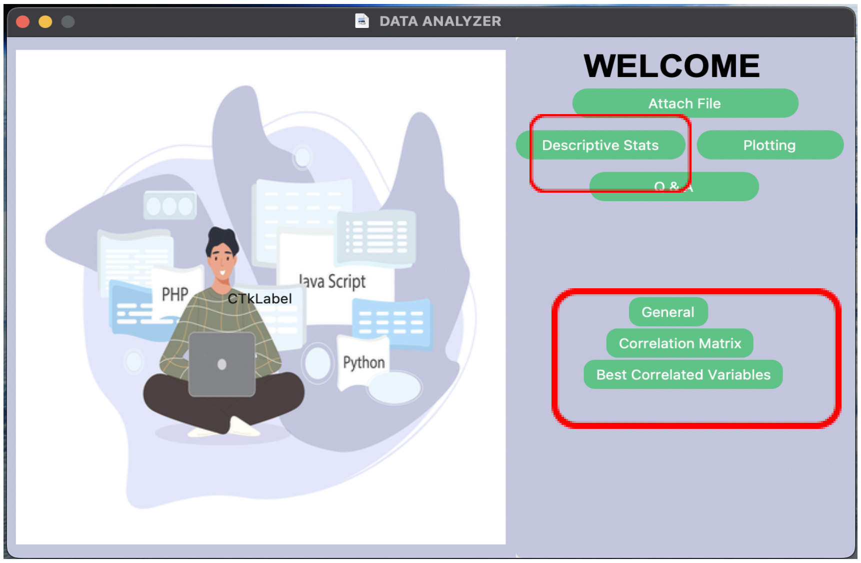
- **User-Friendly Interface:** The intuitive design will make the application accessible to users with varying levels of technical expertise.
- **Comprehensive Analysis:** The combination of statistical analysis, plotting, and Q&A capabilities will provide a holistic view of the data.
- **Interactive Exploration:** The Q&A feature will enable users to engage with the data conversationally and uncover insights.
- **Customizable Visualization:** The choice of plot types will allow users to tailor the visualization to their specific needs.

Visual Implementation :



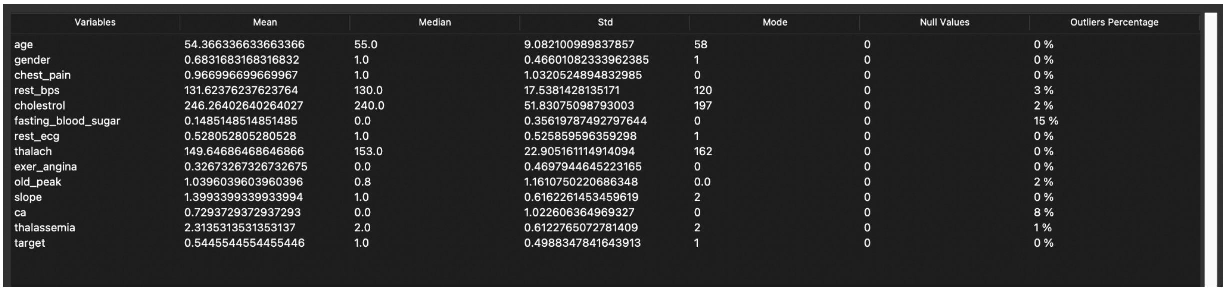
The "Attach File" button is used to upload a CSV file. Once a file is attached, the remaining buttons on the screen become activated.

DESCRIPTIVE STATS BUTTON :



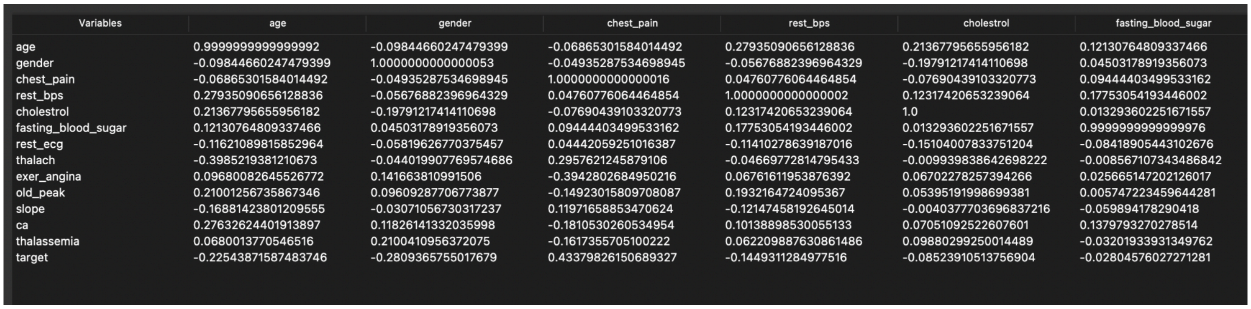
The "Descriptive Stats" button activates three additional buttons for calculating specific statistics. Clicking any other button deactivates these three buttons. The "General" button displays mean, median, mode, count of null values, and outlier percentage. The "Correlation Matrix" button displays the correlation matrix for the dataset. The "Best Correlated Variable" button displays the best-correlated pair of variables along with their values.

GENERAL BUTTON OUTPUT:



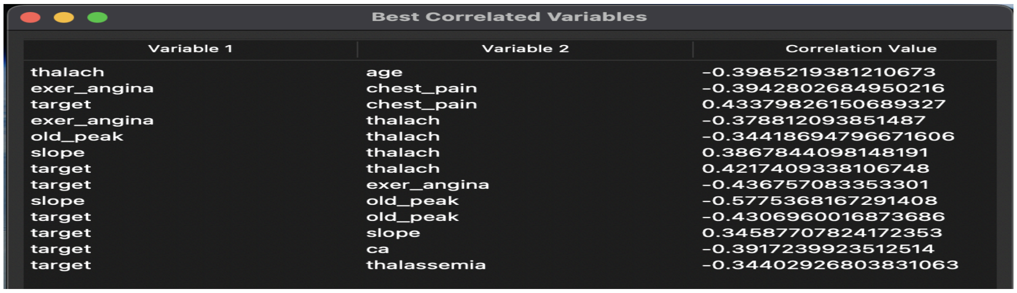
Variables	Mean	Median	Std	Mode	Null Values	Outliers Percentage
age	54.366336633663366	55.0	9.082100989837857	58	0	0 %
gender	0.6831683168316832	1.0	0.46601082333962385	1	0	0 %
chest_pain	0.966996699669967	1.0	1.0320524894832985	0	0	0 %
rest_bps	131.62376237623764	130.0	17.5381428135171	120	0	3 %
cholestrol	246.26402640264027	240.0	51.83075098793003	197	0	2 %
fasting_blood_sugar	0.1485148514851485	0.0	0.35619787492797644	0	0	15 %
rest_ecg	0.528052805280528	1.0	0.525859596359298	1	0	0 %
thalach	149.64686468646866	153.0	22.905161114914094	162	0	0 %
exer_angina	0.32673267326732675	0.0	0.4697944645223165	0	0	0 %
old_peak	1.0396039603960396	0.8	1.1610750220686348	0.0	0	2 %
slope	1.3993399339933994	1.0	0.6162261453459619	2	0	0 %
ca	0.7293729372937293	0.0	1.022606364969327	0	0	8 %
thalassemia	2.3135313531353137	2.0	0.6122765072781409	2	0	1 %
target	0.5445544554455446	1.0	0.4988347841643913	1	0	0 %

CORRELATION MATRIX BUTTON OUTPUT:



Variables	age	gender	chest_pain	rest_bps	cholestrol	fasting_blood_sugar
age	0.9999999999999992	-0.09844660247479399	-0.06865301584014492	0.27935090656128836	0.21367795655956182	0.12130764809337466
gender	-0.09844660247479399	1.0000000000000053	-0.04935287534698945	-0.05676882396964329	-0.19791217414110698	0.04503178919356073
chest_pain	-0.06865301584014492	-0.04935287534698945	1.0000000000000016	0.04760776064464854	-0.07690439103320773	0.09444403499533162
rest_bps	0.27935090656128836	-0.05676882396964329	0.04760776064464854	1.0000000000000002	0.12317420653239064	0.17753054193446002
cholestrol	0.21367795655956182	-0.19791217414110698	-0.07690439103320773	0.12317420653239064	1.0	0.013293602251671557
fasting_blood_sugar	0.12130764809337466	0.04503178919356073	0.09444403499533162	0.17753054193446002	0.013293602251671557	0.9999999999999976
rest_ecg	-0.11621089815852964	-0.05819626770375457	0.04442059251016387	-0.11410278639187016	-0.15104007833751204	-0.08418905443102676
thalach	-0.3985219381210673	-0.044019907769574686	0.2957621245879106	-0.04669772814795433	-0.009939838642698222	-0.008567107343486842
exer_angina	0.09680082645526772	0.141663810991506	-0.3942802684950216	0.06761611953876392	0.06702278257394266	0.025665147202126017
old_peak	0.21001256735867346	0.09609287706773877	-0.14923015809708087	0.1932164724095367	0.05395191998699381	0.005747223459644281
slope	-0.16881423801209555	-0.03071056730317237	0.11971658853470624	-0.12147458192645014	-0.0040377703696837216	-0.059894178290418
ca	0.27632624401913897	0.11826141332035998	-0.1810530260534954	0.10138898530055133	0.07051092522607601	0.1379793270278514
thalassemia	0.0680013770546516	0.2100410956372075	-0.1617355705100222	0.062209887630861486	0.09880299250014489	-0.03201933931349762
target	-0.22543871587483746	-0.2809365755017679	0.43379826150689327	-0.1449311284977516	-0.08523910513756904	-0.02804576027271281

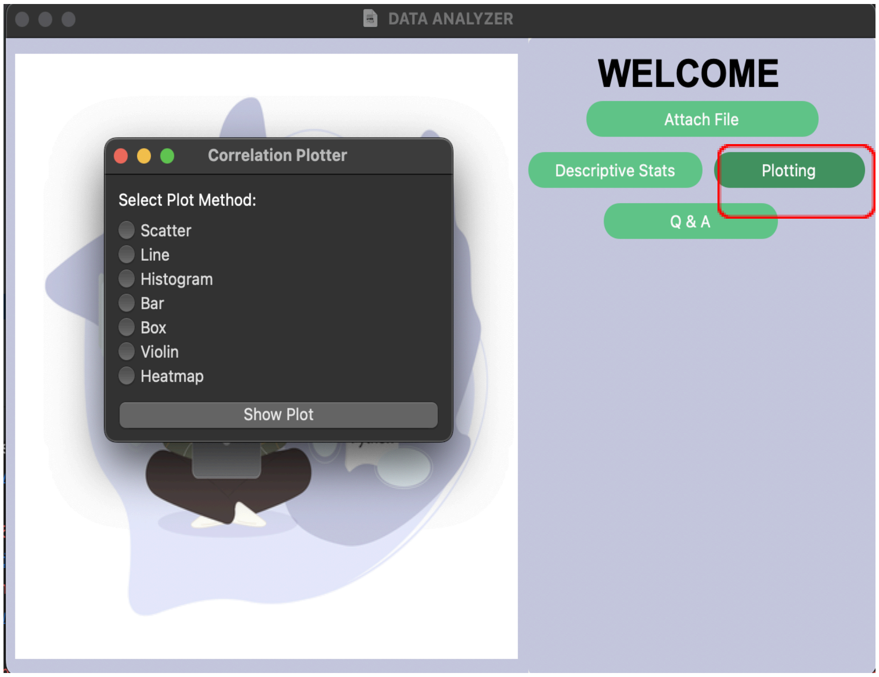
BEST CORRELATED MATRIX OUTPUT:



Best Correlated Variables

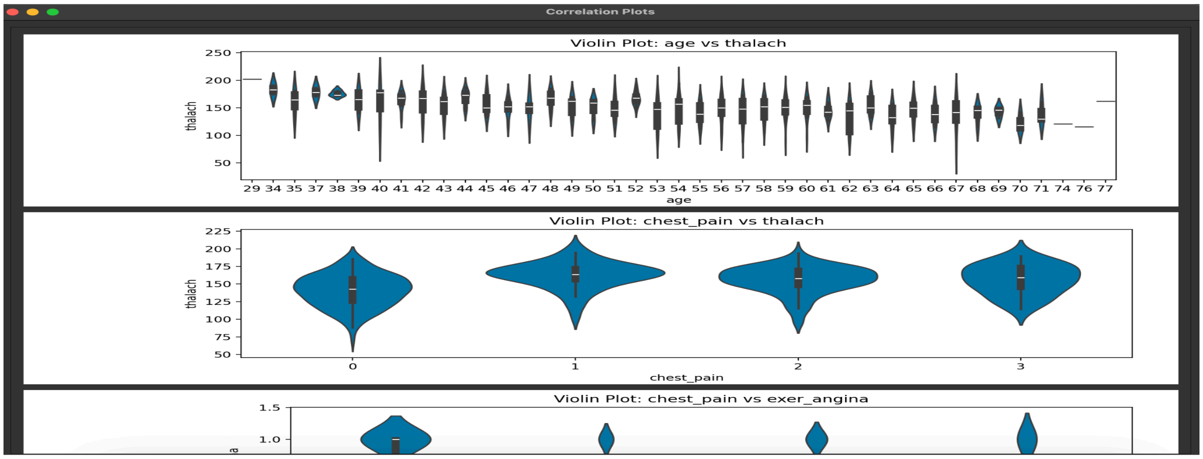
Variable 1	Variable 2	Correlation Value
thalach	age	-0.3985219381210673
exer_angina	chest_pain	-0.3942802684950216
target	chest_pain	0.43379826150689327
exer_angina	thalach	-0.378812093851487
old_peak	thalach	-0.34418694796671606
slope	thalach	0.3867844098148191
target	thalach	0.4217409338106748
target	exer_angina	-0.436757083353301
slope	old_peak	-0.5775368167291408
target	old_peak	-0.4306960016873686
target	slope	0.34587707824172353
target	ca	-0.3917239923512514
target	thalassemia	-0.34402926803831063

PLOTTING BUTTON :

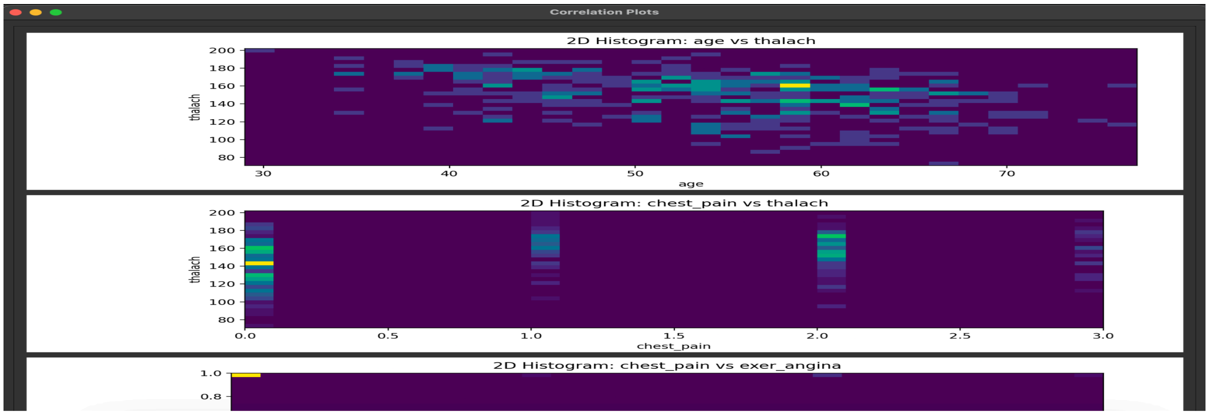


Clicking the "Plotting" button opens a pop-up screen with seven plot options. Selecting an option and clicking the "Show Plot" button displays the best-correlated plot in a new window. The window includes a scrollbar to view all generated plots within the same window.

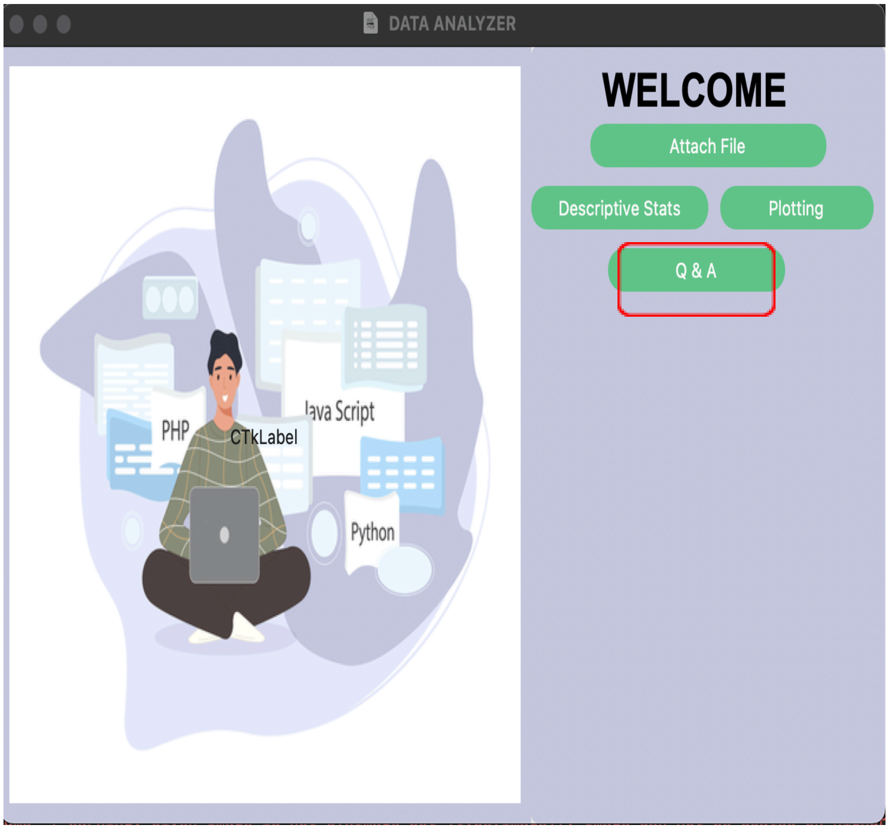
VIOLIN PLOT:



HISTOGRAM PLOT:

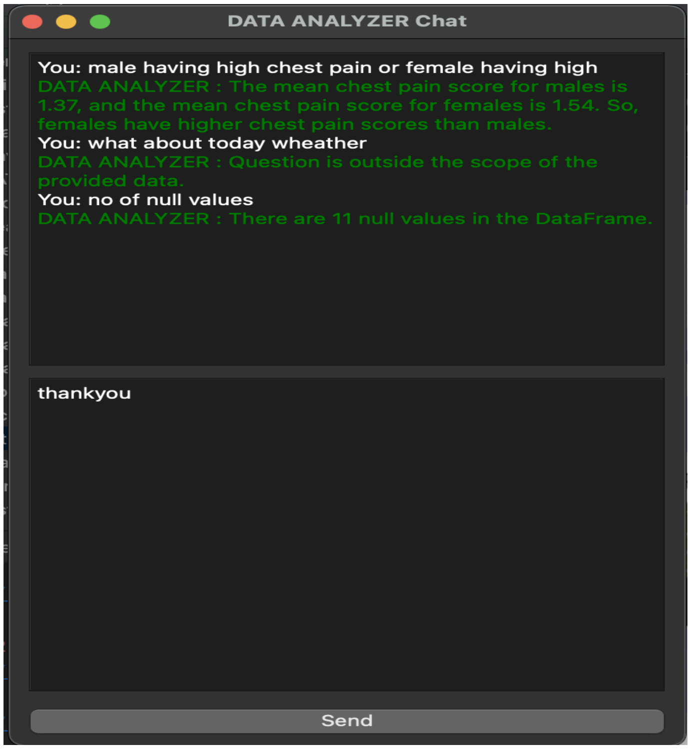


Q & A BUTTON :



Clicking the "Q&A" button opens a new pop-up window where users can type any question related to the uploaded CSV data. The chatbot will provide relevant answers based on the dataset. For questions unrelated to the data, the chatbot will respond with "Out of Scope."

CHAT BOX:



You: male having high chest pain or female having high
DATA ANALYZER : The mean chest pain score for males is 1.37, and the mean chest pain score for females is 1.54. So, females have higher chest pain scores than males.
You: what about today wheather
DATA ANALYZER : Question is outside the scope of the provided data.
You: no of null values
DATA ANALYZER : There are 11 null values in the DataFrame.

thankyou

CONCLUSION:

This project will deliver a powerful and versatile Windows application that empowers users to analyze, visualize, and interact with their data effectively. By combining statistical rigor with intuitive design and interactive features, the application will cater to a wide range of users, from students and researchers to data analysts and business professionals.

LINKS:

PROJECT SOURCE CODE LINK : https://github.com/Viswa792/TENSORGO-ASSIGNMENT

VIDEO DEMONSTRATION LINK :
https://drive.google.com/file/d/1Us7fN1pPp9oYJw3LTGOz6fdqTT0mfNlR/view?usp=sharing