VizWhiz: An Automated Dashboard Creation Solution for Biomedical Data Visualization

Sheraz Hassan
Georgia Institute of Technology
shassan74@gatech.edu

Marianne Al Haj
Georgia Institute of Technology
mariannealhaj@gatech.edu

Temi Owopetu
Georgia Institute of Technology
towopetu3@gatech.edu

Taylor Graham
Georgia Institute of Technology
tgraham47@gatech.edu

Abstract

As organizations rely on data- driven decision making, it is paramount to have clean and accurate data. However, raw datasets have issues like missing values, duplications and more. Also, this data is used to communicate and summarize insights using dashboards. Although there are multiple tools for data cleaning and generating dashboards, it is time consuming, error prone or hard to use. To combat these issues, we developed VizWhiz, an interactive and easy to use tool for automated data cleaning and visualization. To demonstrate the functionality of VizWhiz, it was applied to a dataset containing body measurements. It resulted in clean data as well as created multiple insightful visualizations like histograms, box plots, scatter plots and violin plots.In this paper, we talk about the flexibility and robustness of VizWhiz.

I. Introduction

As the world becomes more data-driven, institutions depend on accurate and clean data to drive decision-making. However, raw data is muddled with errors, missing values, inconsistencies, inaccuracies, or incomplete data. This greatly hinders data analysis and can lead to inaccurate results or poor decision-making.[1,2] Once the data is analyzed, the results can be used for storytelling through dashboards. Dashboards are a powerful tool that allows users to understand and analyze complex data at a glance. It also allows for real-time updates and quicker identification of trends and key performance indicators while saving time as users don't have to sift through multiple reports[3,4]. The dashboard can display inaccurate or misleading stories without clean and reliable data. Consequently, dashboards are only fully effective when accurate and clean data is used to build them.

There are numerous tools for data cleaning and building a dashboard. Due to its manual nature, it tends to be time-consuming and error-prone. For users who lack the technical skills, these tools can be overwhelming and hard to use. Furthermore, non-technical users are often unfamiliar with data visualization principles like the Gestalt principles which can limit their ability to build effective dashboards and efficiently tell a story. To address these challenges, the team has developed an easy-to-use tool, VizWhiz, that automates data cleaning by allowing users to choose different data cleaning methods that will be applied to their raw dataset, ensuring the data is accurate as well as ready for analysis. VizWhiz also allows users to generate dashboards by simply choosing which variables should be displayed. It guides users in selecting from a range of distribution and comparison plots to display their selected variables appropriately. VizWhiz reduces the manual effort, and lowers the technical barrier that some users face while empowering users to transform raw data into meaningful insights.

This paper outlines the development and implementation of VizWhiz, its features, the algorithms built for data cleaning as well as dashboard generating tools. Our objective is to enhance the accessibility of data analytics by making it less time-consuming and easier to use.

II. Methods

A. System Architecture and Overview.

VizWhiz was developed using Python. For the front-end interface, streamlit, an open-source Python library used for making custom web applications for data science, was used.[5] Pandas, Numpy and Scipy were used for data manipulation, numerical and statistical operations.[6] For generating the dashboards, Plotly, an open-source library for creating interactive data visualizations was used.[7]

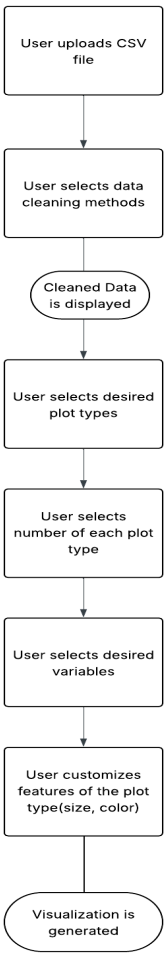

Figure 1: Flowchart of VizWhiz. The user uploads a CSV file, selects automated data cleaning methods, selects desired plots and variables, customizes visualization features, and generates a final visualization.

Figure 1 is a system flowchart that shows how different parts of the tool interact with each other. The user uploads the dataset and selects the data cleaning methods to be applied to their data. The tool outputs the clean dataset so the user can review and ensure all essential information is present. After, the user selects what plot types are desired and how many of each plot should be displayed. Then, the user selects the variables for each plot and customizes different features like size, color, scale, and more. Each step provides real-time previews to improve usability for non-technical users.

B. Automated Data Cleaning

Once a CSV file is uploaded to VizWhiz, it is read and different data-cleaning methods are displayed. It processes the data and determines which variables would be used for grouping and plotting. Currently, VizWhiz supports three methods of data cleaning.

1. Removing Rows with Null Values

Missing data can skew the results of the analysis, which damages its credibility or leads to failure when generating visualizations.[8] It can also negatively affect decision-making by providing false or misleading insights.[8] To prevent this, users can either decide to remove rows that contain empty values or fill in the empty rows with the mean, median and mode value. In VizWhiz, this is achieved using the "dropna" function.

2. Removing Duplicate Rows

Like missing data, duplicate entities can lead to biased results which destroy the reliability of the analysis.[9] It also leads to inaccurate visualizations as the metrics are inflated and trends and distorted, thereby also negatively affecting decision making.[9] Deletion of duplicate rows is accomplished in VizWhiz using the "drop_duplicates" function. It allows users to select which columns should be screened for duplicates

3. Remove outlier

Outliers are values that differ significantly from the other data points in the dataset.[10] They can be unusually small or large, deviating from the overall trend of the dataset.[10] Outliers affect the scale, shape or trends of visualizations thereby leading to misrepresentations and incorrect conclusions. [11] Therefore, VizWhiz can detect and remove outliers based on the Interquartile Range(IQR) method which identifies outliers by determining the upper and lower bounds around the median of the data using a the user's specified threshold.[12]

C. Visualization

After the desired data cleaning methods have been applied, the user can select the preferred plot types and choose how many of each plot they desire. Currently, to accommodate for both categorical and numerical variables VizWhiz offers 5 different plot types.

1. Box Plot

Users will select the variables to be displayed and choose if the mean and outliers are shown. The box will be horizontal if one variable is selected or vertical if multiple variables are chosen. In addition, users can decide whether the whisker extent should be 1.5 IQR or the minimum and maximum values.

2. Histogram

Users can select which variable should be depicted as well as apply a Logarithmic scale if needed. The user has the option to overlay the Kernel Density Estimation(KDE) curve on the histogram.

3. Line Plot

VizWhiz allows users to assign one variable to the x-axis and multiple variables to the y-axis. The axes can be scaled logarithmically and the user can choose to display individual data points with a marker.

4. Scatter Plot

Users can select which variable should be the x-axis and which variables will be the y-axis. Additionally, a regression line can be applied if desired and logarithmic scaling is available for both axes.

5. Violin Plot

VizWhiz allows users to display multiple variables using the violin plot. It will be horizontal if only one variable is being displayed and vertical if multiple variables are being depicted. In addition, users can display an embedded box plot which shows the median and the interquartile range can be present as well as the KDE curve. Moreover, the width of the violin plot can be scaled by either width, area or count.

Moreover, all plot types can be customized by color, size, and transparency while selected variables can be grouped by other variables. To demonstrate the functionality of VizWhiz, it was applied to a sample dataset that contains information about many individual's body measurements. The data was cleaned using various methods and multiple visualizations were generated to showcase the capability of VizWhiz.

III. Results

A. Data Cleaning

VizWhiz users select which data cleaning operations are applied to their data. The top of the VizWhiz dashboard begins with a visualization of the uploaded raw data in the form of a table. Directly underneath VizWhiz displays another table that is populated with the cleaned version of the user's data. Users can see exactly how their data has been cleaned, this helps users determine the right cleaning settings to establish to maintain data volume without sacrificing data integrity.

B. Visualizations

Once a user selects the plot types they want on their VizWhiz dashboard visualizations will appear beneath the tables displaying the user's data. Users also get the choice from within their dashboard to choose how many of a specific visualization type they would like to see. Each visualization type comes with a custom selection of features for users to choose from to customize visualizations to user needs.

1. Box Plot

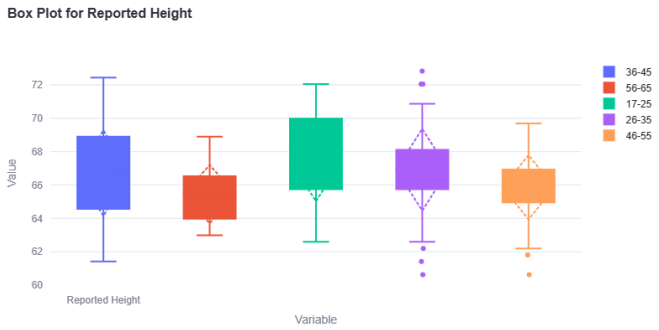

Figure 2: Box plot created and saved from VizWhiz. Box plot displays reported heights grouped by age range.

VizWhiz can produce highly detailed box plots, with just a few user selections. The boxplot shown in *figure 2* was made from the user selecting reported height as the column of interest, allowing all sample points to be displayed, and grouping the reported height data by age range. Data is clearly displayed for each age group, including outliers within the 26-35 and 46-55 age groups. The output displayed here is interactive from within the VizWhiz application but figures can easily be saved and exported as png images from the VizWhiz dashboard directly.

2. Histogram

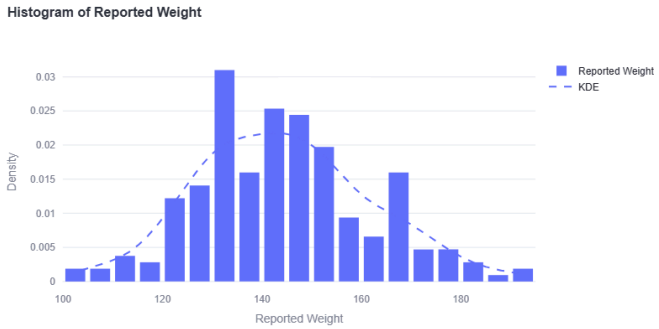

Figure 3: Histogram created and saved from VizWhiz. Histogram plot displays distribution of reported weights.

Histograms produced using VizWhiz are detailed and include optional kernel density estimations (KDE curve) to further inform upon the distribution of the selected data. The histogram shown in *figure 3* was created by choosing "Reported Weight" as the column of interest, selecting to show the KDE curve, and choosing not to group by any additional variables.

3. Line Plot

VizWhiz includes the option to create line plots, as line plots are best utilized to show time series relationships; we have elected not to include an example line plot since the selected dataset does not include time series data.

4. Scatter Plot

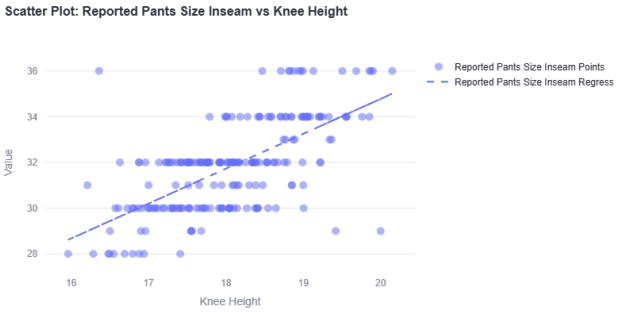

Figure 4: Scatter plot created and saved from VizWhiz. Scatter plot displays Pants Size Inseam and Knee Height and the linear regression line associated with these variable's relationship.

Scatter plots are useful for gaining information on the association between variables. In the scatter plot shown in *Figure 4* users can see a moderate positive relationship between knee height and pants size inseam reflected in the scattered data points. In the creation of this VizWhiz plot the user selected to add a regression line to the scatter plot, in this plot we can see that the regression line reflects the positive relationship that these variables have with one another.

5. Violin Plot

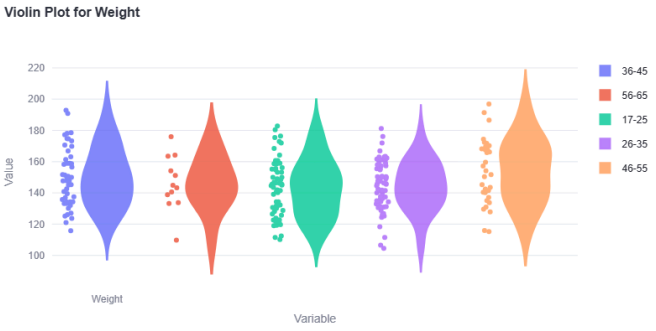

Figure 5: Violin plot created and saved from VizWhiz. Violin plot displays weight data points grouped by age range.

The violin plot displayed in *Figure 5* shows weight data grouped by age range, alongside the violin plots users can see the original data points to provide more insight into the distribution of the data.

IV. Discussion

VizWhiz can be used to create a wide variety of visualizations for a selected dataset. By offering highly customizable visualizations, users can identify different trends within their data. Each visualization type that VizWhiz offers users is specialized to identify certain features of the dataset.

Boxplot visualizations are best suited to identify the spread of the data while gaining information about key metrics like median, min, and max. Another key piece of information that can be visualized on box plots within VizWhiz is the existence of outliers within a dataset or grouping of data. If users were not as interested in the metric values that box plots provide they could choose instead to visualize their data with a violin plot within

VizWhiz which combines the visualization of spread from the boxplots with visualizations of data density.

Users who want to visualize the distribution of a variable may find that their best choice is to utilize the histogram for data visualization, VizWhiz offers the optional inclusion of the KDE curve that further advises users on the distribution of the data within their dataset.

It is very common to look at the association and correlation of 2 or more variables using data visualizations. VizWhiz allows users to view this information with either scatter plots or line plots. VizWhiz offers the option to use line plot visualizations which is most useful for visualizing time series associations. If the data users are interested in the associations between 2 or more variables that are not time series data a scatter plot will reveal the relationship between variables. VizWhiz offers optional linear regression lines for display on scatter plots which is a helpful additional tool for revealing the relationship between variables.

In potential future iterations of VizWhiz there could be developments in a few areas including accepting more data formats and visualization upgrades. VizWhiz could be improved by allowing for Excel sheet uploads (.xlsx) or accepting data from database connections. To improve VizWhiz visualization, a future direction could include automatically creating a dashboard of the created visualizations that is available for user download.

V. Conclusion

This data visualization tool fulfills a need for easy and interactive visualization creation. VizWhiz is easily customizable for biomedical data visualization purposes across a wide range of applications. With built-in features for data cleaning and visualizations of many types, VizWhiz allows users to capture data trends to yield results and conclusions from data in just a few minutes by reducing the technical skills required to create meaningful visualizations. VizWhiz removes a majority of the manual effort of biomedical data visualization by reducing user input to the simple task of variable selection and fully automating the visualization creation process beyond that. VizWhiz is a helpful tool that will make the process of data visualization easier than ever in addition to making biomedical data visualization more accessible to individuals of all technical backgrounds.

Author Contributions

Sheraz Hassan was responsible for coding the application. Marianne Al Haj was responsible for our presentation and demo for presentation to BDV 8813 class. Temi Owopetu and Taylor Graham split the labor required for the compilation of project resources for the paper.

References

1. Gupta A. *The 7 most common data quality issues*. Collibra. 2022. URL: https://www.collibra.com/blog/the-7-most-common-data-quality-issues (Accessed 25 April 2025).

2. Why You Shouldn't Build Reports From Raw Data | Blog | Fivetran. Www.fivetran.com. n.d. URL: https://www.fivetran.com/blog/dont-build-reports-raw-data.

3. Team ThoughtSpot. *Data visualization dashboard: Examples, benefits, and more*. ThoughtSpot. 2024. URL: https://www.thoughtspot.com/data-trends/dashboard/data-visualization-dashboard.

4. Domo. *Domo Resource - Data Visualization Dashboards: Benefits and Examples*. Domo.com. 2025. URL: https://www.domo.com/learn/article/data-visualization-dashboards.

5. Snowflake Documentation. *About Streamlit in Snowflake | Snowflake Documentation*. Snowflake.com. 2024. URL: https://docs.snowflake.com/en/developer-guide/streamlit/about-streamlit.

6. LinkedIn. *What are the differences between pandas, NumPy, and SciPy for data manipulation?* Linkedin.com. 2025. URL: https://www.linkedin.com/advice/3/what-differences-between-pandas-numpy-scipy-data-manipulation-eyvke (Accessed 28 April 2025).

7. Plotly. *Plotly Python Graphing Library*. Plotly.com. 2023. URL: https://plotly.com/python/.

8. Kaur T. *The Impact of Missing Data on Statistical Analysis and How to Fix It*. Medium. 2025. URL: https://medium.com/@tarangds/the-impact-of-missing-data-on-statistical-analysis-and-how-to-fix-it-3498ad084bfe.

9. FanRuan. *Understanding Data Duplication and Its Impact*. Fanruan.com. 2024. URL: https://www.fanruan.com/en/glossary/big-data/data-duplication.

10. GraphPad. *Outlier calculator*. Www.graphpad.com. 2024. URL:https://www.graphpad.com/quickcalcs/grubbs1/.

Simplilearn.

11. *Outliers in Data: Identification and Impact Revealed!* Simplilearn.com. Simplilearn; 2024. URL: https://www.simplilearn.com/outliers-in-data-article.

Interquartile Range to Detect Outliers in Data.

12. GeeksforGeeks. 2020. URL: https://www.geeksforgeeks.org/interquartile-range-to-detect-outliers-in-data/.