Through the Lens of Visualization: Exploring Univariate and Multivariate Data for Better Data Understanding.

***Abstract*—Through Graphical representation, Data visualization can help people to understand information about dataset. It transforms raw data into comprehensible visuals which facilitate easier understanding and interpretation of the dataset. Focusing on both univariate and multivariate data visualization techniques, this paper investigates how visualization aids to get meaningful knowledge about the dataset. By exploring univariant data, insight can be gained about their distribution, trends, and outliers. On the other hand, multivariant data visualization techniques allow to identify the relationships and the comparison within multiple attributes. The purpose of the research paper is to explore the pivotal role of data visualization in comprehensive understanding and exploration of the dataset. The findings will help to gain insights about any complex data set and identify the most efficient ways to visualize data for different data types.**

Keywords— ***Data Visualization, Data Visualization Techniques, Univariant Data, Multivariant Data, Data insights.***

I. INTRODUCTION

In today's age of big data, working with big data volume across various domains is not easy. Navigating through a large quantity of data can be challenging. As datasets grow larger and more complex, it has become very important to explore data in meaningful and efficient ways and understand the data.

That's where data visualization comes in handy. Data visualization is an important part of data exploration in data science. The goal of visualization is to make the data more understandable and to give information about what we are working with. Visualizations can successfully explain data to all kinds of audiences without any technical knowledge. It allows complex results to be presented in a simple manner. Moreover, visualizing data provides a perspective on data by transforming raw data into meaningful charts, graphs, and plots. Data visualization helps to explore the data by identifying patterns within the data, spotting the anomalies which may affect the data, to understand the data distribution and relation between the variables.

Although data visualization can be very helpful to explore the data, it is important to choose the most sufficient technique for a certain sort of with many visualization techniques available. The paper is structured as follows: Literature Review explains the research already done on various data visualization techniques and the importance of data visualization. Methodology explains the techniques used in this research including the explanation of the data type, data visualization techniques chosen for various data types. The Results and Discussion explain the findings of the research. Finally in Conclusion, the outcome of the research is discussed.

II. LITERATURE REVIEW

The main objective of the research is to use different ways of data visualization to extract information about a dataset, the usefulness of the usefulness and the limitation of the techniques. Data Visualization has been used from the ancient age to represent any kind of information. In their research, Saccenti et al. (2013) explored the intricacies of univariate and multivariate analysis in metabolomics research [1]. Through a meticulous examination, they underscore the importance of integrating univariant and multivariant approaches to comprehensively analyze metabolomics data. Univariate analysis allows for the identification of individual metabolites exhibiting significant variations across experimental conditions or groups, while multivariate analysis unveils complex interactions between multiple metabolites, offering a holistic view of metabolic profiles and underlying biological processes. Gosink et al. (2011) aimed to use data visualization for the purpose of query-driven data [2]. They analyzed how to use different data visualization techniques such as KDE, PCA and MDS to gain insights about query-driven data such as hurricane dataset and methane dataset. They mainly focused on uncovering the complex patterns within the dataset. While working on data visualization, they also mentioned why it is important to map out the process of data visualization to do the work efficiently. Qin et al. (2020) divided the data visualization process into three parts to make the work more effective based on data management [3]. Firstly, the discussed why it is important to understand the requirements. Secondly, based on the requirements, they discussed the data visualization approaches for the specific requirement and finally, they gave recommendation based on the visualization approaches. Their work reflects why data visualization is important to the commercial field.

Li et al. (2020) discussed the relation between data visualization approaches and human history [4]. They mentioned the fact that data visualization has been used from the ancient age to express human thoughts. They also

mentioned why data visualization is a more sufficient way to gain information because images can help to gain more insights. Moreover, they discussed the traditional ways to represent data graphically such as using Maps, Scatterplots, Charts, Tables, and diagrams. Furthermore, they examined human perceptions with data visualization examining the importance of data visualization for people. Stengel et al. (2020) discussed how any visualization technique should be presented and the importance of representing any graphical information in a meaningful way [5]. Firstly, they mentioned that the density of data in an image must be low because too much data can hamper the purpose. Secondly, the ink to data ratio must be appropriate. Finally, the visualization must be properly labelled because without labeling it is not possible to give any information about the representation. They also mentioned how visualization can be very important for the scientific community. Overall, it is important to map out the process of data visualization and the techniques before the process.

Mishra et al. (2022) presented the importance of SIV model for data representation of graph. [6]. Summarization can help better analysis of the data and remove unwanted data. Interpretation helps to figure out the hidden values of a graph. They used various GNN (Graph Neural Networks) to represent information about a graph. Finally, they explained the importance of visualization of graphs and discussed how using visualization, a person can gain insight about any graph. Urwin (2020) gave a basic introduction about data visualization and the importance of data visualization [7]. The history of data visualization in the history of mankind is discussed by him. Also, He gave a brief discussion about presentation graphics and exploratory graphics. Presentation graphics use only one graph to represent information and it must be well defined. Exploratory graphics use more than one graph to represent data. Furthermore, he highlighted the significance of visualization in the research field and mentioned how graphical representation can make data more understandable. He also mentioned why understanding about data visualization tools is important because of the nature of information extracted by different representation can be different.

Srivastava (2023) gave an overall discussion of the uses of data visualization in different fields of our daily life [8]. Firstly, she discussed about the different types of tools use for data visualization such as spreadsheet, data programming libraries and data visualization softwires. For various fields such as humanities, sports, healthcare, environment science, she discussed how to use different data visualization tools to extract information. She discussed the importance of using the right tools for graphical representation of data to extract the right information. Yalim et al. (2023) explored why data visualization is important for the research community and how to choose the right technique for specific purpose of gaining insights [9]. All the data visualization techniques do not represent the same information about a dataset. They discussed this topic using three types of data including numerical, categorical and time-series. Although a bar plot may give the frequency of data in a particular range, the density of data can't be particularly found using bar plot. For that the use of density plot is a must. Overall, they illustrated the different techniques of data visualization to extract different information about the data such as bar plot, density plot, box plot, stacked bar plot, pie chart, heatmap, line plot and wavelet analysis. Although they used three different datasets for the purpose of visualization, but they successfully presented how data visualization techniques can differ one from another and how to choose the right visualization technique for information extraction purpose. Many programming languages offer built-in visualization libraries. Han et at. (2023) gave insights about the different tools of data visualization using Python [10]. They mentioned how different built in libraries in data programming languages like Python can help a person to easily represent information. They used the built-in libraries in python like Matplotlib and Seaborn to represent data.

To sum up, there are various techniques to represent data visually and graphically. data visualization techniques can be very useful to extract information about a dataset. The insight of dataset can be gained and beneficial for various purposes. But from the literature review, it can also be said that it is important to choose the right technique because all techniques do not necessarily represent the same information.

III. METHODOLOGY

This research works with two data types: univariant and multivariant. The aim of the research is to show how visualizing data can help to understand the characteristics of these data types. The objectives of the research are to: (1) Understand what are univariant and multivariant data; (2) literature review on existing research on data visualization techniques and their importance; (3) to apply the visualization techniques and identify the appropriate data visualization technique for specific purpose; (4) The information that can be learnt from each data visualization technique.

Univariant data visualization only involves the examination of a single attribute. It solely focuses on the distribution or patterns within one variable and does not consider the relationship with other attributes. Furthermore, Univariant data can be categorized into two types, Continuous and Categorical. In the research histogram, density, dot plotting, boxplot, violin plot, ECDF, and Q-Q analysis were used for the purpose of visualizing continuous data. For categorical data bar plot, pie chart and tree map were used.

Multivariant data visualization involves two or more attributes at a time. It helps to understand the relation and interactions between more than one attribute. This data visualization can help to identify how changes in one attribute can affect the others. In this research, scatter plot, regression plot, correlation heatmap, bubble plot, violin plot and data comparison between two instances using bar plot were used.

For data visualization purposes, the dataset Kaggle Red Wine Quality was used. Although all the attributes of the dataset are continuous, the "quality" attribute which is the target attribute, was converted in categorical using four values "bad", "average", "good" and "excellent" for the purpose of categorical data visualization. As there were no missing value values, the dataset is considered clean.

This research outlines a systematic and simple approach to extract information about the characteristics of univariant and multivariant data though various visualization techniques. Through the application of various visualization methods to the Kaggle Red Wine Quality dataset, ranging from histograms and scatter plots to regression plots and correlation heatmaps, this study seeks to elucidate the information that can be derived from each technique. Moreover, the cleanliness of the dataset ensures the reliability of the findings obtained through visualization analysis.

IV. Results And Discussions

A. *Univariant Data Visualization Techniques:*

Histogram is one of the most basic techniques to represent data. Fig. 1 is the histogram plot of pH attribute showing the symmetric distribution, central tendency and data distribution across various ranges which gives hints about data patterns. It also represents the frequency of data across multiple ranges and potential outliers.

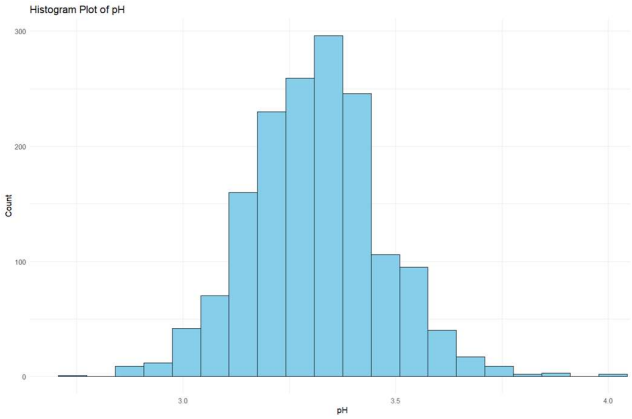

Figure 1: The Histogram Plot of pH Attribute.

But histogram plot is sensitive to the bin size which may hamper the visualization. That's where density plot comes in handy. It is visually more appealing. Fig. 2 shows a better form of data visualization of pH data with the bandwidth or bin size not affecting the data visualization. It can be easily determined that the pH attribute is almost zero skewed meaning perfectly symmetrical. Unfortunately, the density plot does not provide any frequency count on the plot.

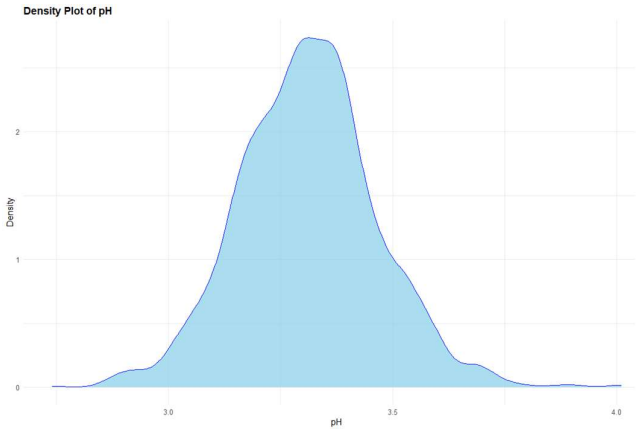

Figure 2: The Density Plot of pH Attribute.

Dot plot is another technique in which data points are shown individually as dots. Fig. 3 represents the dot plot for the data of pH attribute. It shows each data point as a dot and shows how the data is distributed throughout the range of pH and helps to organize and differentiate the data points.

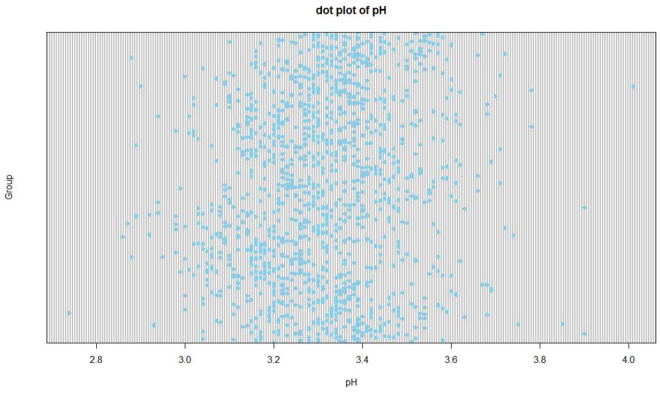

Figure 3: Dot plot of pH Attribute.

In the distribution of data for each attribute, there can be some outliers which can hamper the dataset. It is not possible to detect the outliers by just looking at the data. Fig. 4 uses the boxplot visualization technique to detect the outlier of the Citric Acid attribute using dots. Boxplot visualization technique also includes the median of the attribute, the quartiles, and the interquartile range of the attribute. It can be seen there is an outlier in the boxplot of Citric Acid boxplot. Using boxplot, it is not possible to determine the data distribution of attribute properly. That's where violin plot can offer a more informative and visually appealing alternative. In Fig. 5, the violin plot of Citric Acid attribute indicates the median and the quartiles along with density of the data. A violin plot combines elements of both a density plot and a box

plot to provide a comprehensive visualization of the distribution of data.

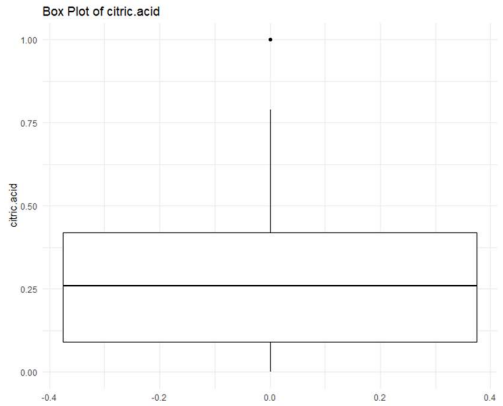

Figure 4: The Boxplot of Citric Acid Attribute.

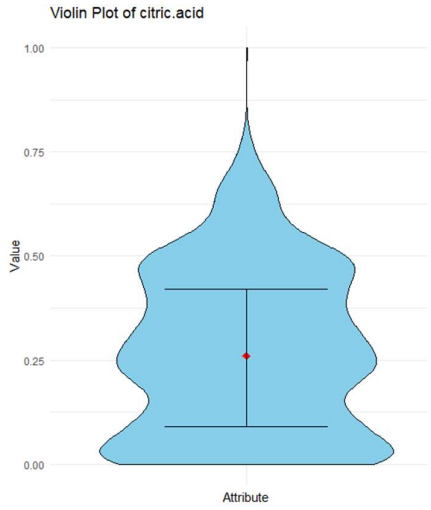

Figure 5: The Violin Plot of Citric Acid Attribute.

Fig. 6 depicts the Empirical Cumulative Distribution Function (ECDF) plot to visualize the distribution of alcohol attribute. The technique allows to see how the data is spread out across its range and provides insights into its shape, spread and their cumulative probability.

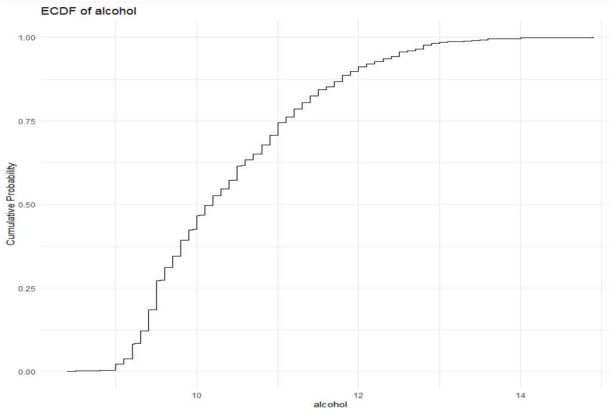

Figure 6: The ECDF Plot of Alcohol Attribute.

Fig 7 shows the Quantile-Quantile (Q-Q) plot of Total Sulfur Dioxide attribute which indicates if the distribution of the data is normal or not comparing with the theoretical distribution. Q-Q plots are commonly used in statistics to assess whether the observed distribution of data matches the expected distribution under a particular assumption and can identify potential outliers. For the Total Sulfur Dioxide attribute, it can be said, the distribution is similar with the theoretical quantiles but there are some outliers.

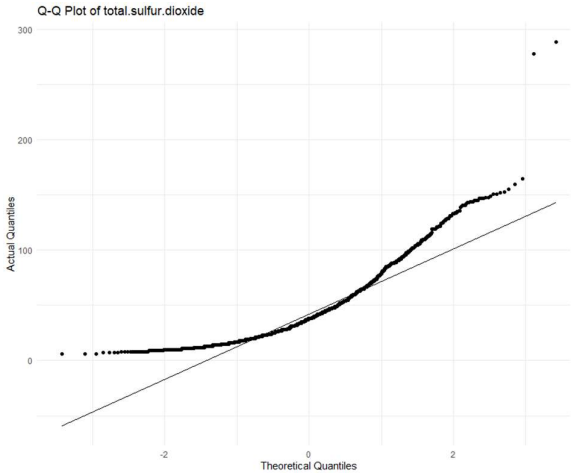

Figure 7: The Q-Q Plot of Total Sulfur Dioxide Attribute.

Bar plot is a traditional type of plotting that presents categorical data with rectangular bar. Fig. 8 displays the distribution of quality attribute across each category. Moreover, the bar plot makes it easy to compare the frequency of different categories. For comparing the proportion of different categories, pie charts or tree maps are very useful. In Fig. 9, the pie chart shows the distribution of quality attribute in a circular format. In Fig. 10, the tree map shows part to whole relationships in a hierarchy of categories using rectangle. Both pie chart and tree map are area-based data visualization techniques for showing proportions, comparing categories, and visualizing hierarchical data.

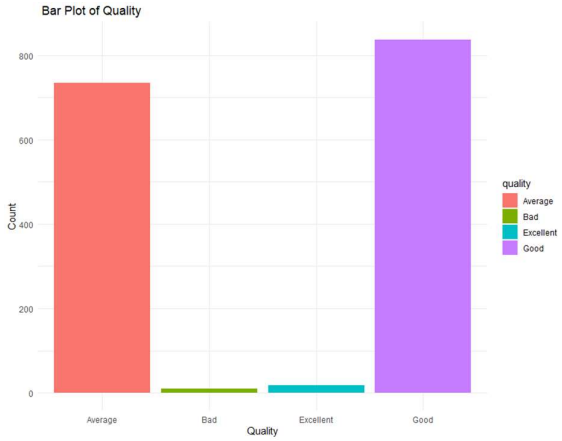

Figure 8: The Bar plot of Quality Attribute.

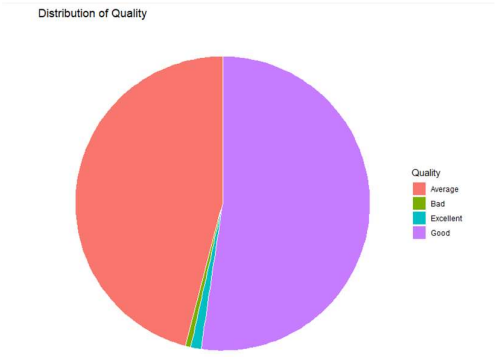

Figure 9: The Pie Chart of Quality Attribute.

Figure 10: The Treemap of Quality Attribute.

B. *Multivariant Data Visualization Technqiues:*

A Scatter plot is one of the most common visualization techniques used to show the relationship between two continuous variables. It represents a datapoint using a dot. Combining the x axis and the y axis, the datapoint is created. It can be used to determine whether two attributes are correlated or not. Sometimes, it can't be determined just by visualizing the scatter plot, regression line is alternative great way to determine if there are any relationship. Regression line is basically the line that represents the correlation between two attributes using a line. Fig. 11 illustrates the positive relation between the Fixed Acidity attribute and the Citric Acid attribute by combining the scatter plot and regression line meaning increasing or decreasing one attribute's value will have the same impact on the other attribute. Fig. 12 shows the negative relation between the Fixed Acid and pH attribute and Fig meaning increasing or decreasing one attribute's value will have the exact opposite impact on the other attribute. 13 displays the zero relationship between the Volatile Acid and Residual Sugar attributes meaning changing one of the attribute's data does not have any effect on the other.

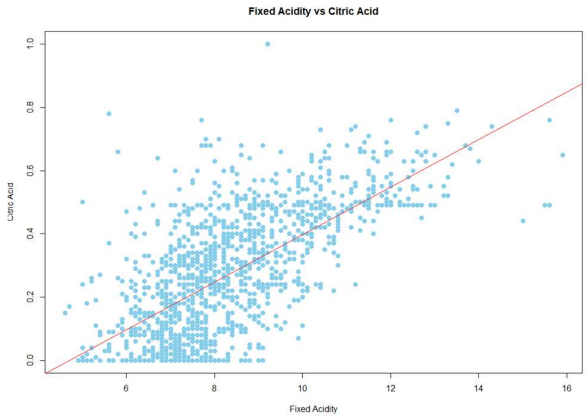

Figure 11: Scatterplot and regression line of Positive Correlation.

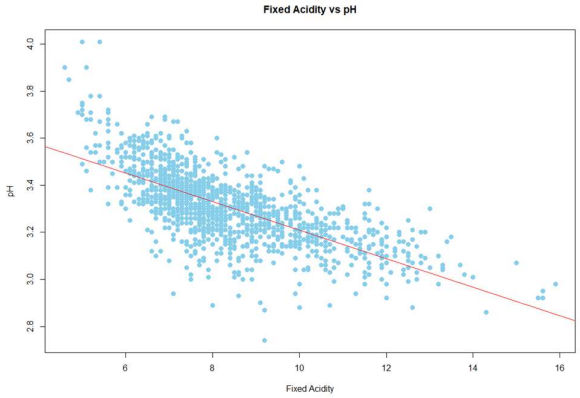

Figure 12: Scatterplot and regression line of Negative Correlation.

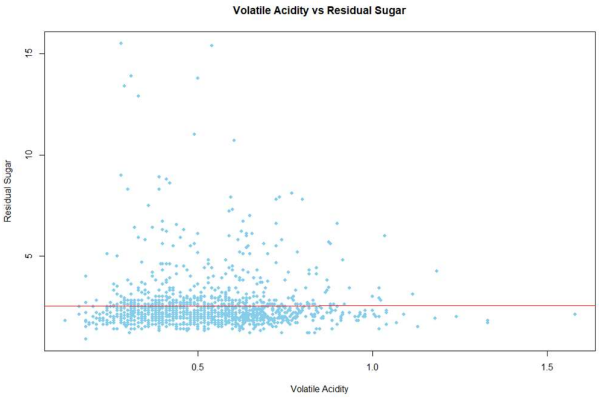

Figure 13: Scatterplot and regression line of zero correlation.

Bubble plots are a type of scatter plot where a third dimension of the data is represented by the size of the markers (bubbles) on the plot. The purpose of a bubble plot is to visualize three-dimensional data in a two-dimensional space. Fig. 14 shows the correlation between Citric Acid, Alcohol and Fixed Acidity attributes. Although it becomes very hard to visualize using bubble plot when the data is very large meaning high number of datapoints.

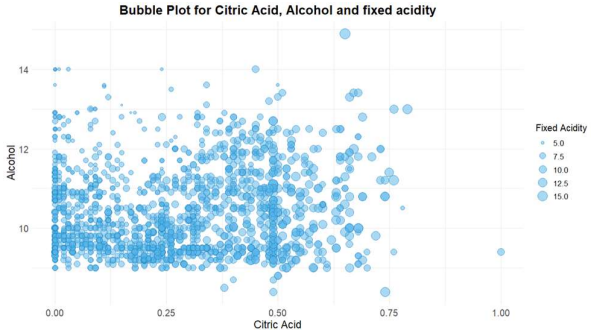

Figure 14: The Bubble Plot of Citric Acid, Alcohol and Fixed Acidity Attributes.

For visualizing all the attributes' correlation with each other at once, heatmap is an efficient technique. A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as individual color colors. It is particularly useful for visualizing large datasets and identifying patterns, trends, and relationships between variables at once. Fig. 15 demonstrates the correlation between all the continuous attributes in heatmap based on colors. Using the heatmap, it can easily determine the correlation between all the attributes at once easily.

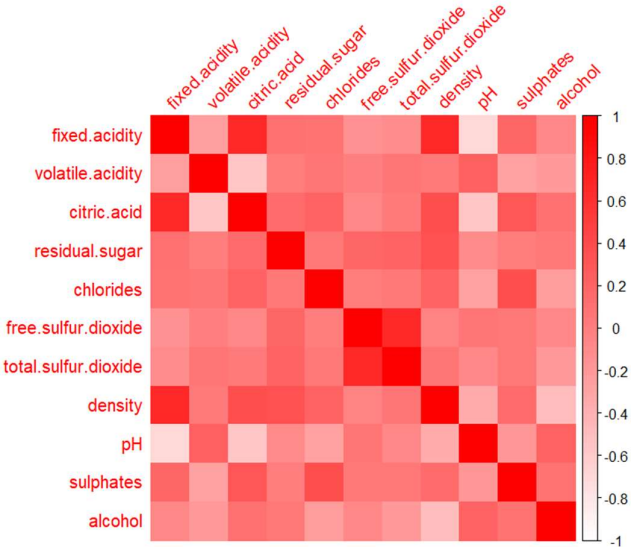

Figure 15: The Heatmap for Correlation between All Attributes.

Besides getting insights about univariant data using violin plot, information about multivariant data can also be easily gained. Fig 16 displays the distribution of pH across the categories of quality. The distribution, median and spread of pH across quality categories is shown. The violin plot shows the median pH is lowest for the quality "Excellent" and highest for the quality "Bad". Also, the density of pH is very high when the quality is "Excellent" at the lower point meaning It can be said the density of pH is the lowest. So, the wine quality is excellent meaning lower the pH level of the wine, the better the quality.

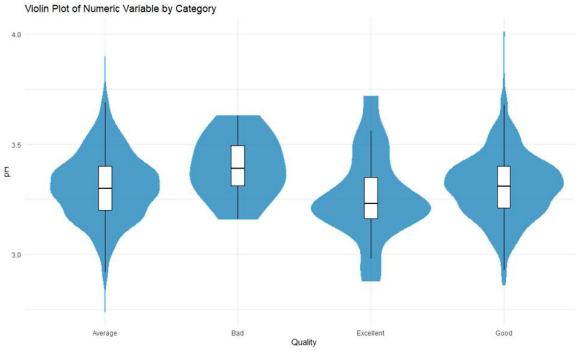

Figure 16: The Violin Plot of Quality and pH attributes.

Data visualization can also be used compare the attributes between two instances. Fig. 17 illustrates the comparison between two random instances of the dataset using bar plot.

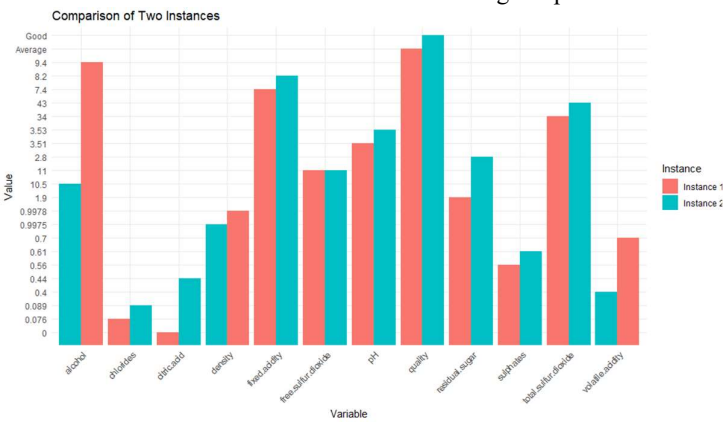

Figure 17: The Bar Plot to Show Comparison of All Attributes between Two Instances.

V. Conclussions And Future Endeavors

This research has evaluated the importance of data visualization techniques for two data types, including univariant and multivariant data. The findings of the research provide insights about the Kaggle Red Wine Quality Dataset. For univariant data, how the data is distributed, the trends and changes within the data, if there are any anomaly within the data were discussed. For multivariant data the relation between multiple attributes, how changes in one attribute may impact another attribute and comparison between two instances were discussed. Nevertheless, these techniques have their own capabilities with strength and weakness to demonstrate data. These details can help the later stages of the data science process. Future work of data visualization could focus on dynamic exploration of dataset with integrating machine learning techniques to automate pattern identification and represent uncertainty visually. Moreover, development of more innovative techniques for representing high dimensional data in a more interpretable manner can help to perform data visualization more efficiently. Research into effective strategies for data visualization can enable rapid insights and decision making.

VI. REFERENCES

[1] E. Saccenti, H. C. J. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. W. B. Hendriks, "Reflections on univariate and multivariate analysis of metabolomics data," pub Dec 2013.

[2] L. J. Gosink, C. Garth, J. C. Anderson, E. W. Bethel, and K. I. Joy, "An Application of Multivariate Statistical Analysis for Query-Driven Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 3, March 2011.

[3] X. Qin, Y. Luo, N. Tang, and G. Li, "Making data visualization more efficient and effective: a survey," VLDB J., vol. 29.

[4] Q. Li and Q. Li, "Overview of data visualization," Embodying Data Chin. Aesthet. Interact. Vis. Gaming Technol., pp. 17–4.

[5] Dirk Stengel, Giorgio Maria Calori, and Peter Glannoudis, “Graphical data presentation”, pub: July 2008.

[6] P. Mishra, S. Kumar, and M. K. Chaube, "Graph Interpretation, Summarization and Visualization Techniques: A Review and Open Research Issues," Multimed. Tools Appl., pp. 1–43, 2022.

[7] A. Unwin, "Why is data visualization important? what is important in data visualization?" Harv. Data Sci. Rev., vol. 2, no. 1, p. 1, 2020.

[8] D. Srivastava, "An Introduction to Data Visualization Tools and Techniques in Various Domains," International Journal of Computer Trends and Technology, vol. 71, no. 4, pp. 125-130, April 2023.

[9] C. Yalim and H. Handly, "The Effectiveness of Visualization Techniques for Supporting Decision-Making," pub: June 2023.

[10] Soyul Han, ll-Youp Kwak, “Mastering data visualization with Python: Practical tips for researchers”, The Journal of Minimally Invasive Surgery, pp. 167-175, Pub: Nov 2023.