JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MULTIVARIATE DATA VISUALIZATION TECHNIQUES USING R

[1]**Tanvi Koyande**

[1]Assistant Professor
[1]Department of Statistics
[1]Ramnarain Ruia Autonomous College, Mumbai, India

Abstract: *In the modern world, the ability to interpret and understand data is becoming ever more essential. Datasets with many variables or features included in each observation or sample point are referred to as multivariate data. Multivariate data visualization's primary goal is to show the relationship between two or more variables. The scatterplot, boxplot violin plot, bubble chart, heat map, and correlation matrix are the most basic methods for multivariate data visualization. Several different visualization methods for multivariate data are covered in this article.*

IndexTerms: Scatterplot, Boxplot, Violin plot, Bubble chart, Heat Map, Correlation matrix

I. INTRODUCTION

Any data analysis process starts with data visualization. A data visualization helps the researcher understand the data better. Also, the data become simple enough for anyone to understand. Since multivariate data contains more than two variables, data visualization is crucial to understanding each variable's characteristics and how it relates to the others.

1) Scatter Plot: Scatter plots are mostly used to observe and demonstrate relationships between two quantitative variables. The variables' values are indicated by dots. This visual aid uses a cartesian coordinate system, in which a dot on a two-dimensional plane represents each data point. One variable, sometimes known as the independent variable, is represented by the horizontal axis (X-axis). The values of the second variable, referred to as the dependent variable, are represented by the vertical axis (Y-axis). It is possible for the relationships to be strong or weak, non-linear or linear, or positive or negative. When examining the data as a whole, the data points, or dots, that show up on a scatter plot allow for the identification of patterns in addition to representing the unique values of each data point.

2) Box Plot: A box and whisker plot, often called a boxplot, is a graph that summarizes a collection of data. It shows a set of data's five-number summary. The smallest observation, first quartile, median, third quartile, and largest observation make up the five-number summary. Box Plots display the data's skewness as well. The boxplot's shape displays both the distribution of the data and any outliers. The points that deviate numerically from the rest of the data are called outliers.

3) Violin Plot: While the box plot is useful for comparing summary statistics, it is unable to display data variances. The distribution of numerical data can be observed using violin plots. Violin plots show the density of each variable as well as summary statistics. In order to visualize the distribution, violin plots combine the advantages of boxplots and kernel density charts. Violin plots offer a more in-depth look at the data distributions and validate the conclusions drawn from the boxplots.

4) Bubble Chart: A bubble chart is a kind of chart where three data dimensions are shown. Every entity is represented as a disk that represents two of the v_i values through the disk's xy location and the third through its size, together with its triplet (v_1, v_2, v_3) of related data. If the data has three data series, each with a set of values, a bubble chart can be used in place of a scatter chart. The values in the third data series dictate the sizes of the bubbles. Financial data is frequently presented using bubble charts.

5) Heat Map: A heat map is a two-dimensional data representation where colors stand in for values. A straightforward heat map offers a visual summary of data instantly. Complex data sets can be understood by the viewer with more intricate heat maps. Heat maps can be shown in a variety of ways, but they all have one thing in common: they use color to convey correlations between data values that would be far more difficult to understand if seen in a spreadsheet as numbers.

6) Correlation Matrix: An effective tool for determining the relationships between several variables is a correlation matrix. In simple terms, a table containing the correlation coefficients for several variables is called a correlation matrix. The matrix displays the relationships between each possible pair of values in a table. The variables are displayed in a table with rows and columns in a correlation matrix. The matrix is a table where each cell has a correlation coefficient, with 1 denoting a strong association, 0 a neutral relationship, and -1 a weak relationship between the variables.

II. DATA

Understanding the data visualization methods stated above is the purpose of this article. Numerous methods and software programs are available to carry out different kinds of analysis and visualization. The R software is the most widely used software for data analysis. The software is freely available and open source. With very little coding, it makes it possible to execute many visualization styles. Numerous diagrammatic and graphical representations are available in R using the ggplot2 package. This article's visual representations are plotted using the ggplot2 tool. This article makes use of the Iris dataset from the R software. A well-known dataset that is frequently used to learn and understand machine learning and data analysis techniques is the Iris dataset. Setosa, Versicolor, and Virginica are the three species of iris flowers that are measured in the Iris dataset. The four characteristics that make up each sample are the length and the width of the sepal and petal.

III. VISUAL REPRESENTATION

To facilitate visual representation, the first 50 data points (Species-setosa) are highlighted in red. Green is used to represent the following 50 data points (Species-versicolor), and blue is used to represent the final 50 data points (Species-virginica).

3.1 Scatter Plot

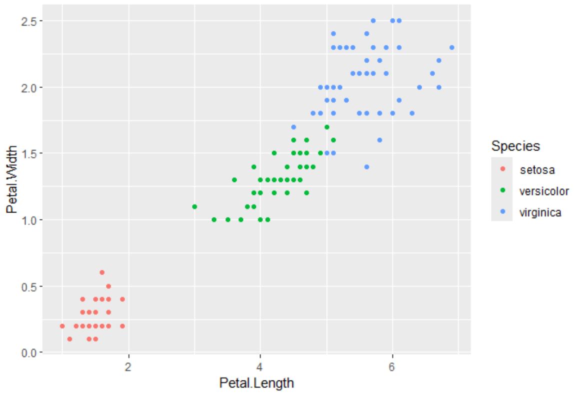


Figure 1: Scatter Plot of Petal Length versus Petal Width

The scatter plot of 150 flowers' petals' length against petals' width is shown in Figure 1. Various species are represented by different colors. Since a scatter plot illustrates the pairwise relationship between two factors, each pair can be studied separately by taking into account all possible pairing combinations. Finding relationships between variables in huge datasets is tough. Scatter plots of every pair of variables can be made simultaneously with R.

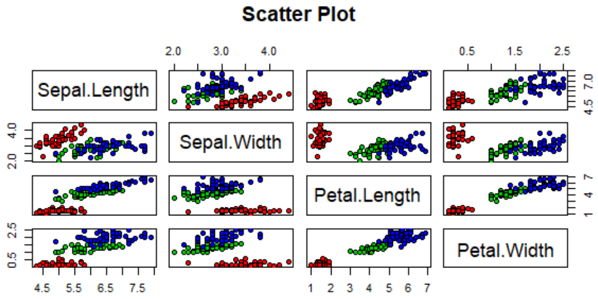


Figure 2: Pair Plot

Pairwise scatterplots of the variables (sepal length, sepal width, petal length, and petal width) against each other are displayed in the Figure 2, with the points colored according to the species. The plot makes it abundantly evident that the species virginica has the largest petal length and width, while the species Setosa has the smallest. It is possible to obtain such information for any other species.

3.2 Box Plot

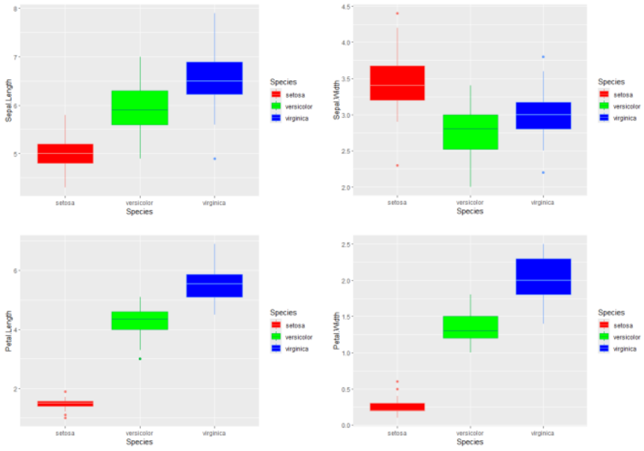


Figure 3: Box Plots

To compare and contrast two or more groups, a boxplot is made. Here, a boxplot is used to compare the three distinct species' various characteristics (petal length, petal width, sepal length, and sepal width). Outliers, median, and quartiles are also displayed in boxplots. The distributions of the three species clearly varied in significant manners depending on the length and width of the petals, as shown by the boxplots in Figure 3.

3.3 Violin Plot

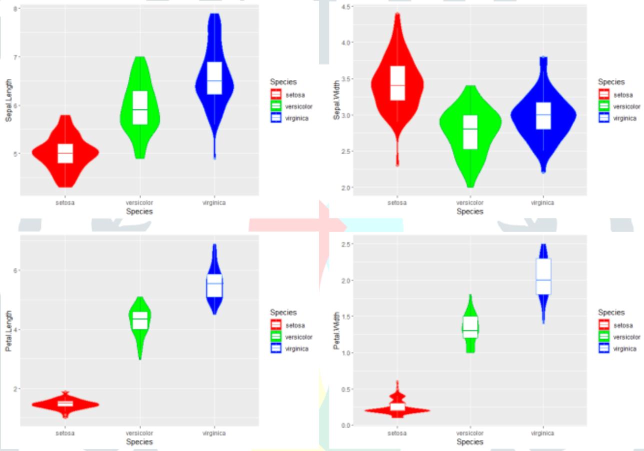
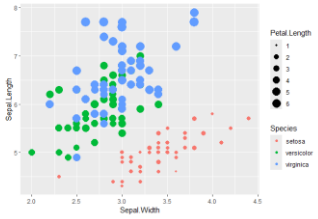


Figure 4: Violin Plots

The relationships between species and petal and sepal lengths and widths are depicted in Figure 4. The box plot parts indicate that Sentosa has shorter median sepal and petal lengths than other species. The distribution's shape—very narrow at each end and wide in the center—indicates that Virginica's Sepal Width is concentrated toward the median.

3.4 Bubble Chart



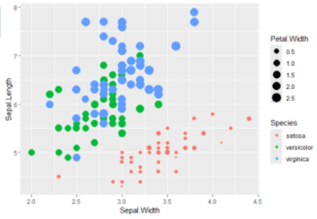


Figure 5(a): Bubble Chart **Figure 5(b):** Bubble Chart

When comparing three or more variables, bubble charts are utilized. The bubble chart displays two variables as the x-y axes, while the third variable is represented by the bubbles' sizes. The features of a single iris flower are represented by each bubble. The vertical location of a bubble indicates the Sepal Length, and the horizontal position indicates the Sepal Width. The size of each bubble represents the length of the petal in Figure 5(a) and the width of the petal in Figure 5(b), with larger bubbles denoting larger petals. Since red-colored bubbles are smaller than blue- and green-colored ones. Comparing Species Sentosa to Versicolor and Virginica, it is evident that the former have smaller petal lengths and widths.

3.5 Heat Map

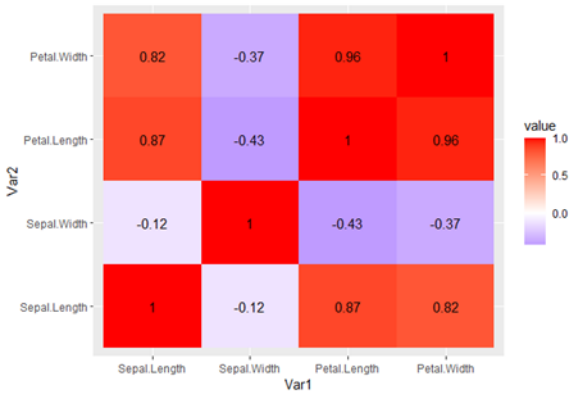


Figure 6: Heat Map

The correlations between the variables are shown visually by the heatmap in Figure 6. Lighter hues reflect weaker or no relationships, while darker hues indicate stronger correlations. Petal Length and Width have a positive association, while Petal Length and Sepal Width have a negative correlation, according to the heatmap analysis.

3.6 Correlation Matrix

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
```

The correlation matrix is the key component of the correlation heatmap. The correlation coefficients between each pair of variables in the dataset are provided by this matrix. Since the correlation between a variable and itself is always perfect all of its diagonal elements must be 1 and the correlation matrix is symmetric. The ggplot2 package in R can visually represent the correlation matrix in a variety of ways, as illustrated in Figure 7, making it easier to interpret.

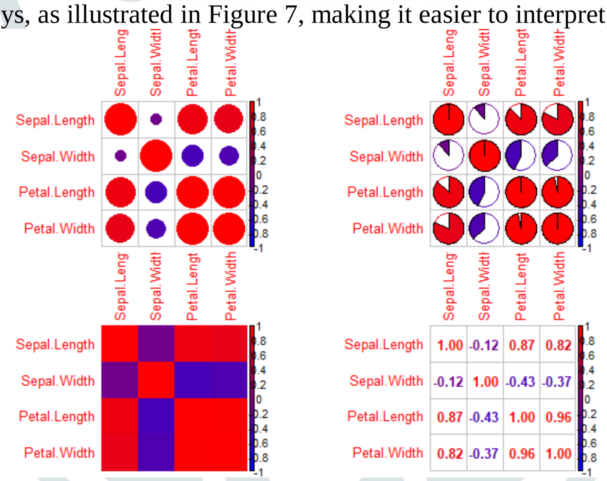


Figure 7: Correlation Matrix

IV. Conclusion

Complex information can be accessed and understood more easily when it is presented visually. The most effective method for comprehending the type and extent of data is data visualization. Consequently, the most important stage in any statistical work is to choose the right technique before beginning any study. It is critical to utilize the best technique to accurately display the data and make it easy to grasp. Visualization can make insights visible that raw data alone might not be able to convey. Analysts can make better decisions and solve problems by more effectively identifying outliers, correlations, and other noteworthy trends when data is presented visually. The simultaneous discovery of patterns and interactions between several variables is made possible by multivariate data visualization.

References

[1] [Deepmala Srivastava. An Introduction to Data Visualization Tools and Techniques in Various Domains, International Journal of Computer Trends and Technology, Volume 71 Issue 4, 125-130, April 2023

[2] Matthew N O Sadiku, Adebowale E. Shadare, Sarhan M. Musa, Cajetan Akujuobi. DATA VISUALIZATION, International Journal of Engineering Research And Advanced Technology(IJERAT), Volume. 02 Issue.12, December– 2016

[3] Fabian Beck and Shahid Latif. Introduction to Multivariate Data Visualization, Technical Report · May 2022

[4] https://www.geeksforgeeks.org/iris-dataset-in-r/

[5] https://www.geeksforgeeks.org/multivariate-data-visualization-with-r/