Vidyavardhini's College of Engineering and Technology

Department of Computer Engineering

Academic Year : 2023-24 (Odd Sem)

Experiment No.4
Perform data visualization
Date of Performance:
Date of Submission:

Vidyavardhini's College of Engineering and Technology

Department of Computer Engineering

Academic Year : 2023-24 (Odd Sem)

Aim: To perform data visualization on the given data set

Objective:- To implement data visualization on the given data set

Theory:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. It aims to communicate data clearly and effectively through graphical representation.

Scatter Plot:

A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane. The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships. Two attributes, X, and Y, are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated).

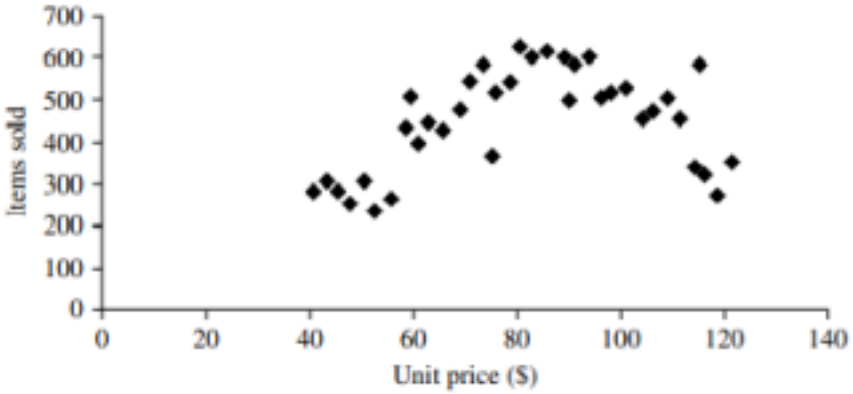


Figure 1: Scatter Plot

Histogram:

Histograms is a graphical method for summarizing the distribution of a given attribute, X . “Histos” means pole or mast, and “gram” means chart, so a

histogram is a chart of poles.If X is nominal, such as automobile model or item type, then a pole or vertical bar is drawn for each known value of X. The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a bar chart.

If X is numeric, the term histogram is preferred. The range of values for X is partitioned into disjoint consecutive subranges. The subranges, referred to as buckets or bins, are disjoint subsets of the data distribution for X. The range of a bucket is known as the width. Typically, the buckets are of equal width. For example, a price attribute with a value range of $1 to $200 (rounded up to the nearest dollar) can be partitioned into subranges 1 to 20, 21 to 40, 41 to 60, and so on. For each subrange, a bar is drawn with a height that represents the total count of items observed within the subrange.

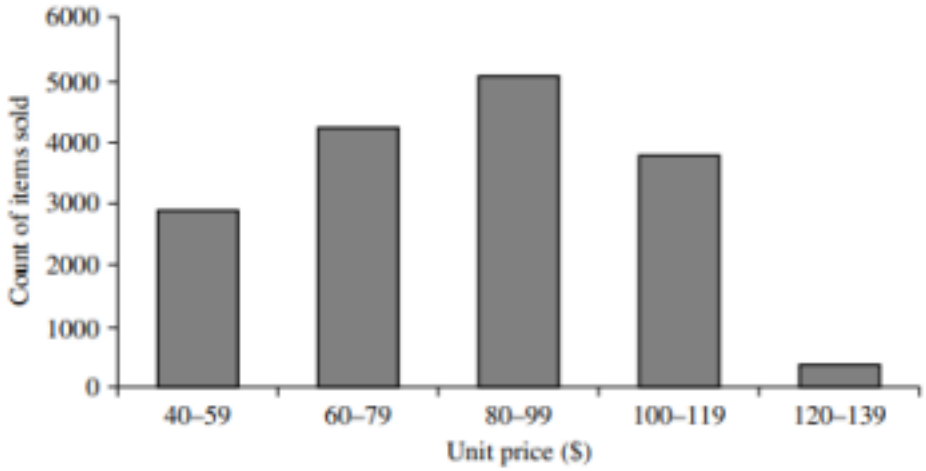


Figure 2: Histogram

Box plot:

Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows: Typically, the ends of the box are at the quartiles so that the box length is the interquartile range. The median is marked by a line within the box. Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

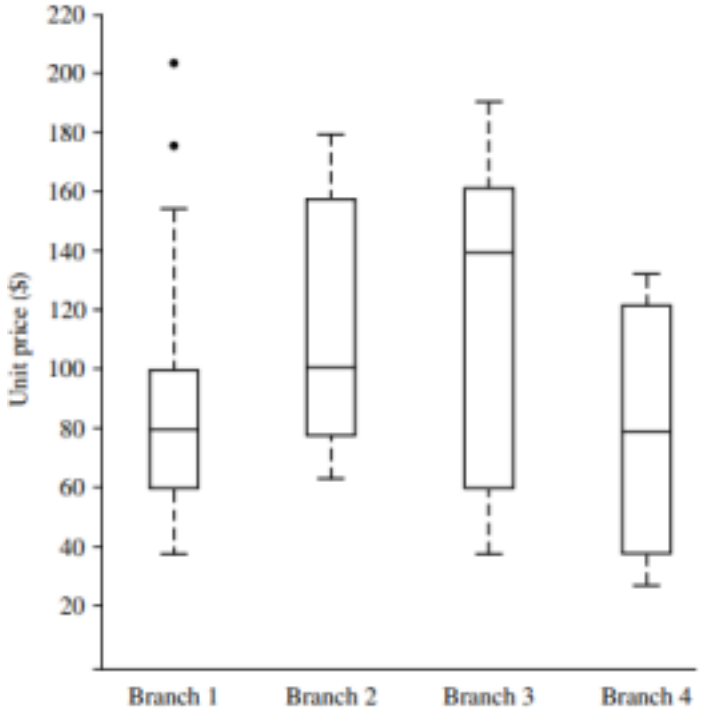
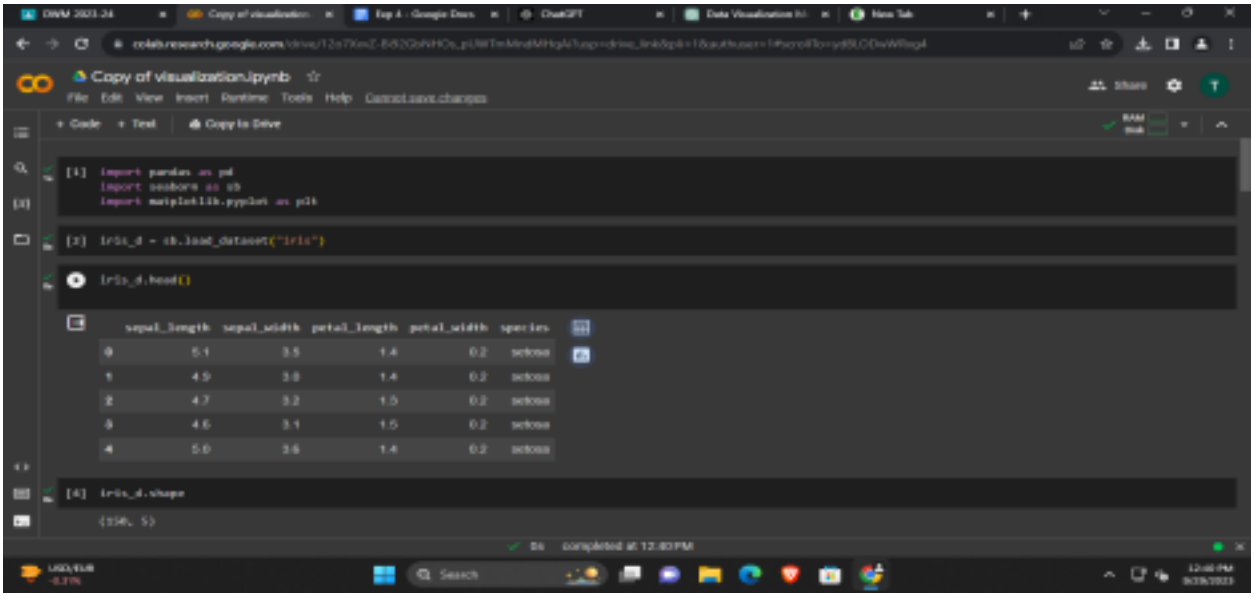
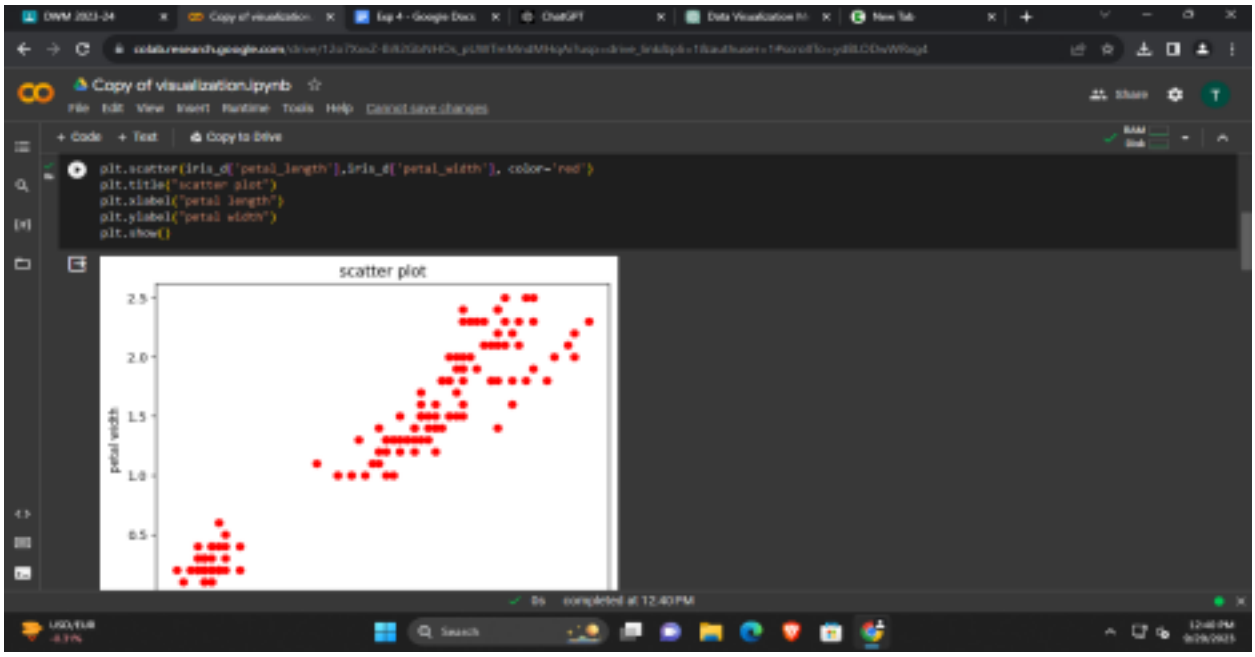
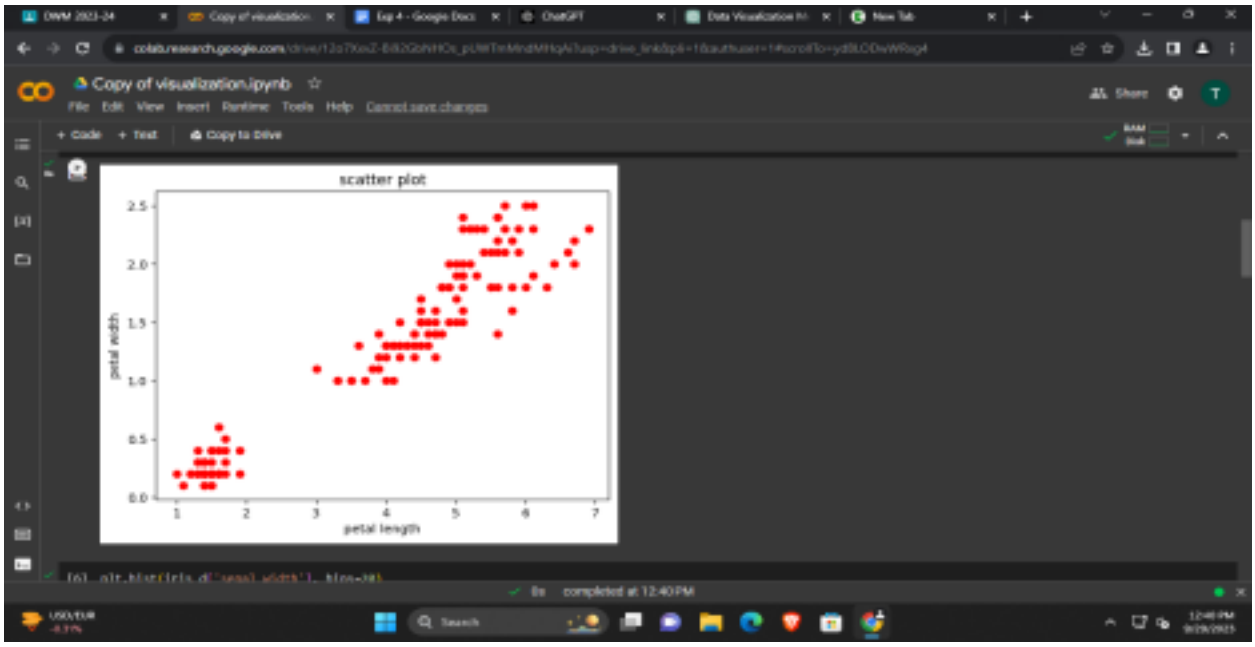


Figure 3: Box Plot

Code and output:



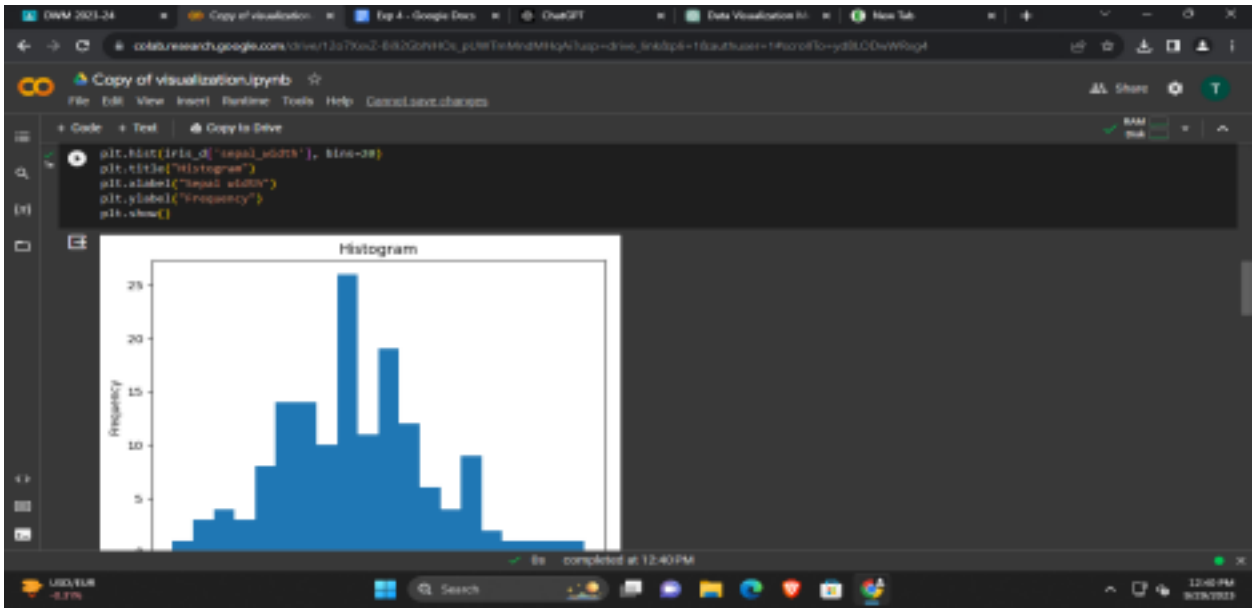


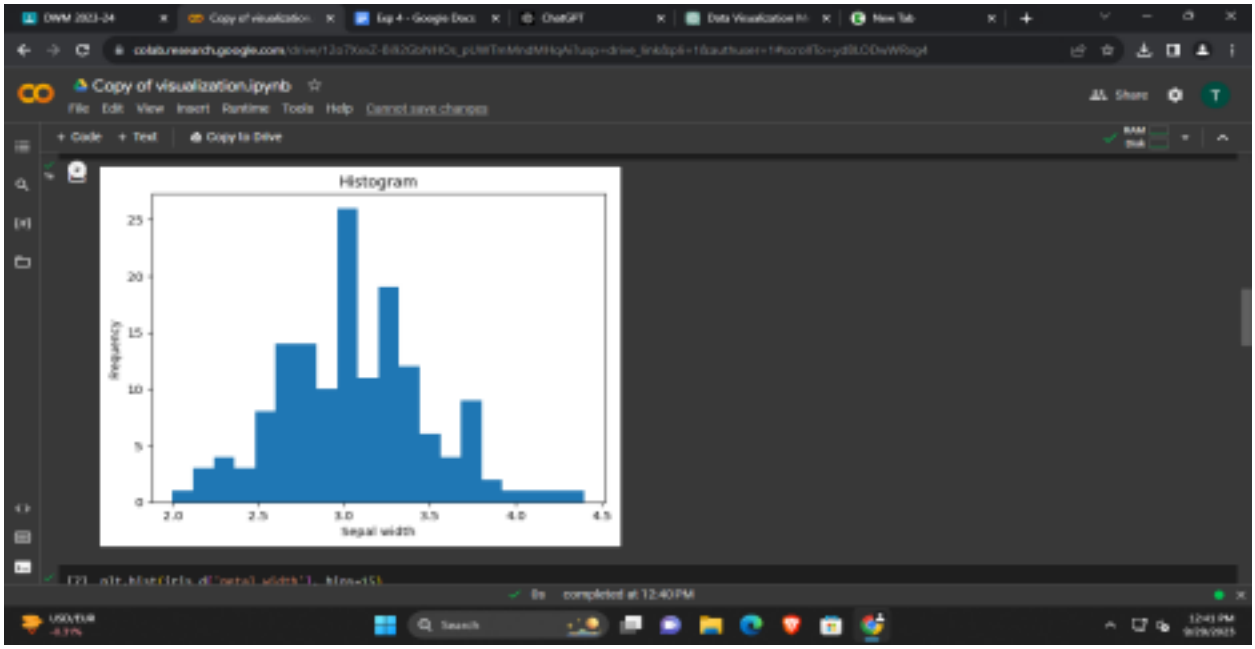


Vidyavardhini's College of Engineering and Technology

Department of Computer Engineering

Academic Year : 2023-24 (Odd Sem)

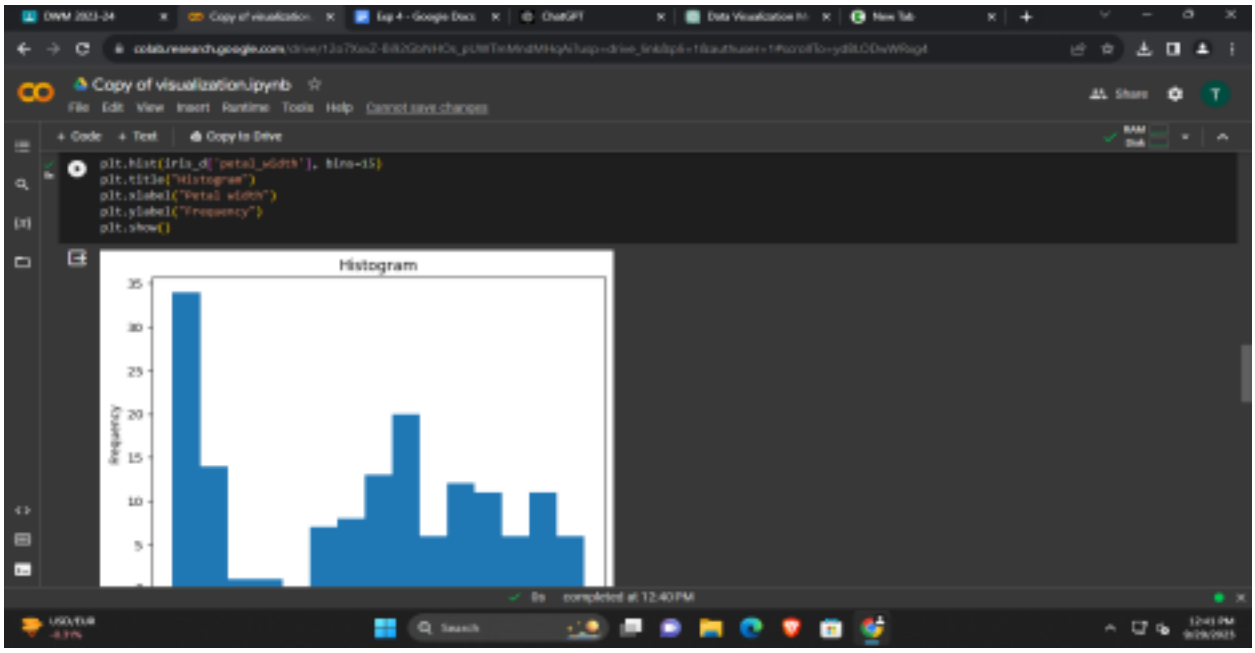




Vidyavardhini's College of Engineering and Technology

Department of Computer Engineering

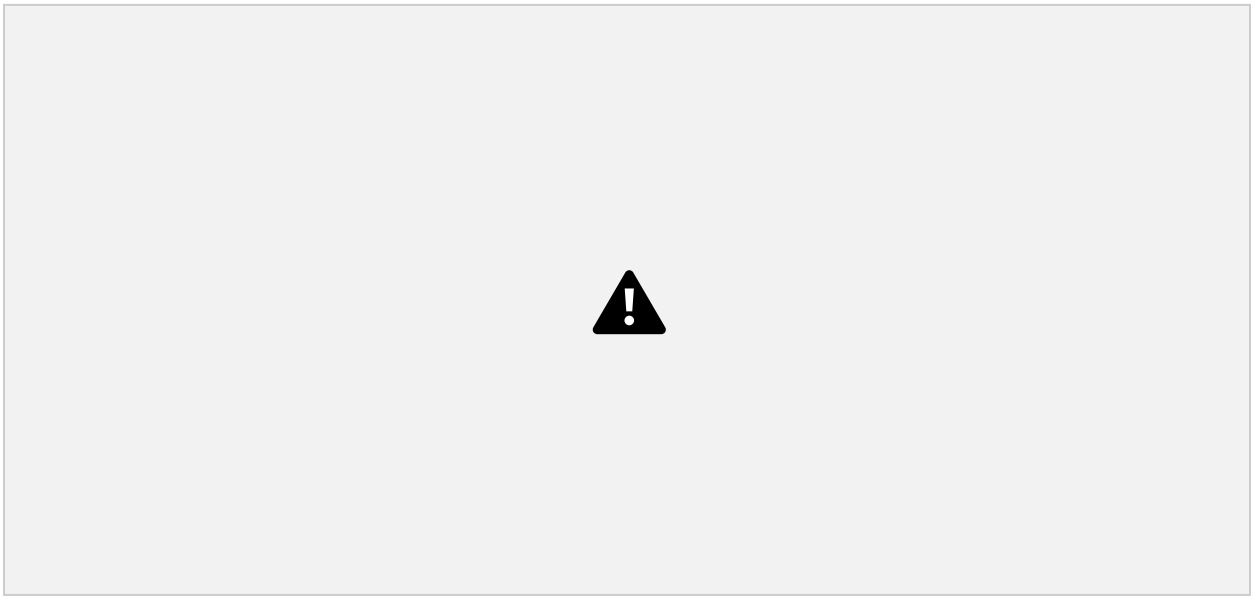
Academic Year : 2023-24 (Odd Sem)

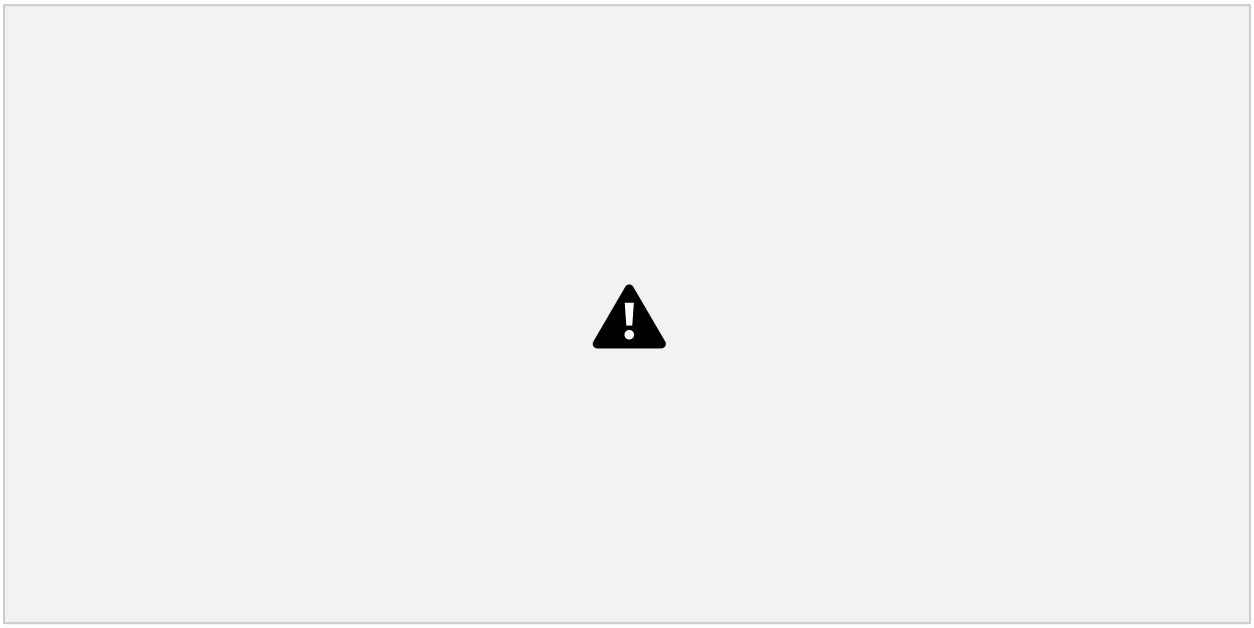


Vidyavardhini's College of Engineering and Technology

Department of Computer Engineering

Academic Year : 2023-24 (Odd Sem)





Conclusion:Scatter plots, histograms, and box plots are valuable tools for visualizing and drawing conclusions from data, including the Iris dataset, which is a well-known dataset in the field of machine learning and data analysis. The Iris dataset contains measurements of four features (sepal length, sepal width, petal

length, and petal width) for three species of iris flowers (setosa, versicolor, and virginica).scatter plots help visualize relationships and species separation, histograms reveal the data distribution and species-specific characteristics, and box plots provide a summary of central tendency, variability, and the presence of outliers. When applied to the Iris dataset, these visualization methods can aid in understanding the characteristics of the iris species and help in feature selection for classification or other analyses.