Unit 5

Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques, Visualizing Complex Data and Relations.

Data visualization is defined as a graphical representation that contains the information and the data.

- By using visual elements like **charts, graphs, and maps,** data visualization techniques provide an accessible way to see and understand trends, outliers, and patterns in data.
- In modern days we have a lot of data in our hands i.e, in the world of **Big Data**, **data visualization tools, and technologies are crucial to analyze massive amounts of information and make data-driven decisions.**

The basic uses of the Data Visualization technique are as follows:

Identifies Patterns and Trends: Visualization helps in making complex datasets more understandable. It translates raw data into visual representations, making it easier for individuals to grasp patterns/trends/insights.

Supports Decision-Making: Data visualization helps decision-makers quickly understand information by presenting it visually. This makes it easier for them to make smart and timely decisions, whether in business or healthcare.

Detects Anomalies and Outliers: Visualizations make it easier to identify anomalies, outliers, or unexpected patterns in the data.

Supports Exploratory Data Analysis: Data visualization is a powerful tool for exploratory data analysis, enabling analysts to explore datasets, generate hypotheses, and gain insights into the underlying structure of the data.

Facilitates Comparative Analysis: Visualizations make it straightforward to compare different datasets, variables, or scenarios.

Types of Analysis for Data Visualization:

- Mainly, there are three different types of analysis for Data Visualization:
  - ✓ **Univariate Analysis:** In the univariate analysis, we will be using a single feature to analyze almost all of its properties.
  - ✓ **Bivariate Analysis:** When we compare the data between exactly two features then it is known as bivariate analysis.

- ✓ **Multivariate Analysis:** In the multivariate analysis, we will be comparing more than two variables.

Data Visualization Tools:

The best data visualization tools include:

- ✓ Data Visualization Tools
- ✓ Tableau
- ✓ R Programming
- ✓ Python Programming
- ✓ PlotlyIBM Watson Analytics
- ✓ Google Charts
- ✓ FusionCharts
- ✓ Datawrapper
- ✓ Infogram, ChartBlocks, and D3.js.

The best tools offer a variety of visualization styles, are easy to use, and can handle large data sets.

- ➢ Data visualization comes in various types, each serving specific purposes. Here are some common types of data visualization:
  - ✓ Bar Chart
  - ✓ Pie Chart
  - ✓ Scatter Plot
  - ✓ Histogram
  - ✓ Boxplot
  - ✓ HeatMap

Difference between Data Visualization and Data Analytics

Data Visualization is the process of representing data visually, to communicate insights and patterns in the data. The goal of data visualization is to make complex data more accessible and easier to understand. This can involve creating charts, graphs, maps, and other visual representations of data. ***Data Visualization can be used in a wide range of contexts, from business to science to journalism.*** For example, a business might use data visualization to understand trends in sales over time, while a scientist might use it to visualize patterns in climate data. Journalists might use data visualization to create interactive graphics that help readers understand complex stories.

Data Analytics, on the other hand, is the process of examining data to conclude and insights from it. This involves using statistical and computational techniques to identify patterns and relationships in the data. Data Analytics can be used in a variety of fields, including business, science, and healthcare. For example, a business might use data analytics to identify which products are selling the most and why. A scientist might use it to analyze genomic data to better understand the causes of disease.

The main difference between **Data Visualization and Data Analytics**, Data Visualization is focused on communicating insights and patterns in the data, while Data Analytics is focused on concluding insights from the data. Data Visualization is often used as a tool for Data Analytics. By creating visual representations of data, it can help analysts better understand the patterns and relationships in the data. For example, a chart or graph might make it easier to see that sales of a particular product are increasing over time

Another key difference between the two is the skill set required. Data Visualization typically requires skills in graphic design and visual communication. Data Analytics, on the other hand, requires skills in statistics, programming, and data analysis.

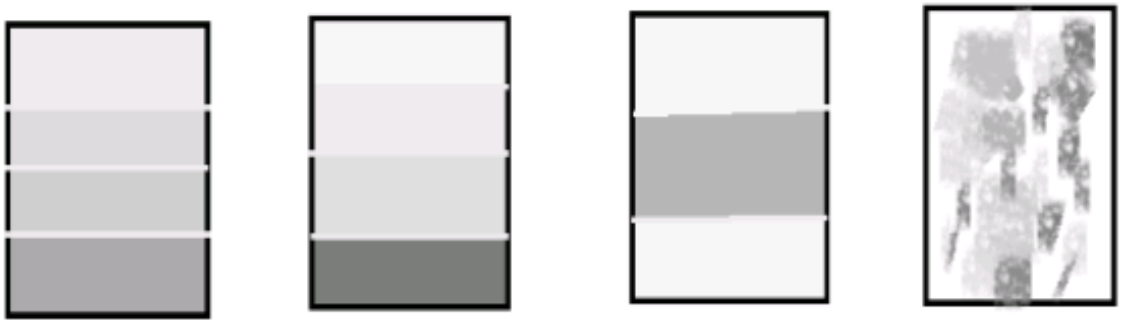
Data Visualization Techniques:

- Pixel Oriented Visualization Techniques
- Geometric Projection Visualization Techniques
- Icon-Based Visualization Techniques
- Hierarchical Visualization Techniques

Pixel Oriented Visualization Techniques:

- Pixel-oriented visualization techniques are a powerful way to represent large, multidimensional datasets by mapping each data value to a single pixel on the screen. This approach allows you to visualize a massive amount of information simultaneously, making it ideal for exploring trends, patterns, and relationships within complex datasets.
- Divide the screen: The screen is divided into multiple sub-windows, one for each dimension of your data. For example, if you have a dataset with three dimensions (e.g., temperature, pressure, and humidity), you would create three sub-windows.

- **Map data to pixels:** Each data value is then mapped to a corresponding pixel within its respective sub-window.
- **Arrange pixels:** The pixels within each sub-window can be arranged in different ways, such as by sorting them based on one of the dimensions or by clustering them based on similarities.
- Few types of pixel-oriented visualization techniques includes: Scatterplot matrices, Parallel coordinates, Image-based techniques.
- All Electronics maintains a customer information table, which consists of 4 dimensions:
    - income
    - credit_limit
    - transaction_volume and age.
  - We analyze the correlation between income and other attributes by visualization.



Geometric Projection Visualization Techniques

- A drawback of pixel-oriented visualization techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space.
- Geometric projection techniques help users find interesting projections of multidimensional data sets.
- Geometric transformations modify the arrangement or characteristics of data points in a plot. For instance, scaling height and weight data to have a mean of zero and a standard deviation of one normalizes the data, aiding in comparing feature importance.
- Projections simplify complex datasets by mapping them onto lower-dimensional spaces while maintaining key relationships between data

points. For instance, projecting three-dimensional data onto a two-dimensional plane enables easier visualization and analysis, preserving essential structural information, which can be done by using principal component analysis.

- A scatter plot displays 2-D data point using Cartesian co-ordinates. A third dimension can be added using different colors of shapes to represent different data points.

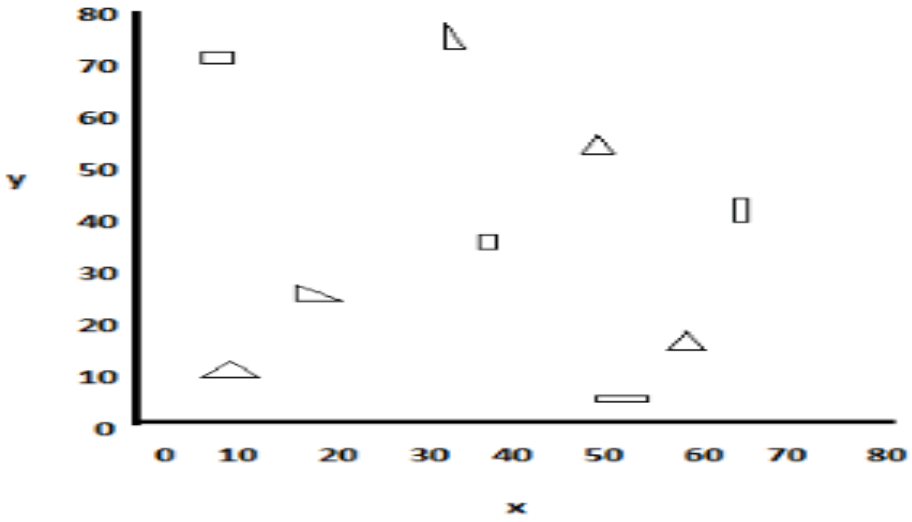
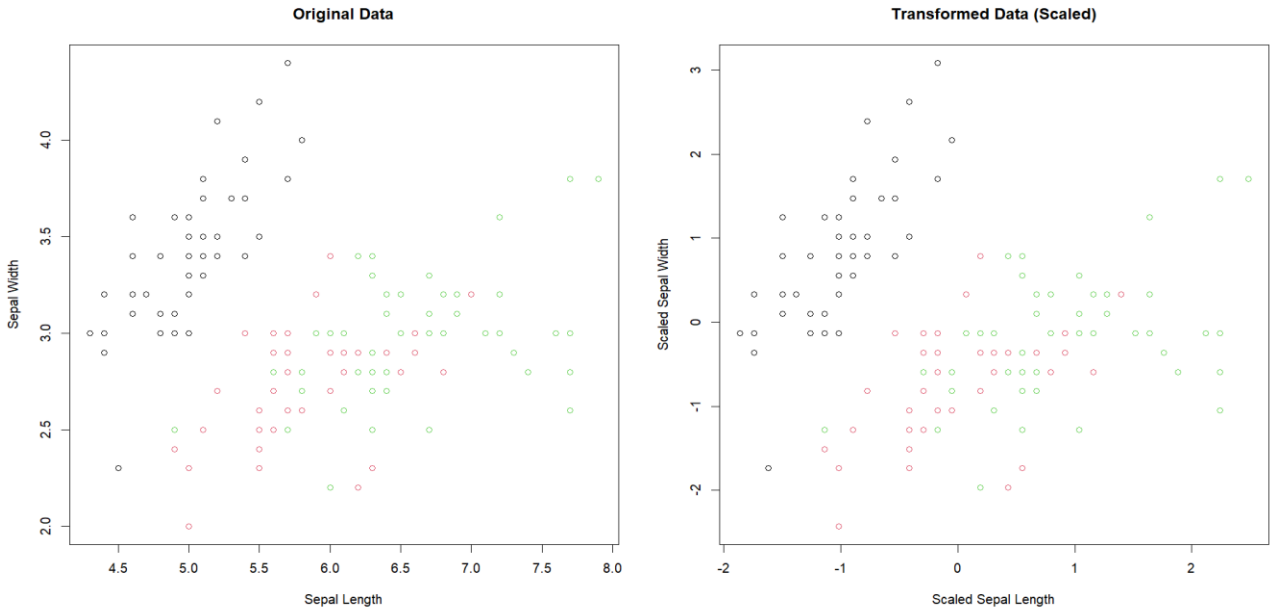


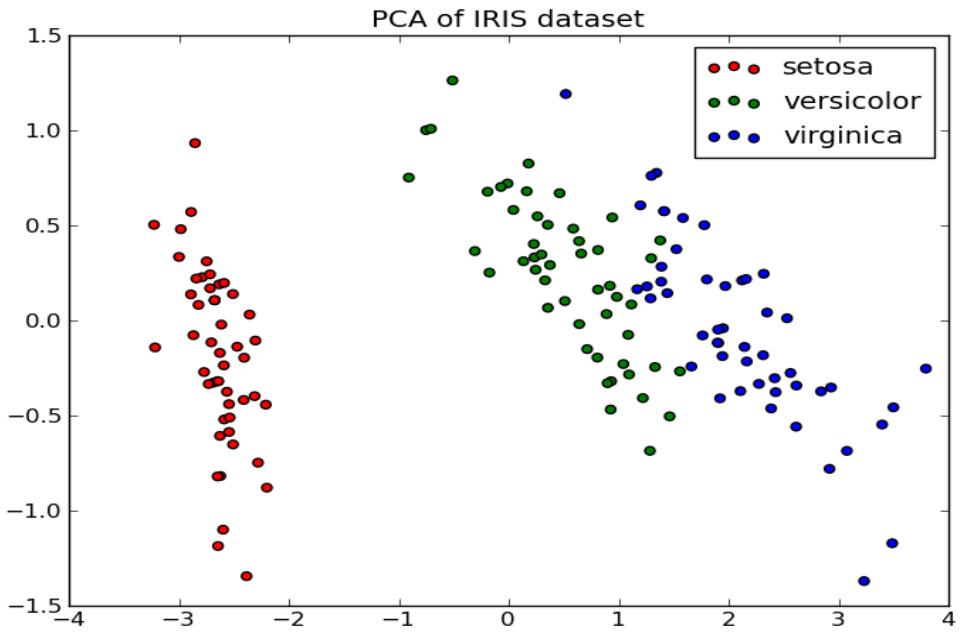
Figure: Geometric Projection

The following example shows the graph between two columns sepal length and sepal width of iris dataset original values and transformed values.



From above graph it is observed that same pattern was observed in both graphs.

Complete iris dataset(four dimensions) is represented two dimensions using principal component analysis.



- Ever growing volume of data and its importance for business make data visualization an essential part of many companies business strategies.
  - Data Visualization
    - ✓ Charts(Line, Bar, Pie)
    - ✓ Plots(Scatter)
    - ✓ Maps(Geographical Maps)

Icon-Based Visualization Techniques:

- It uses small icons to represent multidimensional data values
  - ✓ Two popular icon based techniques:-

Chern off faces: - They display multidimensional data of up to 18 variables as a cartoon human face.

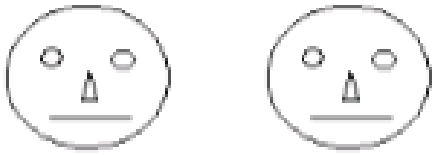
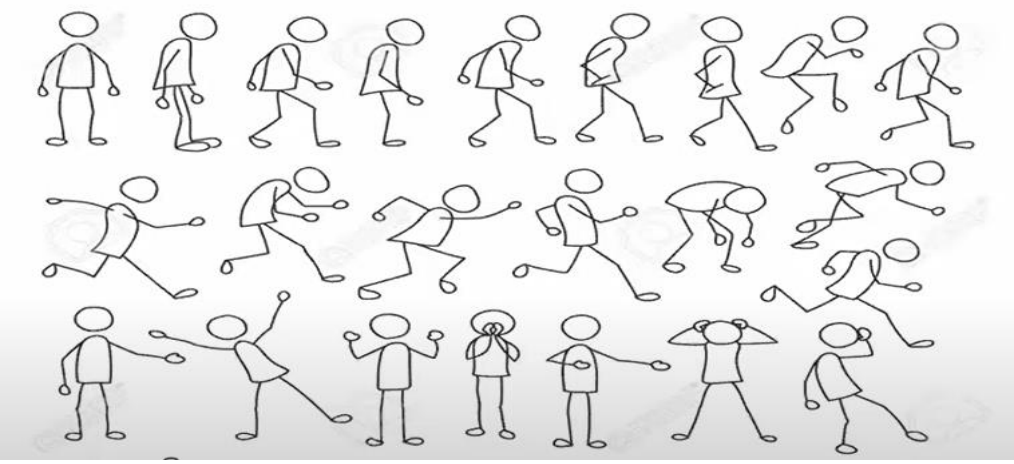


Fig: chern off faces each face represents an 'n' dimensional data points (n<18)

Stick figures: It maps multidimensional data to five –piece stick figure, where each figure has 4 limbs and a body.

- ✓ Two dimensions are mapped to the display axes and the remaining dimensions are mapped to the angle and/ or length of the limbs.



Hierarchical Data visualization techniques:-

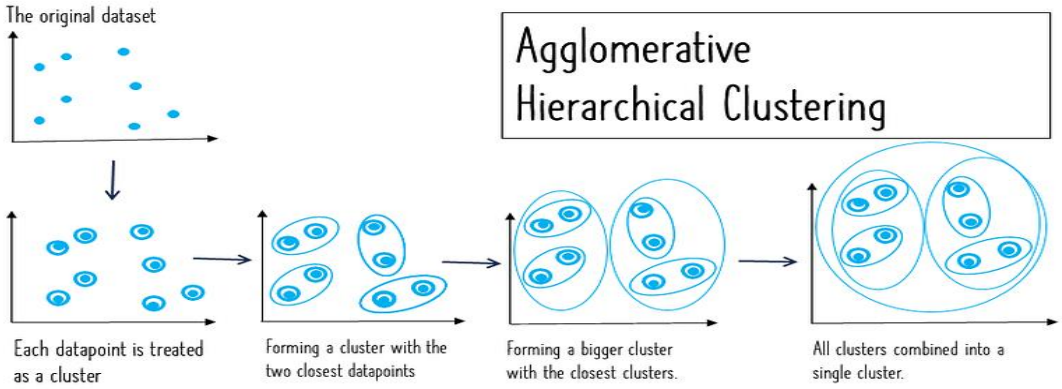
Hierarchical data visualization techniques are used to represent data that has a hierarchical structure, such as organizational structures and clustering results. Here are some common techniques:

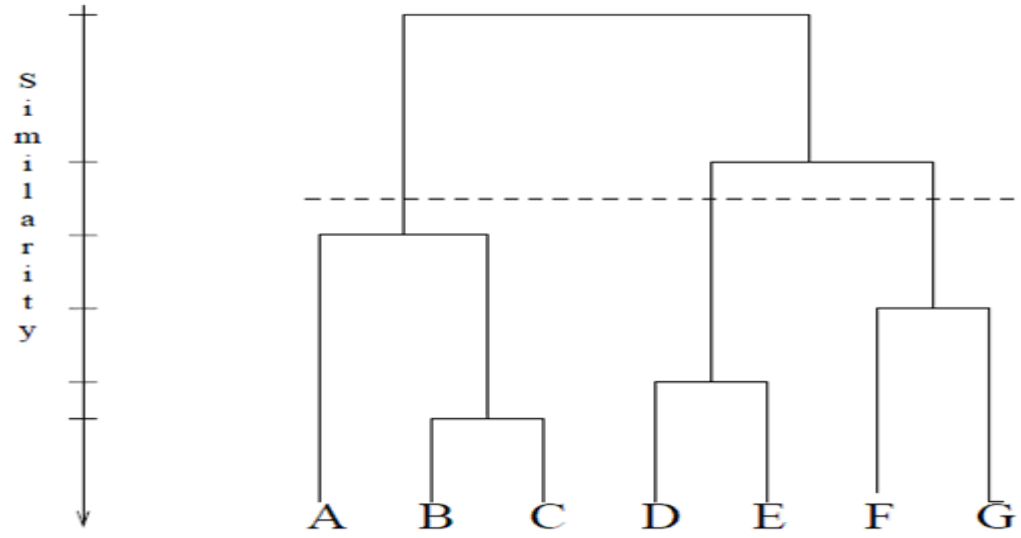
Tree Maps: Tree maps represent hierarchical data as nested rectangles, where each rectangle represents a node in the hierarchy, and its size or color encodes a quantitative or qualitative attribute of the node.

Dendrogram: A dendrogram is a tree-like diagram that shows the arrangement of the clusters produced by hierarchical clustering algorithms**.**

Hierarchical clustering clusters data points into a hierarchy of clusters. There are two techniques of performing this clustering. These are the divisive and agglomerative hierarchical clustering techniques.

Agglomerative hierarchical clustering involves starting with each data point as a separate cluster and then gradually merging the clusters according to similarity to form a hierarchy of clusters.





The figure represents dendrograms which is clustered into single cluster using hierarchical cluster.

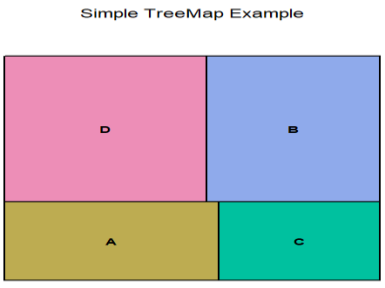
Visualizing complex data and relationships:

Visualizing complex data and relationships involves using graphical representations to make patterns, trends, and connections within the data more understandable.

Network Graphs: Network graphs are used to depict relationships between entities. Nodes represent entities, and edges represent connections between them. For instance, visualizing social networks where nodes are individuals and edges represent friendships.

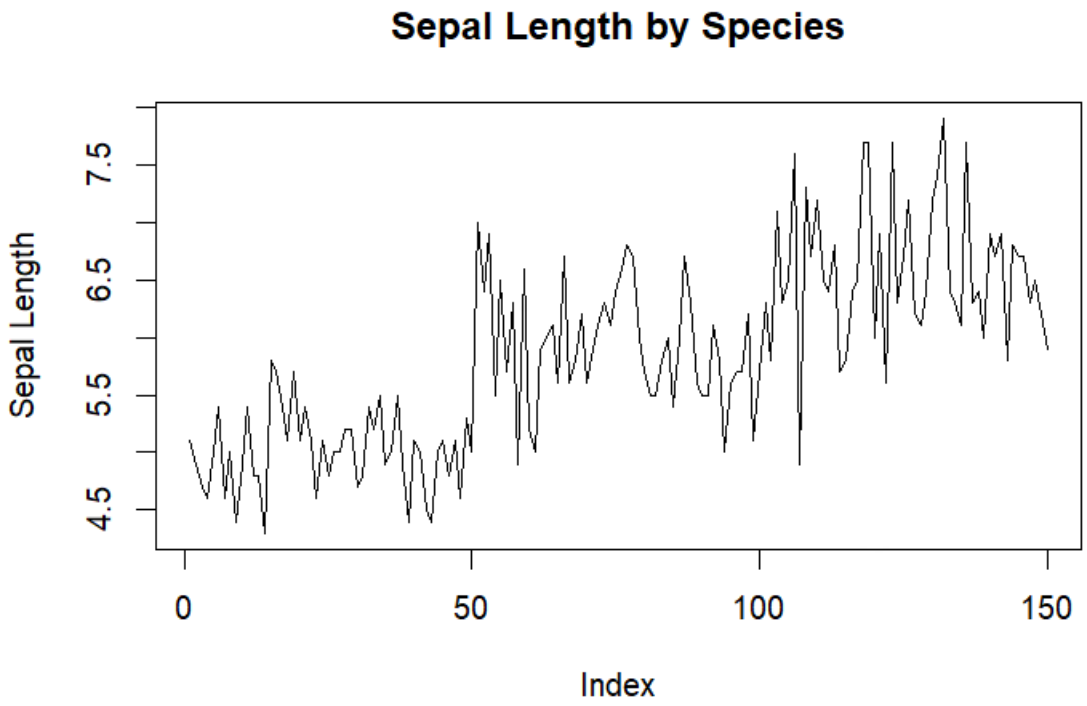
Heatmaps: Heatmaps display data in a matrix where colors represent values. They're handy for identifying patterns in large datasets. An example could be visualizing temperature variations across different regions on a map.

Tree Map: A treemap is a visual method for displaying hierarchical data that uses nested rectangles to represent the branches of a tree diagram. Each rectangle has an area proportional to the amount of data it represents.



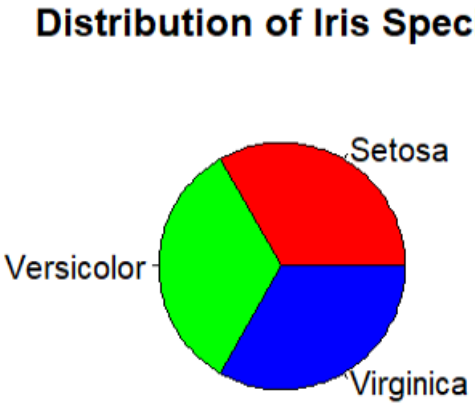
Line Plot: A line plot, also known as a line chart, is a type of graph that displays data as points connected by straight lines. It's commonly used to visualize trends or patterns over time or across different categories.

```
data("iris") # Load the Iris dataset (it's built into R)
# Plotting sepal length by species
plot(iris$Sepal.Length, type = "l", col = iris$Species, main = "Sepal Length by Species", xlab = "Index", ylab = "Sepal Length")
```



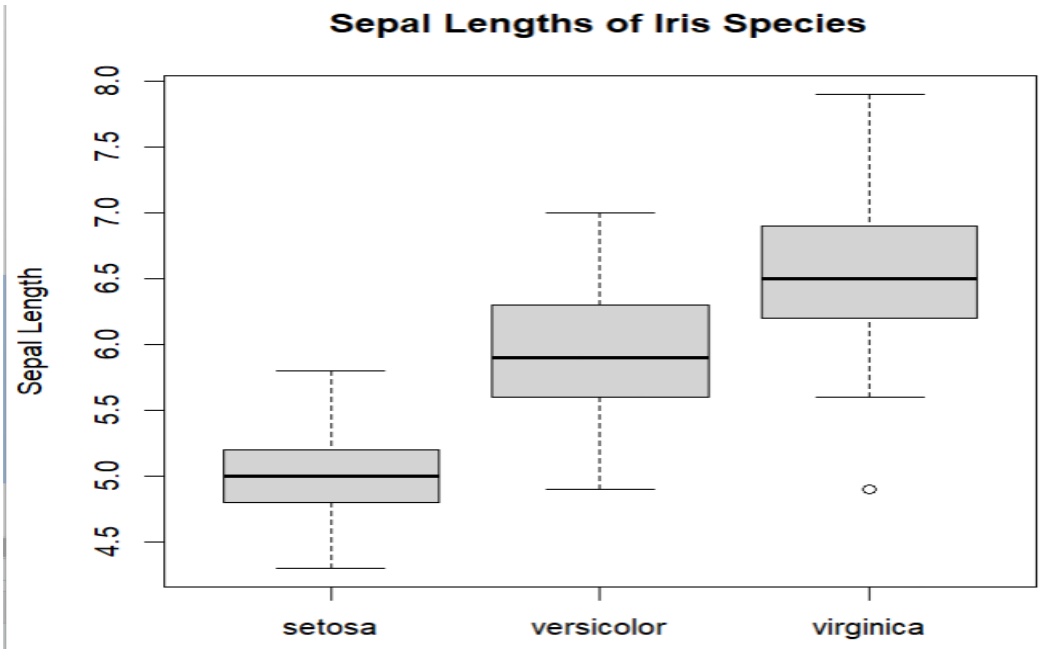
Pie chart: Pie charts are commonly used to represent categorical data, where each category is represented by a slice of the pie. They are particularly useful for showing the proportion of each category relative to the whole.

```
data("iris")
# Calculate the number of each species
species_counts <- table(iris$Species)
# Create a pie chart
pie(species_counts,
    main = "Distribution of Iris Species",
    col = rainbow(length(species_counts)),
    labels = c("Setosa", "Versicolor", "Virginica"))
```



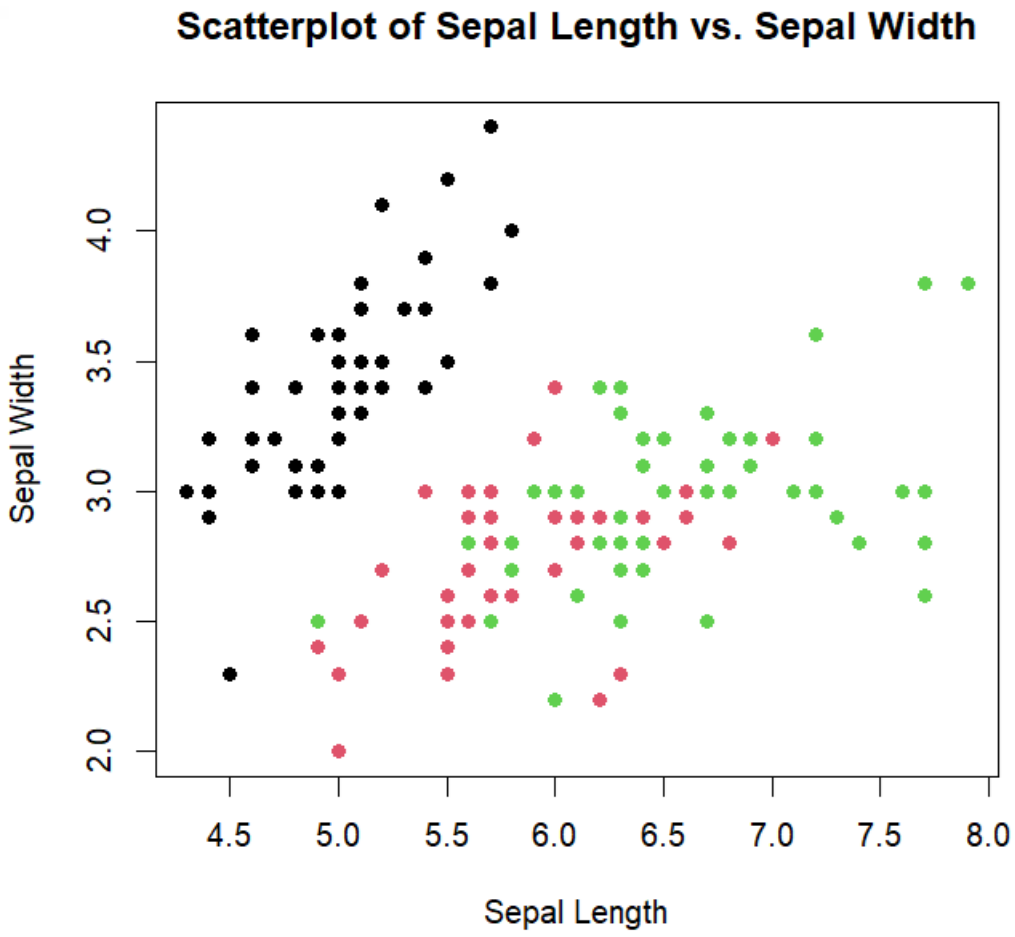
Boxplot: A boxplot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset based on five summary statistics: minimum, first quartile (Q1), median (second quartile, Q2), third quartile (Q3), and maximum. It is particularly useful for visualizing the spread and variability of the data, as well as identifying potential outliers.

```
boxplot(iris$Sepal.Length ~ iris$Species,
        main = "Sepal Lengths of Iris Species",
        xlab = "Species",
        ylab = "Sepal Length")
```

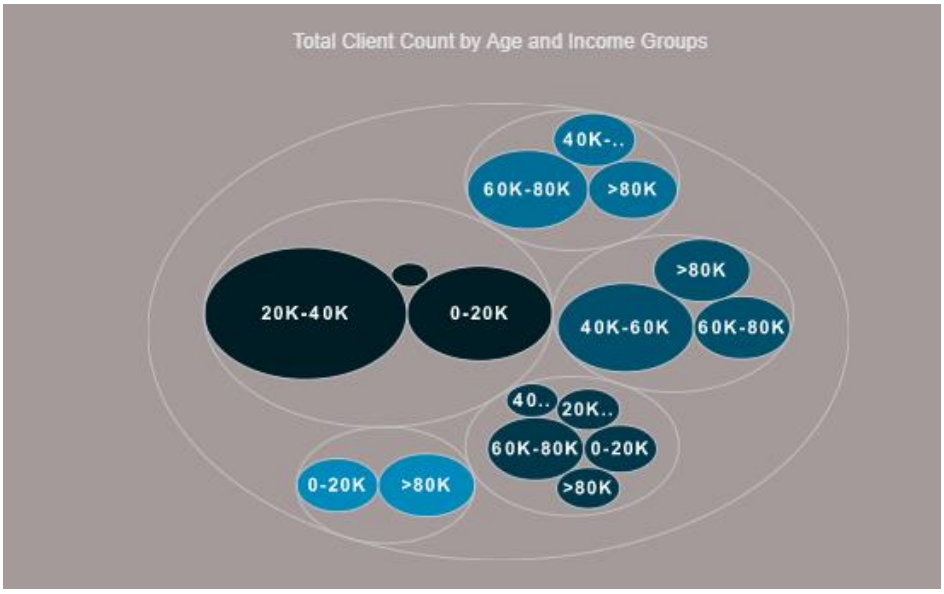


Scatterplot: A scatterplot is a type of data visualization that displays the relationship between two variables by plotting individual data points on a two-dimensional graph. Each point on the plot represents a single observation, with its position determined by the values of the two variables being compared. Scatterplots are useful for identifying patterns, trends, and correlations between variables.

```
plot(iris$Sepal.Length, iris$Sepal.Width,
    main = "Scatterplot of Sepal Length vs. Sepal Width",
    xlab = "Sepal Length", ylab = "Sepal Width",
    pch = 16, col = iris$Species)
```

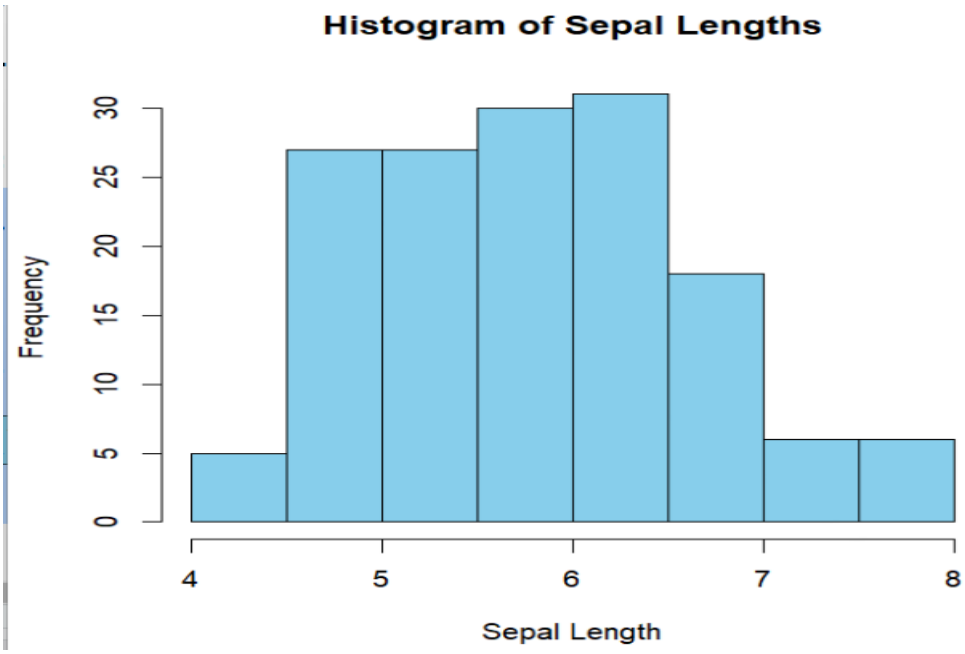


Circle parking: Circular packing or circular treemap allows to visualize a hierarchic organization. It is an equivalent of a treemap or a dendrogram, where each node of the tree is represented as a circle and its sub-nodes are represented as circles inside of it.



Histogram: A histogram is a type of graphical representation that displays the distribution of continuous data by dividing it into intervals called bins and plotting the frequency or count of data points falling into each bin. It is commonly used to visualize the frequency distribution of numerical data.

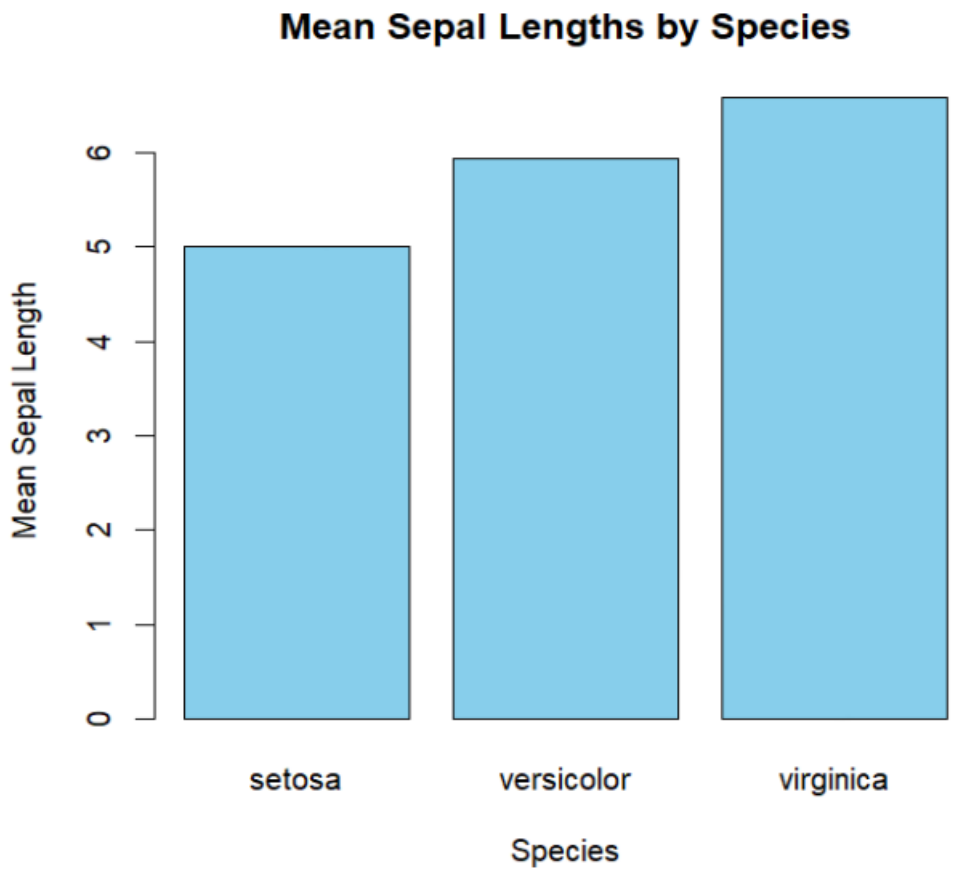
```
hist(iris$Sepal.Length,
    main = "Histogram of Sepal Lengths",
    xlab = "Sepal Length",
    ylab = "Frequency",
    col = "skyblue",
    border = "black")
```



Bargraph:

A bar graph, also known as a bar chart, is a type of data visualization that represents categorical data with rectangular bars. The length or height of each bar is proportional to the frequency or count of the data it represents. Bar graphs are commonly used to compare the values of different categories or to show the distribution of categorical data.

```
mean_sepal_lengths <- tapply(iris$Sepal.Length, iris$Species, mean)
# Create a bar graph of mean sepal lengths by species
barplot(mean_sepal_lengths,
    main = "Mean Sepal Lengths by Species", xlab = "Species", ylab = "Mean Sepal Length", col = "skyblue",border = "black")
```



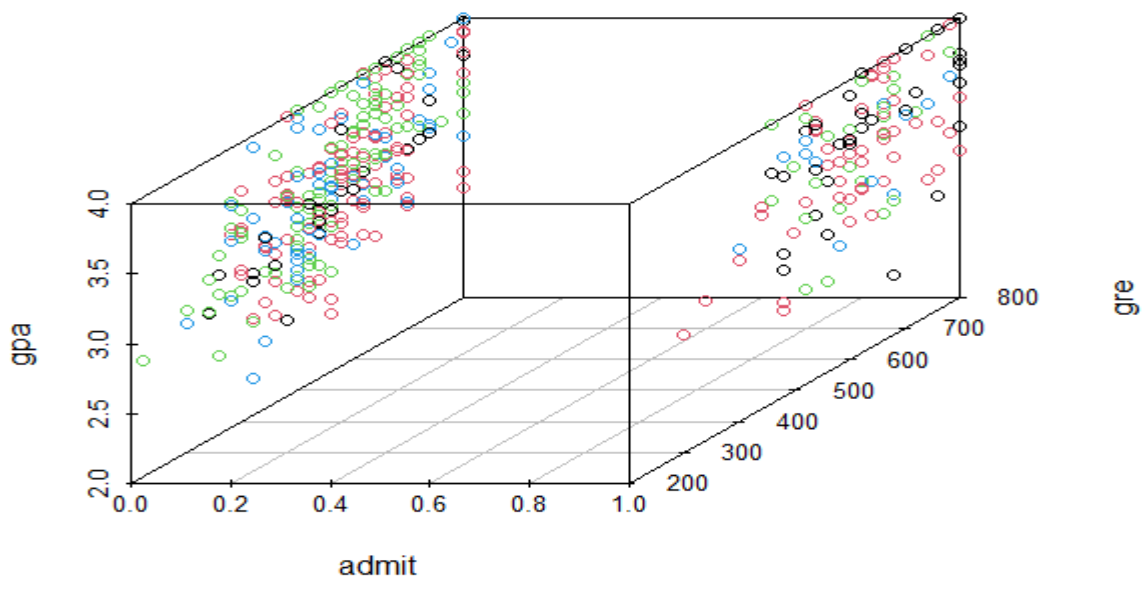
Generating 3d scatter plot:

install.packages("scatterplot3d")

#scatter plot 3d

library("scatterplot3d")

scatterplot3d(d)



ETL (Extract, Transform, Load) is a process used in data warehousing and business intelligence to extract data from various sources, transform it into a consistent format, and load it into a target database or data warehouse. There are several popular tools and platforms available for implementing the ETL process. Here are a few of them:

- Informatica PowerCenter
- IBM InfoSphere DataStage
- Apache Spark
- Pentaho Data Integration (Kettle)

Word Cloud: A word cloud is a visual representation of text data, where the size of each word corresponds to its frequency or importance within the text. In a word cloud, more frequently occurring words are typically displayed in larger font sizes, while less frequent words are displayed in smaller font sizes.

```
install.packages("wordcloud")
library(wordcloud)
# Sample text data
text <- "Hello world hello data science world machine learning R R R data visualization"
# Generate word frequencies
word_freq <- table(unlist(strsplit(text, "\\s+")))
# Create a word cloud
wordcloud(words = names(word_freq), freq = word_freq, min.freq = 1,
          max.words = 100, random.order = FALSE, colors = brewer.pal(8, "Dark2"))
```

Tableau: Tableau is a powerful and widely used data visualization software that allows users to create interactive and shareable visualizations, dashboards, and reports from various data sources. It provides an intuitive drag-and-drop interface that enables users to quickly analyze, visualize, and understand their data without requiring extensive programming or technical skills.

Tableau is designed to handle large datasets efficiently and offers a wide range of visualization options, including bar charts, line charts, scatter plots, maps, histograms, heatmaps etc.

One of Tableau's key strengths is its ability to connect to a variety of data sources, including spreadsheets, databases, cloud services, and web data connectors. It supports live connections as well as in-memory data processing, allowing users to work with real-time data and make informed decisions quickly.

- Connect to Data Source
- Data Preparation
- Build Visualization
- Create Dashboards
- Share and Collaborate

Connect to Data source: Choose Data source from various options like spreadsheets, databases, cloud storage, and more. Tableau supports various data formats like CSV, Excel, Access, SQL databases, etc.

Data Preparation: Once connected, Tableau allows to prepare data for visualization. This may involve filtering out irrelevant data, aggregating or disaggregating data, creating calculated fields, joining or blending data from multiple sources, etc.

Build Visualization: Tableau offers various visualization types, including bar charts, line charts, scatter plots, maps, histograms, etc. Choose the appropriate visualization type that best represents data and insights.

Create Dashboards: Dashboards allow you to combine multiple visualizations into a single interactive display, providing a comprehensive view of data.

Share and Collaborate: Once your visualizations are ready, you can share them with others. Tableau offers various options for sharing, including publishing to Tableau Server or Tableau Online, exporting as static images or PDFs, embedding in web pages or presentations.

Tableau facilitates iterative refinement of visualizations based on stakeholder feedback, while also offering advanced features like advanced calculations, statistical analysis, and integration with R and Python for deeper insights and enhanced decision-making.