UNIT II
Data VIsualization

By

C.Sivamurugan (AP/CSE)

Data Visualization

- Data visualization is a crucial aspect of data analysis and communication.
- It involves creating graphical representations of data to reveal patterns, trends, insights, and relationships that might not be apparent in raw data.
- It helps data scientists and analysts explore the characteristics of the dataset, identify patterns, outliers, and relationships, and gain insights that inform further analysis and decision-making.

Importance of Data Visualization

- Simplifies complex data and makes it more understandable.
- Reveals patterns, correlations, and outliers.
- Communicates insights to non-technical stakeholders.
- Supports data exploration and hypothesis generation

Common Types of Data Visualizations

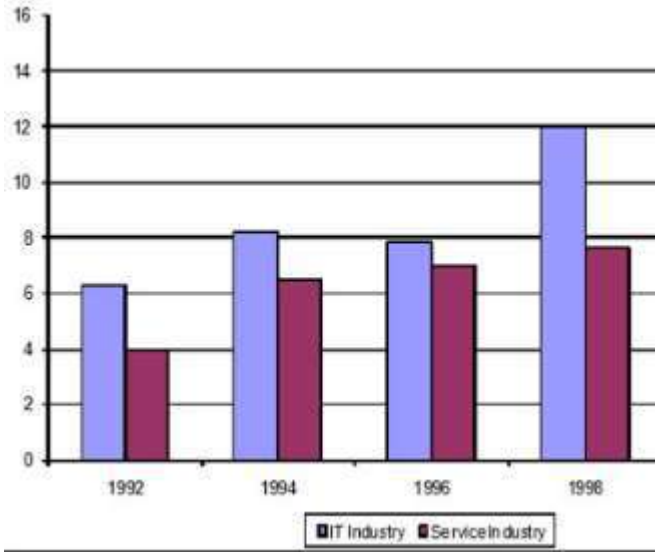
- **Bar Charts:** Used to compare categories or discrete data.
- **Line Charts:** Show trends or changes over time.
- **Scatter Plots:** Display relationships between two variables.
- **Histograms:** Visualize the distribution of a continuous variable.
- **Pie Charts:** Show parts of a whole (use with caution due to limited accuracy).
- **Heatmaps:** Display relationships in a matrix-like format.
- **Box Plots:** Summarize data distribution and identify outliers.
- **Area Charts:** Similar to line charts, often used to show cumulative values

Data Visualizations Libraries

- **ggplot2 (R):** A powerful and flexible plotting system in R.
- **matplotlib (Python):** A widely used plotting library inspired by MATLAB.
- **Seaborn (Python):** Built on top of matplotlib, designed for statistical visualization.
- **Plotly (Python, R, JavaScript):** Interactive and web-based visualizations.
- **Bokeh (Python):** Interactive visualizations for web browsers.
- **Tableau, Power BI:** Data visualization tools with drag-and-drop interfaces.
- **D3.js (JavaScript):** A library for creating custom, interactive visualizations.

1. Bar Charts

- **Bar Charts:** A bar chart is a graphical representation of data using rectangular bars or columns.
- Bar charts are particularly useful for displaying categorical data and making visual comparisons between different data points
- **Library:** ggplot2
- **Function :** geom_bar()
- **Code:**




```
install.packages(“ggplot2"")  # Install library file
 library(ggplot2)
# Create a ggplot object and use geom_bar() to create a bar chart
ggplot(data, aes(x = Category, y = Value)) +   geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Bar Chart using ggplot2", x = "Category", y = "Value")      # Plot Bar Graph
```

2. Line Charts

- **Line Charts:** A line chart, also known as a line plot or line graph, is a type of data visualization that uses lines to represent data points over a continuous interval or time period.
- Line charts are particularly useful for displaying trends and patterns in data that change continuously or sequentially.

- **Library:** ggplot2
- **Function :** geom_line()
- **Code:**

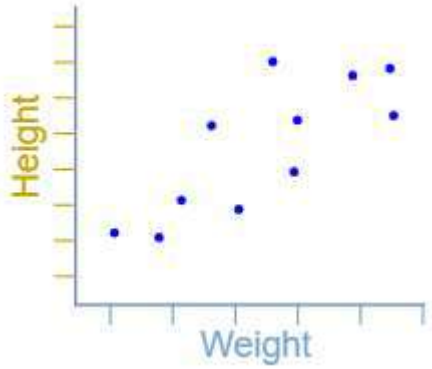
```
install.packages(“ggplot2"")  # Install library file

library(ggplot2)

ggplot(data, aes(x = Time, y = Value)) +   geom_line(color = "blue") +

labs(title = "Line Chart using ggplot2", x = "Time", y = "Value")  # Plot Bar Graph
```

3. Scatter Plots



- **Scatter Plots:** scatterplot or scatter diagram, is a type of data visualization that uses points to represent individual observations or data points in a two-dimensional coordinate system.
- Scatter plots are particularly useful for visualizing the relationship between two continuous variables and for identifying patterns, trends, clusters, or outliers in the data.
- **Library:** ggplot2
- **Function :** geom_point()
- **Code:**

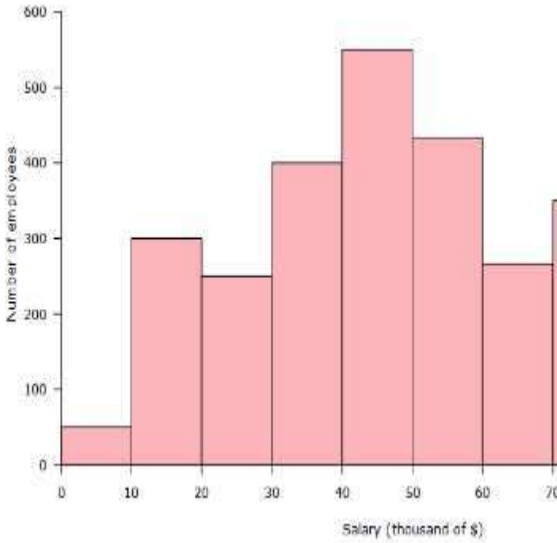
```
install.packages(“ggplot2"")  # Install library file

library(ggplot2)

ggplot(data, aes(x = Variable1, y = Variable2)) +   geom_point(color = "blue") +

labs(title = "Scatter Plot using ggplot2", x = "Variable1", y = "Variable2") # Plot Bar Graph
```

4. Histograms



- **Histograms:** A histogram is a graphical representation of the distribution of a dataset.
- Histograms are particularly useful for understanding the shape, central tendency, spread, and potential outliers of a dataset

- **Library:** ggplot2
- **Function :** geom_histogram()
- **Code:**

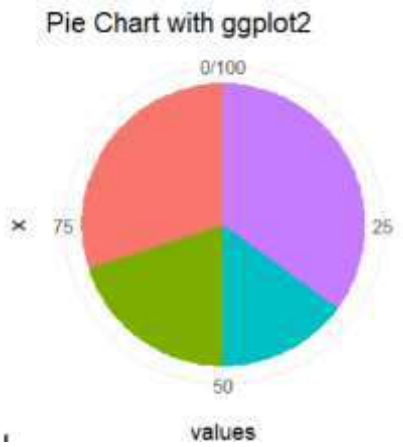
```
install.packages(“ggplot2"")  # Install library file

library(ggplot2)

ggplot(data, aes(x = Variable)) +   geom_histogram(binwidth = 5, fill = "blue", color
= "black") +   labs(title = "Histogram Example", x = "Variable", y = "Frequency")

# Plot Bar Graph
```

5. Pie Charts



- **Pie Charts:** A pie chart is a circular graphical representation that displays the distribution of a categorical variable as proportions or percentages of a whole.

- **Library:** ggplot2
- **Function :** geom_bar()
- **Code:**

```
install.packages(“ggplot2"")  # Install library file

library(ggplot2)

pie_chart <- ggplot(data, aes(x = "", y = value, fill = category)) + geom_bar(stat = "identity", width = 1) + coord_polar(theta = "y") + theme_void() + labs(title = "Pie Chart Example")     # Plot Bar Graph
```

6. Heat Map

- **Heat Map:** A heatmap is a graphical representation of data in a two-dimensional format, where values are represented using a color scale. be easily identified.
- Heatmaps are particularly useful for visualizing data matrices or tables, where each cell's value is depicted as a color, allowing patterns and relationships to be easily identified.
- **Library:** ggplot2
- **Function :** geom_tile()
- **Code:**

```
install.packages(“ggplot2"")  # Install library file

library(ggplot2)

heatmap_plot <- ggplot(data, aes(x = column, y = row, fill = value)) + geom_tile() +

scale_fill_gradient(low = "blue", high = "red") + # Color scale labs(title = "Heatmap

Example")              # Plot Bar Graph
```