5. Big Data Visualisation

1. Data Visualization:-

- Data visualization is the graphical representation of information. In this guide we will study what is Data visualization and its importance with use cases.
-
- Data visualization translates complex data sets into visual formats that are easier for the human brain to understand.
- This can include a variety of visual tools such as:
    - **Charts**: Bar charts, line charts, pie charts, etc.
    - **Graphs**: Scatter plots, histograms, etc.
    - **Maps**: Geographic maps, heat maps, etc.
    - **Dashboards**: Interactive platforms that combine multiple visualizations.
- The primary goal of data visualization is to make data more accessible and easier to interpret allow users to identify patterns, trends, and outliers quickly.

Key Importance:

1. **Simplifies Complex Data:**
   Big Data is often unstructured and complex. Visualization makes it easier to digest by presenting it in charts, graphs, heat maps, etc., which highlight trends and patterns.

2. **Faster Decision-Making:**
   With visual data representation, analysts and business leaders can make quicker, data-driven decisions rather than sifting through raw data.

3. **Reveals Hidden Patterns:**
   Visualization tools can uncover patterns, correlations, and outliers that may not be obvious in raw datasets. This is crucial in areas like fraud detection, marketing analysis, and customer behavior.

4. **Better Communication:**
   Visuals help in communicating insights clearly to both technical and non-technical stakeholders. Dashboards and infographics can be shared across departments for better collaboration.

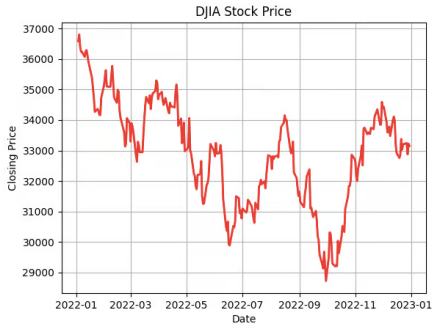
5. **Interactive Exploration:**
   Modern tools allow for real-time, interactive exploration of data, helping users drill down into specifics without needing deep technical skills.

Challenges to Big Data Visualization:

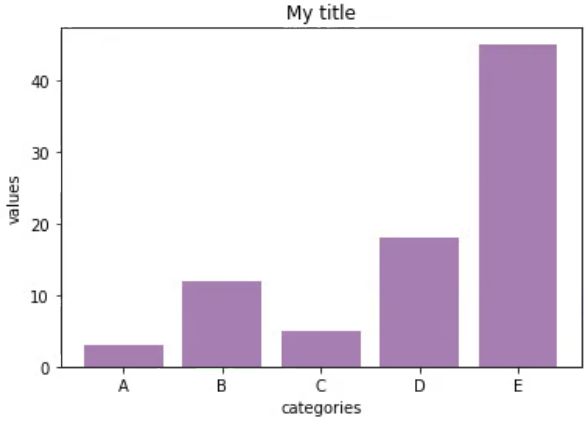
1. **Handling Large Volume of Data:**
    - Big Data involves terabytes to petabytes of data.
    - Rendering and processing such large datasets for visualization is computationally expensive and time-consuming.
2. **High Data Variety:**
    - Data comes in various forms—structured (databases), semi-structured (XML, JSON), and unstructured (videos, images, text).
    - Integrating and visualizing such diverse data types in a single platform is difficult.
3. **Real-Time Visualization:**
    - Big Data often requires real-time analysis (e.g., stock trading, sensor monitoring).
    - Generating real-time visual updates without performance lags is a major challenge.
4. **Scalability Issues:**
    - As the data grows, visualization systems need to scale.
    - Many tools cannot scale effectively, leading to system crashes or long processing times.
5. **Choosing the Right Visualization Technique:**
    - Picking the wrong chart type or visual format can mislead users or hide important insights.
    - There's no one-size-fits-all—visualizations must be tailored to the audience and data context.
6. **Data Quality and Noise:**
    - Big Data may include inaccurate, incomplete, or irrelevant data.
    - Poor data quality can result in misleading visualizations and wrong interpretations.
7. **User Interpretation:**
    - Visualizations are open to interpretation, and non-experts may misread them.
    - This can lead to poor decisions if visuals are not clearly labeled or explained.
8. **Privacy and Security Concerns:**
    - Visualizing personal or sensitive data (like health records or financial information) requires strict privacy controls.
    - Improper access or leaks through visualization dashboards can lead to data breaches.

2. Methods / Techniques of Data Visualization:-

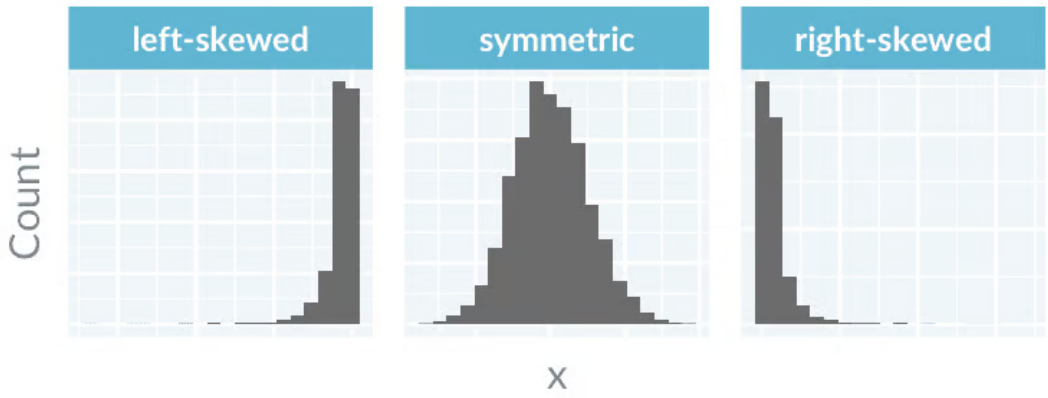
1. **Line Plots**
   Line plots show how something changes over time. You place time on the x-axis and values (like temperature or sales) on the y-axis. They are great for spotting trends, like how stock prices moved during a year.



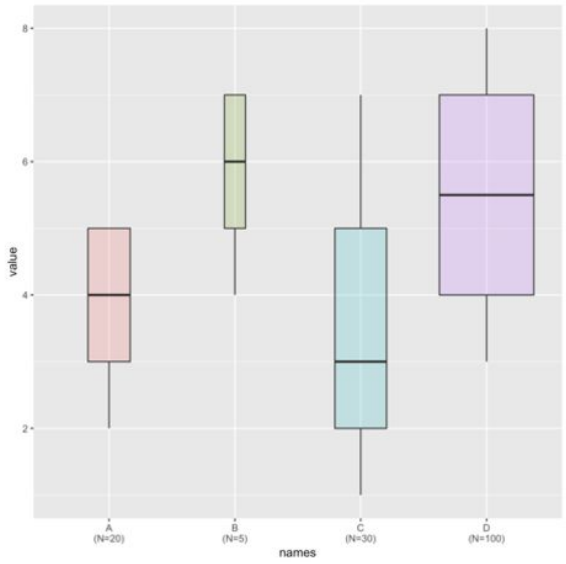
2. **Bar Charts**
   Bar charts use rectangular bars to compare different items. Longer bars mean higher values. They're useful for comparing things like sales in different regions or performance across teams.



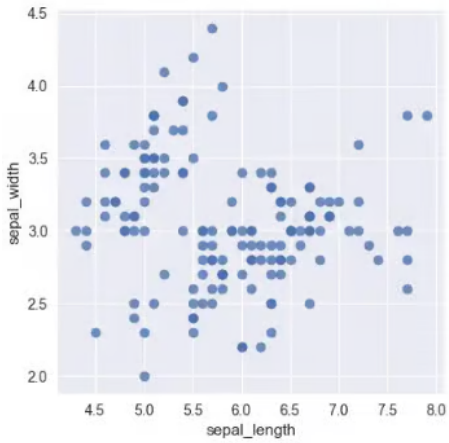
3. **Histograms**
   Histograms show how often numbers appear in a dataset. The data is grouped into ranges (called bins), and each bar shows how many times values fall into each range. It's great for seeing data distribution.



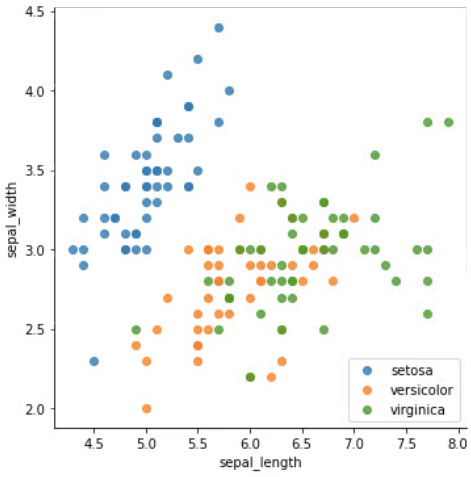
4. **Box Plots**
   Box plots summarize a dataset using its median, upper and lower quartiles, and show outliers. They help you quickly understand the spread and identify unusual values in your data.



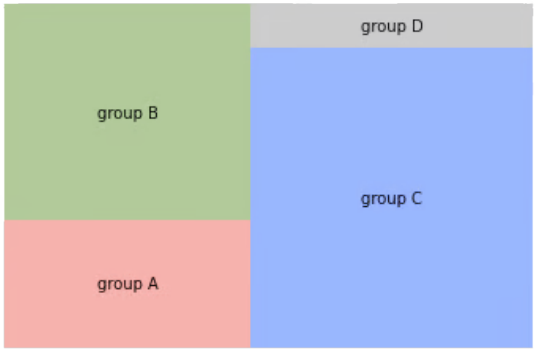
5. **Scatter Plots**
   Scatter plots display the relationship between two continuous variables. Each point represents a pair of values. They help you see if there's a pattern or correlation between the two.



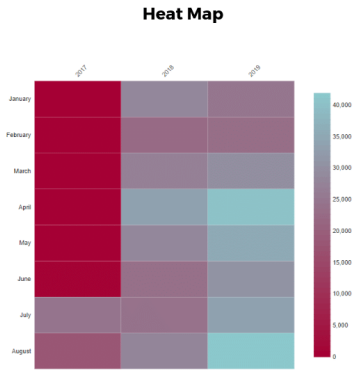
6. **Bubble Plots**
   Bubble plots are like scatter plots but add a third variable using the size of the bubble. They help show more details in a single graph, like GDP, population, and life expectancy.



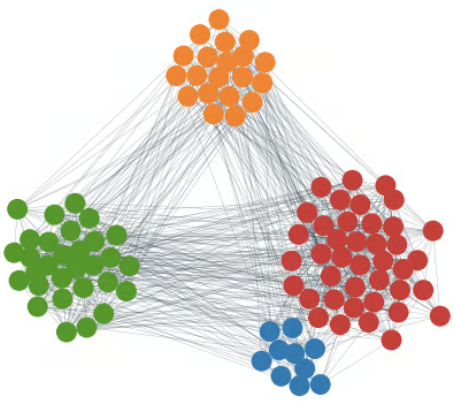
7. **Treemaps**
   Treemaps show parts of a whole using nested rectangles. Each box's size represents the proportion of a category. They're helpful for visualizing things like budget breakdowns or market share.



8. **Heatmaps**
   Heatmaps use colors to show values in a matrix. Darker or brighter colors mean higher or lower numbers. They're often used to show correlations or data density. A heatmap is a common and beautiful matrix plot that can be used to graphically summarize the relationship between two variables. The degree of correlation between two variables is represented by a color code.



9. **Network Diagrams**
   Network diagrams use points (nodes) and lines (edges) to show how things are connected. They're great for visualizing social networks, transportation systems, or data relationships.



3. Type of Data Visualization:-

1. Comparative Visualization

Purpose:
Used to **compare values** across different groups or categories. It helps identify differences, patterns, or similarities between items.

How it works:
Data from different categories is shown side-by-side to allow for easy visual comparison. This is useful in identifying the best-performing or worst-performing groups.

Examples:

- **Bar Charts**: Horizontal or vertical bars are used to represent data values.
  Example: Comparing sales of different products in a company.

- **Column Charts**: A variation of bar charts with vertical columns.
  Example: Comparing yearly revenues of companies.

2. Trend-Based Visualization

Purpose:
To **show how data changes over time**. This helps in identifying patterns, trends, or changes in behavior over specific intervals.

How it works:
Time is placed on the x-axis, and the values of the variables are plotted on the y-axis. These visualizations are helpful in analyzing patterns over days, months, or years.

Examples:

- **Line Graphs**: Display trends using lines that connect data points over time.
  Example: Stock market prices throughout the year.

- **Area Charts**: Similar to line charts but with the area below the line filled with color.
  Example: Website traffic growth month by month.

3. Distribution Visualization

Purpose:
To understand the **spread or frequency of data** across a range. It helps in detecting outliers, clusters, and the shape of the data distribution.

How it works:
The data range is divided into intervals or categories, and the number of data points in each interval is shown visually.

Examples:

- **Histograms**: Show how frequently data falls into specific ranges (bins).
  Example: Distribution of student test scores.

- **Scatter Plots**: Plot individual data points to see how values are distributed.
  Example: Height vs. weight of people.

4. Relationship Visualization

Purpose:
To reveal **correlations or connections** between two or more variables. Helps in identifying whether variables move together or independently.

How it works:
Each variable is represented on a separate axis, and data points are plotted based on their values to see if there's a visible trend or pattern.

Examples:

- **Scatter Plots**: Show the relationship between two variables using dots.
  Example: Study time vs. exam score.

- **Bubble Charts**: Enhance scatter plots by using the size of the dot to represent a third variable.
  Example: GDP vs. life expectancy with bubble size representing population.

5. Part-to-Whole Visualization

Purpose:
To show **how parts contribute to a whole**. It highlights proportions and percentages.

How it works:
Displays the contribution of each part compared to the total value.

Examples:

- **Pie Charts**: A circular chart divided into slices to show proportions.
  Example: Market share of different companies.

- **Stacked Bar Charts**: Bars are divided into segments representing different parts.
  Example: Sales of different product categories within total sales.

6. Geospatial Visualization

Purpose:
To display **location-based data** visually. Helps in identifying regional trends or geographic patterns.

How it works:
Uses maps to show data associated with geographic locations.

Examples:

- **Choropleth Maps**: Use color gradients to show data density across areas.
  Example: Population density across states.

- **Dot Distribution Maps**: Show individual data points by location.
  Example: COVID-19 case locations.

7. Hierarchical Visualization

Purpose:
To show **parent-child or nested relationships** within a dataset.

How it works:
Data is structured in levels, and each level is visualized to show how items are grouped.

Examples:

- **Tree Maps**: Use nested rectangles to represent parts of a hierarchy.
  Example: Storage space usage by folder size.

- **Dendrograms**: Tree-like diagrams used for clustering or hierarchy visualization.
  Example: Taxonomy of animal species.

8. Multivariate Visualization

Purpose:
To **analyze multiple variables** simultaneously in one graphic. Useful for complex datasets.

How it works:
Visualizations use color, size, and shape to represent multiple variables.

Examples:

- **Heatmaps**: Show values using color intensity.
  Example: Correlation matrix between financial indicators.

- **Radar Charts (Spider Charts)**: Show multivariate data on axes starting from the same point.
  Example: Performance comparison of athletes in different skills.

4. Tools of Data Visualization:-

i) Tableau:-

- **Tableau** is a powerful tool that helps people like data analysts and businesses turn raw data into clear, interactive visuals quickly.
- It makes understanding data easy without needing deep technical skills and ensures your data stays secure.
- It's widely used by businesses, analysts, and researchers to quickly analyze data and make informed decisions.

Tableau Products:

1. **Tableau Desktop:**
   A software for creating detailed and interactive charts and dashboards. It connects to many data sources like databases, spreadsheets, or cloud services. It's mainly for individuals or small teams.
2. **Tableau Public:**
   A free version where you can create and share visualizations publicly online. Great for students or bloggers, but the data is not private.
3. **Tableau Online:**
   A fully cloud-hosted version where you can create, share, and access visualizations from anywhere without worrying about installing software or managing servers.
4. **Tableau Server:**
   A company-wide version installed on your own servers to securely share visualizations and integrate with existing company security systems. Best for larger organizations

How Does Tableau Work?

- **Connect to Data:** Tableau can connect to many types of data sources like Excel files, databases, cloud services, and big data platforms.
- **Drag and Drop Interface:** Users don't need to write complex code; they can create visualizations by dragging and dropping fields like sales, dates, or categories onto the canvas.
- **Create Visuals:** Tableau offers many chart types such as bar charts, line charts, maps, scatter plots, and more. You can customize colors, labels, and filters to highlight important insights.
- **Build Dashboards:** Multiple visualizations can be combined into interactive dashboards that provide an overview of key metrics.
- **Share and Collaborate:** Dashboards and reports can be published online or shared within teams so everyone can explore the data and insights.

Why Use Tableau for Data Visualization?

- **User-Friendly:** Easy for beginners but powerful enough for experts.
- **Interactive:** Users can click, filter, and drill down into data to explore it more deeply.
- **Fast Insights:** Quickly uncover patterns, trends, and outliers without waiting for complex reports.
- **Versatile:** Works with almost any data source and supports various industries and use cases.
- **Visual Appeal:** Creates attractive, professional visuals that make data storytelling more effective.

ii) Candela

Candela is a set of web-based visualization components based on the **Vega** visualization grammar. It is designed to make **creating interactive, high-quality visualizations easier**—especially for scientific and analytical data.

Key Features:

- Built on top of **Vega, D3.js, and WebGL**.
- Designed for **scalability** and large datasets.
- Offers several chart types: heatmaps, scatter plots, bar charts, parallel coordinates, and more.
- Easily integrates with Python and Jupyter notebooks using **Python bindings**.

Use Case:

- Ideal for data scientists working in **Python notebooks** who need interactive visualizations.
- Example: Visualizing gene expression data or large matrix-style data.

iii) D3.js (Data-Driven Documents)

D3.js is a powerful **JavaScript library** for building custom, interactive data visualizations directly in the browser using HTML, SVG, and CSS.

Key Features:

- Gives you full control over the final look and feel of the visualization.
- Supports dynamic and interactive visualizations.
- Highly customizable but has a steep learning curve.

Use Case:

- Used for creating **custom dashboards**, **interactive charts**, and **visual storytelling**.
- Example: Animated scatter plots, dynamic bar charts, interactive maps.

iv) Google Chart API

The **Google Chart API** is a **free web service** that creates charts and graphs from user-supplied data and parameters, rendered using **HTML5/SVG**.

Key Features:

- Simple to use with minimal coding knowledge.
- Provides a variety of chart types: line, bar, pie, geo, gauge, etc.
- Built-in support for **interactive dashboards**.
- Easily embeds into websites and web apps.

Use Case:

- Ideal for **quick business dashboards**, **presentations**, or **web-based reports**.
- Example: Pie chart showing sales distribution by region.

5. Explain data visualization with respect to l-D, 2-D, 3-D data:-

i) 1-D Data Visualization

Definition:
1-D (one-dimensional) data visualization represents data involving only a single variable. It focuses on showing the distribution, frequency, or summary statistics of that variable.

Types:

- **Bar Chart:** Shows the count or value for each category.
- **Histogram:** Displays the frequency distribution of numeric data by grouping data into bins.
- **Line Chart:** Shows the trend of a variable over time (time series).
- **Pie Chart:** Represents parts of a whole as slices of a circle.

Example:
A histogram showing the distribution of students' test scores in a class.

ii) 2-D Data Visualization

Definition:
2-D (two-dimensional) data visualization involves two variables, often to examine the relationship or correlation between them. It plots data points based on two dimensions (x and y axes).

Types:

- **Scatter Plot:** Displays individual data points based on two variables to identify correlation or clusters.
- **Line Graph:** Shows changes of one variable relative to another over time or categories.
- **Heatmap:** Uses color intensity to represent values across two dimensions (e.g., correlation matrix).
- **Area Chart:** Shows cumulative values over time.

Example:
A scatter plot showing the relationship between advertising spend (x-axis) and sales revenue (y-axis).

iii) 3-D Data Visualization

Definition:
3-D (three-dimensional) data visualization involves three variables and represents data in a three-dimensional space, adding depth to data analysis and helping to observe interactions among three factors.

Types:

- **3D Scatter Plot:** Plots data points in three dimensions to explore relationships among three variables.
- **Surface Plot:** Visualizes data as a surface in 3D space, often to show trends or peaks.
- **3D Bar Chart:** Bars extend in three dimensions representing multiple variables.

Example:
A 3D scatter plot showing height, weight, and age of a group of people.

6. Write two data visualization functions:-

1) Seaborn:

a) **sns.scatterplot()**
Creates a scatter plot to show the relationship between two variables.
Example:

```
import seaborn as sns

sns.scatterplot(x='total_bill', y='tip', data=tips)
```

b) **sns.boxplot()**
Creates a box plot to show the distribution and detect outliers in data.
Example:

```
import seaborn as sns
sns.boxplot(x='day', y='total_bill', data=tips)
```

2) Matplotlib:

a) **plt.plot()**
Creates a basic line plot, useful for showing trends over a continuous variable (often time).
Example:

```
import matplotlib.pyplot as plt

plt.plot([1, 2, 3, 4], [10, 20, 25, 30])
plt.show()
```

b) **plt.bar()**
Creates a bar chart to compare categorical data.
Example:

```
import matplotlib.pyplot as plt
plt.bar(['A', 'B', 'C'], [10, 20, 15])

plt.show()
```

3)Pandas (built-in plotting) :-

a)DataFrame.plot.line() — Plots a line graph from a DataFrame.

b)DataFrame.plot.bar() — Plots a bar chart from a DataFrame. with example

Example:-

```
import pandas as pd
data = {'Year': [2020, 2021, 2022],
        'Sales': [100, 150, 200],
        'Revenue': [300, 350, 400]}
df = pd.DataFrame(data)
df.plot.line(x='Year', y='Sales')
df.plot.bar(x='Year', y='Revenue'
```