ANNAMACHARYA INSTITUTE OF TECHNOLOGY AND SCIENCES

RAJAMPET

(Autonomous)

Department of ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Lecture Notes

Name of the Faculty: N. Swathi

Class: III B. TECH II SEM

Branch: CSE (DS)

Name of the Course: Data Visualization

Subject Code: 23A3262T

Academic Year: 2025-2026

UNIT I

Introduction to Data Visualization & Perception

1.1 What is Data Visualization?

Data visualization is the graphical representation of data using charts, graphs, and tables to highlight trends, patterns, and outliers more clearly than textual or numeric data. Rooted in humanity's earliest form of communication—visual expression—it remains a powerful and efficient way to convey information. Visualization reduces confusion, captures attention, and emphasizes key messages without requiring users to analyze lengthy data. By transforming complex data into clear visuals, it enables quick understanding and meaningful insights. These insights support informed decision-making, such as identifying sales trends and causes, making data visualization essential in the data analytics process.

1.2 Importance and Scope of Data Visualization

a. **Simplify data**: Reduce the complexity of data by highlighting the key information and removing the noise. For example, a table with hundreds of rows and columns can be simplified into a pie chart that shows the percentage distribution of different categories.
b. **Organize data**: Structure and group data into meaningful categories and hierarchies. For example, a bar chart can show the comparison of different groups or subgroups within a larger dataset.
c. **Summarize data**: Condense large amounts of data into a concise overview. For example, a line graph can show the trend of a variable over time with a single line.
d. **Explore data**: Discover new insights and relationships in data that might not be obvious otherwise. For example, a scatter plot can show the correlation between two variables with dots.
e. **Analyze data**: Perform various types of analysis on data, such as descriptive, diagnostic, predictive, and prescriptive analysis. For example, a histogram can show the frequency distribution of a variable and help us identify outliers or anomalies.
f. **Evaluate data**: Assess the quality and validity of data, such as accuracy, completeness, consistency, and timeliness. For example, a box plot can show the range, median, quartiles, and outliers of a variable and help us detect errors or missing values.
g. **Compare data**: Compare different datasets or scenarios and identify similarities or differences. For example, a stacked bar chart can show the proportion of different segments within each group and help us compare their relative sizes.
h. **Choose data**: Select the most relevant and useful data for our purpose and goals. For example, a dashboard can show the key performance indicators (KPIs) and metrics that matter most for our business or project.
i. **Identify trends**: identify trends in data by showing how variables change over time. For example, a line graph can show the increase or decrease of a variable over time and indicate its direction and magnitude.
j. **Forecast trends**: forecast trends in data by showing how variables are expected to change in the future based on historical or current data. For example, a trendline can show the best fit line for a set of points and project its continuation into the future.
k. **Discover patterns**: discover patterns in data by showing how variables are related or grouped together. For example, a heat map can show the intensity or frequency of a variable across two dimensions and reveal clusters or hotspots.
l. **Classify patterns**: classify patterns in data by showing how variables are divided or categorized into different types or classes. For example, a pie chart can show the percentage of each category within a variable and indicate its composition or diversity.

Pre-attentive processing is the **human brain's ability to quickly detect visual patterns or differences** before conscious attention is applied. It allows viewers to **grasp key insights in a visualization almost instantly**.

Key Features:

- Works **within 200–500 milliseconds**.
- Helps highlight **important data points, trends, or anomalies**.

Visual Attributes that Trigger Pre-Attentive Processing:

| | **Color**
Color differences are detected instantly by the brain. Bright or contrasting colors highlight key data points, helping viewers quickly identify patterns, alerts, or important values.

| | **Orientation**
Elements with different directions stand out immediately. Changes in line or bar orientation help users quickly notice anomalies, trends, or special data categories.

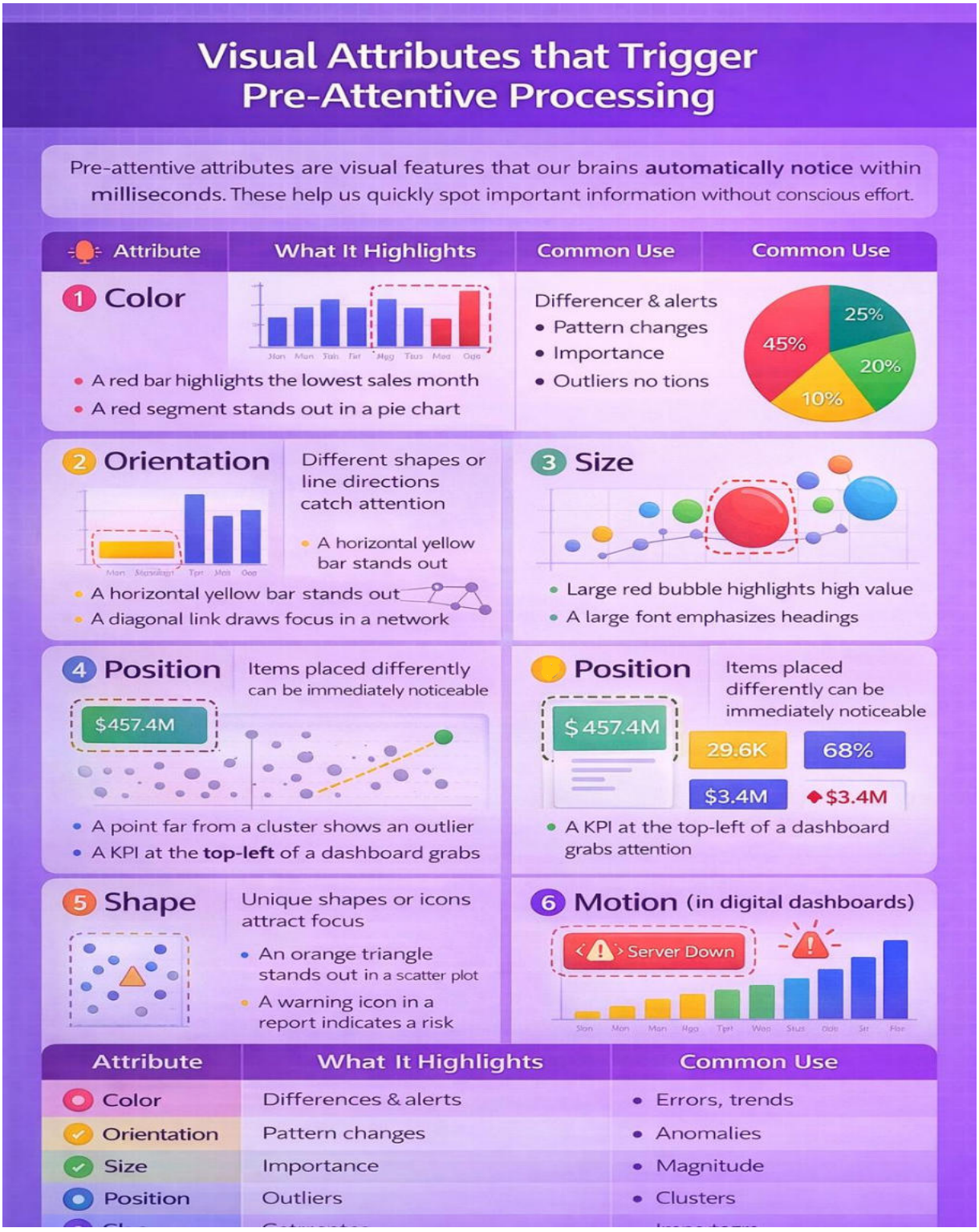
| | **Size**
Larger or smaller elements draw instant attention. Size variations help viewers quickly recognize importance, magnitude, or outliers within visual representations.

| | **Position**
Objects placed away from common alignment are easily noticed. Position helps identify clusters, trends, or unusual values without conscious effort.

| | **Shape**
Unique shapes attract immediate focus. Shape differences help distinguish categories, highlight exceptions, or represent special events in data visualizations.

Motion

Movement captures attention automatically. Animated elements emphasize real-time changes, alerts, or transitions, making critical information noticeable instantly.



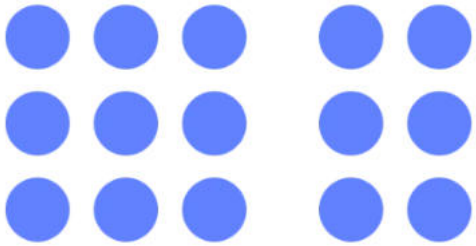
Gestalt Principles in Visualization

Definition:
Gestalt principles describe **how humans naturally perceive visual elements as groups or patterns**, helping make complex data easier to interpret.

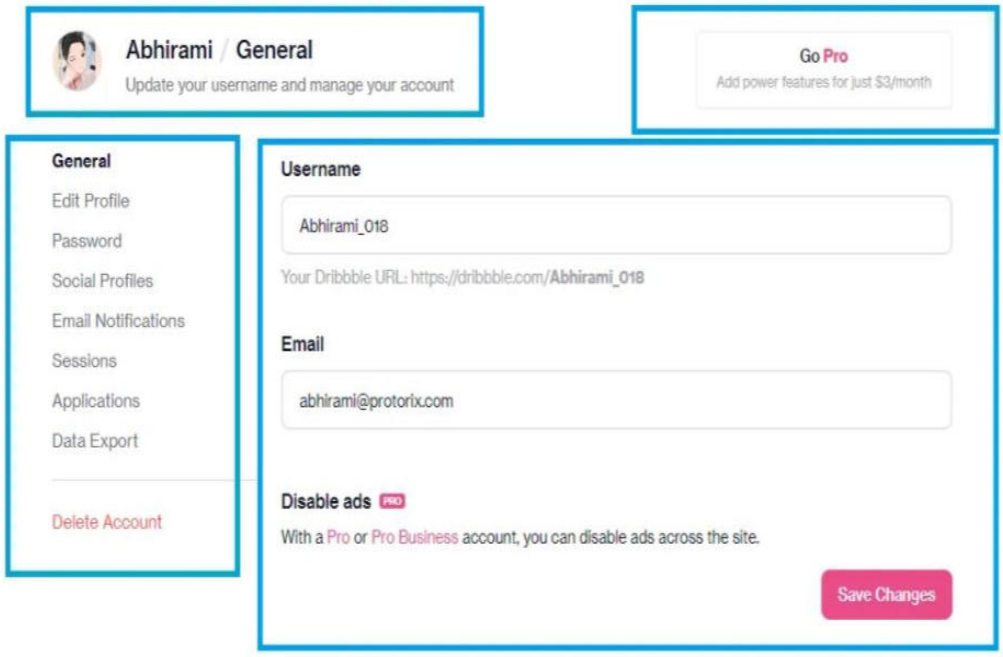
Key Principles:

1. **Proximity:**

   It states that the elements that are close or proximate together are considered more related than those that are not close. With the help of proximity, users can organize the structure and make it understandable.

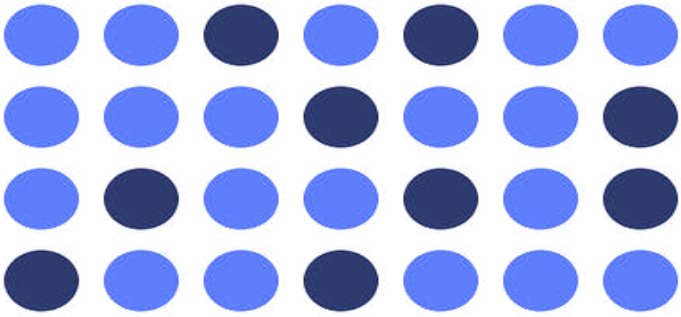


Example: If users to group certain structures more easily than representing them with strong border lines, we can use whitespace between each structure else placing the info that is related as close as possible to the rest.

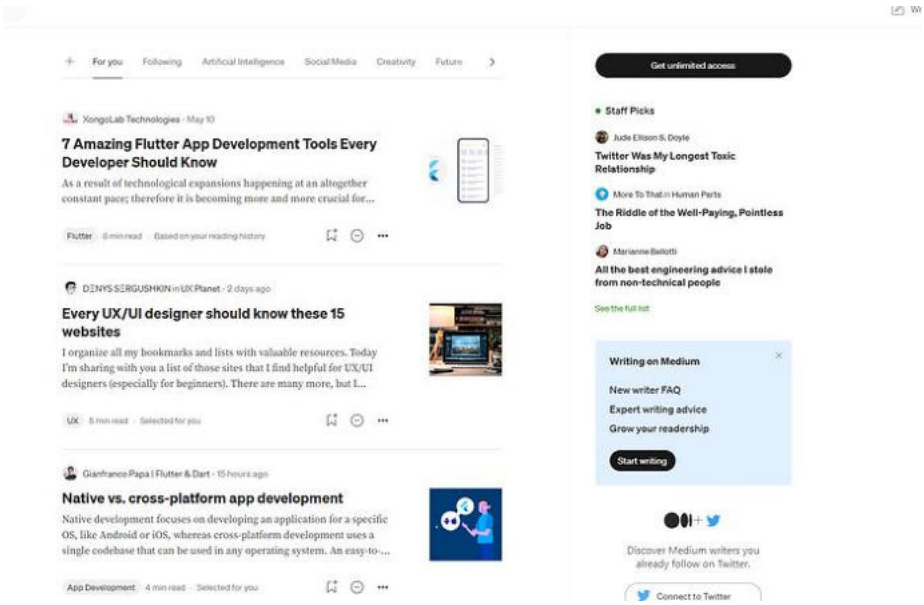


2. Similarity:

Similarity states that similar elements are visually grouped, regardless of their proximity to each other. The visually grouped elements can be color, shape, or size.



Example: In the below example, the law of similarity is explained via button shapes and colors. Highlighting the title text with bold font and color will make users get an easy understanding of what the section is about.



3. **Continuity:**
   - The eye follows lines or curves, making trends easier to detect.
   - Example: Line chart showing recovery rates over time.
4. **Closure:**

It states that whenever we see a complex form of pattern our mind tries to frame a simple and recognizable pattern. Our minds always create a closure pattern.

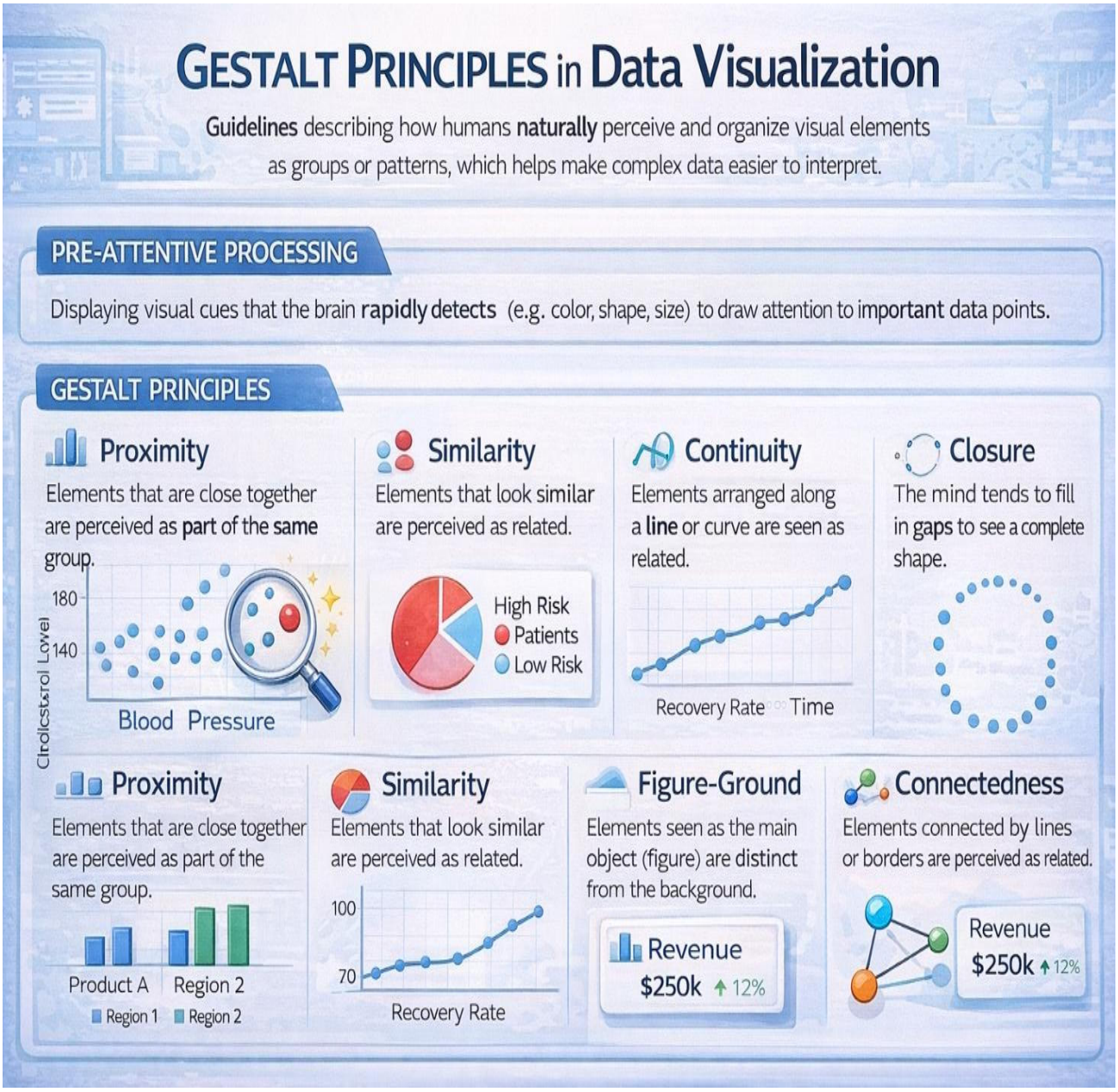
Example: The below example is basically formed of black strokes but our mind fills the gap and tries to look at it as a bear.

5.

   - The mind fills gaps to perceive a complete shape.
   - Example: A partially drawn circle is still seen as a circle.
6. **Figure-Ground:**

- Viewers distinguish objects (figure) from background (ground).
- Example: Highlighting a KPI card against a subtle dashboard background.

7. **Connectedness:**
   - Items connected by lines or borders are perceived as related.
   - Example: Linking nodes in a network graph.



1. Data-Ink Ratio

- The **Data-Ink Ratio** is a concept introduced by Edward Tufte in *The Visual Display of Quantitative Information*.

- It measures how much of the ink (or pixels) in a chart actually **represents data** versus **decorative or non-essential elements**.
- The goal is to **maximize the data-ink ratio**, i.e., use ink efficiently to focus attention on the data itself.

Formula:

$$\text{Data-Ink Ratio} = \frac{\text{Data Ink}}{\text{Total Ink}}$$

Where:

- **Data Ink** = The part of the chart that represents actual data points, values, or trends.
- **Total Ink** = All ink used to produce the graphic, including gridlines, backgrounds, labels, and decorations.

Key Principles:

- Remove **non-data ink** (unnecessary decoration, 3D effects, heavy backgrounds).
- Keep **essential elements** (axes, labels, legends) that help interpret the data.
- Goal: Highlight the **signal**, minimize the **noise**.

Examples:

- **Low Data-Ink Ratio:** 3D pie chart with shadows, gradient backgrounds, extra icons → distracts the viewer.
- **High Data-Ink Ratio:** Clean bar chart with flat colors, minimal gridlines, clear labels → emphasizes the data.

Why it matters:

- Increases clarity and speed of comprehension.
- Reduces misinterpretation caused by decorative clutter.

2. Data Density

- Data density measures how **much information is packed into a visual area**.
- Higher data density allows the viewer to **see patterns and trends** more efficiently.
- Conceptual formula:

$$\text{Data Density} = \frac{\text{Number of data points displayed}}{\text{Graphical area}}$$

Guidelines:

- **High data density:** More data points per chart area → can show trends, correlations, and outliers.
- **Low data density:** Sparse data → easier for beginners but may underutilize space.
- **Balance:** Too high data density can overwhelm, too low wastes screen space.

Examples:

1. **High density:** Scatterplot of 10,000 patient records showing blood pressure vs age → reveals clusters and anomalies.
2. **Low density:** A pie chart with 3 categories → minimal data, simple to read, but limited insight.

Why it matters:

- Efficiently communicates complex datasets.
- Helps analysts detect **patterns, correlations, and trends** quickly.
- Supports **storytelling** by showing both overview and detail.

3. Lie Factor

- The **Lie Factor** evaluates how accurately a visualization represents the underlying data.
- Introduced by Edward Tufte as a measure of **visual distortion**.

Formula:

$$\text{Lie Factor} = \frac{\text{Size of effect shown in graphic}}{\text{Actual size of effect in data}}$$

Where:

- **Size of effect shown in graphic:** Visual representation (e.g., bar height, pie slice angle).
- **Actual size of effect:** Real-world data change or difference.

Interpretation:

- **Lie Factor = 1:** Visualization is truthful.
- **Lie Factor > 1:** Exaggerated effect → misleading, may overstate trends.
- **Lie Factor < 1:** Understated effect → downplays trends.

Examples:

- **Exaggerated:** Revenue increases 5%, but bar height visually shows 50% → Lie Factor > 1.
- **Downplayed:** Stock loss 20%, but bar barely moves → Lie Factor < 1.

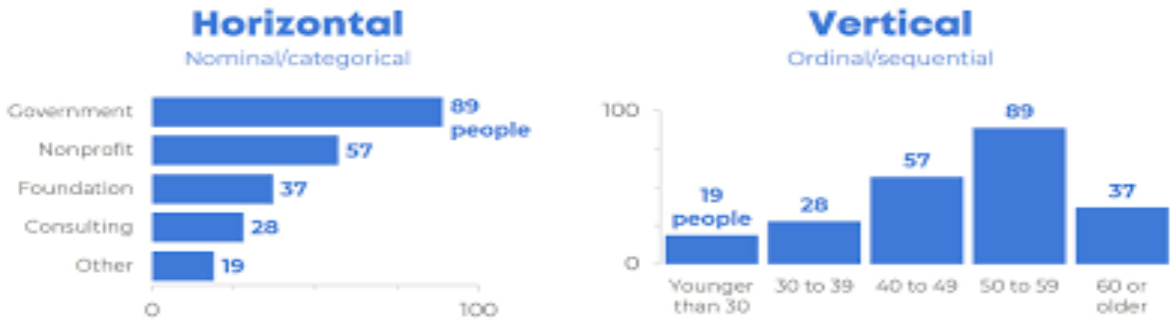
UNIT II

Visualization Techniques for Categorical & Quantitative Data

2.1 Charts for Categorical Data:

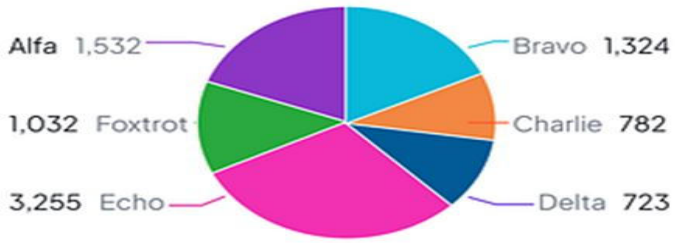
a. Bar Charts / Column Charts

Bar charts are widely used to display categorical data, where bar lengths represent values proportionally. They require two variables and must start at zero to ensure accurate comparisons. Bar charts are ideal for simple category comparisons and are easy to read and interpret. Their clear and compact structure allows variations, such as grouped or stacked bar charts, to effectively present more complex information in a straightforward and understandable manner.



b. Pie Charts

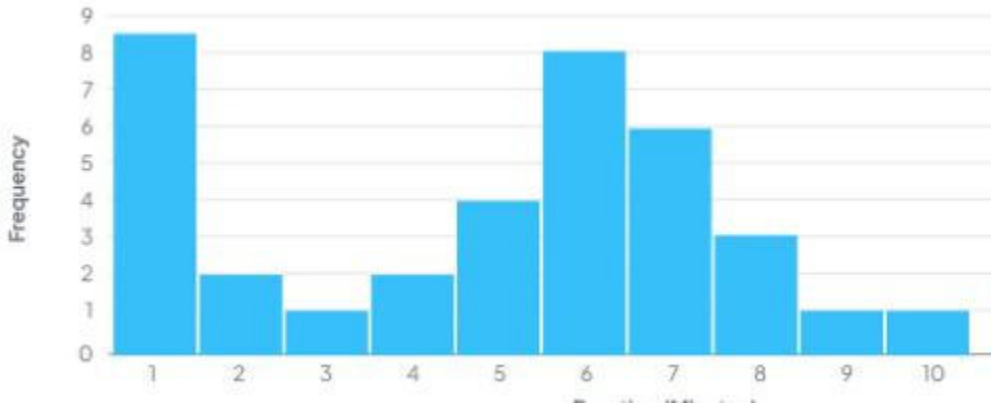
Pie charts are one of the most iconic yet controversial tools in data visualization. While they are commonly used to display proportions, many experts argue that they are not the best method for comparing data. Edward Tufte has famously criticized pie charts for their inefficiency, as some people struggle to compare angles accurately.



2.2 Charts for Quantitative Data:

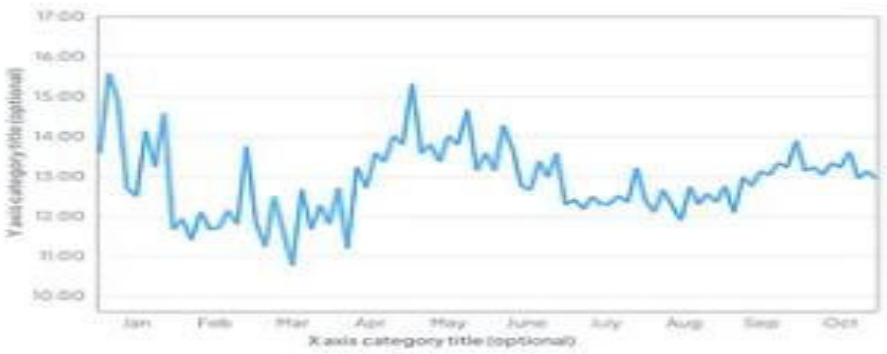
a. Histograms

Histograms display the **frequency of continuous data** by grouping values into intervals called **buckets**. The height of each bar shows how many data points fall into that range. Bars are adjacent to represent continuity, usually with equal widths. Unlike bar charts, which compare categories, histograms reveal the **distribution of data**. They help identify patterns such as symmetric, unimodal, or bimodal shapes, making them useful for spotting trends, predicting frequency behavior, detecting skewness, and understanding overall data patterns.



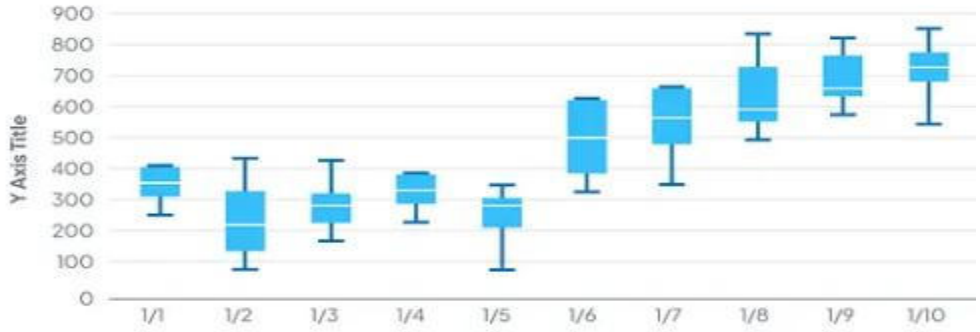
b. Line Charts

Line charts, like bar charts, display data but excel at showing **trends over time**. They handle many data points and small proportional differences, don't require starting at zero, and allow zooming into relevant scales. They are ideal for comparing specific data sets and observing changes or patterns clearly.



c. Box plots

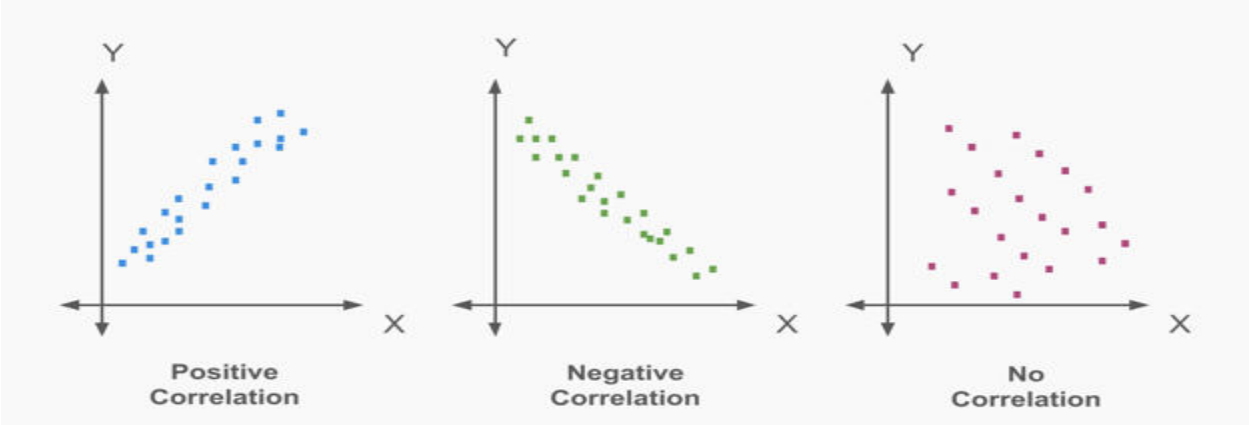
Box plots (or box-and-whisker plots) show data distribution using **five key values**: minimum, first quartile, median, third quartile, and maximum. The **box** represents the middle 50% of data, the **line** inside shows the median, and **whiskers** extend to the lowest and highest values. They clearly reveal trends, spread, and outliers in a compact format.



d. Scatter Plots

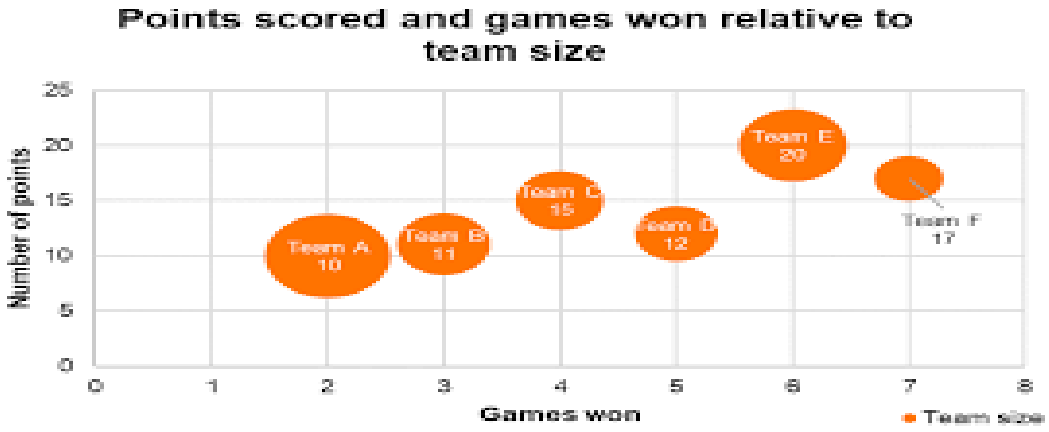
Scatter plots show data points on an **x/y plane**, usually comparing two variables. They help visualize **correlations**, which can be positive, negative, or none. Scatter plots effectively reveal

trends, distributions, and outliers, and work best with multiple data points rather than single values over time.



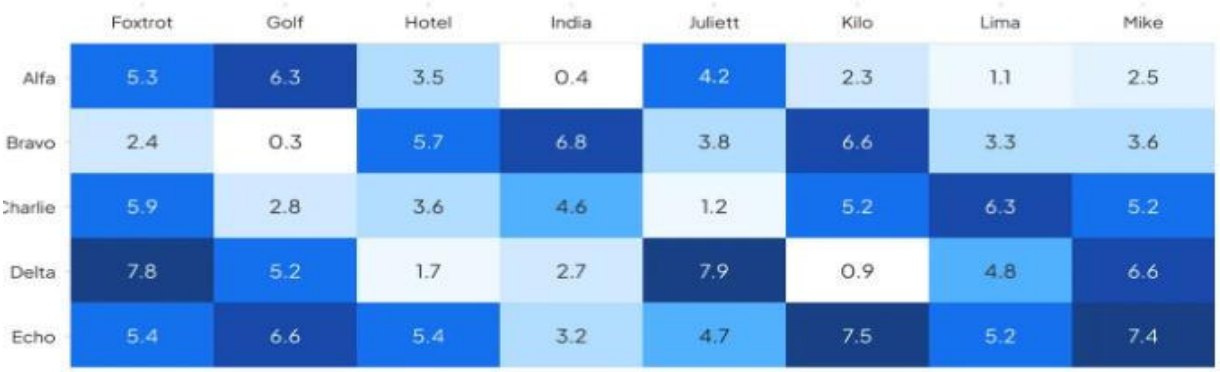
e. Bubble Charts

Bubble charts are like scatter plots but include a **third variable**, represented by **bubble size** (and sometimes color). They plot data on an **x/y plane** while showing additional information, making them useful for analyzing concentrations and complex relationships. They are most effective as a primary visualization when three data variables need comparison.



f. Heatmaps

A heat map visually represents data in which values in a matrix are depicted as colors according to their density. Heat maps make it easy to scan measurements by grouping values into categories and displaying their density through color — the darker the color, the higher the density.



This heat map compares survey results across different criteria (rows) by participants (columns).

2.3 Choosing the Right Chart Type

1. **Understand your audience:**
   Before choosing a chart, it is crucial to know who will view it and what they need to learn. A chart should communicate data clearly, avoiding unnecessary complexity while highlighting the key insights your audience needs. Tailoring visuals to their familiarity and expectations ensures better comprehension and engagement.
2. **Use line graphs or scatter plots for relationships:**
   When the goal is to show how one variable relates to another, line graphs or scatter plots are effective. They clearly illustrate correlations, patterns, or trends between two variables, helping viewers identify positive, negative, or no relationships. Scatter plots are particularly useful for large datasets with many points.
3. **Use line or area graphs for changes over time:**
   Line and area graphs are ideal for tracking changes or trends over specific time periods. They allow easy visualization of increases, decreases, and fluctuations, helping audiences quickly understand patterns, seasonal effects, or growth trends. Marked intervals on both axes enhance precision and clarity.
4. **Use bar charts or pie charts for comparisons:**
   Bar and pie charts are effective when comparing individual values to overall trends or totals. Bar charts are best for side-by-side comparisons, while pie charts show proportional contributions to a whole, such as sales by product type or market share. They simplify understanding distributions and relative sizes.
5. **Match chart type to the data's purpose:**
   Selecting the appropriate chart ensures your visualization communicates insights efficiently. Using the wrong chart type can confuse viewers, obscure trends, or misrepresent the data. Consider the story, data structure, and variables to ensure your chart highlights the intended message clearly.

2.4 Best Practices in Labeling, Coloring, and Scaling

Labels: Labels must be clear and concise, allowing viewers to understand data without external references. They should describe what the values represent, such as "Total Sales Revenue," and include units on each axis. Proper labeling improves comprehension, enables accurate comparisons with other sources, and ensures the chart communicates information effectively.

Coloring, when representing multiple data types, use different colors for clarity but limit to two or three per chart to avoid confusion. Alternatively, use patterns or shades of gray to differentiate categories, especially for a single variable, such as comparing sales revenue across product types.

Scaling: Ensure axes and chart elements are proportionate to accurately represent data. Use consistent scales to avoid misleading interpretations, and adjust the range to focus on relevant values. Proper scaling highlights trends, differences, and patterns while maintaining clarity and preventing distortion of information.

UNIT III

Multidimensional, Temporal and Hierarchical Data Visualization

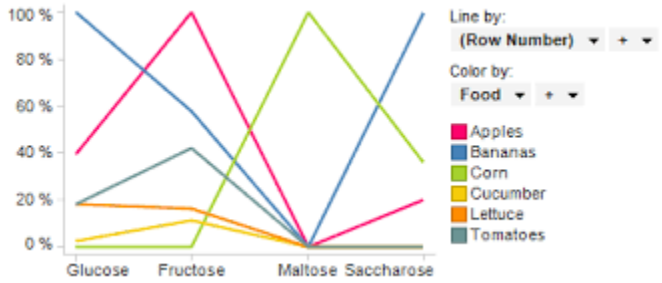
3.1 Visualizing Multivariate Data

Parallel Coordinates: Parallel coordinates are a visualization technique used to display multivariate data. Each variable is represented as a vertical axis, and each data point is plotted as a line that intersects all axes at positions corresponding to its values. This method helps identify patterns, correlations, clusters, and outliers across multiple variables simultaneously. Parallel coordinates are especially useful when comparing many features of a dataset, revealing relationships that may not be visible in standard two-dimensional plots, and supporting exploratory analysis for complex, high-dimensional data.

Example: Identify the food that has consistently low sugar content across all types.

X-axis (horizontal): Represents different types of sugars: Glucose, Fructose, Maltose, Saccharose.

Y-axis (vertical): Represents the percentage content (%) of each sugar in the food.



Observations

1. **Apples (Red line):**
   - High in **Fructose** (~100%)
   - Moderate in Glucose (~40%)
   - Low in Maltose and Saccharose
2. **Bananas (Yellow line):**
   - High in **Maltose** (~100%)
   - Moderate in Glucose
   - Low in Fructose and Saccharose
3. **Corn (Blue line):**
   - Highest in Saccharose (~100%)
   - Lower in other sugars
4. **Cucumber, Lettuce, Tomatoes:**
   - Generally low in all sugar types, but slight variations exist.
   - For example, Cucumbers are slightly higher in Maltose.

Radar Charts: Radar charts (or spider charts) display multivariate data on a circular grid, with each axis representing a different variable. Data points are plotted along each axis and connected to form a polygon. These charts help compare multiple variables for one or more entities, highlighting strengths, weaknesses, patterns, and outliers. They are especially useful for performance analysis, skill assessment, or comparing different items across several attributes, providing a visual overview of multiple dimensions in a compact, easy-to-interpret format.

3.2 Time-Series Visualization

Time Plots: Time plots (or time series plots) are graphs that display data points in chronological order along a horizontal time axis. The vertical axis represents the variable being measured, such as temperature, sales, or power consumption. Time plots are ideal for visualizing trends, patterns, and changes over time, including seasonal variations or long-term growth. They help identify peaks, troughs, and anomalies in the data. By connecting data points with lines, time plots make it easy to track fluctuations and support forecasting or decision-making based on historical trends.

Example:

| | The **horizontal line (X-axis)** shows **years**, from about **1400 to 2000**.

| | The **vertical line (Y-axis)** shows **how many records or events happened** in each time period.

| | The **black curved line** shows the **amount of activity** over time:

- When the line is **low**, there were **few records/events**.
- When the line is **high**, there were **many records/events**.

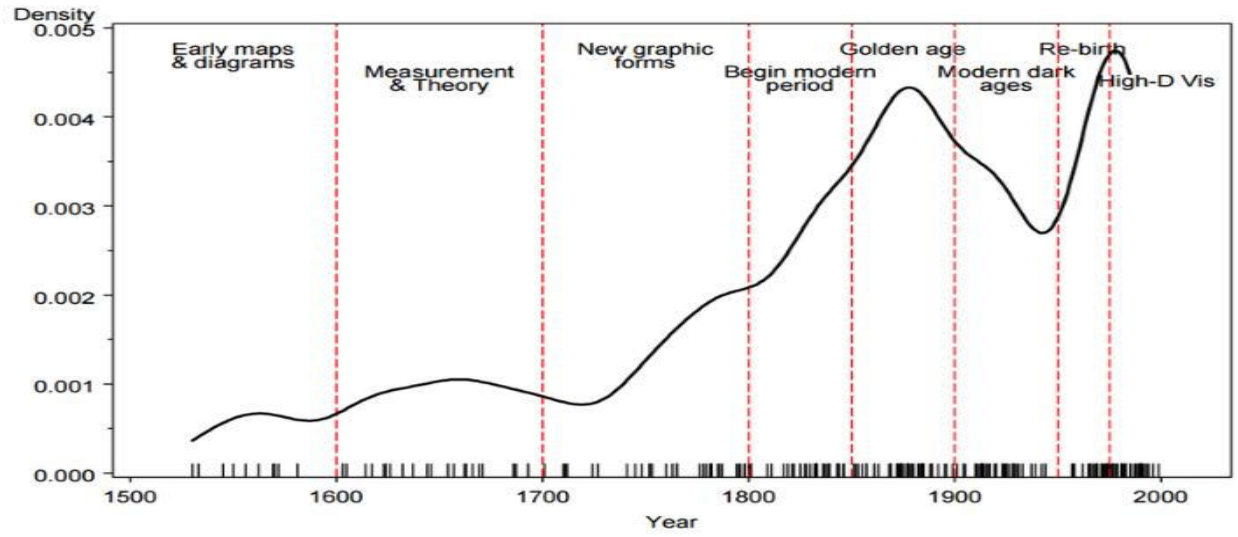
| | In the **early years (before 1600)**, the line is low, meaning **very little information was recorded**.

| | As time moves forward, the line slowly goes up.

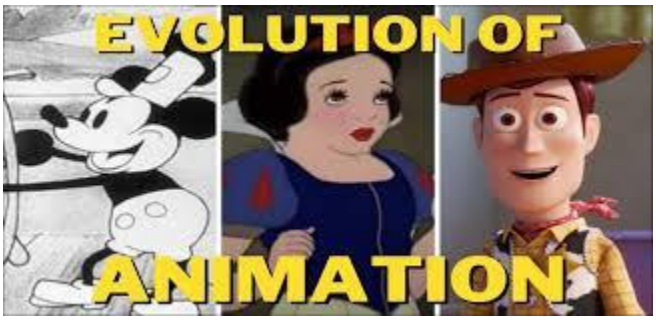
| | After **1800**, the line rises sharply, showing **rapid growth in records and activity**.

| | The **highest point near 2000** means **maximum information and activity in modern times**.

| | The **red dotted lines** divide history into **different periods** like early modern, photographic era, and modern age.



Animation over Time: Animation over time is a visualization technique where data changes are displayed dynamically, showing how values evolve across different timestamps. It helps reveal trends, patterns, and relationships that might be hard to see in static charts. By animating data points, lines, or shapes, viewers can easily observe growth, fluctuations, or movements**,** making complex time-based data more intuitive. This approach is especially useful for tracking progress, simulations, or spatial-temporal changes**,** such as population growth, weather patterns, or energy consumption over time.



The evolution of animation shows how animated visuals have developed over time with the help of technology.

1. **Early Animation**
   Animation began with simple **hand-drawn black-and-white cartoons**. Each frame was drawn manually, and movements were basic.
2. **Traditional 2D Animation**
   Later, animations became **colored and more detailed**. Characters looked smoother, and storytelling improved.
3. **Computer-Assisted Animation**
   Computers were used to help draw, color, and edit animations, making the process faster and more accurate.

4. **3D Animation**
   Modern animation uses **3D computer graphics**. Characters appear realistic with depth, lighting, and smooth movements.
5. **Modern Animation**
   Today, animation includes **AI, motion capture, and virtual reality**, making animations more realistic and interactive.

3.3 Geographical Data Visualization

Maps: Maps are visualizations that display data geographically, linking values to specific locations on a map. They help reveal spatial patterns, trends, and relationships across regions, such as population density, sales distribution, or weather events. Maps can use colors, sizes, or symbols to represent data, making it easier to interpret regional differences and identify hotspots or anomalies. They are widely used in geospatial analysis, urban planning, marketing, and environmental studies**,** providing an intuitive way to combine location and data for meaningful insights.

Choropleths: Choropleth maps are a type of map where regions are shaded or colored based on the values of a variable, such as population, income, or election results. Darker or more intense colors usually represent higher values, while lighter colors indicate lower values. These maps make it easy to visualize spatial patterns, regional differences, and trends across geographic areas. Choropleths are widely used in demographics, economics, public health, and social studies to highlight variations between locations and support decision-making based on regional data comparisons.

Example:

1. This is a choropleth map of India showing population distribution across states.
   Different colors are used to represent different population levels.
2. Darker colors indicate states with higher population.
   Uttar Pradesh, Maharashtra, Bihar, and West Bengal appear most populated.
3. Lighter colors represent states with lower population.
   North-eastern states and smaller regions fall in this category.
4. Southern and western states mostly show medium shades.
   This indicates a moderate level of population compared to other regions.
5. The legend helps in understanding the population range for each color.
   Overall, the map clearly compares population differences between states.

3.4 Hierarchical Data:

Tree Maps: Tree maps visualize hierarchical data using nested rectangles. Each rectangle represents a category, and its size corresponds to a quantitative value, such as sales or population. Colors can show additional variables or categories. Tree maps are useful for comparing proportions within a hierarchy, identifying dominant categories, and understanding part-to-whole relationships in large, complex datasets within a compact visual space.

Example **tree map** that represents **daily food sales** using rectangles of different sizes. The entire chart shows total sales for the day, divided into two main categories: **Breakfast** (blue) and **Lunch** (red).

Each large rectangle represents a category, and the smaller rectangles inside represent **individual food items** such as waffles, eggs, pancakes, tea, coffee for breakfast, and salad, sandwich, soup, pie, iced tea, and coffee for lunch.

The **size of each rectangle** indicates the **sales volume** of that item. Larger blocks mean higher sales. This tree map makes it easy to compare item popularity within and across meal categories at a glance.

Sunburst Charts: Sunburst charts visualize hierarchical data using concentric circular rings. The center represents the top-level category, while outer rings show subcategories and deeper levels. Each segment's size corresponds to a value, such as sales or counts, and colors help distinguish categories. Sunburst charts make it easy to explore part-to-whole relationships and understand how smaller components contribute to the overall structure. They are commonly used in business analytics, file system visualization, and organizational data analysis.

Example **sunburst chart** that represents **hierarchical organizational data** across multiple countries and departments.

- The **center circle** shows the top-level categories, which are **countries** such as *India, USA, UK,* and *Germany*.
- The **first ring** around the center represents **departments** within each country, such as *Tech, HR, Sales, Marketing,* and *Accounts*.
- The **outer rings** show **sub-departments or roles** like *Web, Dev, Test, Analytics, Excel,* and *Windows* under each department.

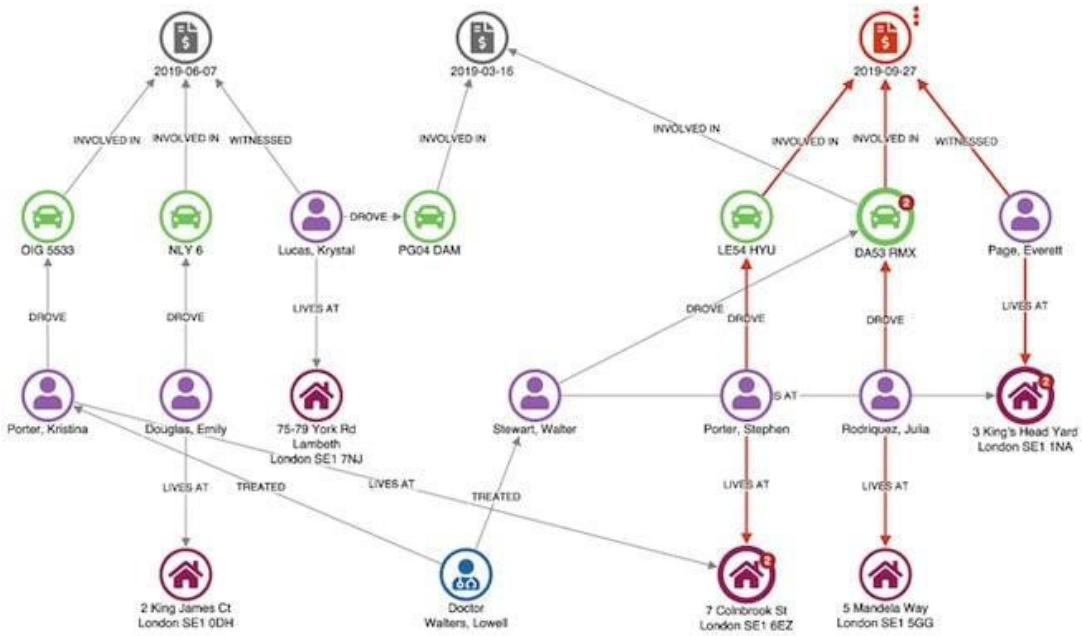
Each segment's **size** indicates the relative proportion or contribution of that category, while **colors** differentiate countries and their subdivisions. This visualization helps quickly understand how workforce or resources are distributed across countries, departments, and roles in a clear part-to-whole hierarchy.

Network and Graph Visualization:

Data for network graphs, referred to as "graph data" is data formatted in node and edge format. Nodes represent discrete things and edges represent the relationships between nodes. Representing data as a **graph** not only makes relationships easy to understand visually but also enables deeper analytical insights. One important concept in graph theory is **centrality**, which helps identify the importance of nodes within a network.

- **Degree centrality** measures how many direct connections a node has. Nodes with higher degree centrality interact more with others and are often highly active or popular.
- **Eigenvector centrality** considers both the number and quality of connections. A node connected to other influential nodes gains higher importance, even if it has fewer direct links.
- **Betweenness centrality** measures how often a node lies on the shortest path between other nodes. Such nodes act as **bridges**, controlling information flow and connecting different clusters in the network.

Example:



Main Elements in the Diagram

People (purple icons):

Represent individuals involved in events (drivers, witnesses, doctors, residents).

Vehicles (green car icons):
Show cars used or driven by people.
Locations (house icons):
Indicate places where people live or events occurred.
Events / Dates (money or case icons with dates):
Represent incidents that happened on specific dates.
Connections (arrows):
Show relationships such as *drove*, *lives at*, *involved in*, *witnessed*, and *treated*.

UNIT IV

Data Visualization Using Python and Dashboards

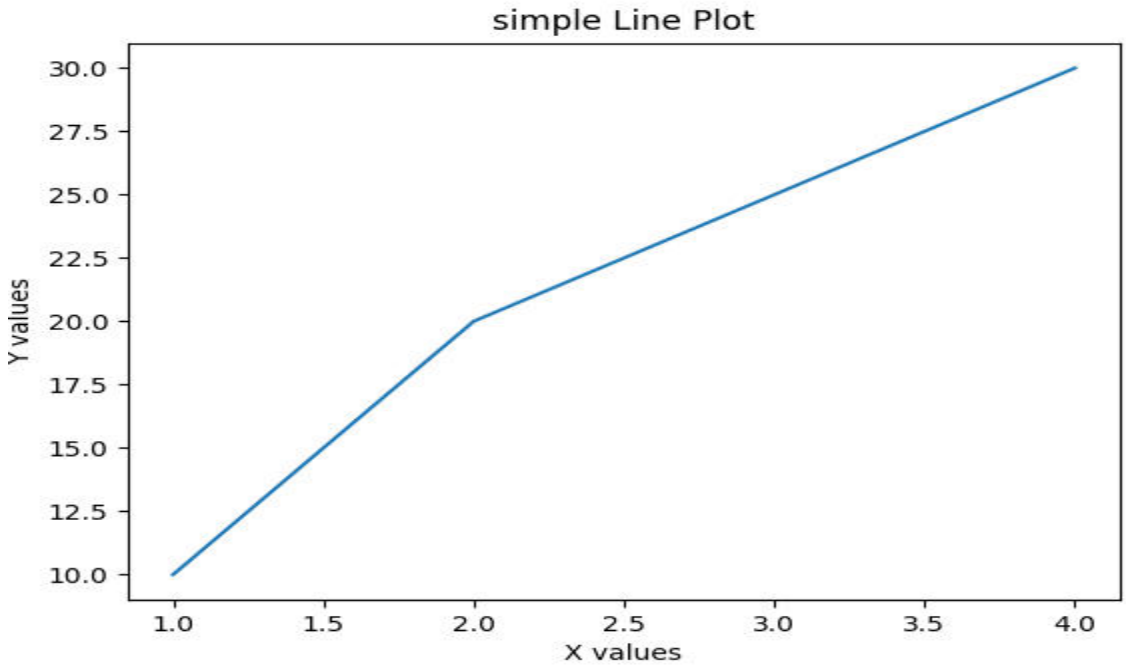
4.1 Introduction to Matplotlib, Seaborn and Plotly:

a. Matplotlib

Matplotlib is the core Python library for data visualization, mainly used for creating static plots. It gives full control over every element of a graph such as axes, labels, titles, legends, and styles. Matplotlib is ideal for understanding trends, comparisons, and distributions in data. It is widely used in academics, research, and basic data analysis. Although visuals are simple by default, it is very powerful for customized plotting and forms the foundation for many other visualization libraries.

Example (Line Plot):

```
import matplotlib.pyplot as plt
x = [1, 2, 3, 4]
y = [10, 20, 25, 30]
plt.plot(x, y)
plt.xlabel("X values")
plt.ylabel("Y values")
plt.title("Simple Line Plot")
plt.show()
```



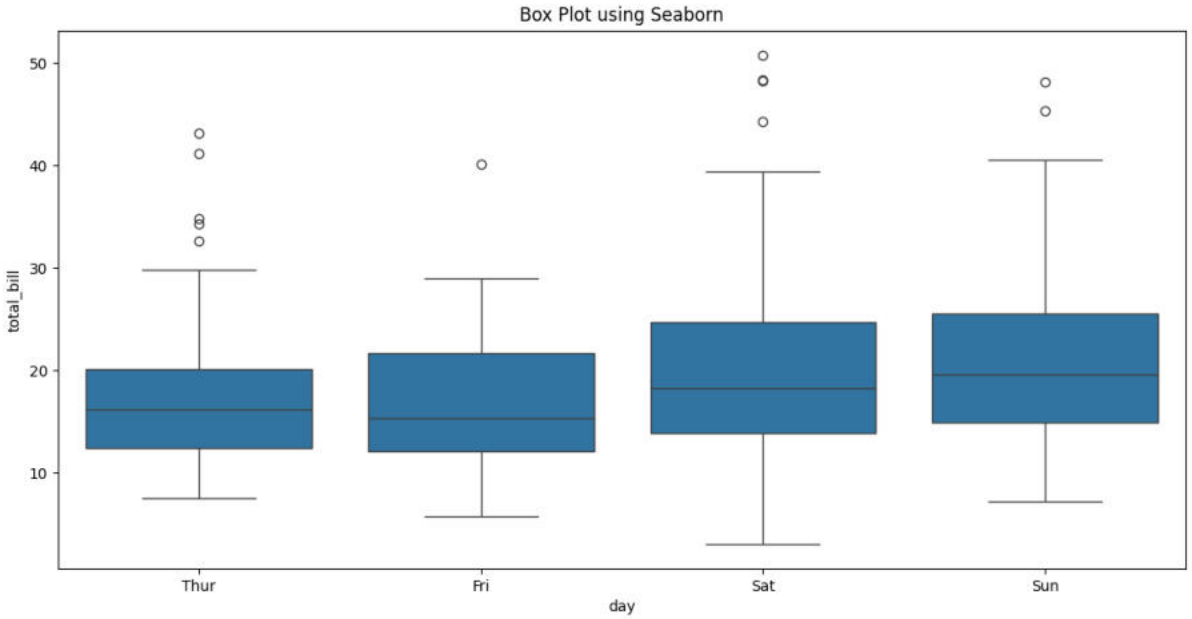
b. Seaborn

Seaborn is a high-level statistical visualization library built on top of Matplotlib. It makes graphs more attractive and informative with less code. Seaborn works directly with Pandas Data Frames and provides built-in themes and color palettes. It is mainly used for exploratory data analysis (EDA), helping analysts understand data distribution, relationships, and correlations. Seaborn supports advanced plots such as box plots, violin plots, heatmaps, and regression plots.

Example (Box Plot):

```
import seaborn as sns
import matplotlib.pyplot as plt

tips = sns.load_dataset("tips")
sns.boxplot(x="day", y="total_bill", data=tips)
plt.title("Box Plot using Seaborn")
plt.show()
```



c. Plotly

Plotly is an **interactive visualization library** used to create dynamic, web-based charts. Unlike Matplotlib and Seaborn, Plotly allows users to **zoom, hover, pan, and interact** with data. It supports 2D, 3D, animated charts, and dashboards, making it popular in **business analytics and real-time applications**. Plotly is ideal when user interaction and presentation-quality visuals are required.

Example (Interactive Scatter Plot):

```
import plotly.express as px

df = px.data.iris()
fig = px.scatter(df, x="sepal_width", y="sepal_length",
                 color="species", title="Interactive Scatter Plot")
fig.show()
```

4.2 Creating Static and Interactive Charts:

a. Static Charts: Static charts are visual representations of data that do not change or respond to user interactions. They are fixed images that display patterns, trends, or comparisons clearly and are commonly used in reports, presentations, and publications. Static charts include bar charts, line graphs, pie charts, and histograms. They are simple to create using libraries like **Matplotlib** or **Seaborn** in Python. The primary advantage is clarity: they provide a quick, straightforward view of the data. However, they cannot display dynamic insights or allow users to explore the data interactively, making them best for summarizing information.

b. Interactive Charts: Interactive charts are dynamic visualizations that allow users to engage with data by hovering, zooming, filtering, or clicking to reveal more details. Unlike static charts, they provide a richer understanding of patterns, trends, and relationships within datasets. Libraries like **Plotly**, **Bokeh**, and **Dash** in Python are commonly used to create interactive charts such as interactive line graphs, scatter plots, bar charts, and heatmaps. These charts are ideal for dashboards and web applications because they enable exploration of large datasets. Interactivity enhances decision-making by letting users focus on specific data points or time periods.

4.3 Pandas Visualization Capabilities:

Pandas, a powerful Python library for data analysis, provides built-in visualization capabilities through its integration with **Matplotlib**. Using the .plot() method, Pandas can create various charts directly from **DataFrames** or **Series**, such as line plots, bar charts, histograms, area charts, scatter plots, and box plots. This simplifies the process of exploring data and identifying trends,

distributions, and outliers without switching between multiple libraries. Pandas visualizations are customizable, allowing control over titles, labels, colors, and figure size. While basic, these visualizations are excellent for quick insights and exploratory data analysis before using more advanced plotting libraries like Seaborn or Plotly.

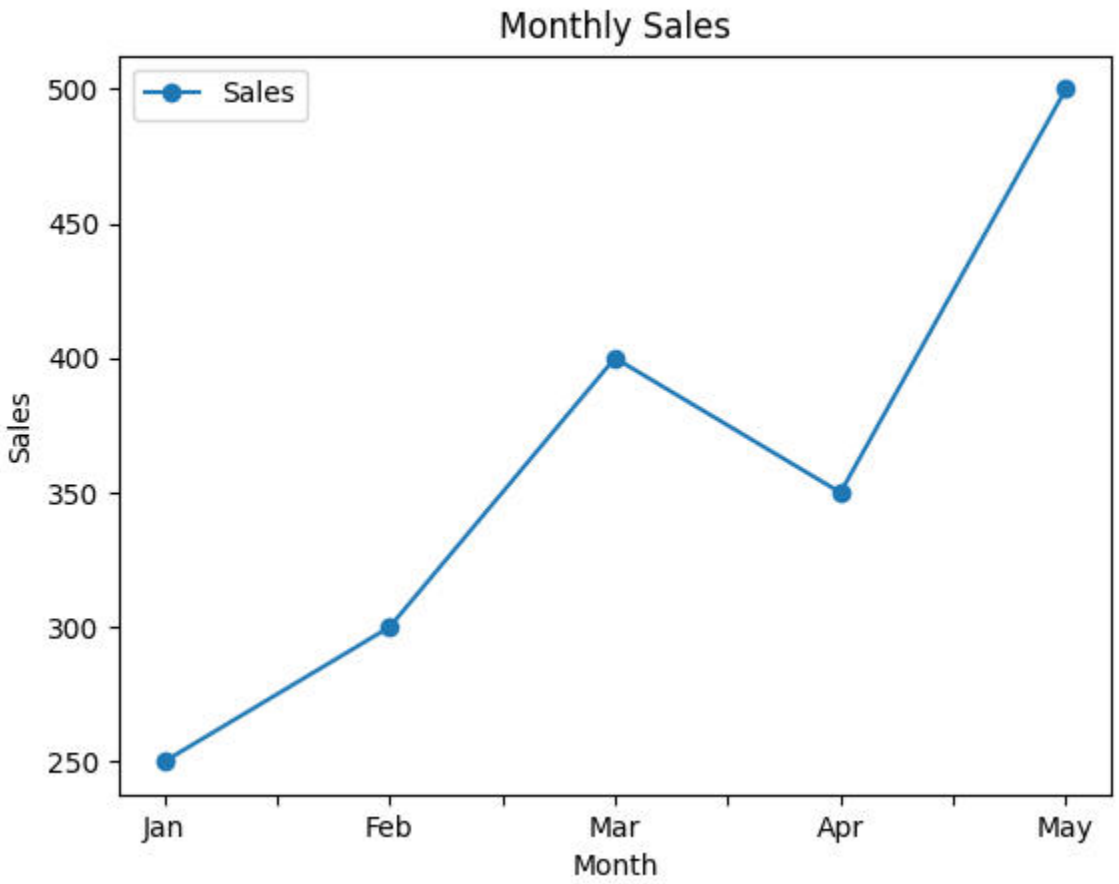
```
import pandas as pd
import matplotlib.pyplot as plt

# Sample data
data = {
   'Month': ['Jan', 'Feb', 'Mar', 'Apr', 'May'],
   'Sales': [250, 300, 400, 350, 500],
   'Profit': [50, 70, 90, 60, 100]
}

df = pd.DataFrame(data)

# Line plot
df.plot(x='Month', y='Sales', kind='line', title='Monthly Sales', marker='o')
plt.ylabel('Sales')
plt.show()

# Bar plot
df.plot(x='Month', y='Profit', kind='bar', color='green', title='Monthly Profit')
plt.ylabel('Profit')
plt.show()
```



4.4 Dashboards with Dash, Streamlit, Power BI

Dashboards are interactive platforms that consolidate data visualizations, metrics, and key performance indicators (KPIs) in a single view for analysis and decision-making. Tools like **Dash**, **Streamlit**, and **Power BI** make building dashboards efficient:

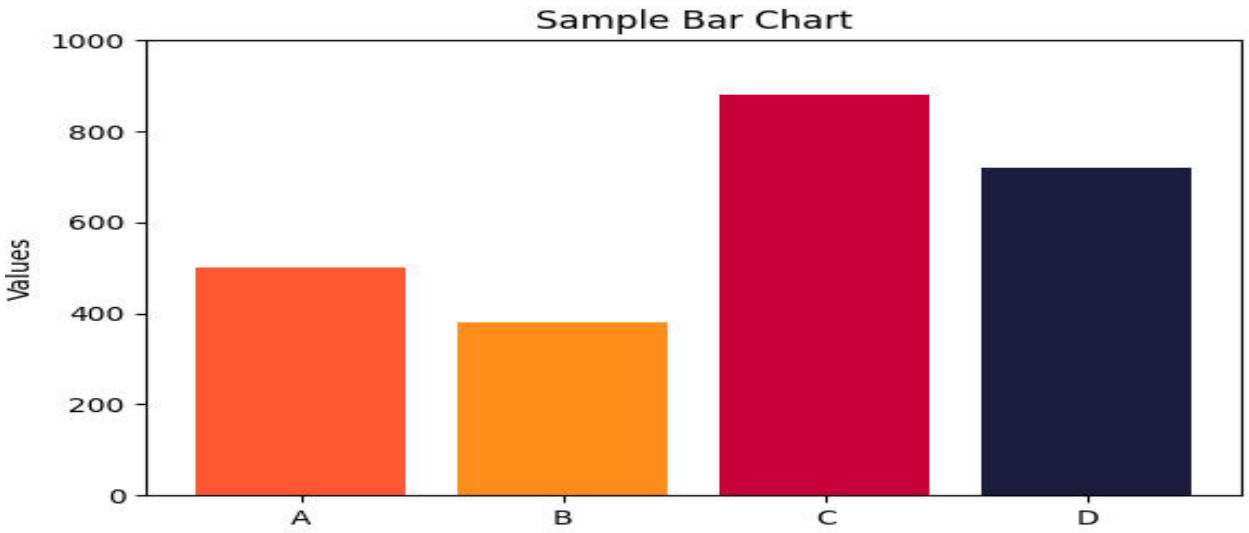
- **Dash**: A Python framework for creating web-based analytical dashboards. It integrates **Plotly** for interactive charts and supports callbacks to make dynamic, responsive dashboards. Ideal for developers with Python knowledge.
- **Streamlit**: Simplifies dashboard creation with minimal code. You can turn Python scripts into interactive apps with widgets like sliders, dropdowns, and buttons. Great for quick prototyping and sharing ML or data insights.
- **Power BI**: A Microsoft tool for enterprise-grade dashboards. It allows importing data from multiple sources, creating interactive visuals, and sharing insights across organizations. It's user-friendly and doesn't require programming.

Together, these tools enable effective **data storytelling** and real-time monitoring of metrics for businesses or projects.

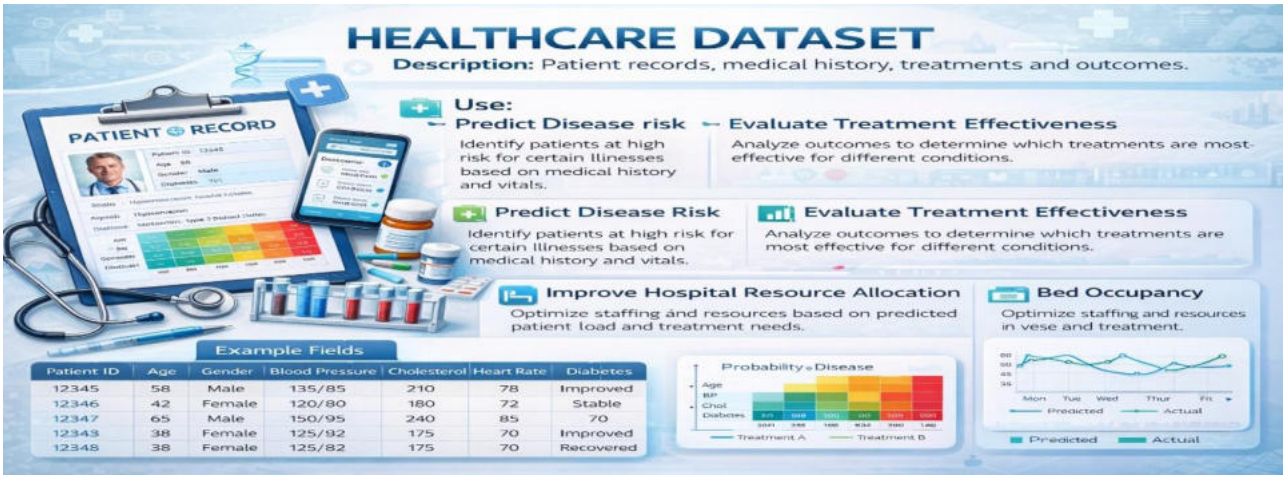
4.5 Case Studies: Real-world Dataset

Real-world datasets are used in case studies to demonstrate practical applications of data analysis, visualization, and predictive modeling. They provide context for understanding how data-driven insights influence decisions in various domains. Here are a few examples:

1. **Sales and Marketing Dataset**
   - **Description:** Contains monthly sales, customer demographics, and product categories.
   - **Use:** Analyze trends, identify high-performing products, forecast future sales, and design targeted marketing campaigns.



2. **Healthcare Dataset**
   - **Description:** Patient records, medical history, treatments, and outcomes.
   - **Use:** Predict disease risk, evaluate treatment effectiveness, and improve hospital resource allocation.



3. **Financial Dataset**
    - **Description:** Stock prices, trading volumes, and economic indicators.
    - **Use:** Build predictive models for stock price movements, risk assessment, and portfolio optimization.
4. **E-commerce Dataset**
    - **Description:** Customer behavior, purchase history, and website activity.
    - **Use:** Perform recommendation system analysis, customer segmentation, and sales funnel optimization.

UNIT V

Storytelling with Data and Ethical Visualization

5.1 Storytelling and Narrative Techniques in Visualization:

Storytelling in data visualization is the practice of **using data to convey a clear, compelling narrative**. Rather than just presenting raw numbers or charts, storytelling guides the audience through the insights, helping them **understand the significance, context, and implications** of the data.

Key Techniques:

1. **Context Setting:**
    - Introduce the background or problem before showing the data.
    - Example: Showing hospital readmission rates without context is confusing; explain why readmission matters first.
2. **Sequence / Flow:**
    - Organize visuals to **follow a logical path**. Start with overview, then drill down into specifics.
    - Example: Begin with overall disease prevalence, then break down by age groups or regions.
3. **Highlighting Key Insights:**

- Use **colors, annotations, or callouts** to emphasize important data points.
- Example: Highlight a spike in diabetes cases in a certain region to draw attention.

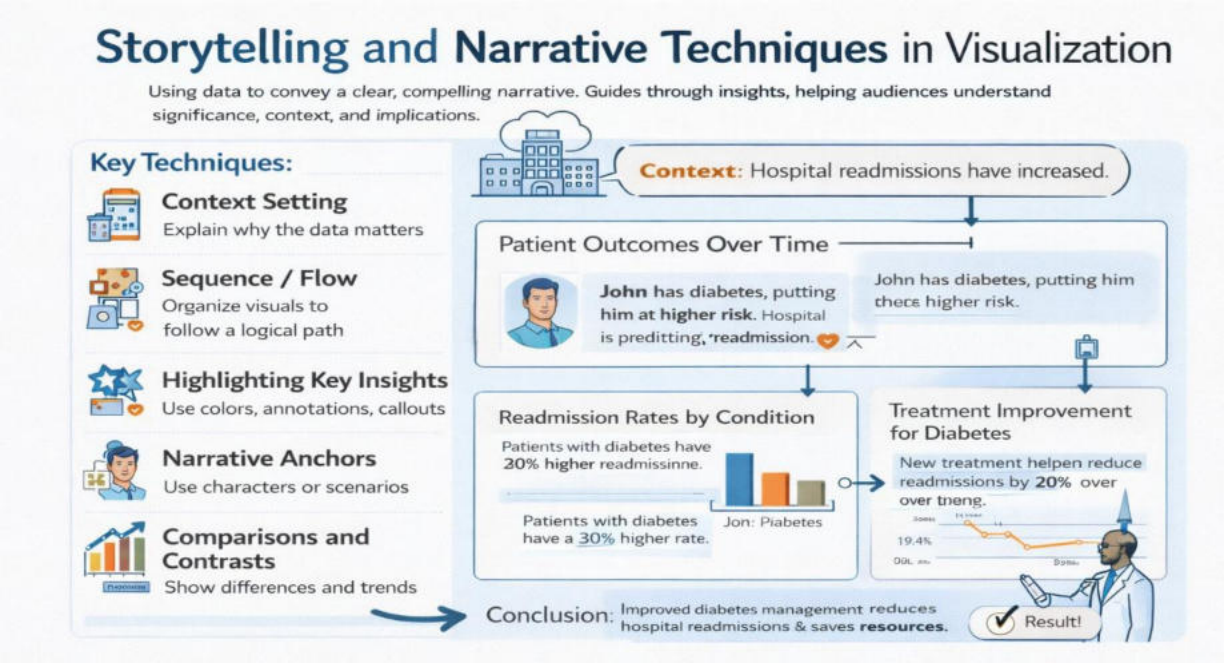
4. **Narrative Anchors:**
   - Use **characters, personas, or scenarios** to make data relatable.
   - Example: "John, a 55-year-old patient, represents the typical risk profile for heart disease."
5. **Comparisons and Contrasts:**
   - Show differences or trends clearly to make the story impactful.
   - Example: Compare treatment outcomes over two years to demonstrate improvement.
6. **Interactivity (Optional):**
   - In digital dashboards, allow users to explore the story themselves through filters and clickable charts.
   - Example: Interactive healthcare dashboard showing outcomes by hospital department or patient age.

Benefits:

- Makes complex data **accessible and memorable**.
- Engages audiences emotionally and intellectually.
- Helps drive **decision-making** and **insight adoption**.

Example in Healthcare:

- Instead of just showing a bar chart of treatment outcomes, a story could follow **a patient cohort**, showing how treatments affected their recovery over time, highlighting trends and anomalies, and explaining the **why behind the numbers**.



5.2 Dashboards and Reporting

- **Dashboards** are **interactive visual interfaces** that display key metrics, KPIs, and trends in real-time or near real-time. They are designed to provide a quick overview and support **decision-making**.
- **Reporting** involves creating **structured summaries** of data, often in static formats, to analyze trends, track performance, and communicate insights periodically.

Key Features of Dashboards:

1. **Interactivity:** Users can filter, drill down, or explore data dynamically.
2. **Real-Time Insights:** Metrics and visuals update automatically for timely decisions.
3. **KPI Focused:** Highlights critical metrics like patient recovery rate, bed occupancy, or disease incidence.
4. **Customizable Views:** Tailored for different users (managers, clinicians, analysts).
5. **Integration:** Pulls data from multiple sources (databases, EHRs, spreadsheets, APIs).

Types of Dashboards:

- **Operational Dashboards:** Track day-to-day operations (e.g., ER admissions, ICU occupancy).
- **Analytical Dashboards:** Support deeper exploration and trend analysis (e.g., treatment outcomes over time).
- **Strategic Dashboards:** Provide high-level insights for executives (e.g., hospital performance scorecards).

Reporting:

- **Purpose:** Summarizes performance, monitors trends, and supports audits or compliance.
- **Formats:** PDF, Excel, PowerPoint, or dashboard exports.
- **Frequency:** Daily, weekly, monthly, or ad-hoc reports.
- **Benefits:** Ensures **consistency**, improves **decision-making**, and communicates results effectively.

Example:
A **corporate finance dashboard** could include:

- A **line chart** showing monthly revenue vs expenses.
- A **bar chart** comparing budgeted vs actual spending by department.
- A **pie chart** showing expense distribution (salaries, rent, marketing, utilities).
- **KPIs** for cash flow, net profit margin, and outstanding invoices.
- Filters for selecting **quarter, department, or region**.

2. Reporting in Finance:
Financial reporting summarizes data for decision-making, compliance, and analysis.

Types of Reports:

- **Income Statement (Profit & Loss):** Shows revenue, expenses, and net profit over a period.
- **Balance Sheet:** Displays assets, liabilities, and equity at a given point.
- **Cash Flow Report:** Shows inflow and outflow of cash for operations, investing, and financing.
- **Budget Reports:** Compare planned vs actual spending.
- **Investment Reports:** Track performance of stocks, bonds, or portfolios.

5.3 Misleading Visualizations and Bias:

Misleading visualizations occur when charts, graphs, or dashboards **distort the data or its interpretation**, intentionally or unintentionally, leading viewers to incorrect conclusions. Bias in visualization arises when the **design, selection, or presentation of data favors certain outcomes** or perspectives.

Common Types of Misleading Visualizations:

1. **Truncated Axes:**
   - Starting a Y-axis at a value other than zero can exaggerate differences.
   - Example: A bar chart showing revenue growth starting at $90k instead of $0 makes small changes look huge.
2. **Inappropriate Chart Types:**
   - Using the wrong chart type can confuse interpretation.
   - Example: Using a 3D pie chart to show percentages may distort the perception of slice sizes.
3. **Cherry-Picked Data:**
   - Selecting only data that supports a conclusion while ignoring contradictory data.
   - Example: Showing only high-performing sales regions to claim overall company growth.

4. **Improper Scaling:**
    - Unequal intervals on axes can exaggerate trends.
    - Example: A line chart with irregular time intervals can make trends look sharper than they are.
5. **Overcomplicated Visuals:**
    - Excessive colors, labels, or clutter can obscure insights.
    - Example: A heatmap with too many categories and colors, making patterns hard to detect.

Bias in Visualizations:

- **Selection Bias:** Only certain data points are visualized.
- **Presentation Bias:** Color, size, or placement influences perception.
- **Confirmation Bias:** Visualizations are designed to confirm pre-existing beliefs.

How to Avoid Misleading Visualizations:

1. Use appropriate chart types for your data.
2. Start axes at zero unless a valid reason exists, and clearly indicate breaks.
3. Show full context, not just selective data.
4. Keep designs clean and avoid unnecessary embellishments.
5. Label data clearly, including units and sources.

5.4 Ethical Principles in Data Visualization:

Ethical data visualization ensures that charts, graphs, dashboards, and reports **accurately represent data** and convey insights **honestly**, without misleading or manipulating the audience. Ethics in visualization is about **responsibility, transparency, and clarity**.

Key Principles:

1. **Accuracy:**
    - Ensure that visualizations correctly represent the data. Avoid misleading scales, truncated axes, or selective data.
2. **Clarity:**
    - Make visuals easy to understand. Use appropriate chart types, labels, legends, and color schemes. Avoid clutter and unnecessary complexity.
3. **Transparency:**
    - Clearly communicate sources, methodology, and any assumptions made. Provide context to avoid misinterpretation.
4. **Objectivity:**
    - Present data without personal or organizational bias. Avoid cherry-picking data or emphasizing misleading trends.

5. **Accessibility:**
    - Make visualizations understandable for diverse audiences, including color-blind or visually impaired individuals.
6. **Respect for Privacy:**
    - Do not reveal sensitive or personally identifiable information without consent. Mask or anonymize data when necessary.

Why Ethics Matters:

- Builds **trust** in data and analysis.
- Prevents **misinterpretation** or manipulation of insights.
- Supports **informed decision-making** in business, healthcare, finance, and research.

5.5 Final Project: Create a Storytelling Dashboard with Real Data

Design an **interactive dashboard** that tells a clear story from a real dataset. The dashboard should **engage users**, highlight **key insights**, and allow exploration of the data.

Steps to Complete the Project:

1. Choose a Dataset

- Pick a real dataset relevant to your interest or domain:
    - **Healthcare:** Patient outcomes, treatment effectiveness, hospital resources
    - **Finance:** Revenue, expenses, investment portfolio performance
    - **Education:** Student performance, attendance, grades
    - **Environment:** Pollution levels, climate data, renewable energy trends
- Ensure your dataset is **clean, complete, and accessible**.

2. Define Your Story

- Decide what **insight or narrative** your dashboard should communicate.
- Example:
    - Healthcare: “How patient age and lifestyle affect disease risk”
    - Finance: “Budget vs Actual spending trends across departments”
    - Education: “Student performance trends across subjects and semesters”

3. Choose Key Metrics & KPIs

- Identify **critical metrics** that support your story.
- Examples:
    - **Healthcare:** Average recovery rate, readmission rate, treatment success rate
    - **Finance:** Revenue, net profit, cash flow, expense breakdown
    - **Education:** Average grades, pass rate, attendance percentage

4. Select Visualizations

- Use **appropriate charts** to tell the story:
  - **Line charts:** Trends over time
  - **Bar/Column charts:** Comparisons between categories
  - **Pie/Donut charts:** Proportions
  - **Heatmaps:** Correlations or intensity patterns
  - **Maps:** Geographical distributions

5. Add Interactivity

- Include **filters, drop-downs, and drill-downs** to allow user exploration:
  - Time periods (monthly, quarterly, yearly)
  - Categories (departments, regions, patient groups)
  - KPIs selection (profit, revenue, recovery rate)

6. Apply Storytelling Principles

- Provide **context, annotations, and highlights** to guide the audience.
- Use **sequence and flow**:
  1. Overview of key insights
  2. Detailed exploration of data
  3. Summary and recommendations

7. Ensure Ethical Visualization

- Avoid **misleading scales, cherry-picked data, or biased visuals**.
- Label axes, units, and sources clearly.
- Protect sensitive information if using real patient or financial data.

8. Tools You Can Use

- **Power BI** – Interactive dashboards, easy storytelling
- **Tableau** – Advanced visualization and analytics
- **Plotly / Dash** (Python) – Customizable dashboards with interactivity
- **Excel / Google Sheets** – Simpler dashboards for small datasets

9. Deliverables

- **Interactive Dashboard:** With filters, charts, and highlights
- **Story Summary:** A brief explanation of the narrative and insights
- **Data Source & Documentation:** Dataset description, preprocessing steps

10. Example Project Idea

Domain: Healthcare
Story: "Impact of Age and Blood Pressure on Heart Disease Risk"
Dashboard Elements:

- Line chart: Average blood pressure over age groups
- Bar chart: Number of patients with/without heart disease by age
- KPI cards: % of high-risk patients, average recovery rate
- Filter: Gender, age range, hospital location