

# **UNIT-III**

## **Basics of Data Visualization**

# Computational Statistics:

- Computational statistics (is the bond between statistics and computer science) refers to the use of computational techniques and algorithms to analyse, model, and interpret data.
- The term 'Computational statistics' may also be used to refer to computationally *intensive* statistical methods including resampling methods, Markov chain Monte Carlo methods, local regression, kernel density estimation, artificial neural networks and generalized additive models.

# Some key aspects of computational statistics include:

- **Data Preparation:** Before any analysis can take place, data often need to be cleaned, transformed, and structured for analysis. This may involve handling missing data, outliers, and data normalization.
- **Statistical Models:** Computational statistics involves the use of statistical models to describe relationships and patterns in data. Common models include linear regression, logistic regression, decision trees, and more complex models like neural networks.
- **Hypothesis Testing:** Hypothesis testing is used to make inferences about populations based on sample data. Computational statistics can automate this process, making it more efficient and robust.

- **Simulation:** Monte Carlo simulations and other computational techniques are used to estimate probabilities and study complex systems, especially when analytical solutions are difficult to obtain.
- **Resampling Methods:** Bootstrap and permutation tests are examples of resampling methods used for estimating sampling distributions and making statistical inferences.
- **Statistical Software:** Computational statistics often relies on specialized software like R, Python (with libraries like NumPy, SciPy, pandas, and scikit-learn), and statistical packages such as SPSS(Statistical Package for Social Science) and SAS(Statistical Analysis System).

# **Data Visualization:**

- Data visualization is the graphical representation of data to help users understand patterns, relationships, and trends.
- Effective data visualization can simplify complex data and make it more accessible.
- Here are some common types of data visualization techniques:

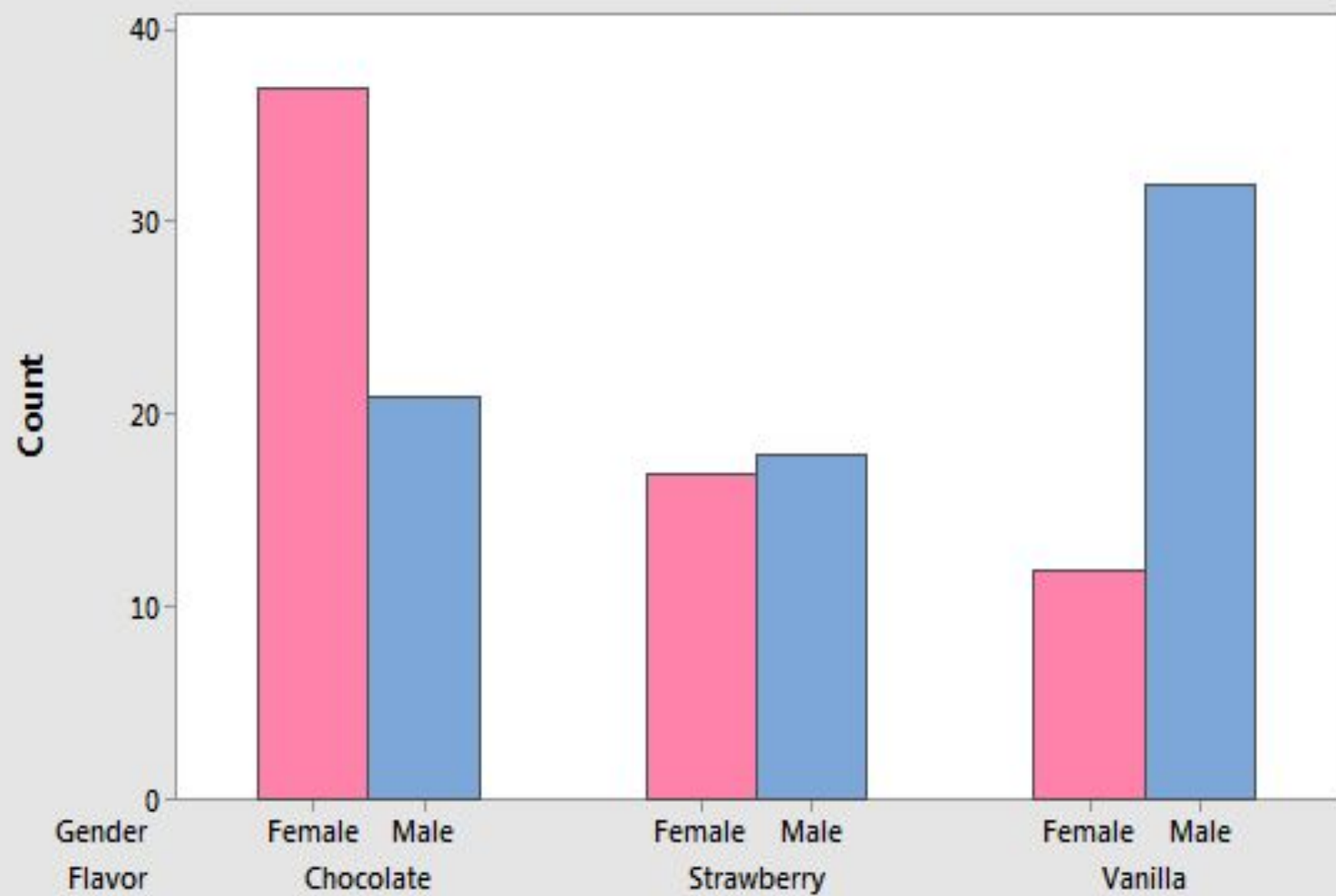
# Types of Data Visualization:

- Certainly, in the domain of computational statistics and data visualization, a wide array of techniques and tools are employed to represent, analyze, and communicate data.
- Here are some common types of data visualizations used in computational statistics:

# Bar Charts:

- Bar charts are used to represent categorical data.
- They display data as bars of varying lengths or heights, with each bar representing a category and the length or height representing the value of that category.
- They are great for comparing different categories.
- A bar chart plots numeric values for levels of a categorical feature as bars.
- Levels are plotted on one chart axis, and values are plotted on the other axis.
- Each categorical value claims one bar, and the length of each bar corresponds to the bar's value.
- Bars are plotted on a common baseline to allow for easy comparison of values.
- <https://www.splashlearn.com/math-vocabulary/geometry/bar-graph>

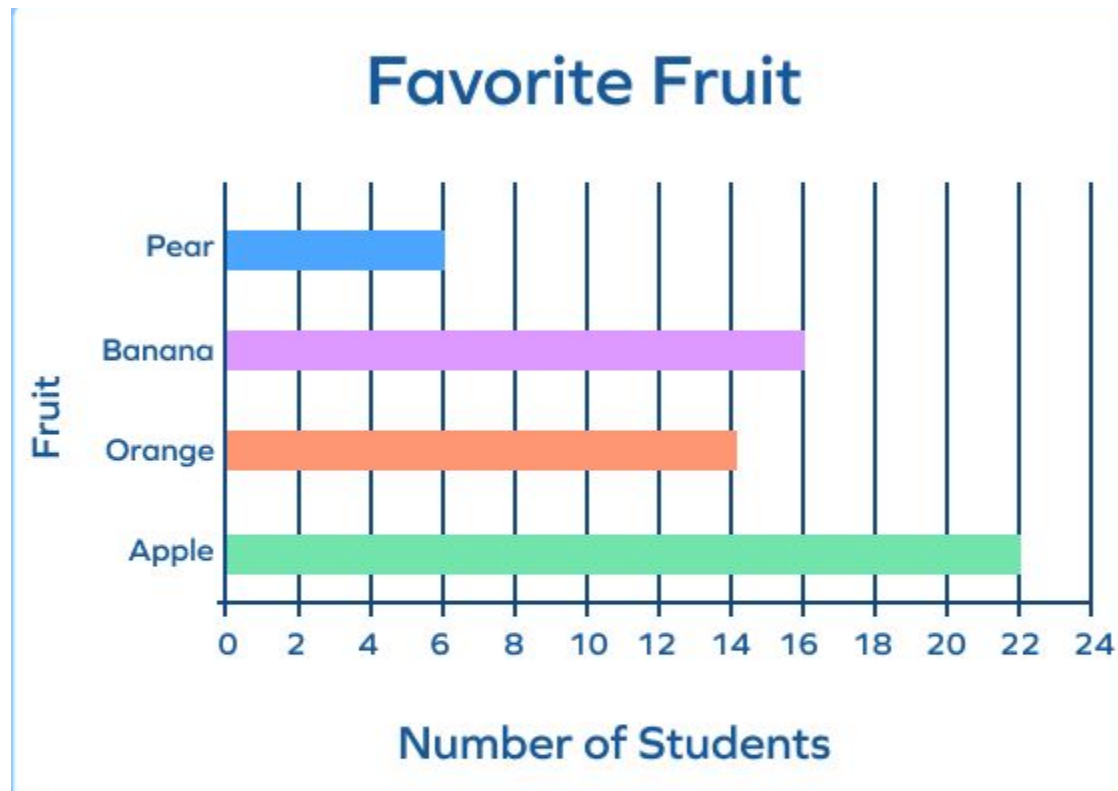
# Flavor Preferences by Gender





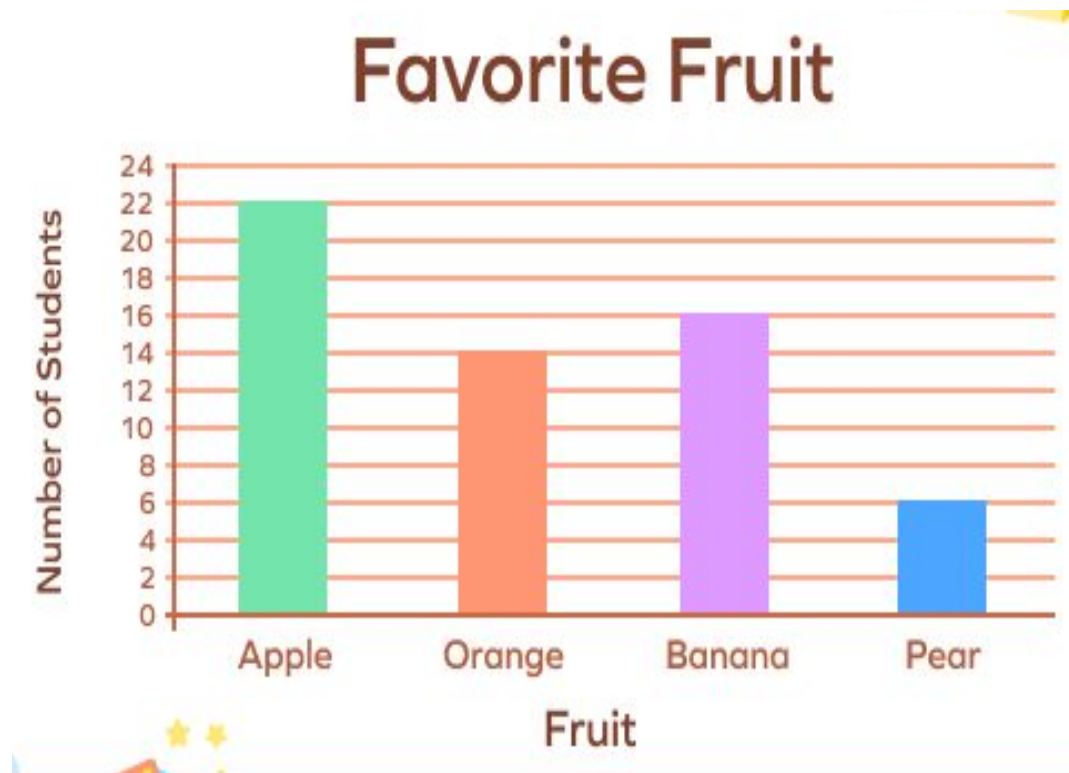
# Types of Bar Graph

- **Horizontal:** Here, the bars are drawn horizontally from left to right.
- The data categories are placed on the vertical axis and numerical values are placed on the graph's horizontal axis.



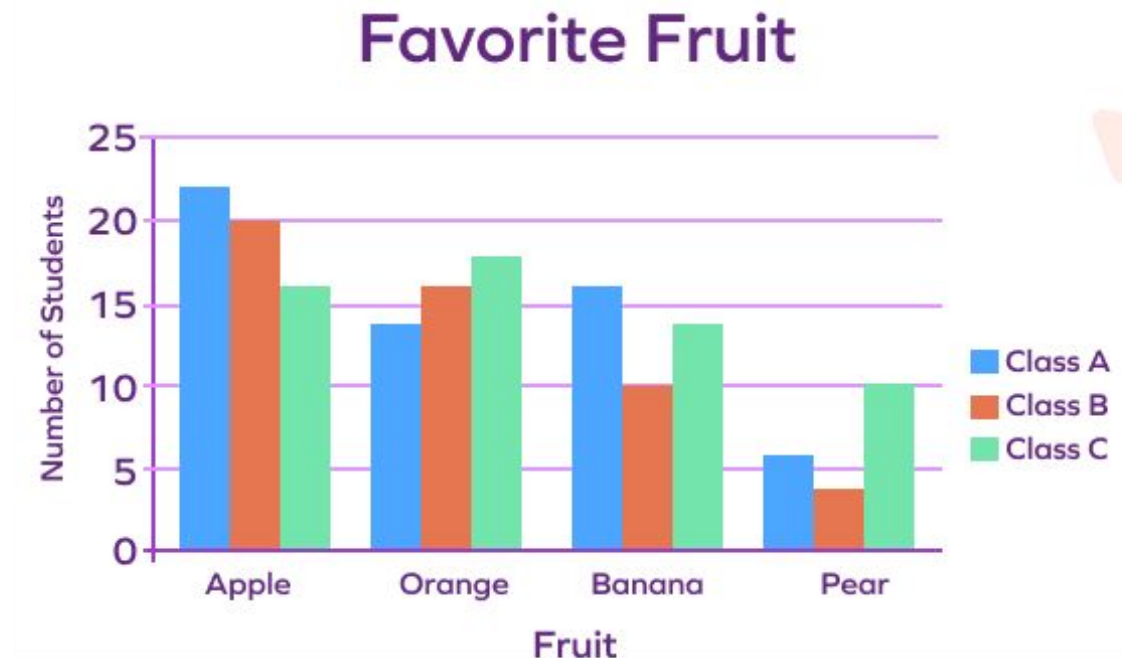
**Vertical:** Here, the bars are drawn vertically from down to top.

- The data categories are placed on the horizontal axis, and the numerical values are placed on the graph's vertical axis.



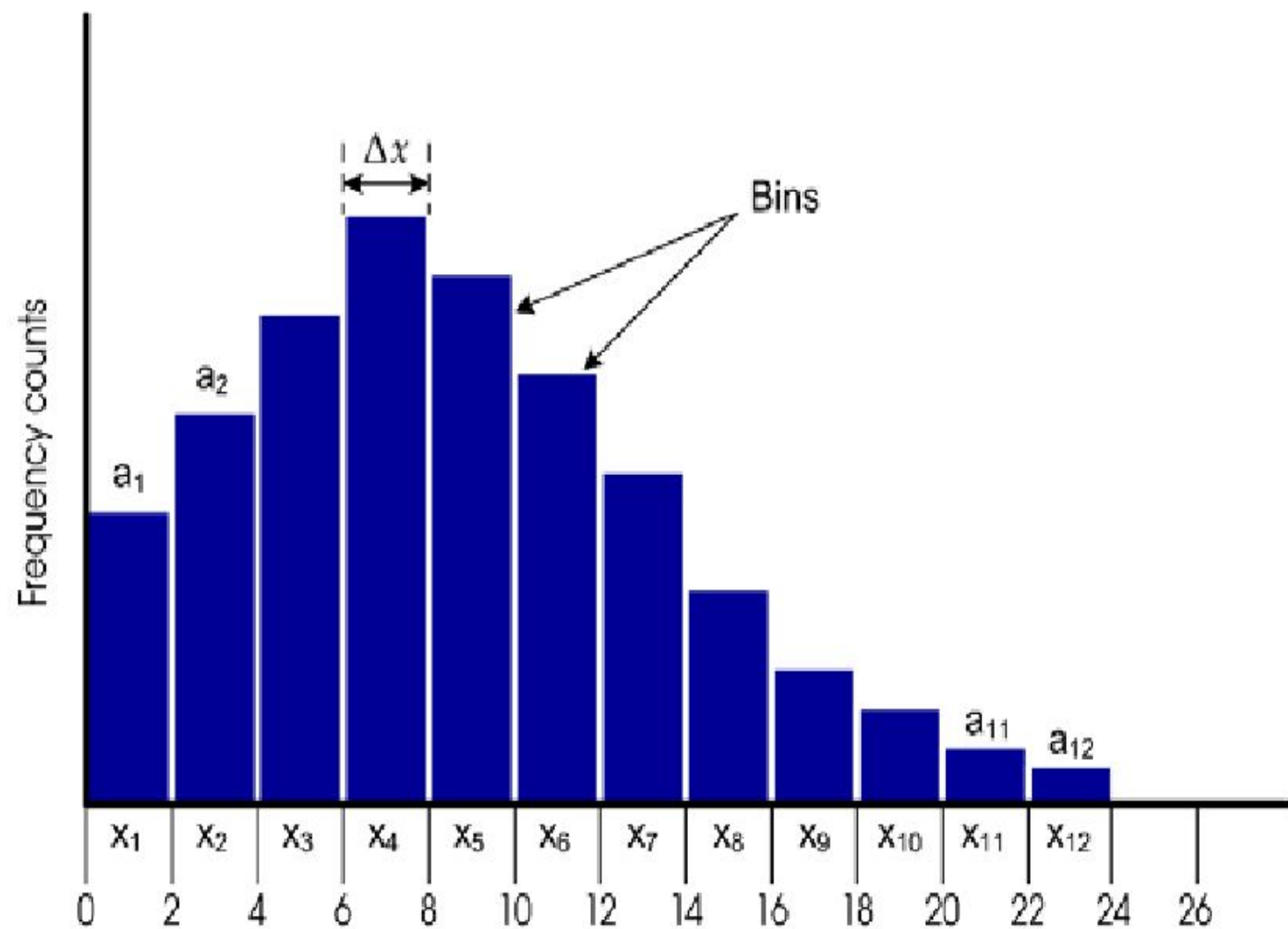
## Grouped:

- This graph represents related sets of data.
- Each set of data is graphed separately but on the same graph.
- The key explains which set of data is shown by the graph.



# Histograms:

- Histograms are used to represent the distribution of a single variable.
- They divide the data into intervals (bins) and display the frequency or probability of data points falling into each bin as bars.
- Histogram is a graphical representation of data points organized into user-specified ranges.
- **Histogram**: a graphical display of data using bars of different heights.
- It is similar to a Bar Chart, but a **histogram** groups numbers into ranges .



- A **histogram** is a graphical representation of a grouped frequency distribution with continuous classes.
- It is an area diagram and can be defined as a set of rectangles with bases along with the intervals between class boundaries and with areas proportional to frequencies in the corresponding classes.
- In such representations, all the rectangles are adjacent since the base covers the intervals between class boundaries.
- The heights of rectangles are proportional to corresponding frequencies of similar classes and for different classes, the heights will be proportional to corresponding frequency densities.
- In other words, a histogram is a diagram involving rectangles whose area is proportional to the frequency of a variable and width is equal to the class interval.

# How to Plot Histogram?

- You need to follow the below steps to construct a histogram.
1. Begin by marking the class intervals on the X-axis and frequencies on the Y-axis.
  2. The scales for both the axes have to be the same.
  3. Class intervals need to be exclusive.
  4. Draw rectangles with bases as class intervals and corresponding frequencies as heights.
  5. A rectangle is built on each class interval since the class limits are marked on the horizontal axis, and the frequencies are indicated on the vertical axis.
  6. The height of each rectangle is proportional to the corresponding class frequency if the intervals are equal.
  7. The area of every individual rectangle is proportional to the corresponding class frequency if the intervals are unequal.

# When to Use Histogram?

- The histogram graph is used under certain conditions. They are:
  1. The data should be numerical.
  2. A histogram is used to check the shape of the data distribution.
  3. Used to check whether the process changes from one period to another.
  4. Used to determine whether the output is different when it involves two or more processes.
  5. Used to analyse whether the given process meets the customer requirements.

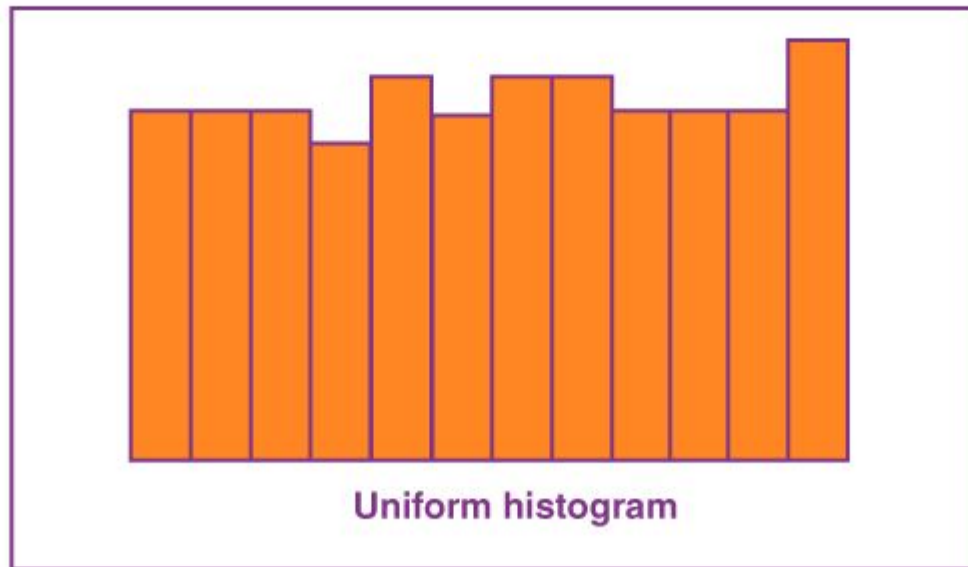


# Uniform Histogram

- Uniform histogram
- Symmetric histogram
- Bimodal histogram
- Probability histogram

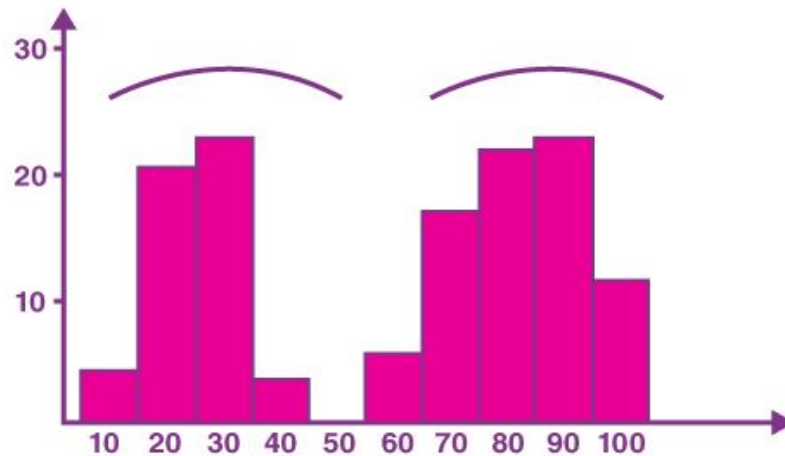
# Uniform Histogram:

A uniform distribution reveals that the number of classes is too small, and each class has the same number of elements. It may involve distribution that has several peaks.



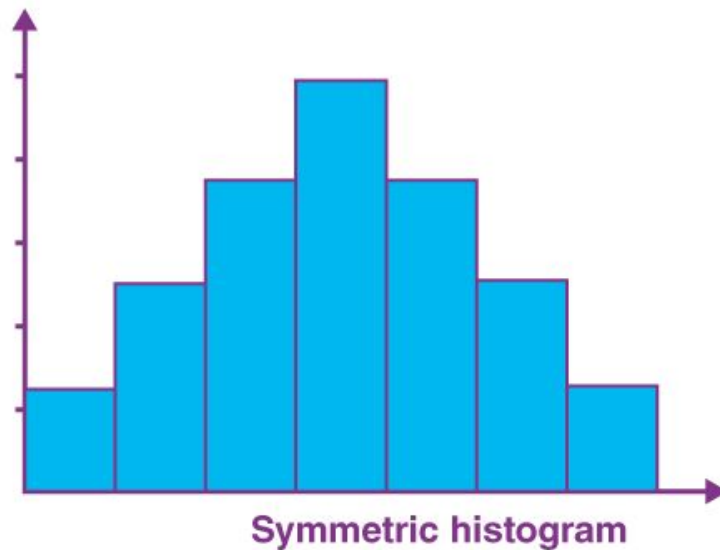
# Bimodal Histogram

- If a histogram has two peaks, it is said to be bimodal.
- Bimodality occurs when the data set has observations on two different kinds of individuals or combined groups if the centers of the two separate histograms are far enough to the variability in both the data sets.



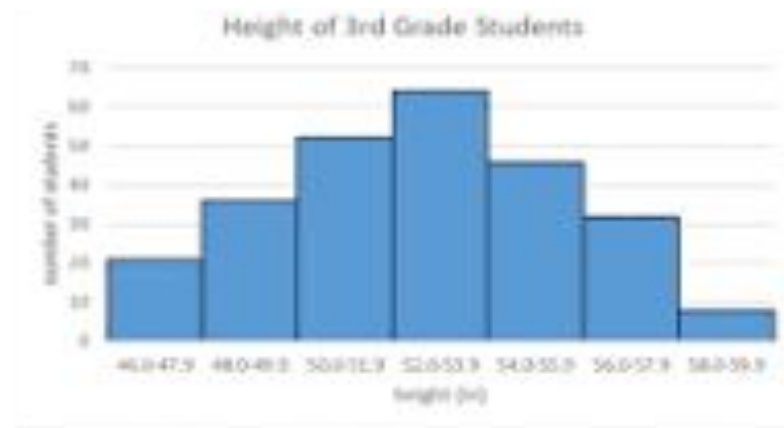
# Symmetric Histogram

- A symmetric histogram is also called a bell-shaped histogram.
- When you draw the vertical line down the center of the histogram, and the two sides are identical in size and shape, the histogram is said to be symmetric.
- The diagram is perfectly symmetric if the right half portion of the image is similar to the left half.
- The histograms that are not symmetric are known as skewed.



# Probability Histogram

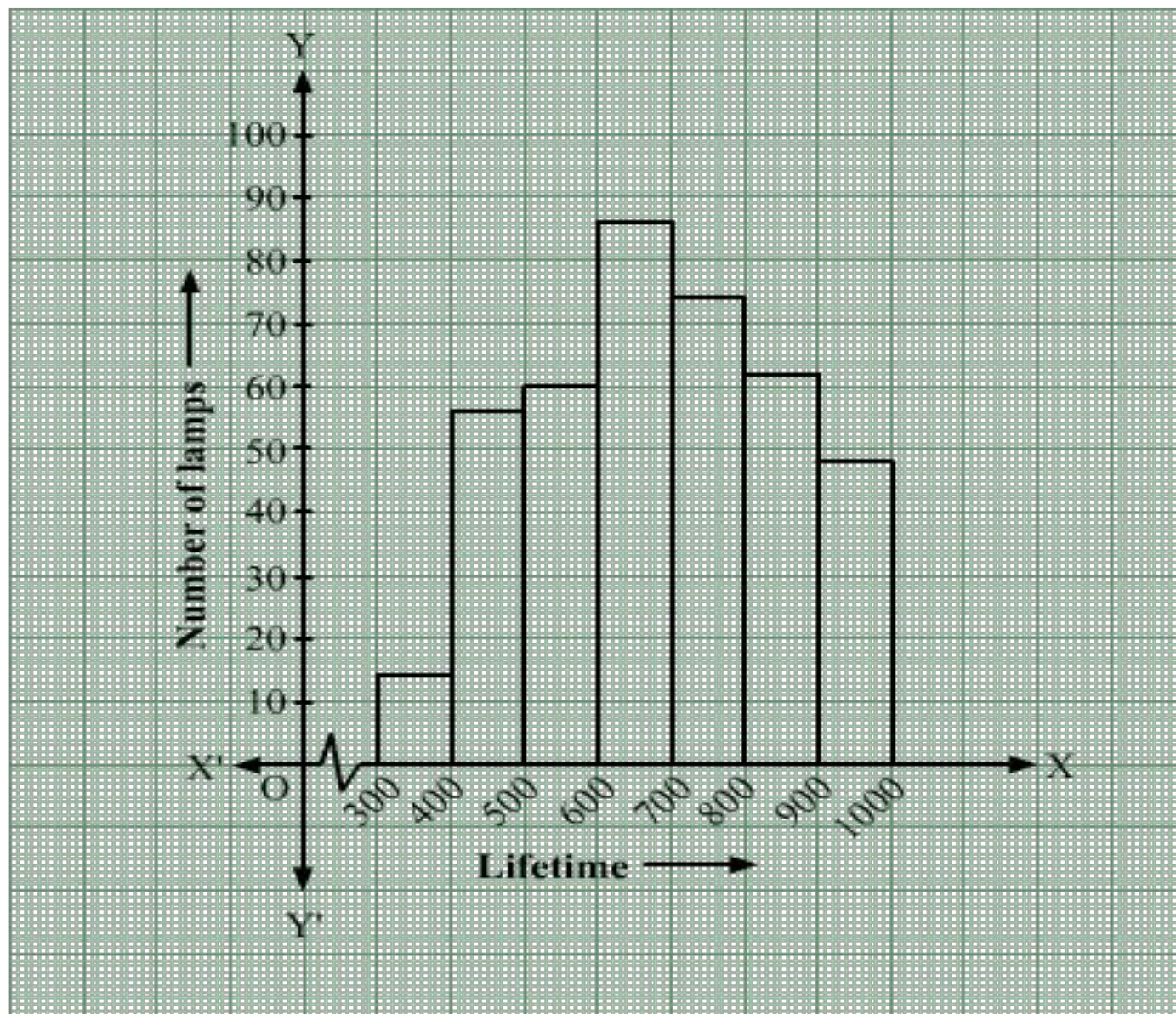
- A Probability Histogram shows a pictorial representation of a discrete probability distribution.
- It consists of a rectangle centered on every value of  $x$ , and the area of each rectangle is proportional to the probability of the corresponding value.
- The probability histogram diagram is begun by selecting the classes.
- The probabilities of each outcome are the heights of the bars of the histogram.



# Histogram Solved Example

- **Question:** The following table gives the lifetime of 400 neon lamps. Draw the histogram for the below data.

Lifetime (in hours)	Number of lamps
300 – 400	14
400 – 500	56
500 – 600	60
600 – 700	86
700 – 800	74
800 – 900	62
900 – 1000	48



# Scatter Plots:

- Scatter plots show the relationship between two continuous variables.
- They are useful for identifying correlations or clusters in data.
- A scatter plot is also called a scatter chart, scatter gram, or scatter plot, XY graph.
- The scatter diagram graphs numerical data pairs, with one variable on each axis, show their relationship.



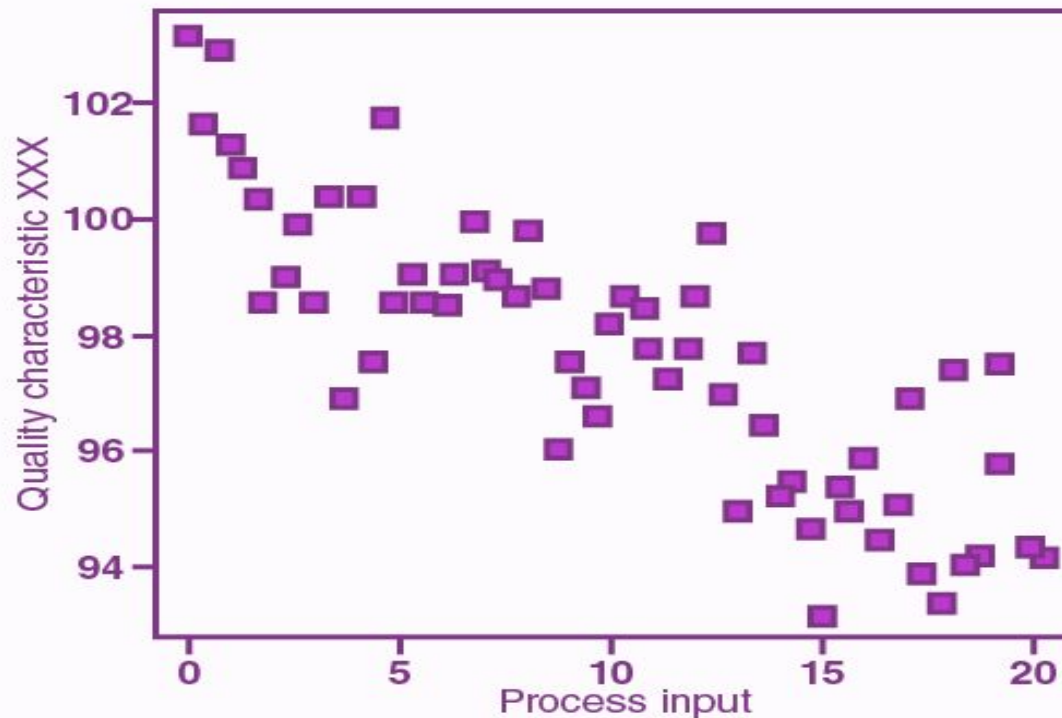
# Scatter plots are used in either of the following situations

1. When we have paired numerical data
2. When there are multiple values of the dependent variable for a unique value of an independent variable
3. In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appear to be related both occur with the exact cause and so on.

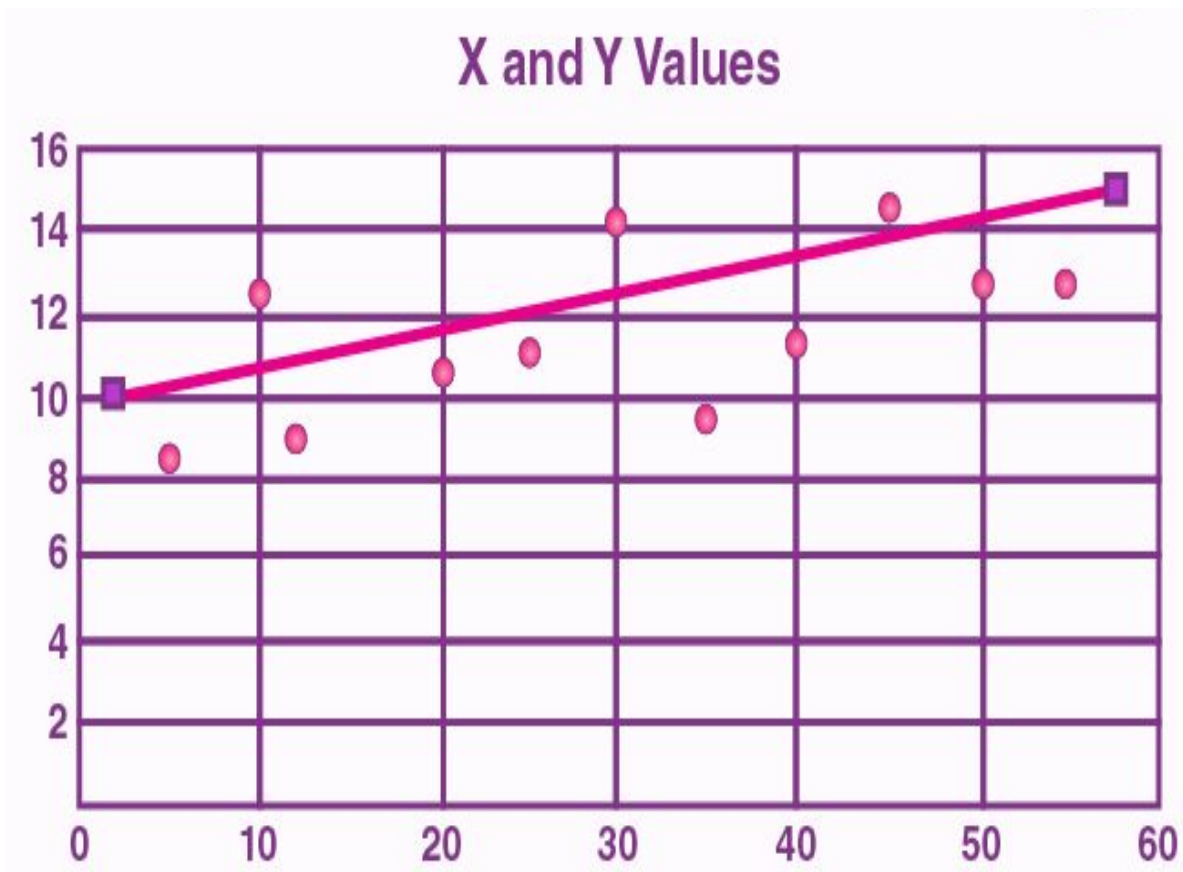
# Scatter Plot Uses and Examples

Scatter plots instantly report a large volume of data. It is beneficial in the following situations

- For a large set of data points given
- Each set comprises a pair of values
- The given data is in numeric form



- The line drawn in a scatter plot, which is near to almost all the points in the plot is known as “**line of best fit**” or “**trend line**”. See the graph below for an example.



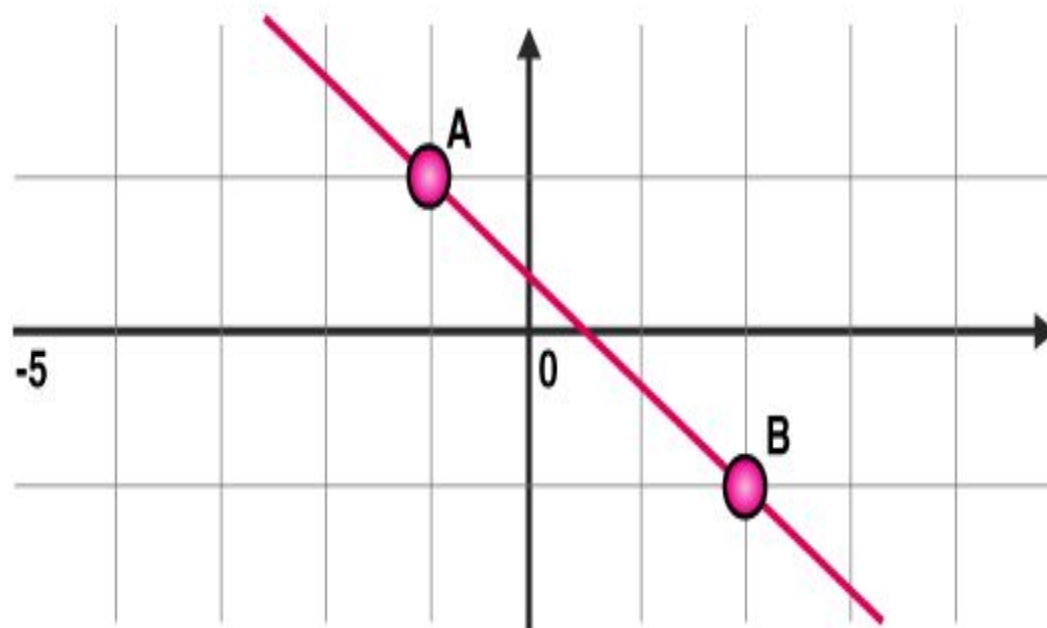
# Line Charts:

- Line charts are ideal for showing trends and changes over time.
- They connect data points with lines, making it easy to see how values evolve over a continuous scale, such as time.
- A line graph or **line chart** or **line plot** is a graph that utilizes points and lines to represent change over time.
- It is a chart that shows a line joining several points or a line that shows the relation between the points.
- The graph represents quantitative data between two changing variables with a line or curve that joins a series of successive data points.
- Linear graphs compare these two variables in a vertical axis and a horizontal axis.

# How to Make a Line Graph?

- If we have created data tables, then we draw linear graphs using the data tables.
- These graphs are plotted as a series of points, which are later joined with straight lines to provide a simple way to review data collected over time.
- It offers an excellent visual format of the outcome data collected over time.

- To plot a linear/line graph follow the below steps:
  1. Use the data from the data-table to choose a suitable scale.
  2. Draw and label the scale on the vertical (y-axis) and horizontal (x-axis) axes.
  3. List each item and place the points on the graph.
  4. Join the points with line segments.



# Types of Line Graphs

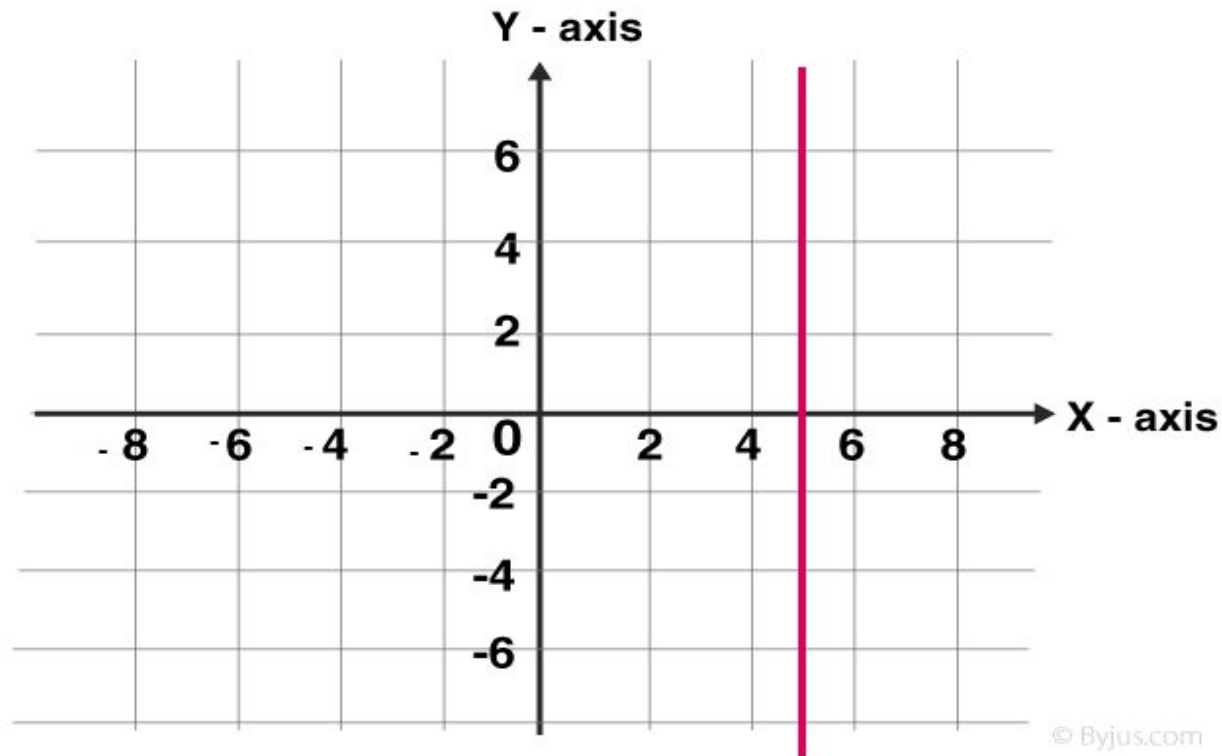
The following are the types of the line graph They are:

1. **Simple Line Graph:** Only one line is plotted on the graph.
2. **Multiple Line Graph:** More than one line is plotted on the same set of axes. A multiple line graph can effectively compare similar items over the same period of time.
3. **Compound Line Graph:** If information can be subdivided into two or more types of data. This type of line graph is called a compound line graph. Lines are drawn to show the component part of a total. The top line shows the total and line below shows part of the total. The distance between every two lines shows the size of each part.



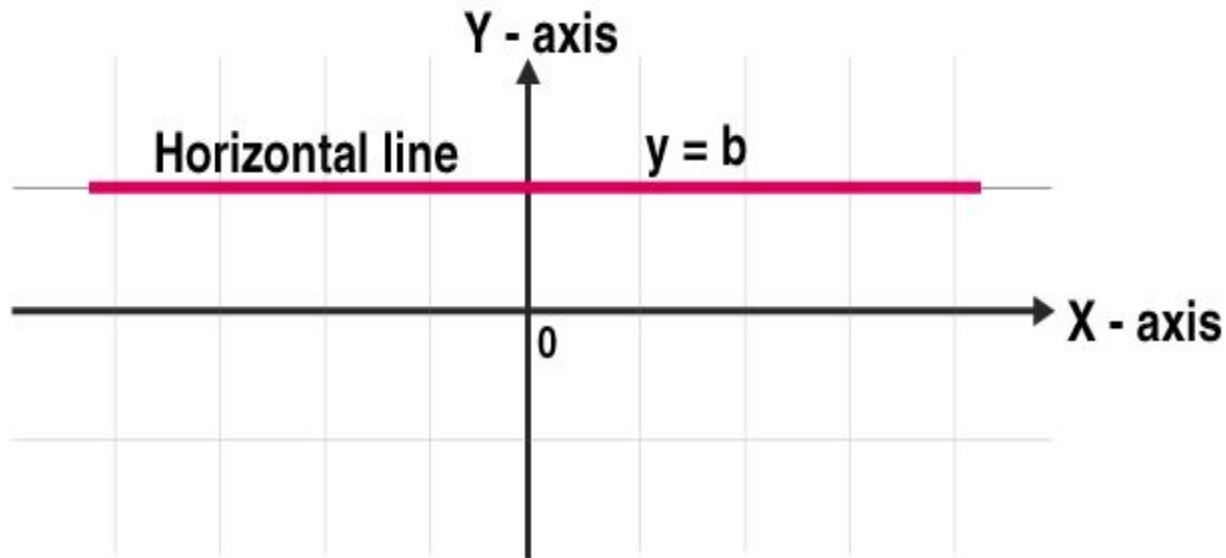
# Vertical Line Graph

- Vertical line graphs are graphs in which a vertical line extends from each data point down to the horizontal axis.
- Vertical line graph sometimes also called a column graph.
- A line parallel to the y-axis is called a vertical line.



# Horizontal Line Graph

- Horizontal line graphs are graphs in which a horizontal line extends from each data point parallel to the earth.
- Horizontal line graph sometimes also called a row graph.
- A line parallel to the x-axis is called a vertical line.



# Straight Line Graph

- A line graph is a graph formed by segments of straight lines that join the plotted points that represent given data.
- The line graph is used to solve changing conditions, often over a certain time interval.
- A general linear function has the form  $y = mx + c$ , where  $m$  and  $c$  are constants.


## Straight Line Graphs

A **straight line graph** is a visual representation of a linear function.

A straight line has a general equation of

$$y = mx + c$$

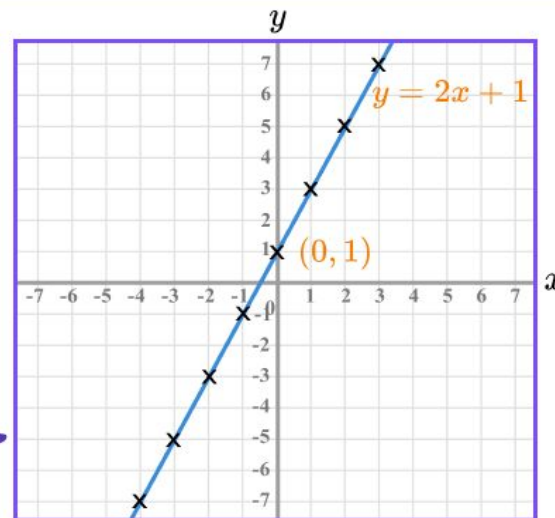
gradient      y-intercept

 Example

$$y = 2x + 1$$

$$m = 2, \text{ and } c = 1$$

The graph of this equation looks like this:



# Pie Charts:

- Display parts of a whole, where each slice represents a proportion of the total.
- The “**pie chart**” is also known as a “circle chart”, dividing the circular statistical graphic into sectors or sections to illustrate the numerical problems.
- Each sector denotes a proportionate part of the whole.
- To find out the composition of something, Pie-chart works the best at that time.
- In most cases, pie charts replace other graphs like the bar graph, line plots, histograms, etc.

## Formula

- The pie chart is an important type of data representation.
- It contains different segments and sectors in which each segment and sector of a pie chart forms a specific portion of the total(percentage).
- The sum of all the data is equal to 360°.

**The total value of the pie is always 100%.**

- To work out with the percentage for a pie chart, follow the steps given below:
- Categorize the data
- Calculate the total
- Divide the categories
- Convert into percentages
- Finally, calculate the degrees
- Therefore, the pie chart formula is given as

**$(\text{Given Data} / \text{Total value of Data}) \times 360^\circ$**

# How to Create a Pie Chart?

- Imagine a teacher surveys her class on the basis of favorite Sports of students:

Football	Hockey	Cricket	Basketball	Badminton
10	5	5	10	10

- The data above can be represented by a pie chart as following and by using the circle graph formula, i.e. the pie chart formula given below. It makes the size of the portion easy to understand.
- Solution:

**Step 1:** First, Enter the data into the table.

Football	Hockey	Cricket	Basketball	Badminton
10	5	5	10	10

- **Step 2:** Add all the values in the table to get the total.
- I.e. Total students are 40 in this case.
- **Step 3:** Next, divide each value by the total and multiply by 100 to get a percent:

Football	Hockey	Cricket	Basketball	Badminton
$(10/40) \times 100$ =25%	$(5/40) \times 100$ =12.5%	$(5/40) \times 100$ =12.5%	$(10/40) \times 100$ =25%	$(10/40) \times 100$ =25%

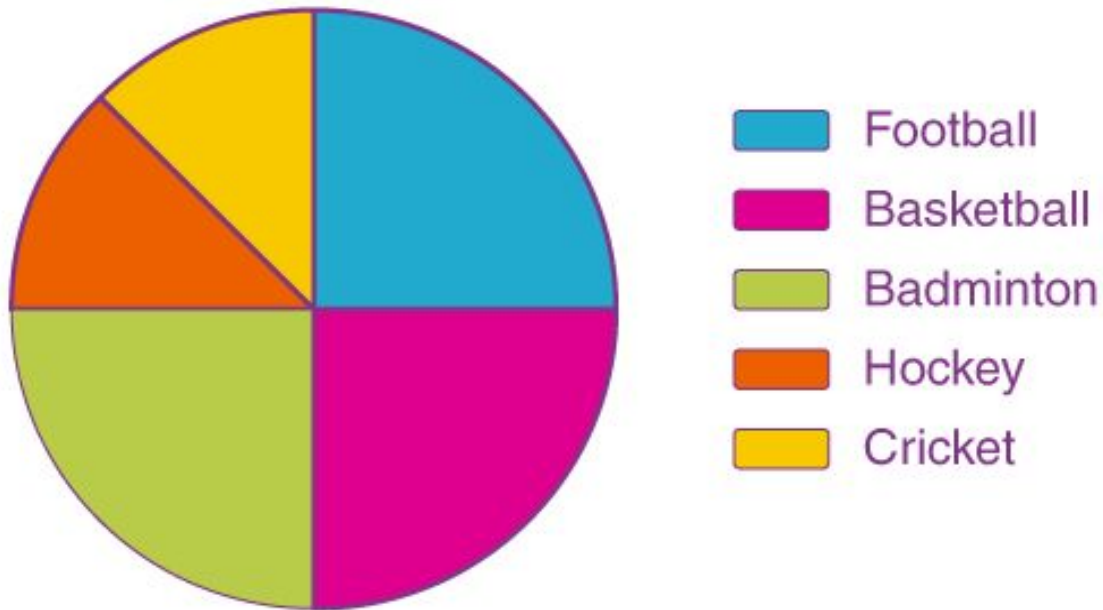
- **Step 4:** Take a full circle of 360° and follow the calculations below:
- The central angle of each component = (Value of each component/sum of values of all the components)  $\times 360^\circ$

Football	Hockey	Cricket	Basketball	Badminton
$(10/40) \times 360^\circ$ =90°	$(5/40) \times 360^\circ$ =45°	$(5/40) \times 360^\circ$ =45°	$(10/40) \times 360^\circ$ =90°	$(10/40) \times 360^\circ$ =90°

Now you can draw a pie chart.

- **Step 5:** Draw a circle and use the protractor to measure the degree of each sector.

## Favourite Sports Percentage





# Uses of Pie Chart

- Within a business, it is used to compare areas of growth, such as turnover, profit and exposure.
- To represent categorical data.
- To show the performance of a student in a test, etc.

# Heatmaps:

- Use color intensity to represent values in a matrix, often used for visualizing correlations or patterns in large datasets.
- A heatmap is a graphical representation of data that uses a system of color coding to represent different values.
- Heatmaps are used in various forms of analytics but are most commonly used to show user behavior on specific web pages or webpage templates.
- Heatmaps can be used to show where users have clicked on a page, how far they have scrolled down a page, or used to display the results of eye-tracking tests.

# Benefits of heatmaps

- Visualize a variety of data points, including mouse clicks, scroll depth, and eye movements.
- Identify areas of a page that are most (or least) engaging.
- Test different design changes to see how they affect user behavior.
- Troubleshoot problems with a page's usability.

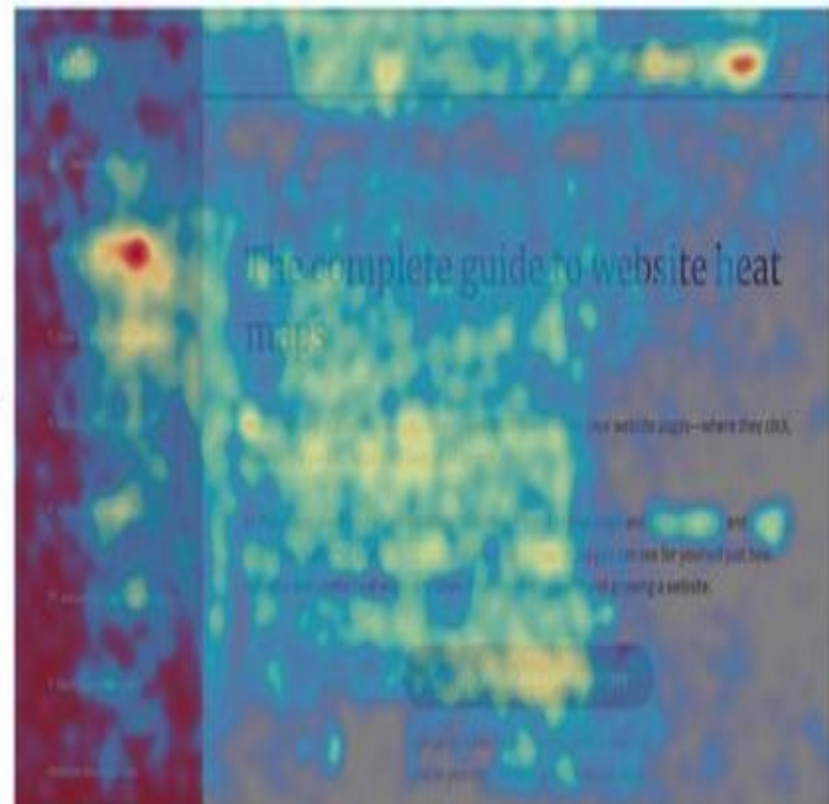
# What are the different types of heat maps?

There are many different types of heatmaps, but some of the most common ones include:

- **Click heatmaps:** These heatmaps show where users click on a webpage. Identify the integral elements on a page and see how users interact with different features.
- **Scroll heatmaps:** See how far users scroll down a webpage with this type of heatmap. See which parts of a page are most engaging and how users find the information they are looking for.
- **Mouse movement heatmaps:** These heatmaps show the path of a user's mouse as they move the cursor around a webpage. Know where users are looking and how they interact with different elements on the page.

- **Eye tracking heatmaps:** This heatmap shows the path of a user's eye movements as they look at a webpage. Understand where users are paying attention and how they process different elements on the page.
- **Conversion heatmaps:** Get a view of all the steps your users take to complete a desired action, such as when making a purchase, clicking on the call to action (ctas), or signing up for a newsletter. Use this information to identify bottlenecks in the conversion process and guide users to take the desired action.

ELEMENT	VISIBLE	INTERACTION POINTS	N OF TOTAL
#hs_cos_wrapper_Main_Content	Yes	4,982	5.3%
u content-column>div.hub-intro-section-content-text>	Yes	3,987	4.2%
u n-content-column>div.hub-intro-section-content-text>	Yes	3,051	3.2%
u olumn-wrapper>div.hub-intro-section-content-column>	Yes	2,959	3.1%
u ow-normal.hub-content-image-row>div.image-row-ima>	Yes	2,456	2.6%
div#hs_menu_wrapper_module_14701478960212019>ul(1	Yes	2,280	2.4%
u content-column>div.hub-intro-section-content-image>	Yes	2,268	2.4%
#hub-content	Yes	2,081	2.2%
u 553172750295>div.hub-content-row.hub-content-row>	Yes	2,007	2.1%



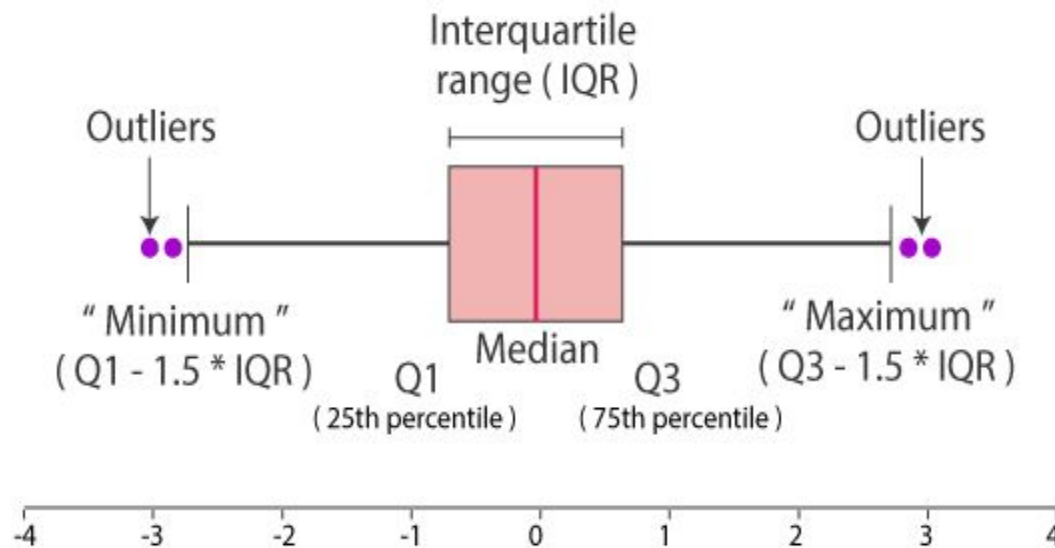
# Box Plots

- Summarize the distribution of a variable by showing its quartiles, outliers, and overall spread.
- The method to summarize a set of data that is measured using an interval scale is called a box and whisker plot.
- These are maximum used for data analysis.
- We use these types of graphs or graphical representation to know:
  - Distribution Shape
  - Central Value of it
  - Variability of it
- A box plot is a chart that shows data from a five-number summary including one of the measures of [central tendency](#).
- It does not show the distribution in particular as much as a stem and leaf plot or histogram does.

- In simple words, we can define the box plot in terms of descriptive statistics related concepts.
- That means box or whiskers plot is a method used for depicting groups of numerical data through their quartiles graphically.
- These may also have some lines extending from the boxes or whiskers which indicates the variability outside the lower and upper quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram.
- Outliers can be indicated as individual points.



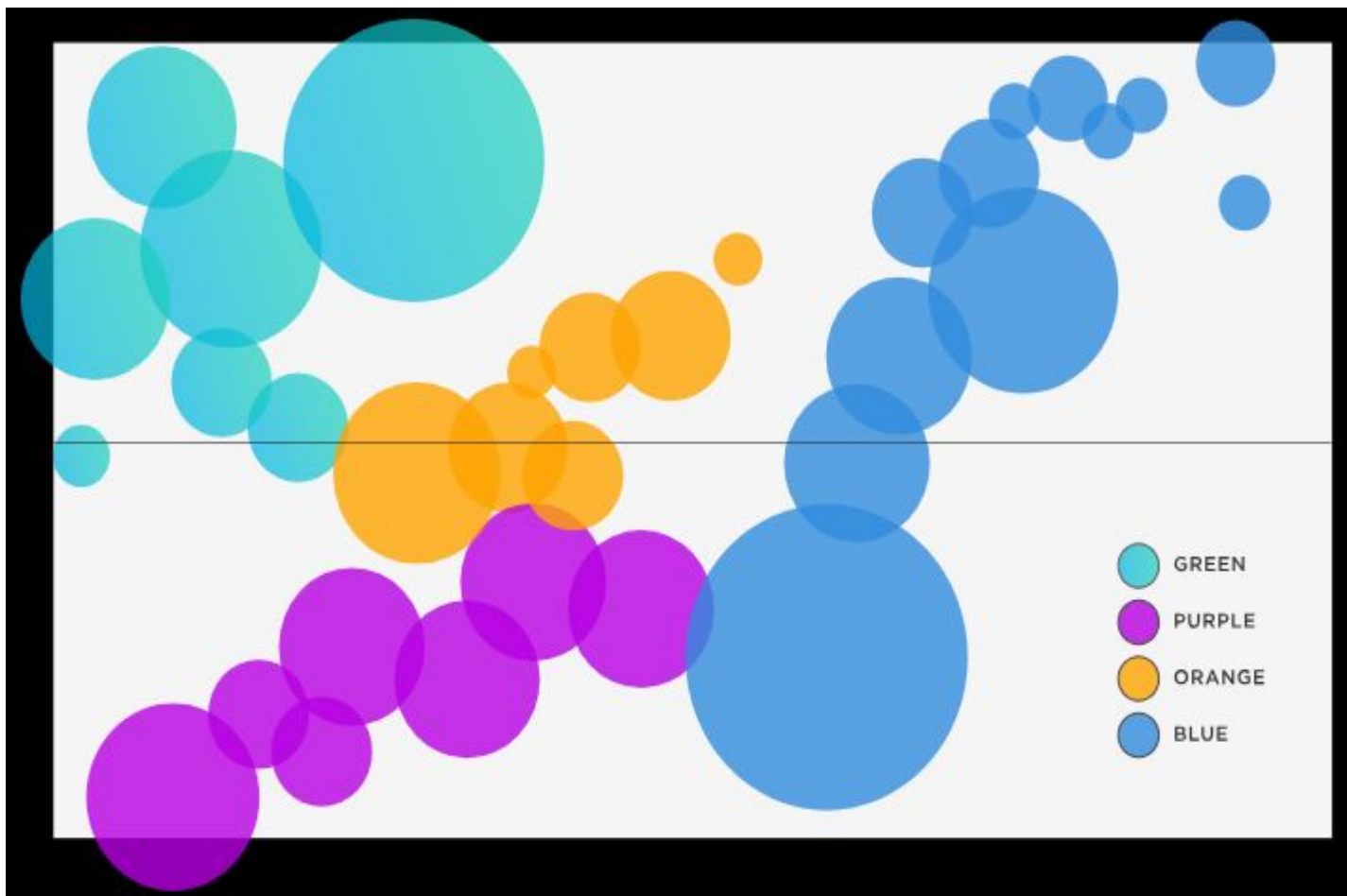
Check the image below which shows the minimum, maximum, first quartile, third quartile, median and outliers.



**Different parts of boxplot**

# Bubble Charts:

- Bubble charts are a variation of scatter plots where the size of the data points is used to represent a third variable.
- They are useful for visualizing three-dimensional data.
- **Bubble charts**, also known as **bubble plots** or **bubble graphs**, are used when data needs a third dimension to provide richer information to viewers.
- A bubble plot is a relational chart designed to compare three variables.
- Unlike other three-dimensional charts that process and represent data across three axes (usually x, y, and z), a bubble chart is represented on two axes (x and y), and the size of the bubble communicates the third, vital piece of information.

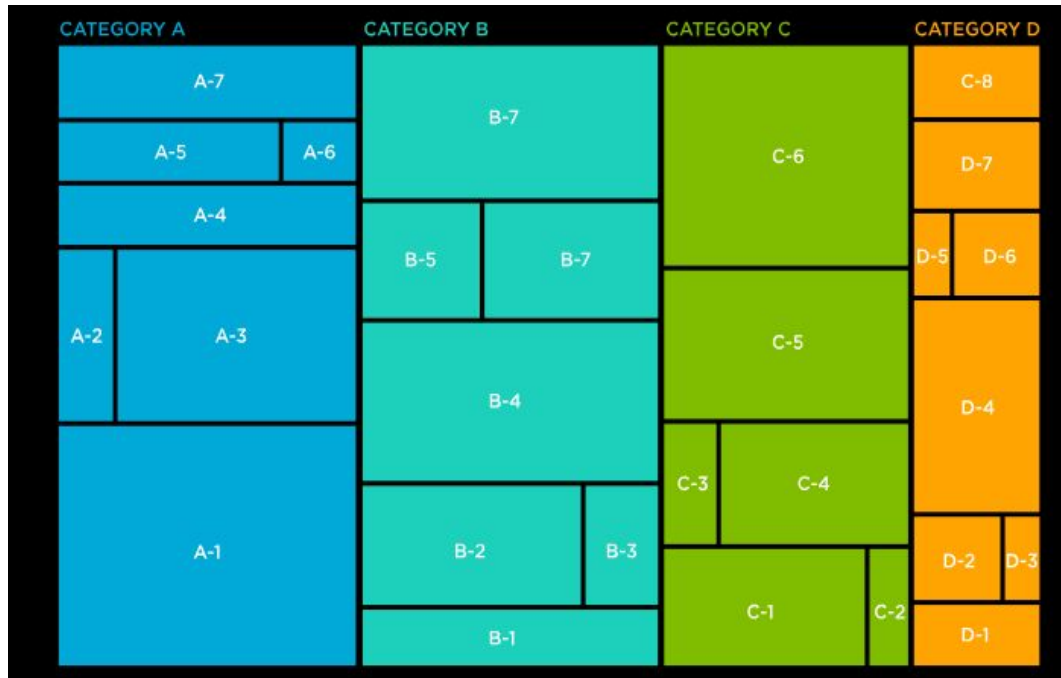


# Example

- For example, a country's population could be large or small to begin with but grow faster when health and sanitation facilities are available. It also falls rapidly when population control measures are implemented. This scenario would show bubble size increasing over time as infrastructure improved. In this socio-economic example, a bubble chart can help us understand how different parameters move over time.
- Bubble charts can also be useful in a business context. In fact, a bubble chart is often used in one of today's core financial processes: valuation and investments. For example, the cost of valuation can be studied against risk by using the standard axes to represent cost and value and the bubble sizes to represent risk.

# Treemaps:

- Treemaps are hierarchical visualizations that display data as nested rectangles.
- They are often used to represent hierarchical data structures and show the breakdown of a whole into its parts.
- Treemapping is a [data visualization](#) technique that displays [hierarchical data](#) using rectangles of decreasing sizes, often called nesting, to create **treemap charts**.



# Characteristics and Components of a Treemap Chart

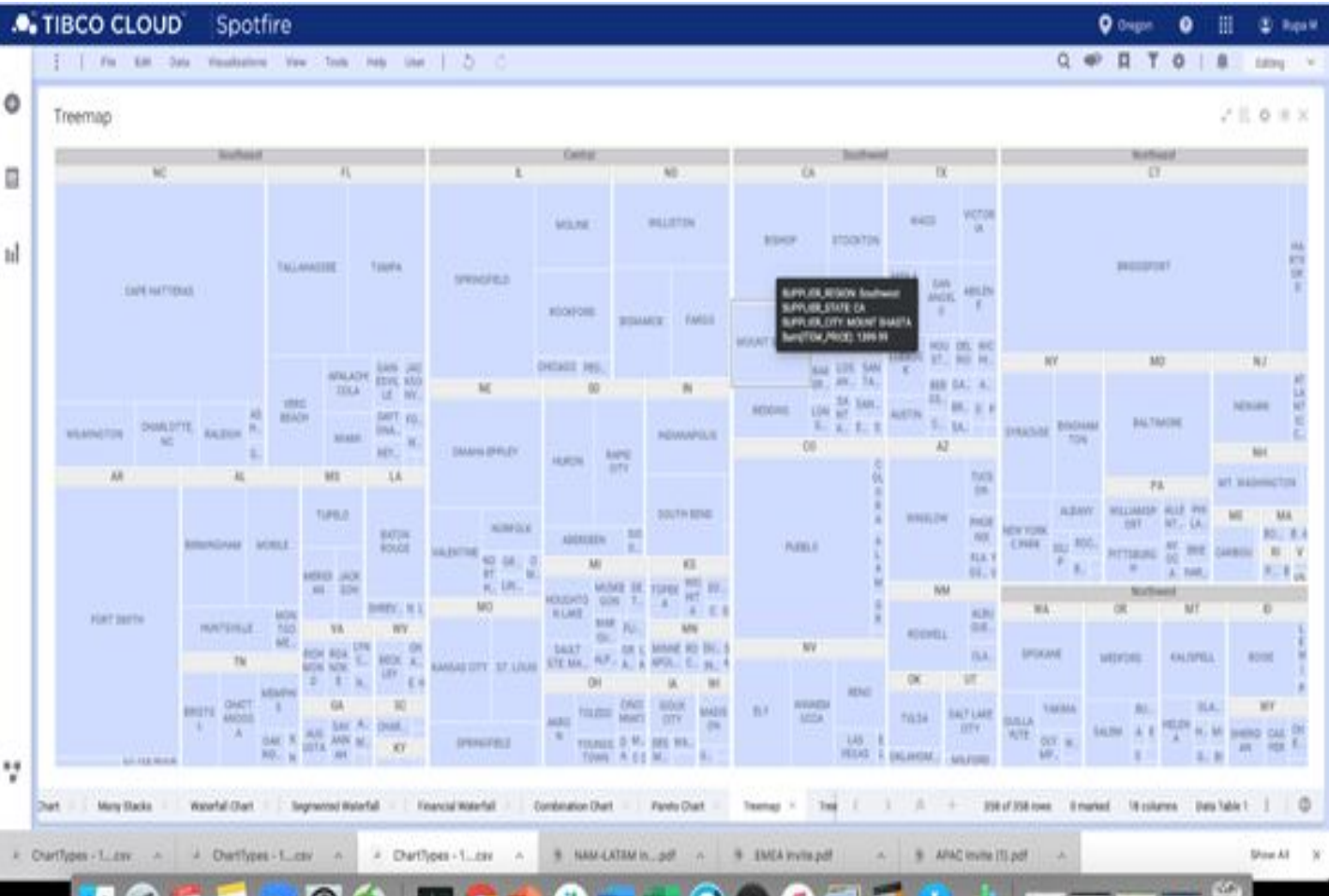
- Data is represented using rectangles.
- Each rectangle represents two numerical values. The rectangles are sometimes referred to as “nodes” or “branches.” The nested datasets within are then referred to as “leaves.”
- The dimensions and plot colors of the rectangles are calculated based on the quantitative variables associated with the respective rectangles.
- The data can be multi-layered: hierarchically organized data is depicted in a set of nested rectangles, with the “parent elements” tiled along with their “child elements.”
- When a quantity is assigned to a category, the area size of the rectangle is proportionate to that quantity.
- The parent category’s area is made up of the sum of its subcategories.

- The rectangles in the treemap are arranged according to size. The standard format is that the rectangles range in size from the chart's top left corner and flow to the bottom right corner. Therefore, the top left corner of the treemap has the largest rectangle, whereas the bottom right corner of the chart has the smallest rectangle.
- With hierarchical data, i.e. with nested rectangles, the same order is followed, with the lower level rectangles stacked within each higher level rectangle in the treemap.
- The size and position in the chart of the parent rectangle containing the nested rectangle depends on the sum of the areas of the nested rectangles.

# Examples of Treemap Charts and Their Variants

- There are a wide number of ways that a treemap can be used across industries, areas of study, and presentation types:
- Comparing the sales numbers of different brands or models over a certain period can make up a great two-dimensional treemap chart
- Literacy rates within districts belonging to a certain geographic area over a specific period
- Relative population densities of the top 10 highly populated countries
- Inventory of various birds, animals, and fish (including types) in a pet store is the perfect example of a nested treemap chart.
- Below is an image of a treemap chart that uses data referring to sales of electronic items in several cities.





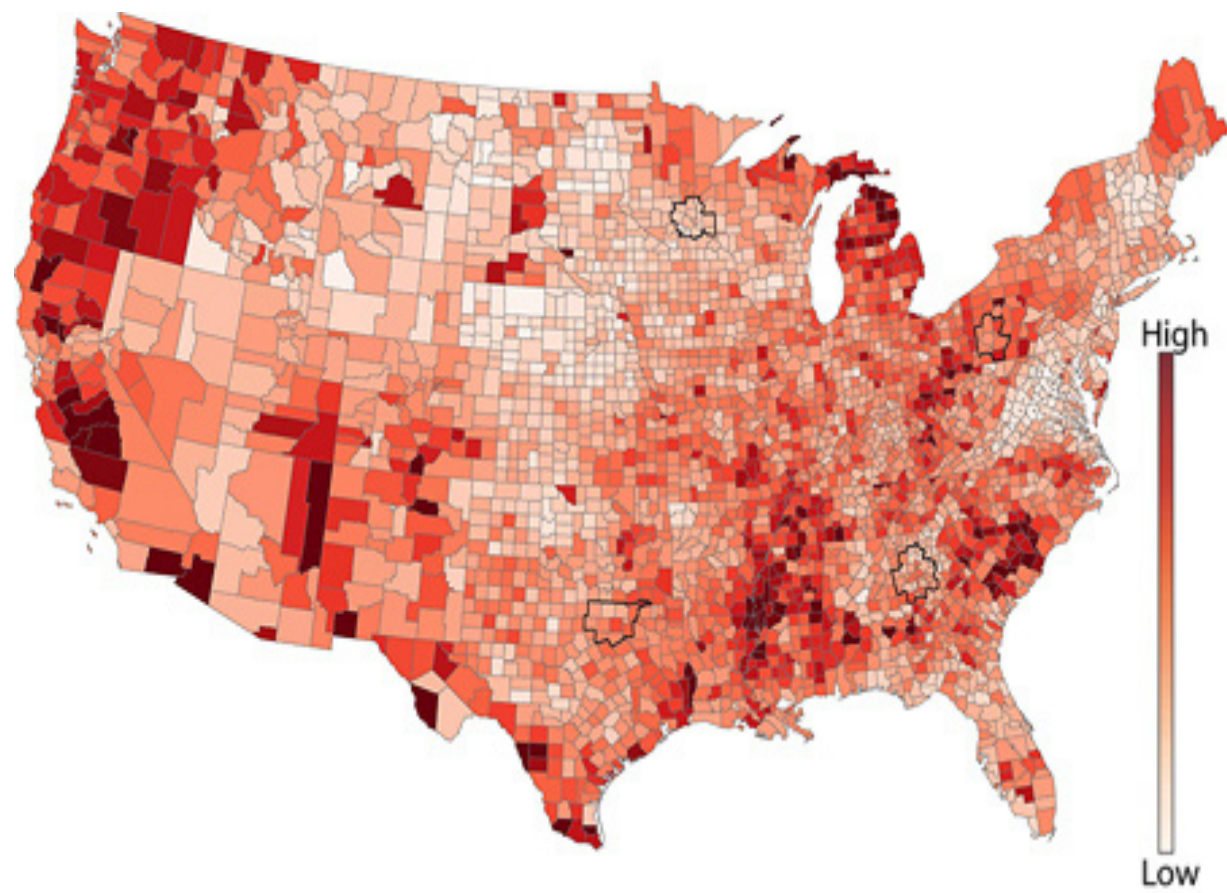
# Choropleth Maps:

- Choropleth maps use color-coding to represent data values for different geographic regions.
- They are commonly used to visualize regional or spatial patterns.
- A choropleth map is a thematic map that is used to represent statistical data using the color mapping symbology technique.
- It displays enumeration units, or divided geographical areas or regions that are colored, shaded or patterned in relation to a data variable.
- To show variation or patterns across the displayed location, choropleth maps provide a way to visualize values over a geographical area.

# What is a choropleth map used for?

Choropleth maps are useful when the data you want to convey are:

- Attached to enumeration units like cities, counties and countries ,standardized to show rates or ratios, and in a continuous statistical surface, meaning measurable values occur everywhere within the area of study and not just at specific locations.



- **Programming Languages and Libraries for Computational Statistics and Data Visualization:**

- **Python:**
  - Libraries: matplotlib, seaborn, Plotly, Pandas (for data manipulation), NumPy (for numerical computing), and SciPy (for scientific computing).
- **R:**
  - R is a statistical programming language with built-in data visualization capabilities. It has packages like ggplot2, lattice, and base R plotting functions.
- **Julia:**
  - Julia is a high-performance language for technical computing, and it has packages like Plots.jl and Gadfly.jl for data visualization.
- **JavaScript:**
  - Libraries like D3.js and Chart.js are commonly used for web-based data visualization.
- **Tableau:**
  - Tableau is a popular data visualization tool for creating interactive and shareable dashboards.
- **Power BI:**
  - Microsoft Power BI is another powerful tool for data visualization and business intelligence.
- **Excel:**
  - Microsoft Excel provides basic data visualization capabilities with charts and graphs.

# Presentation Graphics

- Presentation and exploratory graphics are essential components of data visualization, serving distinct purposes in the process of data analysis and communication.

## **Presentation Graphics:**

- Presentation graphics are visuals created with the primary goal of communicating insights and findings to a broader audience, such as stakeholders, clients, or the general public.
- These graphics are typically polished, refined, and designed for clarity and impact.
- Key characteristics of presentation graphics include:

# Key characteristics of presentation graphics include:

- **Clarity and Simplicity:** Presentation graphics must be clear, concise, and easy to understand for a non-technical audience. They avoid unnecessary complexity or jargon.
- **Aesthetics:** Aesthetics play a significant role in presentation graphics. These visuals often prioritize design elements such as color choices, fonts, and layout to make them visually appealing.
- **Storytelling:** Presentation graphics often follow a narrative structure. They are used to tell a data-driven story, emphasizing key insights and supporting the intended message.
- **Limited Interactivity:** Presentation graphics are typically static or minimally interactive. They are intended for passive consumption and may not include extensive interactive features.
- **Visualization Types:** Common types of presentation graphics include bar charts, line charts, pie charts, infographics, and well-designed dashboards. The choice of visualization type depends on the data and the message you want to convey.
- **Audience-Centric:** Presentation graphics are tailored to the needs and knowledge level of the target audience, ensuring that the data is accessible and meaningful to them.
- Tools commonly used for creating presentation graphics include Microsoft PowerPoint, Adobe Illustrator, Tableau, and other data visualization software designed for non-technical users.



# Exploratory Graphics:

- Exploratory graphics, on the other hand, are visuals created during the data exploration phase, primarily for the benefit of data analysts and researchers.
- These graphics are dynamic, often changing as analysts explore and understand the data better.
- Key characteristics of exploratory graphics include:

- **Flexibility and Interactivity:** Exploratory graphics are flexible and interactive. They allow analysts to explore various facets of the data, zoom in on details, and experiment with different visual representations.
- **Richness of Information:** These graphics may contain more detailed information, multiple plots, and overlays to support in-depth data exploration. They help analysts uncover patterns, outliers, and relationships in the data.
- **Hypothesis Testing:** Exploratory graphics are used for quick prototyping and hypothesis testing. Analysts use them to generate hypotheses about the data, which can then be further tested.

- **Raw Data Examination:** They often involve the examination of raw data to identify issues, missing values, or anomalies that require data preprocessing or cleaning.
- **Statistical Analysis:** Computational statistics is closely integrated with exploratory graphics. Analysts calculate summary statistics, correlations, and other statistical measures to guide the creation of visualizations.
- **Data Iteration:** As insights emerge during exploration, exploratory graphics can evolve to accommodate new findings or questions raised during the analysis process.

- Tools commonly used for creating exploratory graphics include programming languages like R, Python (with libraries such as matplotlib, seaborn, and Plotly), and specialized data exploration platforms.

# Graphics and Computing

- Graphics and computing play crucial roles in data visualization, enhancing the effectiveness and interactivity of visual representations of data.
- Here's how they are integrated into the field of data visualization:

# Graphics in Visualization:

- Graphics play a vital role in computational statistics and data visualization.
- They are essential for representing complex data, revealing patterns, and communicating findings effectively.
- Here are some key aspects of graphics in computational statistics and data visualization:

- **Exploratory Data Analysis (EDA):** Graphics are often the first step in understanding data. EDA uses various graphical representations like scatter plots, histograms, box plots, and density plots to visualize data distributions, identify outliers, and explore relationships between variables. These graphics help analysts gain insights into the structure and characteristics of the dataset.
- **Statistical Graphics:** Statistical graphics, such as bar charts, line charts are used to visualize the results of statistical analyses. They allow for the representation of relationships between variables, the distribution of data, and trends over time. Statistical graphics are instrumental in conveying the results of hypothesis tests, regression analyses, and other statistical procedures.
- **Data Visualization Libraries:** Specialized data visualization libraries and tools are commonly used in computational statistics. Examples include ggplot2 (R), Matplotlib (Python), and D3.js (JavaScript). These libraries offer a wide range of customizable graphics and make it easier for analysts to create informative visualizations.
- **Interactive Visualization:** Interactive graphics go beyond static images and allow users to interact with data visualizations dynamically. Interactive dashboards, tooltips, zooming, and panning provide a more immersive exploration experience. Libraries like Plotly and Bokeh enable the creation of web-based interactive visualizations for data exploration and presentation.
- **Geospatial Visualization:** For data with a geographic component, maps and geospatial visualizations are crucial. Geographic Information Systems (GIS) software and libraries like Leaflet (JavaScript) and Folium (Python) enable the creation of maps that display spatial data, making it easier to analyze and communicate location-based information.

- **3D Visualization:** In some cases, three-dimensional graphics are used to represent complex datasets with depth and volume. 3D plots and visualizations can be created for scientific and engineering applications to visualize 3D data structures and simulations.
- **Time Series Visualization:** Time series data, common in computational statistics, is effectively represented using line charts, area charts, and heatmaps. Time series graphics allow analysts to detect trends, seasonal patterns, and anomalies in time-varying data.
- **Heatmaps:** Heatmaps are used to display matrices of data, such as correlation matrices, confusion matrices, or data matrices. They use color intensity to represent values, making it easy to identify patterns and relationships within large datasets.
- **Parallel Coordinates:** Parallel coordinates plots are helpful for visualizing high-dimensional data by representing each data point as a line that intersects parallel axes. They can reveal patterns and clusters in multivariate datasets.
- **Statistical Reports:** Graphics are often integrated into statistical reports and research papers. Tools like LaTeX and R Markdown allow for the inclusion of dynamic and high-quality graphics within documents, making it easier to communicate statistical findings.
- **Data-Driven Storytelling:** Data visualization is an essential tool for telling data-driven stories. Effective storytelling using graphics helps convey the significance and implications of statistical analyses to a non-technical audience.



# Computing in Visualization:

- Computing plays a pivotal role in computational statistics and data visualization, as it provides the computational power and tools needed to perform complex data analyses, generate visualizations, and gain insights from data.
- Here's how computing is integrated into these fields:

- **Data Processing and Management:** Computing is used to preprocess, clean, and manage datasets. It involves tasks such as data extraction, transformation, and loading (ETL), handling missing data, and merging data from multiple sources. Tools like Python, R, SQL, and data wrangling libraries simplify these processes.
- **Statistical Computations:** Statistical analyses often require substantial computational resources. Computing platforms provide the ability to perform statistical tests, calculations of summary statistics, hypothesis testing, and regression analyses. Specialized statistical software packages like R, SAS, and SPSS offer a wide range of statistical functions.
- **Machine Learning and Predictive Modeling:** Computational statistics leverages machine learning algorithms to build predictive models, classification systems, clustering, and dimensionality reduction techniques. Computing resources are essential for training, validating, and optimizing these models, especially when dealing with large datasets or complex algorithms.
- **Numerical Optimization:** Many statistical procedures involve numerical optimization techniques to find model parameters that minimize or maximize specific criteria (e.g., maximum likelihood estimation). Computing is crucial for solving optimization problems efficiently.

- **Simulation Studies:** Monte Carlo simulations and bootstrapping techniques are frequently employed in computational statistics to estimate parameters and assess the uncertainty of statistical estimates. These simulations involve running statistical models or procedures repeatedly, which requires significant computing power.
- **Resampling Techniques:** Bootstrap resampling and permutation tests are resampling techniques commonly used for hypothesis testing and model validation. These techniques rely on extensive computational iterations to generate distributions of test statistics.
- **Parallel and Distributed Computing:** High-performance computing (HPC) clusters and distributed computing frameworks (e.g., Apache Spark) are used to accelerate computationally intensive statistical analyses. Parallelization techniques distribute tasks across multiple processors or nodes, improving computational efficiency.
- **Data Visualization Libraries:** Data visualization libraries and tools, such as Matplotlib, ggplot2, and Plotly, are designed to create various types of visualizations. Computing is required to process and transform data, generate graphics, and adjust visual elements to create informative visual representations.

- **Interactive Data Exploration:** Interactive computing environments like Jupyter Notebooks and R Markdown allow analysts to perform exploratory data analysis (EDA) collaboratively and interactively. They enable the integration of code, text, and visualizations to create data-driven narratives.
- **Geospatial Analysis:** Geographic Information Systems (GIS) and geospatial libraries use computing to analyze and visualize spatial data, perform spatial statistics, and create maps for data visualization.
- **High-Performance Graphics Rendering:** When dealing with large datasets or real-time data visualizations, computing power is essential for rendering complex graphics efficiently. Graphics processing units (GPUs) are often employed to accelerate rendering tasks.
- **Interactive Dashboards:** Building interactive dashboards for data visualization and exploration requires computing resources to handle user interactions and update visualizations dynamically. Tools like Shiny (R) and Dash (Python) integrate computing and visualization for interactive dashboard development.
- **Data Simulation and Generation:** Computing is used to simulate synthetic datasets for testing statistical methods or generating data for scenarios where real data is unavailable or limited.

# Statistical Historiography

- "Statistical historiography" and "Scientific in Data Visualization" are two distinct concepts in the fields of history and data visualization, respectively.
- **Statistical historiography** refers to the application of statistical methods and quantitative analysis to historical research and the study of historical events.
- This approach seeks to enhance our understanding of history by using data, numbers, and statistical techniques to analyze patterns, trends, and relationships within historical datasets.

# Key aspects of statistical historiography include:

- **Data Collection:** Gathering historical data, documents, and records is a fundamental step. Historians may digitize archival materials, collect data from historical sources, or use existing datasets.
- **Quantitative Analysis:** Statistical techniques such as regression analysis, time series analysis, and hypothesis testing are employed to analyze historical data. This allows historians to test hypotheses and draw conclusions based on empirical evidence.
- **Visualization:** Data visualization is often used to present historical data in a comprehensible and visually appealing manner. Charts, graphs, and maps can help historians and researchers communicate their findings effectively.
- **Interdisciplinary Approach:** Statistical historiography often involves collaboration between historians, statisticians, data scientists, and domain experts to combine historical expertise with statistical methodologies.
- **Historical Context:** Despite its quantitative focus, statistical historiography recognizes the importance of historical context and the need to interpret data within the framework of historical events, cultural influences, and societal changes.

# Scientific Data Visualization:

- **Scientific data visualization** refers to the use of visual representations to explore and communicate scientific data and complex information.
- It is a crucial tool in various scientific disciplines, enabling researchers to analyze, understand, and present their findings effectively.

# Key aspects of scientific data visualization include:

- **Data Representation:** Scientific data, which can be multidimensional and complex, is represented visually using various types of charts, graphs, diagrams, and maps. The choice of visualization depends on the nature of the data and the research objectives.
- **Exploratory Data Analysis:** Scientists use data visualization to explore datasets, detect patterns, outliers, and correlations, and generate hypotheses. Visualization tools allow for interactive exploration and manipulation of data.
- **Communication:** Visualizations serve as a powerful means of communicating research findings to both scientific peers and the broader public. Well-designed visuals can convey complex scientific concepts in an accessible manner.
- **Simulation and Modeling:** Scientific data visualization is often used in conjunction with simulations and models to visualize the behavior of systems, phenomena, and processes. This aids in hypothesis testing and predictive analysis.
- **3D and Multivariate Visualization:** In fields like biology, chemistry, geology, and engineering, 3D and multivariate data visualization techniques are employed to represent data in three dimensions and display relationships among multiple variables.
- **Big Data Visualization:** With the advent of big data in scientific research, advanced visualization techniques and high-performance computing are used to handle and visualize large and complex datasets.



# Higher-dimensional Displays and Special Structures

- Higher-dimensional displays and special structures in data visualization are techniques and approaches used to represent and make sense of data with more than three dimensions or data that has unique structural characteristics.
- Here are some common methods and considerations for visualizing higher-dimensional data and special structures:

## **1. Dimensionality Reduction:**

- When dealing with data that has many dimensions, consider using dimensionality reduction techniques like PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), or UMAP (Uniform Manifold Approximation and Projection) to project the data into lower-dimensional spaces while preserving essential patterns and relationships.

## **2. Scatterplot Matrices:**

- Scatterplot matrices are useful for visualizing relationships between multiple variables simultaneously. In a scatterplot matrix, each combination of variables is plotted against one another, allowing you to identify patterns, correlations, and outliers.

## **3. Parallel Coordinates:**

- Parallel coordinates plots are particularly effective for visualizing multivariate data. Each variable is represented by a vertical axis, and data points are connected by lines, making it easy to spot patterns and trends across multiple dimensions.

## **4. Hyperdimensional Visualization:**

- Techniques like hyperdimensional visualization and projection methods can be used to create 2D or 3D representations of high-dimensional data by mapping the data points into a lower-dimensional space while preserving key characteristics.

## **5. Heatmaps:**

- Heatmaps are useful for visualizing data with both row and column dimensions, such as gene expression data or correlation matrices. They use color intensity to represent values, making it easy to identify clusters or patterns.

## **6. Tree Maps and Sunburst Charts:**

- Tree maps and sunburst charts are excellent for displaying hierarchical or tree-like structures. Each level of the hierarchy is represented by nested rectangles or segments, allowing you to explore the data's structure and distribution.

## **7. Network Graphs:**

- For data with network or graph structures, network graphs can help visualize connections and relationships between nodes. Techniques like force-directed layouts can help untangle complex networks.

## **8. Chernoff Faces and Star Plots:**

- Chernoff faces and star plots are unconventional techniques that use visual attributes like facial features or radial axes to represent multivariate data. They can be effective for certain types of data, such as representing characteristics of individuals.

## **9. Time Series Visualization:**

- When dealing with temporal data, consider using time series visualization techniques, such as line charts, stacked area charts, or Gantt charts, to display how data changes over time.

## **10. Color and Interaction:**

- Use color and interactivity thoughtfully. Color can represent additional dimensions or categories, but it should be chosen carefully to avoid confusion. Interactivity, such as tooltips or filtering, can help users explore and analyze the data in higher-dimensional displays.

## **11. User-Centered Design:**

- Always consider your target audience and their needs when designing higher-dimensional displays. Ensure that the visualization is intuitive and provides meaningful insights without overwhelming the viewer.

## **12. Evaluation and Feedback:**

- Regularly evaluate the effectiveness of your higher-dimensional displays by soliciting feedback from domain experts or end-users. Make adjustments based on their input to improve the visual representation of the data.

# Static Graphics: Complete Plots, Customization, Extensibility,

- Static graphics, such as plots and charts, are visual representations of data that do not involve user interactivity or animations. Creating effective static graphics involves considerations related to completeness, customization, and extensibility.

# Complete Plots:

- **Data Representation:** Ensure that your plot accurately represents the underlying data. Verify that the data points are correctly plotted and that there are no omissions or duplications.
- **Axes and Labels:** Provide clear and appropriately labeled axes with units of measurement. Include axis titles, tick marks, and gridlines where relevant to aid in data interpretation.
- **Legend:** Include a legend if your plot contains multiple data series or categories. The legend should describe what each element in the plot represents.
- **Title and Caption:** Include a concise and informative title for your plot. Consider adding a caption or additional context to help the viewer understand the significance of the data.

# Customization:

- **Color Scheme:** Choose a color scheme that complements your data and enhances readability. Use color effectively to distinguish data categories or highlight key points.
- **Style Elements:** Customize line styles, markers, and data point shapes to differentiate between multiple data series or data points. Adjust their size and style for emphasis.
- **Fonts and Typography:** Carefully select fonts for labels, titles, and annotations. Ensure that the text is legible and appropriately sized. Utilize font weights and styles to create visual hierarchy.
- **Background and Borders:** Customize the background color, borders, and padding of your plot to align with the overall design or document where it will be placed.
- **Annotations:** Add annotations, arrows, and labels to draw attention to specific data points, trends, or events. Annotations should enhance understanding without cluttering the plot.

# Extensibility:

- **Output Formats:** Save your static graphic in multiple formats (e.g., PNG, JPEG, SVG) to ensure compatibility with various documents, presentations, and platforms.
- **Vector Graphics:** Whenever possible, use vector graphics (e.g., SVG) for scalability without loss of quality, especially if the graphic may need to be resized for different purposes.
- **Template or Style Guide:** Create templates or style guides that define the consistent design elements for all your static graphics. This promotes a uniform look and maintains brand or project consistency.
- **Automation:** Develop scripts or workflows for generating static graphics, especially if you need to create many similar plots with varying data. Automation can save time and reduce errors.

# User-Centric Customization:

- Consider allowing users to customize certain aspects of the plot, such as color schemes or data ranges. Interactive tools or options can enhance the viewer's experience.
- Provide users with the ability to download or save the static graphic in their preferred format or resolution.



# Testing and Documentation:

- Test your static graphic with potential viewers or colleagues to gather feedback on clarity and effectiveness.
- Document the data sources, methodologies, and any assumptions made during the creation of the graphic. Clear documentation enhances transparency and trustworthiness.

# 3-D Plots

Creating 3D plots is a valuable technique for visualizing and analyzing data in three dimensions, adding depth and complexity to your visual representations.

Here's a guide on how to create 3D plots:

## 1. Choose a Suitable Library or Tool:

- Depending on your programming language of choice and specific needs, there are various libraries and tools available for creating 3D plots. Some popular options include:
  - **Matplotlib** (Python): Offers versatile 3D plotting capabilities.
  - **Plotly** (Python, JavaScript): Provides interactive 3D visualizations.
  - **ggplot2** (R): Useful for creating 3D plots in R.
  - **D3.js** (JavaScript): Allows for highly customizable 3D graphics.

## 2. Prepare Your Data:

- Ensure your data is organized in a way that's compatible with 3D plotting. In most cases, you'll have a set of x, y, and z values representing the coordinates of points in 3D space.
- If your data isn't in this format, consider transforming or reshaping it to fit the requirements.

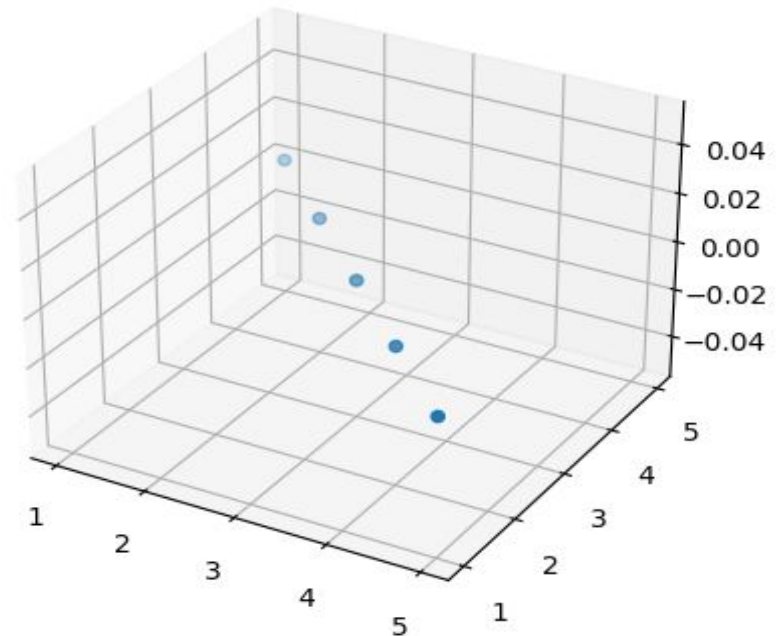
- **Basic 3D Scatter Plot:**

A simple starting point for 3D plotting is a scatter plot. In Matplotlib (Python), for example, you can create one using the **scatter** function.

```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
# Create a figure and a 3D axis
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
# Generate example data
x = [1, 2, 3, 4, 5]
y = [5, 4, 3, 2, 1]
z = [0, 0, 0, 0, 0]

# Create the 3D scatter plot
ax.scatter(x, y, z)

plt.show()
```



# Speed:

- **Optimized Algorithms:** Choose plotting libraries and tools that implement optimized algorithms for 3D rendering. Efficient algorithms can significantly improve the speed of generating 3D plots.
- **Data Reduction:** For large datasets, consider data reduction techniques like downsampling or aggregation to reduce the number of data points plotted. This can significantly improve rendering speed.
- **Caching:** Use caching mechanisms to store pre-rendered 3D plots, especially if the data or plot configuration doesn't change frequently. This can help reduce computation time for repeated access.

# Output Formats:

- **Vector vs. Raster Graphics:** Decide whether vector or raster graphics are more suitable for your needs. Vector formats (e.g., SVG) are scalable without loss of quality, making them ideal for print or high-resolution displays. Raster formats (e.g., PNG, JPEG) are better for images with complex shading and textures but may suffer from quality loss when scaled up.
- **Resolution:** Adjust the resolution of the output to match the intended use. Higher resolutions are suitable for printing, while lower resolutions work well for web or screen display.

# Data Handling:

- **Data Preparation:** Ensure that your data is in a suitable format for 3D plotting. Organize it into arrays or matrices that can be easily processed by your chosen plotting library or tool.
- **Data Scaling:** Scale your data appropriately to fit within the plotting space. This ensures that your 3D plot is not distorted or elongated along any axis.
- **Data Interpolation:** In cases where data points are sparse or irregularly spaced, consider using interpolation techniques to create a smooth 3D surface or mesh.
- **Data Filtering:** If your dataset contains noise or outliers, apply data filtering or smoothing techniques to improve the visual clarity of the 3D plot.
- **Data Exploration:** Use interactive tools to allow users to explore 3D data more effectively. Features like zooming, panning, and rotating the plot can aid in understanding complex 3D structures.

- When creating 3D plots, it's essential to strike a balance between visual richness and performance. Be mindful of the complexity of your data, the capabilities of your hardware, and the expectations of your audience. Efficient data handling and output optimization can lead to more responsive and visually appealing 3D visualizations.

**Thank You**