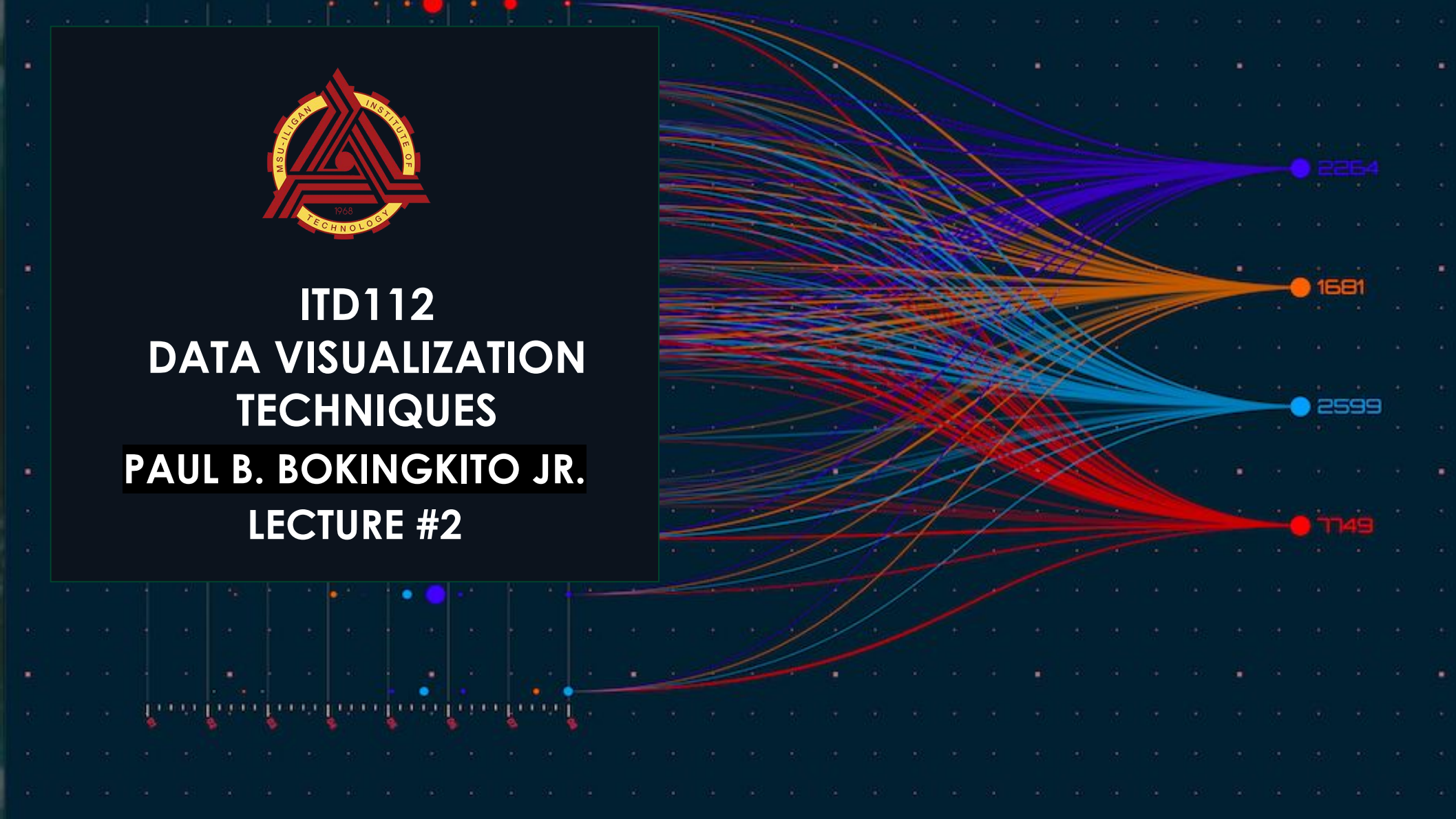


ITD112
DATA VISUALIZATION TECHNIQUES

PAUL B. BOKINGKITO JR.

LECTURE #2

Principles of Effective Visualization

Clarity
Keep It Simple

- Eliminate "chart junk" - unnecessary decorative elements
- Focus on the data story, not flashy graphics
- Use clear, readable fonts and appropriate sizing
- Principle: If it doesn't help communicate the data, remove it

Accuracy
Represent Data Truthfully

- Start axes at zero when appropriate
- Use consistent scales across comparisons
- Maintain proportional relationships
- Show uncertainty and limitations when relevant

Ethics
Avoid Bias and Manipulation

- Present complete context, not just favorable data
- Avoid cherry-picking time periods or data points
- Use neutral colors and design choices
- Acknowledge data sources and methodology

Perception
How Humans Process Visual Information

- Shape - Circles vs. squares vs. triangles
- Size - Larger objects draw attention first
- Color - Bright colors stand out, similar colors group together
- Position - Higher/rightward positions often perceived as "more"

3 simple step to choosing the right chart.

Data Types Classification

1. Categorical data - qualitative data representing distinct groups or categories with no numerical meaning.

🏷️ **Nominal**

No natural order or ranking between categories

📈 **Ordinal**

Natural order or hierarchy exists

Gender | **Colors** | **Education Level** | **Product types** | **Survey Ratings** | **Departments**

Data Types Classification

2. Numerical Data - Quantitative data that can be measured and calculated

🎯 **Discrete**

Countable, finite values (usually whole numbers)

〰️ **Continuous**

Infinite possible values within a range

Height **Sales Revenue** **Student Count** **Temperature** **Age** **Response Time**

Data Types Classification

3. Time-Series Data - Temporal data showing changes and trends over time

📅 **Regular Intervals**
Consistent time gaps between measurements

🎲 **Irregular Intervals**
Varying time gaps between observations

Stock Prices	Website user sessions	Weather Data	Population Growth	Earthquake occurrences	Email arrivals

Data Types Classification

4. Spatial Data - Geographic data representing locations and spatial relationships

📍 **Point Data**

Specific coordinates or location markers

🗺️ **Area Data**

Regions, boundaries, or polygonal areas

Population Density	Crime Incidents	Earthquake Epicenters	Weather Patterns	Air Quality by City	Hospital Locations

3 simple step to choosing the right chart.

Identify the Purpose/Objective of your Visualization

a. **Comparison** - Showing differences or similarities

b. **Composition** - Displaying how parts make up a whole or showing proportions

c. **Distribution** - Revealing how data points are spread across a range of values or showing frequency patterns

d. **Relationship** - Illustrating connections, correlations, or associations

e. **Trend** - Displaying changes or patterns over time or sequence

f. **Ranking** - Ordering items from highest to lowest or showing hierarchical positions

g. **Flow/Process** - Mapping movement, progression, or sequential steps through a system or workflow

Common tools of Descriptive Statistics

- Measures of central tendency
- Measures of dispersion or spread

Measures of central tendency tell us where most values fall, or the typical value of a data set.

A	B	C	D	E
5	4	3	5	8

Mode is the value that shows up the most number of times in a data set.

A	B	C	D	E
5	4	3	5	8

Mean, or average, is the sum of all the numbers divided by the number of elements in the sample

5+4+3+5+8 = 25

25/5 = 5

Median is the middle number of the data set.

A	B	C	D	E
5	4	3	5	8

A	B	C	D	E
3	4	5	5	8

Measures of dispersion or spread

Data Set A

A	B	C	D	E
3	4	5	5	8
+2	+1	+0	+3	

Data Set B

A	B	C	D	E
1	2	5	5	12
+1	+3	+0	+7	

Common measures of dispersion

- Range
- Variance
- Standard deviation

Range is the difference between the high and low values.

Data Set A	A	B	C	D	E
	3	4	5	5	8

Range: 8 – 3 = 5

Data Set B	A	B	C	D	E
	1	2	5	5	12

Range: 12 – 1 = 11

Variance

Sample Variance

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Standard Deviation

Sample Standard Deviation

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

Data Set A

A	B	C	D	E
3	4	5	5	8

Mean = 5

Median = 5

Mode = 5

Range = 5

Variance = 3.5

Standard deviation = 1.87

Data Set B

A	B	C	D	E
1	2	5	5	12

Mean = 5

Median = 5

Mode = 5

Range = 11

Variance = 8.5

Standard deviation = 4.30

Variance measures how far a data set is spread out.

Standard deviation measures how far each observed value is from the mean.

If the standard deviation is small, the data values are close to the mean value. If it is high, the data values are widely spread out from the mean value.

Data Set A

Data Set B

1. Comparison - Analyzing differences and similarities between distinct categories, groups, or entities to determine which performs better, has higher values, or exhibits different characteristics.

Data Type	Variables	Objective	Example Question	Best Chart Type(s)
Categorical	1 Variable	Comparison	*"Which product category has the highest sales?" • "How do departments compare in performance?"*	**Bar Chart** · Column Chart · Pictogram Chart · Lollipop Chart · Bullet Graph
Categorical	2 Variables	Comparison	*"How do sales compare across regions and product types?" • "What's the performance by department and quarter?"*	**Multi-set Bar Chart** · Grouped Bar Chart · Stacked Bar Chart · Heatmap
Categorical	3 Variables	Comparison	*"How do sales compare across region, product, and quarter?" • "What's performance across three categorical dimensions?"*	**Small Multiples** · Faceted Charts · 3D Bar Chart · Treemap
Numerical	2 Variables	Comparison	*"How do test scores compare between schools by subject?" • "What's the performance across two numerical dimensions?"*	**Grouped Bar Chart** · Multi-set Bar Chart · Box & Whisker Plot · Violin Plot

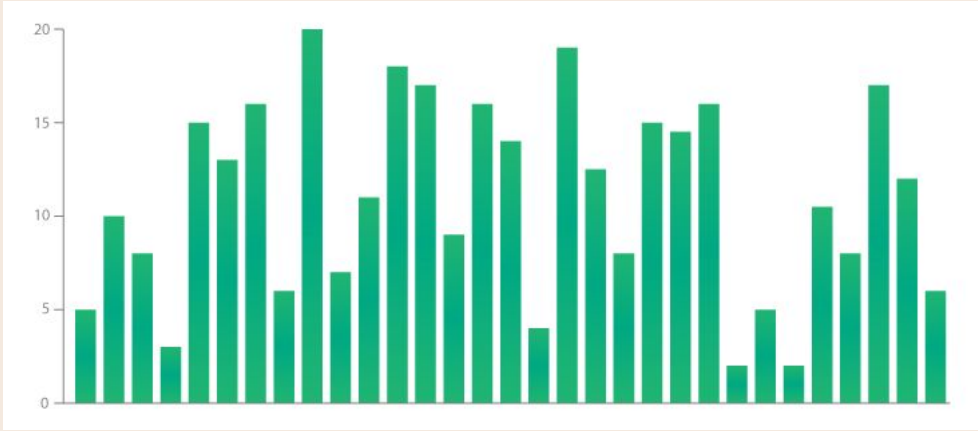
Bar Chart - Compare categorical data.

Also known as Bar Graph or Column Graph.

- A Bar Chart uses either horizontal or vertical bars (column chart) to show discrete, numerical comparisons across categories.

- Bar Charts are distinguished from *Histograms*, as they do not display continuous developments over an interval.

- Instead, **Bar Chart's discrete data is categorical** and therefore answers the question of **"how many?"** in each category.

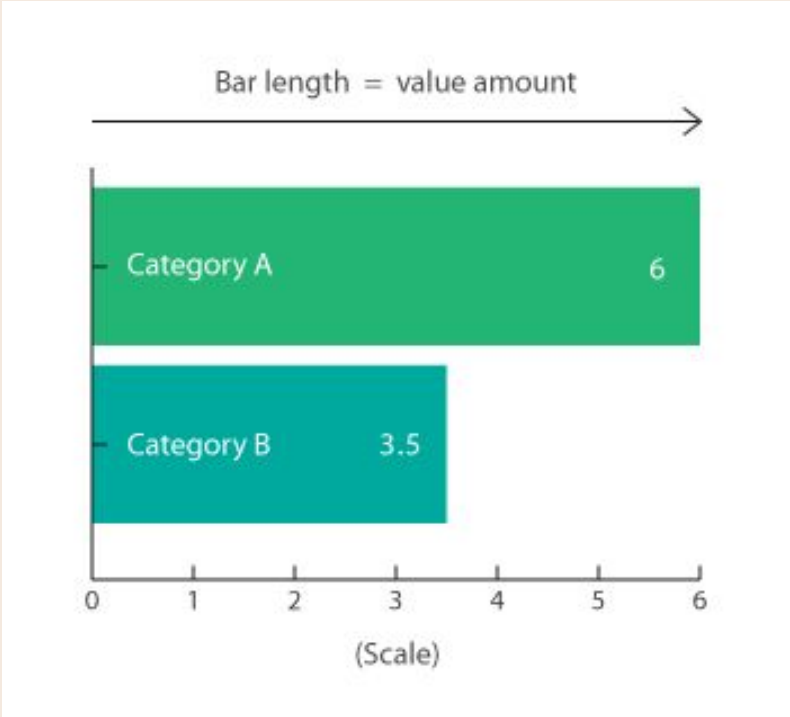


Bar Chart - Compare categorical data.

Also known as Bar Graph or Column Graph.

Design Practices

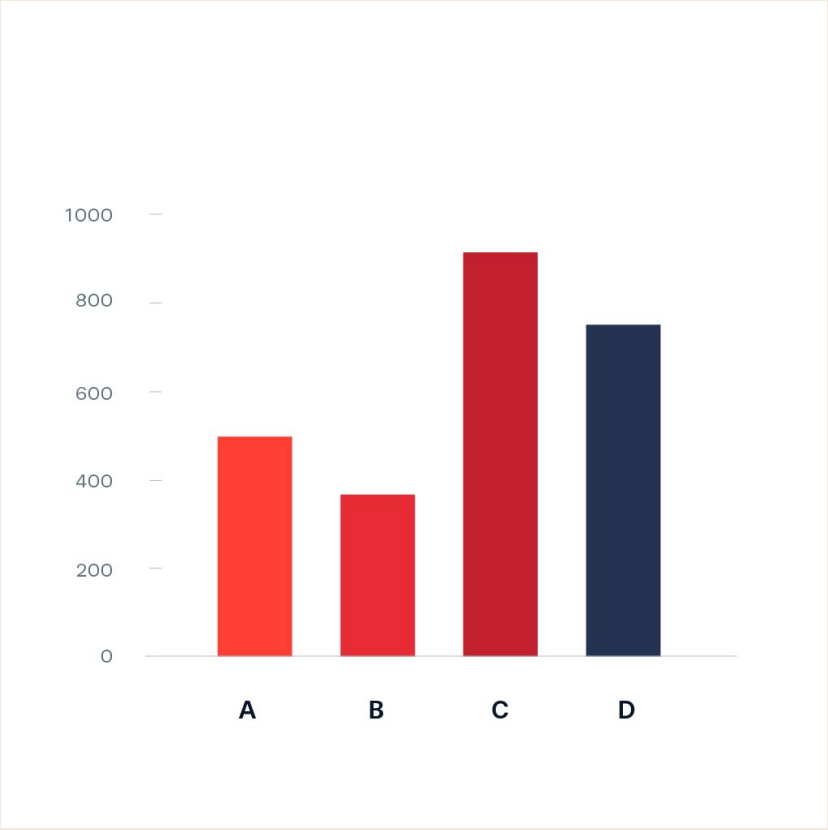
- **Sort Bars Logically** - Sort bars by value or logically.
- **Consistent Bar Widths** - Keep bar widths and spacing uniform.
- **Avoid Clutter -** Limit the number of bars; consider alternatives for large datasets.
- **Bar Orientation** - Use horizontal bars for long category names.
- **Axis Scaling -** Ensure the y-axis starts at zero to avoid distortion.



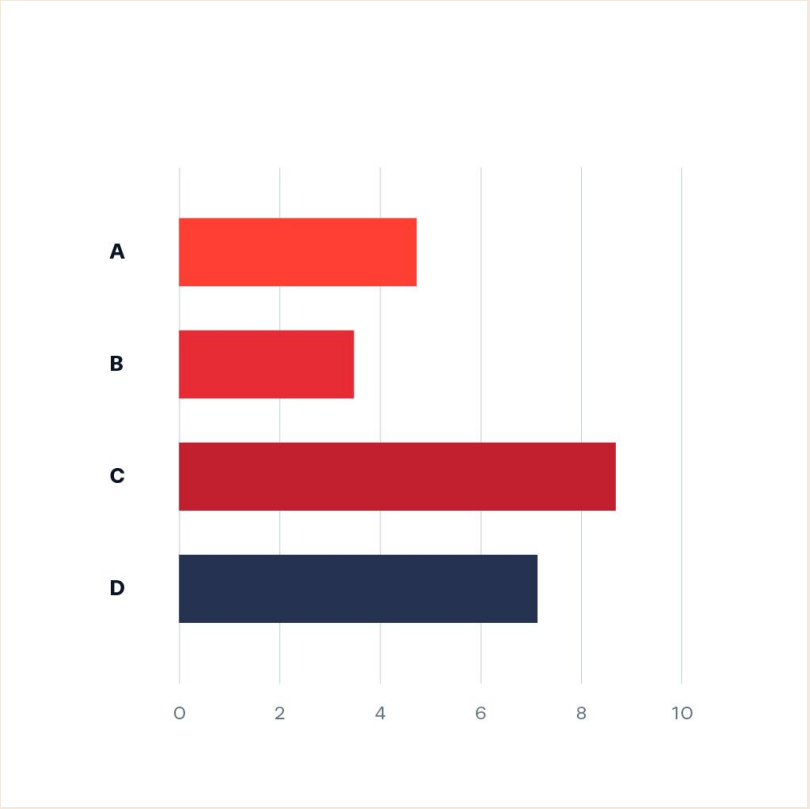
Bar Chart - Compare categorical data.

Also known as Bar Graph or Column Graph.

Vertical Bar chart

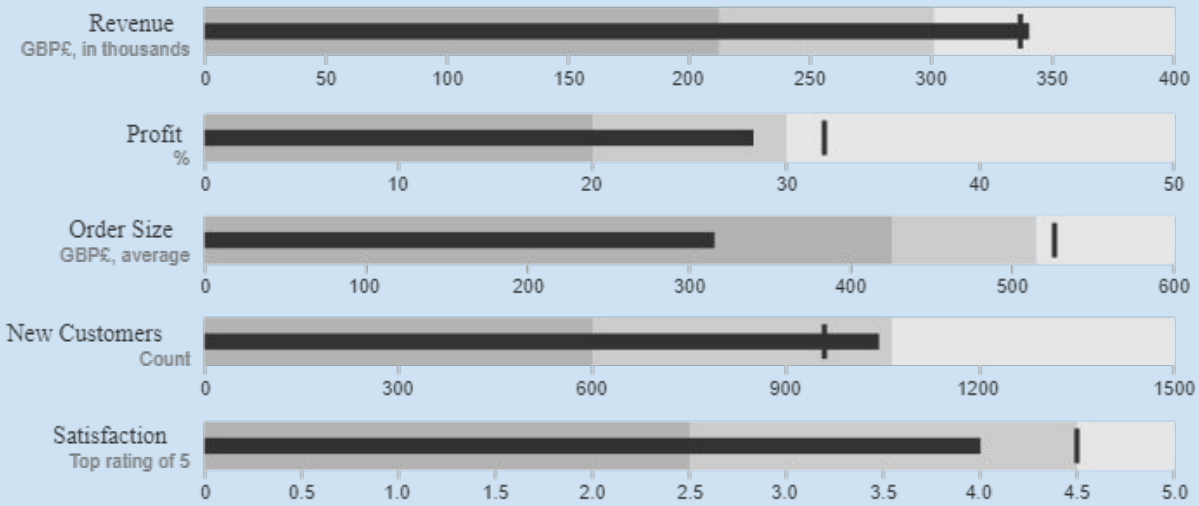


Horizontal Bar chart



Bullet Graph - Compare a single measure against a target or benchmark.

The main data value is encoded by length with the bar in the centre of the chart, which is known as the **Feature Measure.** The line marker that runs perpendicular to the orientation of the graph is known as the **Comparative Measure** and is used as a target marker to compare against the Feature Measure value. So if the main bar has passed the position of Comparative Measure, you know you've hit your goal.

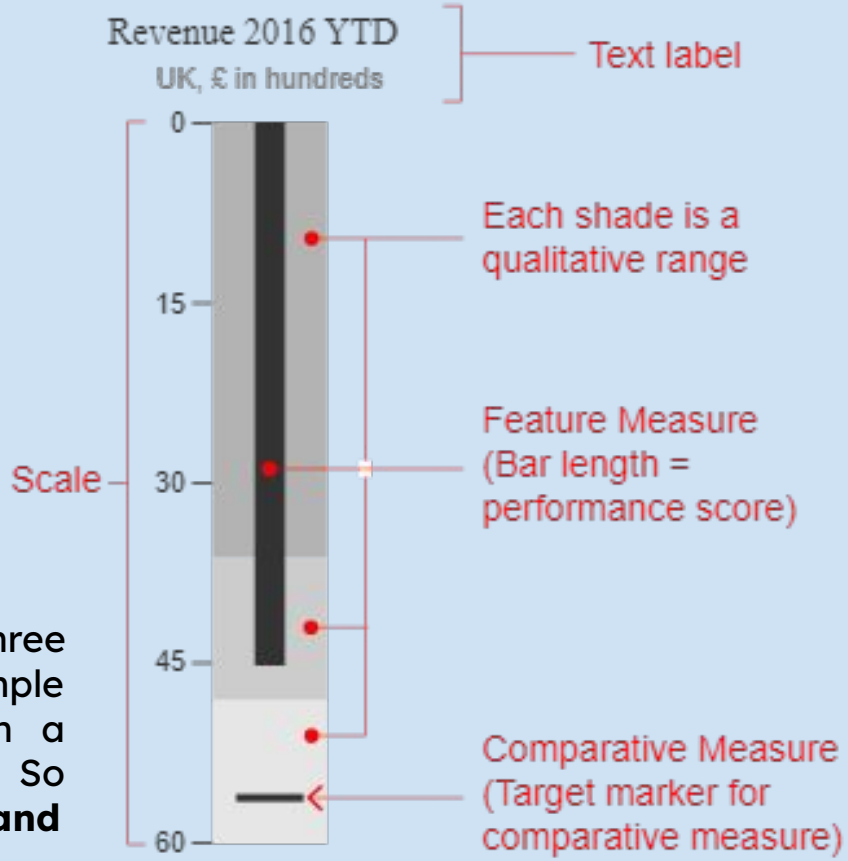


Bullet Graph - Compare a single measure against a target or benchmark.

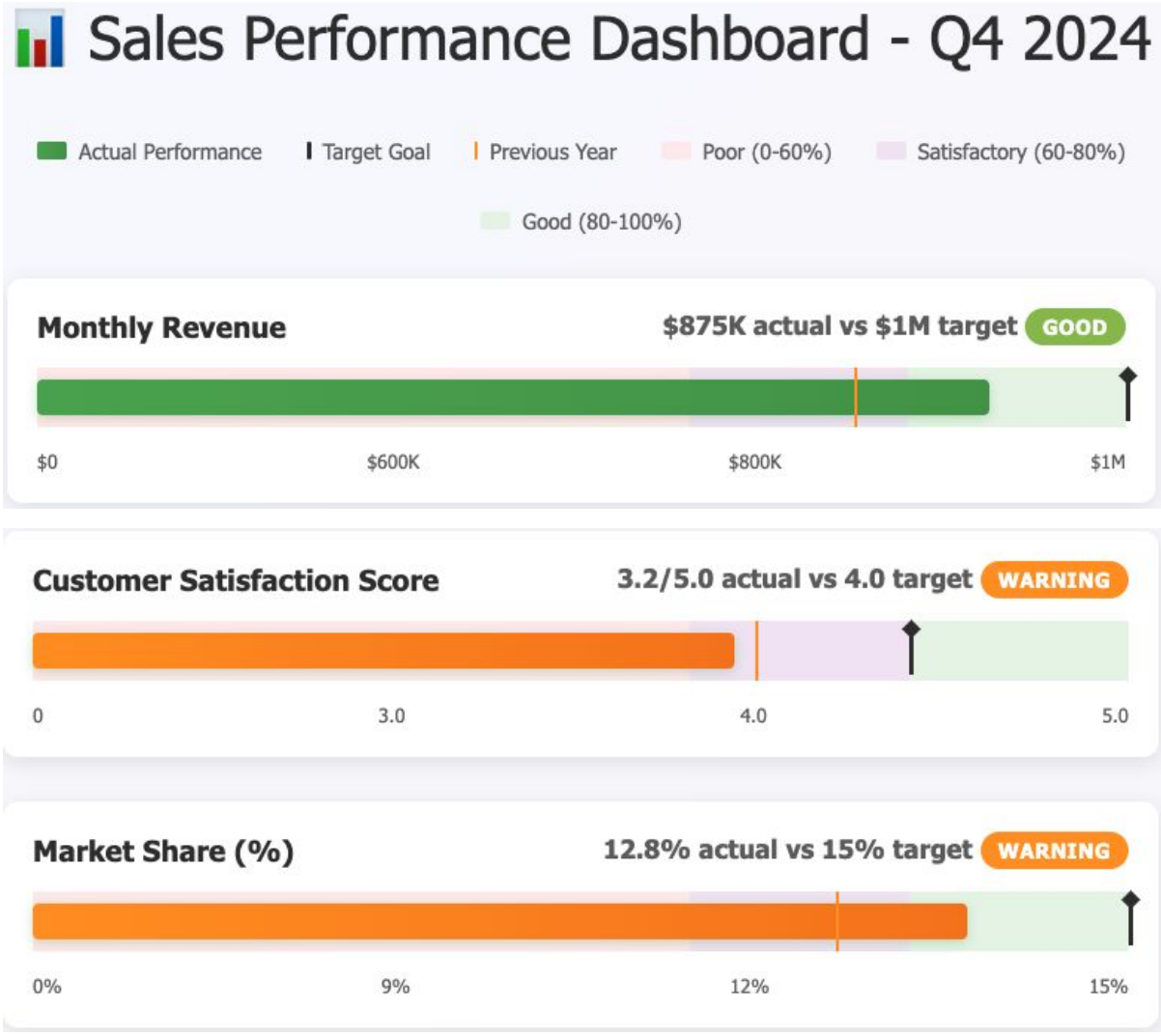
Design Practices

- **Use Color to Differentiate -** Show different performance ranges (e.g., poor, average, good).
- **Highlight Target -** Clearly mark the target value for comparison.
- **Keep it Simple -** Maintain a clean design to focus on the key performance indicators.

Each colour shade (the three shades of grey in the example above) are used to assign a performance range rating. So for example, **poor, average and great.**



Sales Performance Dashboard - Q4 2024



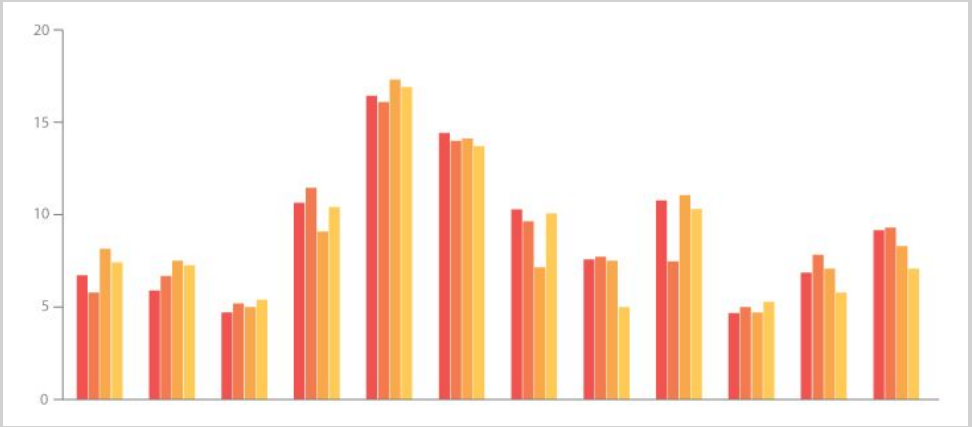
Multi-set Bar Chart - Compare multiple categories across different groups.

Also known as a Grouped Bar Chart or Clustered Bar Chart.

- **This variation of a Bar Chart** can be used when two or more data series need to be plotted all on the same axis and grouped into parent categories.

- Like on a Bar Chart, the length of each bar on a Multiset Bar Chart is used to show discrete, numerical comparisons amongst categories.

- Each bar for a data series is assigned a colour to distinguish them apart. Bars in the same group are placed together and are then spaced apart from other bar groupings.



Multi-set Bar Chart - Compare multiple categories across different groups.

Design Practices

- **Distinct Colors -** Use different colors or patterns for each set of bars within a group to ensure clarity and differentiation between subcategories.
- **Clear Grouping -** Ensure that bars within each group are positioned close together, with adequate spacing between groups, to make comparisons easier.
- **Legend or Labels -** Provide a clear legend or direct labeling to indicate what each bar represents.
- **Avoid Overcrowding -** Limit the number of groups and bars per group to prevent the chart from becoming cluttered and difficult to read.

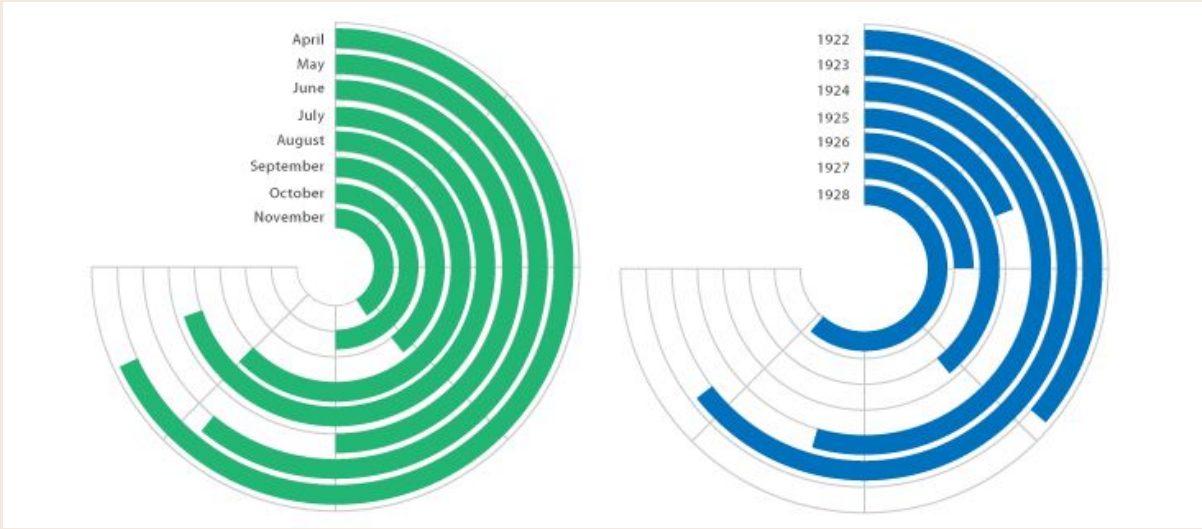
1. Comparison - Analyzing differences and similarities between distinct categories, groups, or entities to determine which performs better, has higher values, or exhibits different characteristics.

Time-Series	2 Variables	Comparison	*"How do sales trends compare between different regions over time?" • "How do multiple time series compare?"*	**Line Graph** · Multi-line Chart · Small Multiples · Dual-axis Chart
Spatial	2 Variables	Comparison	*"Which regions have highest sales vs population?" • "How do geographic areas compare on two metrics?"*	**Bubble Map** · Choropleth Map · Bivariate Map · Proportional Symbol Map
Spatial	3 Variables	Comparison	*"How do regions compare on population, income, and education?" • "What are multi-dimensional spatial relationships?"*	**Bubble Map** · Multivariate Map · Small Multiple Maps · Proportional Symbol Map
Mixed	3 Variables	Comparison	*"How do cities compare on population, income, and satisfaction?" • "How do entities perform across multiple mixed metrics?"*	**Bubble Chart** · Radar Chart · Parallel Coordinates · Scatterplot Matrix
Mixed	4+ Variables	Comparison	*"How do business units compare across revenue, costs, employees, and satisfaction?" • "What's multi-dimensional performance?"*	**Parallel Coordinates** · Radar Chart · Star Plot · Heatmap

Radial Bar Chart - Circular version of a bar chart to compare data points.

- A Radial Bar Chart is essentially a Bar Chart plotted on a polar coordinates system, rather than on a Cartesian one.

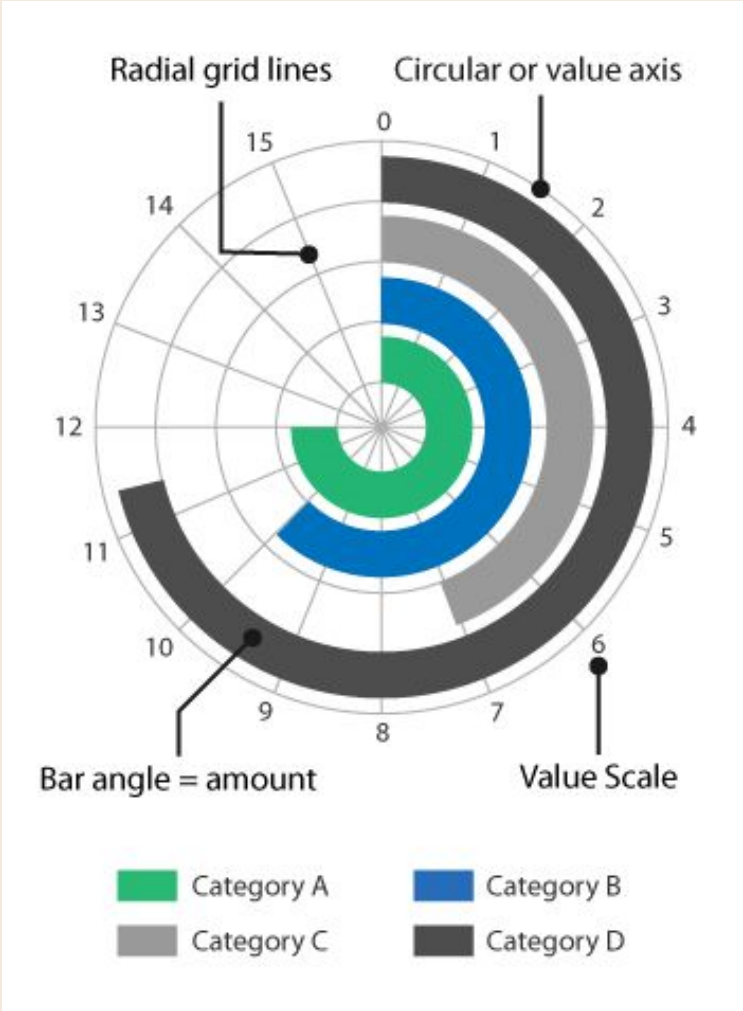
- The problem with Radial Bar Charts is that **the bar lengths can be misleading.** Each bar on the outside gets relatively longer than the previous bar, even if they represent the same value. This is because each bar is at a different radii.



Radial Bar Chart - Circular version of a bar chart to compare data points.

Design Practices

- **Clear Labels -** Ensure each bar and category is clearly labeled since the circular layout can make interpretation harder.
- **Limit Data Series -** Use a limited number of categories to avoid clutter and maintain readability.
- **Use for Emphasis -** Radial bar charts can emphasize specific data points, but they should be used carefully, as they can be harder to read compared to traditional bar charts.

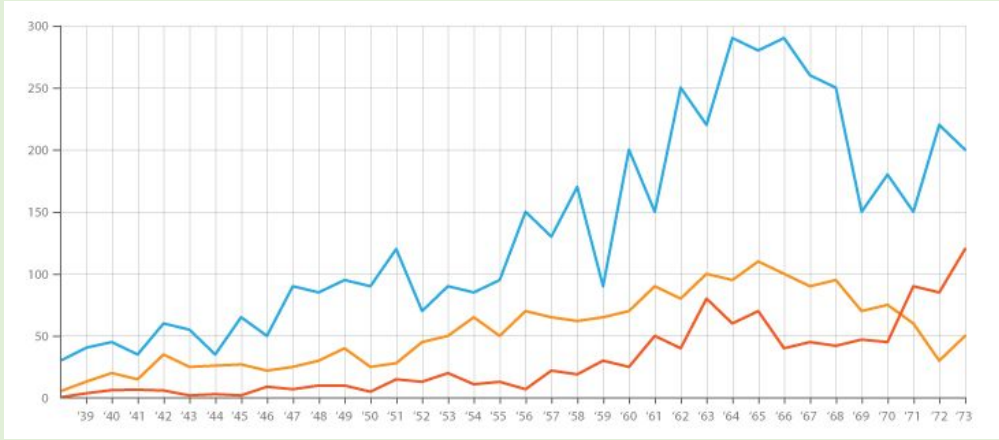


Line Chart - Compare trends over time for multiple series.

- This chart is used to **display quantitative values over a continuous interval or time period.** A Line Graph is most frequently used to **show trends** and analyse how the data has changed over time.

- Line Graphs are drawn by first plotting data points on a Cartesian coordinate grid, and then connecting a line between all of these points.

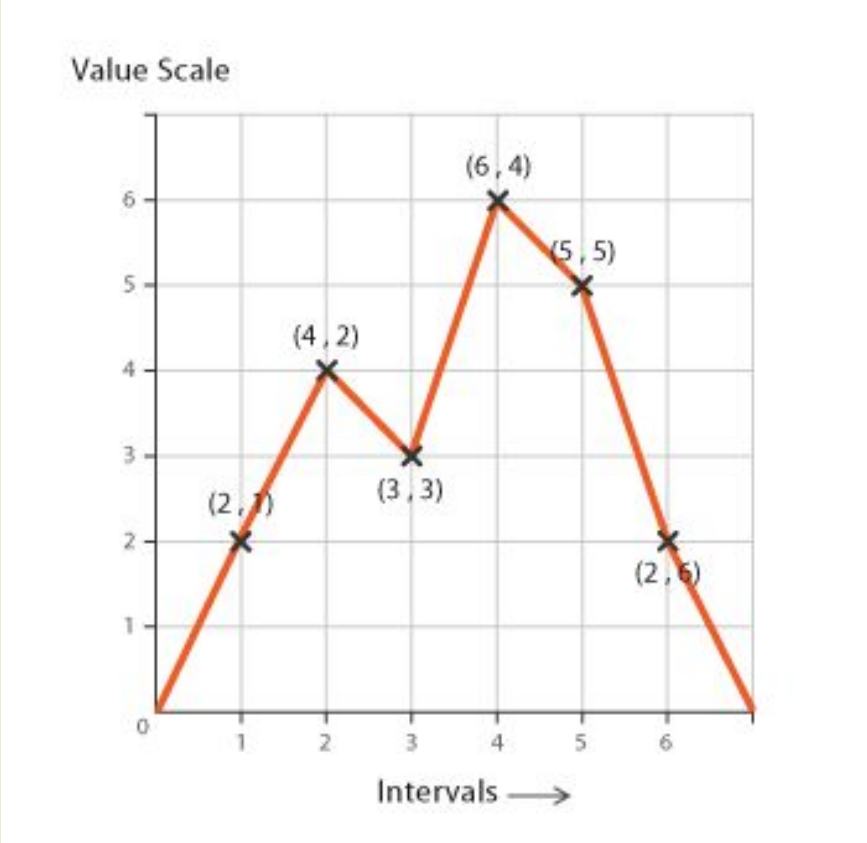
- Typically, the **y-axis has a quantitative value**, while the **x-axis is a timescale** or a sequence of intervals. Negative values can be displayed below the x-axis.



Line Chart - Compare trends over time for multiple series.

Design Practices

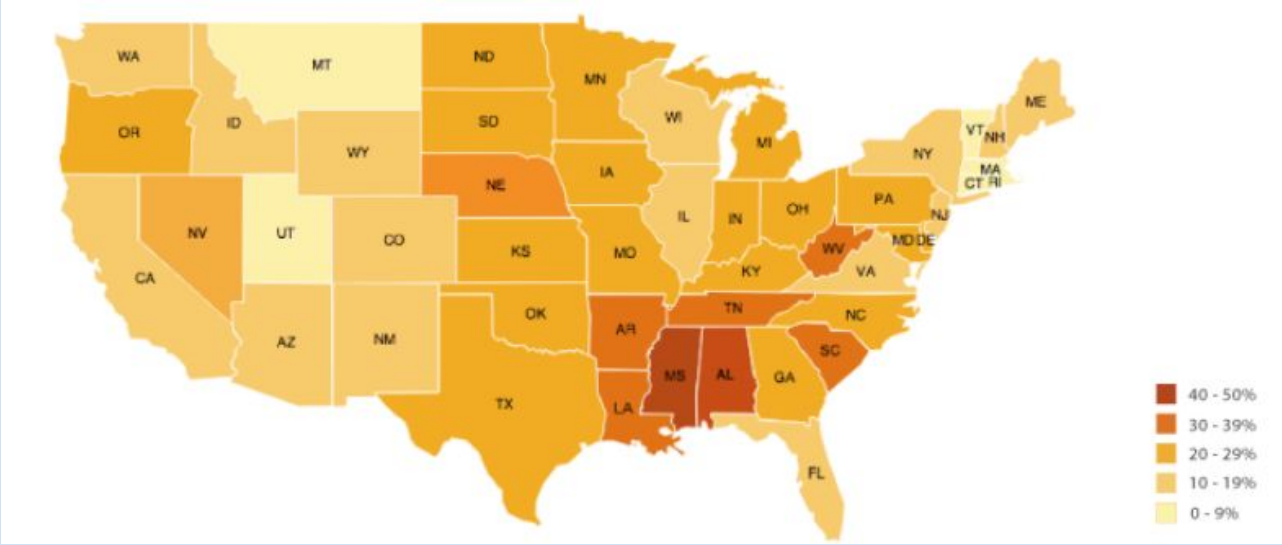
- **Distinct Lines -** Use different colors or styles for clarity.
- **Clear Labels -** Label x-axis (time) and y-axis (values) clearly.
- **Highlight Trends -** Use annotations or markers to emphasize significant trends.
- **Avoid Clutter -** Limit the number of lines and avoid overcrowding.
- **Appropriate Intervals** - Choose time intervals that best represent the data.

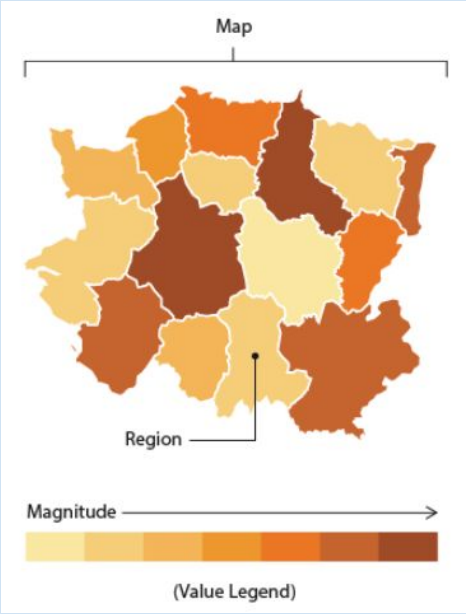


Choropleth Map

- Choropleth Maps display divided geographical areas or regions that are coloured, shaded or patterned in relation to a data variable.

- The data variable uses colour progression to represent itself in each region of the map.





Choropleth Map

A common error when producing Choropleth Maps is to encode raw data values (such as population) rather than using normalized values (calculating population per square kilometre for example) to produce a density map.

1. **Population Density:**

   Instead of mapping raw population numbers, normalize by dividing the population by the area of the region to get population per square kilometer.

   - **Raw data:**
     - City A population: 1,000,000
     - City B population: 500,000
     - City A area: 250 km²
     - City B area: 50 km²
   - **Normalized value** (population density):
     - City A: $\frac{1,000,000}{250} = 4000$ people per km²
     - City B: $\frac{500,000}{50} = 10,000$ people per km²

Choropleth Map

A common error when producing Choropleth Maps is to encode raw data values (such as population) rather than using normalized values (calculating population per square kilometre for example) to produce a density map.

2. **GDP per Capita**:
   Instead of showing raw GDP numbers, divide the total GDP by the population to show wealth per person.
   - **Raw data**:
     - Country A GDP: $1 trillion
     - Country B GDP: $500 billion
     - Country A population: 100 million
     - Country B population: 25 million
   - **Normalized value** (GDP per capita):
     - Country A: $\frac{1{,}000{,}000{,}000{,}000}{100{,}000{,}000} = 10{,}000$ USD
     - Country B: $\frac{500{,}000{,}000{,}000}{25{,}000{,}000} = 20{,}000$ USD

Choropleth Map

A common error when producing Choropleth Maps is to encode raw data values (such as population) rather than using normalized values (calculating population per square kilometre for example) to produce a density map.

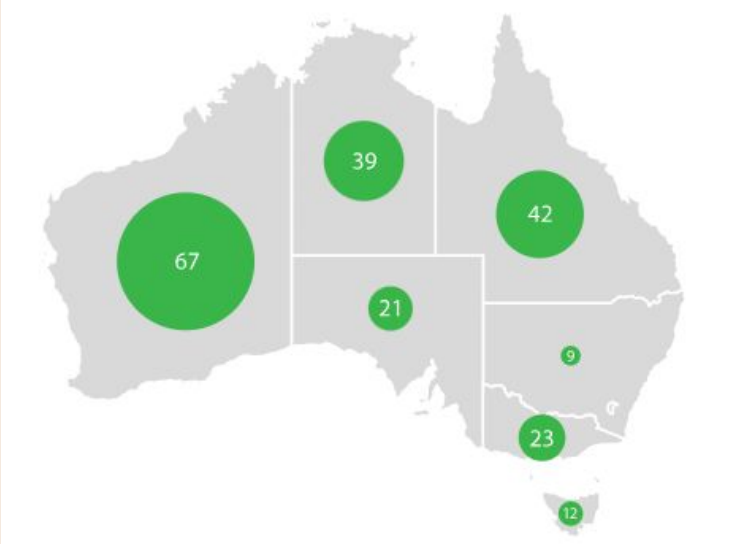
3. **Crime Rate per 1,000 People**:
   Normalize the total number of crimes by the population size to compare crime rates between regions.
   - **Raw data**:
     - City A crimes: 2,000
     - City B crimes: 1,000
     - City A population: 200,000
     - City B population: 50,000
   - **Normalized value** (crime rate per 1,000 people):
     - City A: $\frac{2000}{200,000} \times 1000 = 10$ crimes per 1,000 people
     - City B: $\frac{1000}{50,000} \times 1000 = 20$ crimes per 1,000 people

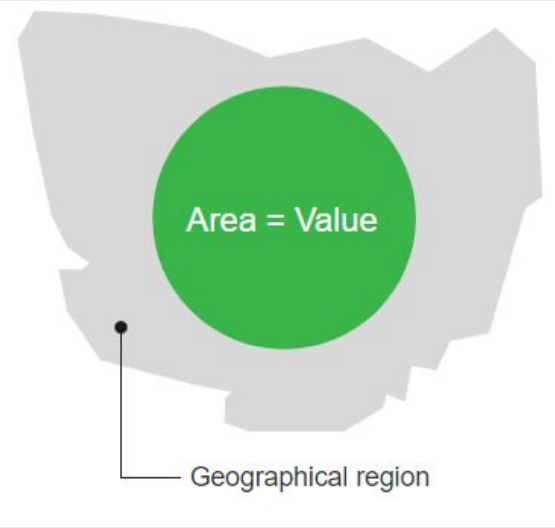
Bubble Map

- With this data map, circles are displayed over a designated geographical region with the area of each circle being proportional to its value in the dataset.

- Bubble Maps are good for comparing proportions over geographic regions without the issues caused by regional area size, as seen on **Choropleth Maps.**

- However, a major flaw with Bubble Maps is that overly large bubbles can overlap other bubbles and regions on the map, so this needs to be accounted for.



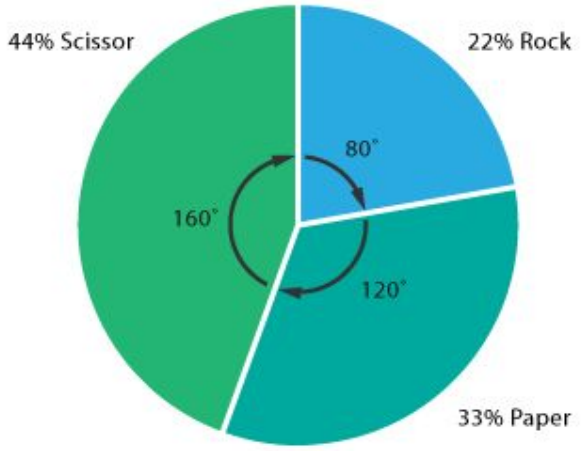


2. Composition - Showing how individual parts contribute to a whole, revealing the proportional makeup of a dataset and understanding the relative size of each component.

Data Type	Variables	Objective	Example Question	Best Chart Type(s)
Categorical	1 Variable	Composition	*"What percentage of budget goes to each department?" • "How is our customer base distributed by segment?"*	**Pie Chart** · Donut Chart · Treemap · Nightingale Rose Chart
Categorical	2 Variables	Composition	*"How does budget allocation vary by department and project type?" • "What's the breakdown across two categorical dimensions?"*	**Stacked Bar Graph** · Marimekko Chart · Mosaic Plot · Sunburst Diagram
Categorical	Special	Composition	*"What's the hierarchical structure of our organization?" • "How are nested categories organized?"*	**Sunburst Diagram** · Circle Packing · Tree Diagram · Icicle Chart
Time-Series	2 Variables	Composition	*"How has revenue breakdown by category evolved over time?" • "How do proportions change temporally?"*	**Stacked Area Graph** · Stream Graph · Stacked Bar Graph · Alluvial Diagram

Pie Chart - show proportions and percentages between categories, by dividing a circle into proportional segments.

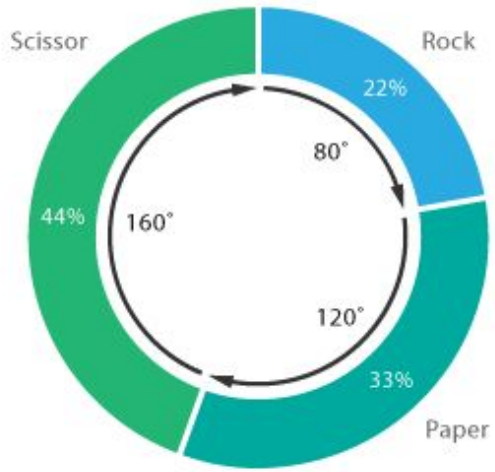
- Pie Charts are ideal for giving the reader a quick idea of the proportional distribution of the data.
- Each arc length represents a proportion of each category, while the full circle represents the total sum of all the data, equal to 100%.
- Conventionally, the **largest slice is placed starting at the 12 o'clock position**, with subsequent slices arranged in a clockwise direction. This makes interpretation more intuitive.



Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9=22%	3/9=33%	4/9=44%	100%
Degrees for each "pie slice"			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

Donut Chart - is essentially a Pie Chart but with the area of the centre cut out.

- A Donut Chart partly addresses this problem by de-emphasising the use of area, to make the viewer focus more on the changes in overall values.
- You are focused on reading the length of the arcs, rather than comparing the proportions between slices.
- Like pie charts, donut charts work best with a limited number of slices (usually no more than 6-8). Too many slices can make the chart difficult to read.

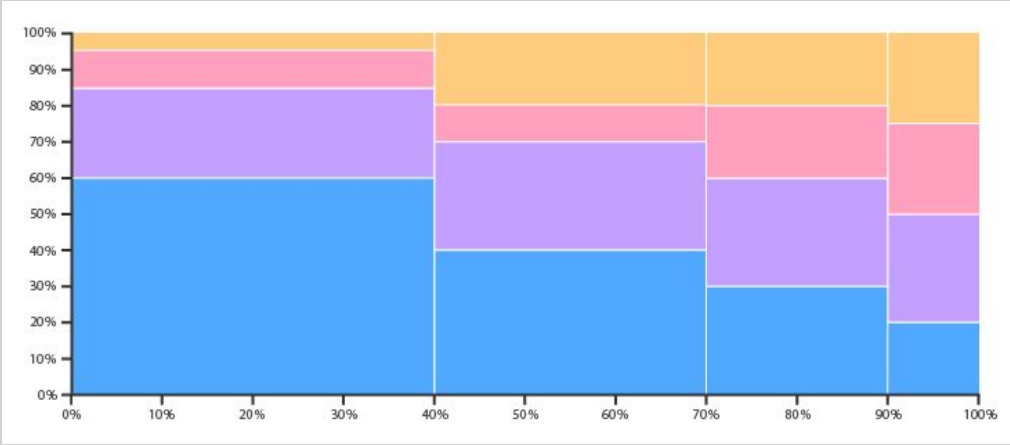


Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9 = **22%**	3/9 = **33%**	4/9 = **44%**	100%
Degrees for each "donut slice"			
(2/9) x 360 = **80°**	(3/9) x 360 = **120°**	(4/9) x 360 = **160°**	360°

Marimekko Chart - are used to visualise categorical data over a pair of variables.

Also known as a Mosaic Plot.

- Marimekko Charts are used to visualise categorical data over a pair of variables.
- In a Marimekko Chart, both axes are variables with a percentage scale, that determines both the width and height of each segment.
- So Marimekko Charts work as a kind of two-way 100% Stacked Bar Graph. This makes it possible to detect relationships between categories and their subcategories via the two axes.

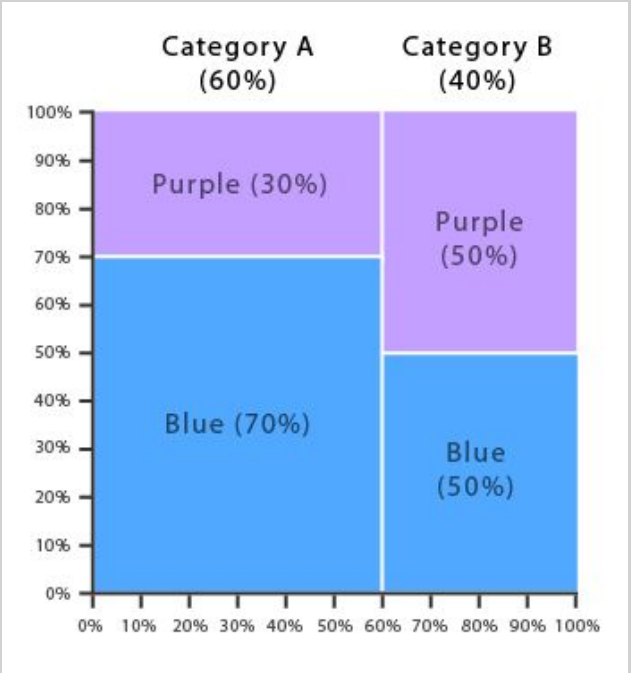


Marimekko Chart - are used to visualise categorical data over a pair of variables.

Also known as a Mosaic Plot.

How to Interpret a Marimekko Chart

- **Bar Width (x-axis)** - The width of each bar represents the proportion of one category relative to the entire dataset (e.g., market share by region).
- **Segment Height (y-axis)** - Each bar is divided into segments, and the height of each segment represents the proportion of subcategories within that main category (e.g., sales within a region broken down by product type).
- **Segment Area -** The area of each segment (width × height) reflects the overall contribution of that category and subcategory combination to the total dataset.



Global Smartphone Market Analysis

Marimekko Chart showing Market Share by Brand across Different Regions (2024)

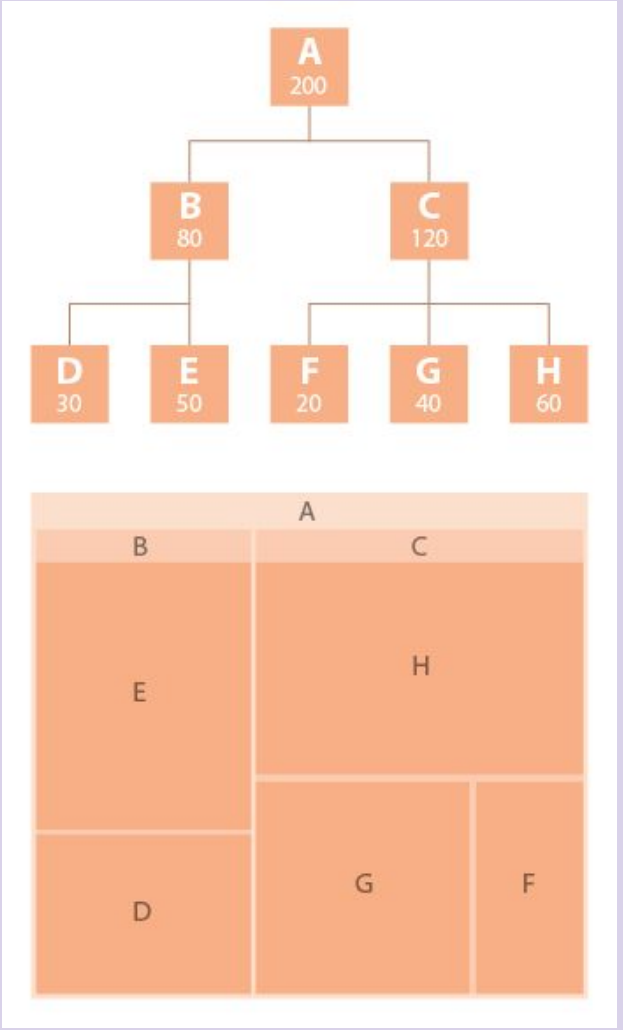
North America	Europe	China	India	Rest of Asia	Other Regions
$120B (27.3%)	$85B (19.3%)	$95B (21.6%)	$45B (10.2%)	$65B (14.8%)	$35B (8.0%)

Apple Samsung Xiaomi Oppo Vivo Others

Treemap - used to display hierarchical data using nested rectangles.

- Each rectangle represents a category, and its size is proportional to the data value it represents.
- Treemaps are particularly effective for visualizing parts-to-whole relationships and are used when displaying large amounts of data in a compact and space-efficient manner.

Treemap - used to display hierarchical data using nested rectangles.

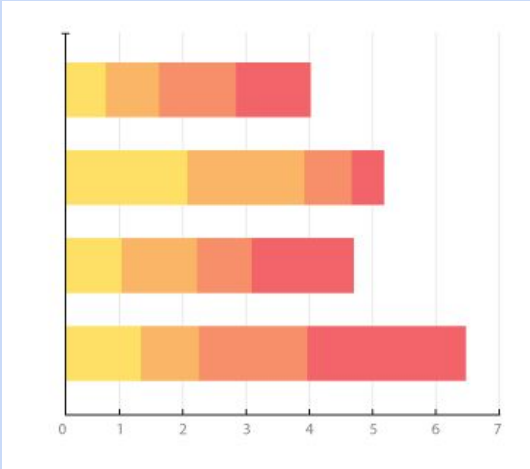


How to Read a Treemap

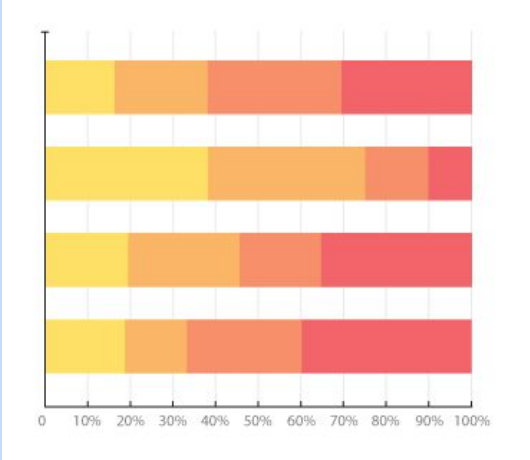
- **Rectangles -** Each rectangle represents a category or subcategory. The area of the rectangle corresponds to the value of that category.
- **Hierarchy -** Treemaps can display multiple levels of hierarchy by nesting smaller rectangles (subcategories) within larger rectangles (parent categories).
- **Color -** Often, treemaps use color to represent additional information, such as performance, growth, or type of category. For example, shades of green could indicate positive growth, while shades of red could indicate decline.

Stacked Bar Graph - Compare total values and their breakdown into subcomponents.

- **Simple Stacked Bar Graphs** place each value for the segment after the previous one. The total value of the bar is all the segment values added together. Ideal for comparing the total amounts across each segmented bar.



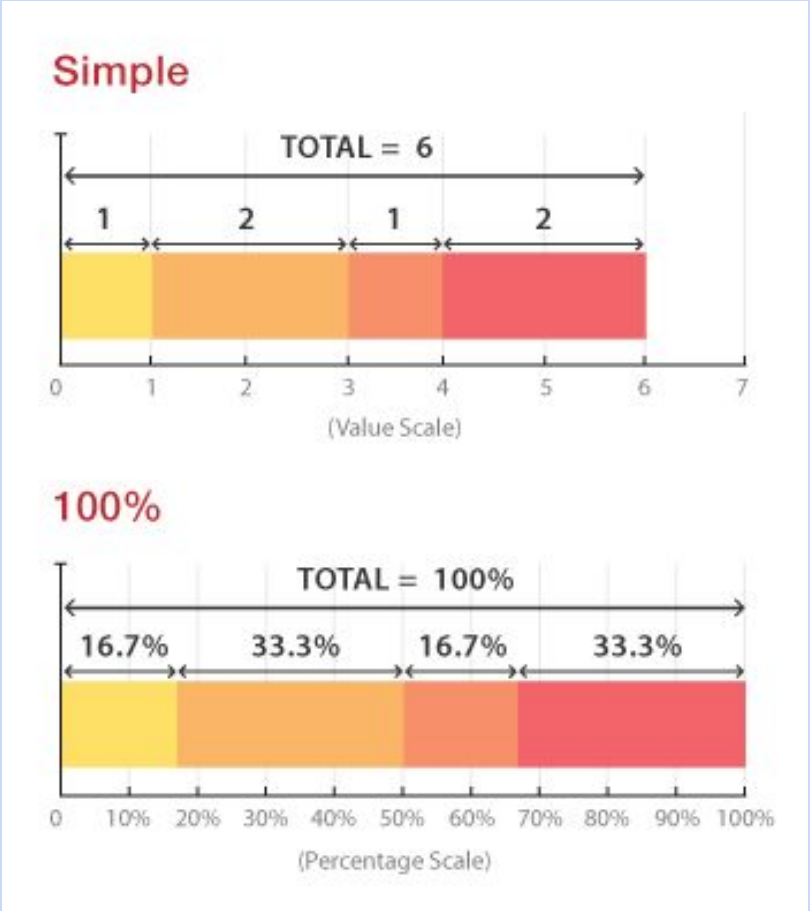
- **100% Stack Bar Graphs** show the percentage-of-the-whole by plotting the percentage of each value to the total amount in each group. This makes it easier to see the relative differences between quantities in each group.



Stacked Bar Graph - Compare total values and their breakdown into subcomponents.

Design Practices

- **Color Coding -** Use distinct colors for each subcomponent to ensure clear differentiation within the stack.
- **Limit Categories -** Limit the number of categories and subcomponents to avoid overcrowding, which can make the chart difficult to interpret.
- **Consistent Scale -** Ensure a consistent scale across all bars so that comparisons between totals and subcomponents are accurate.



Word Cloud - representation of text data, where the frequency or importance of words is represented by their size.

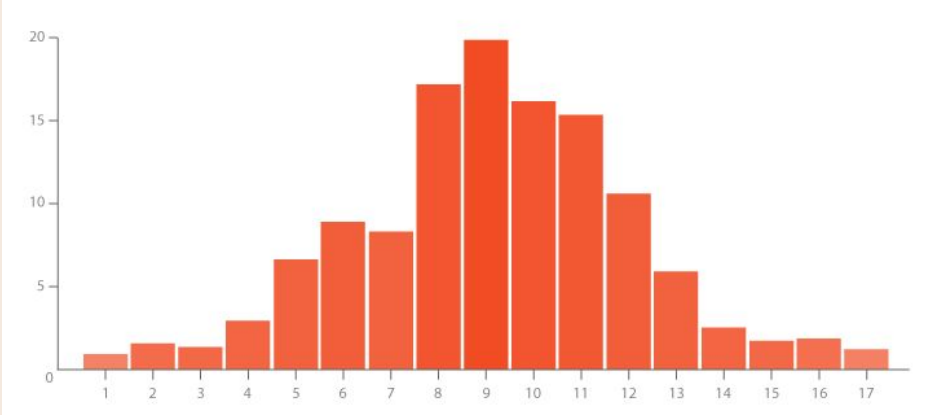
- In text analysis, word clouds highlight the most frequently used words, helping users quickly identify key themes, sentiments, or topics in the data.
- The size of each word in the cloud is proportional to its frequency or importance in the dataset. More frequent words appear larger.
- Words are placed randomly in various orientations, often centered, with larger words standing out.

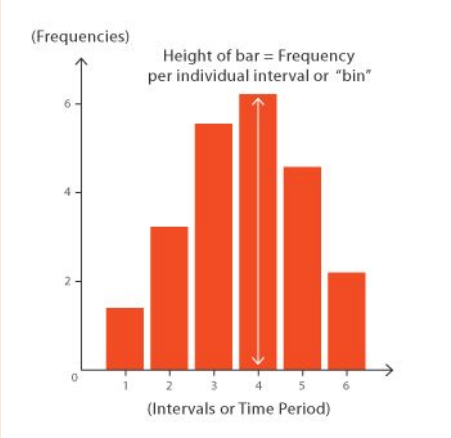
3. Distribution - Examining how data values are spread across a range, revealing patterns like central tendency, variability, skewness, and the frequency of different values.

Data Type	Variables	Objective	Example Question	Best Chart Type(s)
Categorical	1 Variable	Distribution	*"How are survey responses distributed across rating categories?" • "What's the frequency of each category?"*	**Bar Chart** · Histogram · Tally Chart · Dot Matrix Chart
Numerical	1 Variable	Distribution	*"How are employee salaries distributed?" • "What's the shape and spread of our numerical data?"*	**Histogram** · Box & Whisker Plot · Violin Plot · Density Plot · Stem & Leaf Plot
Numerical	2 Variables	Distribution	*"What's the age and gender distribution of our population?" • "How are numerical values distributed across two categorical dimensions?"*	**Population Pyramid** · Back-to-Back Bar Chart · Grouped Histogram · Split Violin Plot
Numerical	4+ Variables	Distribution	*"How are multiple numerical variables distributed across groups?" • "What are multi-dimensional data patterns?"*	**Parallel Coordinates** · Small Multiples · Box Plot Matrix · Violin Plot Matrix
Time-Series	1 Variable	Distribution	*"How do daily sales vary throughout the year?" • "What are the seasonal patterns in our data?"*	**Calendar** · Heatmap · Spiral Plot · Box & Whisker Plot
Spatial	1 Variable	Distribution	*"Where are our customers located?" • "What's the geographic spread of events?"*	**Dot Map** · Point Map · Pin Map · Symbol Map

Histogram - visualises the distribution of data over a continuous interval.

- Each bar in a histogram represents the tabulated frequency at each interval/bin.
- Histograms show how values in a dataset are **distributed across intervals**, revealing patterns such as skewness.
- Highlight central tendency. The shape of the histogram can provide insight into where most data points lie, **helping to identify the mean or median visually.**

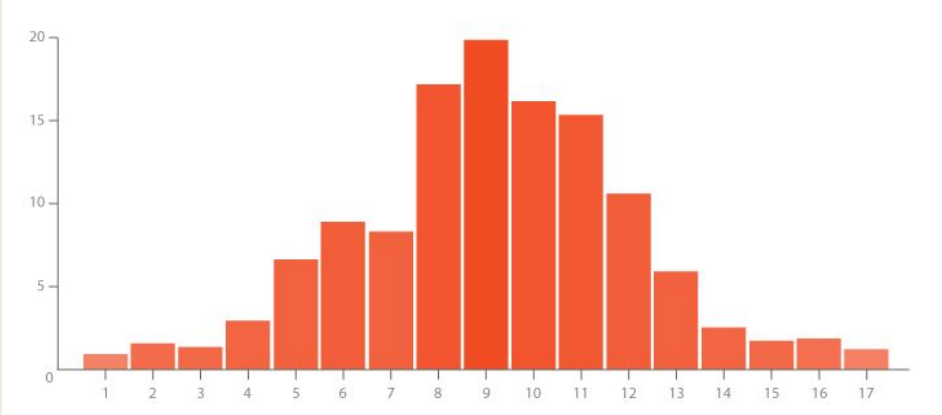


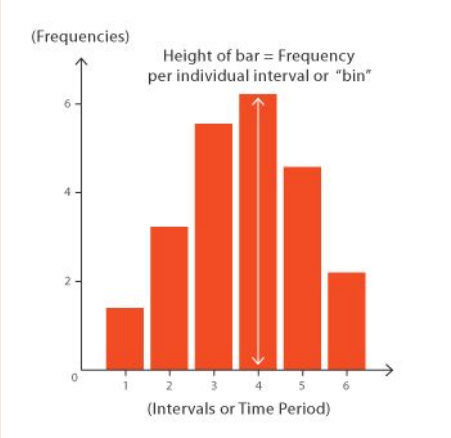


Histogram - visualises the distribution of data over a continuous interval.

How to Read a Histogram

- **X-Axis (Horizontal)-** Represents the bins, or ranges of values. Each bin covers a specific interval of the dataset.
- **Y-Axis (Vertical)-** Represents the frequency of data points within each bin.
- **Bars-** The height of each bar indicates how many data points fall into that bin. For example, if a bar reaches a height of 10, that means there are 10 data points within that specific range.





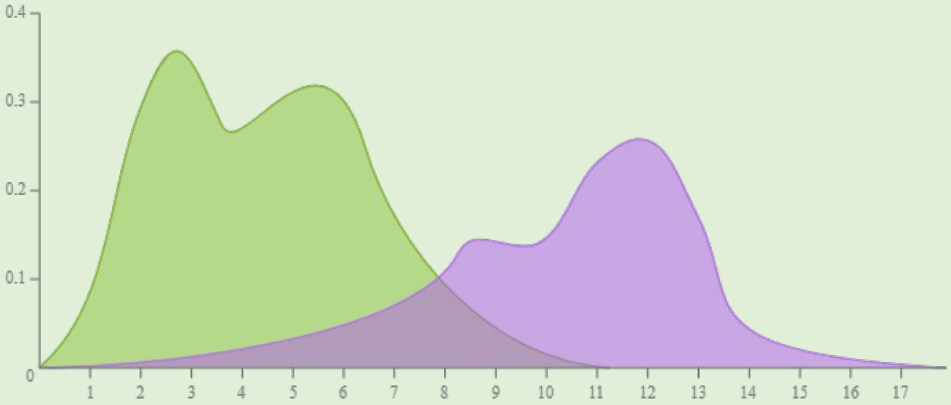
Density Plot - visualises the distribution of data over a continuous interval or time period.

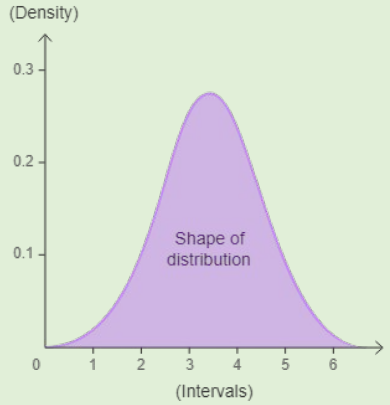
Also known as a Kernel Density Plot or Density Trace Graph.

- Density plots show a smooth curve, which is a continuous approximation of the distribution of data, unlike the blocky representation of histograms.

- This chart is a variation of a Histogram that uses kernel smoothing to plot values, allowing for smoother distributions by smoothing out the noise.

- The peaks of a Density Plot help display where values are concentrated over the interval.



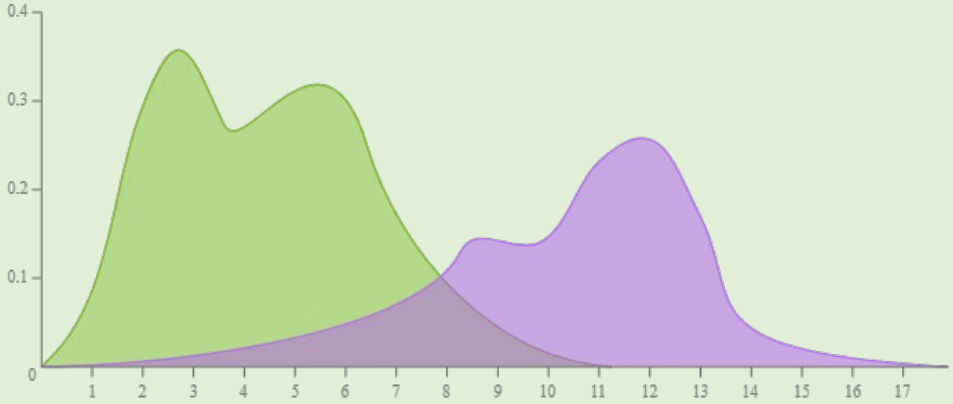


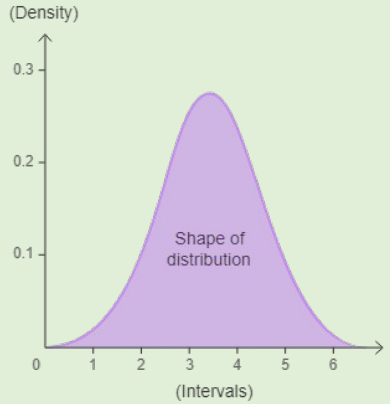
Density Plot - visualises the distribution of data over a continuous interval or time period.

Also known as a Kernel Density Plot or Density Trace Graph.

How to Read a Density Plot

- **X-Axis (Horizontal) -** Represents the range of values of the variable being analyzed.
- **Y-Axis (Vertical) -** Represents the estimated probability density; values indicate how dense the data points are around a particular value.
- **Curve -** The shape of the curve indicates where data points are concentrated. Peaks represent high-density regions, while valleys represent low-density regions.

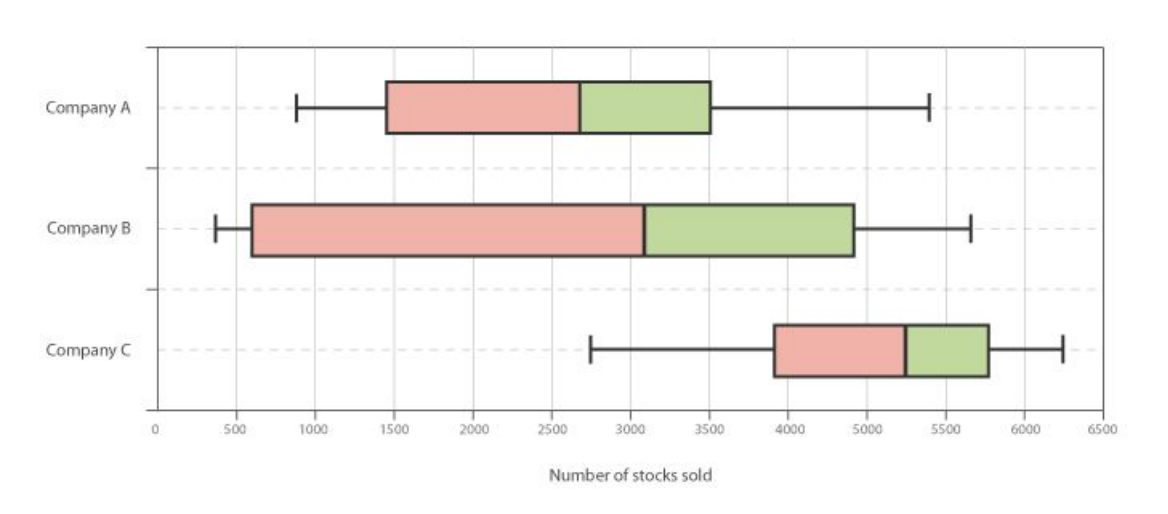


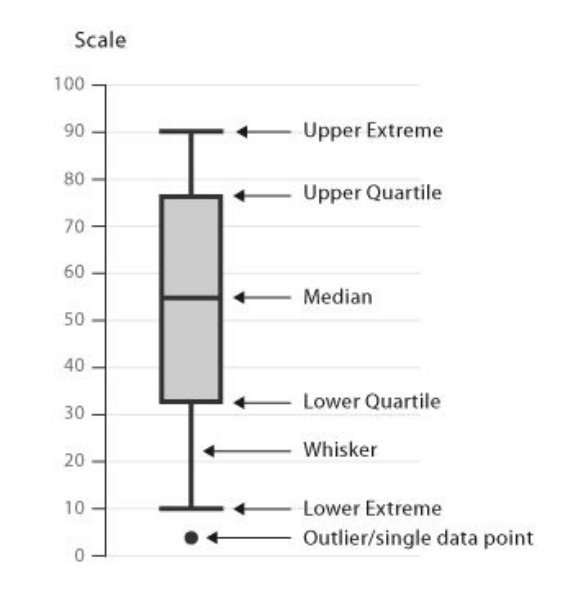


A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles.

The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally.

When multiple box plots are placed side by side, they allow for easy comparison of data distributions between different groups.





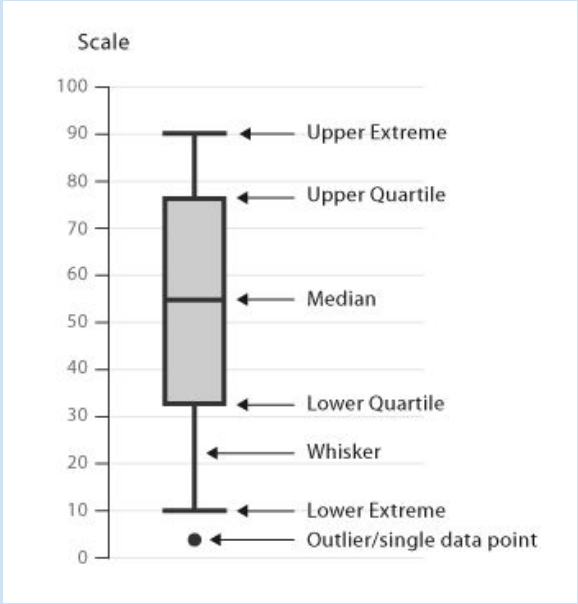
A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles.

- The **box represents the interquartile** range (IQR), which contains the middle 50% of the data.

- **Lower Quartile (Q1) -** The 25th percentile of the data (first quartile).

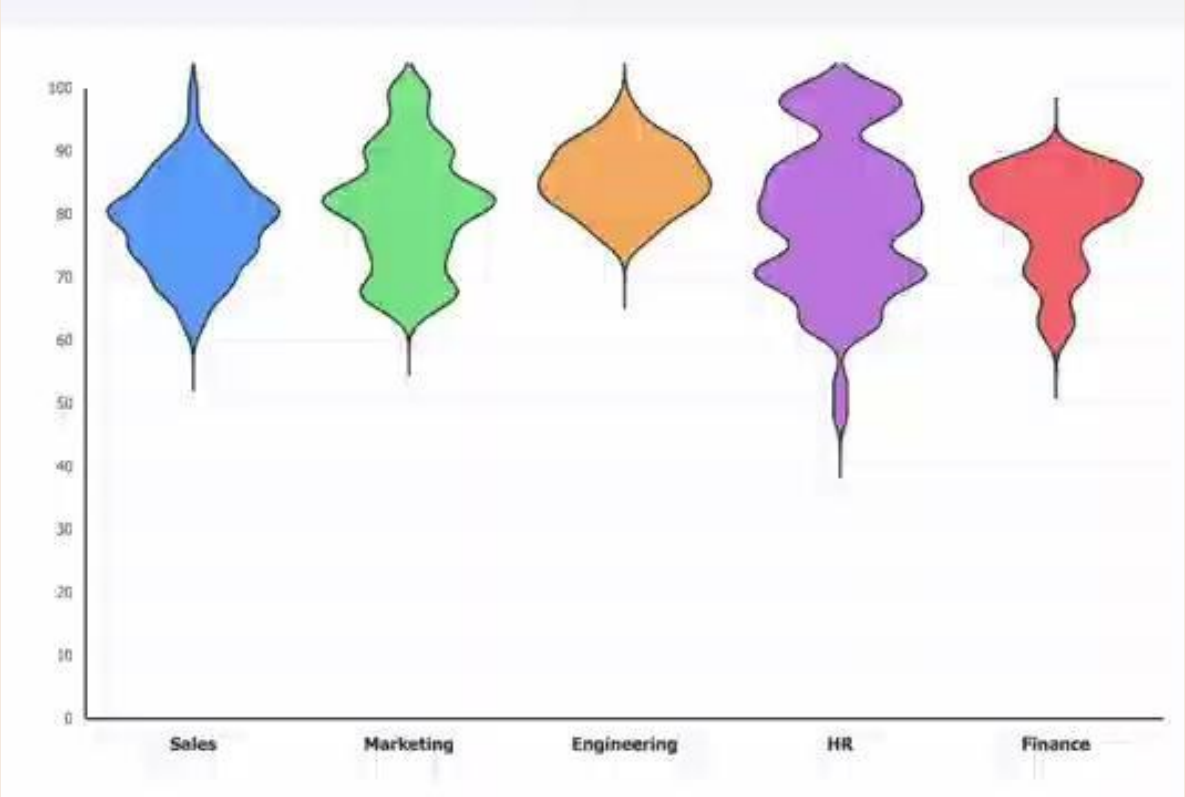
- **Upper Quartile (Q3) -** The 75th percentile of the data (third quartile).

- **Median (Q2) -** A line inside the box indicates the median (50th percentile), which shows the central point of the dataset.
- **X-Axis (Horizontal)-** Represents different groups or categories (if multiple box plots are displayed).
- **Y-Axis (Vertical)-**Represents the values of the variable being analyzed.



A Violin Plot is used to visualise the distribution of the data and its probability density.

- This chart is a combination of a **Box Plot and a Density Plot** that is rotated and placed on each side (to show the distribution shape of the data).



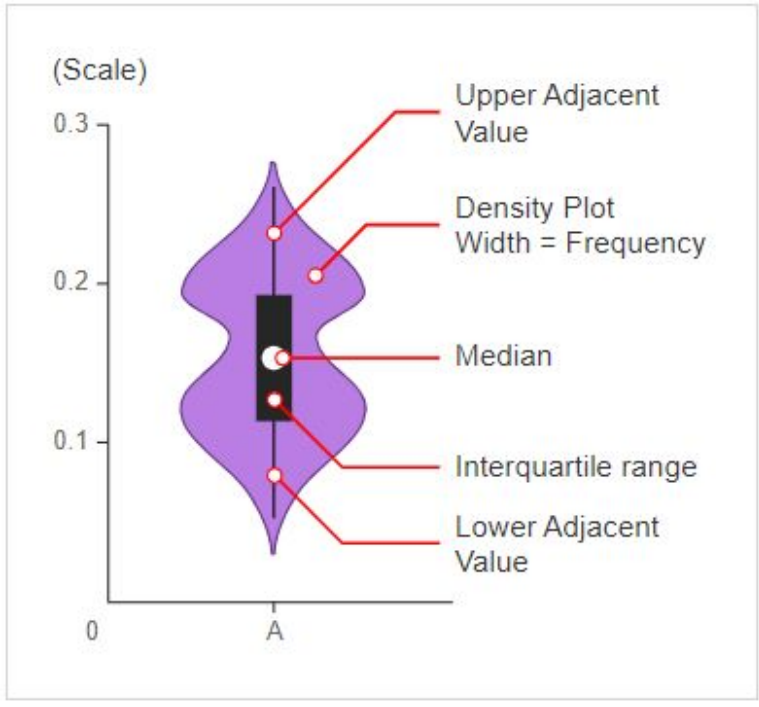
A Violin Plot is used to visualise the distribution of the data and its probability density.

- The **white dot** in the middle is the median value and the **thick black bar** in the centre represents the interquartile range (IQR), highlighting where the middle 50% of the data points lie.

- The **thin black line** extending from it represents the upper (max) and lower (min) adjacent values in the data. Sometimes the graph marker is clipped from the end of this line.

- **X-Axis (Horizontal)-** Represents different categories or groups being compared.

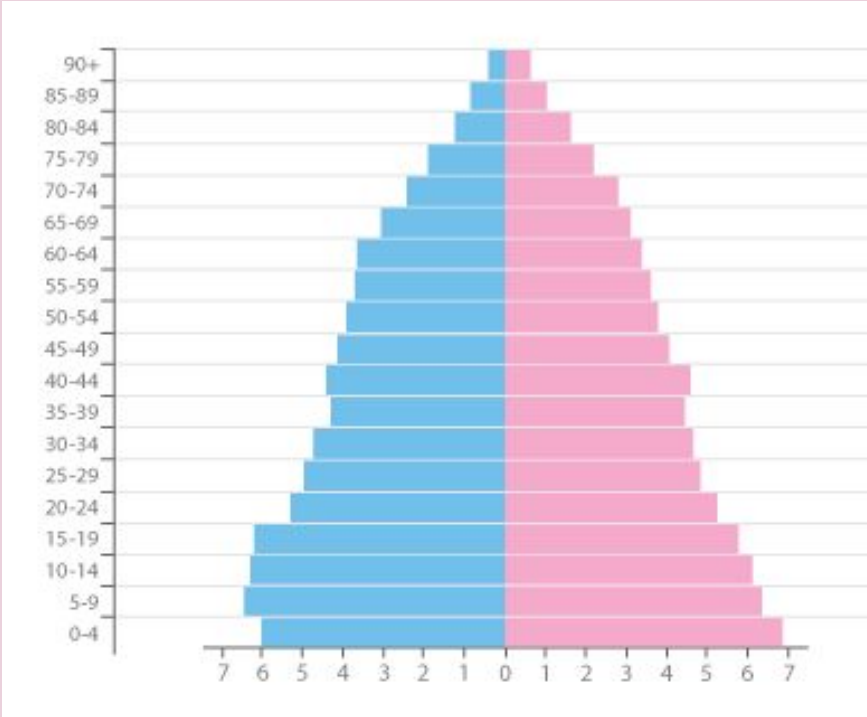
- **Y-Axis (Vertical)-** Represents the values of the variable being analyzed.



Population Pyramid - Compare distributions across age groups for different categories, typically male vs. female.

Uses

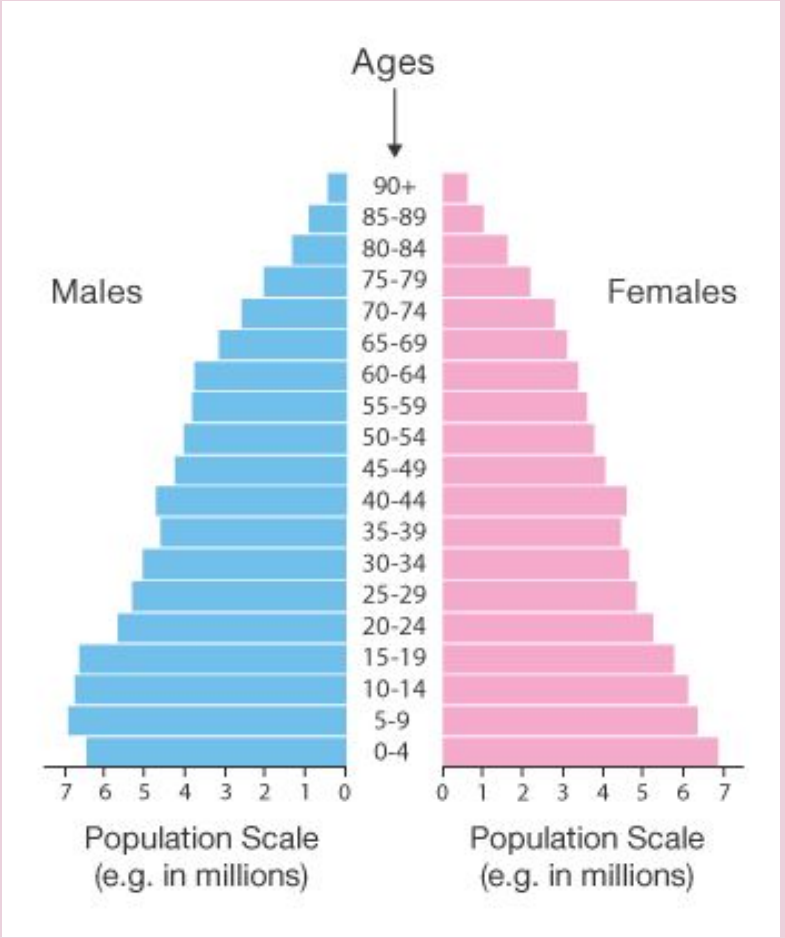
- **Demographic Comparison -** Population pyramids are used to compare age and gender distributions within a population. They show the population structure by displaying age groups (cohorts) on the vertical axis and the population count or percentage on the horizontal axis, typically with males on one side and females on the other.
- **Population Dynamics -** Useful for analyzing population trends, such as growth, decline, or aging populations, and predicting future demographic shifts.



Population Pyramid - Compare distributions across age groups for different categories, typically male vs. female.

Design Practices

- **Clear Labels -** Label age cohorts and population counts/percentages clearly for easy interpretation.
- **Symmetry Focus -** The layout emphasizes visual symmetry (or asymmetry) between male and female populations, so balancing both sides is key for clarity.
- **Use Color or Shading -** Distinguish male and female sides with different colors or shading to enhance readability.

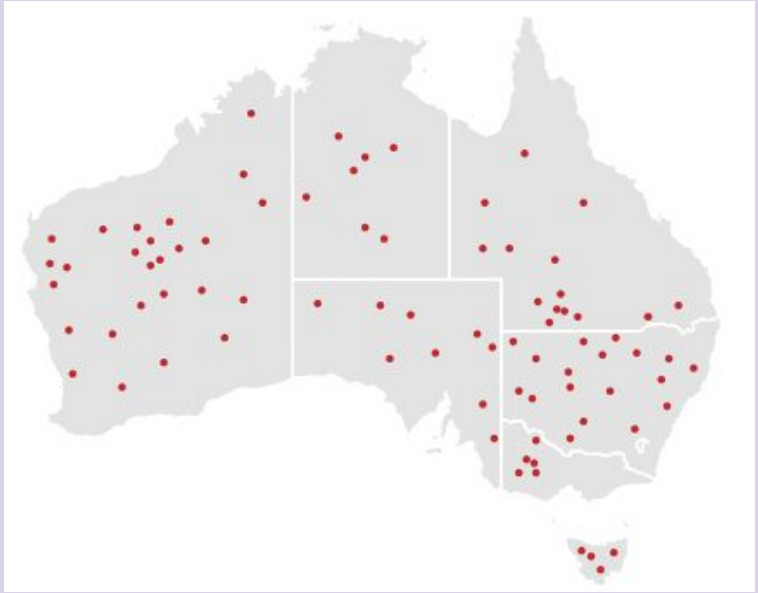


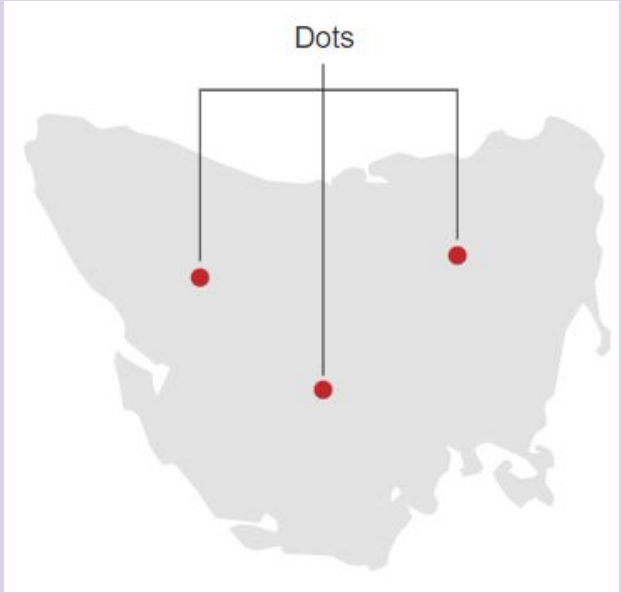
Dot Map

Also known as a Point Map, Dot Distribution Map, Dot Density Map.

- Dot Maps are a way of detecting spatial patterns or the distribution of data over a geographical region, by placing equally sized points over a geographical region.

- There are two types of Dot Map: one-to-one (one point represents a single count or object) and one-to-many (one point represents a particular unit, e.g. 1 point = 10 trees).





4. Relationship - Exploring connections, correlations, and dependencies between two or more variables to understand how changes in one variable affect another.

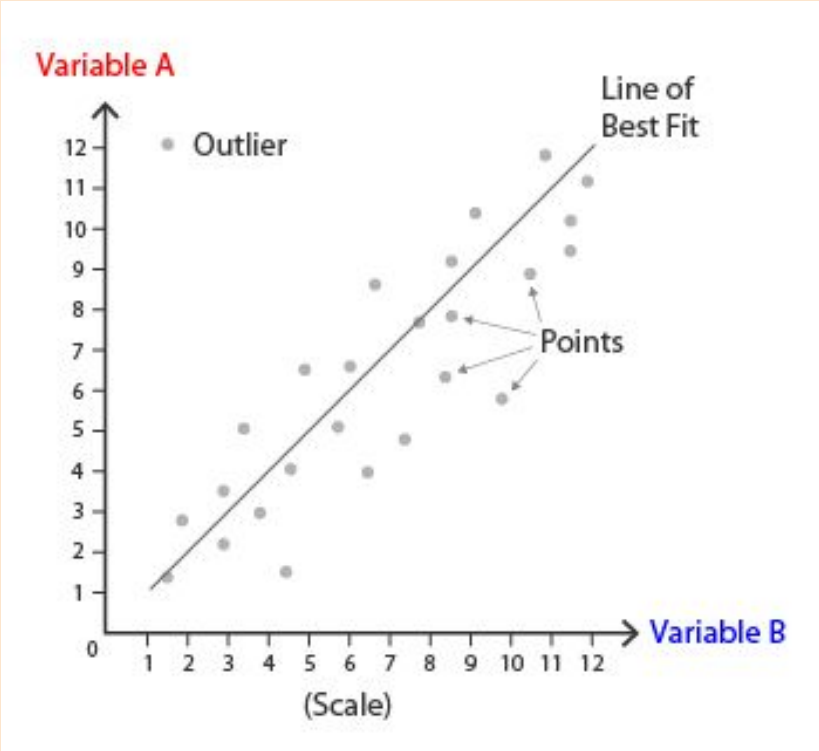
Data Type	Variables	Objective	Example Question	Best Chart Type(s)
Categorical	2 Variables	Relationship	*"How are departments connected through shared projects?" • "What are the relationships between different categories?"*	**Network Diagram** · Chord Diagram · Arc Diagram · Heatmap · Venn Diagram
Numerical	2 Variables	Relationship	*"Is there correlation between advertising spend and sales?" • "How do two numerical variables relate?"*	**Scatterplot** · Line Graph · Heatmap · Regression Plot
Numerical	3 Variables	Relationship	*"How do price, quality, and sales volume relate?" • "What's the relationship between three numerical variables?"*	**Bubble Chart** · 3D Scatterplot · Bubble Map · Animated Scatterplot
Time-Series	2 Variables	Relationship	*"How do multiple time series correlate?" • "What's the relationship between time-based variables?"*	**Line Graph** · Scatterplot · Cross-correlation Plot · Lag Plot
Mixed	4+ Variables	Relationship	*"How do multiple variables interact across categories?" • "What are complex multivariate relationships?"*	**Parallel Coordinates** · Scatterplot Matrix · Correlation Matrix · Network Diagram

Scatterplot - Show relationships or correlations between two continuous variables.

Also known as a Scatter Graph, Point Graph, X-Y Plot, Scatter Chart or Scattergram.

- The strength of the correlation can be determined by how closely packed the points are to each other on the graph. Points that end up far outside the general cluster of points are known as **outliers.**

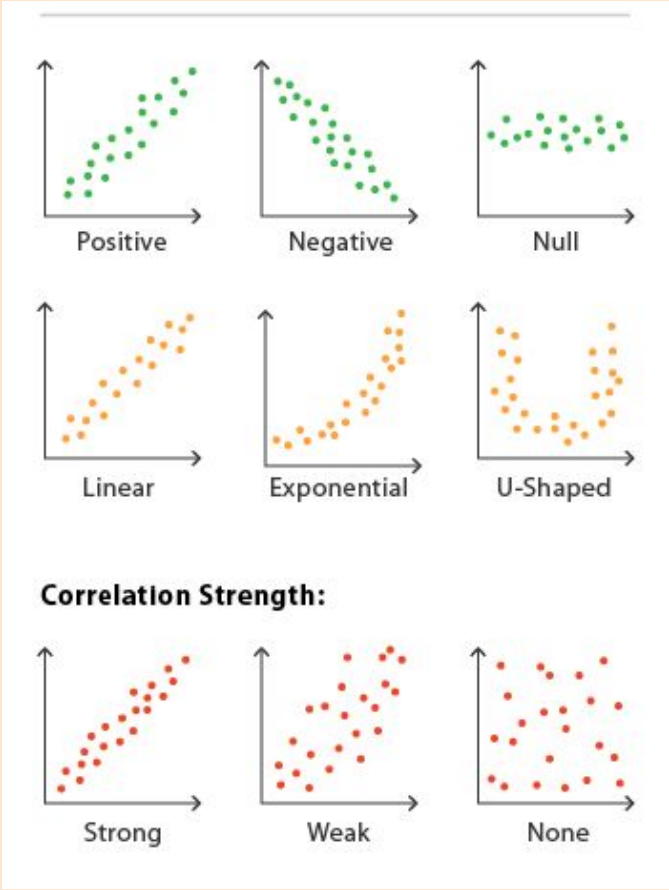
- Lines or curves can be displayed over the graph to aid in the analysis. This is typically known as the Line of Best Fit or Trend Line and can be used to make estimates via interpolation. A Line of Best Fit is drawn as close to all the points as possible to show how it would look if all the points were condensed together into a single line.



Scatterplot - Show relationships or correlations between two continuous variables.

Also known as a Scatter Graph, Point Graph, X-Y Plot, Scatter Chart or Scattergram.

- Scatterplots are ideal when you have paired numerical data and you want to see if one variable impacts the other.
- However, do remember that correlation is not causation and another unnoticed or indirect variable may be influencing the results.



Scatterplot - Show relationships or correlations between two continuous variables.

Also known as a Scatter Graph, Point Graph, X-Y Plot, Scatter Chart or Scattergram.

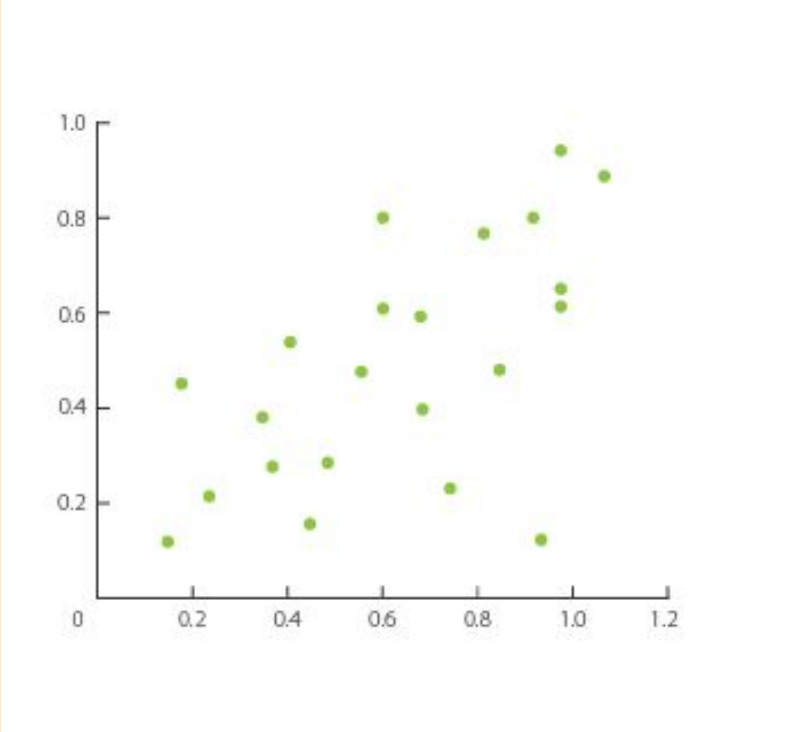
Design Practices

Clear Axes Labels - Label both axes clearly to indicate what each dimension represents.

Trend Lines - Consider adding a trend line or regression line to highlight overall trends or correlations.

Avoid Overlapping Points - If many data points overlap, consider using transparency, jitter, or other methods to make the density of points clearer.

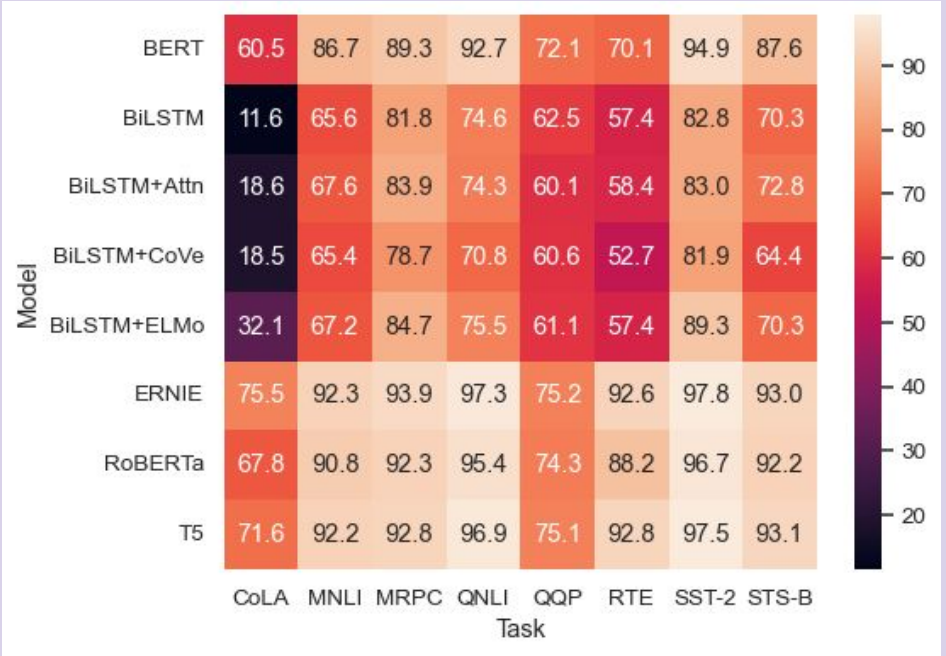
Minimal Gridlines - Use gridlines sparingly to maintain focus on the data points.



Heatmap - Show data intensity with color gradients, often in grid or map layouts.

- **Data Intensity Visualization -** Heatmaps are used to visualize data intensity or concentration across a matrix or grid. They represent data values through color gradients, making it easy to identify patterns, clusters, and variations.

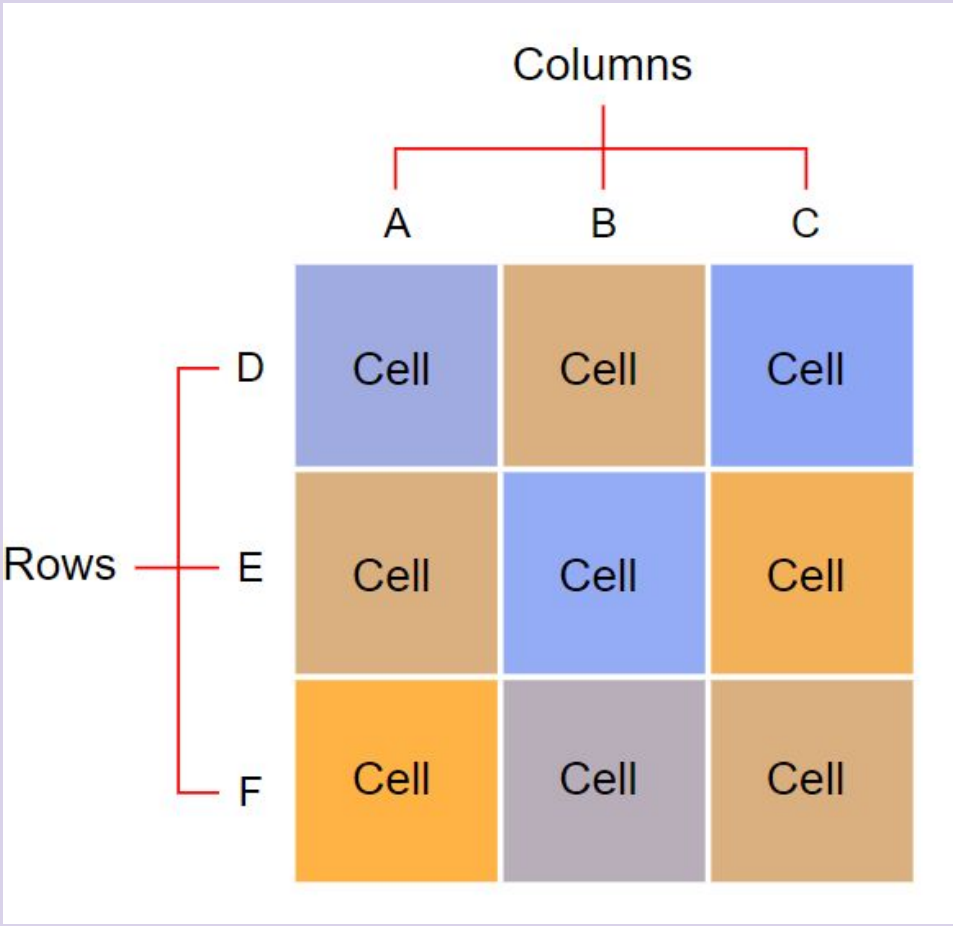
- **Correlation Analysis -** Useful for examining the relationships between variables or showing correlations in large datasets.



Heatmap - Show data intensity with color gradients, often in grid or map layouts.

Design Practices

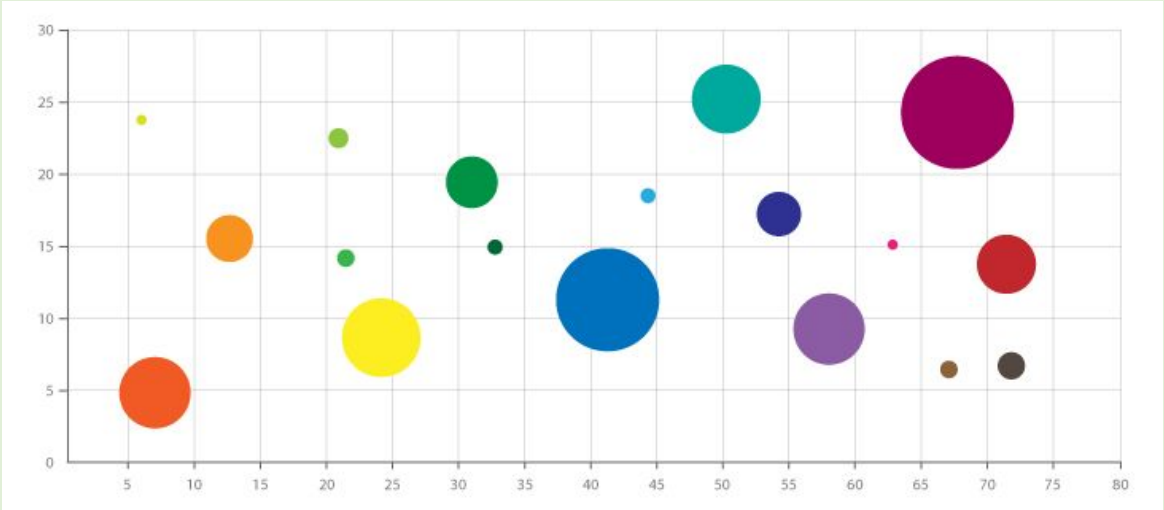
- **Color Gradient -** Choose a clear and intuitive color gradient to represent data values, ensuring that differences in intensity are easily distinguishable.
- **Interactive Features -** Consider interactive elements, like tooltips or zooming, to explore data in more detail, especially for large datasets.



Bubble Chart - Show relationships with an additional variable represented by the size of bubbles.

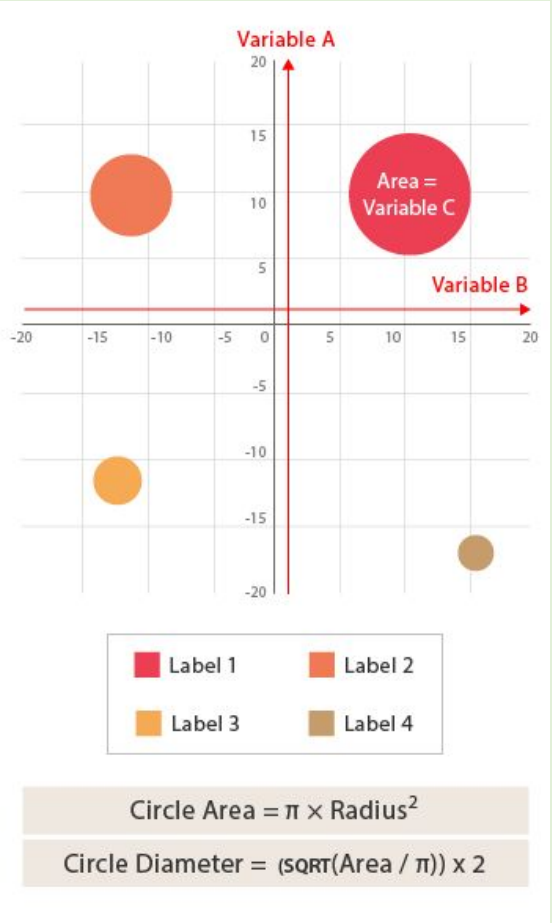
- A Bubble Chart is a **multi-variable graph that is a cross between a Scatterplot** and a Proportional Area Chart.

- Like a Scatterplot, Bubble Charts use a Cartesian coordinate system to plot points along a grid where the X and Y axis are separate variables.



Bubble Chart - Show relationships with an additional variable represented by the size of bubbles.

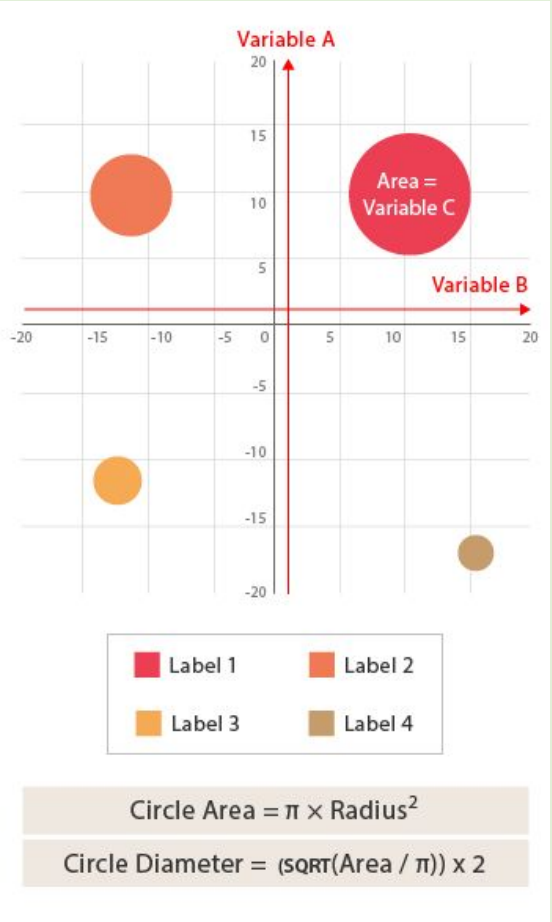
- Each plotted point then represents a **third variable by the area of its circle.**
- **Colours** can also be used to distinguish between **categories** or used to represent an additional data variable.
- **Time** can be shown either by having it as a variable on one of the axis or by animating the data variables changing over time.



Bubble Chart - Show relationships with an additional variable represented by the size of bubbles.

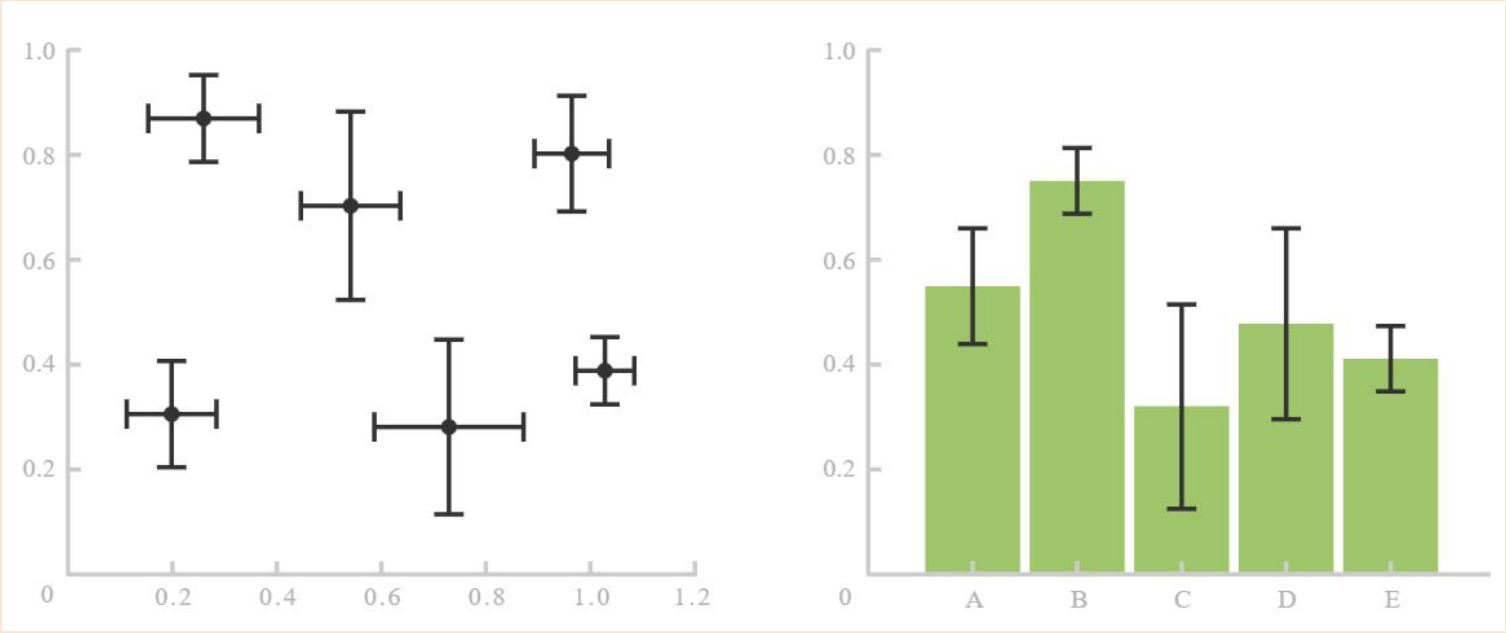
Design Practices

- **Distinct Bubble Sizes -** Ensure that bubble sizes are distinct enough to reflect the third variable accurately and make comparisons easier.
- **Use Color for Clarity -** Differentiate data categories or groups with different colors for better visualization.
- **Limit Overlap -** Minimize overlapping bubbles by adjusting layout or transparency to maintain clarity.
- **Legend or Tooltip -** Provide a legend or interactive tooltips to clarify what the bubble size represents for better interpretation.



Error Bars - Indicate the uncertainty or variability in the relationship between data points.

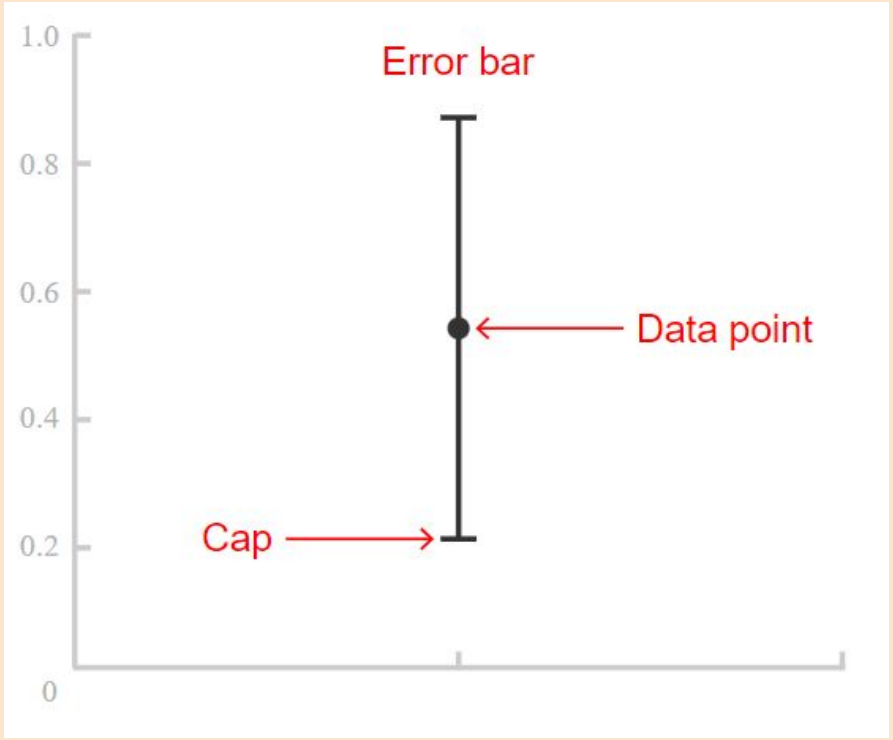
- Error Bars function as a **graphical enhancement** that visualises the variability of the plotted data on a Cartesian graph.
- Error Bars **can be applied to graphs such as Scatterplots, Dot Plots, Bar Charts or Line Graphs,** to provide an additional layer of detail on the presented data.



Error Bars - Indicate the uncertainty or variability in the relationship between data points.

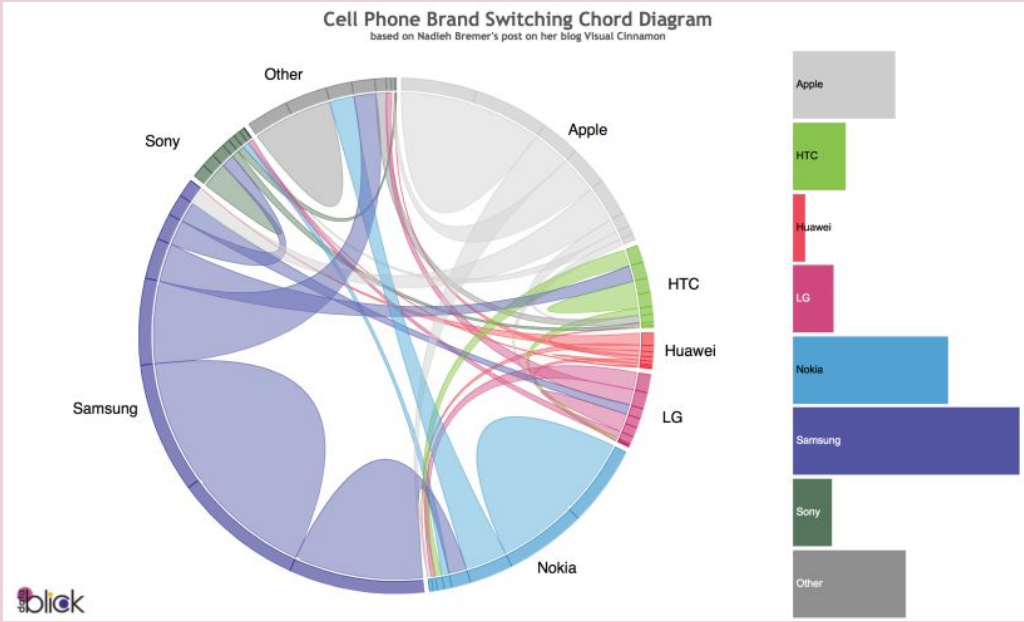
- Typically, Error bars are used to display either the standard deviation, standard error, confidence intervals or the minimum and maximum values in a ranged dataset.

- The length of an Error Bar helps reveal the uncertainty of a data point: **a short Error Bar shows that values are concentrated**, signalling that the plotted average value is more likely, while a long Error Bar would indicate that the values are more spread out and less reliable.



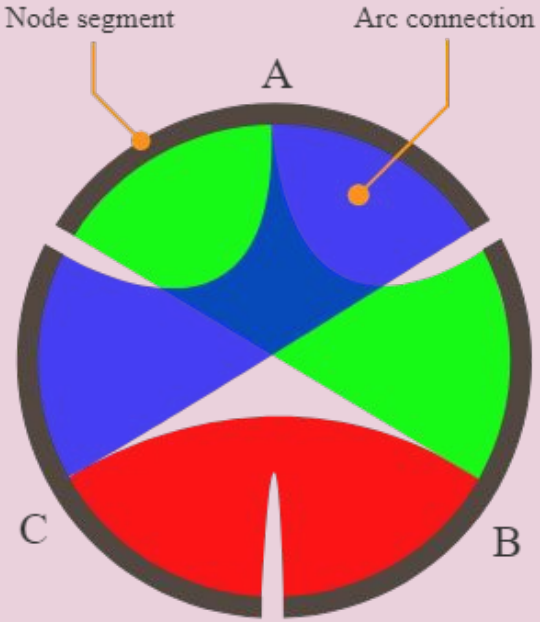
Chord Diagram - Show the relationship between categories in a circular layout, often used for flow or network data.

- This type of diagram **visualises the inter-relationships between entities.** The connections between entities are used to display that they share something in common.
- This makes Chord Diagrams **ideal for comparing the similarities** within a dataset or between different groups of data.
- Useful for analyzing network connections or interactions between nodes.



Chord Diagram - Show the relationship between categories in a circular layout, often used for flow or network data.

- Nodes are arranged along a circle, with the relationships between points connected to each other either through the use of arcs or Bézier curves.
- **Values are assigned to each connection**, which is represented proportionally by the size of each arc.
- **Colour can be used to group the data into different categories,** which aids in making comparisons and distinguishing groups.



	A	B	C
A		10	10
B	10		10
C	10	10	

A $\xrightarrow{10}$ B

B $\xrightarrow{10}$ C

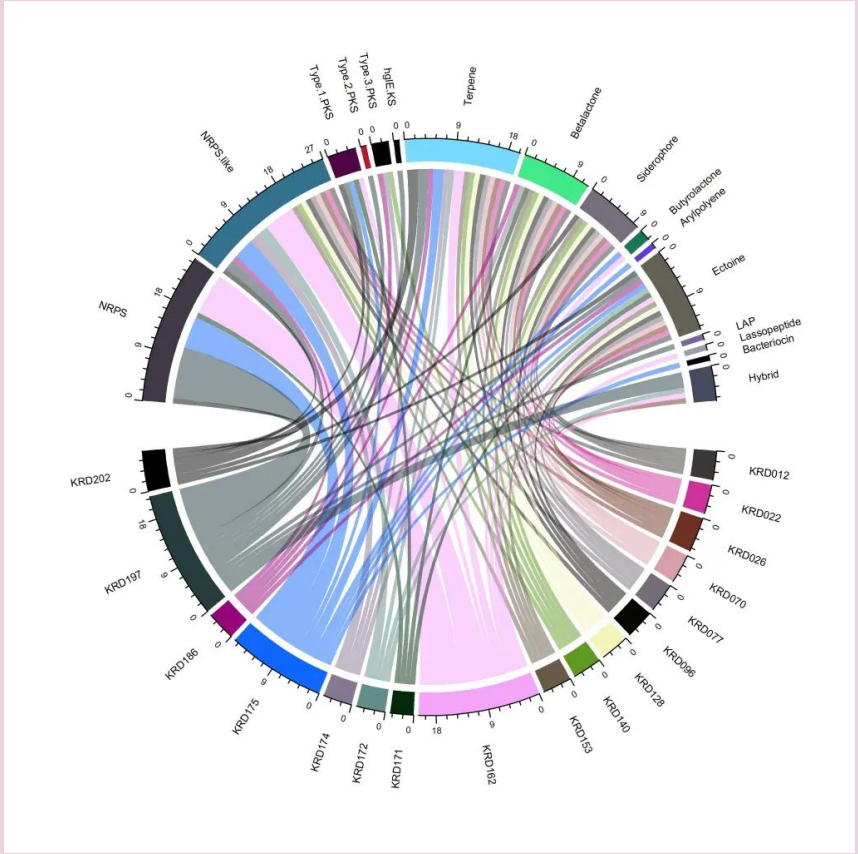
C $\xrightarrow{10}$ A

Chord Diagram - Show the relationship between categories in a circular layout, often used for flow or network data.

Design Practices

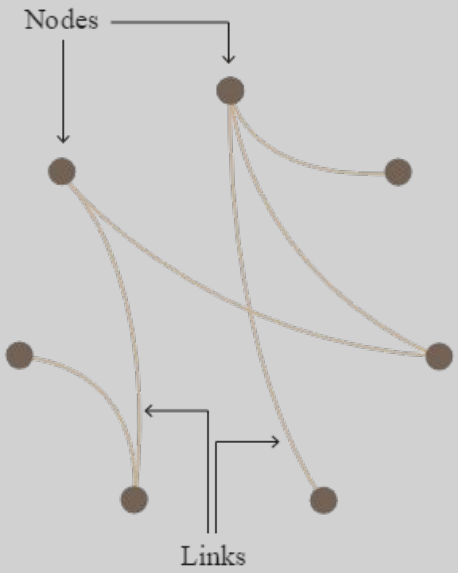
- **Manage Complexity -** Limit the number of categories and connections to avoid clutter and improve readability. Complex diagrams may benefit from interactive features for detailed exploration.

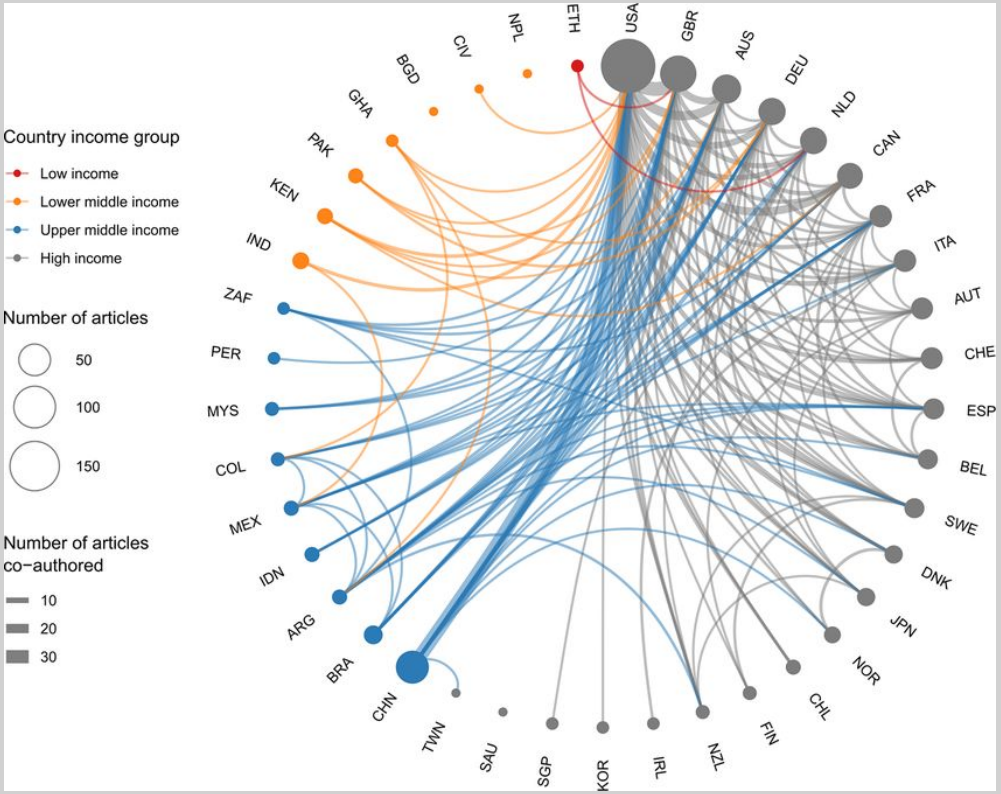
- **Consistent Proportions -** Ensure that the width of the chords (connections) accurately represents the magnitude of the relationships to avoid misleading interpretations.



Non-ribbon Chord Diagram - Similar to a chord diagram but without ribbons connecting categories.

- Non-ribbon Chord Diagrams provide **more emphasis on the connections** within the data.



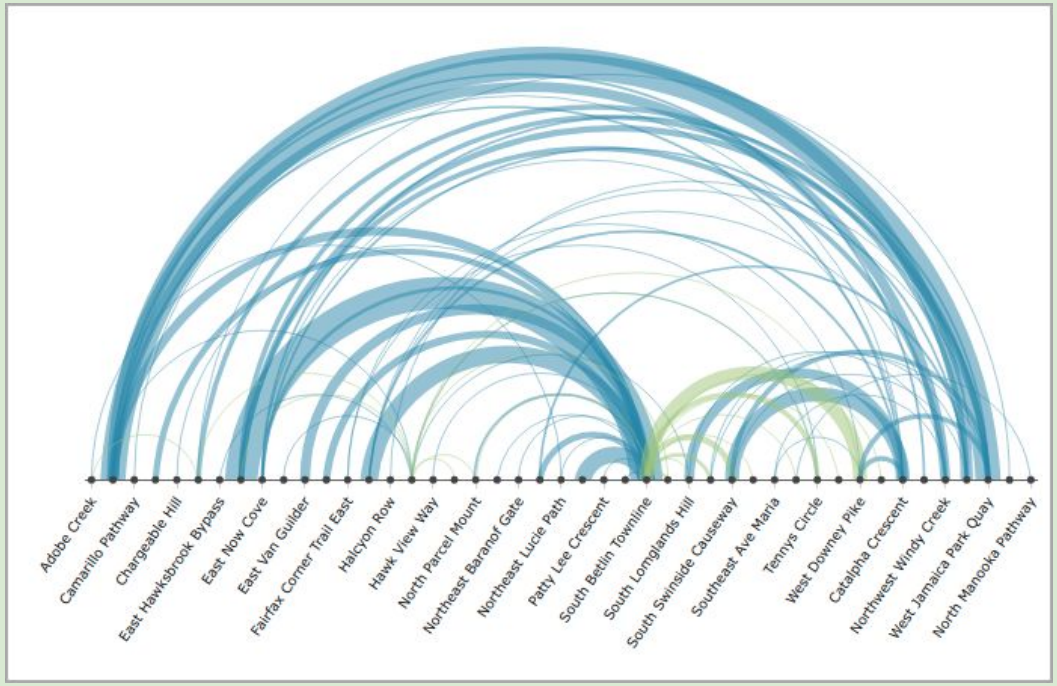


Arc Diagram - Visualize relationships between entities using arcs.

- Arc Diagrams are an alternate way of representing two- dimensional Network Diagrams.

- In Arc Diagrams, **nodes are placed along a single line (a one-dimensional axis)** and arcs are used to show connections between those nodes.

- The **thickness of each arc line can be used to represent frequency between the source and target node.** Arc Diagrams can be useful in finding the co-occurrence within the data.



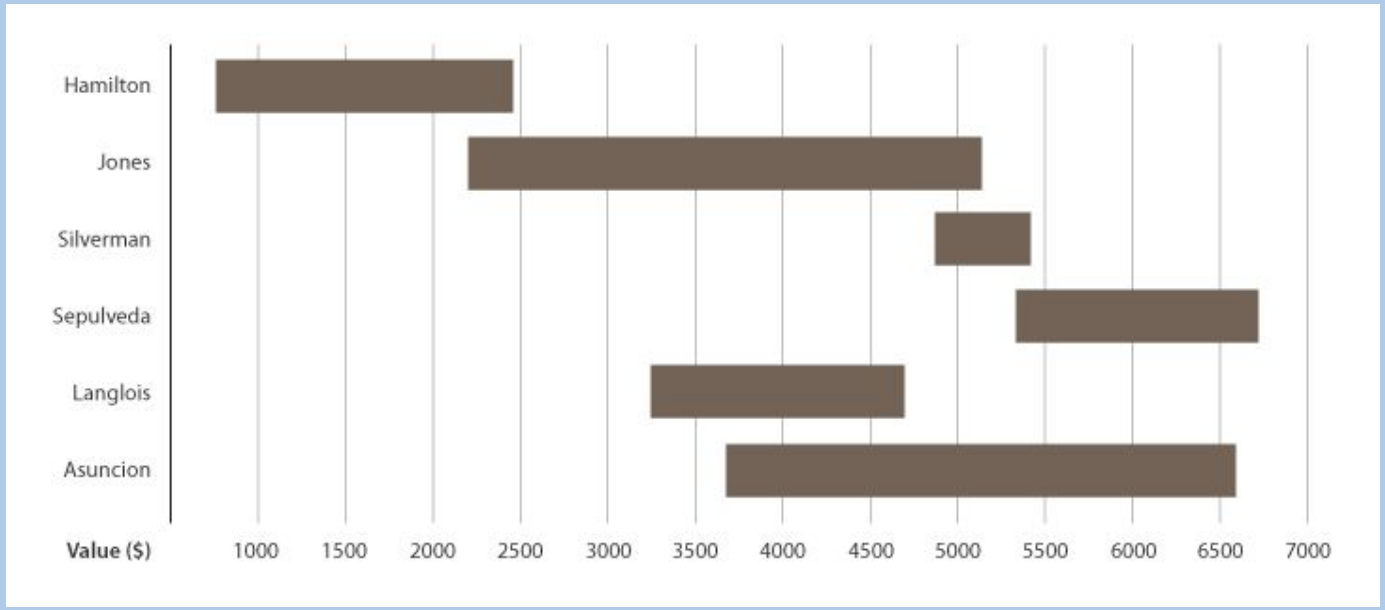
5. Trend - Analyzing how data changes over time to identify patterns, growth, decline, cycles, or seasonal variations in temporal datasets.

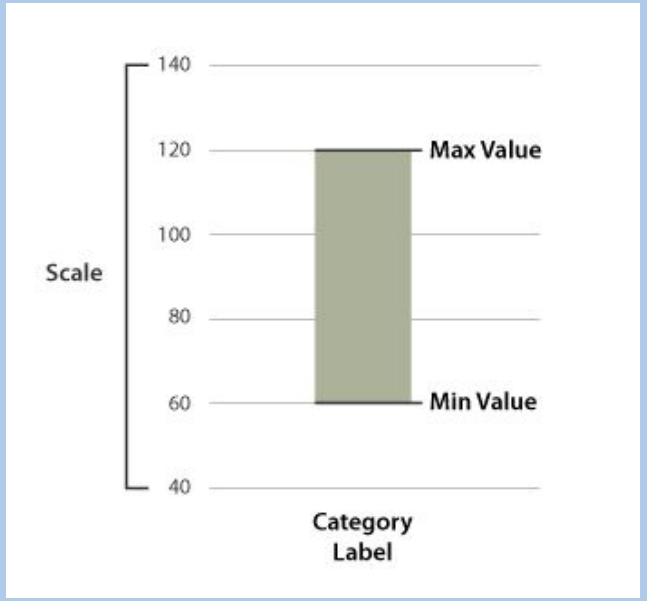
Data Type	Variables	Objective	Example Question	Best Chart Type(s)
Time-Series	1 Variable	Trend	*"How have website visitors changed over time?" • "What's the trend in our KPI over months?"*	**Line Graph** · Area Graph · Sparkline · Calendar
Time-Series	2 Variables	Comparison	*"How do sales trends compare between different regions over time?" • "How do multiple time series compare?"*	**Line Graph** · Multi-line Chart · Small Multiples · Dual-axis Chart
Time-Series	3 Variables	Trend	*"How do revenue, costs, and profit trend together over time?" • "What are the multi-dimensional time trends?"*	**Multi-line Chart** · Dual/Triple-axis Chart · Bubble Timeline · Small Multiples
Time-Series	Special	Trend	*"How do financial prices behave with OHLC data?" • "What are complex financial patterns over time?"*	**Candlestick Chart** · OHLC Chart · Kagi Chart · Point & Figure Chart

Span Chart - Show the range between two data points to illustrate relationships.

Also known as a Range Bar/Column Graph, Floating Bar Graph, Difference Graph, High-Low Graph.

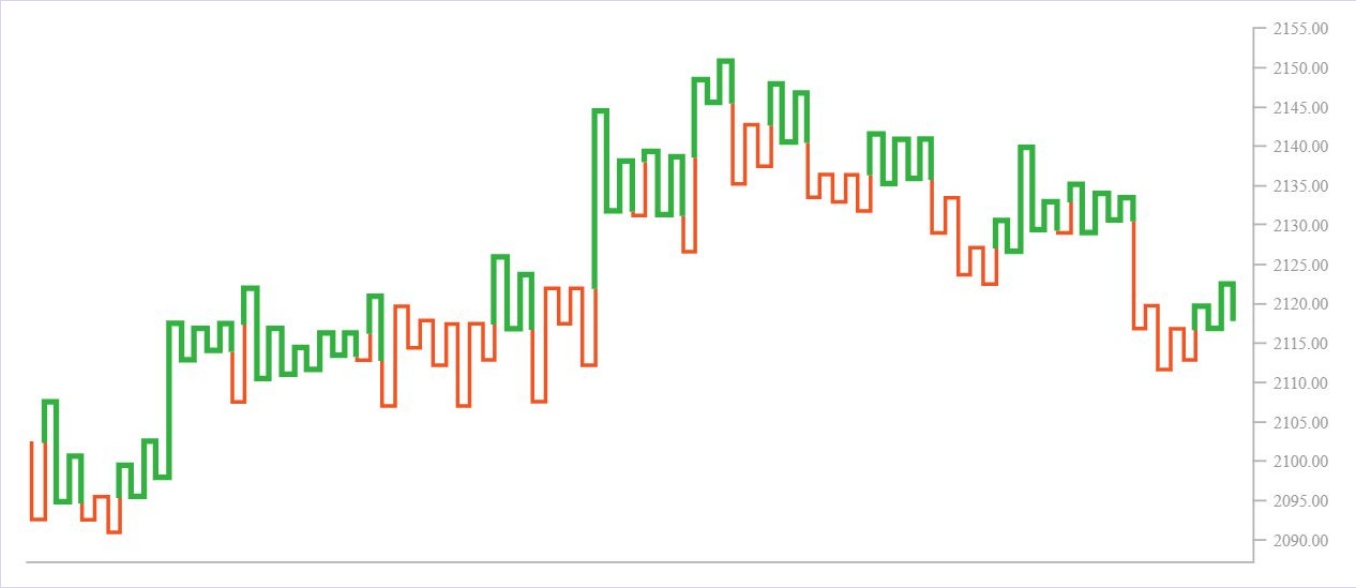
- A chart used to display dataset **ranges between a minimum and maximum value.** Span Charts are ideal for comparing ranges, typically between categories.





Kagi Chart - Show relationships between price levels and their reversals (often used in financial data).

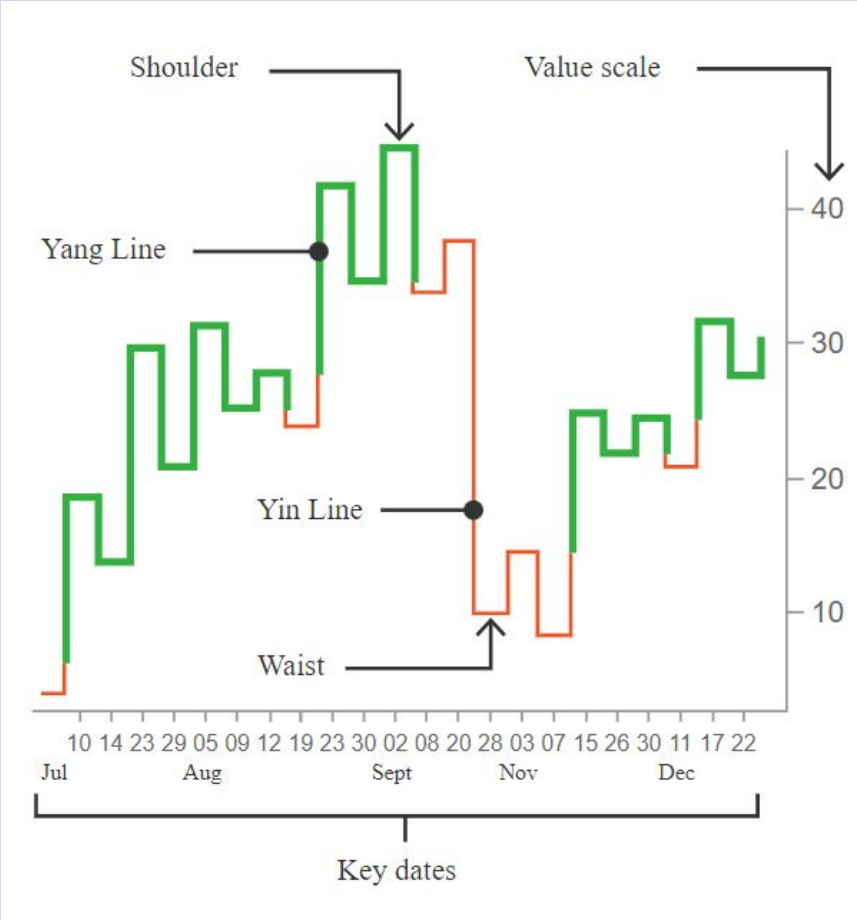
- Used to **display the general levels of supply and demand** of a particular asset by visualising the price actions through a series of line patterns.
- Kagi Charts are **time-independent** and help filter out the noise that can occur on other financial charts.
- This is so that **important price movements** are displayed more clearly. Recognising the patterns that occur in Kagi Charts is key to understanding them.



Kagi Chart – Show relationships between price levels and their reversals (often used in financial data).

- While Kagi Charts do display dates or time on their x-axis, these are in fact markers for the key price action dates and are not part of a timescale. The y-axis on the right-hand side is used as the value scale.

- When a horizontal line joins a rising line with a plunging line it's known as a **"shoulder"**, while a horizontal line connecting a plunging line with a rising line is known as a **"waist".**

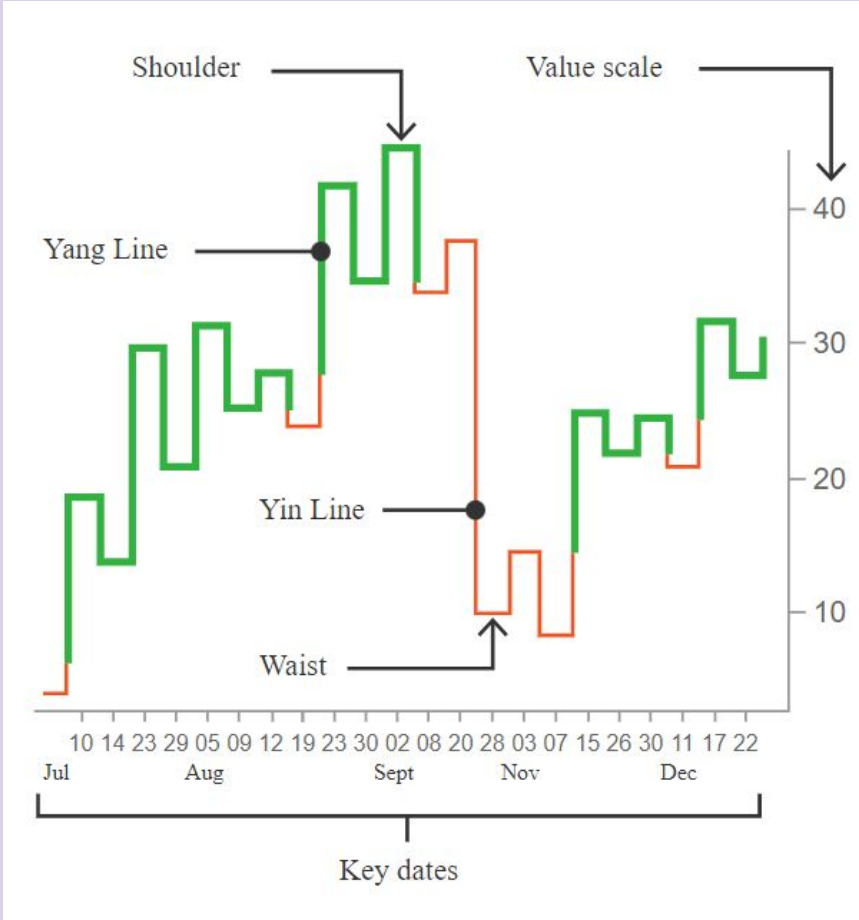


Kagi Chart - Show relationships between price levels and their reversals (often used in financial data).

- When the price goes higher than a previous "shoulder" reversal, **the line becomes thicker (and/or green) and is known as a "Yang line"**. This can be interpreted as an increase in demand over supply for the asset and as a bullish upward trend.

- Alternatively, when the price breaks below a previous "waist" reversal, **the line becomes thinner (and/or red) and is known as a "Yin line".**

- Traders use the shift from thin (Yin) to thick (Yang) lines (and vice versa) as signals to buy or sell an asset. A Yin to Yang shift indicates to buy, while a Yang to Yin shift indicates to sell.



Open-High-Low-Close Chart (OHLC) Chart - Display open, high, low, and close prices to show relationships in financial markets.

Also known as OHLC Chart, Price Chart, Bar Chart

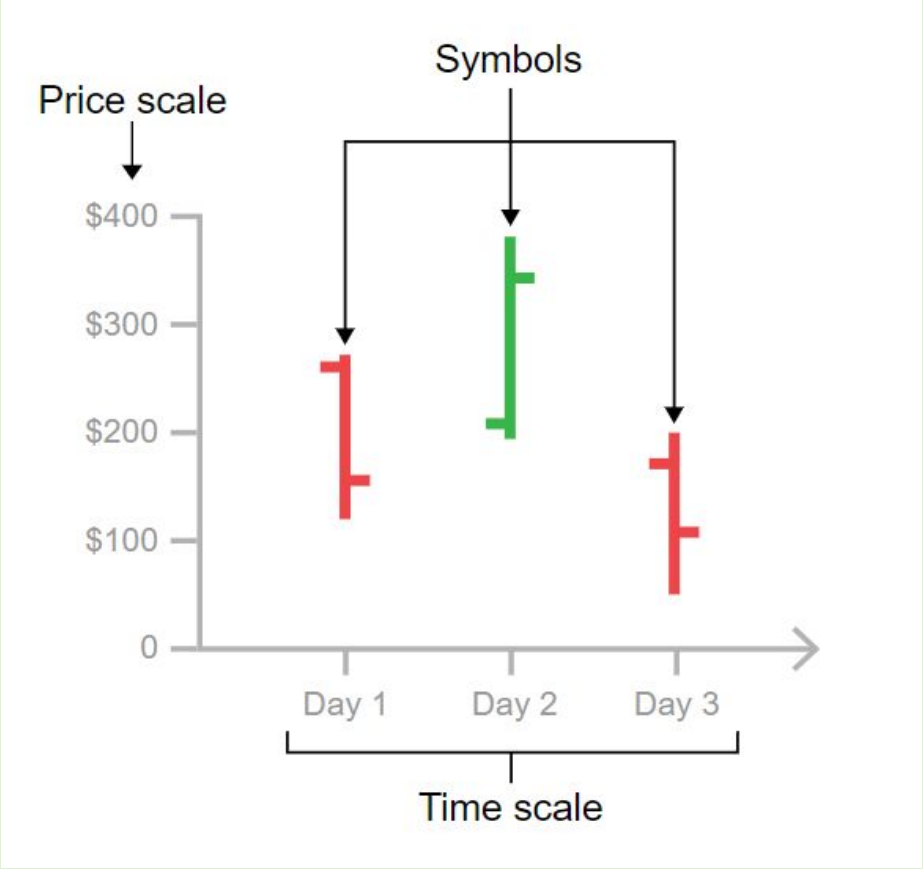
- Open-high-low-close Charts (or OHLC Charts) are used as a trading tool to visualise and analyse the **price changes over time for securities, currencies, stocks, bonds, commodities, etc.**

- OHLC Charts are useful for interpreting the **day-to-day sentiment** of the market and forecasting any future price changes through the patterns produced.

Open-High-Low-Close Chart (OHLC) Chart - Display open, high, low, and close prices to show relationships in financial markets.

Also known as OHLC Chart, Price Chart, Bar Chart

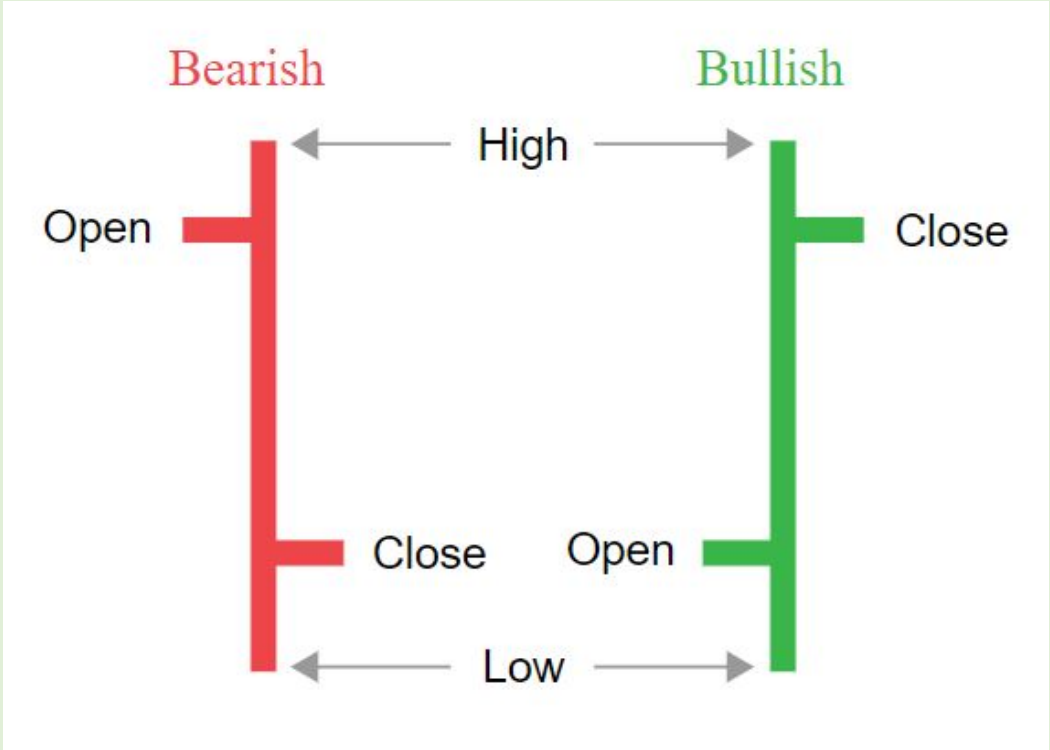
- On the **range symbol**, the high and low price ranges are represented by the length of the main vertical line.
- The open and close prices are represented by the **vertical positioning of tick-marks** that appear on the left (representing the open price) and on right (representing the close price) sides of the high-low vertical line.



Open-High-Low-Close Chart (OHLC) Chart - Display open, high, low, and close prices to show relationships in financial markets.

Also known as OHLC Chart, Price Chart, Bar Chart

- Colour can be assigned to each OHLC Chart symbol, to distinguish whether the market is **"bullish"** (the closing price is higher than it opened) or **"bearish"** (the closing price is lower than it opened).



Candlestick Chart - Show relationships between open, high, low, and close prices with a specific focus on trends.

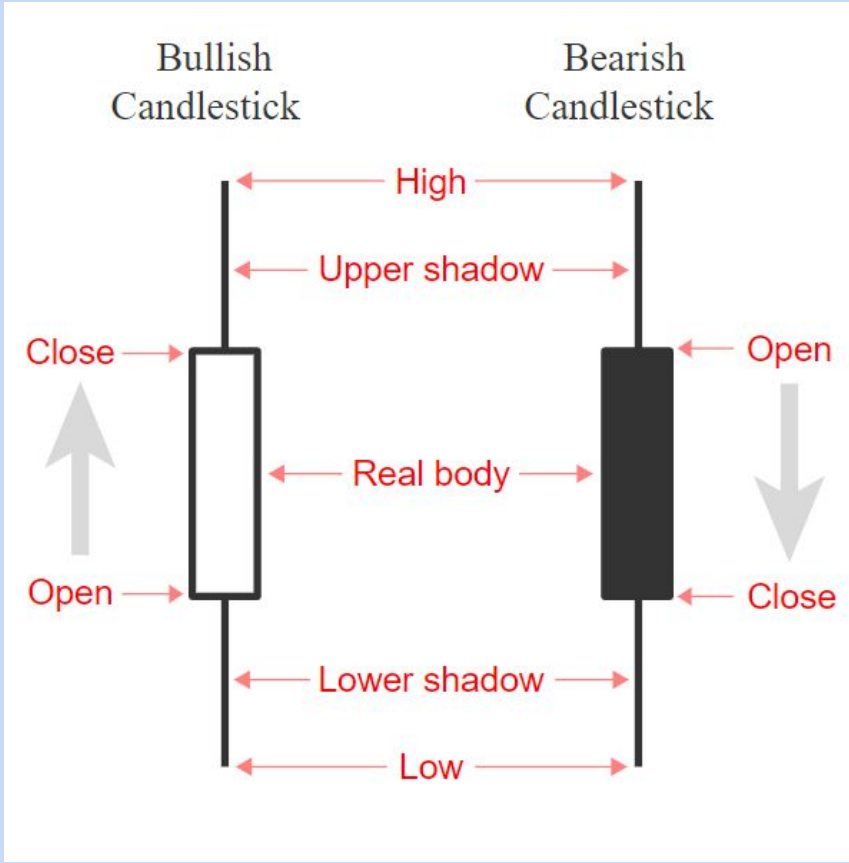
Also known as a Japanese Candlestick Chart.

- This type of chart is used as a trading tool to **visualise and analyse the price movements over time** for securities, derivatives, currencies, stocks, bonds, commodities, etc.

- Although the symbols used in Candlestick Charts resemble a Box Plot, they function differently and therefore, are not to be confused with one another.

Candlestick Chart - Show relationships between open, high, low, and close prices with a specific focus on trends.
Also known as a Japanese Candlestick Chart.

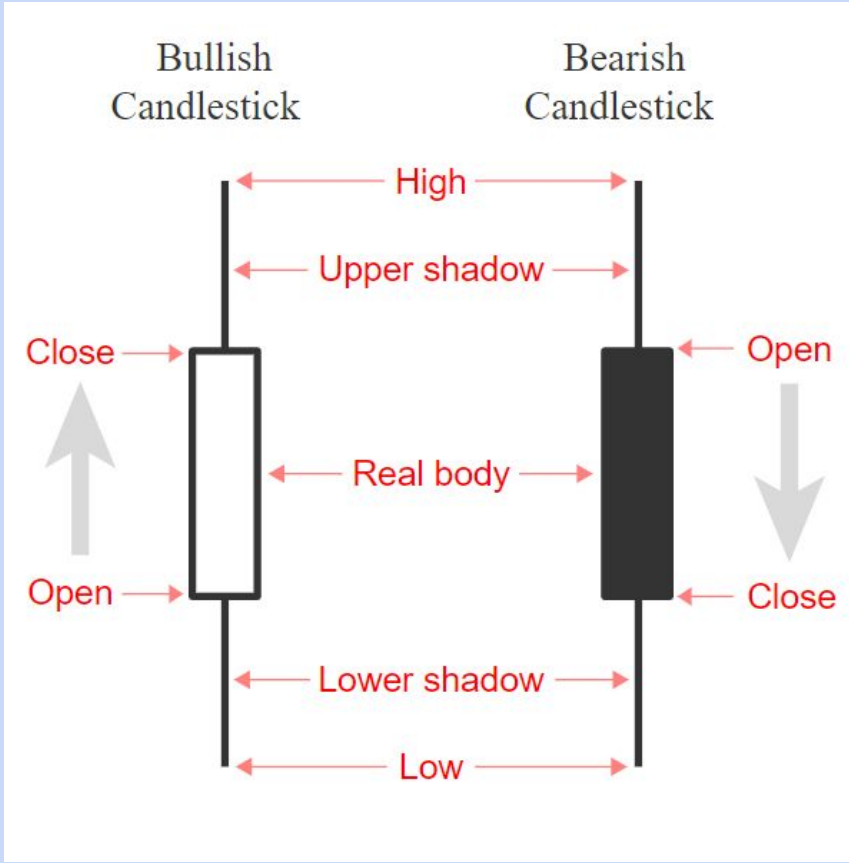
- Candlestick Charts display multiple bits of price information such as the **open price, close price, highest price and lowest price** through the use of candlestick-like symbols.
- Each symbol represents the compressed trading activity for a single time period (a minute, hour, day, month, etc).
- Each Candlestick symbol is plotted along a time scale on the x-axis, to show the trading activity over time.



Candlestick Chart - Show relationships between open, high, low, and close prices with a specific focus on trends.

Also known as a Japanese Candlestick Chart.

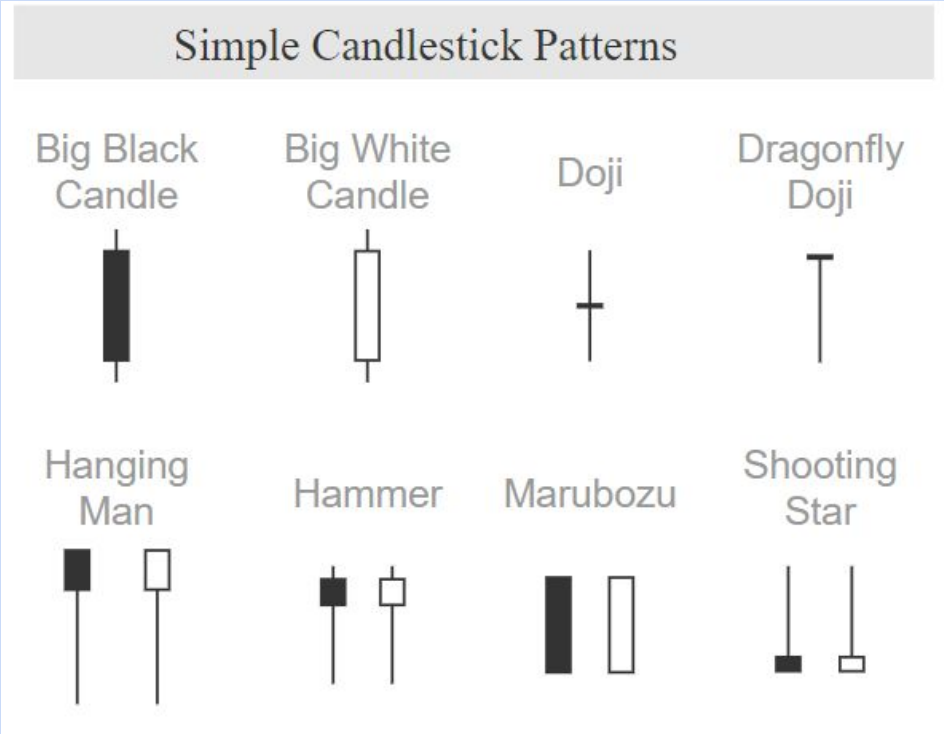
- The main rectangle in the symbol is known as the real body, which is used to display the range between the open and close price of that time period.
- While the lines extending from the bottom and top of the **real body** is known as the **lower and upper shadows** (or wick). Each shadow represents the highest or lowest price traded during the time period represented.
- When the market is **Bullish** (the closing price is higher than it opened), then the body is coloured typically white or green.
- But when the market is **Bearish** (the closing price is lower than it opened), then the body is usually coloured either black or red.



Candlestick Chart - Show relationships between open, high, low, and close prices with a specific focus on trends.
Also known as a Japanese Candlestick Chart.

- Candlestick Charts are great for detecting and predicting market trends over time and are useful for interpreting the day-to-day sentiment of the market, through each candlestick symbol's colouring and shape. For example, **the longer the body is, the more intense the selling or buying pressure is.**

- While, a **very short body, would indicate that there is very little price movement** in that time period and represents consolidation.

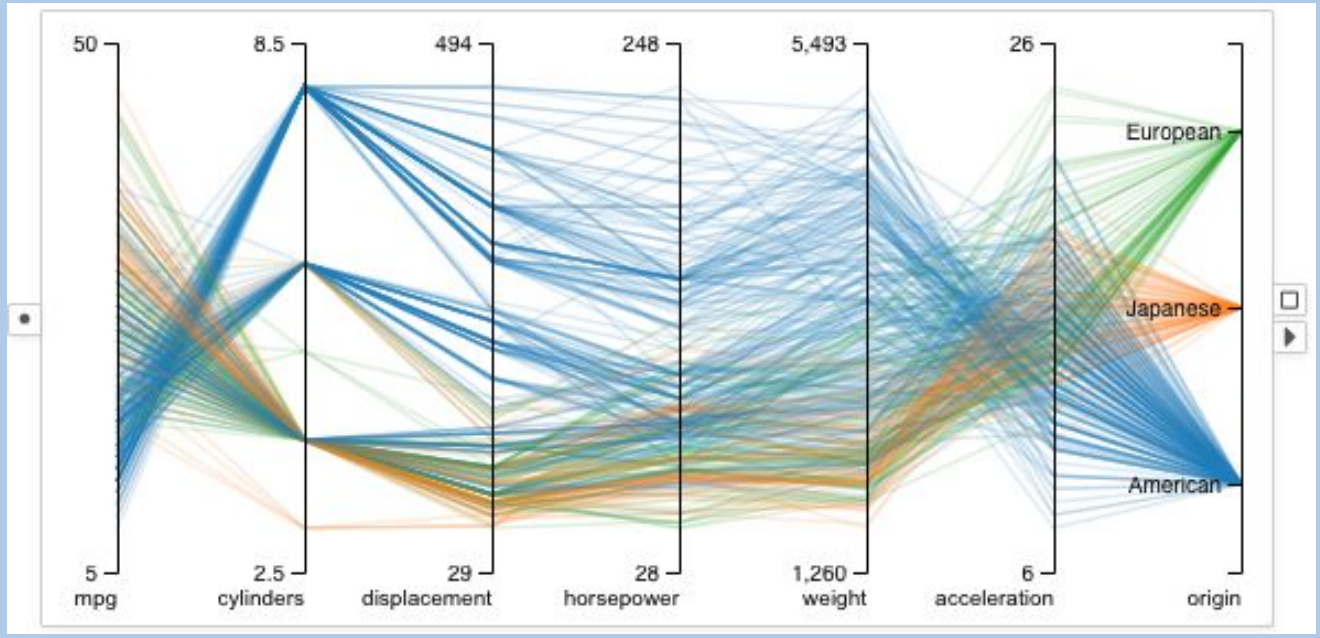


6. Ranking - Ordering items from highest to lowest (or vice versa) based on specific criteria to identify top performers, priorities, or hierarchical relationships.

Data Type	Variables	Objective	Example Question	Best Chart Type(s)
Categorical	1 Variable	Ranking	*"Which sales reps rank highest to lowest?" • "What's the order from best to worst performing category?"*	**Bar Chart** · Bullet Graph · Dot Plot · Slope Chart
Mixed	Special	Ranking	*"How do cities rank considering multiple livability factors?" • "What's the multi-criteria ranking?"*	**Parallel Coordinates** · Radar Chart · Slope Chart · Multi-criteria Chart

Parallel Coordinates Plot - Show relationships between multiple variables using parallel axes.

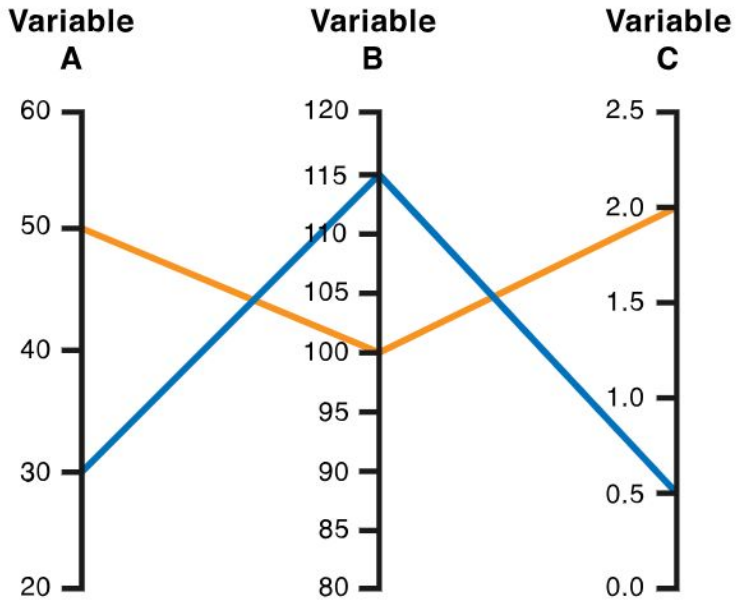
- This type of visualisation is used for plotting multivariate, numerical data. Parallel Coordinates Plots are **ideal for comparing many variables together** and seeing the relationships between them.
- For example, if you had to **compare an array of products with the same attributes** (comparing computer or car specs across different models).



Parallel Coordinates Plot - Show relationships between multiple variables using parallel axes.

- **Each axis can have a different scale,** as each variable works off a different unit of measurement, or all the axes can be normalised to keep all the scales uniform.

- Values are plotted as a series of lines that are connected across all the axes. This means that **each line is a collection of points placed on each axis, that have all been connected.**



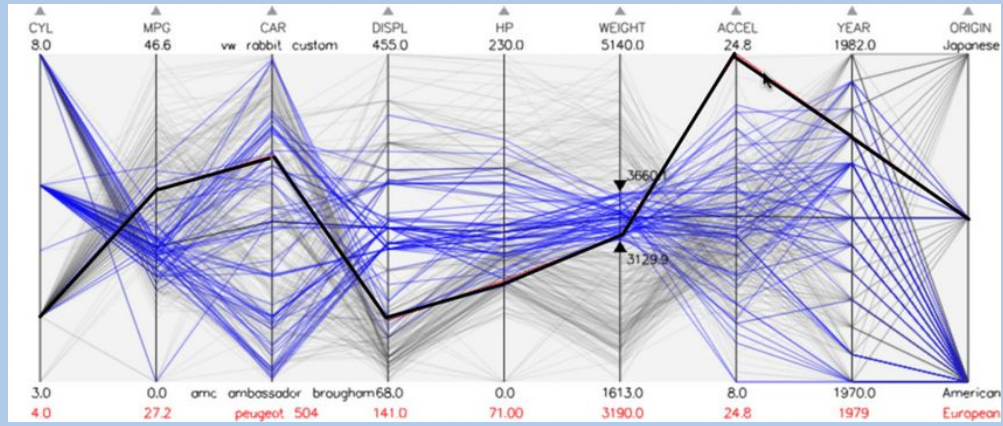
Data			
	Variable A	Variable B	Variable C
Item 1	50	100	2.0
Item 2	30	115	0.5

Parallel Coordinates Plot - Show relationships between multiple variables using parallel axes.

- The downside to Parallel Coordinates Plots, is that they can become over-cluttered and therefore, illegible when they're very data-dense.

- The best way to remedy this problem is through interactivity and a technique known as **"brushing".**

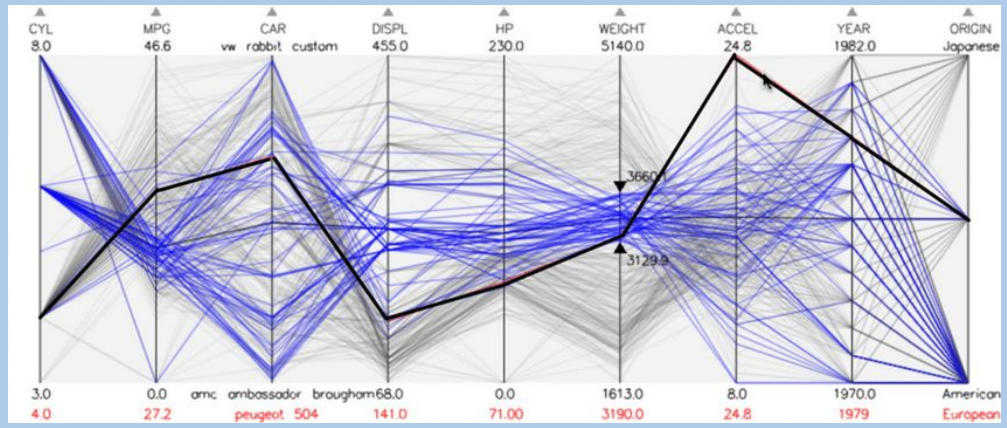
- **Brushing** highlights a selected line or collection of lines while fading out all the others. This allows you to isolate sections of the plot you're interested in while filtering out the noise.



Parallel Coordinates Plot - Show relationships between multiple variables using parallel axes.

Design Practices

- **Use Colors or Styles -** Differentiate data series or categories using colors or line styles to enhance readability.
- **Manage Line Density -** Adjust transparency or line thickness if many lines overlap to avoid clutter and improve clarity.
- **Interactive Features -** Consider interactive features (like brushing or filtering) to allow users to explore specific data subsets more effectively.

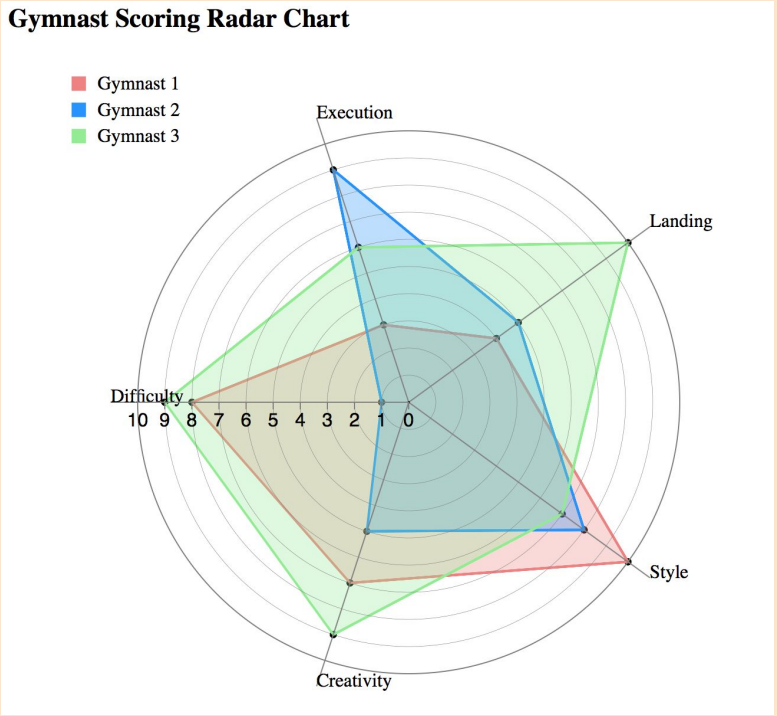


Radar Chart - Show relationships between several variables on a radial grid.

Also known as Spider Chart, Web Chart, Polar Chart, Star Plots.

- Radar Charts are a way of comparing multiple quantitative variables. This makes them **useful for seeing which variables have similar values** or if there are any outliers amongst each variable.

- Radar Charts are also useful for **seeing which variables are scoring high or low** within a dataset, making them suited for displaying performance.

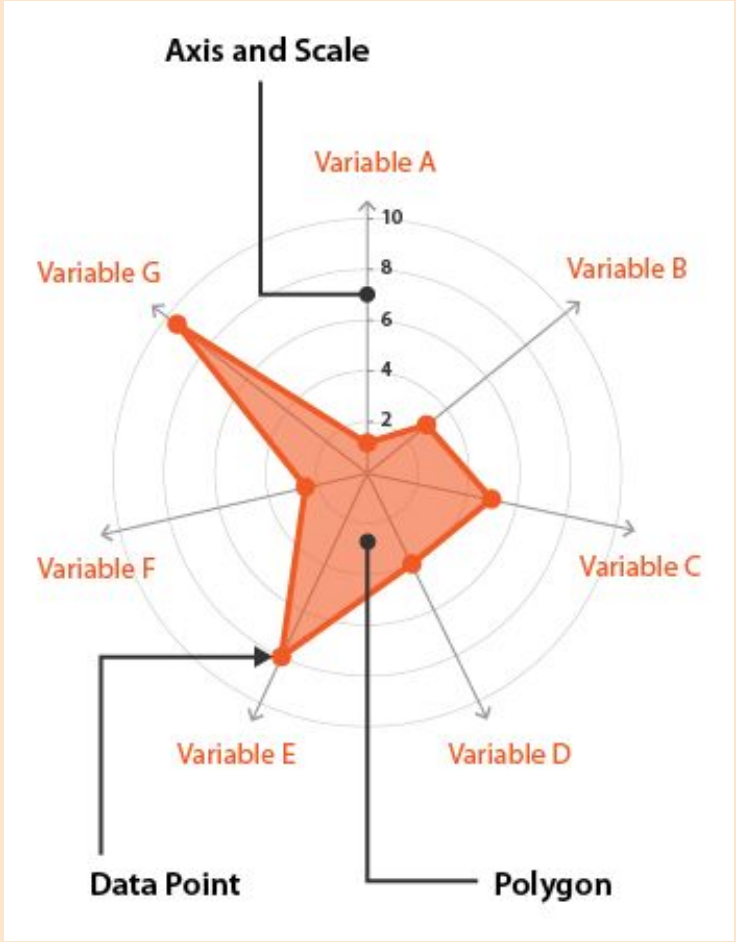


Radar Chart - Show relationships between several variables on a radial grid.

Also known as Spider Chart, Web Chart, Polar Chart, Star Plots.

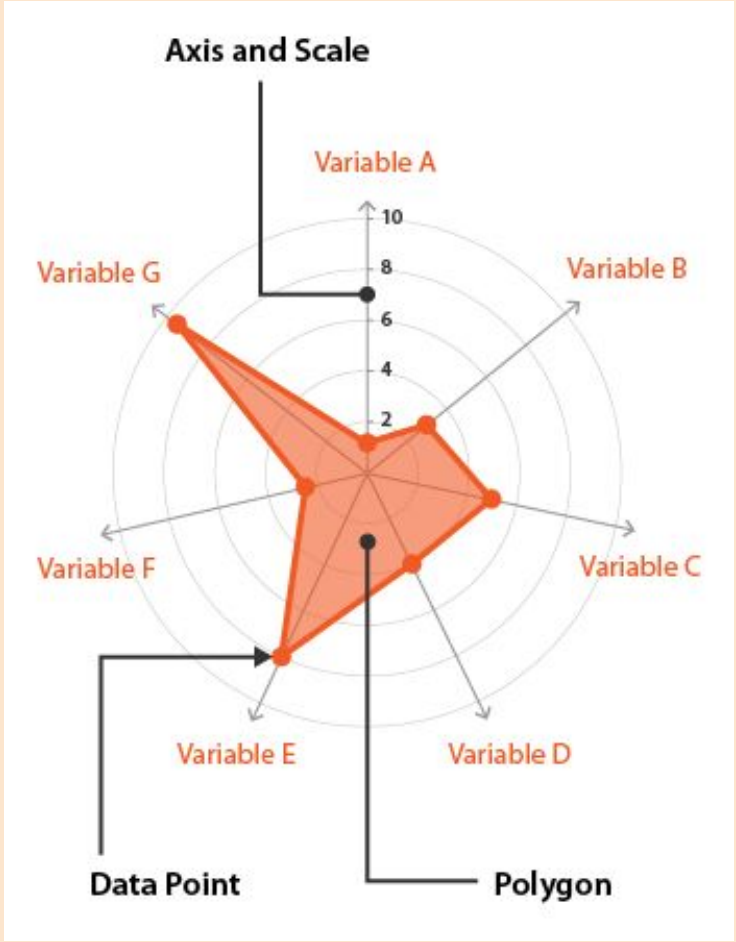
- Each variable is provided with an **axis that starts from the centre.** All axes are arranged radially, with equal distances between each other, while maintaining the same scale between all axes.

- Grid lines that connect from axis to axis are often used as a guide. Each variable value is plotted along an individual axis and all the variables in a dataset and connected to form a polygon.



Radar Chart - Show relationships between several variables on a radial grid.

Also known as Spider Chart, Web Chart, Polar Chart, Star Plots.



Design Practices

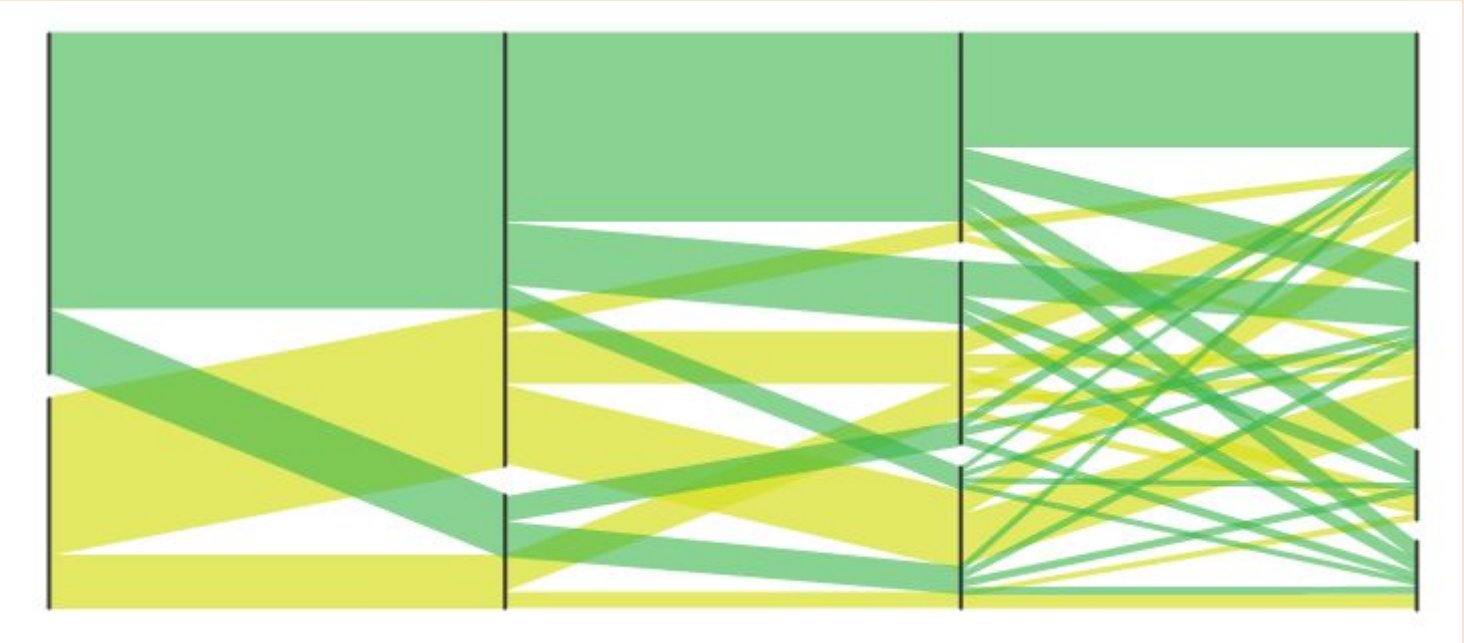
- **Distinct Lines or Areas -** Use different colors or patterns for each data series to ensure clarity and differentiation.
- **Manage Data Overlap -** Avoid overcrowding by limiting the number of variables and series to maintain readability.

7. Flow/Process - Tracking movement, transformation, or progression through stages, systems, or processes, showing how quantities flow from one state to another.

Data Type	Variables	Objective	Example Question	Best Chart Type(s)
Numerical	Special	Flow/Process	*"How do financial values cascade through processes?" • "What's the step-by-step transformation of values?"*	**Waterfall Chart** · Bridge Chart · Cascade Chart · Funnel Chart
Time-Series	Special	Flow/Process	*"What's our project timeline with dependencies?" • "How do tasks flow through time?"*	**Gantt Chart** · Timeline · Timetable · PERT Chart
Spatial	2 Variables	Flow/Process	*"How do people migrate between cities?" • "What are the flows between geographic locations?"*	**Flow Map** · Connection Map · Arc Diagram · Sankey Diagram
Mixed	2 Variables	Flow/Process	*"How does budget flow between departments?" • "How do categorical flows change over time?"*	**Sankey Diagram** · Alluvial Diagram · Parallel Sets · Flow Chart

Parallel Sets - visually represent proportions and categories through bands, making it easier to follow how data is distributed across multiple categorical variables.

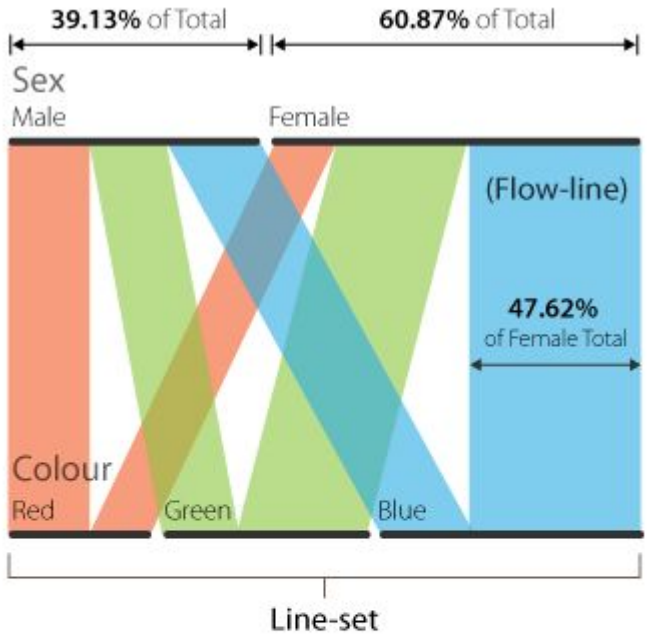
- Each line set corresponds to a dimension, and its values and categories are represented in each line divide in that line set.



Parallel Sets - visually represent proportions and categories through bands, making it easier to follow how data is distributed across multiple categorical variables.

How to Read a Parallel Sets Chart

- **Axes -** Each vertical line or axis represents a categorical variable. The categories within each axis are aligned horizontally.
- **Bands -** Bands flow between the axes, connecting categories from one axis to another. The width of a band is proportional to the number of data points or the percentage of the total data belonging to that category.
- **Flow -** The flow of bands between categories shows how data transitions from one category to another. For example, in a dataset of car buyers, you might see how different age groups (axis 1) prefer certain types of cars (axis 2), and further, which models they choose (axis 3).

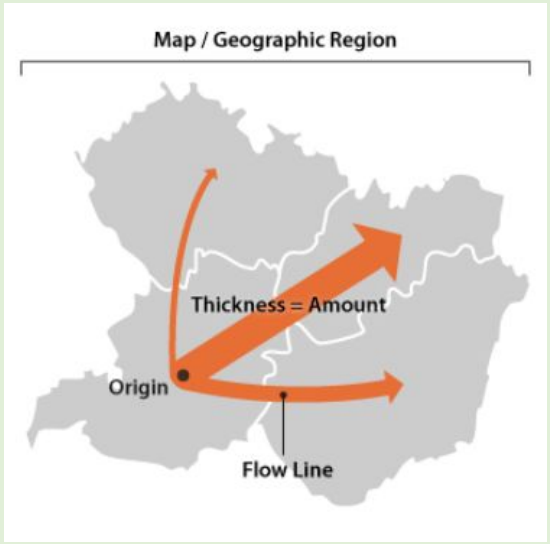


Sex	Colour	Count	% of Sex TOTAL
Male	Red	35	32.41
	Green	33	30.56
	Blue	40	37.04
Female	Red	28	16.67
	Green	60	35.71
	Blue	80	47.62

Flow Map

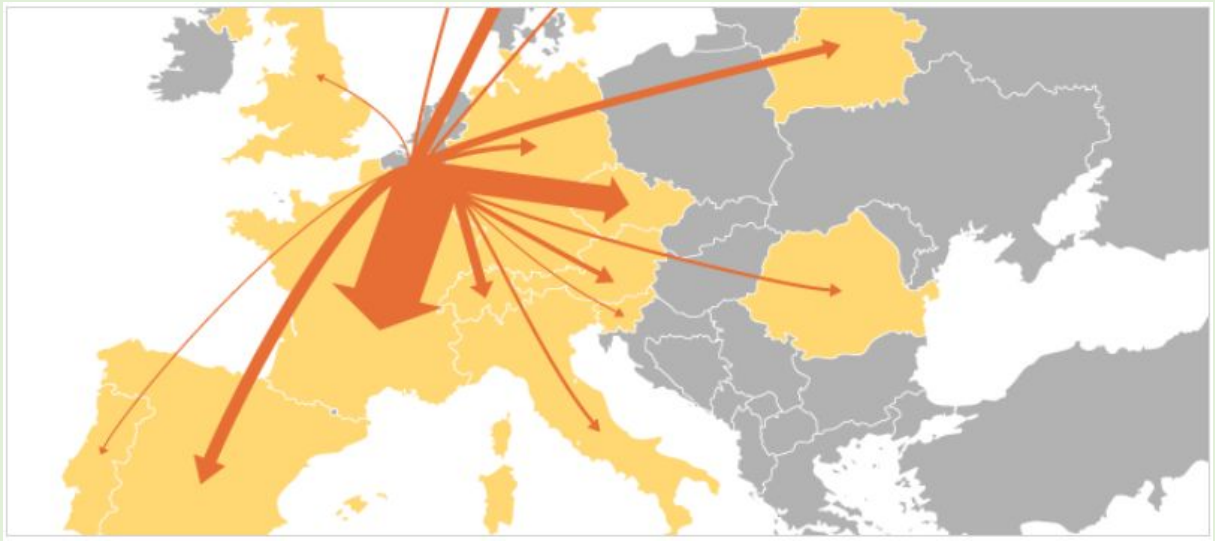
- Flow Maps geographically show the movement of information or objects from one location to another and their amount.

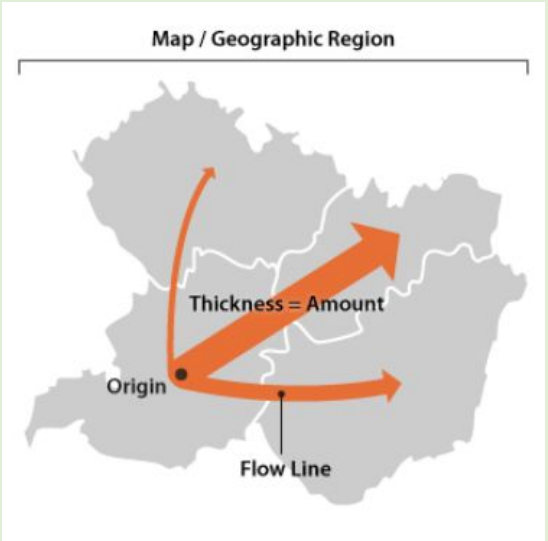
- Typically Flow Maps are used to show the migration data of people, animals and products. The magnitude or amount of migration in a single flow line is represented by its thickness.



Flow Map

- Flow Maps are drawn from a point of origin and branch out of their “flow lines”. Arrows can be used to show direction, or if the movement is incoming or outgoing.
- Drawing flow lines without arrows can be used to represent trade going back-and-forth.
- Merging/bundling flow lines together and avoiding crossovers can help to reduce visual clutter on the map.





3 simple step to choosing the right chart.

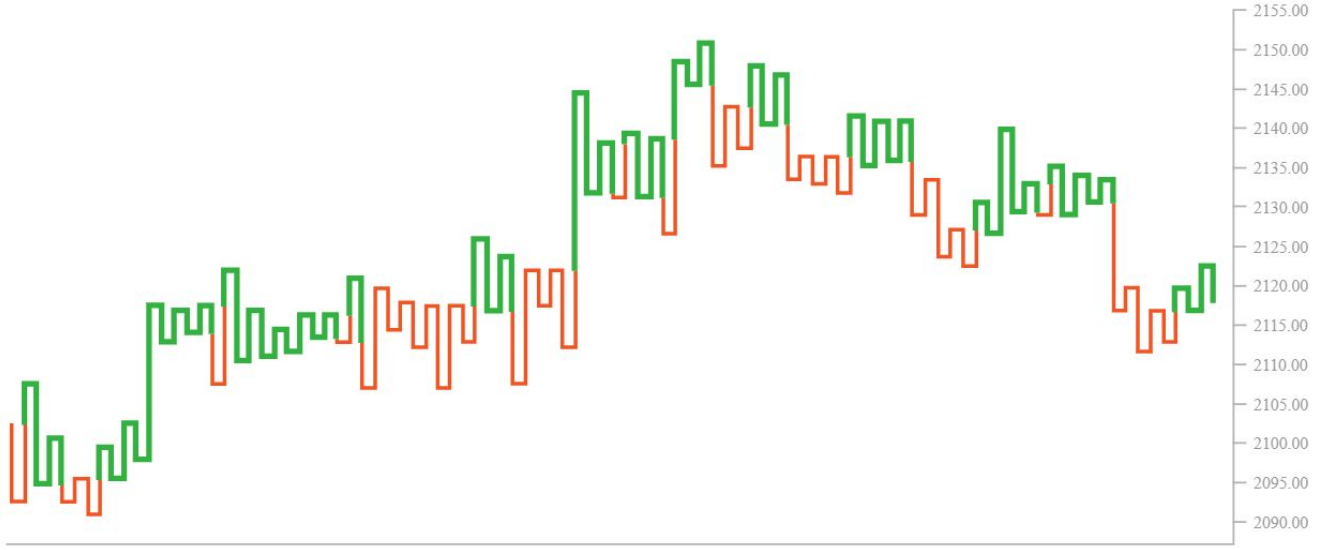
more graphs here: https://datavizcatalogue.com/

QUIZ

1.It is most frequently used to show trends and analyse how the data has changed over time.

a) Scatter Plots
b) Parallel Coordinates Plot
c) Line Graph
d) Bullet Graph

2. Identify the graph



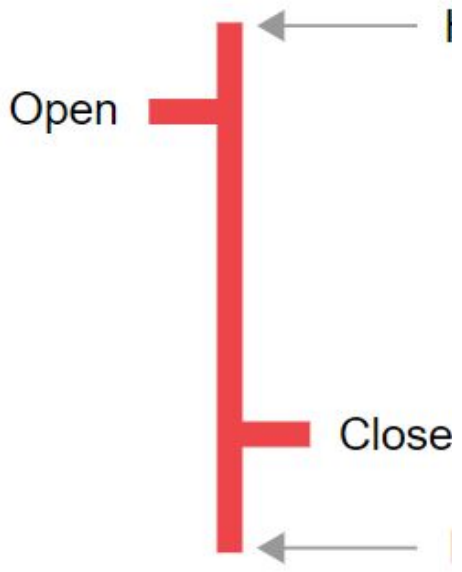
a) OHLC Chart

b) Kagi Chart

c) Candlestick Chart

3. The closing price is lower than it opened

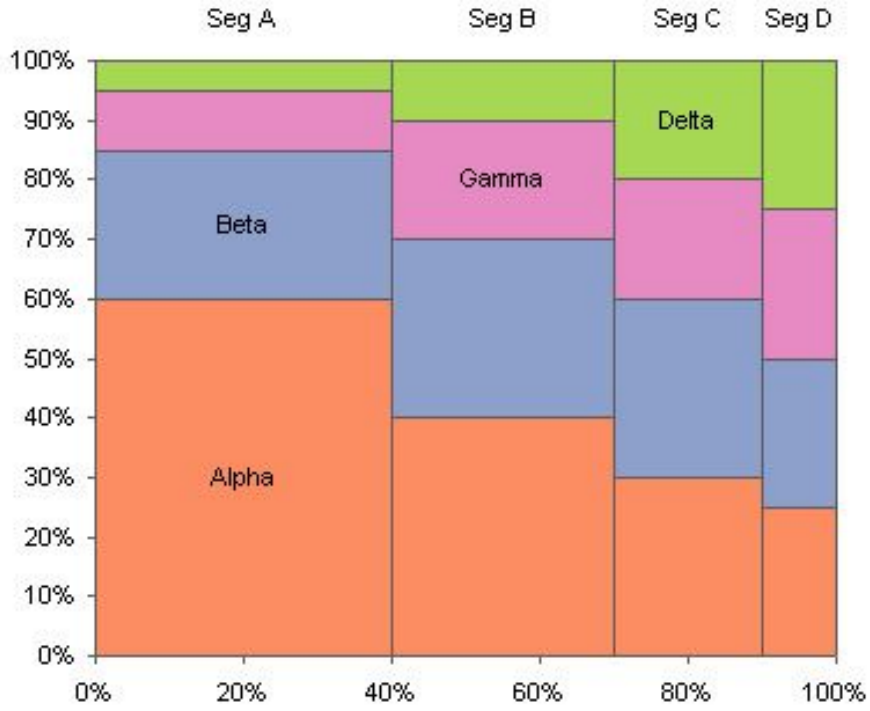
a) bearish
b) bullish
c) shoulder



4. In a pie chart, how is the largest slice positioned?

A) At the left (9 o'clock position)
B) At the right (3 o'clock position)
C) At the bottom (6 o'clock position)
D) At the top (12 o'clock position)

5. Identify the graph



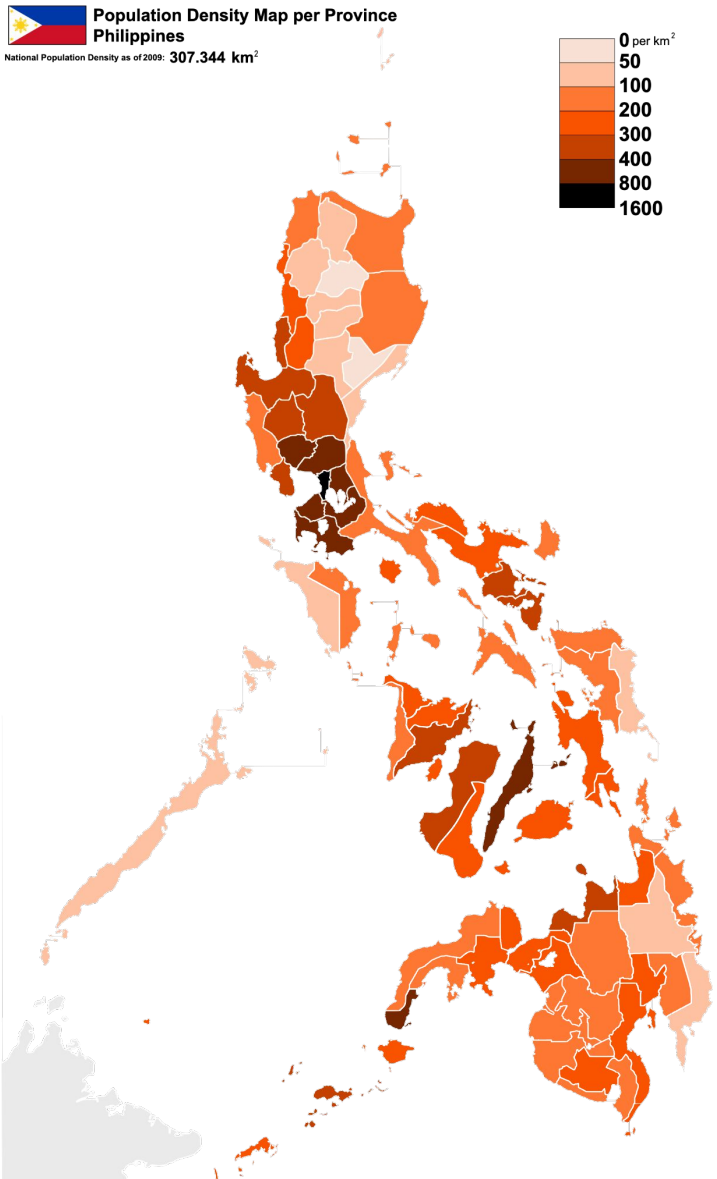
a) Parallel Set
b) Marimekko Chart
c) Treemap
d) Box Plot

6. What does a violin plot combine in its representation?

A) Box plot and Histogram chart
B) Boxplot and Line graph
C) Box plot and Scatter plot
D) Box plot and Density plot

7. Identify the graph

a) Correlation Plot
b) Choropleth Map
c) Heat Map
d) Connection Map



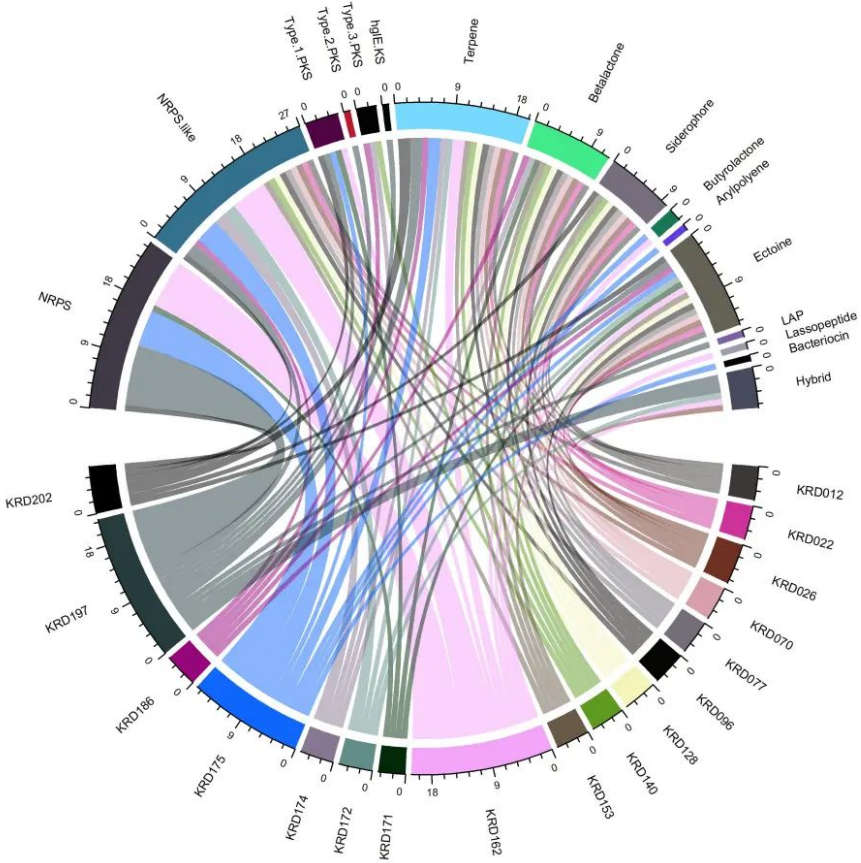
8. What does each rectangle represent in a Treemap?

a) A time period
b) A geographical location
c) A category with size proportional to data value
d) A relationship between variables

9. In graphs, brushing refers to:

a) Using more colors
b) Adding more axes
c) Highlighting selected lines while fading others
d) Making lines thicker

10. Identify the graph



a) Circular diagram
b) Connection diagram
c) Arc diagram
d) Chord diagram

Data Visualization Case Study

You are a data analyst tasked with creating visualizations from a provided dataset. Your goal is to tell a clear story with the data using appropriate charts and design principles from the lecture.

- **Analyze the Dataset**
- **Create Visualizations**
- **Write a Summary Report**

Detailed instructions will be provided in MOLE.