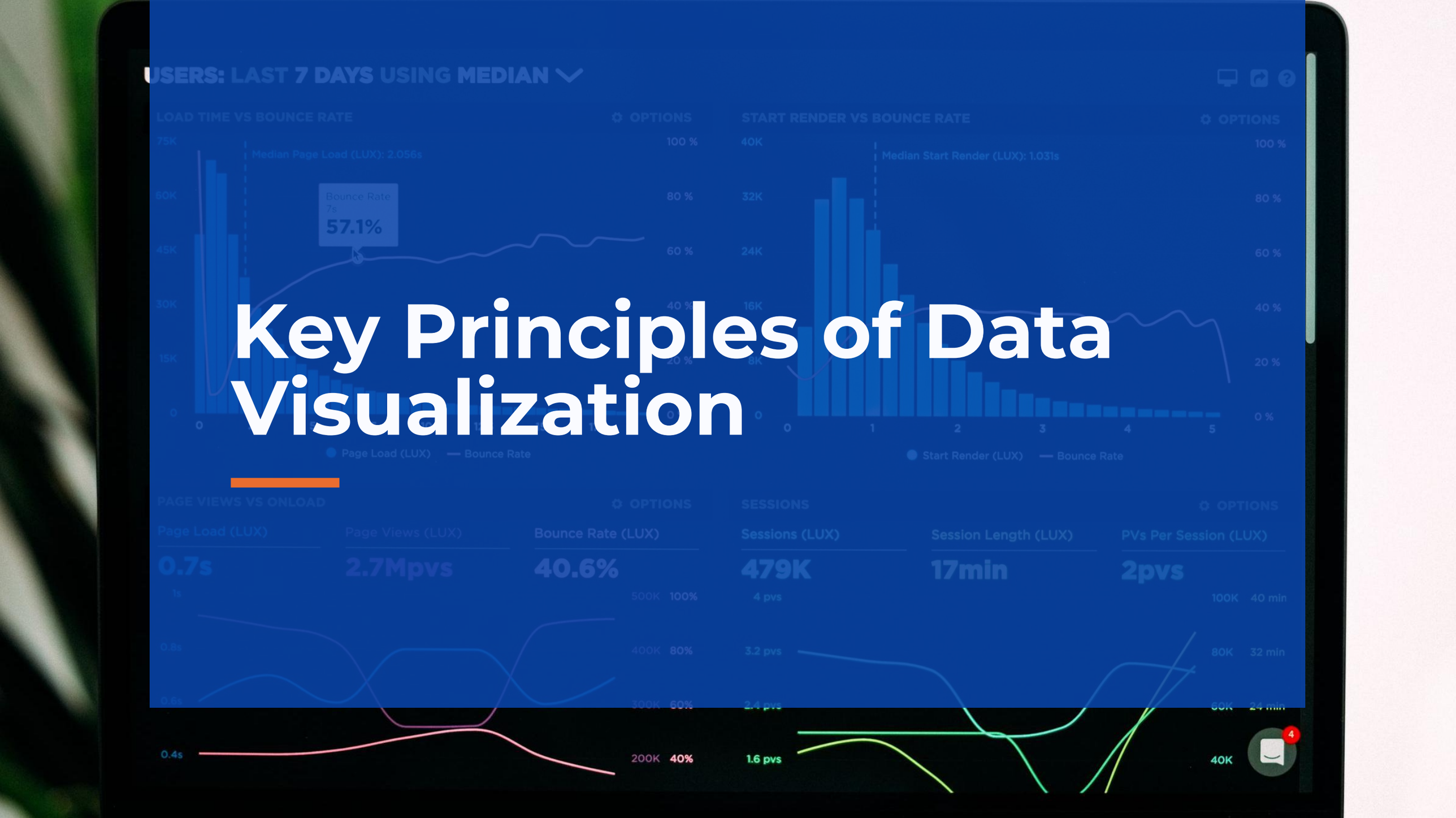


Key Principles of Data Visualization

Research by 3M Corporation found:

Human brains process visual data 60,000 times faster.

KEY PRINCIPLES OF DATA VISUALIZATION

Strive for **CLARITY** & **SIMPLICITY**

- Maximize **impact**, minimize **noise**
- If it doesn't **add value** or **serve a purpose**, get rid of it

Focus on creating a **NARRATIVE**

- Don't just show data, **tell a story**
- Communicate key insights **clearly, quickly** and **powerfully**

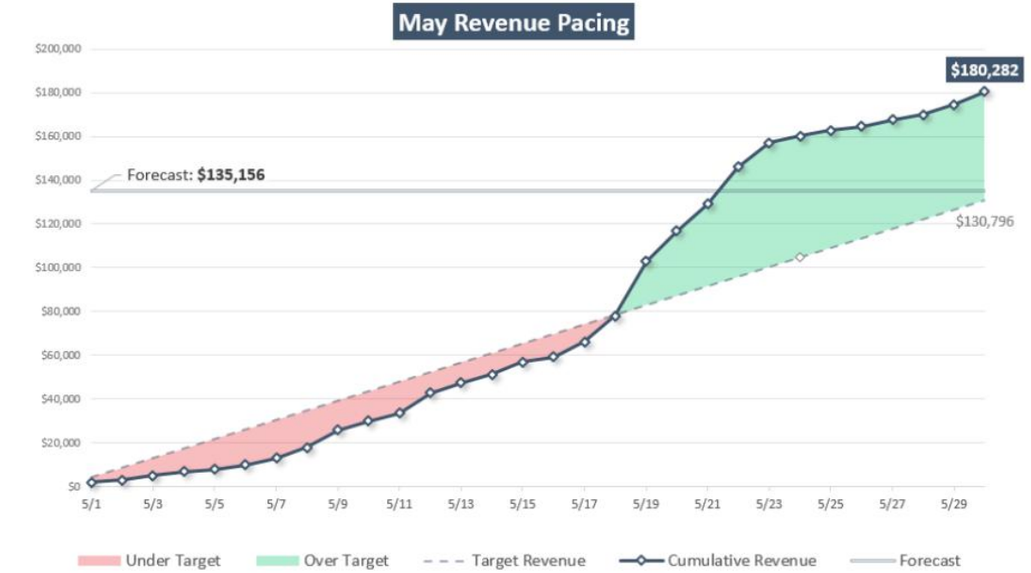
Strike a balance between **DESIGN** & **FUNCTION**

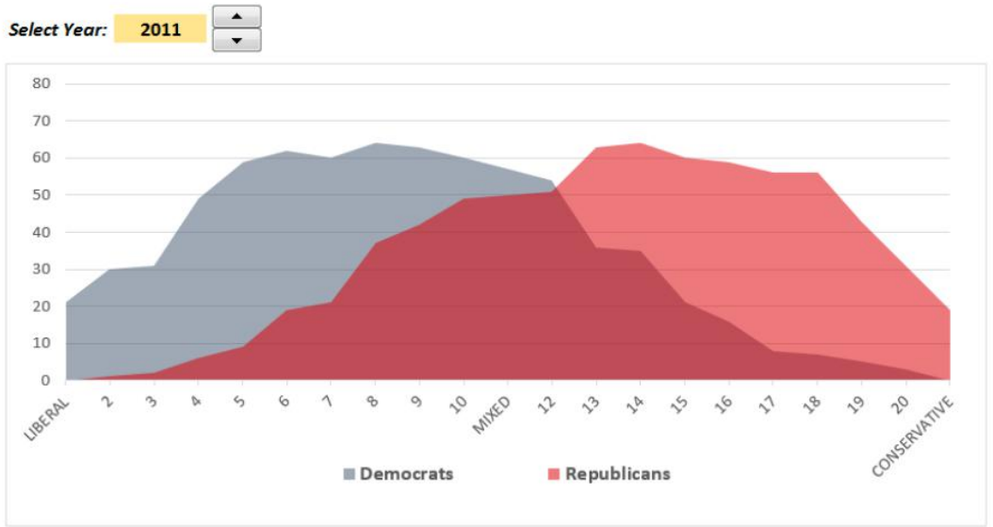
- *Selecting the right type of chart is **critical***
- ***Beautiful** is good, **functional** is better, **BOTH** is ideal*

The GOOD, the BAD, and the UGLY

THE GOOD

Dynamic formatting helps to strengthen the story





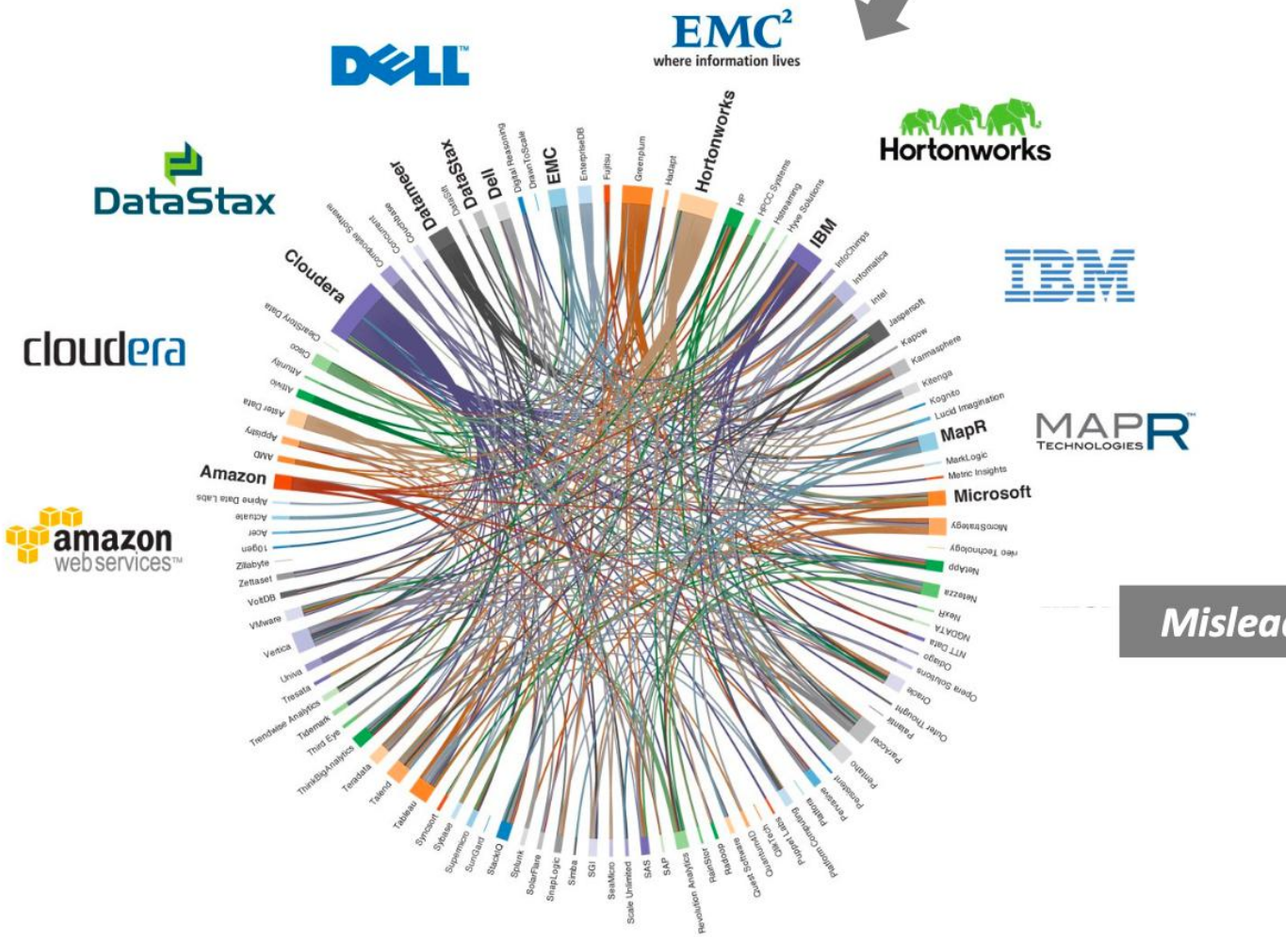
Clean, simple visualization with animation over time

Simple, intuitive custom chart design

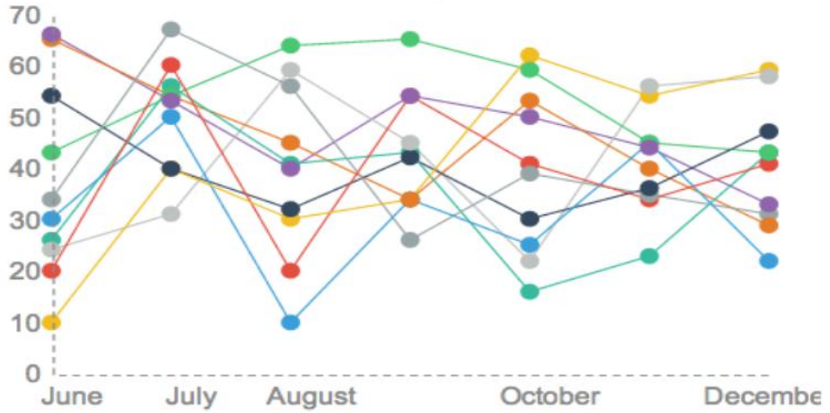
The GOOD, the BAD, and the UGLY

THE BAD

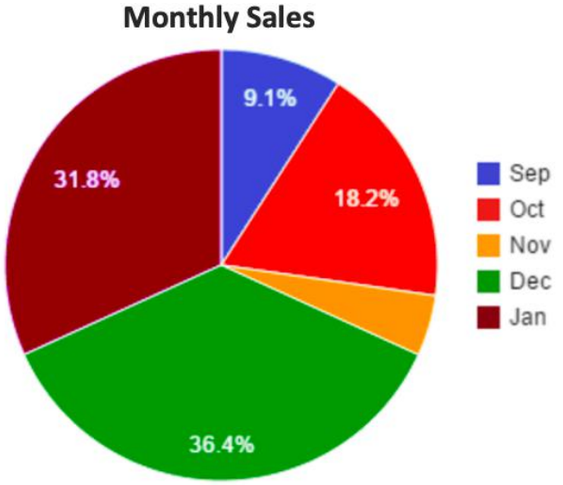
All design, no function



Busy, no clear narrative



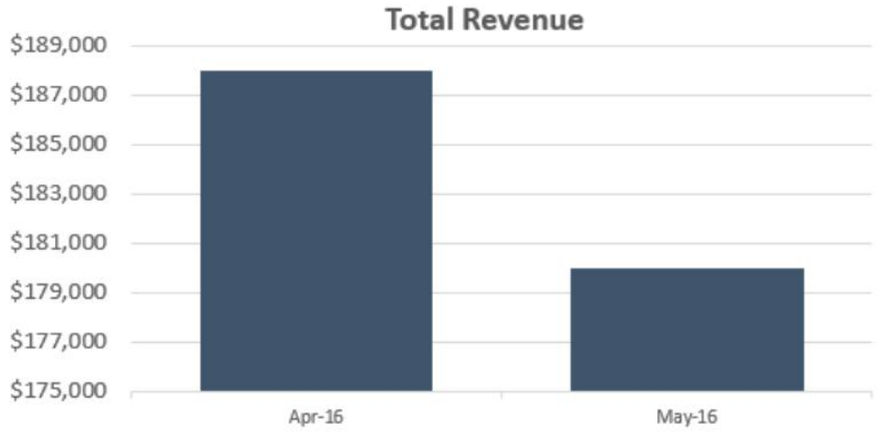
Misleading chart type



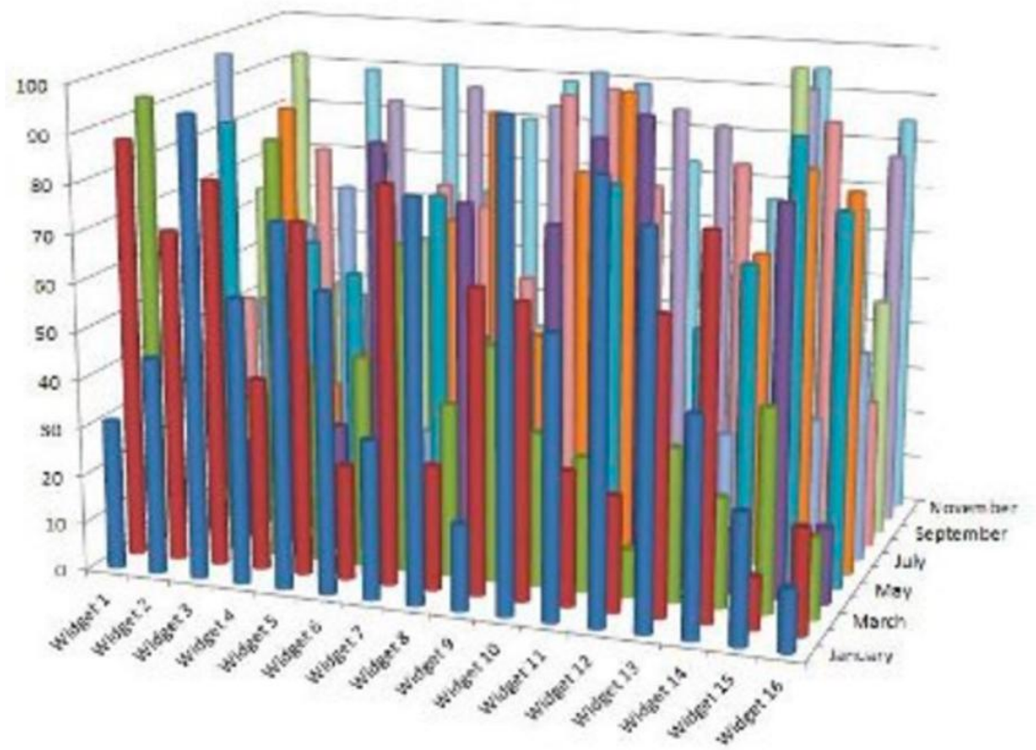
The GOOD, the BAD, and the UGLY

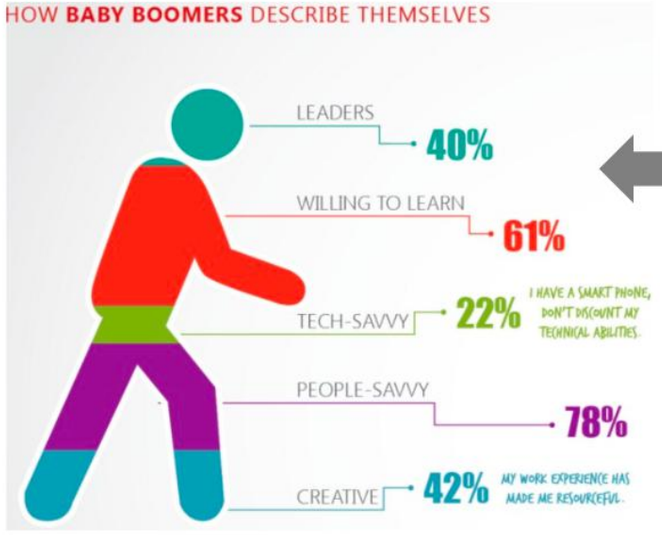
THE UGLY

Misleading y-axis scale



Too many elements, distracting 3D design





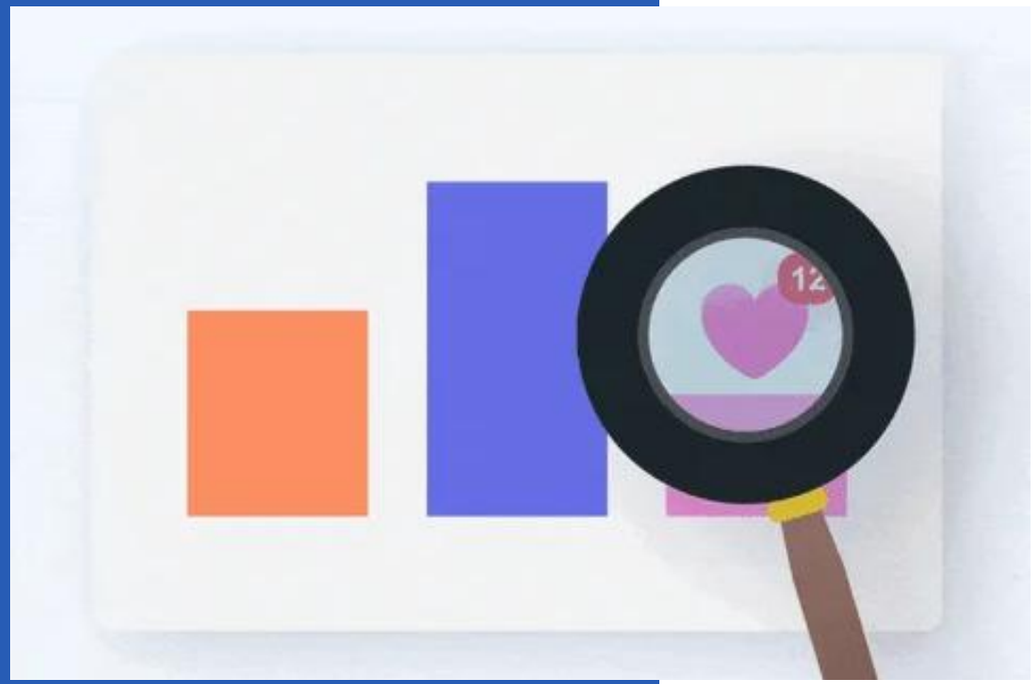
Improper use of percentages & inconsistent scaling

The 3 Key Questions:

1. What **type of data** are you working with?
   - Integer, real, categorical, time-series, geo-spatial, etc.

2. **What** are you trying to communicate?
   - Relationship, comparison, composition, distribution, trending, etc.

3. Who is the **end user** consuming this information?
   - Analyst, CEO, client, intern, etc.



BAR & COLUMN CHARTS

Bar and **column** charts are used to compare the quantity, frequency, or other measure (e.g., mean) for different categories or groups. The x-axis represents categories, and the y-axis represents quantities. Bar charts are *horizontal*, and column charts are *vertical*.

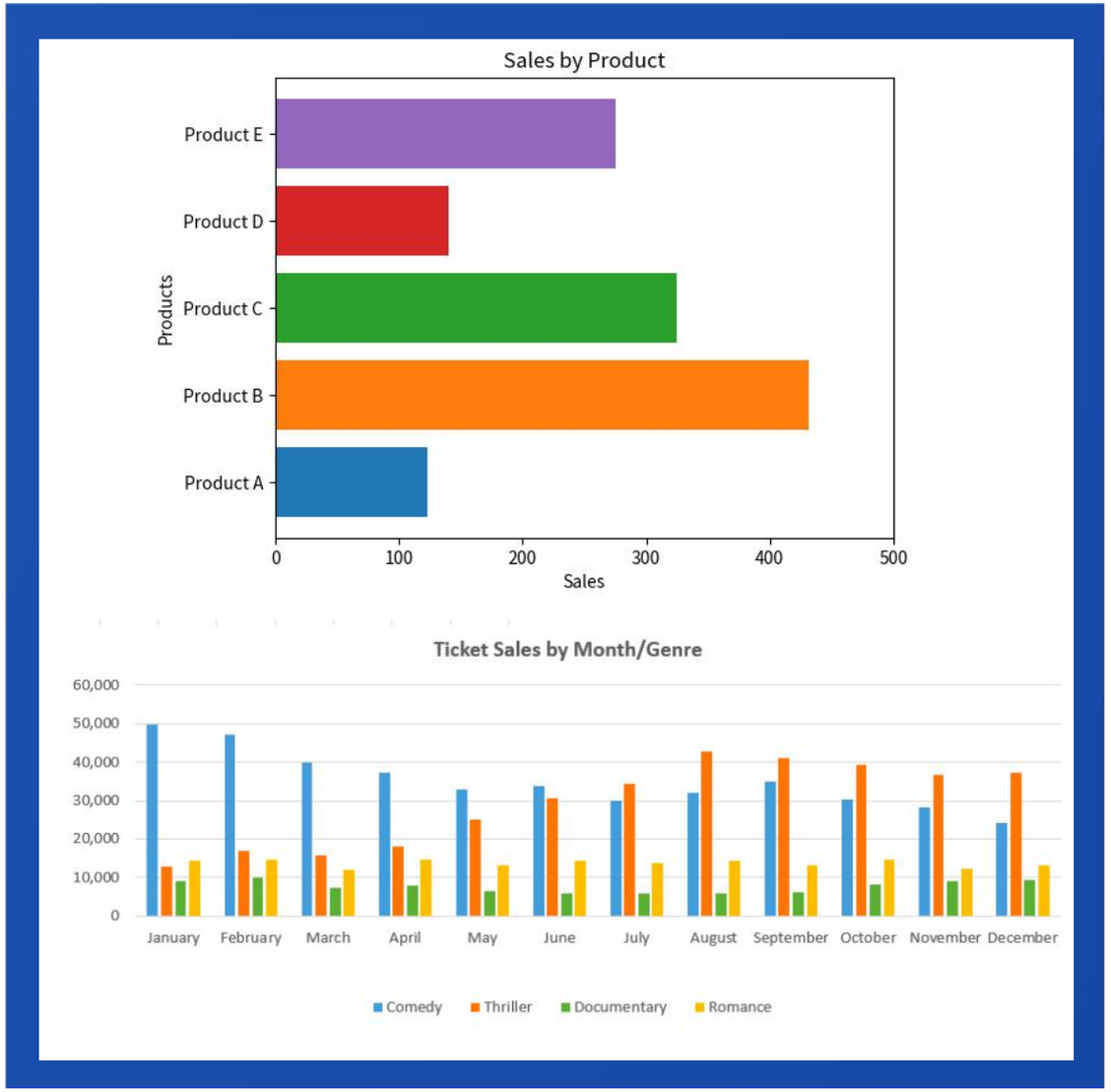
BAR & COLUMN CHARTS

Commonly used for:

- Comparing numerical data across categories

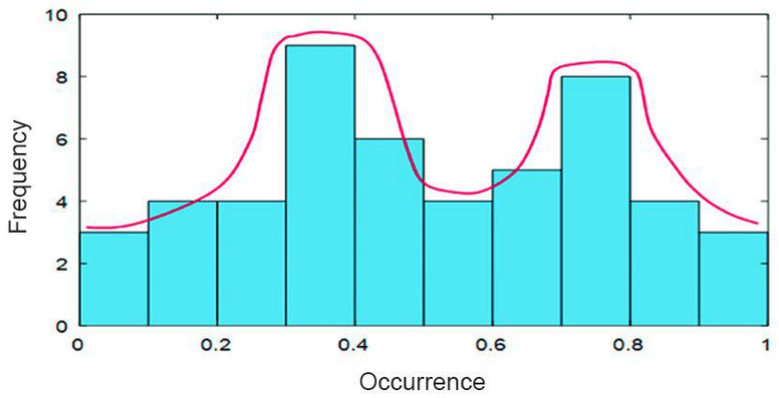
Examples:

- Total sales by product type
- Population by country
- Revenue by department, by quarter



Pro Tips:

- Use **stacked** or **clustered** bars/columns to group by subcategory or compare multiple metrics
- Create custom formatting rules to color-code bars/columns based on their values
- Avoid using these charts when dealing with continuous data or when you have too many categories, which can make the chart hard to read.

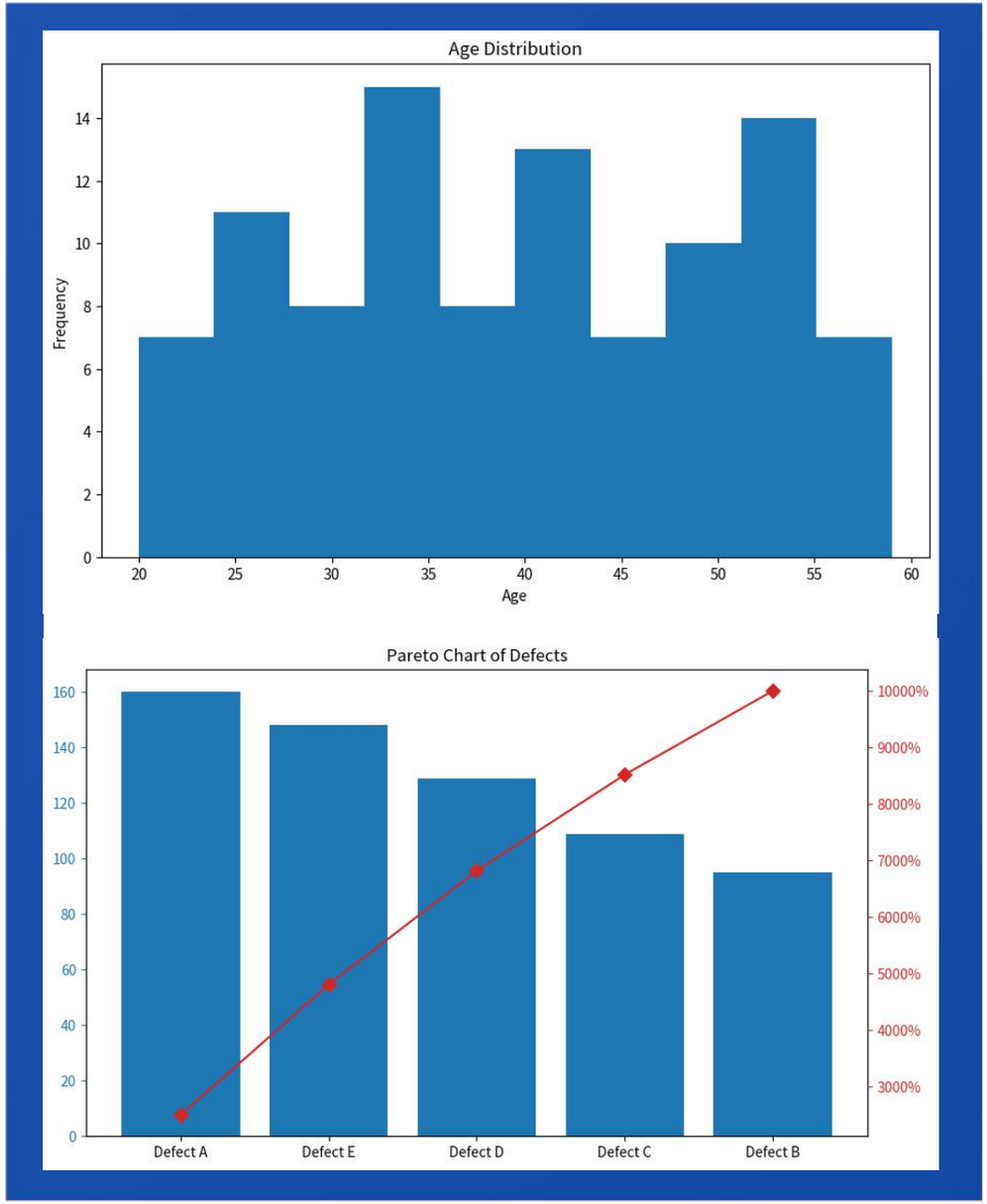


HISTOGRAMS & PARETO CHARTS

A **histogram** is a graphical representation that organizes a group of data points into a specified range. It is an accurate representation of the distribution of numerical data.

A **Pareto** chart, named after *Vilfredo Pareto*, is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line.

HISTOGRAMS & PARETO CHARTS



Commonly used for:

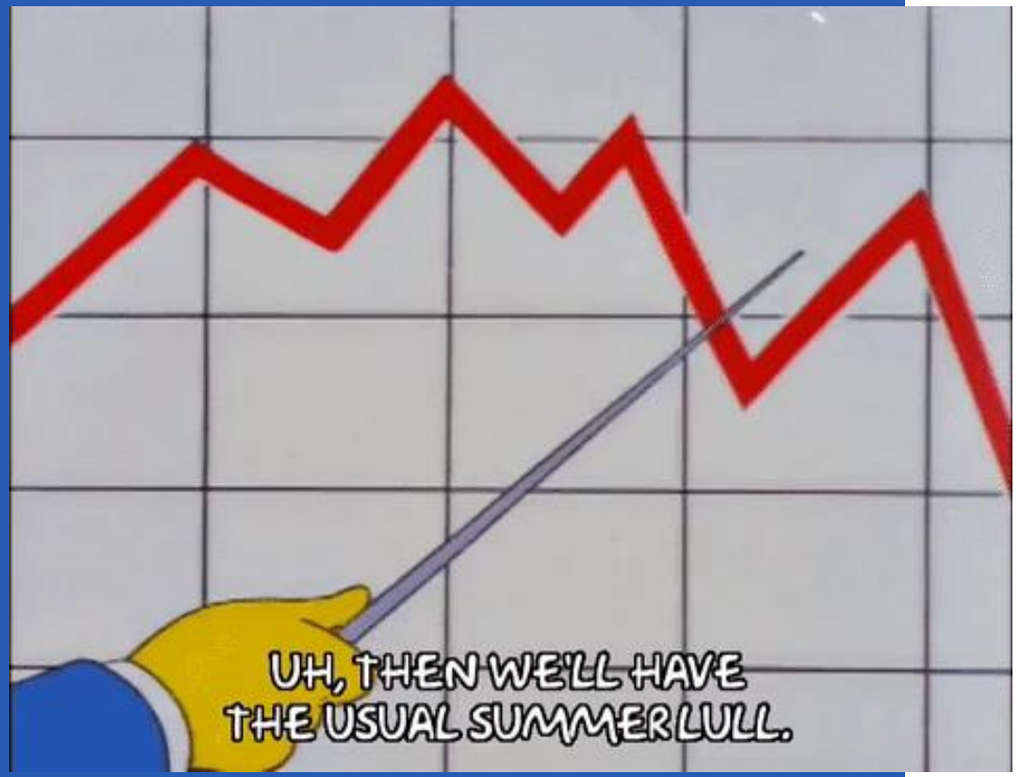
- Showing the distribution of a **continuous** data set
- Use a Pareto chart when you want to highlight the ***most important*** among a (typically large) set of factors

Examples:

- Frequency of test scores among students
- Distribution of population by age group
- Distribution of heights or weights
- In quality control, a Pareto chart often represents the most common sources of defects

Pro Tips:

- Adjust the bucket size to customize the grouping of values
- Use Pareto Charts to show the cumulative impact of each bucket, ordered by significance
- Avoid using a histogram when the data is nominal or the data ranges are not continuous
- Pareto charts are not suitable when there are many categories with similar frequencies



LINE CHARTS

A **line** chart or line plot is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields.

When to use: Use a line chart when you want to observe trends over time or other continuous variable. For example, observing the trend of sales over a year.

When not to use: Avoid using a line chart when you have nominal data or when your data does not have a logical order.

LINE CHARTS

Commonly used for:

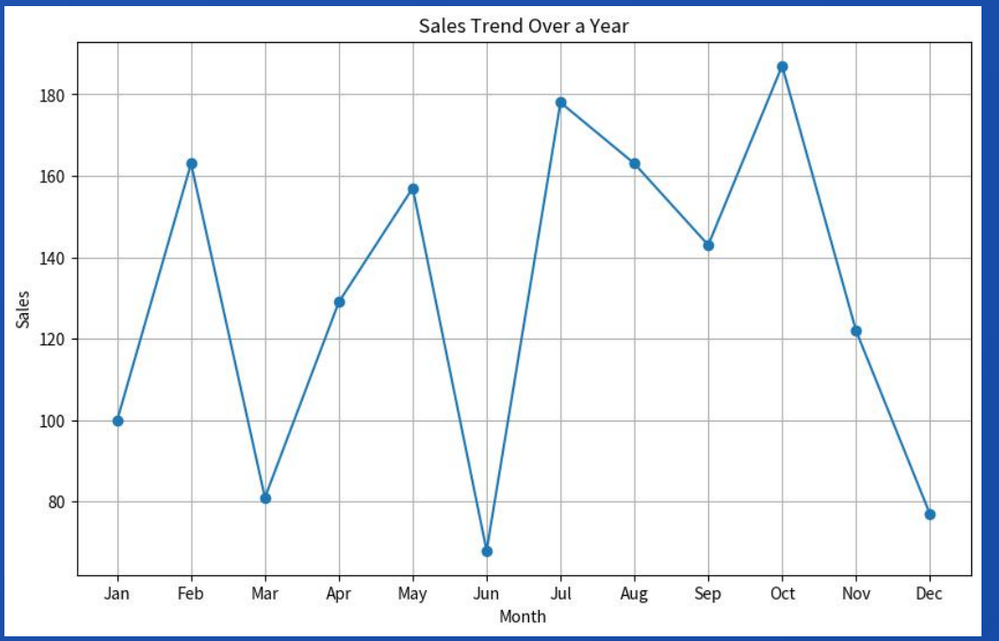
- Visualizing trends over time

Examples:

- Stock price by hour
- Average temperature by month
- Profit by quarter

Pro Tips:

- Use **linear** or **polynomial** trendlines to visualize patterns or forecast future periods
- Combine line & column charts to trend two variables on different scales



AREA CHARTS

An **area** chart or area graph displays graphically quantitative data. It is based on the line chart. The area between the line segments and the x-axis is filled with color or patterns. It represents the evolution of a numerical variable following another numerical variable.

When to use: Use an area chart when you want to show or compare a quantity or multiple quantities over time. For example, observing the trend of sales and costs over a year.

When not to use: Avoid using an area chart when you have categories that do not have a logical order or when the data is not continuous.

AREA CHARTS

Commonly used for:

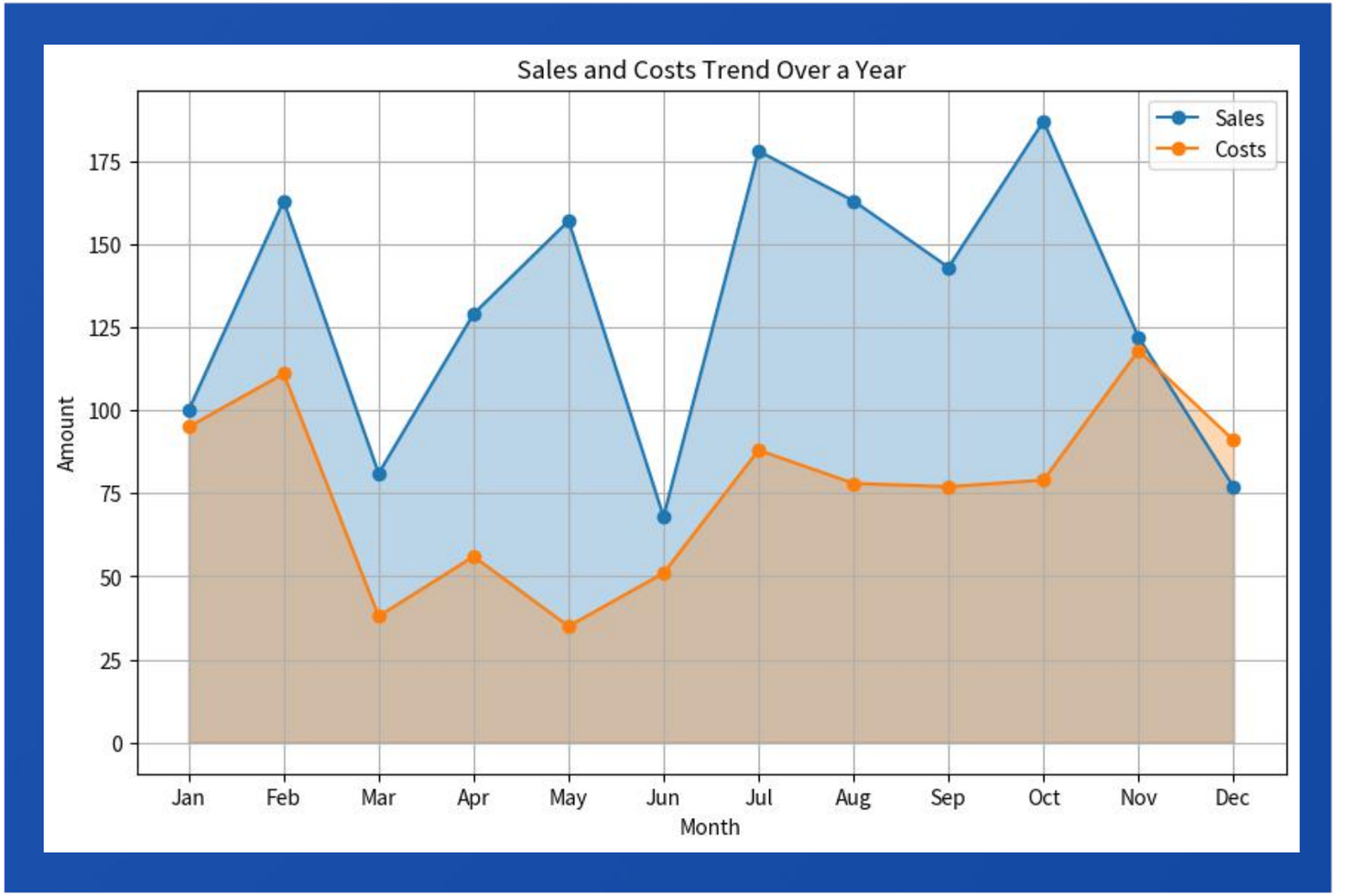
- Showing changes in data composition over time

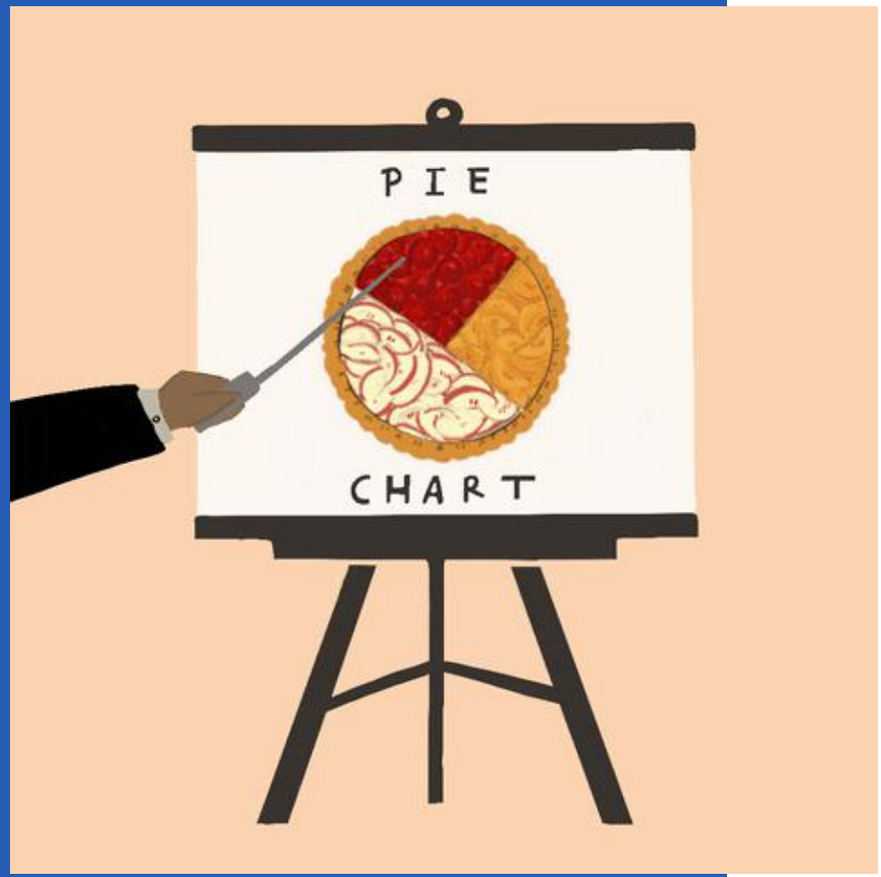
Examples:

- Sales by department, by month
- % of total downloads by browser, by week
- Population by continent, by decade

Pro Tips:

- Keep the number of unique categories relatively low (<6) to maintain clarity
- Use data validation and custom formatting to dynamically highlight specific data series





PIE & DONUT CHARTS

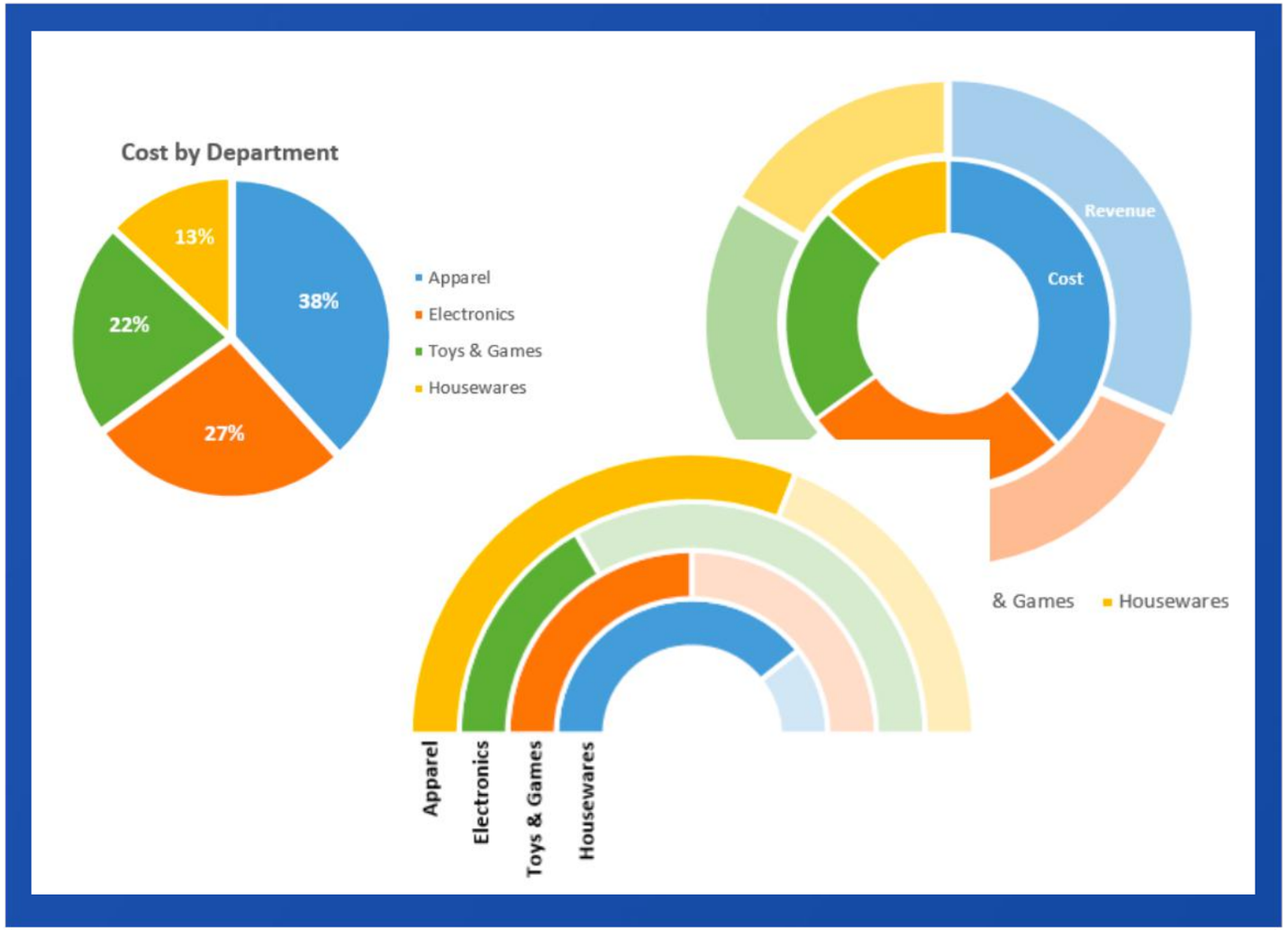
A **pie** chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.

A **donut** chart is a variation of the pie chart, with a blank center allowing for additional information about the data as a whole to be included.

When to use: Use a pie or donut chart when you want to show the proportion of parts to a whole. For example, showing the market share of different companies in an industry.

When not to use: Avoid using a pie or donut chart when you have too many categories, which can make the chart hard to read, or when the proportions are similar, which can make differences hard to see.

PIE & DONUT CHARTS



Commonly used for:

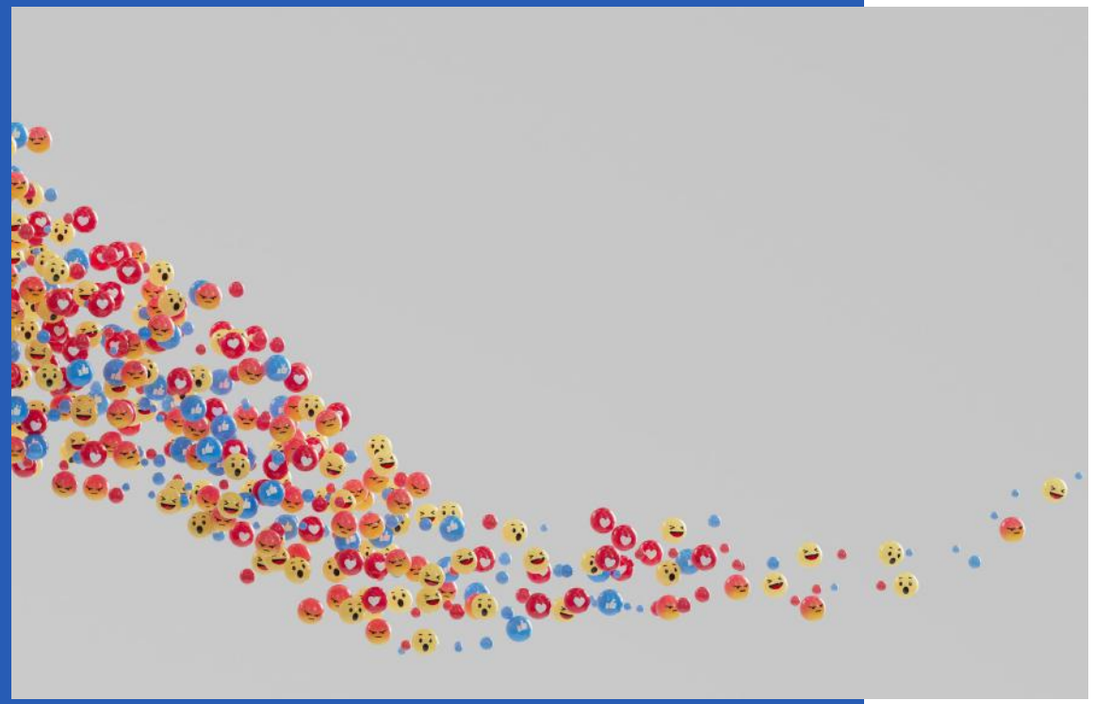
- Comparing proportions totaling 100%

Examples:

- Percentage of budget spent by department
- Proportion of internet users by age range
- Breakdown of site traffic by source

Pro Tips:

- Try to keep the number of slices small (<6) to maximize readability
- Use a donut chart to visualize more than one series at once, or use transparent segments to create a custom “race track” visualization



SCATTER PLOTS

A **scatter plot** uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

When to use: Use a scatter plot when you want to understand the relationship between two variables. For example, understanding the relationship between the age and income of individuals.

When not to use: Avoid using a scatter plot when the variables are not numeric or when the data does not have a logical order.

SCATTER PLOTS

Commonly used for:

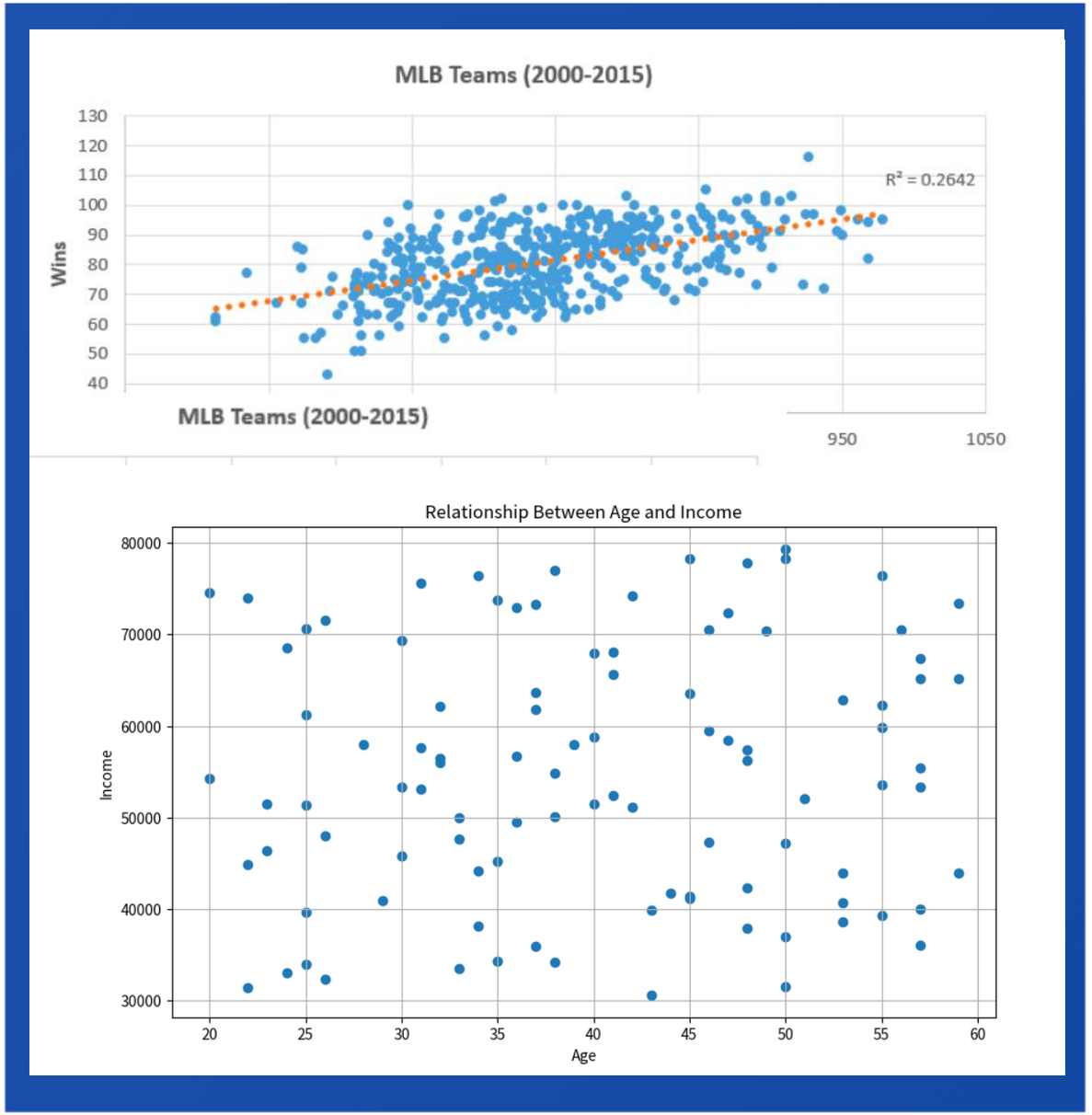
- Exploring correlations or relationships between series

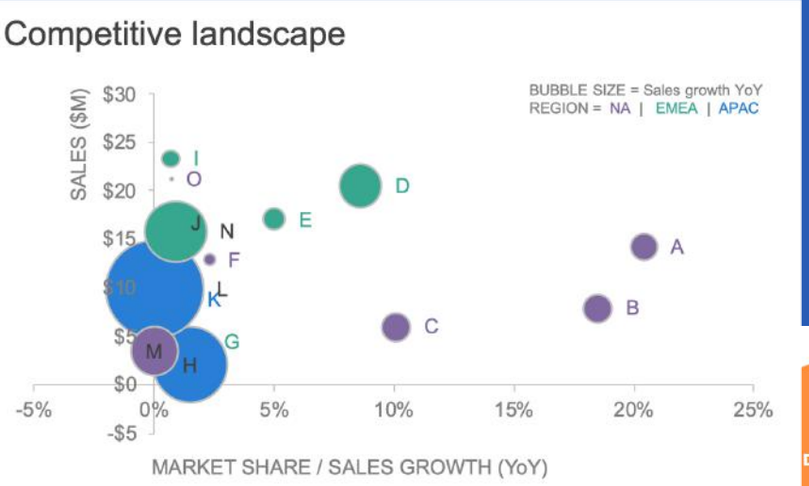
Examples:

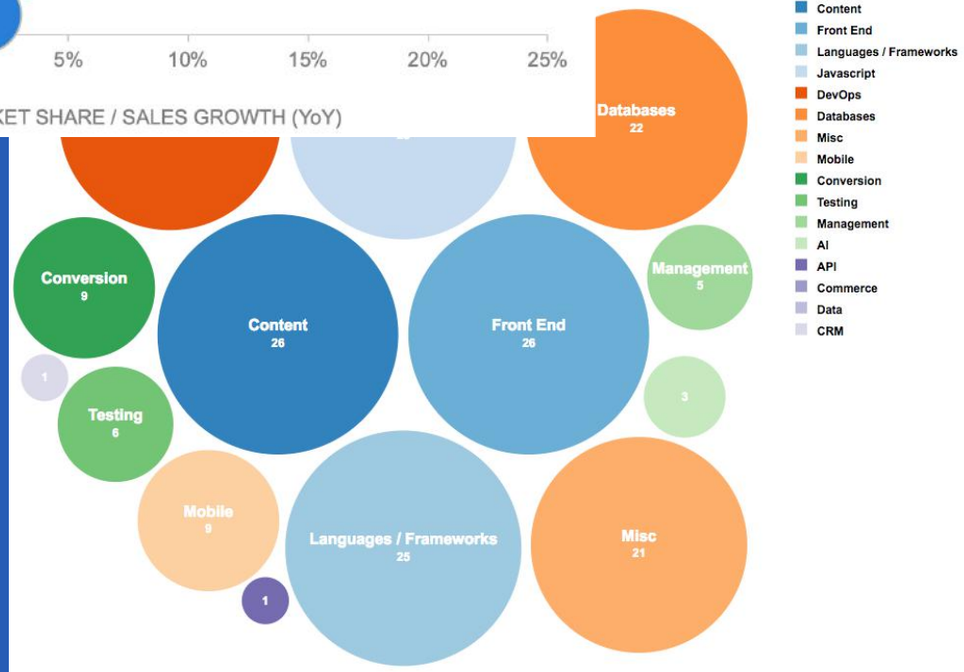
- Number of home runs and salary by baseball player
- Ice cream sales and average temperature by day
- Hours of television watched by age

Pro Tips:

- Add a trendline or line of best fit to quantify the correlation between variables
- Remember that correlation **does not** imply causation







BUBBLE CHARTS

A **bubble** chart is an extension of the scatter plot used to look at relationships between three numeric variables. Each dot in a bubble chart corresponds with a single data point, and the variables' values for each point are indicated by horizontal position, vertical position, and dot size.

When to use: Use a bubble chart when you want to visualize three dimensions of data. For example, visualizing the age, income, and education level of individuals.

When not to use: Avoid using a bubble chart when the variables are not numeric or when the data does not have a logical order.

BUBBLE CHARTS

Commonly used for:

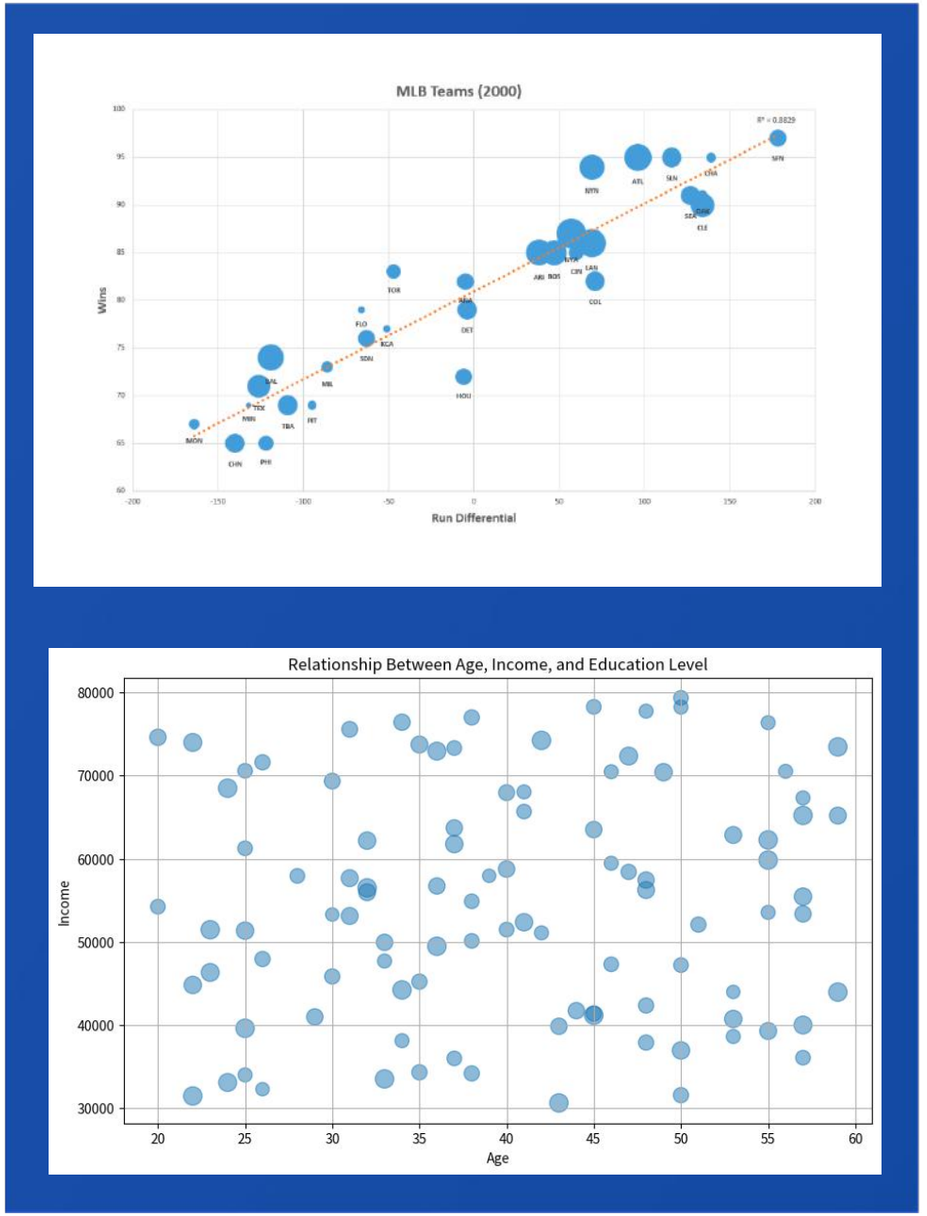
- Adding a third dimension (size) to a scatter plot format

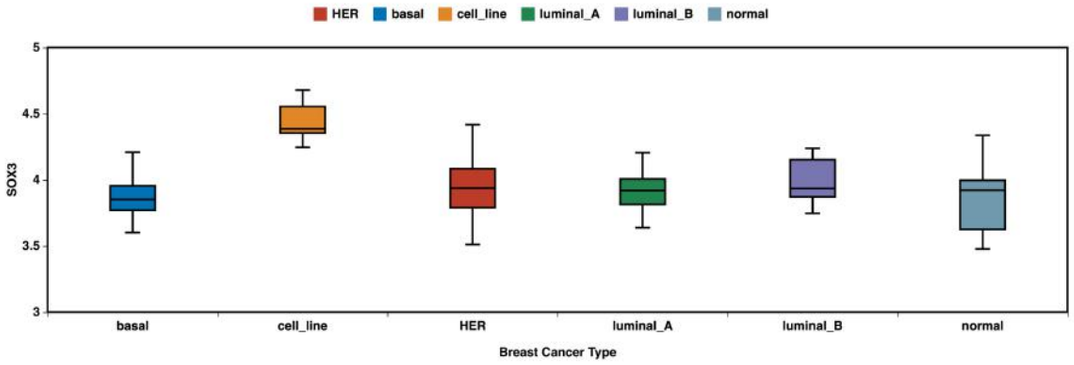
Examples:

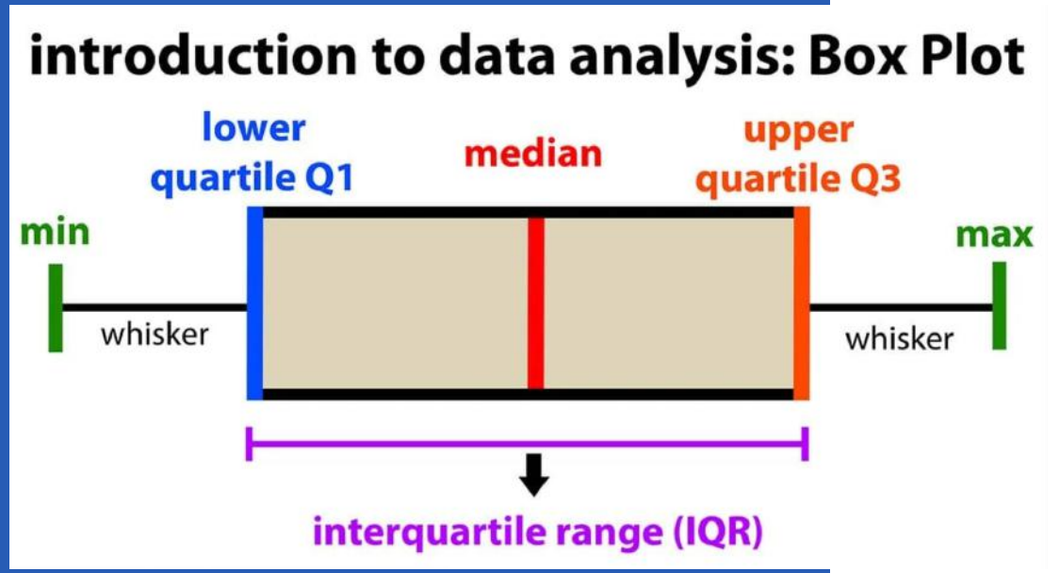
- Product sales (X), Revenue (Y), and Market Share (size) by Company
- Income per Capita (X), Life Expectancy (Y) and Population (size) by Country

Pro Tips:

- Use color as a fourth dimension to differentiate between categories
- Use cell formulas and form controls to create a dynamic, animated bubble chart







BOX & WHISKER CHARTS

A **box and whisker plot** (in the style of Tukey) or boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines (whiskers) extending from the boxes indicating variability outside the upper and lower quartiles.

When to use: Use a box and whisker plot when you want to see the distribution of data and identify outliers. For example, observing the distribution of incomes.

When not to use: Avoid using a box and whisker plot when the data is nominal or when the data ranges are not continuous.

BOX & WHISKER CHARTS

Commonly used for:

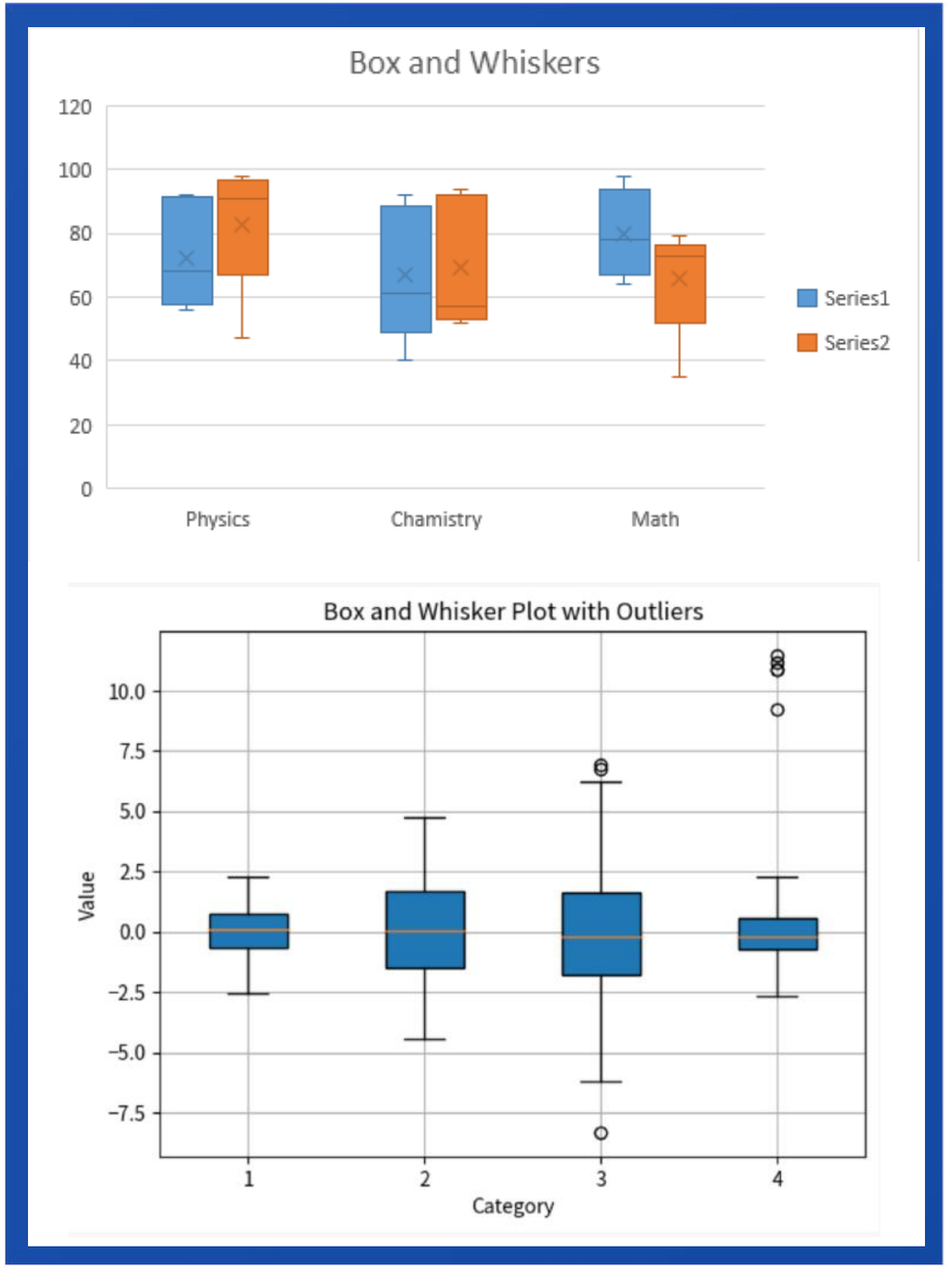
- Visualizing statistical characteristics across data series
- Identifying **outliers**

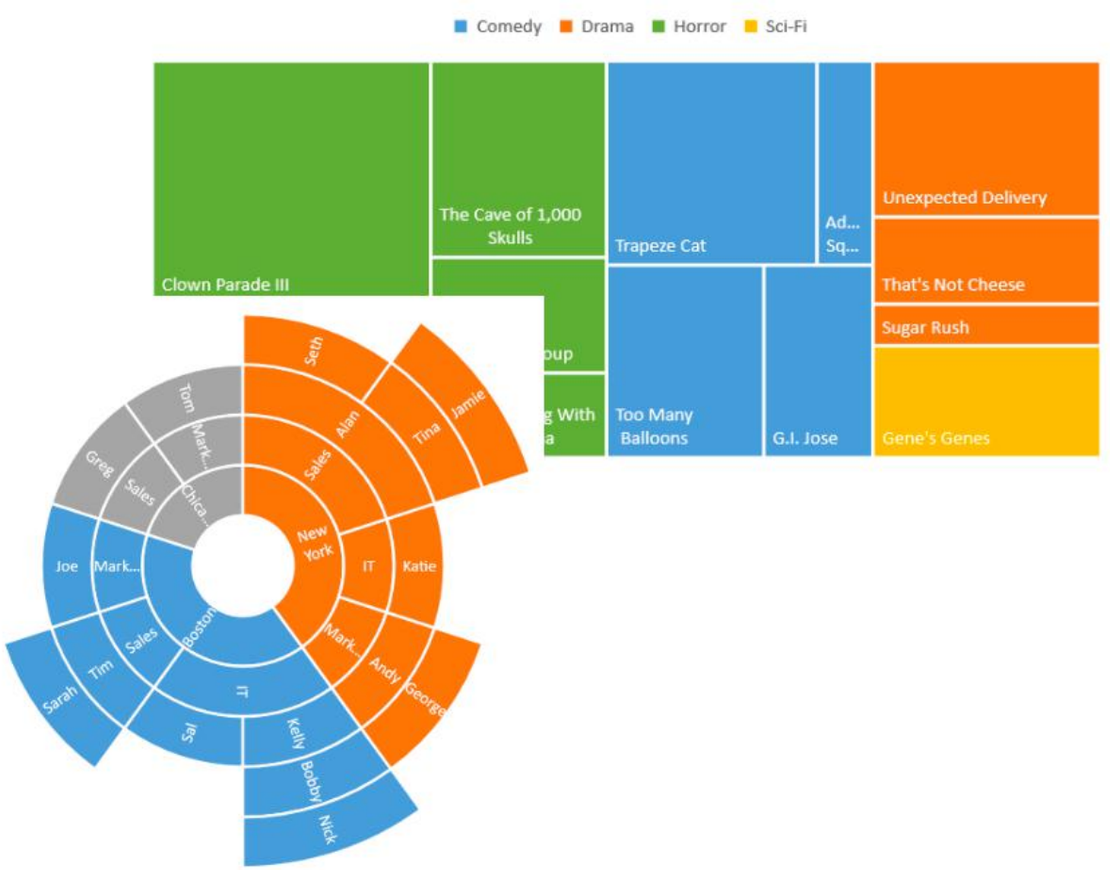
Examples:

- Comparing historical annual rainfall across cities
- Analyzing distributions of values and identifying outliers
- Comparing mean and median height/weight by country

Pro Tips:

- By default, quartiles are calculated by excluding the median; this calculation can be adjusted to include the median, but may significantly change the result (particularly for smaller data samples)





TREE MAPS & SUNBURST CHARTS

A **treemap** is a chart for displaying hierarchical data by using nested rectangles; the size of each rectangle is proportional to the quantity it represents. A **sunburst** chart is similar to a treemap, but it uses a radial layout. The hierarchy of the data is visualized through a series of concentric rings.

When to use: Use a treemap or sunburst chart when you want to display hierarchical data. For example, showing the distribution of sales by region and then by product.

When not to use: Avoid using a treemap or sunburst chart when the hierarchy of the data is not clear or when the data does not have a logical order.

TREE MAPS & SUNBURST CHARTS

Commonly used for:

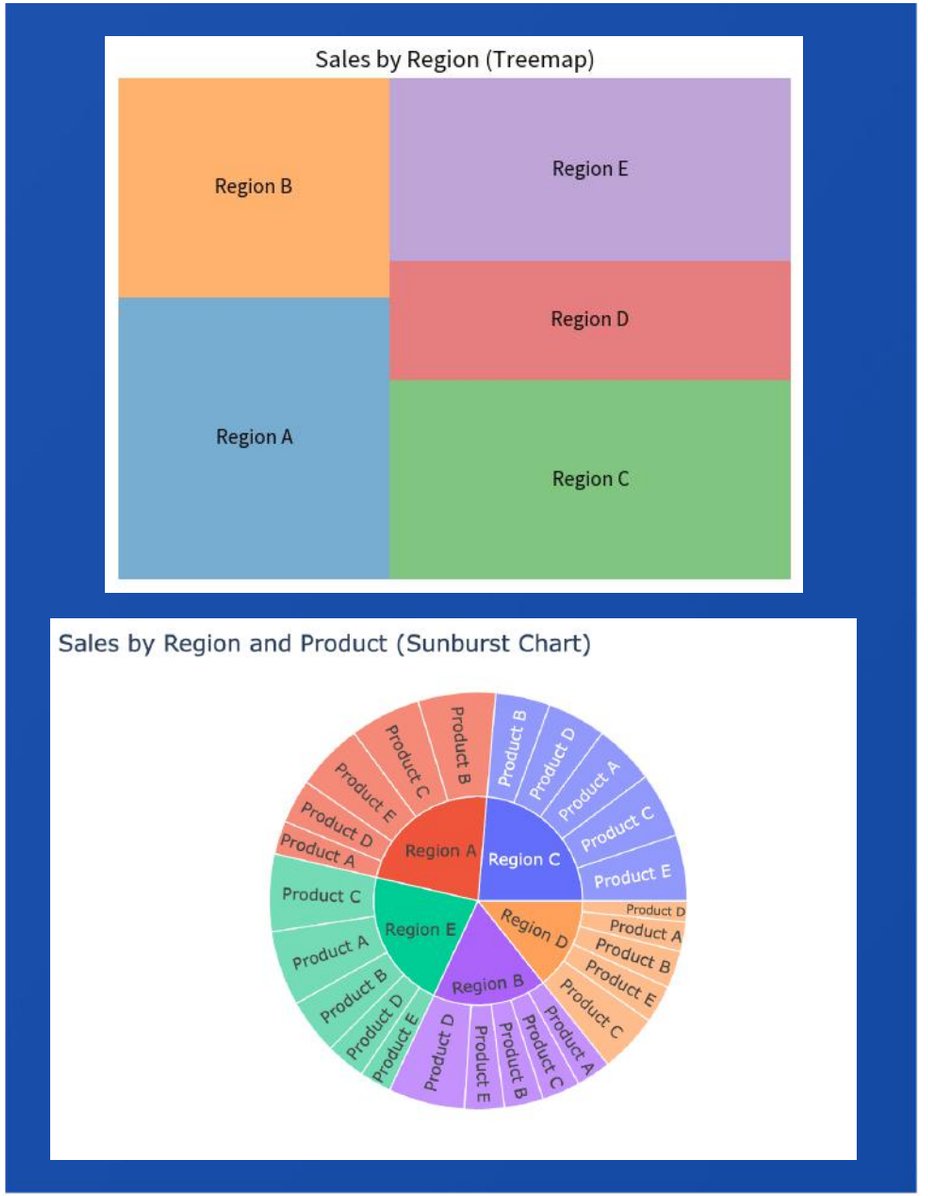
- Visualizing hierarchical data with natural groups/sub-groups

Examples:

- Revenue by Book Title, Sub-Genre, and Genre
- Number of Employees by Department and Office
- Population by City, State, and Region

Pro Tips:

- Use Tree Maps when you are only visualizing 1 or 2 hierarchical levels (i.e. topic & sub-topic) or when relative sizes are important, and Sunburst charts to visualize the depth of multiple hierarchical levels
- Make sure your raw source data is grouped and sorted before creating hierarchical charts



WATERFALL CHARTS

A **waterfall** chart is a form of data visualization that helps in understanding the cumulative effect of sequentially introduced positive or negative values. These intermediate values can either be time-based or category-based.

When to use: Use a waterfall chart when you want to visualize the cumulative effect of sequential data. For example, showing the progression of a financial metric through different contributions.

When not to use: Avoid using a waterfall chart when the data does not have a logical order or when there are too many categories, which can make the chart hard to read.

WATERFALL CHARTS

Commonly used for:

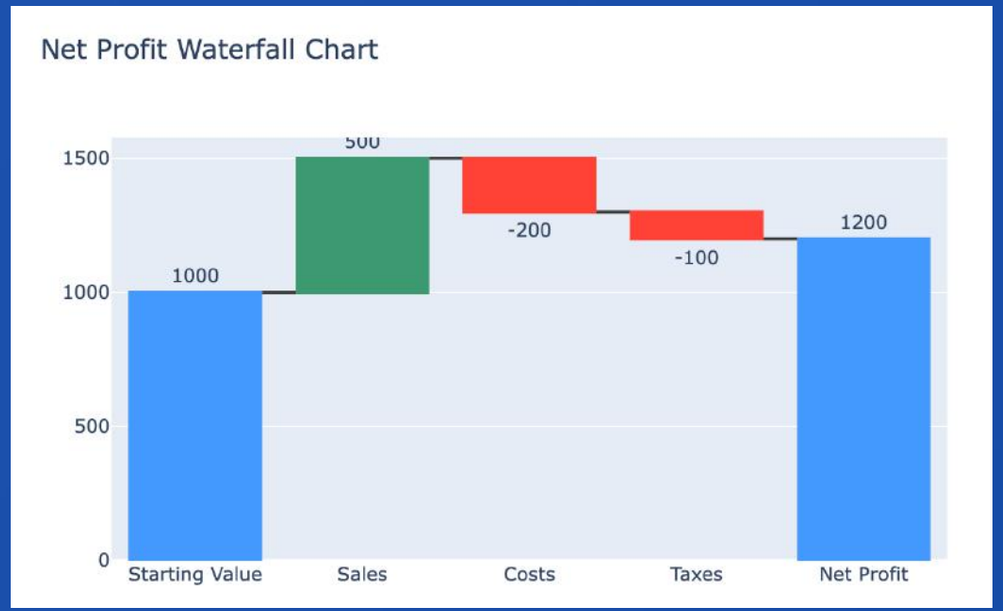
- Showing the net value after a series of positive and negative contributions

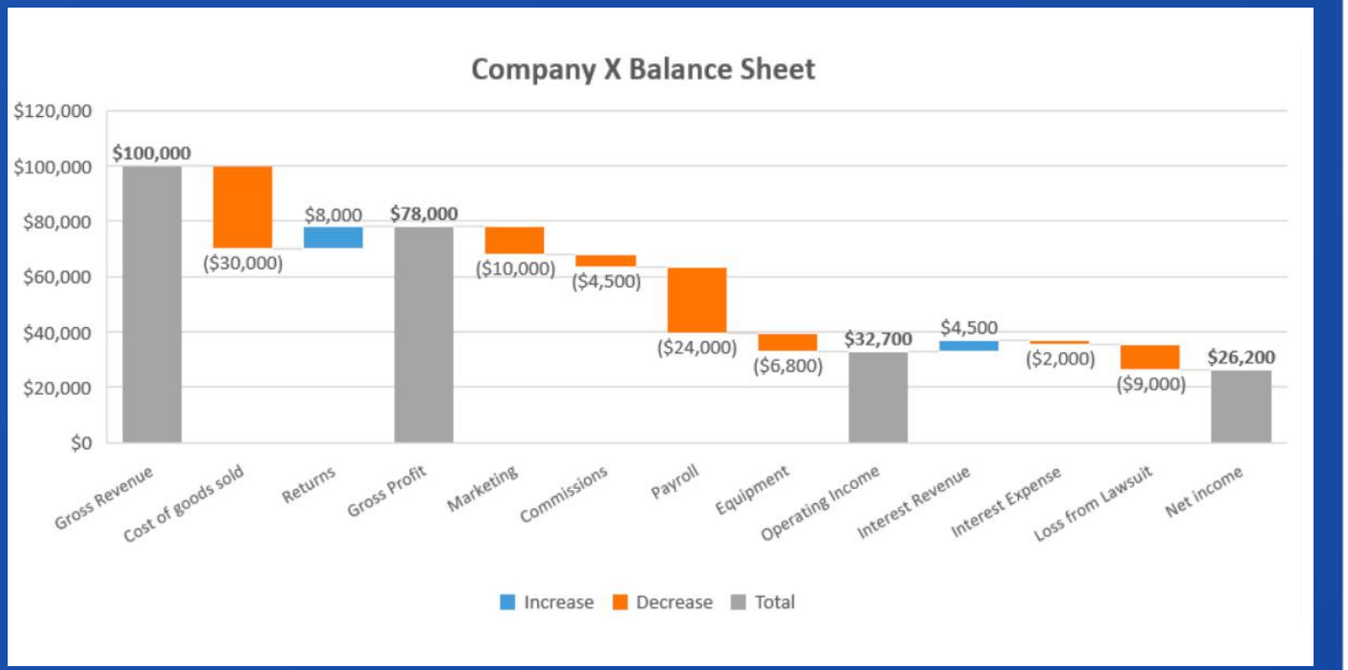
Examples:

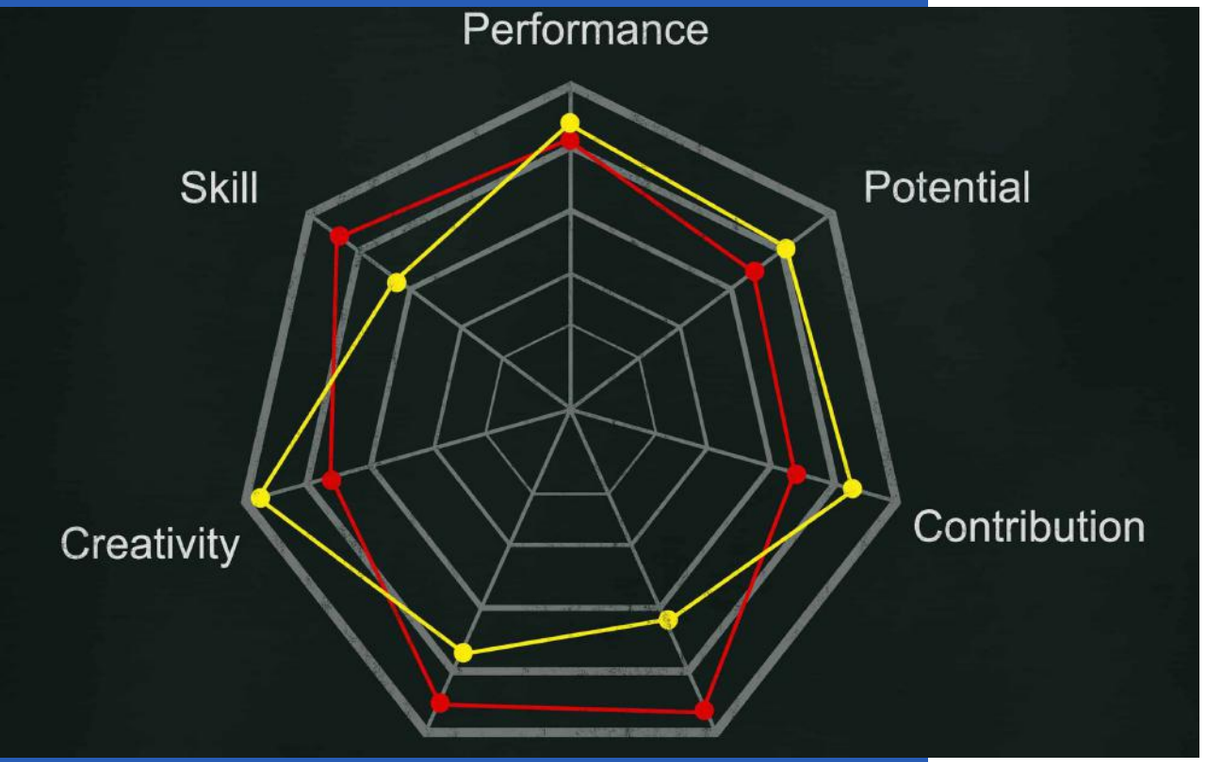
- Corporate balance sheet analysis
- Personal income and spending

Pro Tips:

- Use sub-totals to create “checkpoints” and split up certain types of gains/losses (i.e. Gross Revenue - Cost of Goods Sold = Gross Profit, Gross Profit - Operating Expenses = Operating Income, etc.)







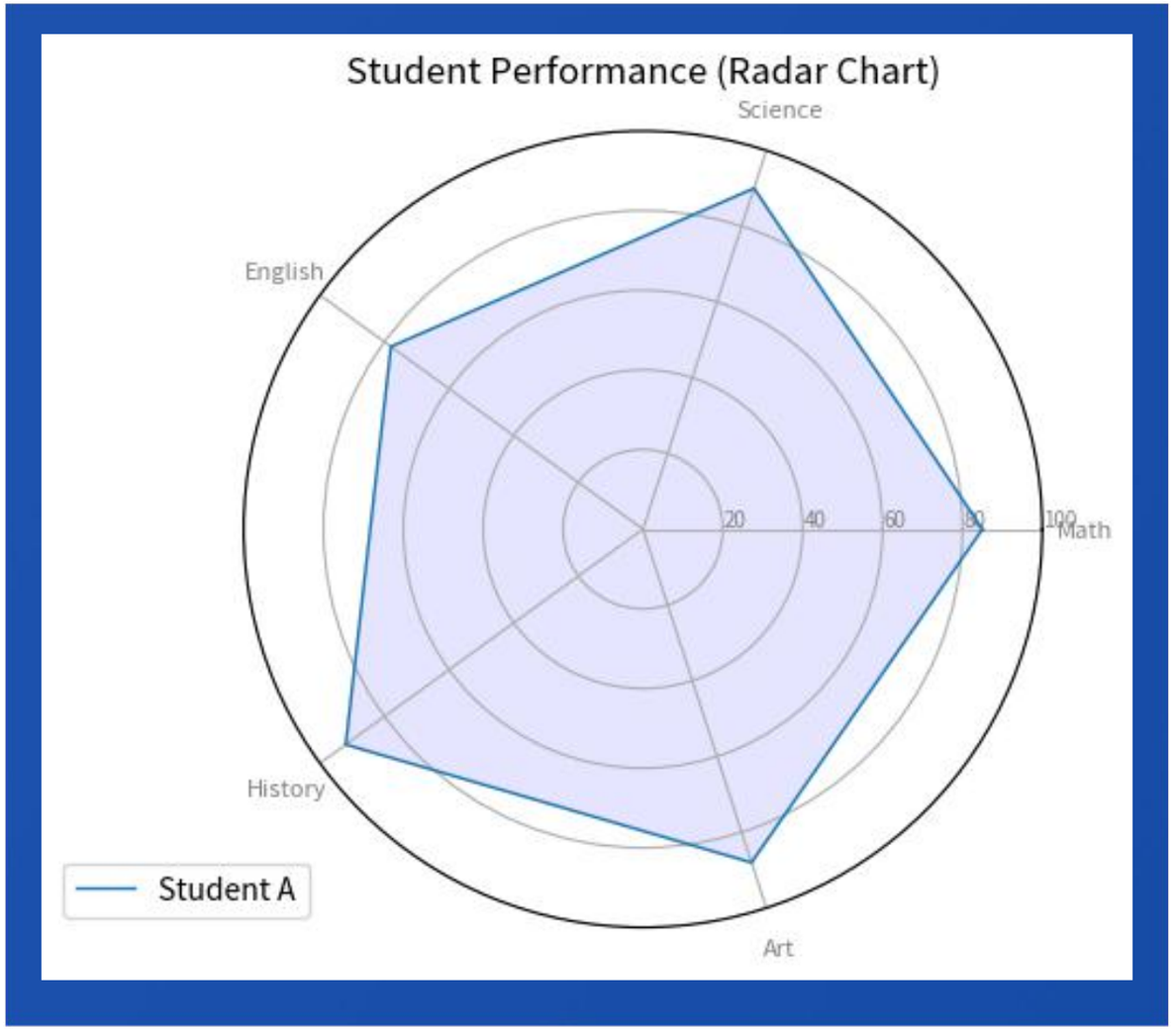
RADAR CHARTS

A **radar** chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point.

When to use: Use a radar chart when you want to compare several quantitative variables. For example, comparing the performance of a student in different subjects.

When not to use: Avoid using a radar chart when the data does not have a logical order or when there are too many categories, which can make the chart hard to read.

RADAR CHARTS



Commonly used for:

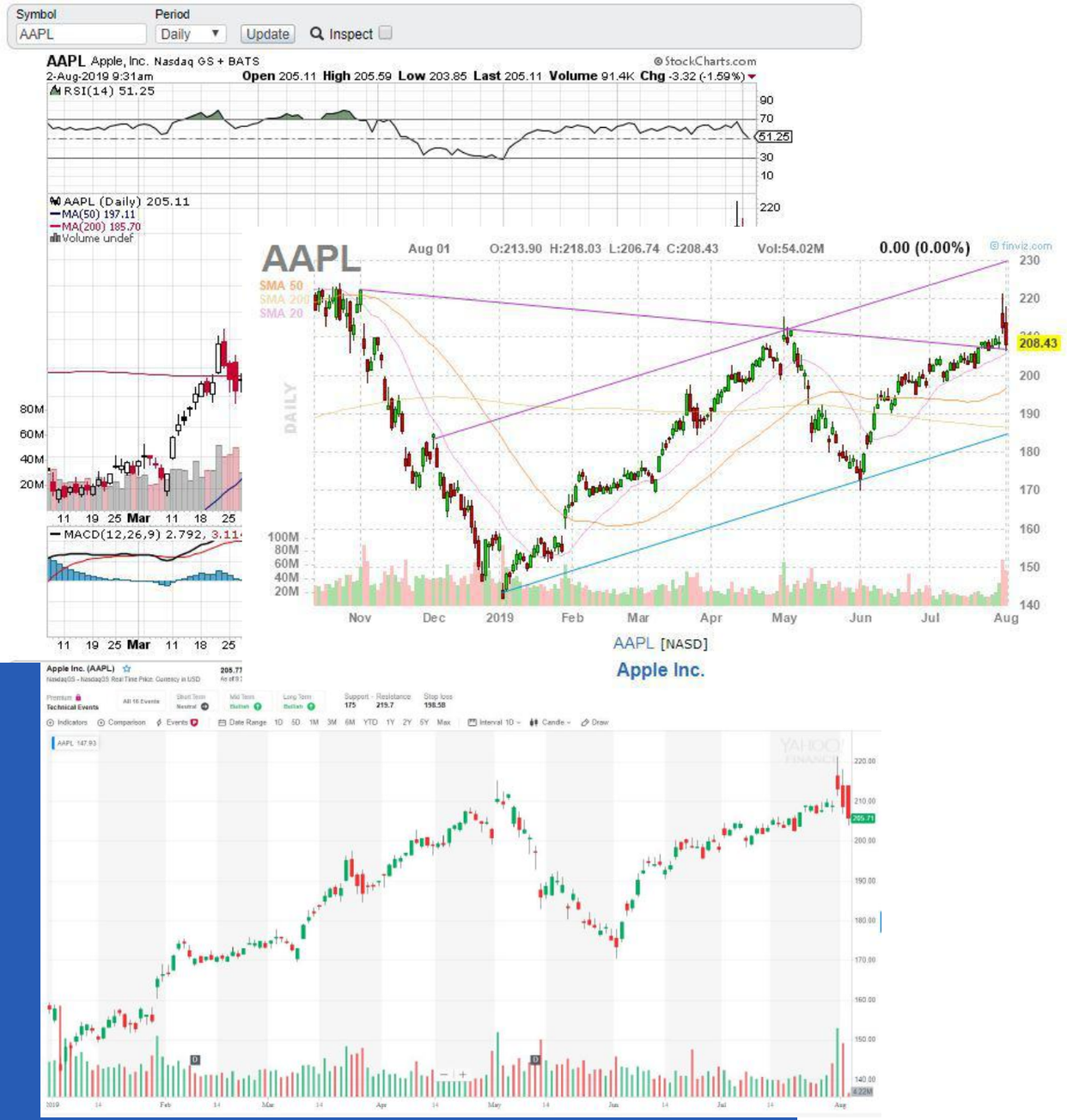
- Plotting three or more quantitative variables on a two-dimensional chart, relative to a central point

Examples:

- Comparing test scores across multiple subjects
- Sales of different types of vegetables, by month
- Visualizing personality test results across subjects

Pro Tips:

- Normalize each metric to the same scale (i.e. 0-1, 1-10, 1-100) to improve readability and create more intuitive comparisons across data series
- Limit the number of categories or data series to minimize noise and maximize impact



STOCK CHARTS

A **stock** chart or stock graph is a chart that depicts the past pricing of a specific stock. The stock chart may also depict the volume of shares traded.

When to use: Use a stock chart when you want to visualize the price movement of a stock over time. For example, showing the historical price of a company's stock.

When not to use: Avoid using a stock chart when the data does not have a logical order or when the data is not time-series data.

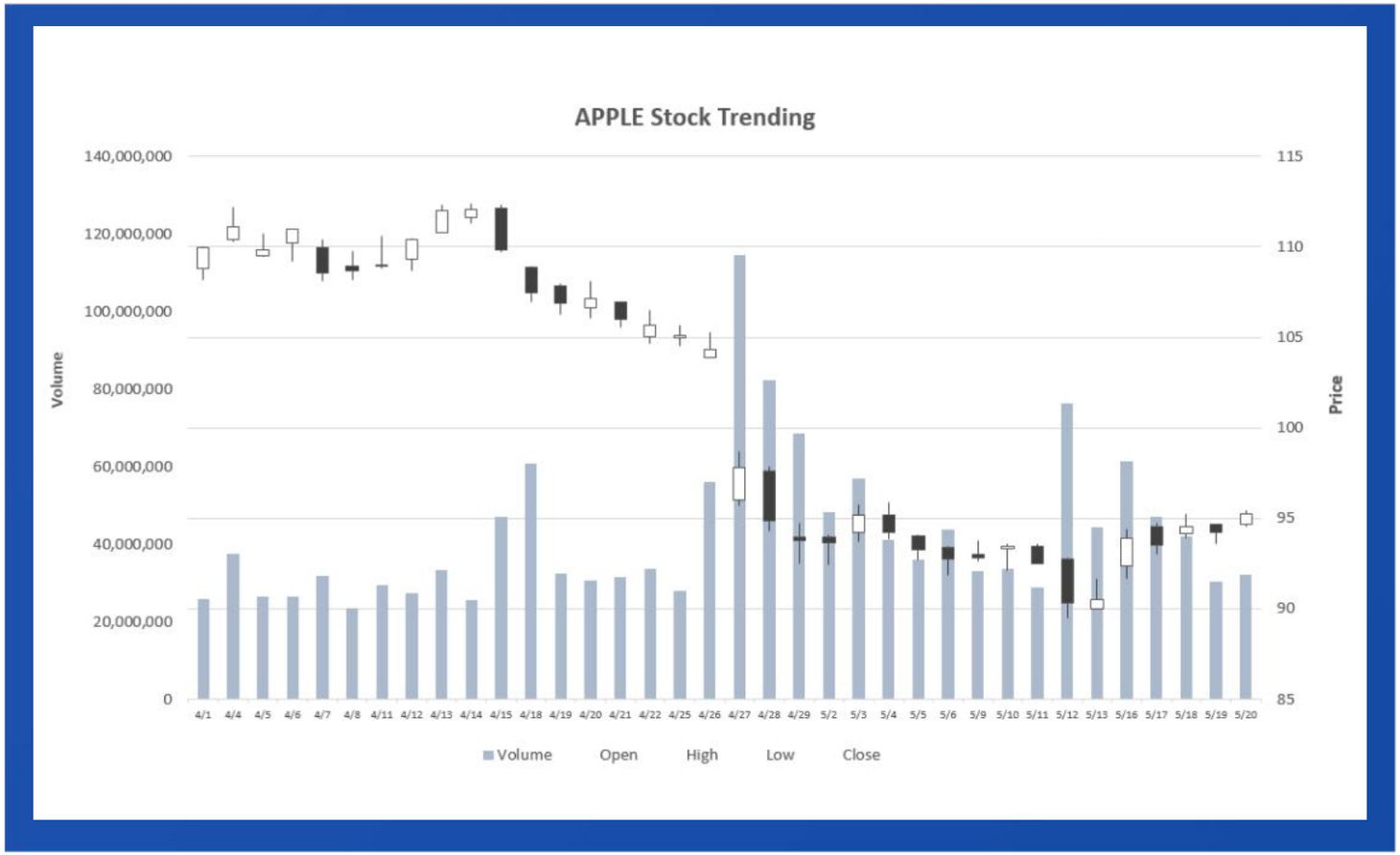
STOCK CHARTS

Commonly used for:

- Visualizing stock market data, including volume, high, low, open, and closing prices

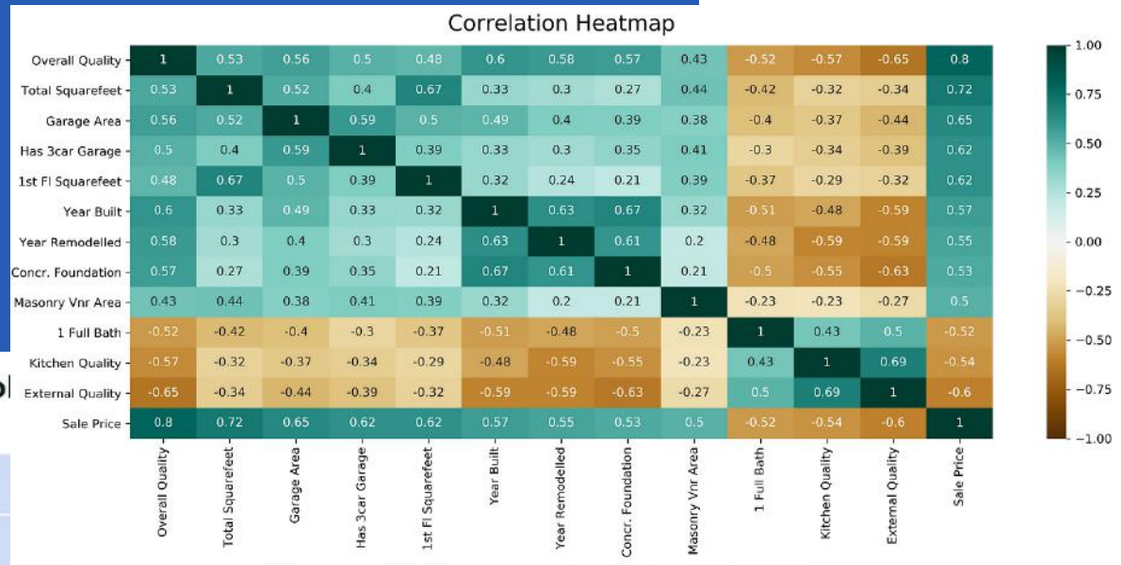
Examples:

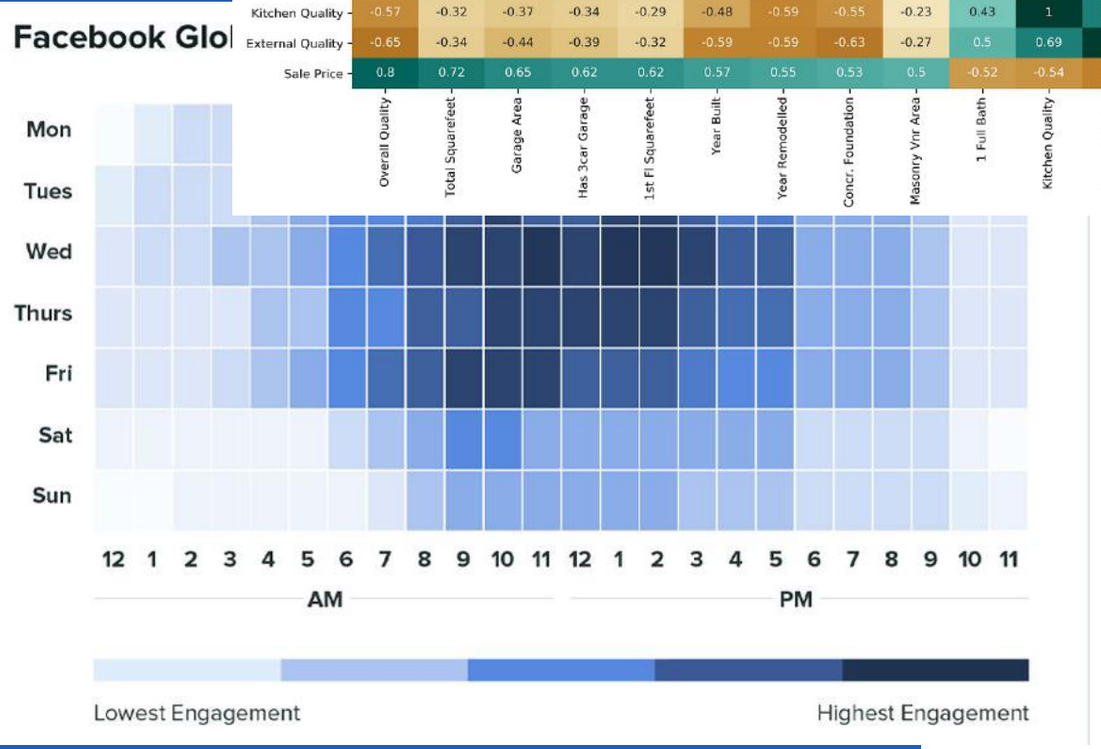
- Facebook's daily stock performance in 2022
- High, low, and closing prices for Google in Q1
- Relative performance across multiple stocks



Pro Tips:

- Manually set axis minimum/maximum values to enhance readability
- Switch from a date to a text axis to eliminate gaps when markets are closed





HEATMAPS

A **heatmap** is a graphical representation of data where individual values contained in a matrix are represented as colors.

When to use: Use a heatmap when you want to understand the density of points in a dataset and identify patterns or correlations. For example, showing the correlation between different features in a dataset.

When not to use: Avoid using a heatmap when the data is not numerical or when the data does not have a logical order.

HEATMAPS

Commonly used for:

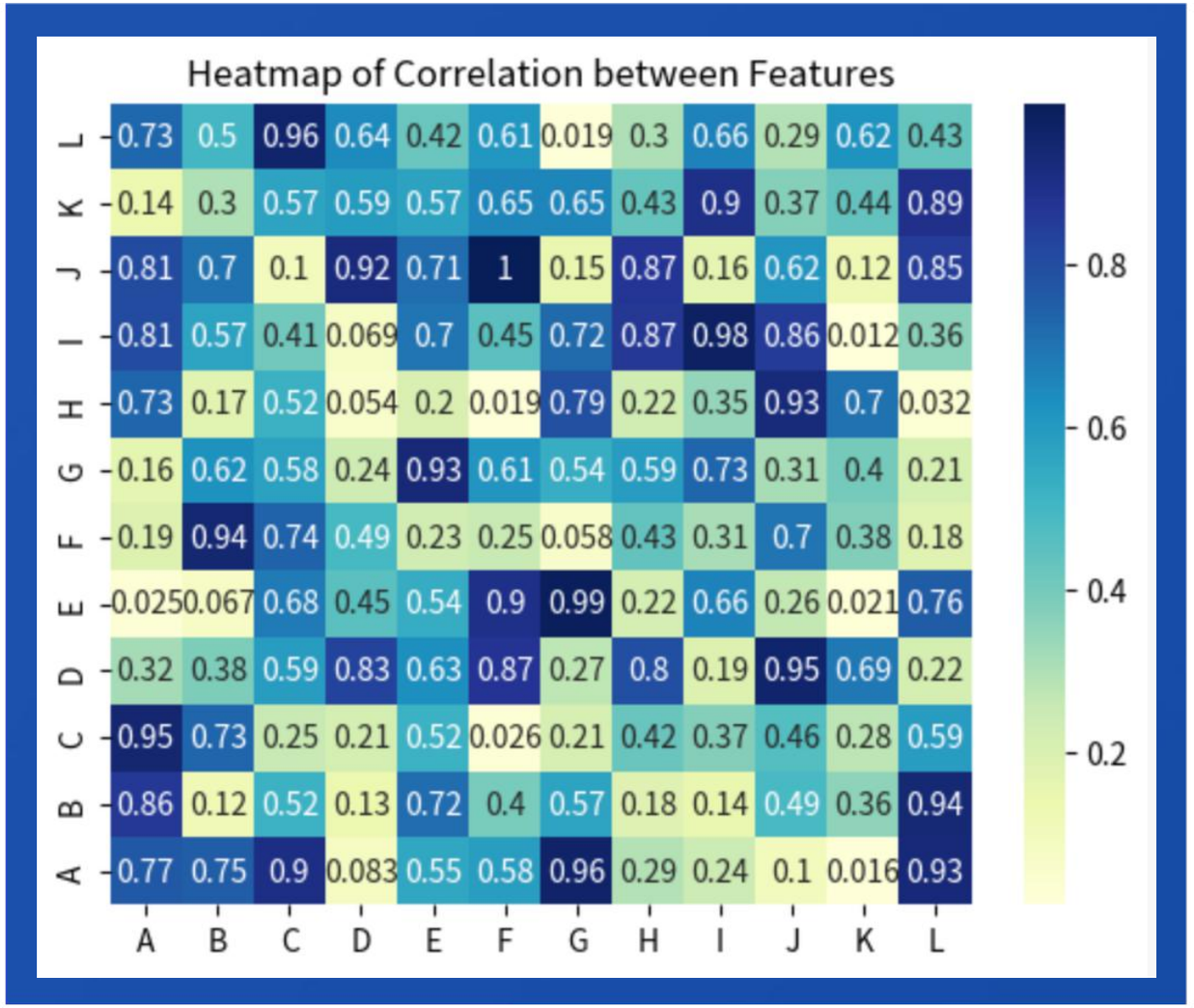
- Visualizing trends or relationships using color scales

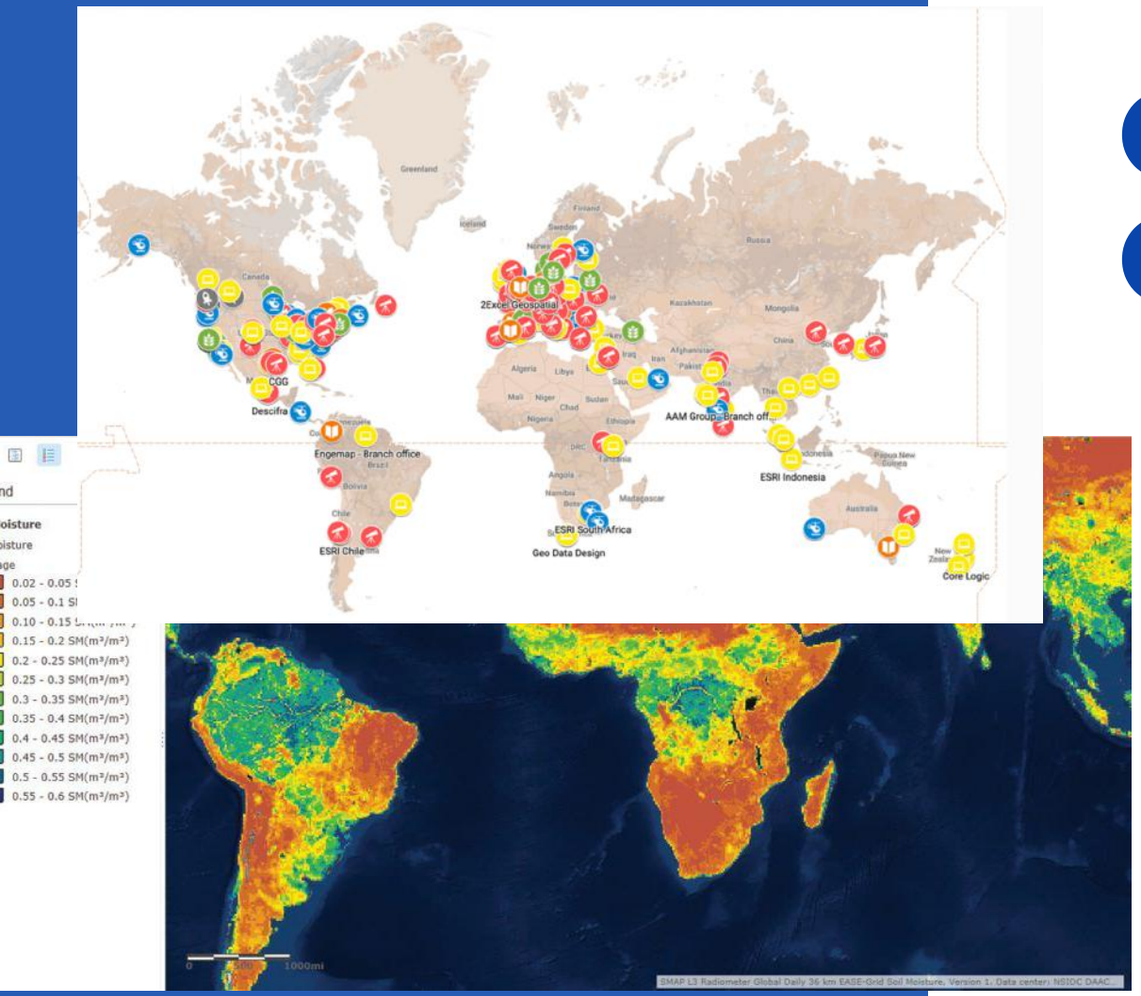
Examples:

- Accident rates by time of day and day of week
- Average temperature by city, by month
- Average sentiment by hashtag

Pro Tips:

- Use intuitive color scales (i.e. red to green) and apply custom formatting to hide cell values (;;;)
- Use data validation and cell formulas to create dynamic heat maps based on user-entered values





GEOSPATIAL/ CHOROPLETH MAP

A **choropleth** map is a type of thematic map in which areas are shaded or patterned in relation to a statistical variable that represents an aggregate summary of a geographic characteristic within each area, such as population density or per-capita income.

When to use: Use a choropleth map when you want to visualize how a measurement varies across a geographic area. For example, showing the population density across different regions.

When not to use: Avoid using a choropleth map when the data is not geographical or when the data does not have a logical order.

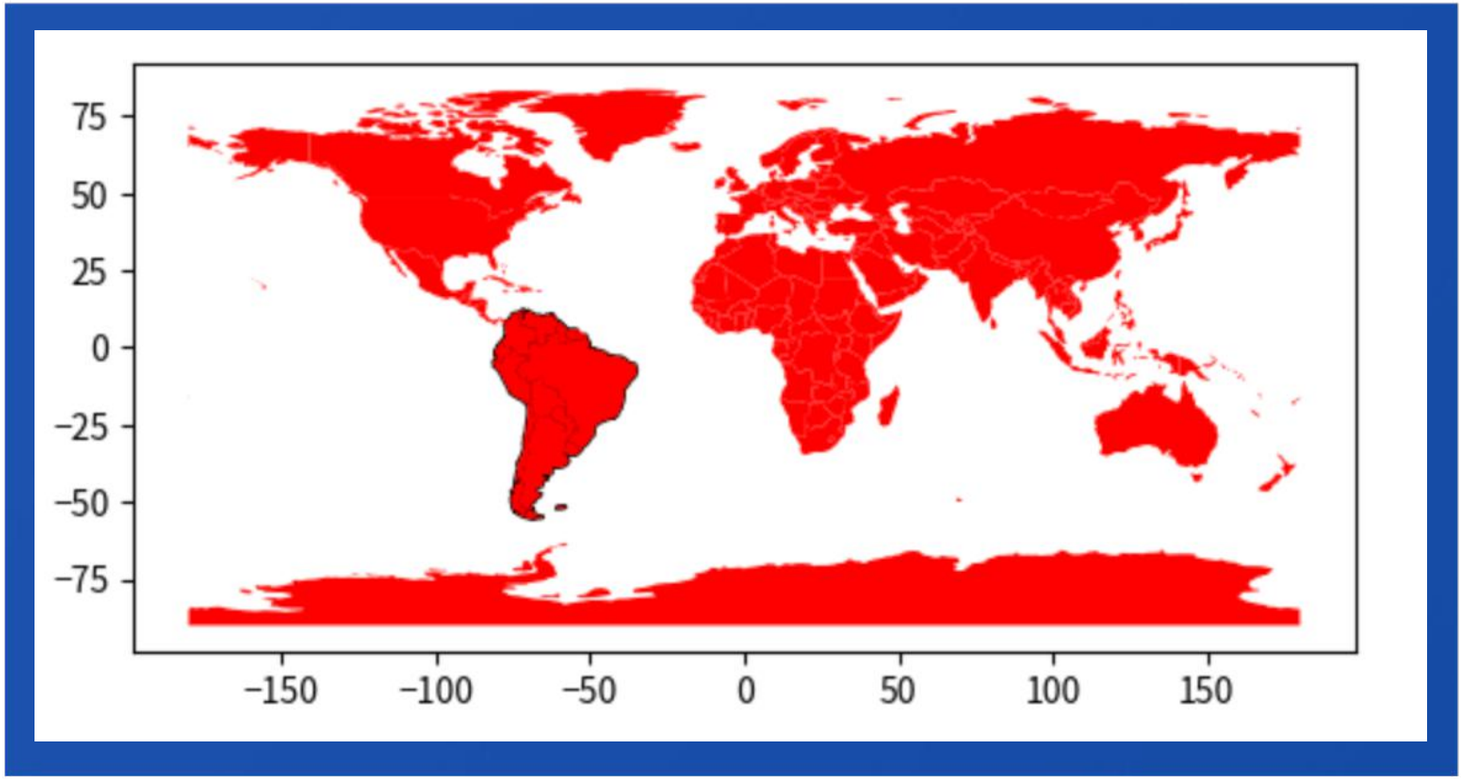
GEOSPATIAL/ CHOROPLETH MAP

Commonly used for:

- Visualizing location-based data

Examples:

- Frequency of accidents by street address
- Unemployment rate by country
- Average rainfall by state



Pro Tips:

- Use Excel's Power Map plug-in to create geo-spatial visualizations and animate changes over time
- Utilize attributes like color and size to visualize multiple attributes at once

DATA VIZ RESOURCES

- Data Visualisation Catalogue: The Data Visualisation Catalogue is an on-going project developed by Severino Ribecca. It is a great reference for information on different chart types and methodology.
  http://datavizcatalogue.com/index.html
- Data Journalism Handbook — The Data Journalism Handbook offers detailed information for journalists who want to use data to enhance their work and tell better news stories.
  http://datajournalismhandbook.org/1.0/en/index.html
- Infogram — Infogram offers a wide variety of helpful data visualization resources.
  https://infogr.am/
- Visualoop — Visualoop is a digital environment dedicated to all things data visualization. They have an incredible selection of infographics for you to enjoy.
  http://visualoop.com/
- Visualising Data — Andy Kirk's Visualising Data is an award-winning site providing readers with a rich variety of content that charts the development of the data visualization field.
  http://www.visualisingdata.com/