Choosing the right chart or graph depends on the type of data you have and the story you want to tell. Here's a guide to help you select the appropriate chart or graph based on your needs:

1. Bar Chart

- **When to Use**: To compare different categories or show frequency distributions.
- **Best For**: Categorical data (e.g., product sales, demographic data).
- **Example**: Comparing the sales of different product categories.

Chart 5.2.1
Number of police officers in Crimeville, 2011 to 2019

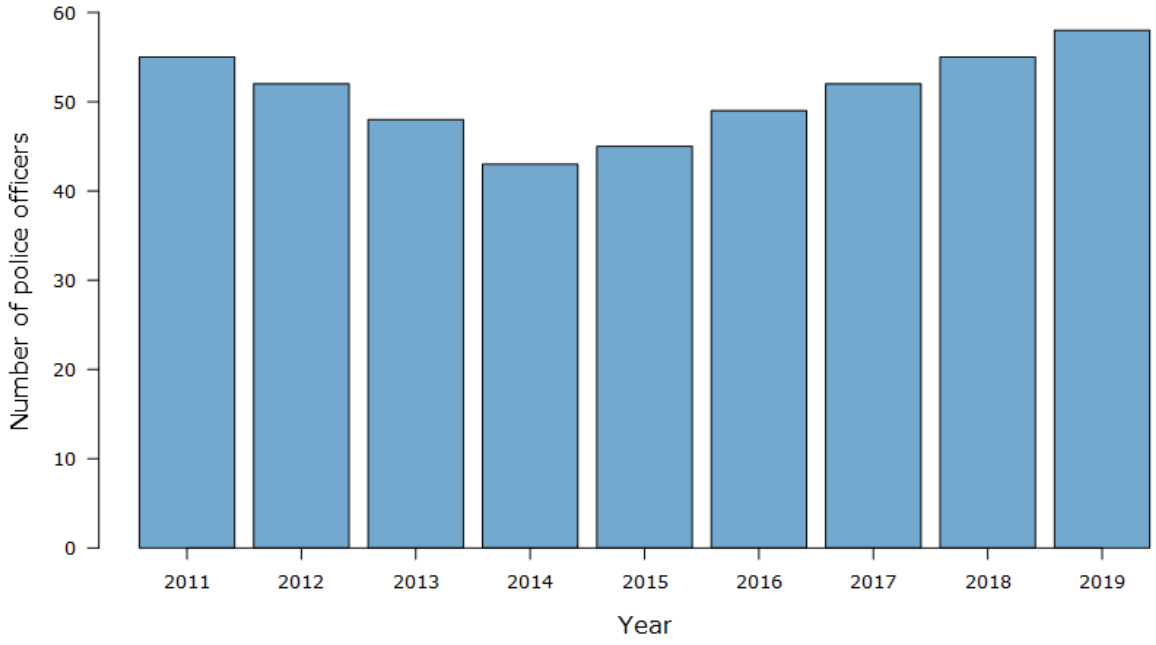


Chart 5.3.1
Number of students who like chocolate chip cookies best

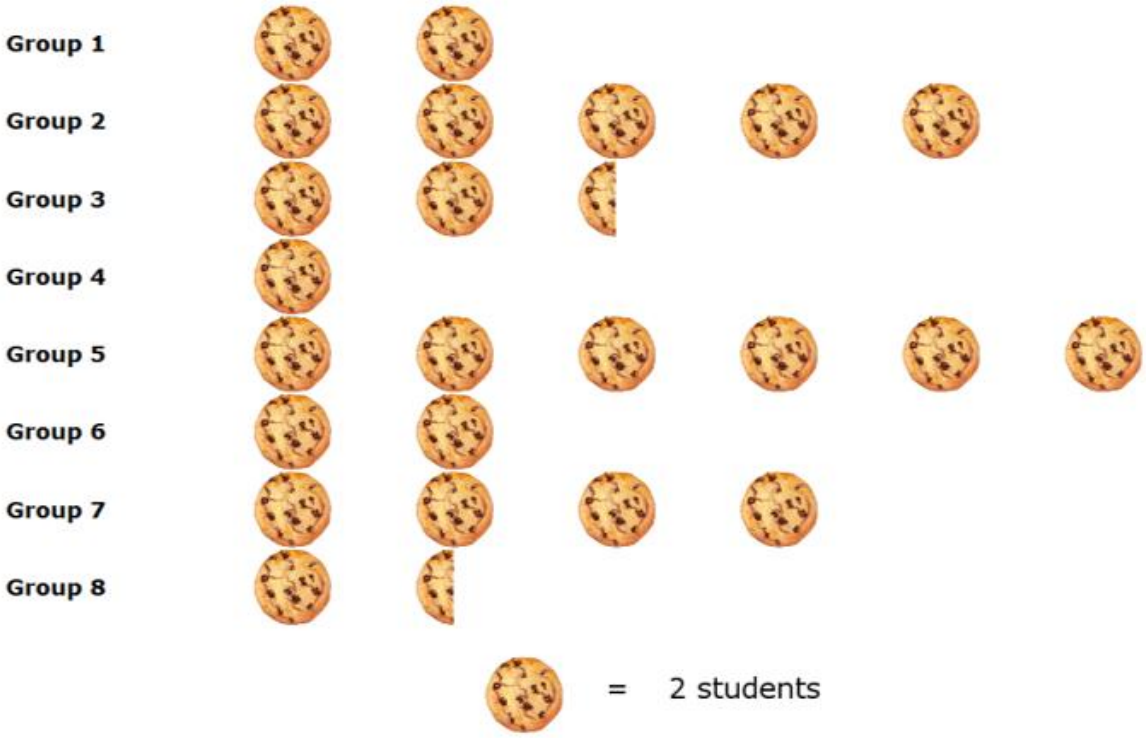
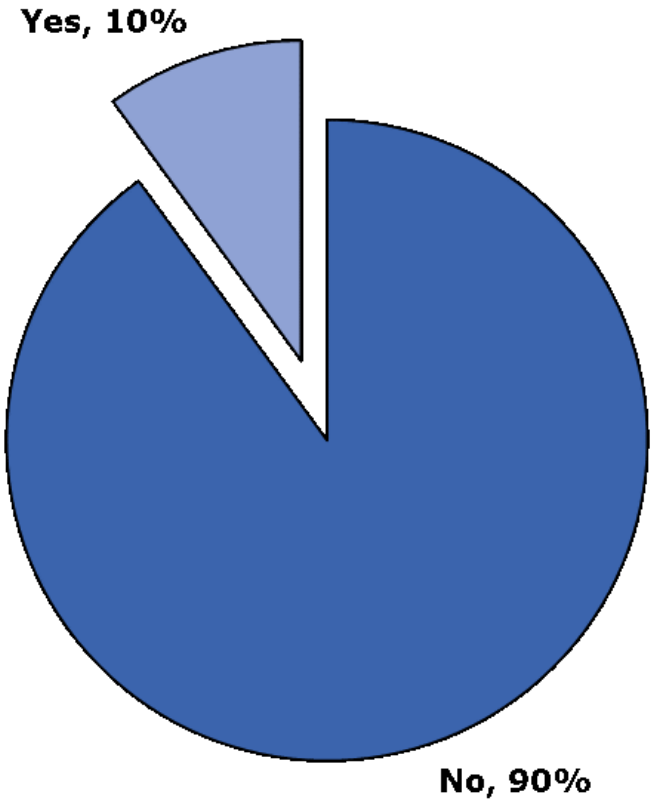


Chart 5.3.2
Purchasing power of the Canadian dollar, 2000 to 2020

2. Pie Chart

- **When to Use**: To show proportions or percentages of a whole.
- **Best For**: Categorical data with few categories, ideally under six.
- **Example**: Displaying the market share of different companies.

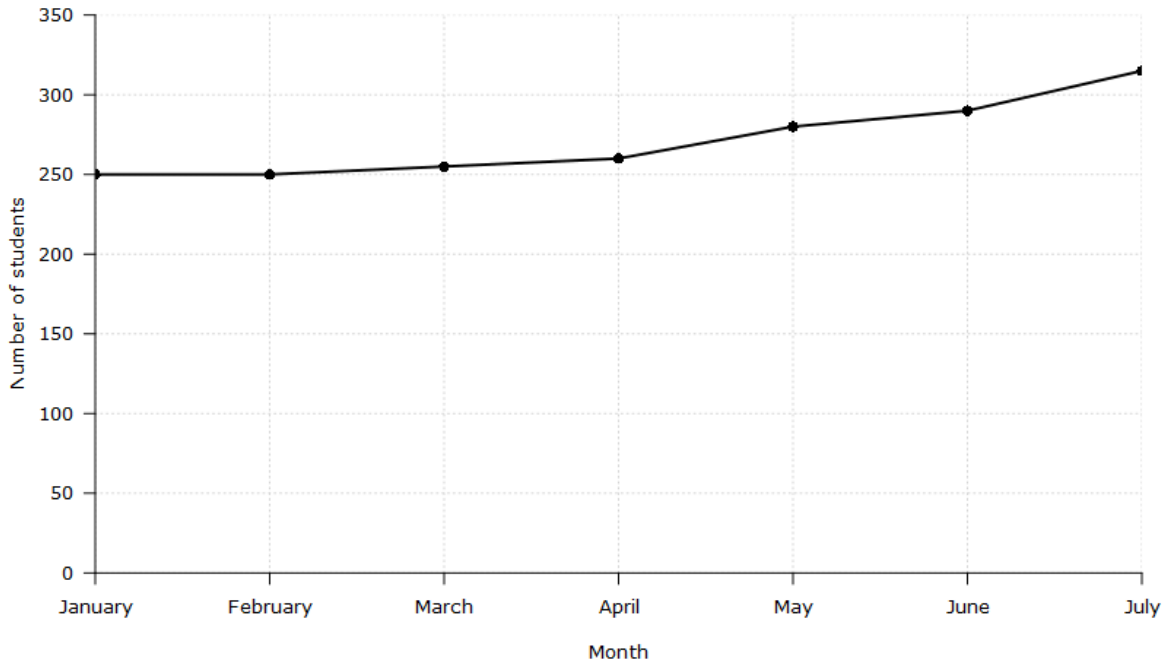
Chart 5.4.1
Student and faculty response to the poll "Should Avenue High School adopt student uniform?"



3. Line Chart

- **When to Use**: To show trends over time.
- **Best For**: Continuous data, especially time-series data (e.g., stock prices, sales over time).
- **Example**: Showing a company's revenue growth over several years.

Chart 5.5.1
Labour force participation in Andrew's high school



4. Scatter Plot

- **When to Use**: To show relationships or correlations between two continuous variables.
- **Best For**: Comparing pairs of numerical data (e.g., height vs. weight, temperature vs. sales).
- **Example**: Examining the relationship between advertising budget and sales revenue.

https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/scatter-nuages/5214827-eng.htm

5. Histogram

- **When to Use**: To show the distribution of a continuous dataset.
- **Best For**: Grouping continuous data into ranges (e.g., ages, salaries).
- **Example**: Displaying the distribution of test scores among students.

Chart 5.7.1
Distribution of salaries of the employees of ABC Corporation

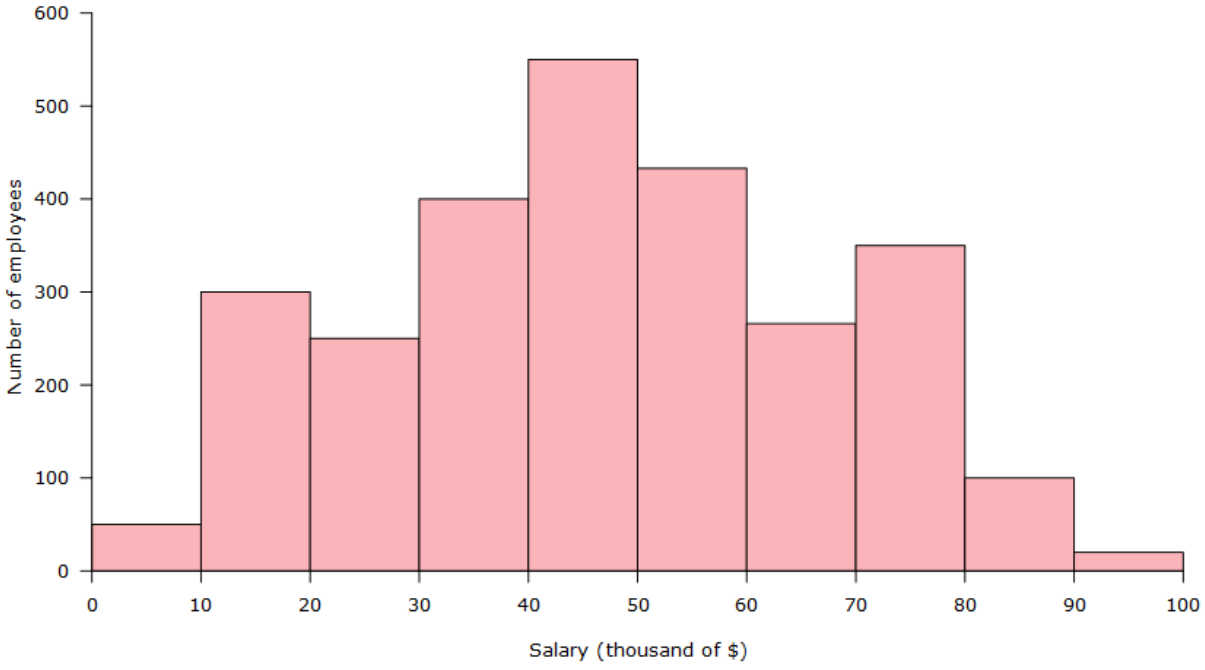
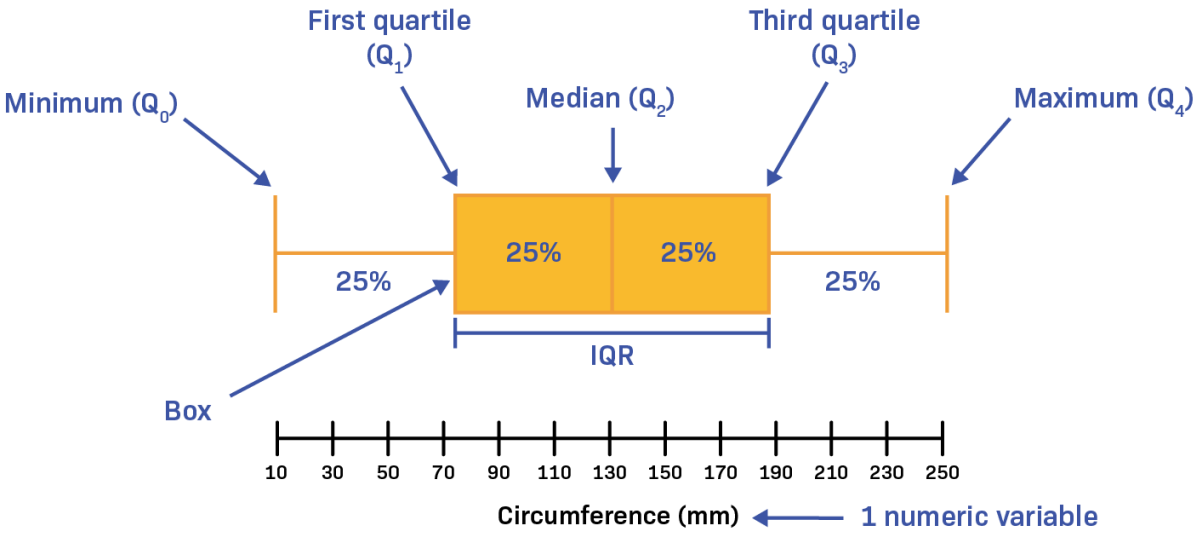


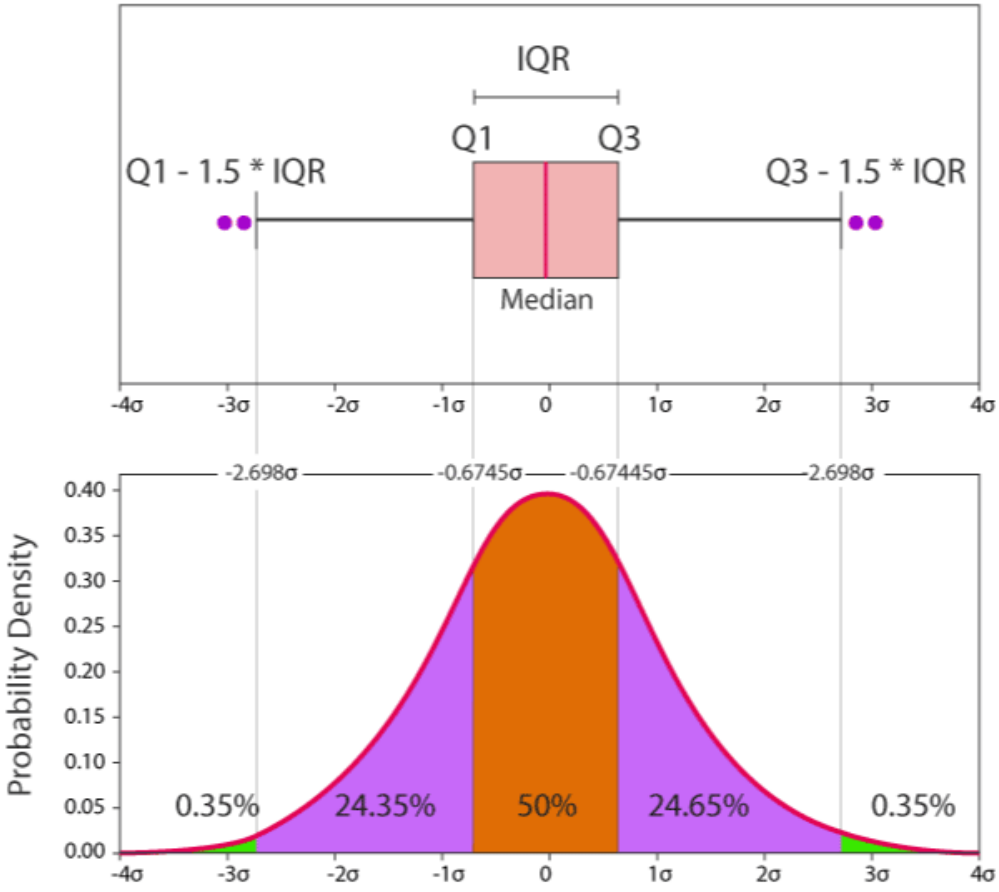
Table 5.7.1
Differences between bar chart and histogram

Comparison terms	Bar chart	Histogram
Usage	To compare different categories of data.	To display the distribution of a variable.
Type of variable	Categorical variables	Numeric variables
Rendering	Each data point is rendered as a separate bar.	The data points are grouped and rendered based on the bin value. The entire range of data values is divided into a series of non-overlapping intervals.
Space between bars	Can have space.	No space.
Reordering bars	Can be reordered.	Cannot be reordered.

6. Box Plot (Box-and-Whisker Plot)

- **When to Use**: To show the spread and skewness of data, highlight outliers.
- **Best For**: Continuous data, comparing distributions between groups.
- **Example**: Comparing the distribution of salaries across different departments in a company.





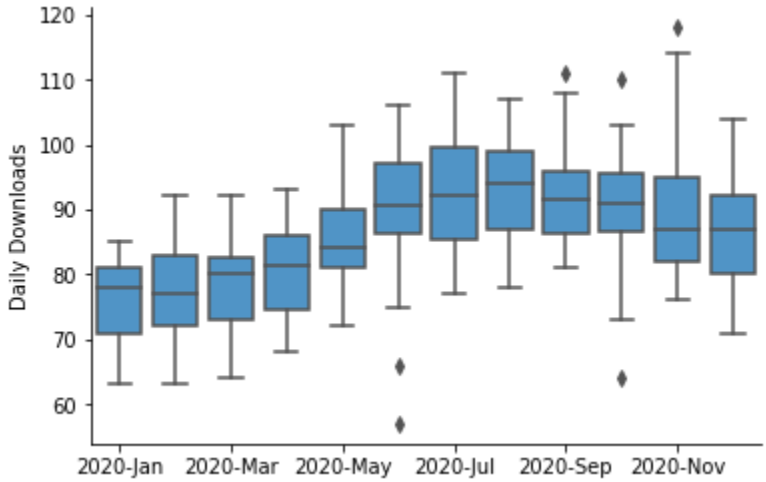
Boxplot on a normal distribution

More on Box Plot

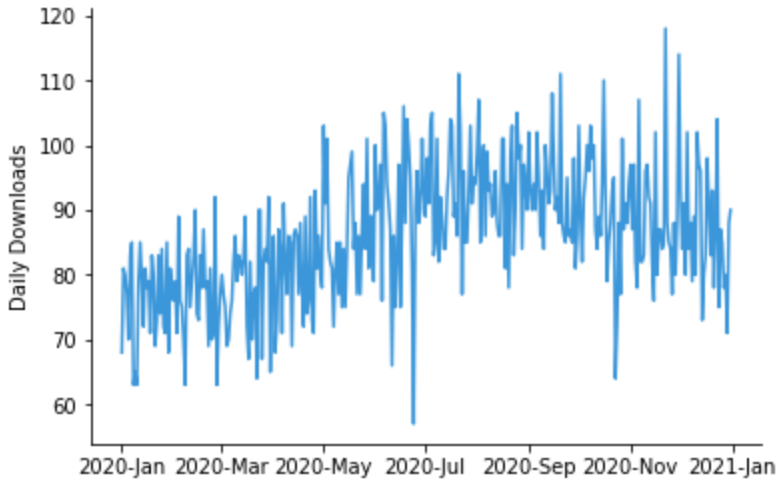
What is a box plot?

Source: Atlassian

A box plot (aka box and whisker plot) uses boxes and lines to depict the distributions of one or more groups of numeric data. Box limits indicate the range of the central 50% of the data, with a central line marking the median value. Lines extend from each box to capture the range of the remaining data, with dots placed past the line edges to indicate outliers.



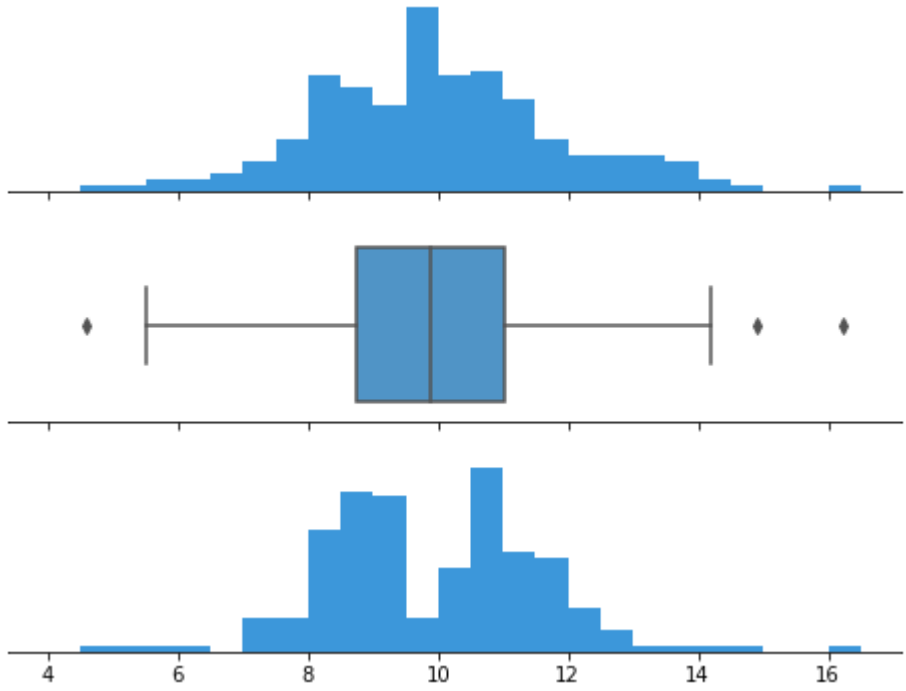
The example box plot above shows daily downloads for a fictional digital app, grouped together by month. From this plot, we can see that downloads increased gradually from about 75 per day in January to about 95 per day in August. There also appears to be a slight decrease in median downloads in November and December. Points show days with outlier download counts: there were two days in June and one day in October with low downloads compared to other days in the month. The box and whiskers plot provides a cleaner representation of the general trend of the data, compared to the equivalent line chart.



When you should use a box plot

Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups. They are built to provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers. It is easy to see where the main bulk of the data is, and make that comparison between different groups.

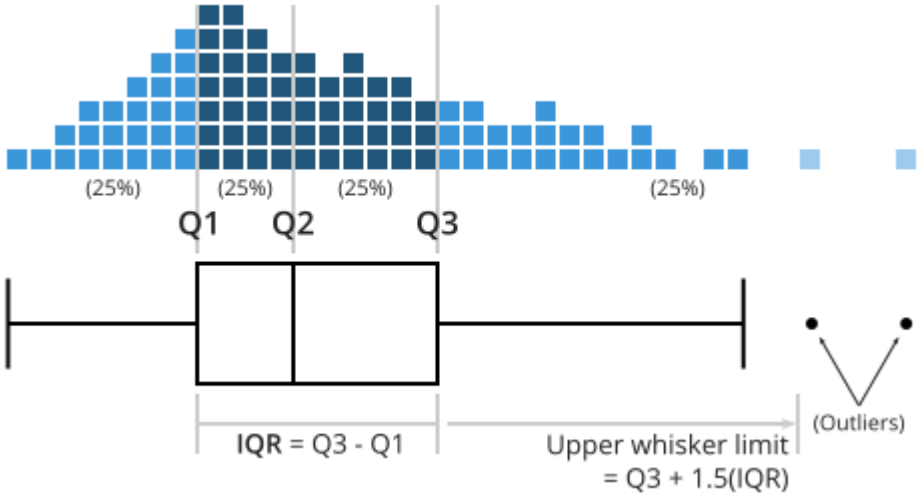
On the downside, a box plot's simplicity also sets limitations on the density of data that it can show. With a box plot, we miss out on the ability to observe the detailed shape of distribution, such as if there are oddities in a distribution's modality (number of 'humps' or peaks) and skew.



The datasets behind both histograms generate the same box plot in the center panel.

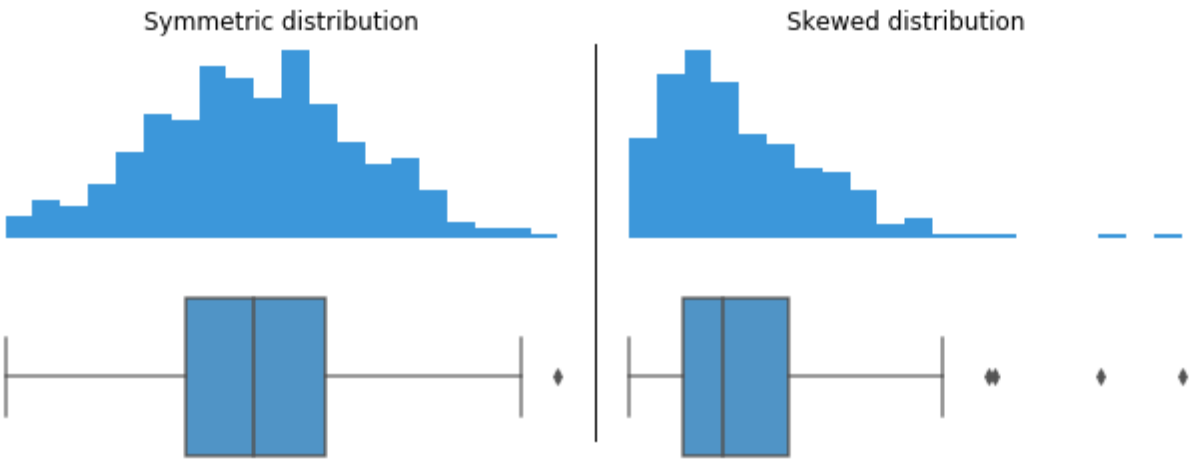
Interpreting a box and whiskers

Construction of a box plot is based around a dataset's quartiles, or the values that divide the dataset into equal fourths. The first quartile (Q1) is greater than 25% of the data and less than the other 75%. The second quartile (Q2) sits in the middle, dividing the data in half. Q2 is also known as the median. The third quartile (Q3) is larger than 75% of the data, and smaller than the remaining 25%. In a box and whiskers plot, the ends of the box and its center line mark the locations of these three quartiles.



The distance between Q3 and Q1 is known as the interquartile range (IQR) and plays a major part in how long the whiskers extending from the box are. Each whisker extends to the furthest data point in each wing that is within 1.5 times the IQR. Any data point further than that distance is considered an outlier, and is marked with a dot. There are other ways of defining the whisker lengths, which are discussed below.

When a data distribution is symmetric, you can expect the median to be in the exact center of the box: the distance between Q1 and Q2 should be the same as between Q2 and Q3. Outliers should be evenly present on either side of the box. If a distribution is skewed, then the median will not be in the middle of the box, and instead off to the side. You may also find an imbalance in the whisker lengths, where one side is short with no outliers, and the other has a long tail with many more outliers.



Example of data structure

DATE	...	MONTH	DOWNLOADS
2020-01-30	...	2020-01	81
2020-01-31	...	2020-01	78
2020-02-01	...	2020-02	76
2020-02-02	...	2020-02	79
...	...	...	...

Visualization tools are usually capable of generating box plots from a column of raw, unaggregated data as an input; statistics for the box ends, whiskers, and outliers are automatically computed as part of the chart-creation process. When a box plot needs to be drawn for multiple groups, groups are usually indicated by a second column, such as in the table above.

Best practices for using a box plot

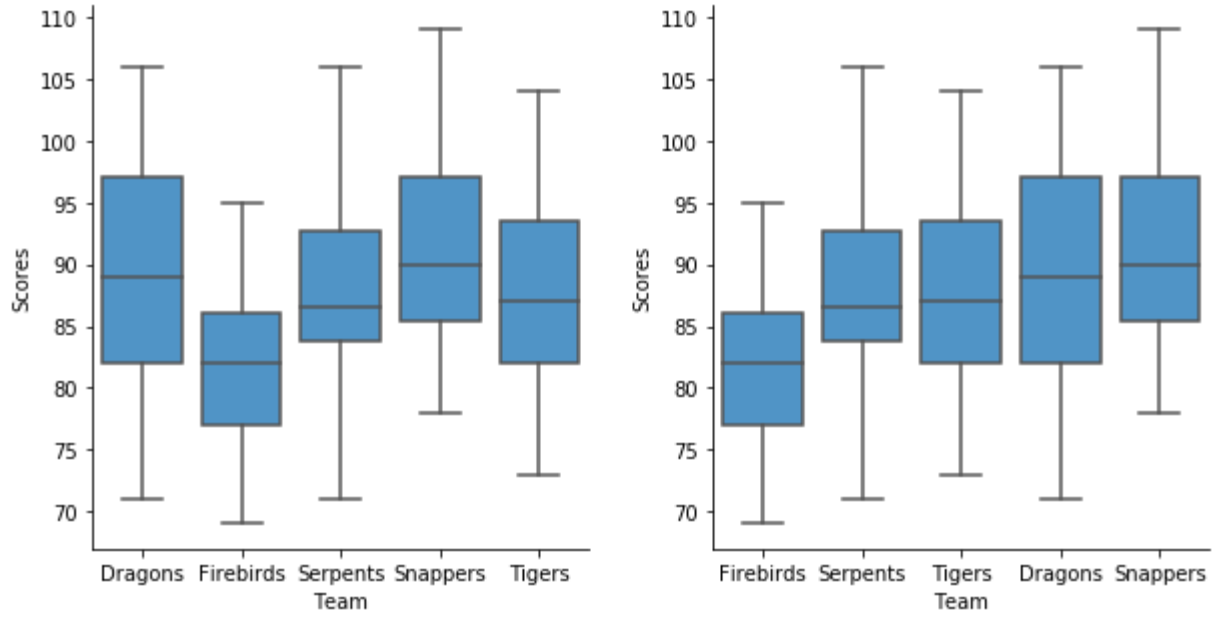
Compare multiple groups

Box plots are at their best when a comparison in distributions needs to be performed between groups. They are compact in their summarization of data, and it is easy to compare groups through the box and whisker markings' positions.

It is less easy to justify a box plot when you only have one group's distribution to plot. Box plots offer only a high-level summary of the data and lack the ability to show the details of a data distribution's shape. With only one group, we have the freedom to choose a more detailed chart type like a histogram or a density curve.

Consider the order of groups

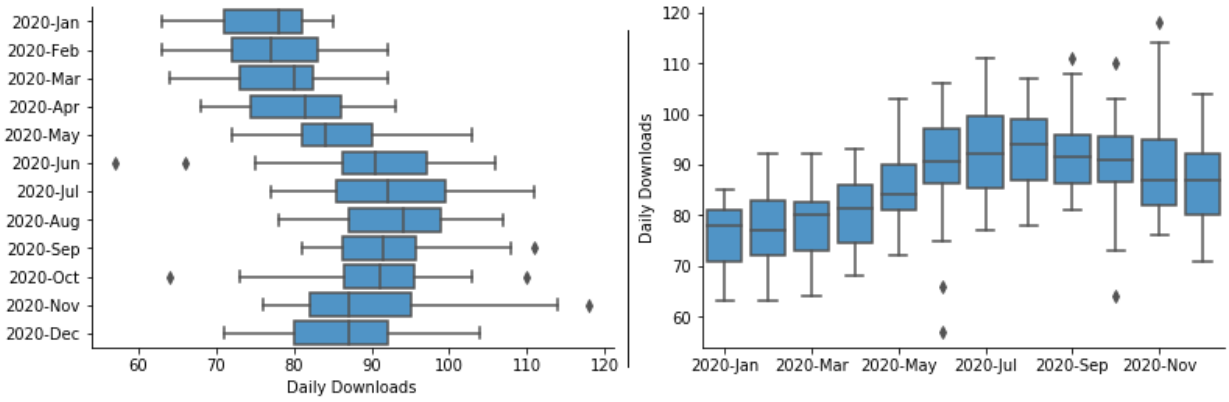
If the groups plotted in a box plot do not have an inherent order, then you should consider arranging them in an order that highlights patterns and insights. One common ordering for groups is to sort them by median value.



Common box plot options

Vertical vs. horizontal box plot

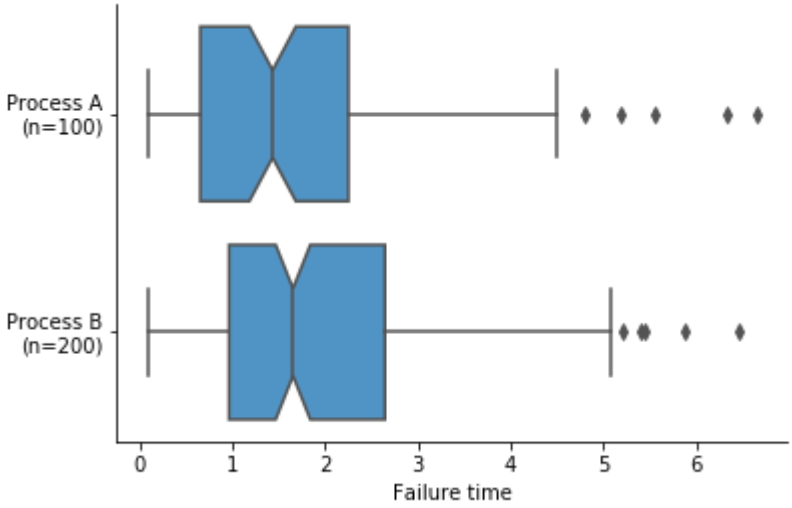
As observed through this article, it is possible to align a box plot such that the boxes are placed vertically (with groups on the horizontal axis) or horizontally (with groups aligned vertically). The horizontal orientation can be a useful format when there are a lot of groups to plot, or if those group names are long. It also allows for the rendering of long category names without rotation or truncation. On the other hand, a vertical orientation can be a more natural format when the grouping variable is based on units of time.



Variable box width and notches

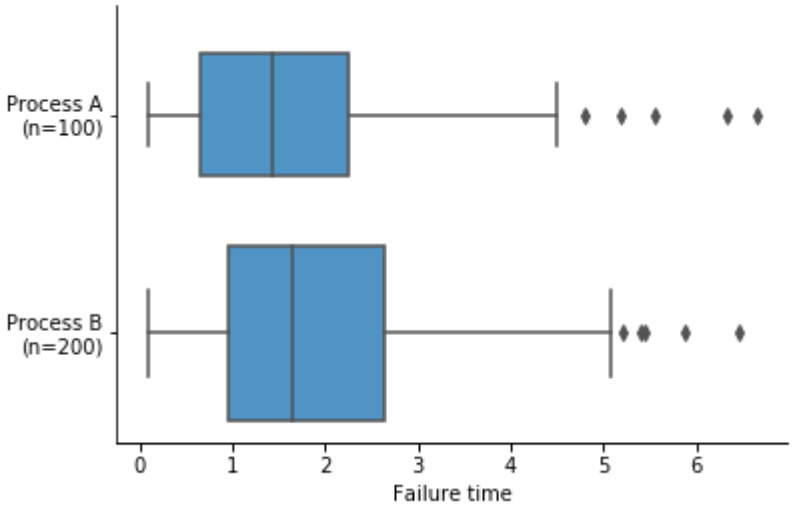
Certain visualization tools include options to encode additional statistical information into box plots. This is useful when the collected data represents sampled observations from a larger population.

Notches are used to show the most likely values expected for the median when the data represents a sample. When a comparison is made between groups, you can tell if the difference between medians are statistically significant based on if their ranges overlap. If any of the notch areas overlap, then we can't say that the medians are statistically different; if they do not have overlap, then we can have good confidence that the true medians differ.



This plot suggests that Process B creates components with better (higher) failure times, but the overlapping notches indicate the difference in medians is not statistically significant.

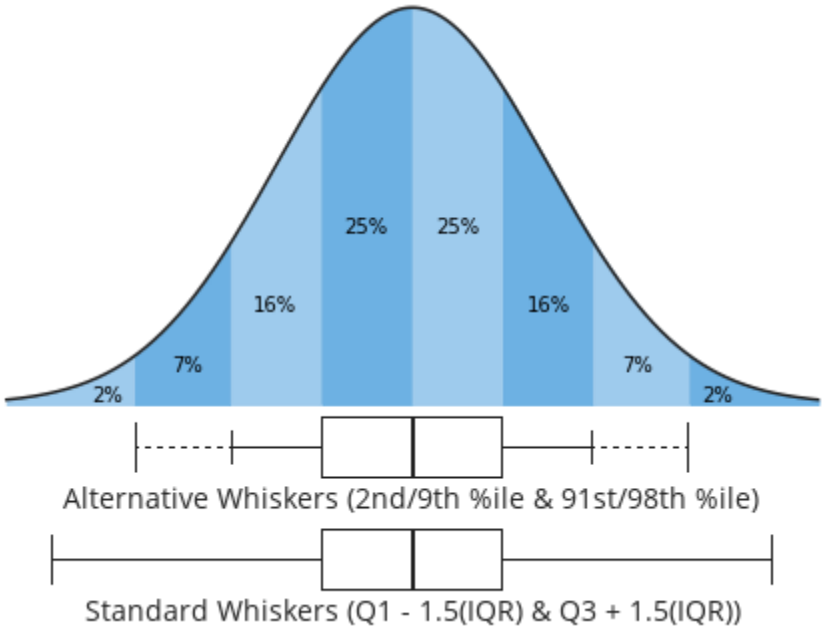
Box width can be used as an indicator of how many data points fall into each group. Box width is often scaled to the square root of the number of data points, since the square root is proportional to the uncertainty (i.e. standard error) we have about true values. Since interpreting box width is not always intuitive, another alternative is to add an annotation with each group name to note how many points are in each group.



Whisker range and outliers

There are multiple ways of defining the maximum length of the whiskers extending from the ends of the boxes in a box plot. As noted above, the traditional way of extending the whiskers is to the furthest data point within 1.5 times the IQR from each box end. Alternatively, you might place whisker markings at other percentiles of data, like how the box components sit at the 25th, 50th, and 75th percentiles.

Common alternative whisker positions include the 9th and 91st percentiles, or the 2nd and 98th percentiles. These are based on the properties of the normal distribution, relative to the three central quartiles. Under the normal distribution, the distance between the 9th and 25th (or 91st and 75th) percentiles should be about the same size as the distance between the 25th and 50th (or 50th and 75th) percentiles, while the distance between the 2nd and 25th (or 98th and 75th) percentiles should be about the same as the distance between the 25th and 75th percentiles. This can help aid the at-a-glance aspect of the box plot, to tell if data is symmetric or skewed.



When one of these alternative whisker specifications is used, it is a good idea to note this on or near the plot to avoid confusion with the traditional whisker length formula.

Letter-value plots

As developed by Hofmann, Kafadar, and Wickham, letter-value plots are an extension of the standard box plot. Letter-value plots use multiple boxes to enclose increasingly-larger proportions of the dataset. The first box still covers the central 50%, and the second box extends from the first to cover half of the remaining area (75% overall, 12.5% left over on each end). The third box covers another half of the remaining area (87.5% overall, 6.25% left on each end), and so on until the procedure ends and the leftover points are marked as outliers.

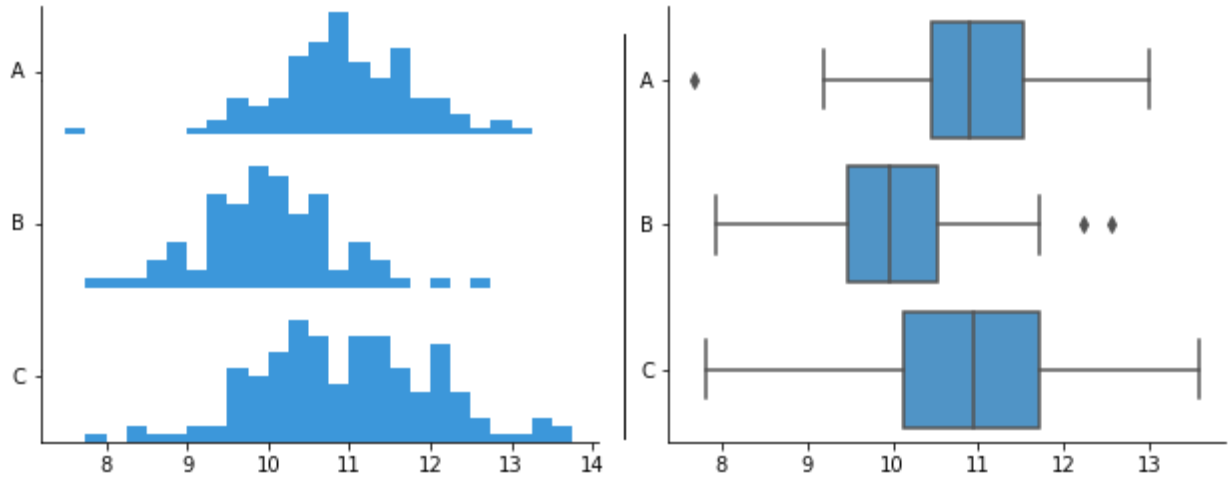
The letter-value plot is motivated by the fact that when more data is collected, more stable estimates of the tails can be made. In addition, more data points mean that more of them will be labeled as outliers, whether legitimately or not. While the letter-value plot is still somewhat lacking in showing some distributional details like modality, it can be a more thorough way of making comparisons between groups when a lot of data is available.

Related plots

Histogram

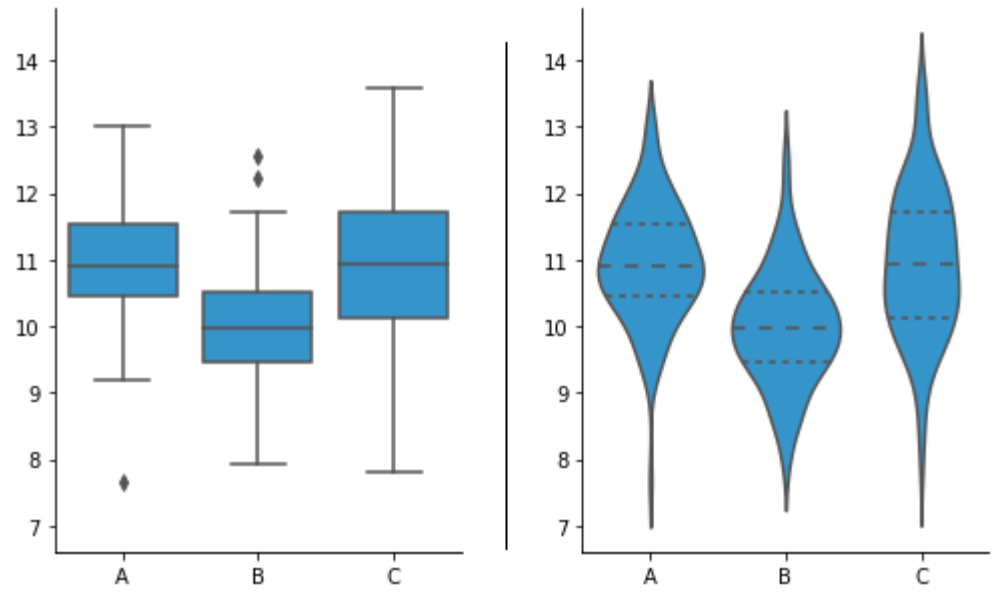
As noted above, when you want to only plot the distribution of a single group, it is recommended that you use a histogram rather than a box plot. While a histogram does not include direct indications of quartiles like a box plot, the additional information about distributional shape is often a worthy tradeoff.

With two or more groups, multiple histograms can be stacked in a column like with a horizontal box plot. Note, however, that as more groups need to be plotted, it will become increasingly noisy and difficult to make out the shape of each group's histogram. In addition, the lack of statistical markings can make a comparison between groups trickier to perform. For these reasons, the box plot's summarizations can be preferable for the purpose of drawing comparisons between groups.



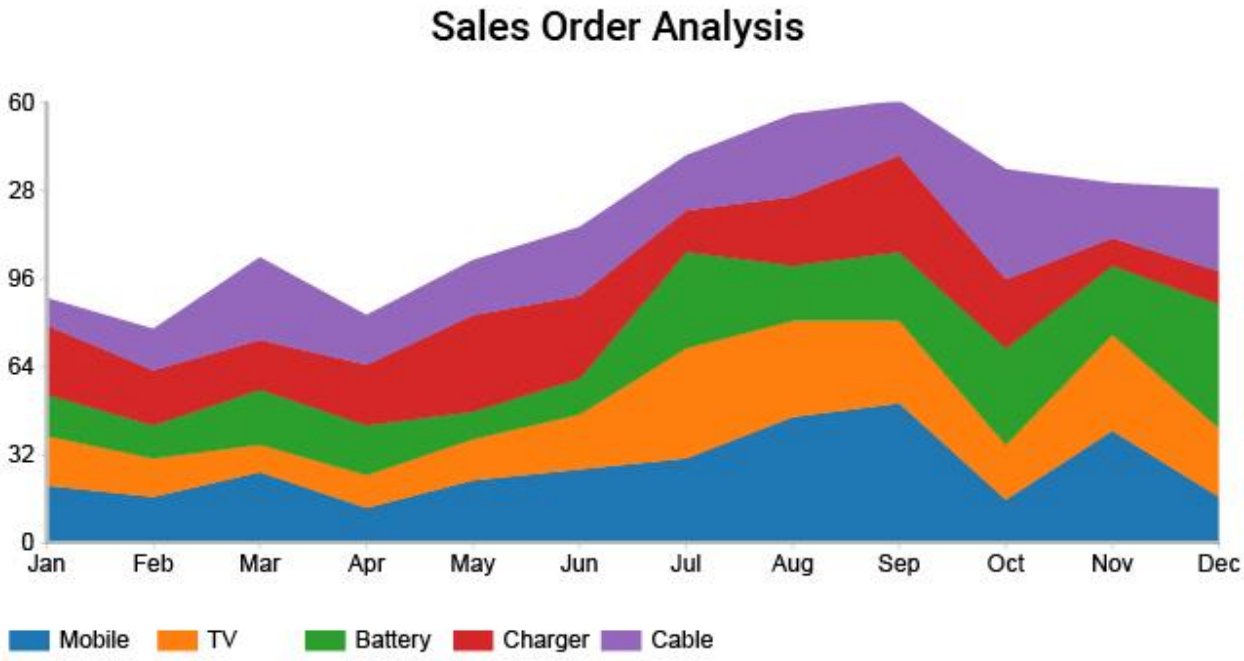
Violin plot

One alternative to the box plot is the violin plot. In a violin plot, each group's distribution is indicated by a density curve. In a density curve, each data point does not fall into a single bin like in a histogram, but instead contributes a small volume of area to the total distribution. Violin plots are a compact way of comparing distributions between groups. Often, additional markings are added to the violin plot to also provide the standard box plot information, but this can make the resulting plot noisier to read.



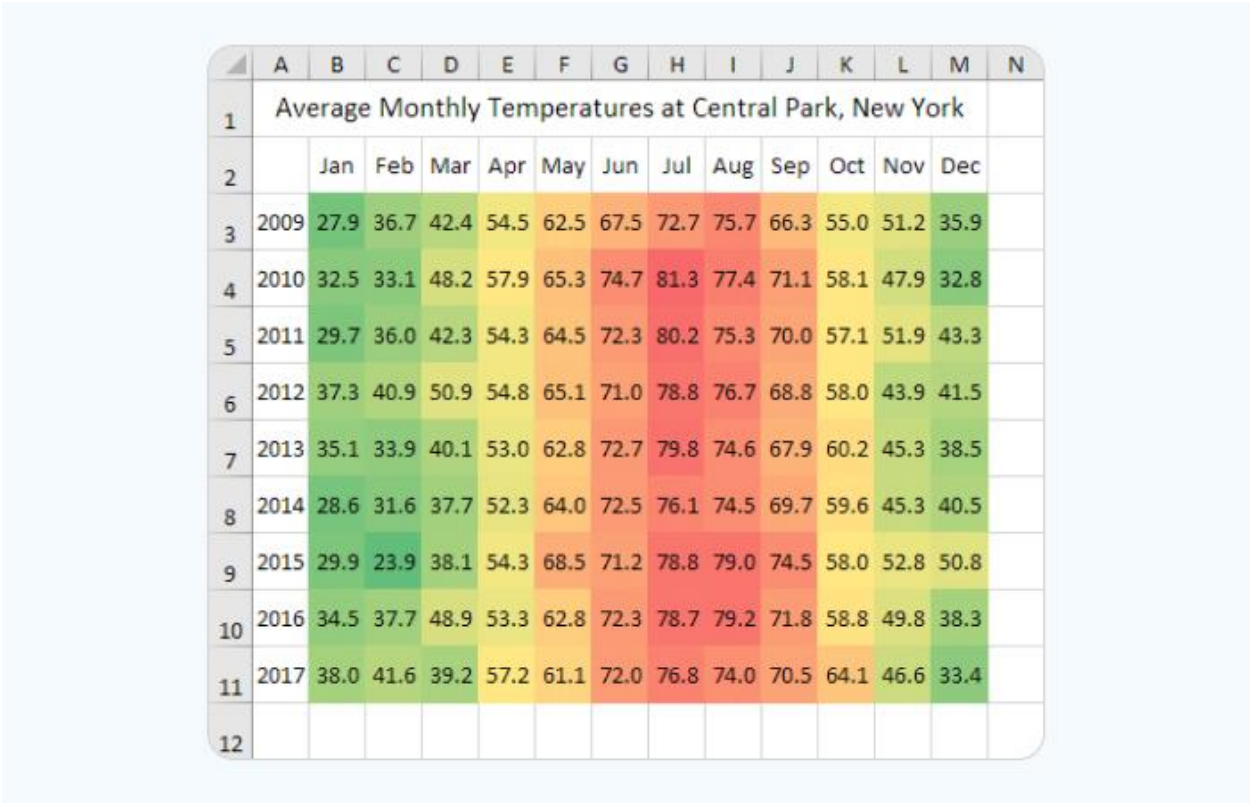
7. Area Chart

- **When to Use**: To show trends over time and also compare the contribution to the total for multiple categories.
- **Best For**: Continuous data, cumulative totals (e.g., revenue contribution by region).
- **Example**: Displaying cumulative sales over time, with each area representing a different product.



8. Heatmap

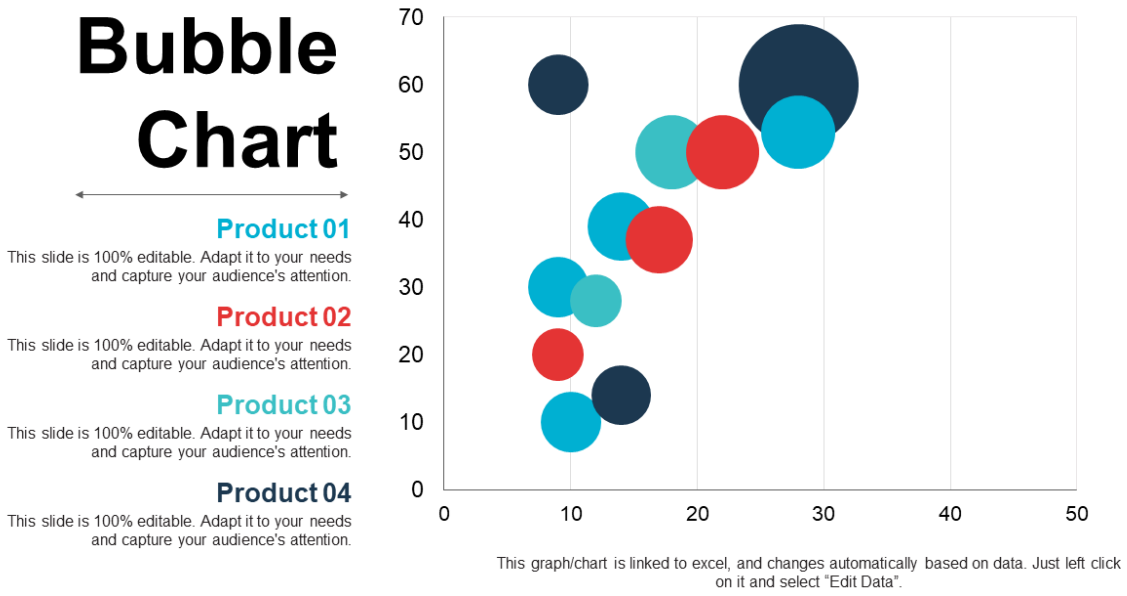
- **When to Use**: To show the intensity of data across two dimensions, often in a grid format.
- **Best For**: Visualizing large datasets, showing patterns or correlations.
- **Example**: Showing the correlation between various variables or the intensity of website user activity.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Average Monthly Temperatures at Central Park, New York													
2		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
3	2009	27.9	36.7	42.4	54.5	62.5	67.5	72.7	75.7	66.3	55.0	51.2	35.9	
4	2010	32.5	33.1	48.2	57.9	65.3	74.7	81.3	77.4	71.1	58.1	47.9	32.8	
5	2011	29.7	36.0	42.3	54.3	64.5	72.3	80.2	75.3	70.0	57.1	51.9	43.3	
6	2012	37.3	40.9	50.9	54.8	65.1	71.0	78.8	76.7	68.8	58.0	43.9	41.5	
7	2013	35.1	33.9	40.1	53.0	62.8	72.7	79.8	74.6	67.9	60.2	45.3	38.5	
8	2014	28.6	31.6	37.7	52.3	64.0	72.5	76.1	74.5	69.7	59.6	45.3	40.5	
9	2015	29.9	23.9	38.1	54.3	68.5	71.2	78.8	79.0	74.5	58.0	52.8	50.8	
10	2016	34.5	37.7	48.9	53.3	62.8	72.3	78.7	79.2	71.8	58.8	49.8	38.3	
11	2017	38.0	41.6	39.2	57.2	61.1	72.0	76.8	74.0	70.5	64.1	46.6	33.4	
12														

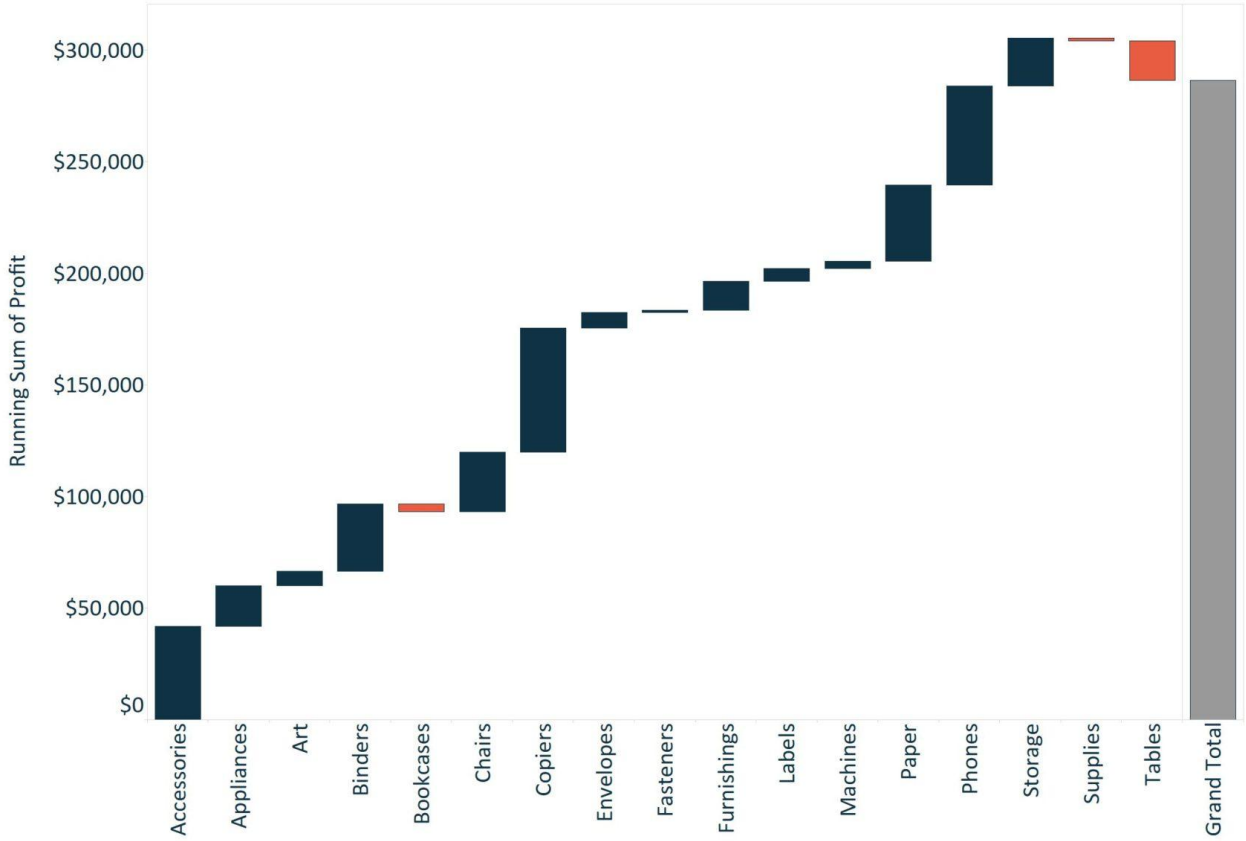
9. Bubble Chart

- **When to Use**: To show relationships between three variables (with the size of the bubble representing the third variable).
- **Best For**: Showing multi-variable comparisons.
- **Example**: Visualizing the relationship between sales, profits, and market size.



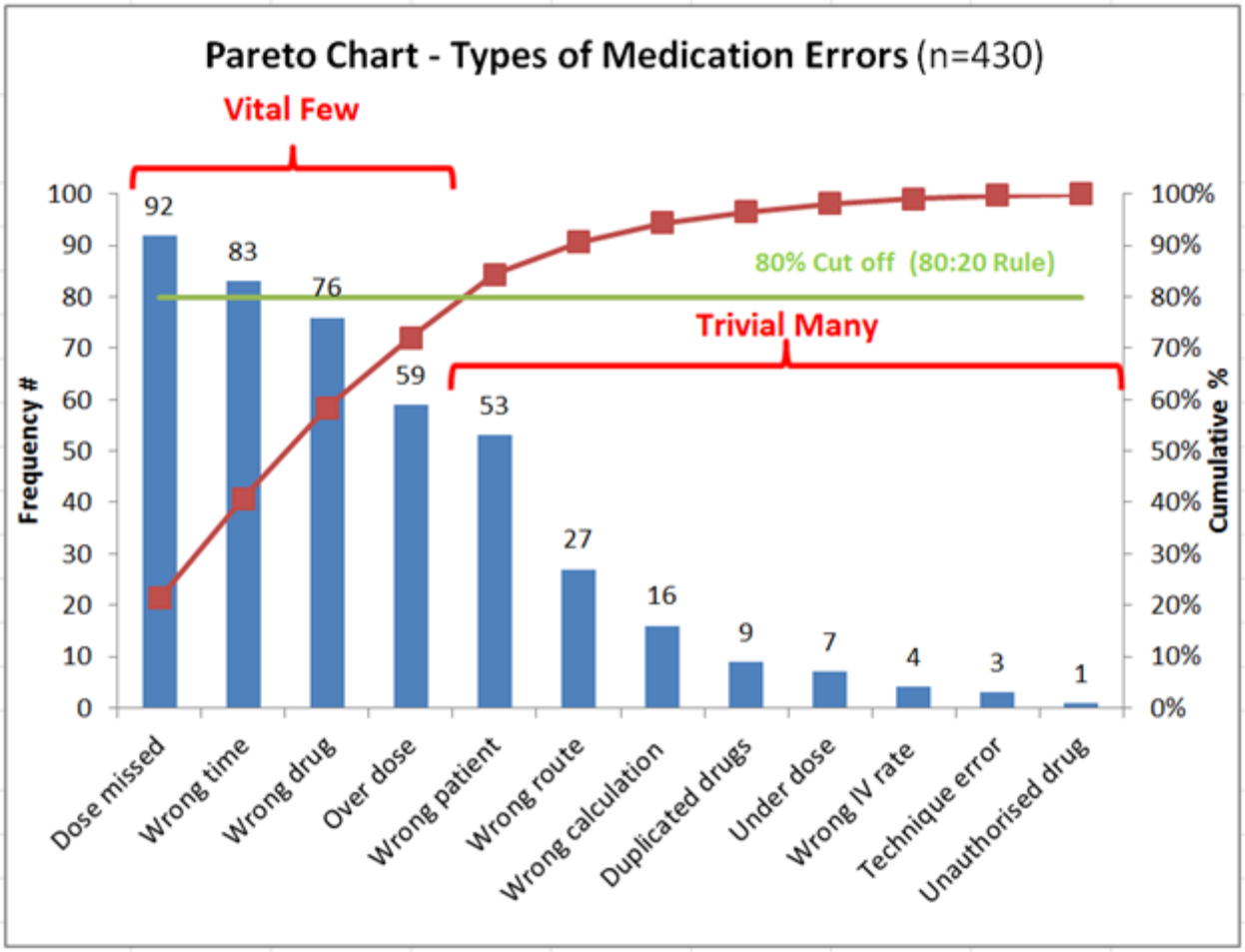
10. Waterfall Chart

- **When to Use**: To show how an initial value is increased and decreased by a series of intermediate values, leading to a final result.
- **Best For**: Financial data (e.g., profit and loss statements), showing cumulative impact.
- **Example**: Displaying how revenue changes from gross sales to net income by adding/removing expenses and taxes.



11. Pareto Chart

- **When to Use**: To identify the most significant factors in a dataset.
- **Best For**: Highlighting the 80/20 rule (Pareto principle), focusing on the few key causes contributing to the majority of problems or effects.
- **Example**: Showing the most common causes of product defects in a factory.



80/20 Rule – The Pareto Principle

The 80/20 Rule (also known as the **Pareto principle** or the law of the vital few & trivial many) states that, for many events, roughly 80% of the effects come from 20% of the causes. Joseph Juran (a well regarded Quality Management consultant) suggested the principle and named it after the Italian economist Vilfredo Pareto, who noted the 80/20 connection in 1896.

Vilfredo Pareto showed that approximately 80% of the land in Italy was owned by 20% of the population. Pareto also observed that 20% of the peapods in his garden contained 80% of the peas. According to the Pareto Principle, in any group of things that contribute to a common effect, a relatively few contributors account for the majority of the effect. Commonly, it is found that:

- 80% of complaints come from 20% of customers
- 80% of sales come from 20% of clients
- 80% of computer crashes come from 20% of IT bugs

The ordering in a Pareto Chart helps identify the 'vital few' (the factors that warrant the most attention i.e. factors whose cumulative per cent (dots) fall under the 80% cut off line) from the 'trivial many' (factors that, while useful to know about, have a relatively smaller effect i.e. cumulative per cent dots that fall above the 80% cut off line).

Using a Pareto diagram helps a team concentrate its efforts on the factors that have the greatest impact. It also helps a team communicate the rationale for focusing on certain areas.

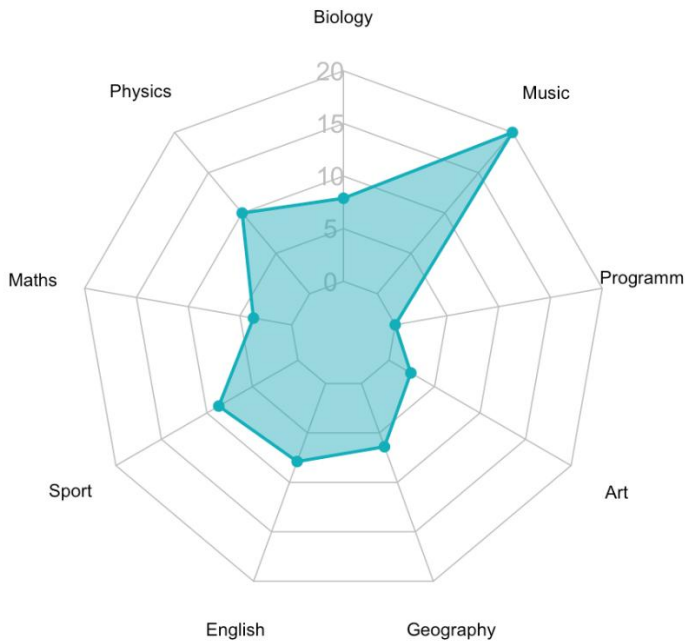
The example in Figure 1 (above) shows a Pareto Chart of types of medication errors. An audit of 430 medication errors was conducted to determine the categories (types) of errors and their frequency. The results were collected initially in a Tally Sheet then the data was placed in descending order of frequency in a Pareto Chart Template in Excel.

The types of errors that fall under the 80% cut off line indicate the 'vital few' types of medication error that should be addressed as a priority as they contribute most to the problem ie:

- Dose missed
- Wrong time
- Wrong drug
- Over dose

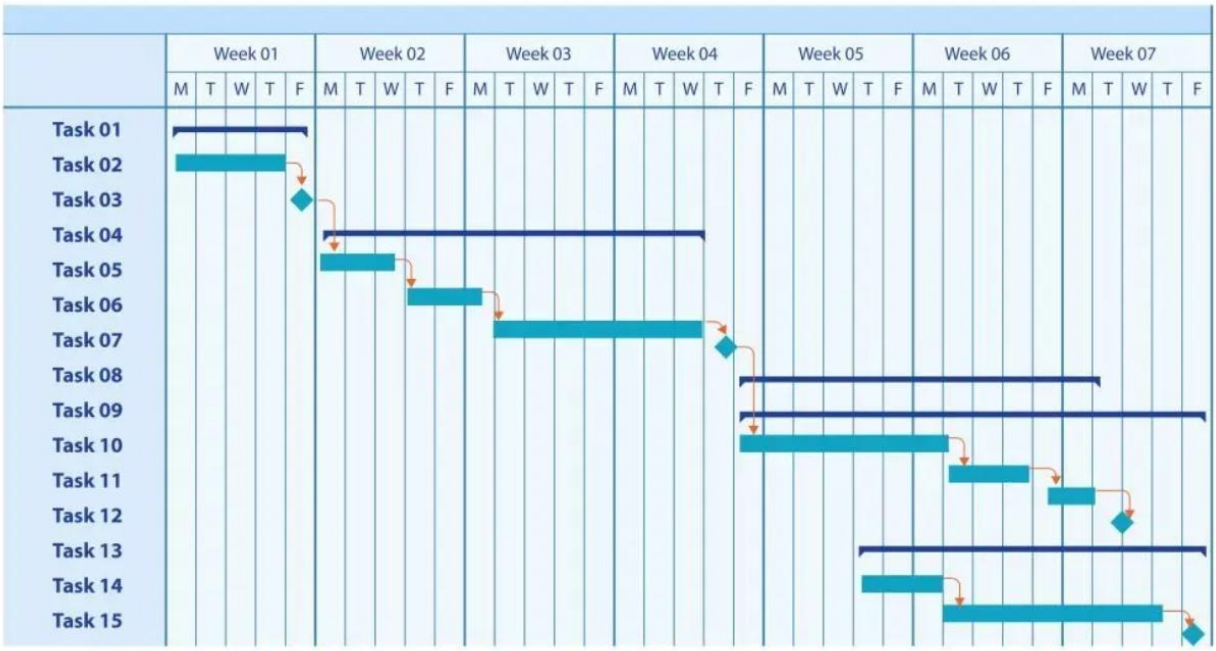
12. Radar Chart (Spider Chart)

- **When to Use**: To compare multiple variables for different categories on a single chart.
- **Best For**: Showing strengths and weaknesses across multiple dimensions.
- **Example**: Comparing the performance of various products across different metrics (e.g., price, quality, and durability).



13. Gantt Chart

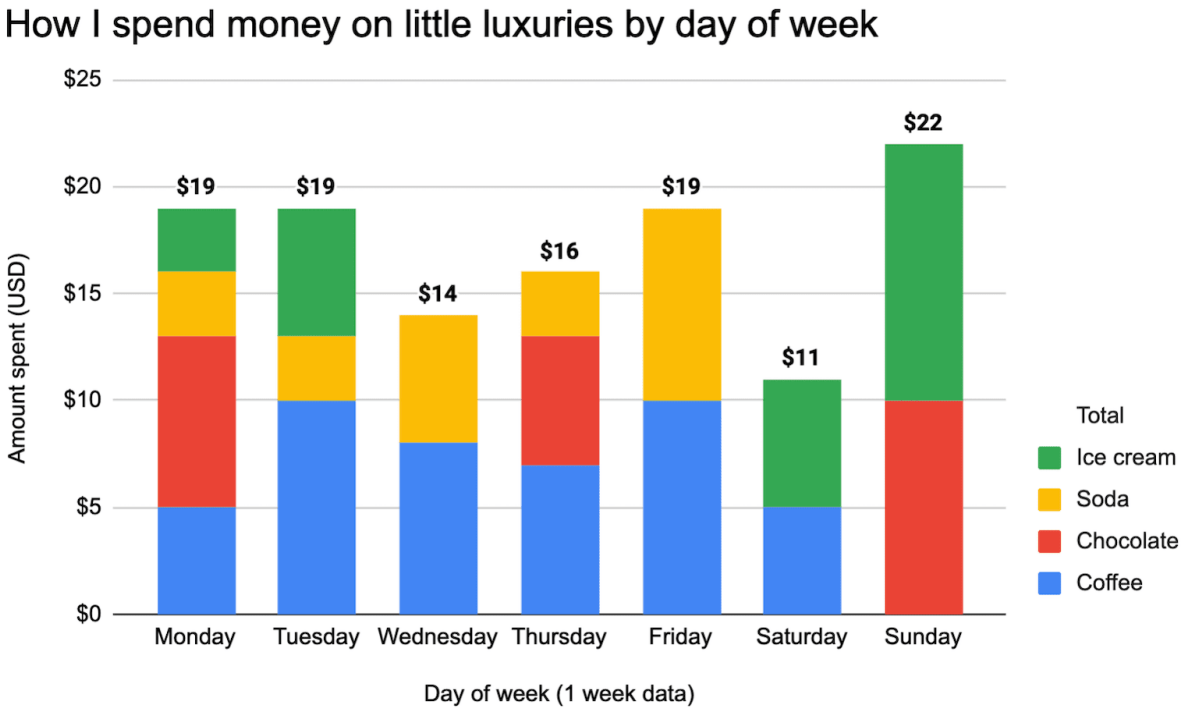
- **When to Use**: To manage project timelines by showing tasks, their duration, and dependencies.
- **Best For**: Project management, tracking the progress of tasks over time.
- **Example**: Visualizing project milestones and task schedules.



14. Stacked Bar

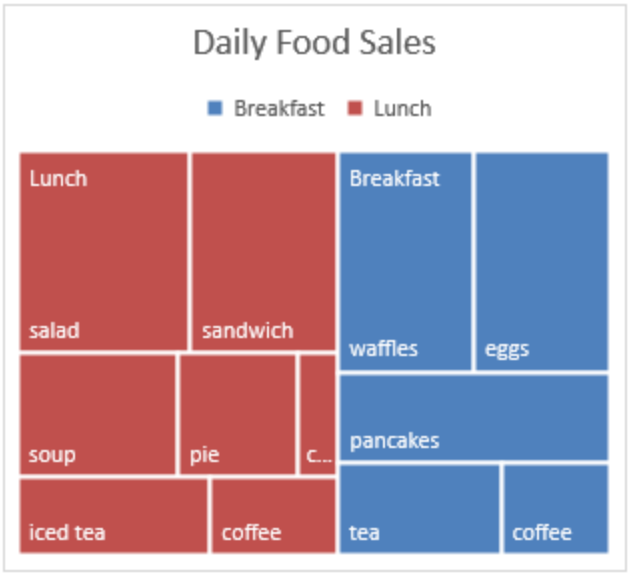
When to Use: To show the composition of categories and how they contribute to the total.

- **Best For**: Comparing parts of a whole across different categories.
- **Example**: Showing the breakdown of expenses for different departments over time.



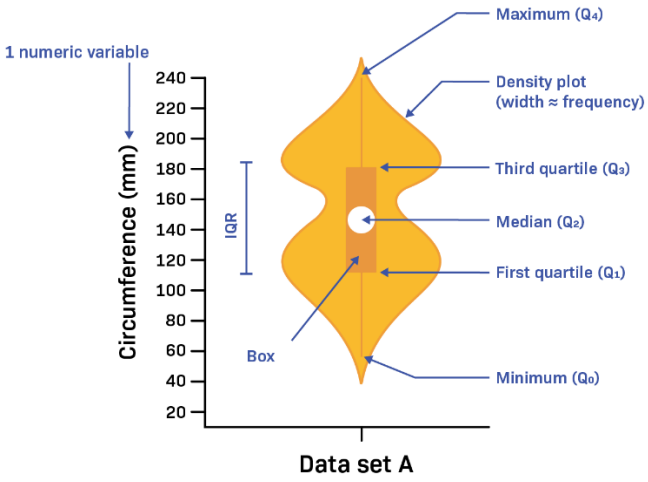
15. Tree Map

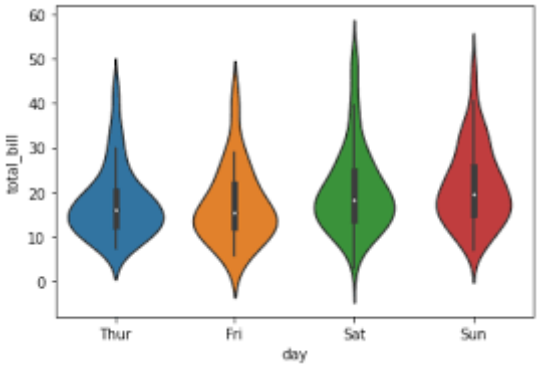
- **When to Use**: To show hierarchical data as a set of nested rectangles.
- **Best For**: Displaying proportions of parts to the whole, especially when you have a lot of categories.
- **Example**: Visualizing the proportion of different segments of a business within the total revenue.



16. Violin Plot

- **When to Use**: To show the distribution of the data and its probability density.
- **Best For**: Comparing the distribution of data between different categories.
- **Example**: Displaying the distribution of exam scores across different classes.





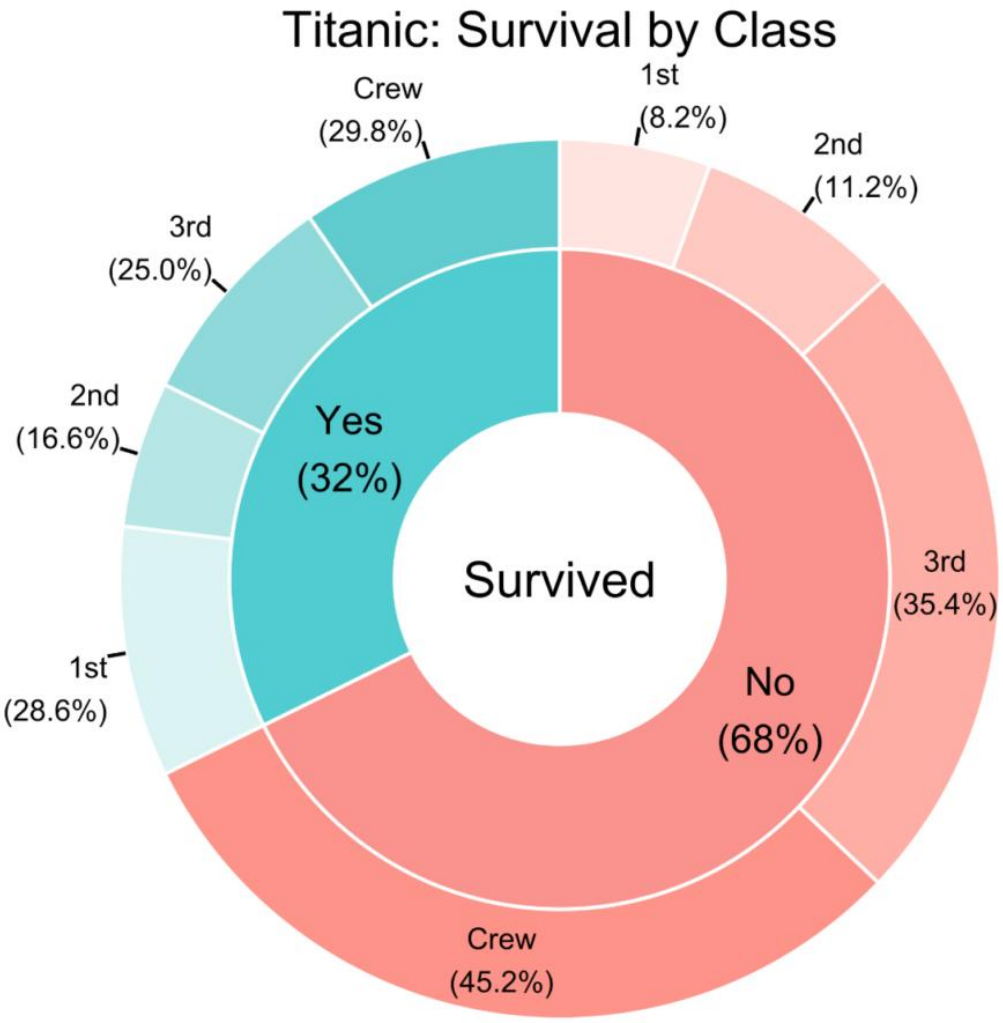
17. Funnel Chart

- **When to Use**: To represent stages in a process, showing how data progresses through stages.
- **Best For**: Visualizing sales or conversion funnels.
- **Example**: Tracking leads through stages of the sales pipeline from initial contact to final sale.

Sales Pipeline	
Stage	Amount
Prospects	500
Qualified prospects	425
Needs analysis	200
Price quotes	150
Negotiations	100
Closed sales	90

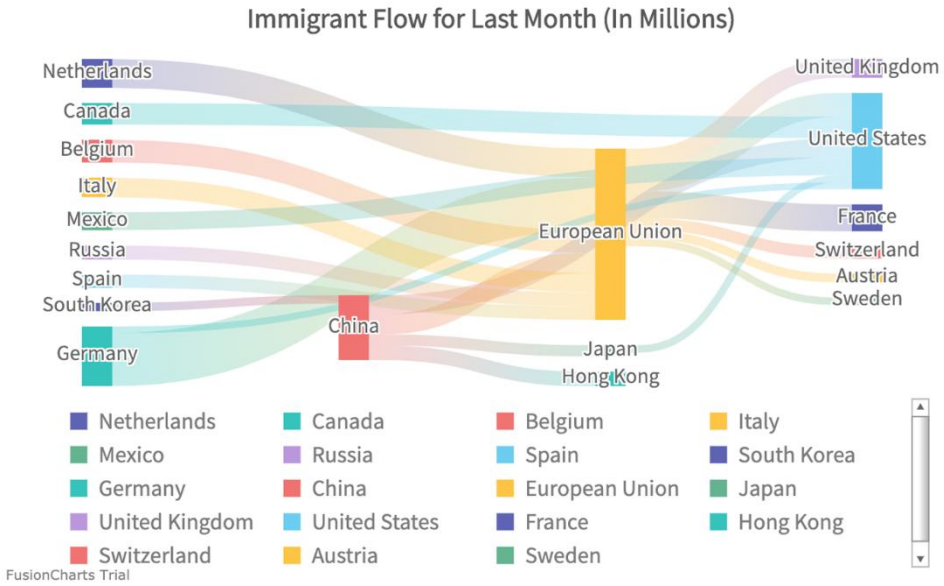
18. Donut Chart

- **When to Use**: Similar to a pie chart, but with the ability to display an additional layer of information in the center.
- **Best For**: Showing parts of a whole with some additional detail in the middle.
- **Example**: Displaying the breakdown of revenue by region with a summary or total in the center.



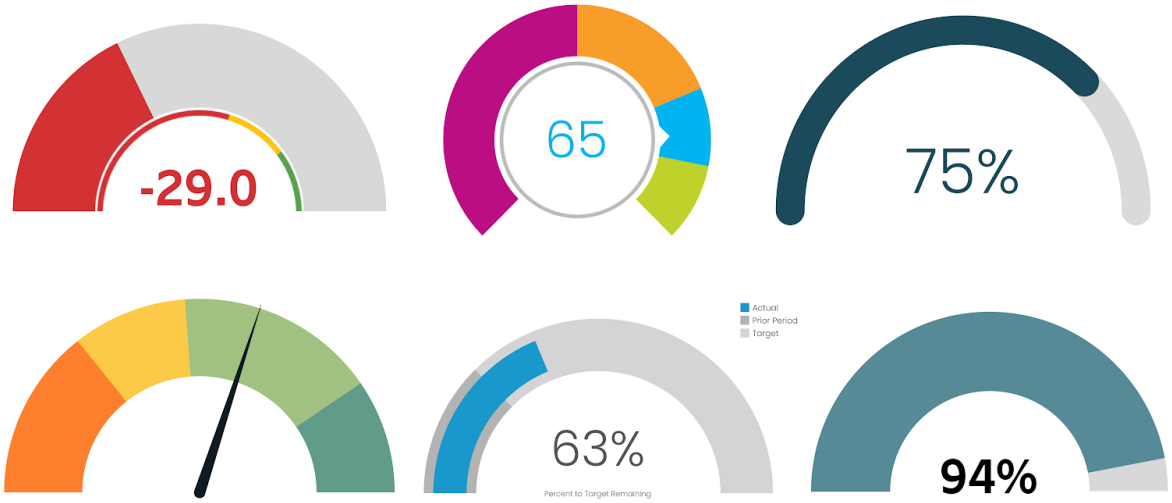
19. Sankey Diagram

- **When to Use**: To visualize the flow and relationship between different entities or processes.
- **Best For**: Energy or resource flow, financial transactions, or showing the movement of quantities.
- **Example**: Visualizing energy consumption from sources to applications.



20. Gauge Chart

- **When to Use**: To show a single measure within a range, often used for KPIs.
- **Best For**: Displaying performance metrics against a target (e.g., speed, customer satisfaction).
- **Example**: Showing progress toward a goal or quota.



General Guidelines:

- **Comparisons**: Use bar charts or line charts for easy comparisons between categories or over time.
- **Relationships**: Use scatter plots or bubble charts to show relationships or correlations between variables.
- **Trends**: Use line charts or area charts to illustrate trends over time.
- **Proportions**: Use pie charts or stacked bar charts to show parts of a whole.
- **Distributions**: Use histograms, box plots, or violin plots to show the distribution of data.
- **Process/Flow**: Use Gantt charts, Sankey diagrams, or funnel charts to show processes or flows.

By understanding the nature of your data and the story you want to tell, you can choose the most effective chart or graph to convey the message clearly.