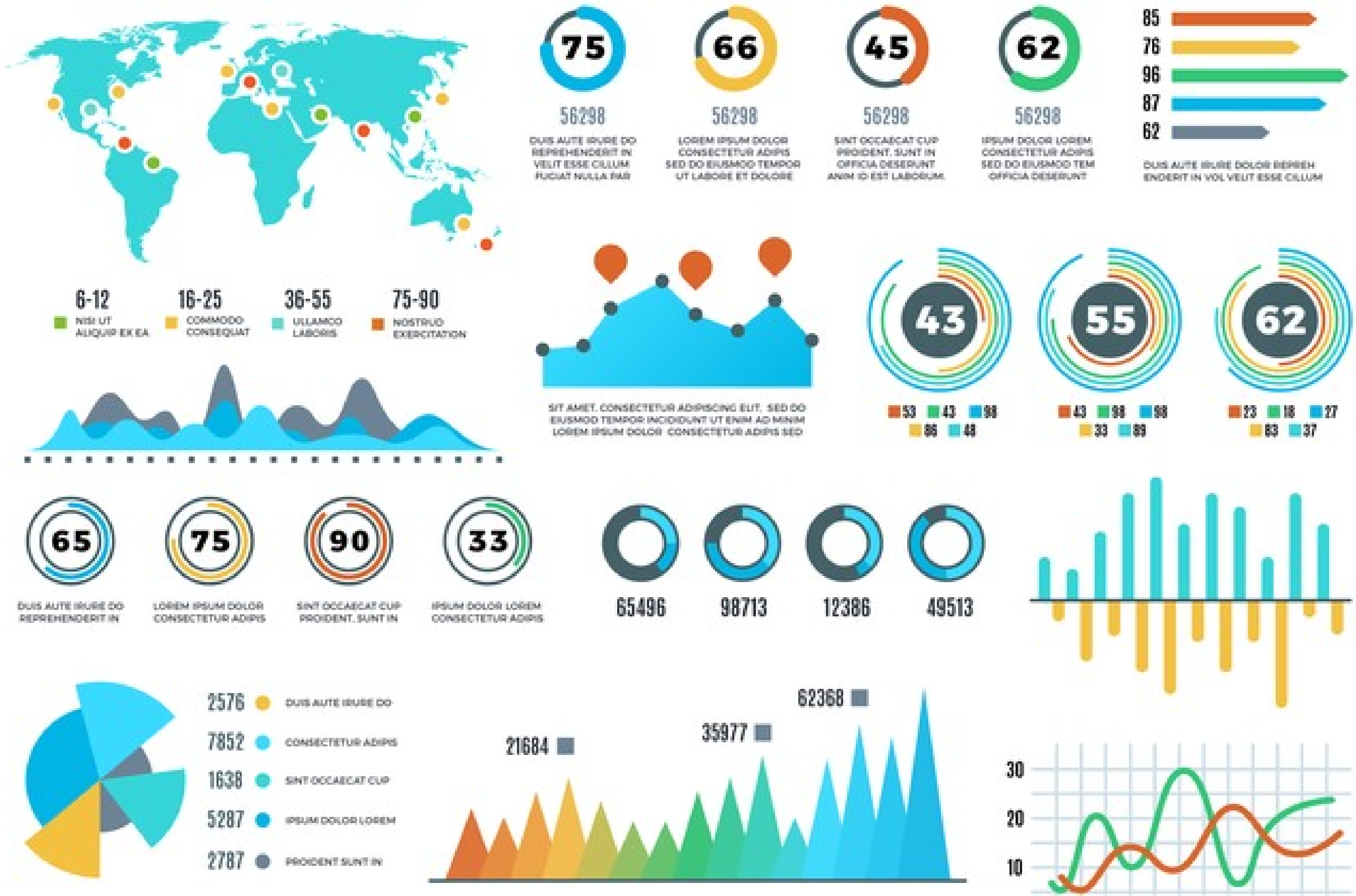
Data Visualization in Data Science

Tushar B. Kute,
http://tusharkute.com

Agenda

- Data visualization
  - What? Why?
  - Benefits
  - Techniques
  - Who uses it?
- Types of Graphs
- Tools
- Techniques in programming
- Best resources

Just check...


75
56298
DUIS AUTE IRURE DO REPREHENDERIT IN VELIT ESSE CILLUM FUGIAT NULLA PAR
66
56298
LOREM IPSUM DOLOR CONSECTETUR ADIPIS SED DO EIUSMOD TEMPOR UT LABORE ET DOLORE
45
56298
SINT OCCAECAT CUP PROIDENT, SUNT IN OFFICIA DESERUNT ANIM ID EST LABORUM.
62
56298
IPSUM DOLOR LOREM CONSECTETUR ADIPIS SED DO EIUSMOD TEM OFFICIA DESERUNT
85
76
96
87
62
DUIS AUTE IRURE DOLOR REPREH ENDERIT IN VOL VELIT ESSE CILLUM
6-12
NISI UT ALIQUIP EX EA
16-25
COMMODO CONSEQUAT
36-55
ULLAMCO LABORIS
75-90
NOSTRUD EXERCITATION
SIT AMET, CONSECTETUR ADIPISCING ELIT. SED DO EIUSMOD TEMPOR INCIDIDUNT UT ENIM AD MINIM LOREM IPSUM DOLOR CONSECTETUR ADIPIS SED
43
53
43
98
86
48
55
43
98
98
33
89
62
23
18
27
83
37
65
DUIS AUTE IRURE DO REPREHENDERIT IN
75
LOREM IPSUM DOLOR CONSECTETUR ADIPIS
90
SINT OCCAECAT CUP PROIDENT, SUNT IN
33
IPSUM DOLOR LOREM CONSECTETUR ADIPIS
65496
98713
12386
49513
2576
DUIS AUTE IRURE DO
7852
CONSECTETUR ADIPIS
1638
SINT OCCAECAT CUP
5287
IPSUM DOLOR LOREM
2787
PROIDENT SUNT IN
21684
35977
62368
30
20
10

Data Visualization

- Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.
- The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.
- The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data Visualization

- Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made.
- Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Data Visualization

- Data visualization is important for almost every career.
- It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders.
- It also plays an important role in big data projects. As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to get an overview of their data quickly and easily.
- Visualization tools were a natural fit.

Benefits of Data Visualization

- The ability to absorb information quickly, improve insights and make faster decisions;
- An increased understanding of the next steps that must be taken to improve the organization;
- An improved ability to maintain the audience's interest with information they can understand;
- An easy distribution of information that increases the opportunity to share insights with everyone involved;

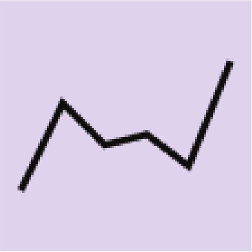
Benefits of Data Visualization

- Eliminate the need for data scientists since data is more accessible and understandable; and
- An increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes.

Data Visualization Roles

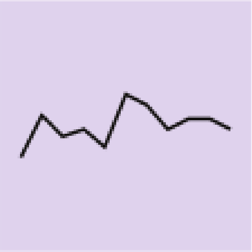
- Showing change over time
- Showing a part-to-whole composition
- Depicting flows and processes
- Looking at how data is distributed
- Comparing values between groups
- Observing relationships between variables
- Looking at geographical data

Change over time



Line chart ● +Comparisons

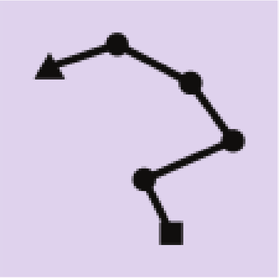
Most common chart type for showing change over time. A point is plotted for each time period from left to right; each point's vertical position indicates the feature's value. Points are connected by line segments to emphasize progression across time.



Sparkline ● +Comparisons

A miniature line chart with little to no labeling, designed to be placed alongside text or in tables. Provides a high-level overview without attracting too much attention. Can also be seen in a sparkbar form, or miniature bar chart (see below).

Change over time



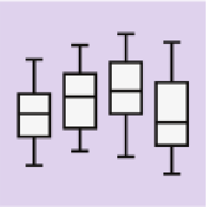
Connected scatter plot ◆ *+Relationships*

Shows change over time across two numeric variables (see scatter plot in *Relationships*). Line segments still connect points across time, but they may not consistently go from left to right like in a line chart.

Bar chart ● *+Distributions* *+Comparisons*

Each time period is associated with a bar; each bar's value is represented in its height above (or below) a zero-baseline. Works best when there aren't too many time periods to show.

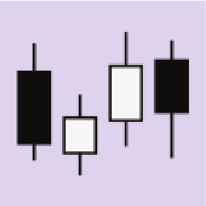
Change over time



Box plot 🟦 *+Distributions* *+Comparisons*

Each time period is associated with a box and whiskers; each set of box and whiskers shows the range of the most common data values. Best when there are multiple recordings for each time period and a distribution of values needs to be plotted.

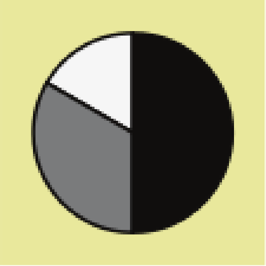
Tracking change over time is of key interest in the financial domain. One specialist chart developed for this field includes the following:



Candlestick chart ◆

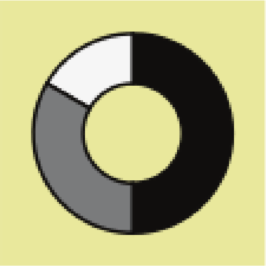
Looks like a box plot, but each box and whiskers encodes different statistics. The box ends indicate opening and closing prices, while color indicates the direction of change.

Part-to-whole composition



Pie chart ●

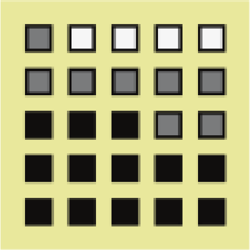
The whole is represented by a filled circle. Parts are proportional slices from that circle, one for each categorical group. Best with five or fewer slices with distinct proportions.



Doughnut chart ●

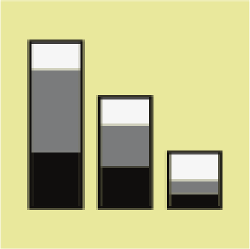
A pie chart with a hole in the center. This central area can be used to show a relevant single numeric value. Sometimes used as an aesthetic alternative to a standard progress bar (see stacked bar chart below).

Part-to-whole composition



Waffle chart / grid plot 🟦

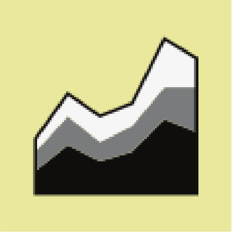
Squares laid out in a (typically) 10 x 10 grid; each square represents one percent of the whole. Squares are colored based on categorical group size.



Stacked bar chart 🟢

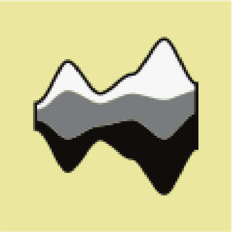
A bar chart (see *Change over time* or *Distributions*) where each bar has been divided into multiple sub-bars to show a part-to-whole breakdown. A single stacked bar can be used as an alternative to the pie or doughnut chart; people tend to make more precise judgments of length over area or angle.

Part-to-whole composition



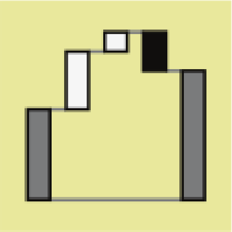
Stacked area chart ●

A line chart (see *Change over time*) where shaded regions are added under the line to divide the total into sub-group values.



Stream graph ◆

Modified version of the stacked area chart where areas are stacked around a central axis. Highlights relative changes instead of exact values.

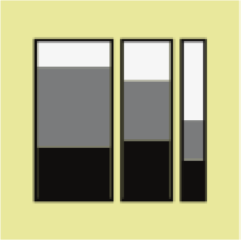


Waterfall chart ◆

Augments a change over time with a part-to-whole decomposition. Bars on the ends depict values at two time points, and lengths of intermediate floating bars' show the decomposition of the change between points.

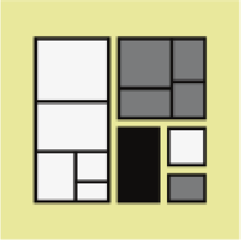
Part-to-whole composition

Certain part-to-whole compositions follow a hierarchical form. In these cases, each part can be divided into finer parts on lower levels. Here are a couple of more specialized chart types for visualizing this type of data:



Mosaic plot / Marimekko chart 🟦

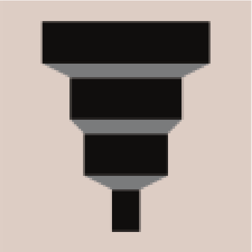
Can be thought of as a stacked bar divided on both axes. A box is divided on one axis based on one categorical variable, then each sub-box is divided in the other axis based on a second categorical variable.



Treemap ◆

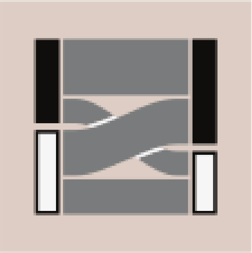
Can be thought of as a more generalized Marimekko plot. Sub-boxes do not need to have a consistent cut direction at a particular hierarchy level, and there can be more than two levels of hierarchy.

Flows and processes



Funnel chart 🟦

Seen in business contexts, showing how people encounter a product and eventually become users or customers. One bar is plotted for each stage, whose lengths reflect the number of users. Connecting regions emphasize connections in stages and give the chart type's namesake shape.



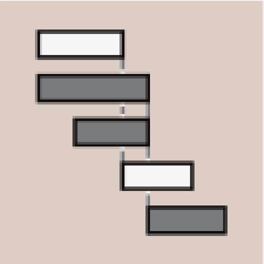
Parallel sets chart ◆

Multiple part-to-whole divisions on different dimensions are depicted as parallel stacked bars. Connecting regions show how different subgroups relate to one another between dimensions.

Flows and processes

Sankey diagram ◆

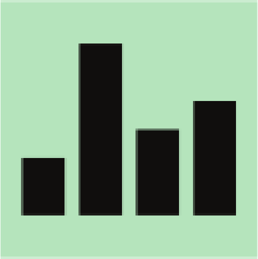
The width of the colored region shows the relative volume at each part of a process. Allows for multiple sources of inputs and outputs to be visualized.



Gantt chart 🟦

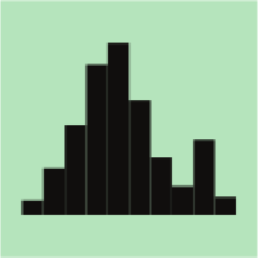
Used for project scheduling, breaking them down into individual tasks. Each task is associated with a bar, providing a timeline for when each task should begin and end.

How data is distributed?



Bar chart ● +Change over time +Comparisons

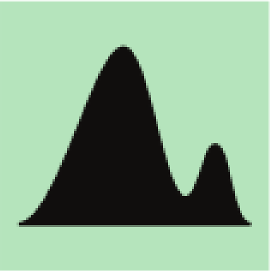
Used when a variable is qualitative or takes discrete values. The height of each bar indicates the amount of each categorical group.



Histogram ●

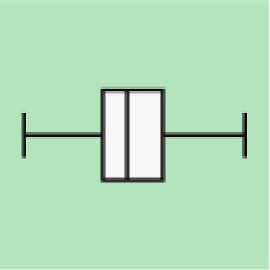
Similar to a bar chart, but used when a variable takes continuous numeric values. The variable's numeric range is divided into bins for aggregating counts. Bars are plotted flush against each other to emphasize the variable's continuous nature.

How data is distributed?



Density curve 🟢

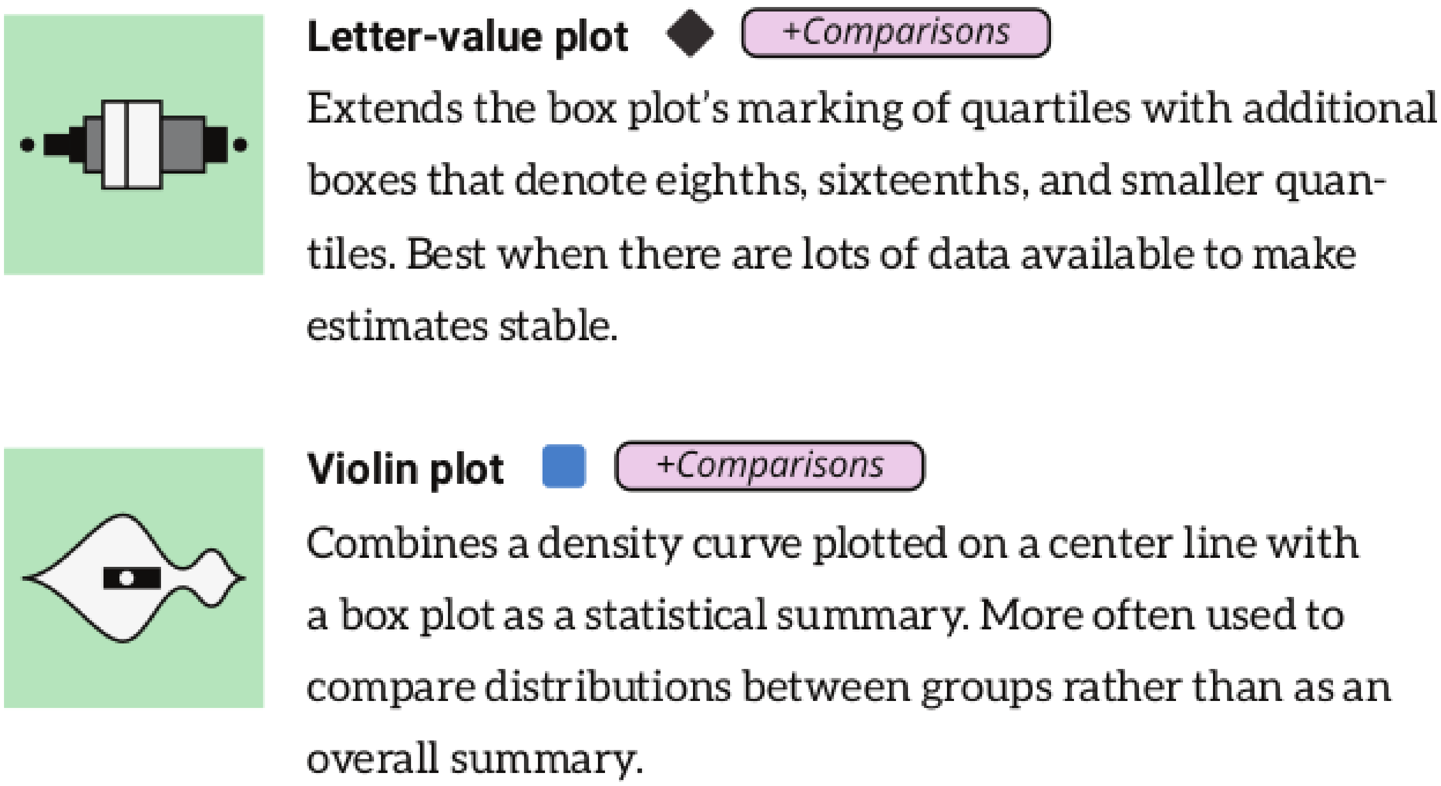
An alternative to the histogram when a variable takes numeric values. Each data point contributes a small amount of local area; the areas are summed across all points to form the full curve.



Box plot 🟦 *+Change over time* *+Comparisons*

A box and whiskers shows the range of the most common data values. The ends of the box outline the central 50% of the data. More often used to compare distributions between groups rather than as an overall summary.

How data is distributed?



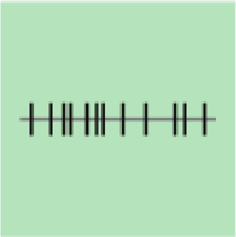
Letter-value plot ◆ *+Comparisons*

Extends the box plot's marking of quartiles with additional boxes that denote eighths, sixteenths, and smaller quantiles. Best when there are lots of data available to make estimates stable.

Violin plot 🟦 *+Comparisons*

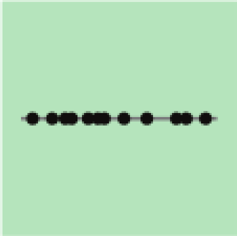
Combines a density curve plotted on a center line with a box plot as a statistical summary. More often used to compare distributions between groups rather than as an overall summary.

How data is distributed?



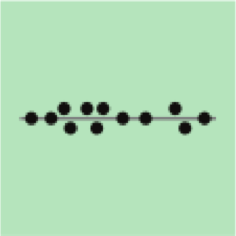
Rug plot 🟦

All data points are plotted as tick marks on a straight line with value corresponding precisely with position.



Strip plot 🟦

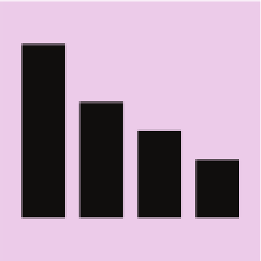
Like a rug plot, but with dots instead of tick marks. Sometimes plotted with points randomly jittered up or down to reduce overlapping.



Swarm plot ◆

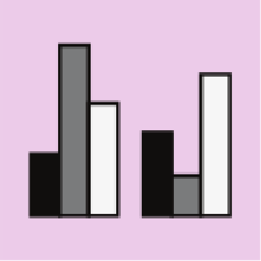
Like a strip plot, but deliberate shifting is performed to prevent overlapping. Some horizontal jitter may be needed in order to keep the dot swarm compact.

Comparing values between groups



Bar chart ● +Change over time +Distributions

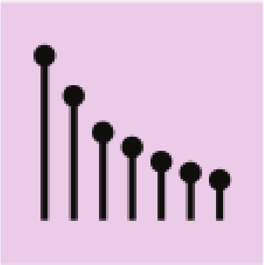
Most basic way of comparing numeric values between groups or categories. Each group is assigned a bar; each bar's value is represented in its height above (or below) a zero-baseline.



Grouped bar chart ● +Relationships

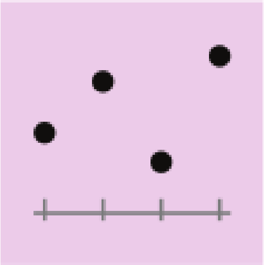
Extends a bar chart to compare data across two categorical variables. Each bar corresponds to an intersection of variable levels: categories for one variable are indicated by the bar cluster positions, while the second variable is indicated by bar color or position within each cluster.

Comparing values between groups



Lollipop chart

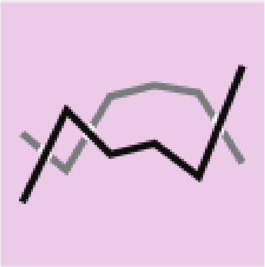
Replaces the bars of a bar chart with lines and dots. Useful for when there are a lot of groups or categories to plot.



Dot plot

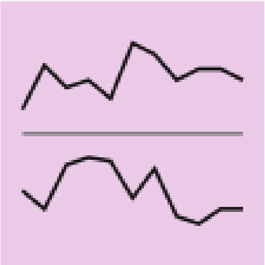
Replaces the bars of a bar chart with just dots. Since value is indicated by position instead of length, the dot plot can be good when a zero baseline is not useful.

Comparing values between groups



Line chart ● +Change over time

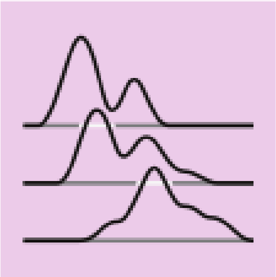
Each line in a line chart shows how values (vertical position) change across time (horizontal). One line is plotted for each group to be compared. Best when there are five or fewer groups to plot.



Sparkline ● +Change over time

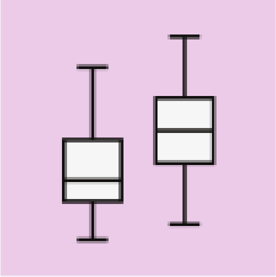
Smaller line charts typically with little to no labeling. Designed to show a high-level overview inline with text or tables, but also useful when there are many groups to plot.

Comparing values between groups



Ridgeline 🟦

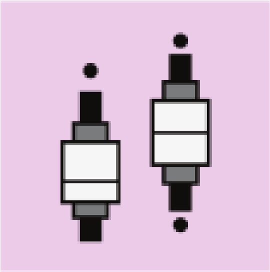
A series of line charts or density curves (see *Distributions*) with partially offset axes used to compare distributions between groups. Best when there are distinct patterns across groups.



Box plot 🟢 *+Change over time* *+Distributions*

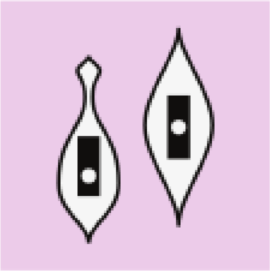
Compares a statistical summary of numeric values between groups. A set of box and whiskers depicting the range of the most common data values (see *Distributions*) is assigned to each group or category.

Comparing values between groups



Letter-value plot ◆ *+Distributions*

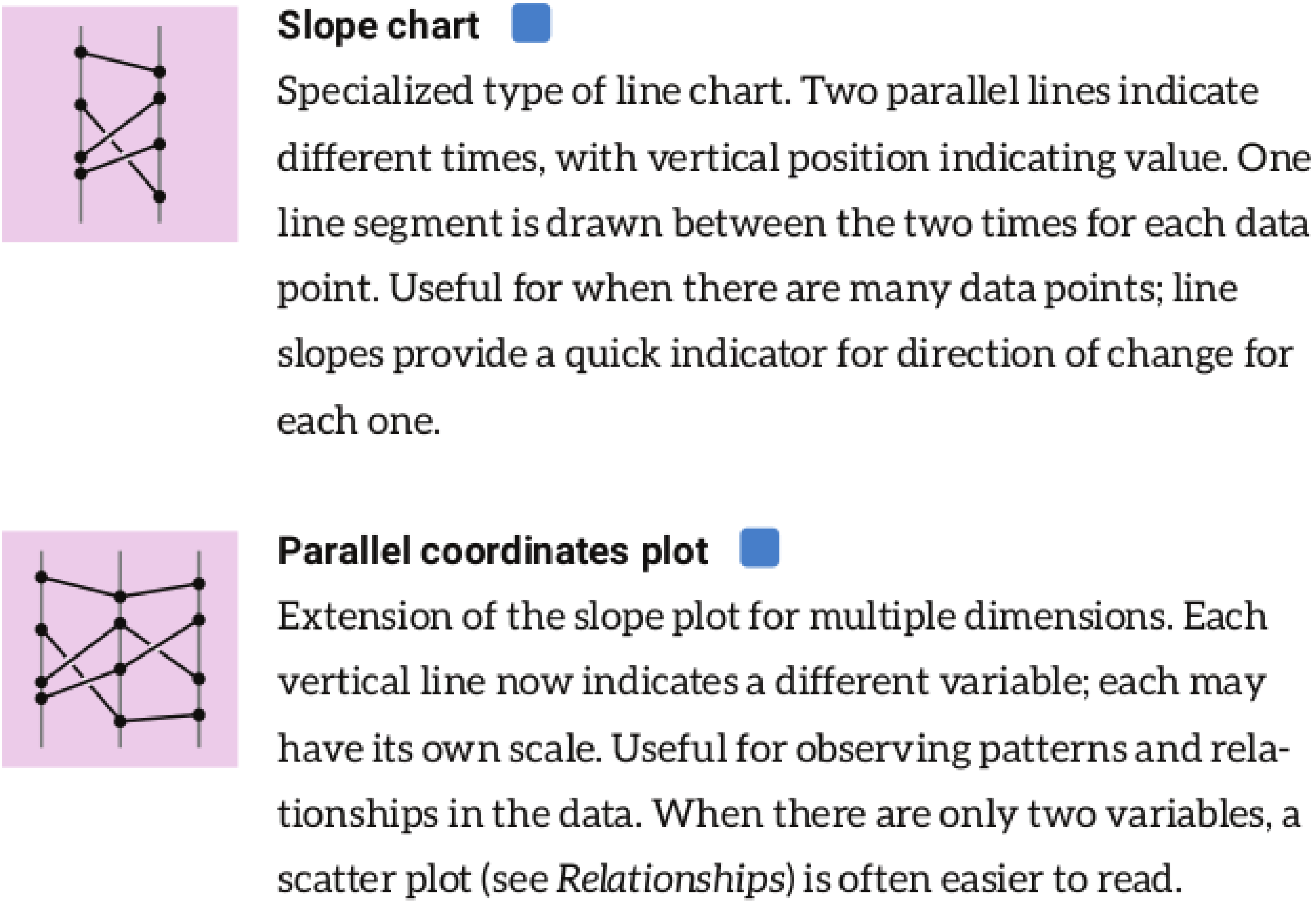
Used in a similar way as the box plot, but a letter-value plot (see *Distributions*) is assigned to each group instead. Best used when there are lots of data in each group so that statistical estimates are stable.



Violin plot ● *+Distributions*

Compares distributions between groups. A violin assembly of density curve and box plot (see *Distributions*) is assigned to each group or category.

Comparing values between groups



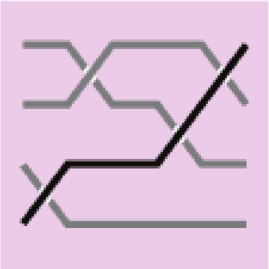
Slope chart

Specialized type of line chart. Two parallel lines indicate different times, with vertical position indicating value. One line segment is drawn between the two times for each data point. Useful for when there are many data points; line slopes provide a quick indicator for direction of change for each one.

Parallel coordinates plot

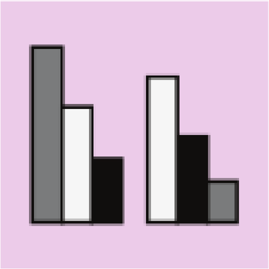
Extension of the slope plot for multiple dimensions. Each vertical line now indicates a different variable; each may have its own scale. Useful for observing patterns and relationships in the data. When there are only two variables, a scatter plot (see *Relationships*) is often easier to read.

Comparing values between groups



Bump chart

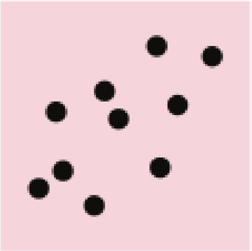
Modified version of a line chart where vertical position corresponds to rank rather than value. This change allows it to support more categories than a standard line chart.



Grouped bar chart

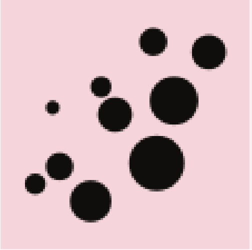
Normally, grouped bar charts will plot the bars within each group in a consistent order. However, they can instead be sorted by value within each group to emphasize ranking, at the cost of making it more difficult to find each sub-category.

Relationships between variables



Scatter plot

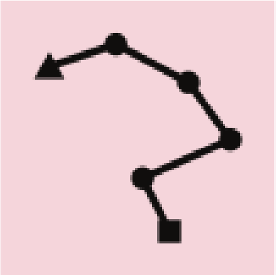
Standard chart type for showing relationships between two numeric variables. Each point's position on the horizontal and vertical axes indicate value on the associated variable.



Bubble chart

Scatter plot with point size dictated by a third numeric variable. Scatter plots can be extended in other ways: point shapes can encode a categorical variable, and color can be used to indicate either categorical or numeric data. It is best to keep a scatter plot to a maximum of three variables to maintain understandability.

Relationships between variables



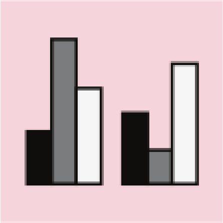
Connected scatter plot ◆

When a third variable represents time, points in a scatter plot can be connected with line segments to show progression in values across time.

Dual-axis bar-line plot ◆

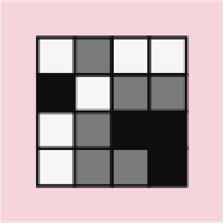
A bar-line plot shares a horizontal axis (typically time) across two chart types: the bar chart and line chart. Useful for when the variables plotted with each chart type are related, but are on different numeric scales.

Relationships between variables



Grouped bar chart ● +*Comparisons*

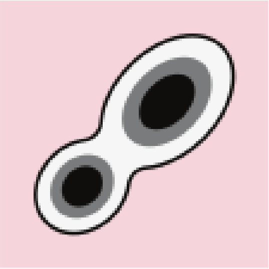
Extension of bar chart (see *Comparisons* or *Distributions*) to two categorical variables. Bar clusters are associated with levels of one variable, while color or position in each cluster indicates levels of the second variable. The length of each bar at the corresponding intersection of levels indicates a value for that group, like data frequency or a summary of a third numeric variable.



Heatmap ●

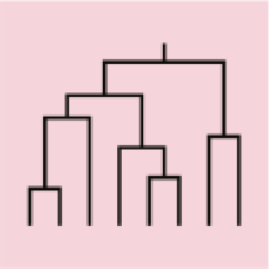
Extension of bar charts and histograms (see *Distributions*) to two variables, each of which can be categorical or numeric. Each axis represents groups or bins of values for one of the variables, forming a grid. Cell colors indicate data frequency or a summary of a third variable for each intersection of axis variables.

Relationships between variables



2-d density curve ◆

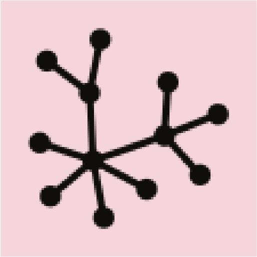
Extension of density curves (see *Distributions*) to two numeric variables. Colors are mapped to values like in a heatmap, but applied smoothly across the plotted area rather than in discrete bins. Somewhat confusingly, this chart is sometimes also known as a heatmap.



Dendrogram ◆

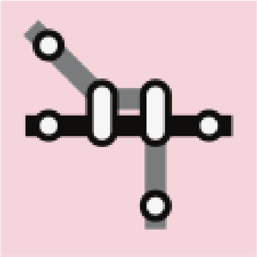
Specialized chart type to show similarity between data points. The lower the branch connecting two data points is, the more similar they are. Sometimes plotted with an accompanying heatmap to depict the underlying data.

Relationships between variables



Network diagram ◆

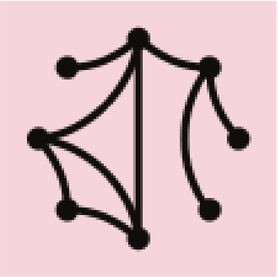
Points (nodes or vertices) represent individual entities. Lines (edges) connect entities with a particular relationship. Line thickness may be used to encode value. Vertex positions do not necessarily have any inherent meaning, and may simply be placed just to make connections as clear as possible.



Transit map ◆

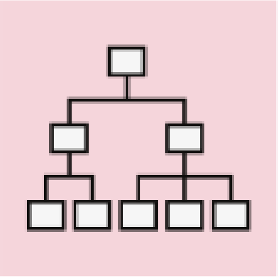
Practical application of network diagrams for train and subway systems. Frequently, these take a fair level of abstraction, emphasizing connections between stations rather than their actual geographical locations.

Relationships between variables



Chord diagram ◆

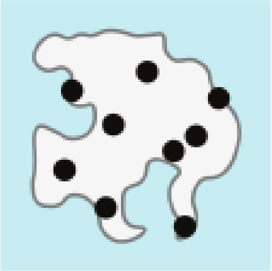
Like a standard network diagram, but vertices are arranged in a circle.



Tree diagram ◆

A network diagram organized to show hierarchical relationships. The direction of each edge corresponds to a relationship between the connected nodes, such as parent-child or senior-junior relationships.

Geographical data



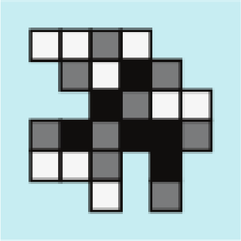
Scatter map 🟢

Scatter plot built on top of a geographical map, using geographic coordinates as point positions.

Bubble map 🟢

Bubble chart built on top of a geographic map, where point size is an indicator of value. Can also be used to group together points in a scatter map if they are too dense.

Geographical data

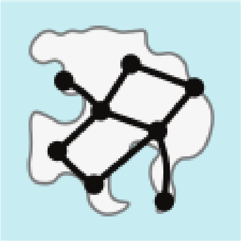


2-d histogram ●

Heatmaps can be built on top of geographic areas. Sometimes seen with a hexagon-shaped grid rather than a rectangular grid. May distort the geography on its edges.

Isopleth / contour map ◆

2-d density curve built on top of a geographic map.



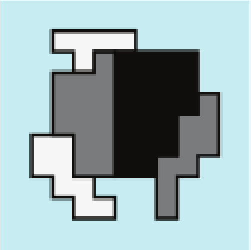
Connection map ◆

Network information and flows built on top of a geographic map.

Geographical data

Choropleth ●

Similar to a heatmap, but colors are assigned to geopolitical regions rather than an arbitrary grid. Values are often in the form of rates or ratios to avoid distortion due to population density.



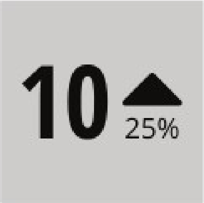
Cartogram ◆

Geopolitical regions sized by value. This necessarily requires distortion in shapes and topology.

Raw Numbers

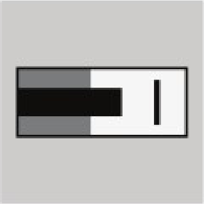
Single Value Chart

Show a **raw singular value**



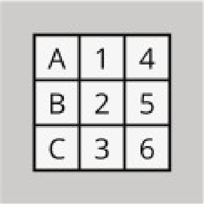
Single Value w/ Indicator

Comparison of a **single value** against a **previous value**



Bullet Chart

Comparison of a **single value** against a **benchmark value**



A	1	4
B	2	5
C	3	6

Table

Show raw values for **multiple data points** on **multiple variables**

Data Visualization Tools

- Tableau
- Infogram
- ChartBlocks
- D3.js
- Google Charts
- Fusion Charts
- Chart.js

Tableau

- Tableau has a variety of options available, including a desktop app, server and hosted online versions, and a free public option.
- There are hundreds of data import options available, from CSV files to Google Ads and Analytics data to Salesforce data.
- Output options include multiple chart formats as well as mapping capability. That means designers can create color-coded maps that showcase geographically important data in a format that’s much easier to digest than a table or chart could ever be.

Infogram

- Infogram is a fully-featured drag-and-drop visualization tool that allows even non-designers to create effective visualizations of data for marketing reports, infographics, social media posts, maps, dashboards, and more.
- Finished visualizations can be exported into a number of formats: .PNG, .JPG, .GIF, .PDF, and .HTML. Interactive visualizations are also possible, perfect for embedding into websites or apps.
- Infogram also offers a WordPress plugin that makes embedding visualizations even easier for WordPress users.

ChartBlocks

- ChartBlocks claims that data can be imported from “anywhere” using their API, including from live feeds. While they say that importing data from any source can be done in “just a few clicks,” it’s bound to be more complex than other apps that have automated modules or extensions for specific data sources.
- The app allows for extensive customization of the final visualization created, and the chart building wizard helps users pick exactly the right data for their charts before importing the data.
- Designers can create virtually any kind of chart, and the output is responsive—a big advantage for data visualization designers who want to embed charts into websites that are likely to be viewed on a variety of devices.

D3.js

- D3.js is a JavaScript library for manipulating documents using data.
- D3.js requires at least some JS knowledge, though there are apps out there that allow non-programming users to utilize the library.
- Those apps include NVD3, which offers reusable charts for D3.js; Plotly’s Chart Studio, which also allows designers to create WebGL and other charts; and Ember Charts, which also uses the Ember.js framework.

Google Charts

- Google Charts is a powerful, free data visualization tool that is specifically for creating interactive charts for embedding online.
- It works with dynamic data and the outputs are based purely on HTML5 and SVG, so they work in browsers without the use of additional plugins. Data sources include Google Spreadsheets, Google Fusion Tables, Salesforce, and other SQL databases.
- There are a variety of chart types, including maps, scatter charts, column and bar charts, histograms, area charts, pie charts, treemaps, timelines, gauges, and many others. These charts can be customized completely, via simple CSS editing.

FusionCharts

- FusionCharts is another JavaScript-based option for creating web and mobile dashboards. It includes over 150 chart types and 1,000 map types.
- It can integrate with popular JS frameworks (including React, jQuery, React, Ember, and Angular) as well as with server-side programming languages (including PHP, Java, Django, and Ruby on Rails).
- FusionCharts gives ready-to-use code for all of the chart and map variations, making it easier to embed in websites even for those designers with limited programming knowledge.

Chart.js

- Chart.js is a simple but flexible JavaScript charting library. It’s open source, provides a good variety of chart types (eight total), and allows for animation and interaction.
- Chart.js uses HTML5 Canvas for output, so it renders charts well across all modern browsers. Charts created are also responsive, so it’s great for creating visualizations that are mobile-friendly.

Visualization using Programming

- Python
  - matplotlib
  - seaborn
  - plotly
  - pylab
- R
  - graphics
  - ggplot2

Best Resources to Learn

- https://python-graph-gallery.com
- https://www.r-graph-gallery.com
- https://chartio.com

Thank you

@mitu_skillologies

@mITuSkillologies

@mitu_group

@mitu-skillologies

@MITUSkillologies

kaggle

@mituskillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

@mituskillologies

contact@mitu.co.in

tushar@tusharkute.com