Data Visualization: Best practices for Creating Visuals

Importance of data visualization

Allows you to represent data in a manner that’s interpretable by giving it a visual context.

Easier to identify trends, patters, outliers & anomalies.

A picture is worth a thousand words

Best practices for data visualization

Identify your target audience

Resist the temptation to cater to everyone. Something for everyone is for no one.

What decision do they expect to make after viewing the presentation?

Clean your data – identify inaccuracies like missing values, outliers or malformed data

Ensure data isn’t skewed & computations are accurate

Select the right visualization type

Provide context & label your chart effectively – use captions, provide value labels & do not depend on chart element sizes are people often misinterpret those

Keep visualizations & dashboards simple

Use color effectively but don’t overwhelm them

Use intuitive colors – green for positives, red for negatives

Explicitly clarify what different colors mean & what they represent

Use grouping to simplify visualization

Make your data readable in any format – small, large devices, etc.

Use contrast in layouts so elements are clearly visible

Presentation types

Presentation types can be compared into 5 broad categories:

Comparison – comparison of values across categories

Distribution – what is the normal tendency & distribution

Relationship – how does 1 variable relate to others

Composition - how do individual parts make up a whole

Trends – how did values perform over a certain time period

Points to consider before choosing a visualization:

How many variables do you want to display in a single chart?

How big is your data? How many data points will you display for each variable?

Will you display values over a period of time, or in groups?

Comparison charts:

Bar charts/Column charts

Line charts

Pie charts

Composition charts

Pie charts – have fallen out of favor because it’s hard to view proportions of circular shapes using angles

Stacked bar charts

Area charts

Tree maps

Distribution charts

Histogram

Line chart

Scatter plot

Relationship charts

Scatter charts

Bubble charts

Trend charts

Line charts

Circular area charts

Tables – Use cases

Usually represent source data used for analysis & visualization

If the size of the data is really small, we might choose to use the table itself for visualization

Can be used for comparison, composition, relationship analysis

Used when there are very few variables in input data, for large sizes tables are overwhelming

Use tables when:

You need to drill down into actual, precise values in your data

You need to visualize values that involve multiple units of measurements (revenue, count, etc.)

You are interested in quantitative information but not distribution or trends

Do not use tables when:

The shape of the data conveys important information

You want to show the general form of a relationship between multiple variables

Column charts a.k.a bar charts

Used to compare values across categories or changes in a value over a period of time for a single category

Users are interested in the actual, precise value for each category

Compare heights of the bars for relative measures

Assign individual labels to each bar in the chart

Associate each quantitative value to each bar instead of expecting users to figure it out from size

Grouped bar charts

Multiple sets of data items are compared – each set belongs to a certain category

Usually use a single color to represent the same sub-category across all sets

Use bar charts when:

Number of categories you’re comparing is quite small

Always start the numerical axis at 0 so bars are not truncated. If you don’t start it from 0, this is considered as misrepresentation of data as the scale will be wrong.

Label individual bars to improve readability

Use solid colors for bars.

Don’t use patters on bars like polka dots, lines, etc.

Line charts

Used for displaying trends over time

Used with a continuous dataset over a long period of time where the dataset has many data points

Emphasis on the flow or continuous nature of the data visualized

Can be used for comparisons as an alternative to column charts where there are just a few data points

Best practices

Use solid lines

Don’t plot more than 4-5 lines per chart

Time axis should always run from left to right

Time intervals should be of equal size

The axis range should allow all line values to be displayed i.e, the line shouldn’t go out of the chart

Pie charts

Have fallen out of favor but are still useful to see how individual proportions make up a whole.

Typically represent data in form of percentages rather than absolute values

Usually not used to compare precise values, rather than comparing proportions

Now, Bar charts are preferable for comparison purposes & tree maps for composition purposes

Best practices

Make sure that all sections of the pie add up to 100%

Use only when you have a few categories i.e, not more than 4-5 sections of the pie

Make sure the category values are not identical or very close. E.g: 4 sections where each is 25% of the pie

Histograms

Used to represent univariate distribution of data

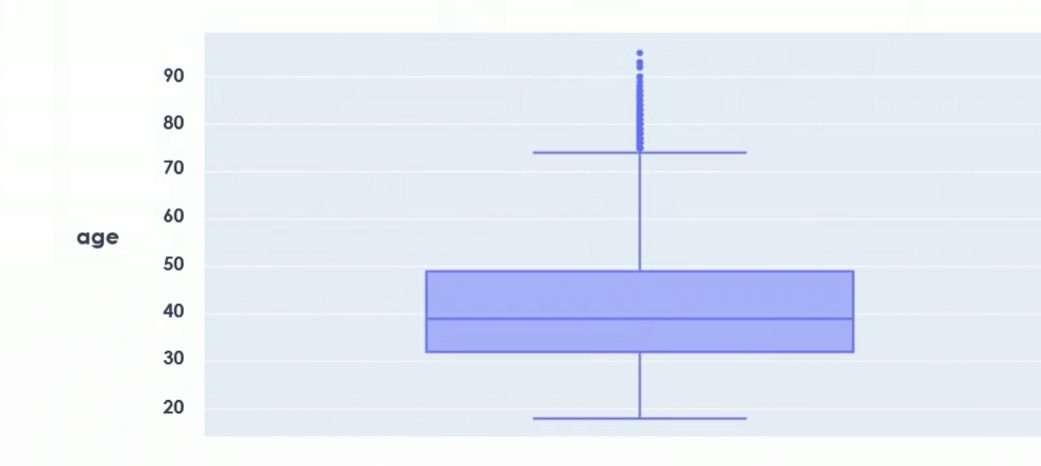
Shows the relationship of a single variable over multiple categories or buckets

Displays how often the different categories occur in a set of data points

Histogram are based on the area of the bar, not the height of the bar. Don’t confuse with bar charts

Box plots a.k.a box & whisker plot

Displays summaries statistics of the data such as median, inter-quartile reach etc.



Center line of the box -> median age = 37 yrs

Size of the box – Interquartile reach

Lower edge of the box -> value at 25th percentile of data

Upper edge of the box -> value at 75th percentile of data

Whiskers – the lines drawn outside the box – usually drawn to 1.5 times the interquartile reach. Useful to identify any outliers in the data set.

Outliers are represented using individual data points drawn above or below the whiskers. They quickly help us identify the inaccuracies in the data.

Box plots can be used to represent distribution of data across multiple categories as well.



Used for explanatory data analysis, to quickly show the center & spread of a distribution.

Provide a statistical summary of the data – mean, median, interquartile range.

Help identify the outliers in the data that are either very large or very small.

Scatter charts

Used to represent the relationship between 2 variables

Allow us to interpret the functional form of the relationship i.e., linear, exponential, non-linear, independent

Represent data distribution or clusters that may be present in the data

Help identify outlier points or anomalies in the data

Bubble chart

It’s a basically a scatter plot with additional information.

The size of the bubble convey information about a 3rd dimension

It represents yet another dimension to the data using the color of the bubbles

Best practices

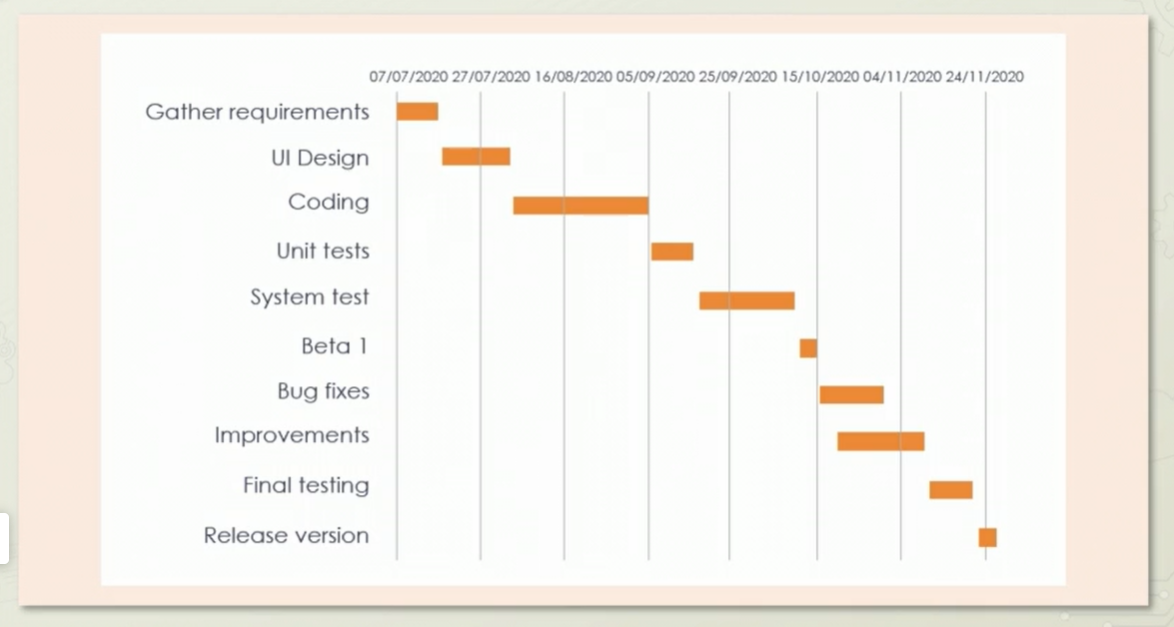
Relationship should be clearly interpretable

Use size or color to incorporate additional dimensions

Start axes at 0 to represent data accurately

For bubble charts, represent the dimension using area & not diameter of bubbles

Gnatt charts



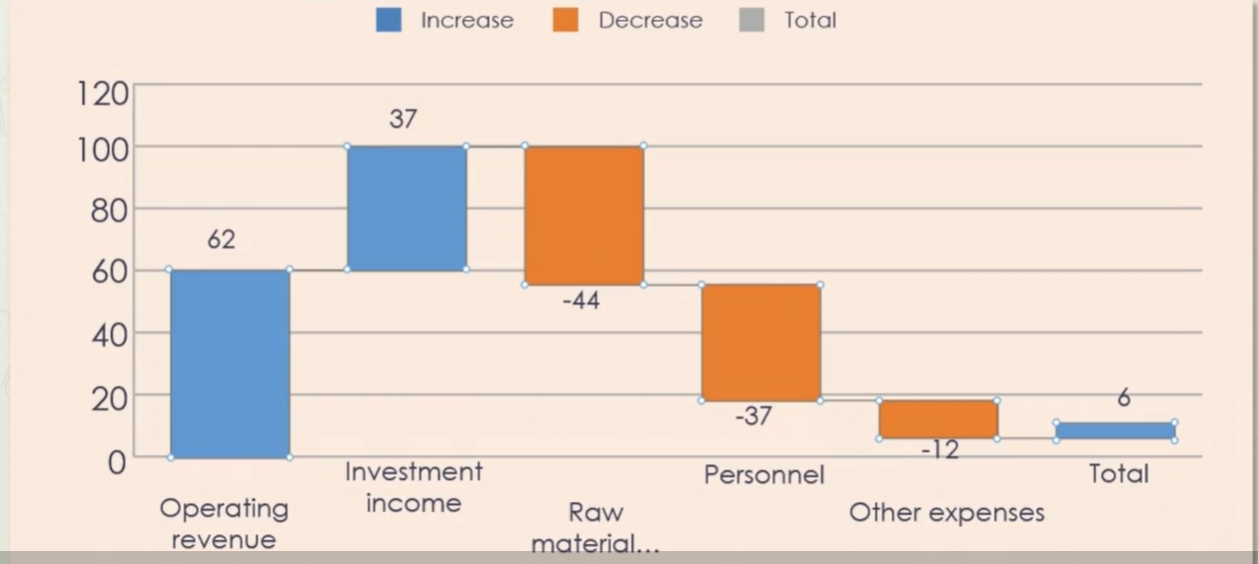
The left side lists out all the steps involved in the completion of the project. Some of these steps might overlap in time, others might not. Along the horizontal axis we list out the date. The contents of the Gantt chart are bars that track how long it will take to complete a particular step in the project.

Used for planning & scheduling projects

Track the work that needs to be done & decide deadline

Can be used to track the percentage completion of tasks, resources required, dependencies, etc.

Waterfall charts



We display all positive accumulations towards final result in 1 color & negative accumulations in another.

The revenue bars are projected upwards as they add up.

The cost bars are projected downwards as they reduce the result.

Doing this for each bar will eventually lead us to the total or result.

Used to reveal the composition of a numeric value or the accumulation towards a final value

Represent how some initial value is updated by intermediate values to give us a final result

Most common use case is for displaying revenues, costs & then a final profit number for a company

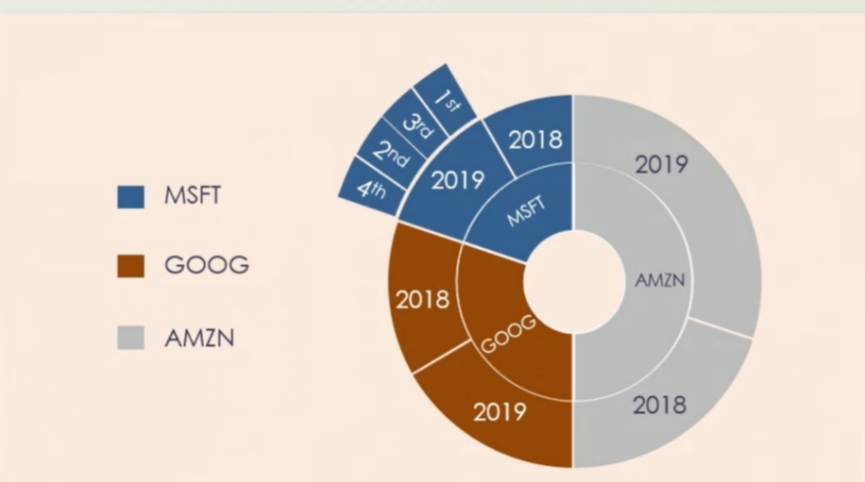
Sunburst chart

Used to represent hierarchical information. Every concentric circle represents different information. The innermost circle represents the company, The 2nd concentric circle correspond to the year where the revenues are computed & the segment correspond to their total revenues. The third circle corresponds to the quarter of the year.

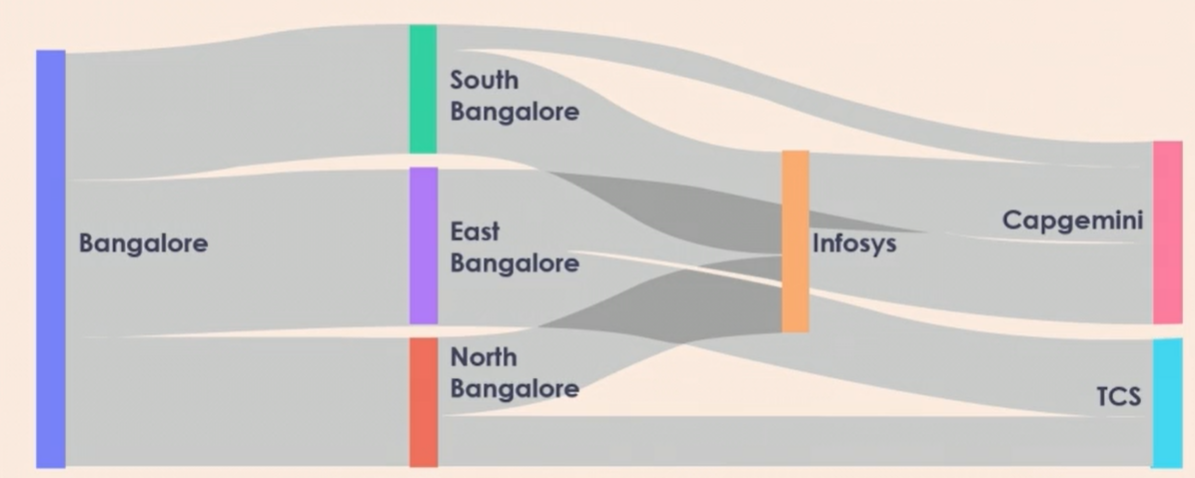
Used for displaying hierarchical data where the hierarchy rather than the actual values are important

Every level of the hierarchy is one ring of the circle

Outer rings represent lower levels of the hierarchy



Sankey charts



Used to visualize how values flow between different nodes.

In the above diagram, each node represents a location. The left-most nodes represent whole of Bangalore. The intermediate nodes represent the sections of Bangalore & final nodes represent location of headquarters of various companies in Bangalore. The flow represents how many people travel to the companies & from which parts of Bangalore.

Visualization of material, cost or energy flows

Directed flow is always between 2 nodes in the chart

The value of the flow is denoted by the thickness of the link between 2 nodes.