Anscombe's Quartet

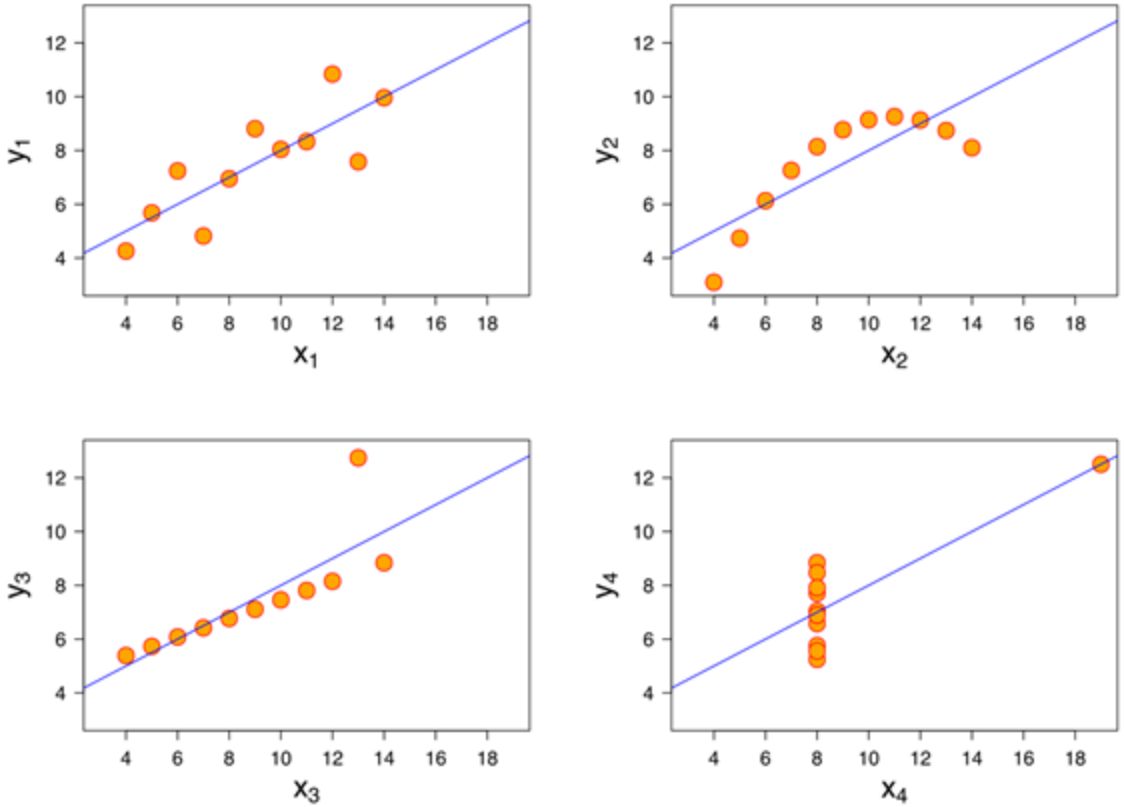
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
Each dataset consists of eleven (*x*,*y*) points.

Activity 1: Choose a data set and plot it as you like in 4 different ways

Activity 2: Make a scatterplot for each data set



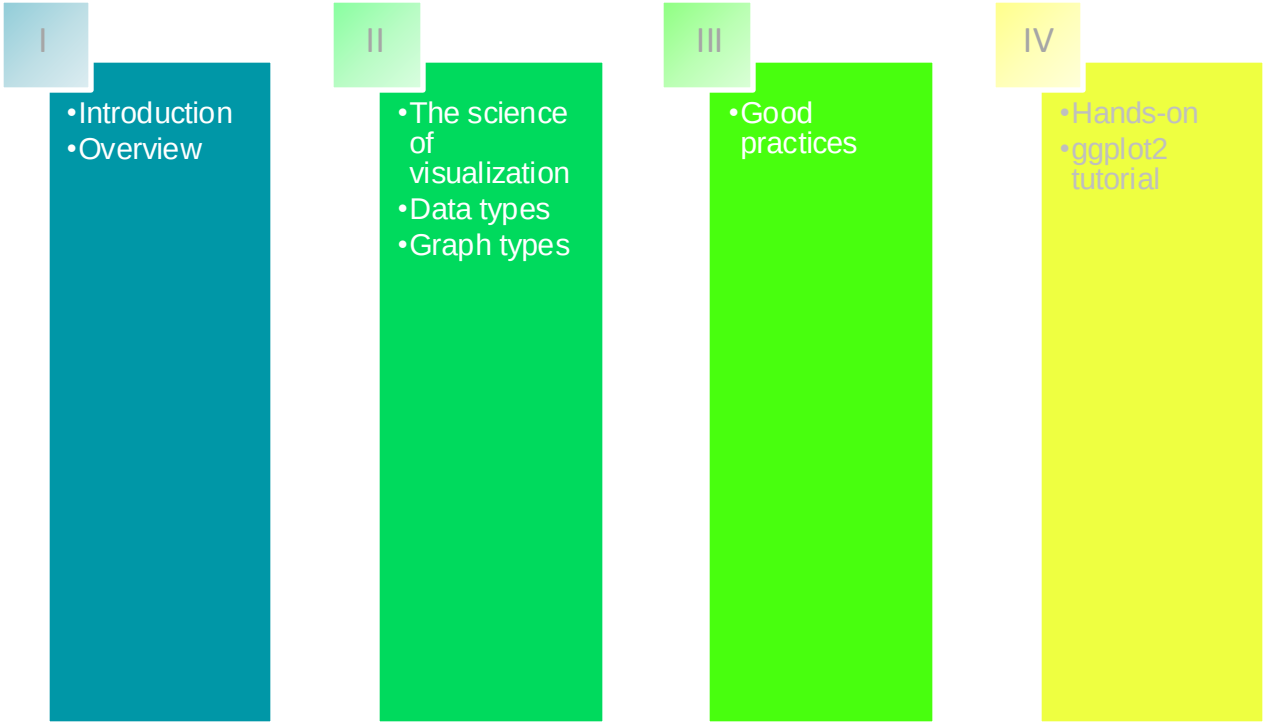
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Data Schools

Data Visualisation

Hugh Shanahan Hugh.Shanahan@rhul.ac.uk
Materials created by Dr. Sara El jadid<s.eljadid@qub.ac.uk>

Course Overview

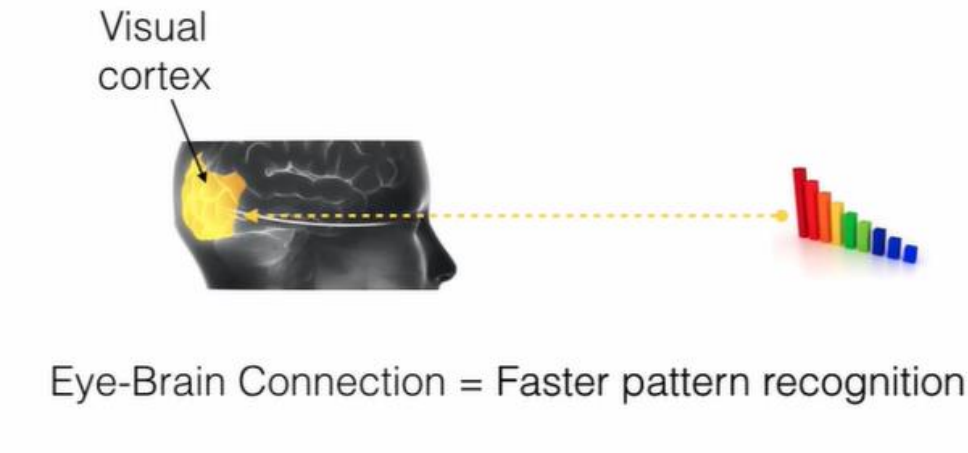


What is Data Visualisation

- Graphic representation of data that aims to communicate numerous **heavy data** in an efficient way that is **easier to grasp and understand**.
- Data visualization describes **any effort to help** people understand the **significance** of data by placing it in a **visual context**.
- A branch of **Descriptive statistics** but it requires both **design**, **computer**, and **statistical skills**.
- Aesthetics and functionality go hand in hand to communicate complex statistics in an **intuitive way**.
- **Data Viz tools and technologies** are essential for making **data-driven decisions**.

How does it work?

- Humans are **visual creatures**, it's a fact that visualization is much more effective and comprehensive.
- **Patterns**, **trends** and **correlations** that might go **undetected** in text-based data can be exposed and **recognized** easier with data visualization software.



The sense of sight is the most dominant

The sense of sight is the most dominant

90%

of information transmitted to the brain is *visual*

(sources: 3MCorporation and Zabisco)

40%

of people will respond better to *visual information* than to plain text

(source: Zabisco)

96.7%

believe than visual content engages best on social media

(source: Newburyport Mass)

Why data visualization is important?

Data Viz is a form of storytelling with the purpose to help us make decisions based on data.

- Tracking sales
- Identifying trends
- Identifying changes
- Monitoring goals
- Monitoring results
- Combining data

When to use it?

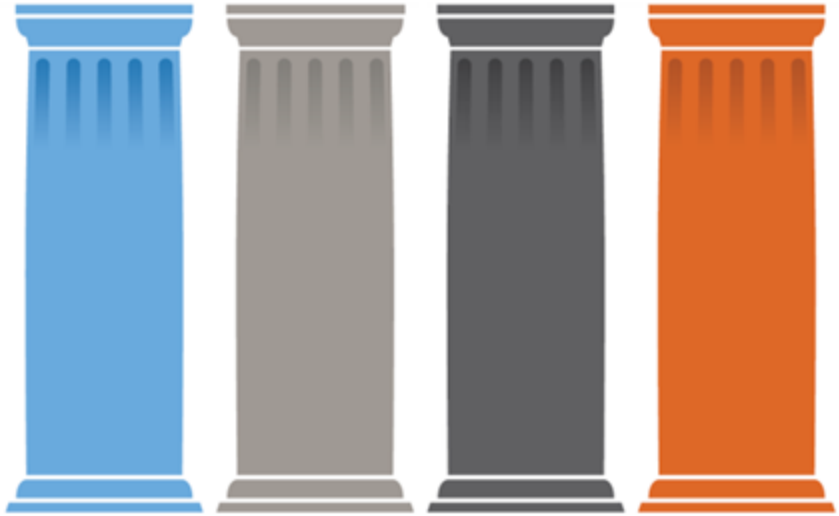
Data Viz is suitable for:

- Annual reports
- Presentations
- Social media micronarratives
- Informational brochures
- Research
- Trend-trafficking
- SciViz
- Candlestick chart for financial analysis
- Determining routes

Why Use it?

- Identifying correlations between the relationship of variables.
- Getting market insights about audience behavior.
- Determining value vs risk metrics.
- Monitoring trends over time.
- Examining rates and potential through frequency.
- Ability to react to changes.

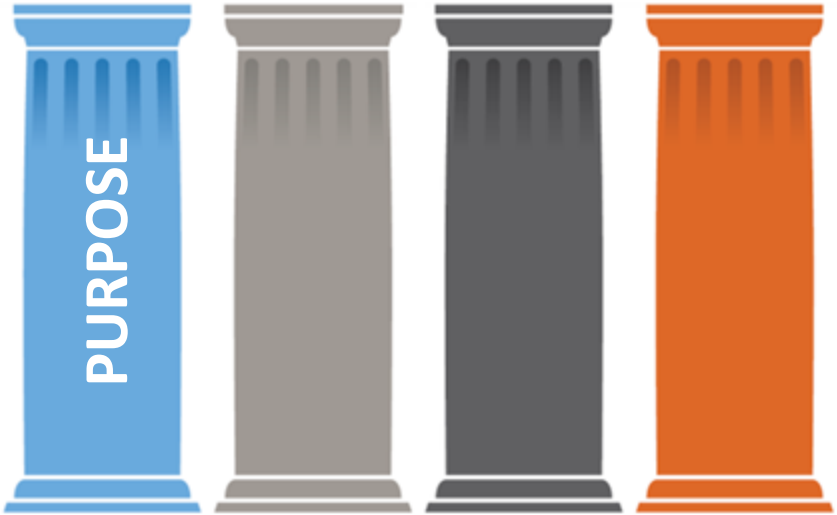
The four pillars of data visualization



1. Purpose - why this visualization

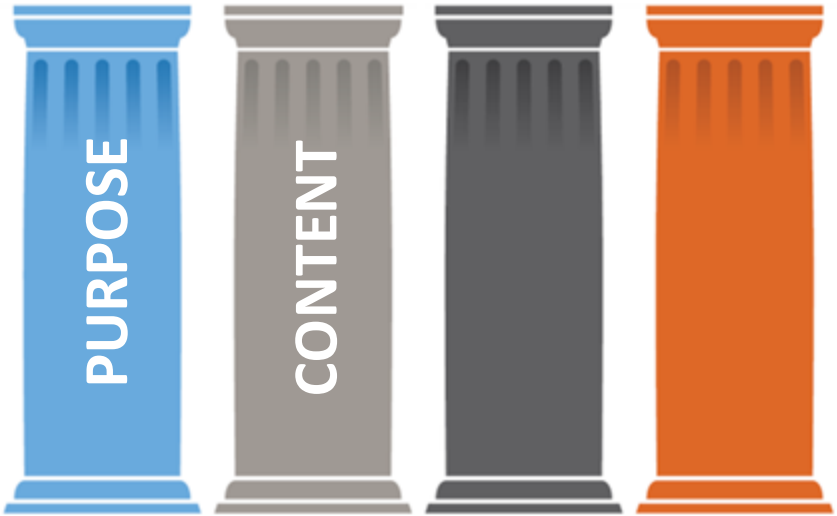
For the creators:

- **Why** am I doing this visualization?
- Who is it for?
- What do they need to understand?
- What actions do you need to enable?
- How it will be consumed?
- What is the most important takeaway message?



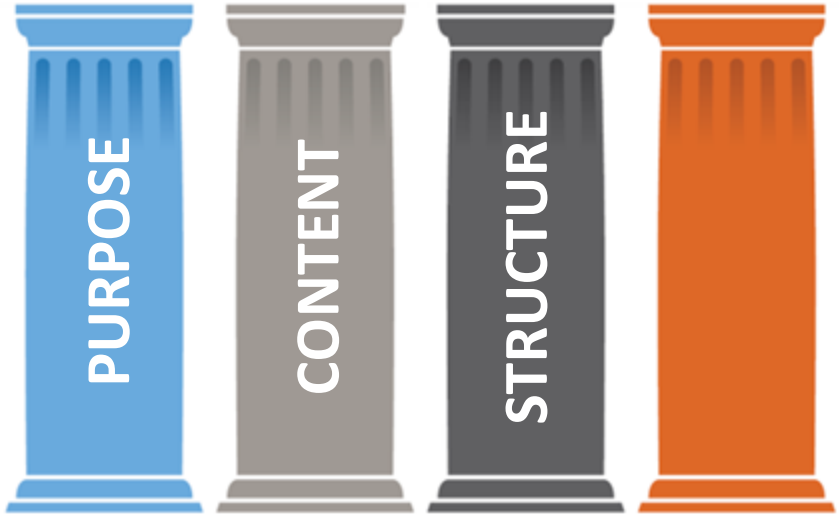
2. Content - what to visualize

- **What** data matters?
- What **relationships** (in the data) matter?
- Informed by the purpose!
- What's *excluded* is as important as what's *included*



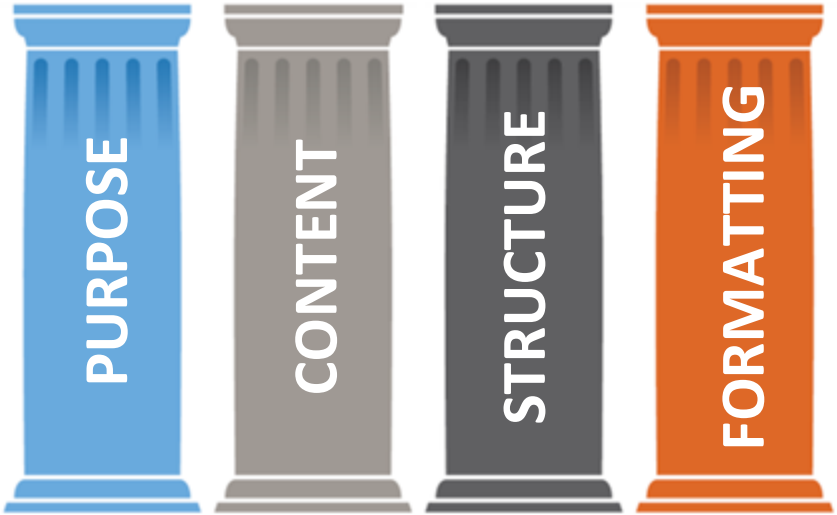
3. Structure - how to visualize it

- **How** can the most important data and relationships be revealed the best?
- **Choose** meaningful layout and axes!
- Use both axes (both, not three..)
- Informed by purpose and content!



4. Formatting - how to make it appealing

- **How** it should look and feel?
- How will it be consumed?
- Makes data and relationships accessible
- Makes importance visible
- Informed by purpose, content and structure!



Guide to data type

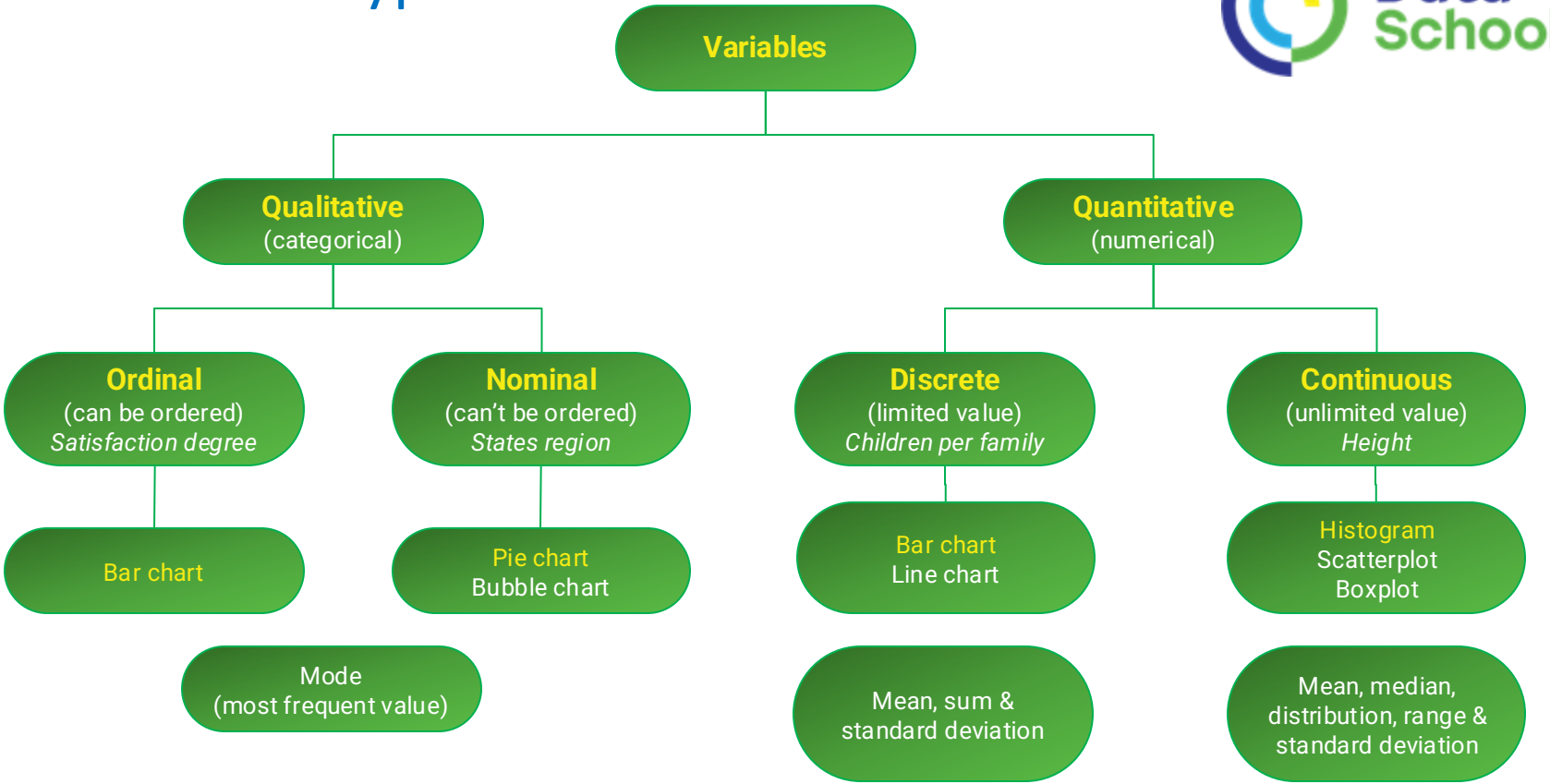


Chart Types

Scatterplot 

Line Chart 

Bar Chart 

Bubble Chart 

Area Chart 

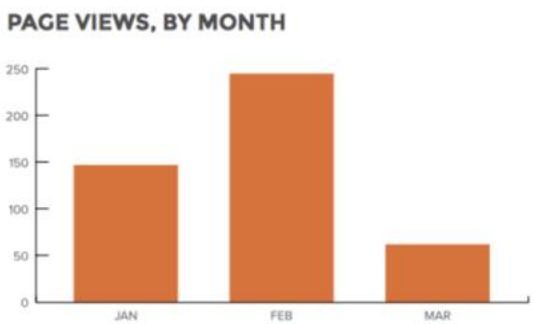
Pie Chart 

Heat Map 

Bar chart

Bar charts are very versatile. They are best used to show change over time, compare different categories, or compare parts of a whole

Common Bar chart variations include Stacked, 100% stacked versions. Usually these variations are used to compare multiple part-to-whole relationships. i.e Monthly online traffic analysis by different sources.



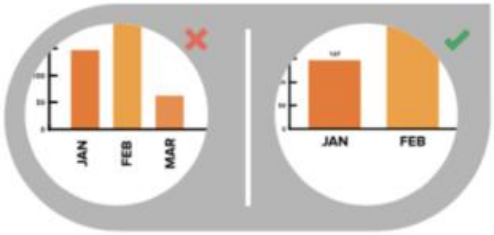
VERTICAL
(Column Chart)

It is best used for chronological data (time-series should always run left to right) or when visualizing negative values below the axes.

HORIZONTAL

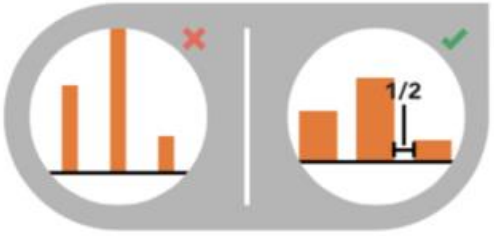
It is best used when data with long categories are to be labelled

Bar chart best practices



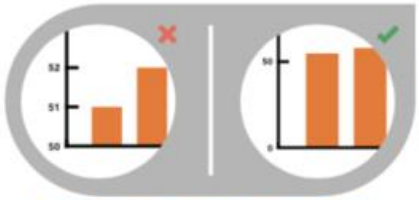
Use Horizontal Labels

Avoid steep diagonal or vertical type, as it can be difficult to read



Space Bars Appropriately

Space between the bars should be at least ½ bar width



Start the y-axis value at Zero

Starting at a value above zero truncates the bars and doesn't accurately reflect the full value.

Use Consistent Colors

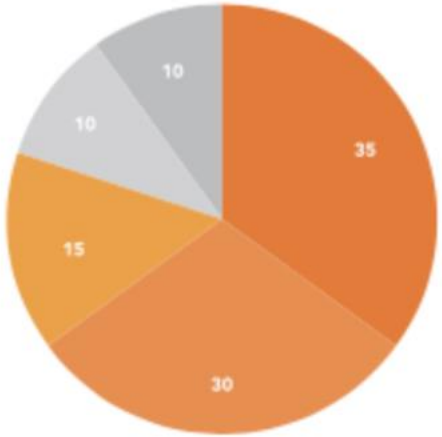
Use one color for bar charts. You may use an accent color to highlight a significant data point.

Order Data Appropriately

Order the categories alphabetically, sequentially or by the values.

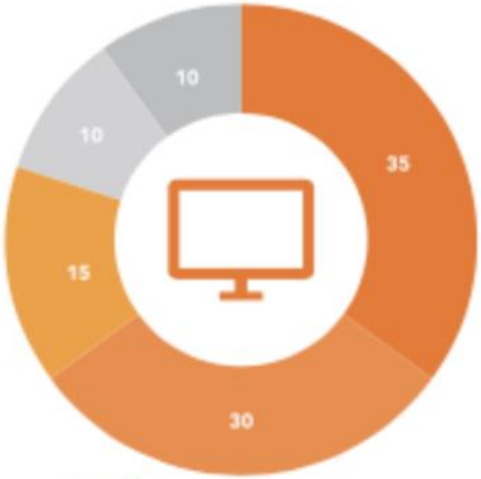
Pie chart

Pie charts are best used for making portion to whole comparisons with discrete or continuous data. They are most impactful with a small data set.



STANDARD

It is used to show part-to-whole relationships.



DONUT

A stylistic variation of the original pie chart with an inclusion of a total value or design element in the center.

Pie chart best practices

Visualize no more than 5 Categories per Chart

It is difficult to differentiate between the small values; depicting to many slices makes it complex and decreases the visualization impact. If needed, multiple small slices may be categorized as “Miscellaneous” or “Other”

Don’t use Multiple Pie charts for Comparison

Sliced sizes are very complex to compare side by side. Hence, if required; use a stacked bar chart instead.

Total Data Count must be 100%

Make sure that total values sum up to 100% and that pie slices are sized proportionate to their corresponding value

Order the slices Correctly

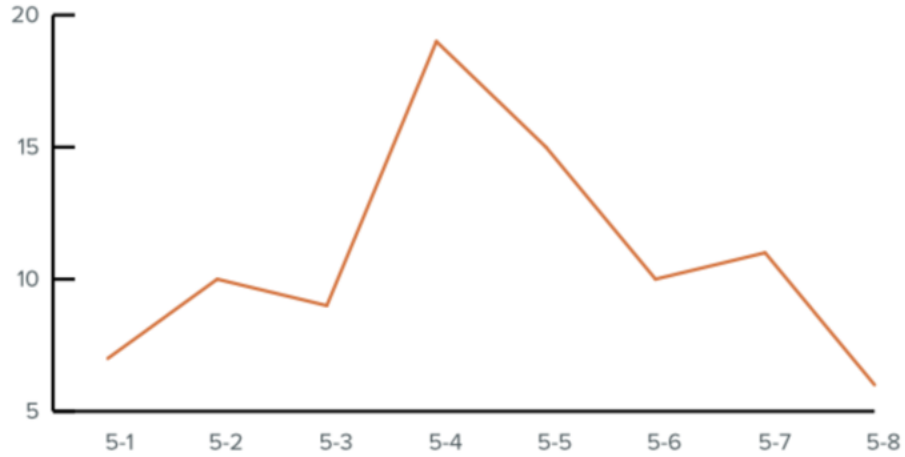
Option-1: Place the largest section at 12 o’clock going clockwise and second largest at 12 o’clock counterclockwise.

Option-2: Place the largest section at 12 o’clock going clockwise. Place remaining sections in the descending order, going clockwise.

Line chart

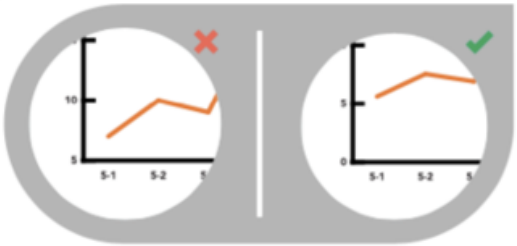
Line charts are used to show time-series relationships with continuous data. They help show trend, acceleration, deceleration, and volatility.

Line chart itself doesn't offer any variations. It may be used to track or identify changing trends in bar chart but it itself doesn't have any variants.



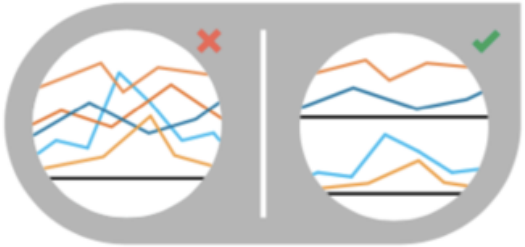
Direct Marketing Views, By Date

Line chart best practices



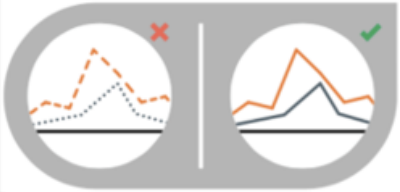
Inclusion of Zero Baseline

Although a Line chart doesn't have to start with a 0 value; it should be included whenever possible.



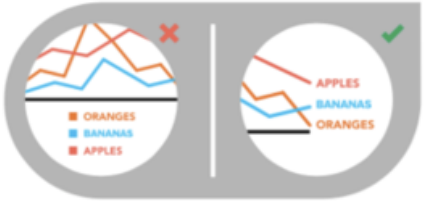
Don't plot more than 4 lines

If you need to display more than 4 lines, break them into separate charts for better comparison



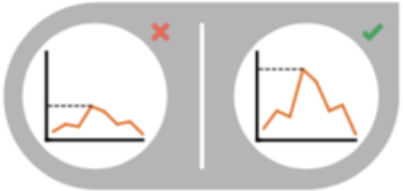
Solid Lines ONLY

Use of dashed and dotted lines can be distracting



Label Directly

This allows readers quickly identify lines.

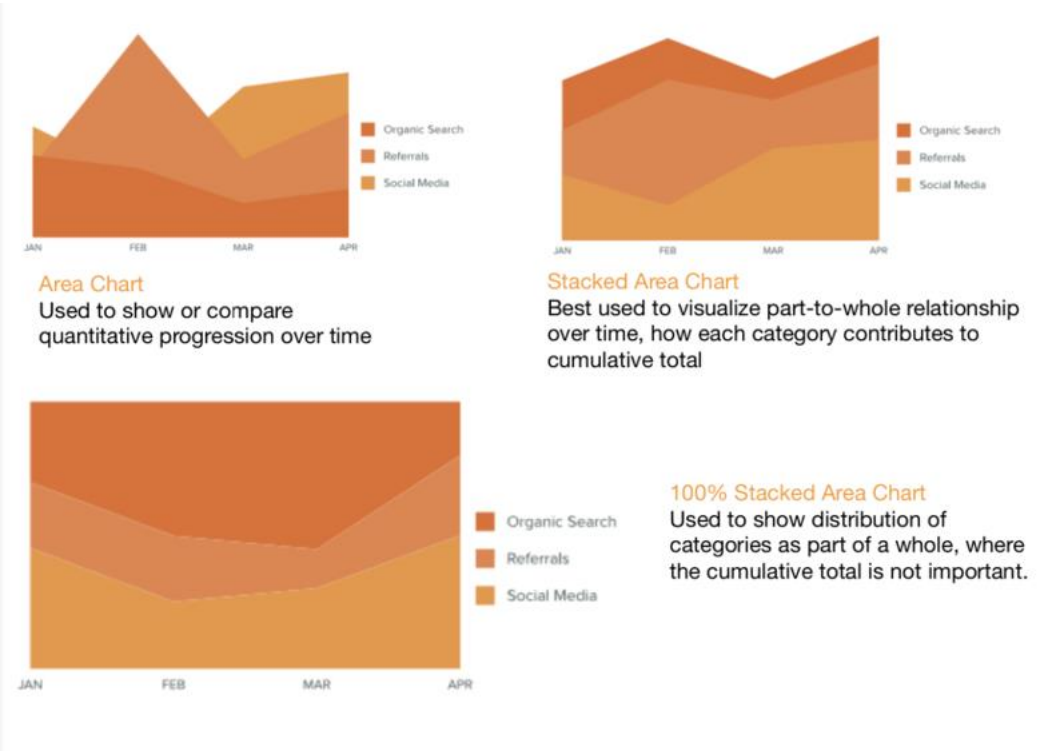


Use the right Height

Plot all lines so that the line chart takes approximately two-thirds of the y-axis's total scale.

Area chart

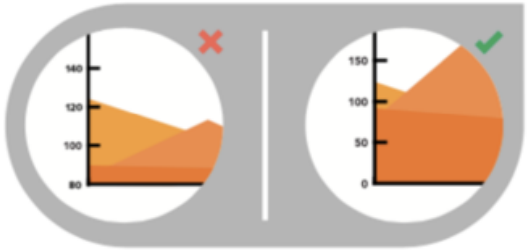
Area charts depict a time-series relationship, but they are different than line charts in that they can represent volume.



Area chart best practices

It should be easy to read

In stacked area charts, arrange data to position categories with highly variable data on the top of chart and low variability on the bottom.



Start y-axis value at 0

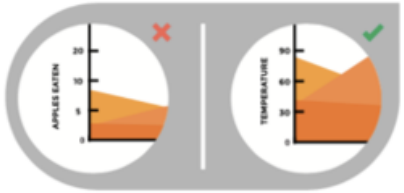
Starting above zero truncates the visualization of values.

Don't display more than 4 categories

It will result in a complex cluster visual

Use Transparent Colors

Use of transparency must be ensured for clear visibility

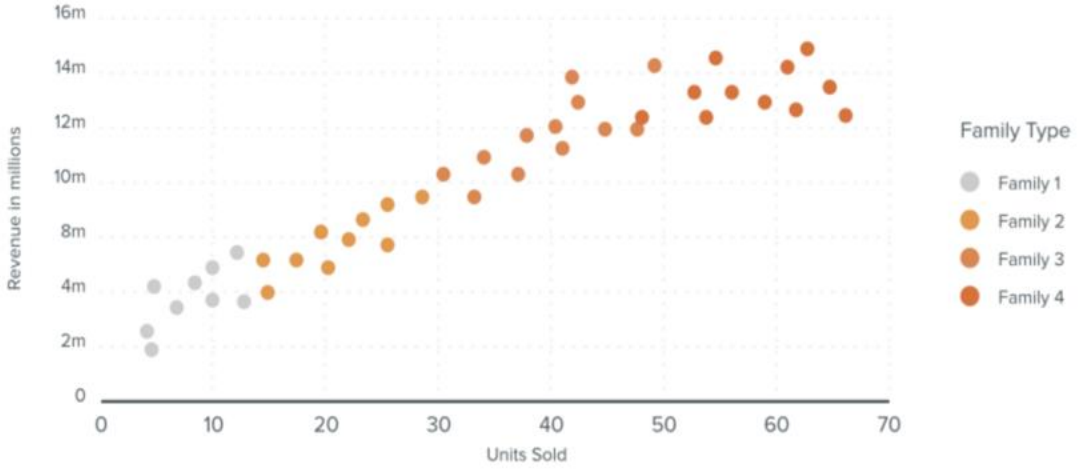


Don't use for Discrete Data

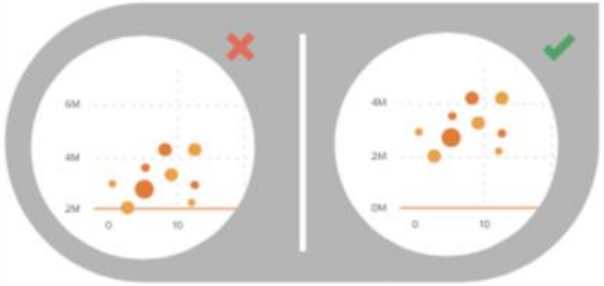
The connected lines imply intermediate values, which only exist in continuous data

Scatterplot chart

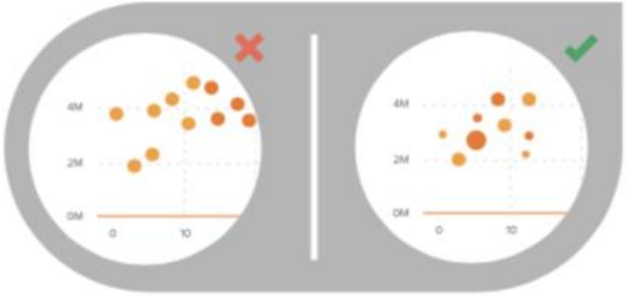
Scatter plots show the relationship between items based on two sets of variables. They are best used to show correlation in a large amount of data.



Scatterplot chart best practices

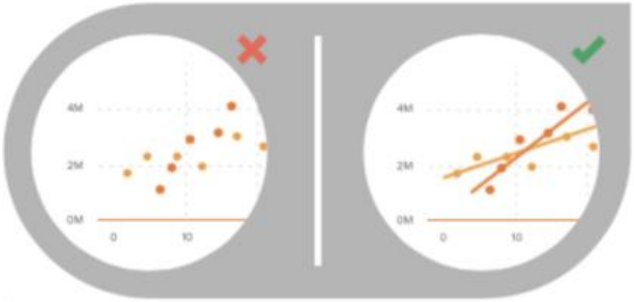


Start with y-axis value at 0



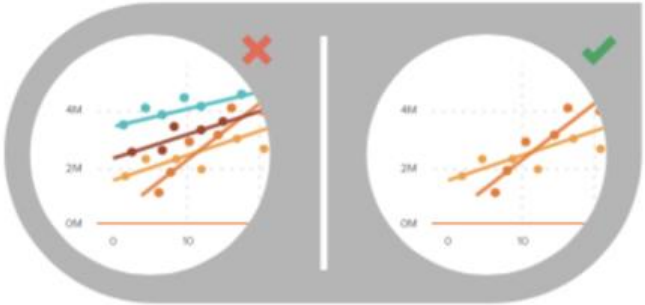
Include more Variables

Use size and dot color to encode additional data variables



Use Trend Lines

These lines help draw correlation between the trending variables

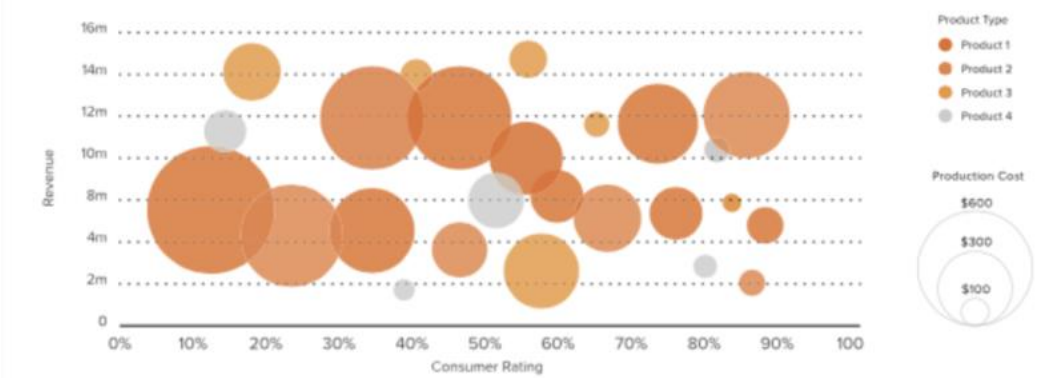


Don't Compare more than 2 Trend Lines

Too many lines make it difficult to interpret

Bubble chart

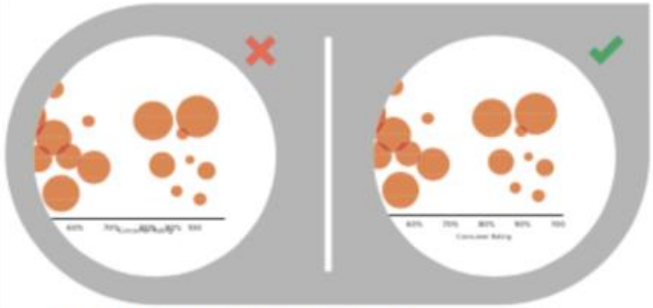
Bubble charts are good for displaying nominal comparisons or ranking relationships.



Bubble Plot is a scatterplot with bubbles best used to display an additional variable.

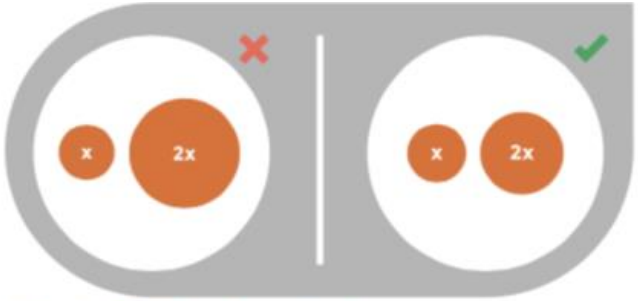
Bubble map is best used to visualize values for specific geographic regions.

Bubble chart best practices



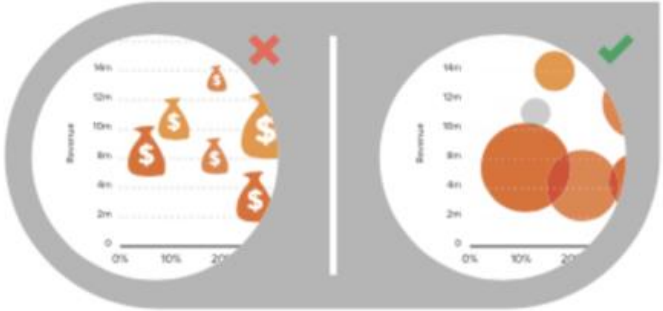
Label Visibility must be ensured

Make sure the labels are visible, easily identifiable and unobstructed



Size the Bubbles Appropriately

Bubbles should be scaled according to the area and not the diameter.

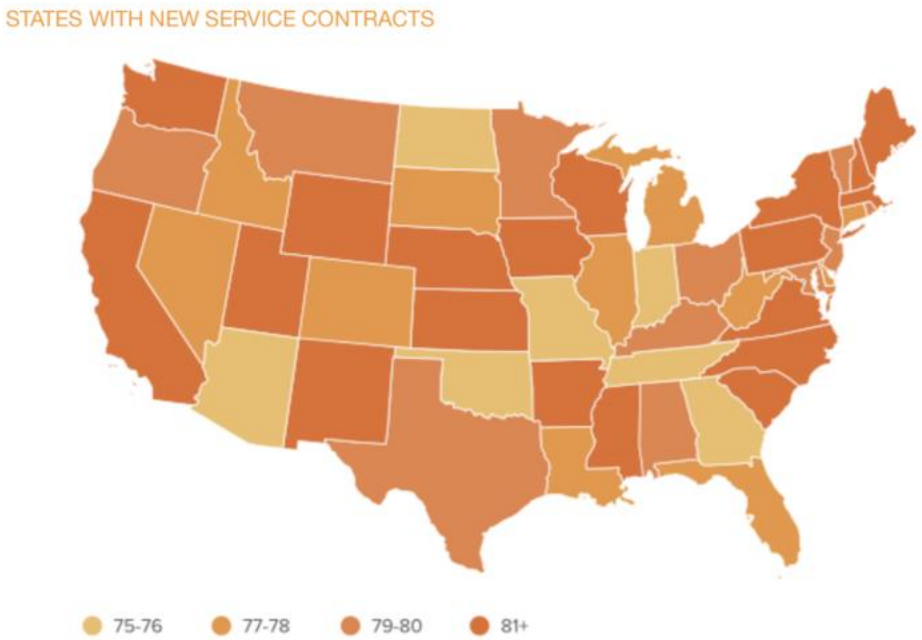


Avoid using Odd shapes

Avoid adding too much details or using shapes that are not entirely circular, this can lead to inaccuracies.

Heat Map

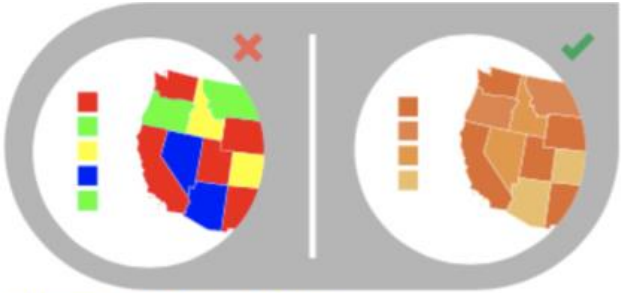
Heat maps are used to display categorical data, using intensity of color to represent values of geographic areas or data tables.



Heat Map best practices

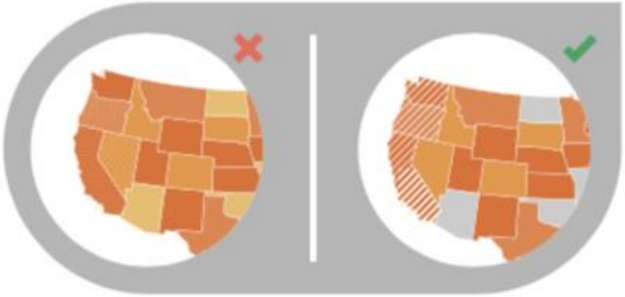
Use a simple Map outline

These lines are meant to frame the data



Appropriate Choice of Colors

Use a single color with varying shades. This will not only make it soothing and appealing visually but also present the results correctly..



Use of Patterns

Use patterns to indicate second variable. But using multiple patterns is overwhelming and distracting

Appropriate Date Ranges

Select 3 to 5 numerical ranges that enable fairly data distribution. Use +/- signs to indicate high and low ranges

Important Tips

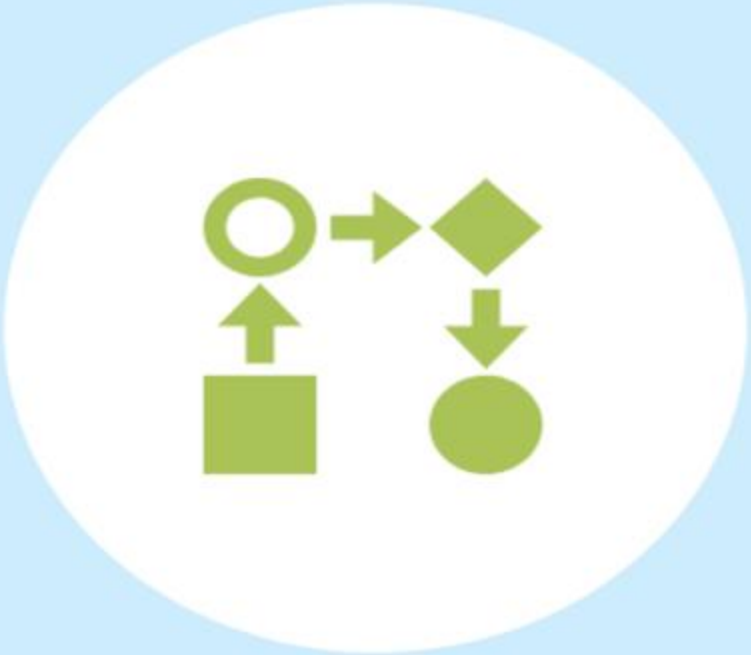
Choice of colors – use color blind friendly palette

Choose a color scheme that can be identified by everyone

Pick a color scheme that people with various kind of color vision deficiencies will be able to see.

http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/#a-colorblind-friendly-palette

Choice of colors – use color blind friendly palette



2

Don't focus your chart design on just using different colors

Focus also on shapes, positions, coloring patterns and line types. This will allow you to make sure that the information is conveyed to those who cannot distinguish the colors.

Choice of colors – use color blind friendly palette

3 Avoid hard-to-see color combinations

Certain color combinations may present an issue for color blind people, such as red & green, green & brown, green & blue, blue & grey, blue & purple, green & grey and green & black.

Choice of colors – use color blind friendly palette

4

Make your chart monochrome

Work with different shades of a single color. This will help you avoid color combinations that are hard to see.

Choice of colors – use color blind friendly palette

5

Use highly contrasting colors

The majority of color blind people can perceive high contrast.

Do’s and don’ts in data design & visualization

- Use one color to present each category.
- Use order data sets using logical hierarchy.
- Use high contrast color combinations such as Red/Green or Blue/Yellow.
- Use callouts to highlight important or interesting information.
- Use 3D charts.
- Visualize your data in a way that it’s easy for readers to compare values.
- Add chart junk.
- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy.
- Use high contrast color combinations such as Red/Green or Blue/Yellow.
- Use callouts to highlight important or interesting information.
- Use 3D charts.
- Visualize your data in a way that it’s easy for readers to compare values.
- Add chart junk.
- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow.
- Use callouts to highlight important or interesting information.
- Use 3D charts.
- Visualize your data in a way that it’s easy for readers to compare values.
- Add chart junk.
- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information.
- Use 3D charts.
- Visualize your data in a way that it’s easy for readers to compare values.
- Add chart junk.
- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information. 

- Use 3D charts.
- Visualize your data in a way that it’s easy for readers to compare values.
- Add chart junk.
- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information. 

- Use 3D charts. 

- Visualize your data in a way that it’s easy for readers to compare values.
- Add chart junk.
- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information. 

- Use 3D charts. 

- Visualize your data in a way that it’s easy for readers to compare values. 

- Add chart junk.
- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information. 

- Use 3D charts. 

- Visualize your data in a way that it’s easy for readers to compare values. 

- Add chart junk. 

- Use more than 6 colors in a single layout.
- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information. 

- Use 3D charts. 

- Visualize your data in a way that it’s easy for readers to compare values. 

- Add chart junk. 

- Use more than 6 colors in a single layout. 

- Use icons to enhance comprehension.
- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information. 

- Use 3D charts. 

- Visualize your data in a way that it’s easy for readers to compare values. 

- Add chart junk. 

- Use more than 6 colors in a single layout. 

- Use icons to enhance comprehension. 

- Use italic, bold or underline text.

Do’s and don’ts in data design & visualization

- Use one color to present each category. 

- Order data sets using logical hierarchy. 

- Use high contrast color combinations such as Red/Green or Blue/Yellow. 

- Use callouts to highlight important or interesting information. 

- Use 3D charts. 

- Visualize your data in a way that it’s easy for readers to compare values. 

- Add chart junk. 

- Use more than 6 colors in a single layout. 

- Use icons to enhance comprehension. 

- Use italic, bold or underline text. 

Next part: Tutorial

Data Visualisation

A picture is worth a thousand of words

Thank you