Data Visualization and Summarization

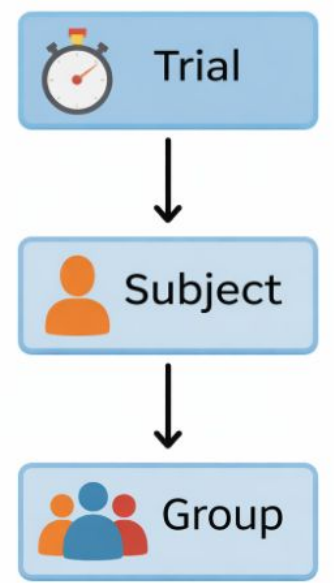
BRSM

Gargi Shukla

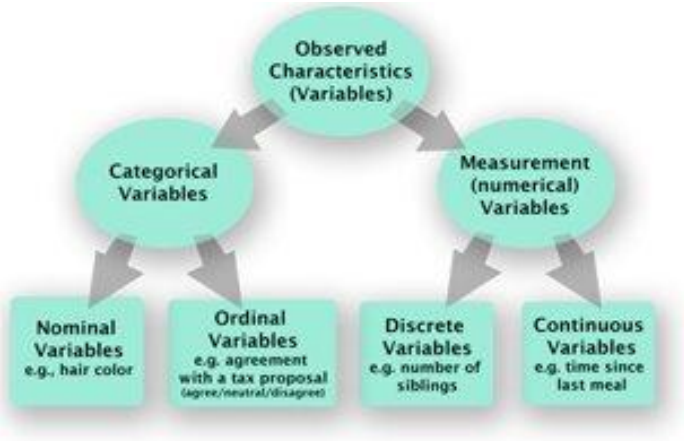
How visual choices shape interpretation, inference, and scientific conclusions

Before You Visualize: Organizing and Summarizing Data

- Behavioral data has **structure** (trial → subject → group)
- Visualization does **not start with plots.**



STEP 1: Identify Variable and variable type



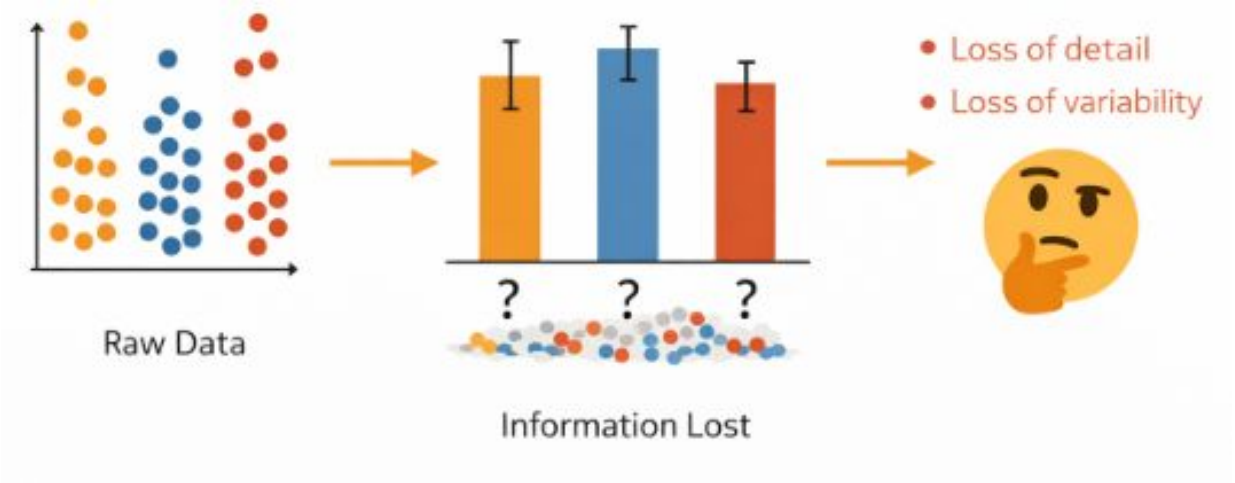
Step 2: Identify unit of analysis

Trial, subject, group?

Summarization is already a choice

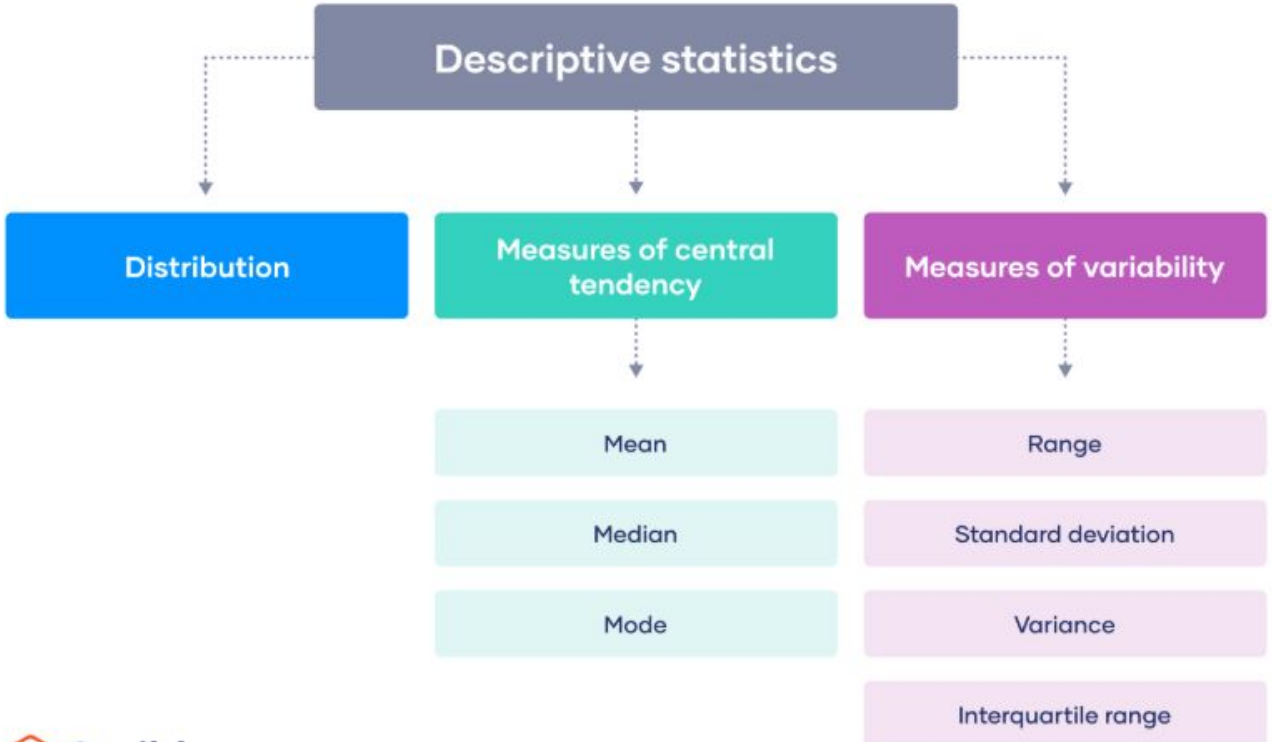
Any summary (mean, median, %) already removes information.

- Means hide **distribution shape**
- Aggregation can hide **individual differences**

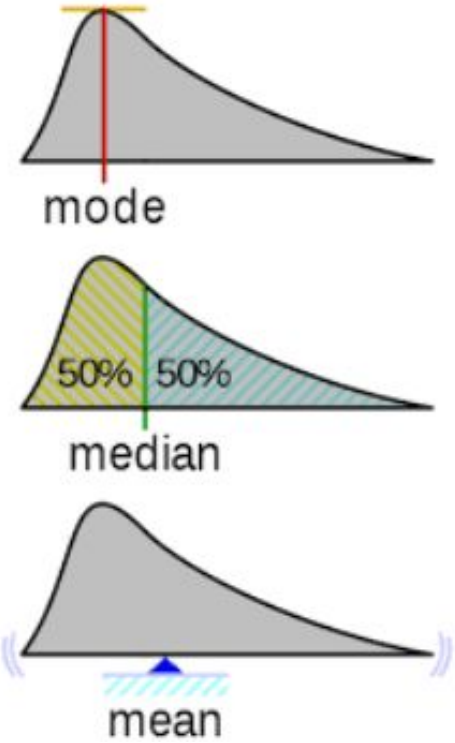


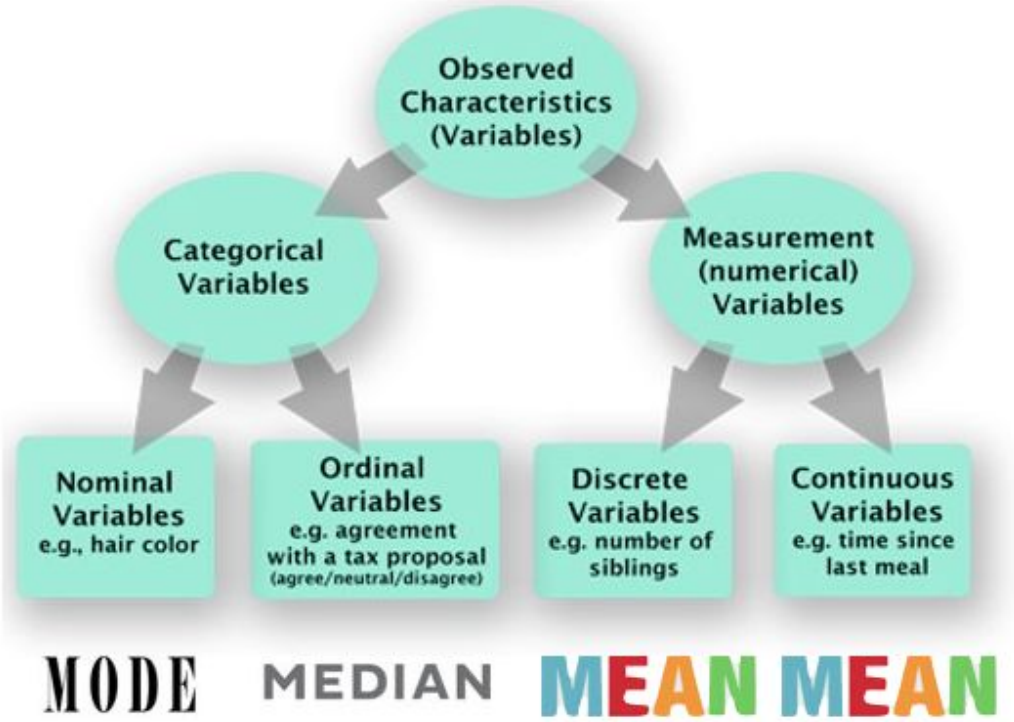
The moment you average across trials or participants, you've already made a decision about what matters in the data. Visualization helps us check whether those decisions are reasonable.

Summary Statistics/ Descriptive Statistics

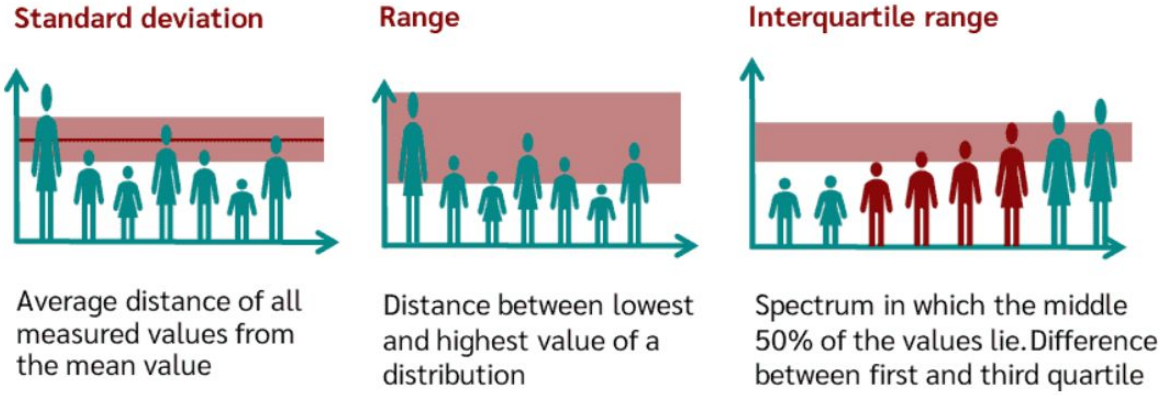


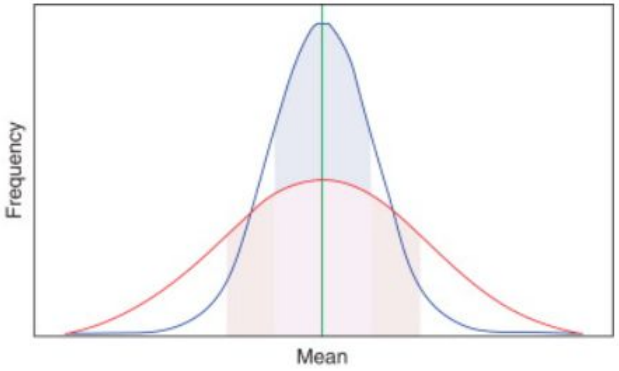
Central tendency





Measure of Variability



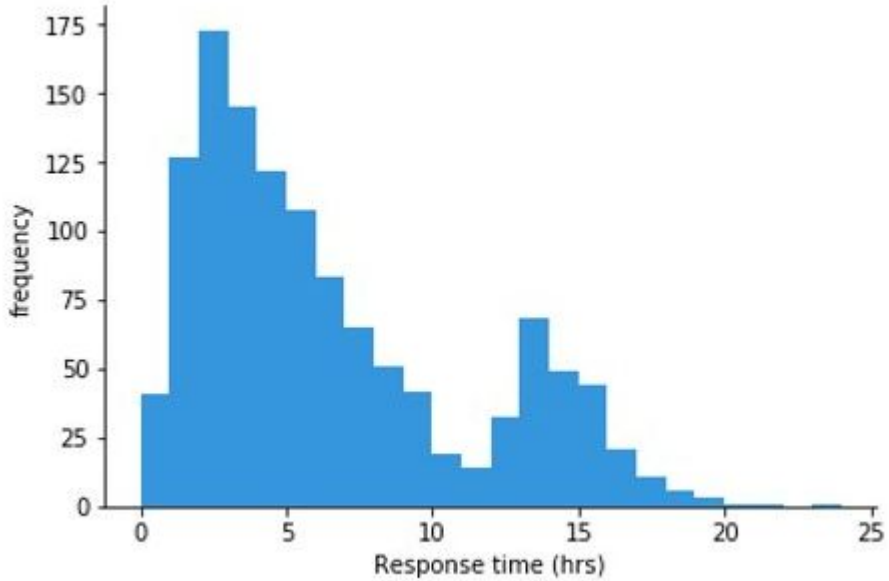


Variance: How far each point in data is from the mean of the data?

Variance is similar to mean deviation except that the deviations are squared rather than being expressed in absolute values. Why squared?

Data Distribution

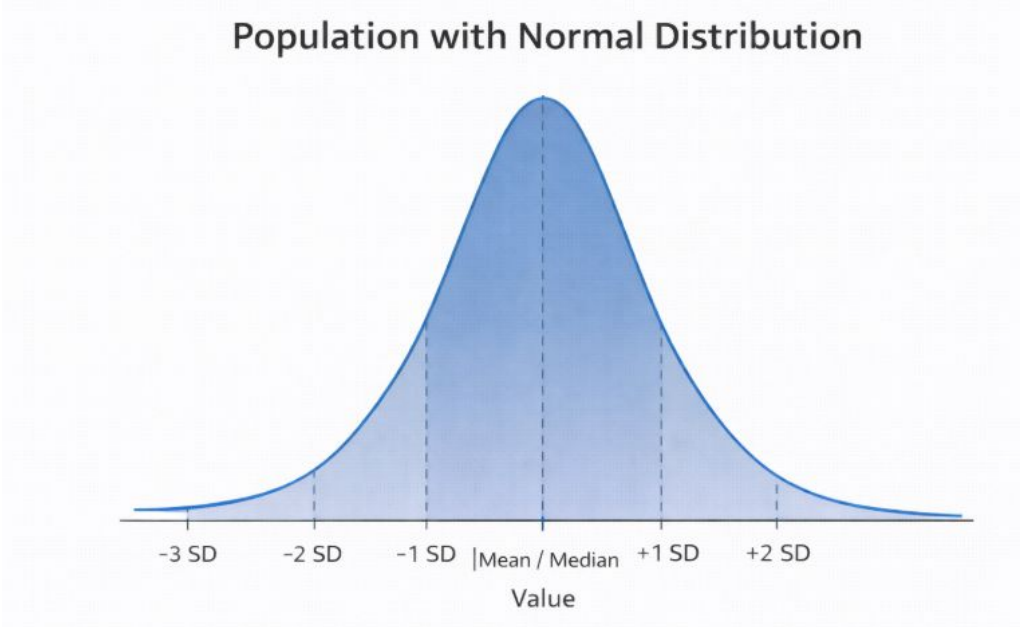
- A distribution describes how often different values occur
- Indicative of potential groups or group differences



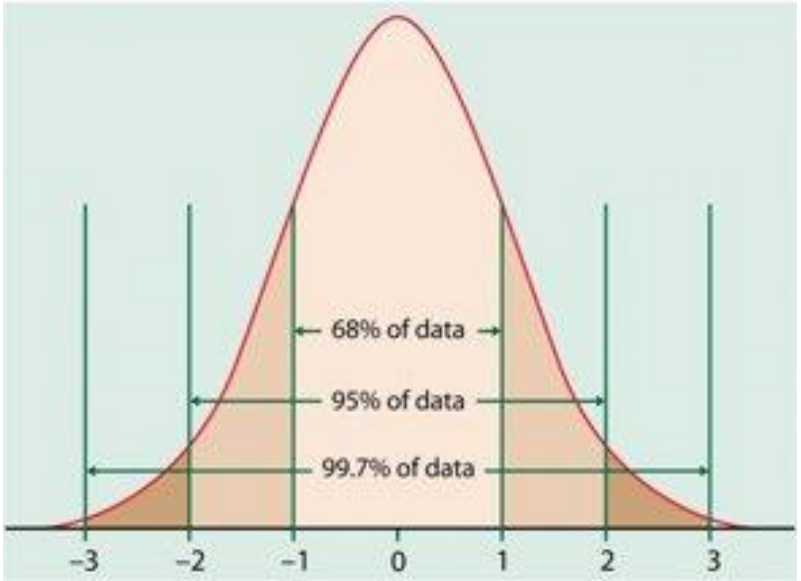
Normal Distribution

A bell-shaped mathematical curve describing how values are distributed

Data taken from a sample is assumed to be ‘normally distributed’, and must approximate this shape in order to use parametric tests of significance

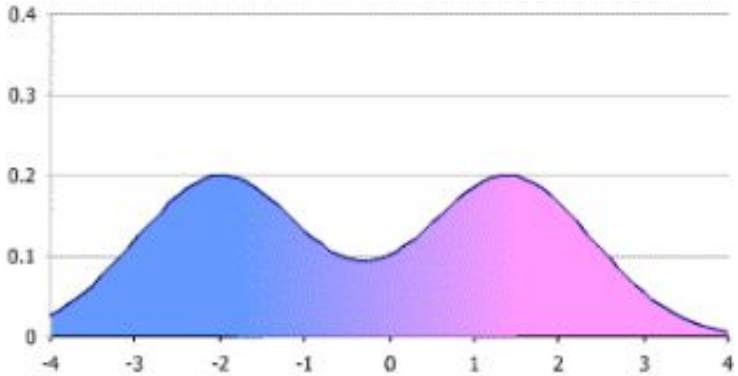


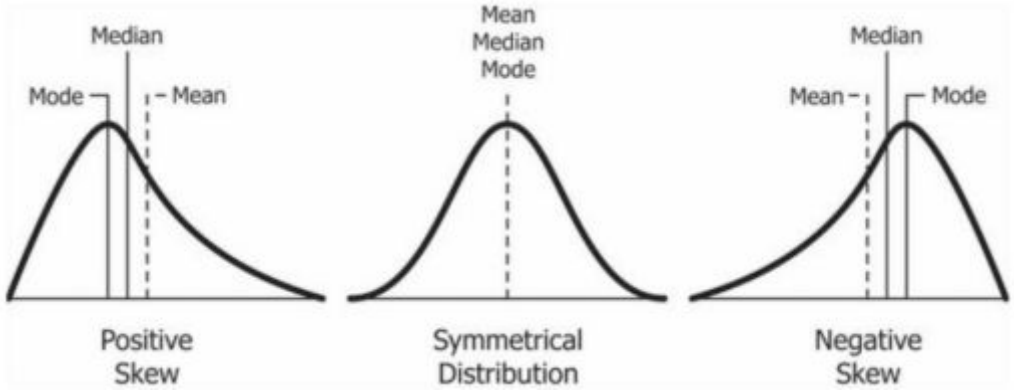
- symmetrical about the horizontal axis midpoint
- mean, median, and mode all fall on the midpoint
- No matter what μ and σ are, the area between
  - μ-σ and μ+σ is about 68%;
  - μ-2σ and μ+2σ is about 95%;
  - μ-3σ and μ+3σ is about 99.7%
- Almost all values fall within 3 standard deviations



Skewed Distribution

- Resembles an exponential distribution
- Lots of extreme values far from mean or mode
- Not straightforward to do useful statistical tests with this type of distribution





Right-skewed (positively skewed)

- Long tail to the right, Mean > median
  Example: Reaction times

Left-skewed (negatively skewed)

- Long tail to the left, Mean < median
  Example: Accuracy with ceiling effects

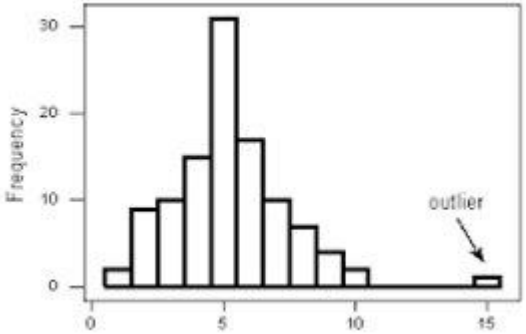
Bimodal / Multimodal

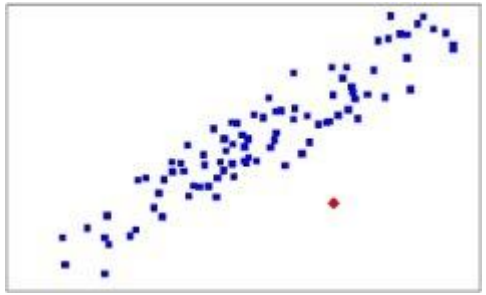
- Two or more peaks
  Example: Mixed strategies or responder subgroups

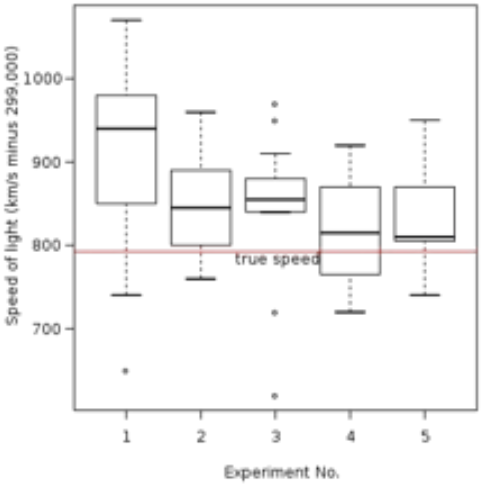
Why Visualize at all?

Visualization helps us see what summary statistics cannot.

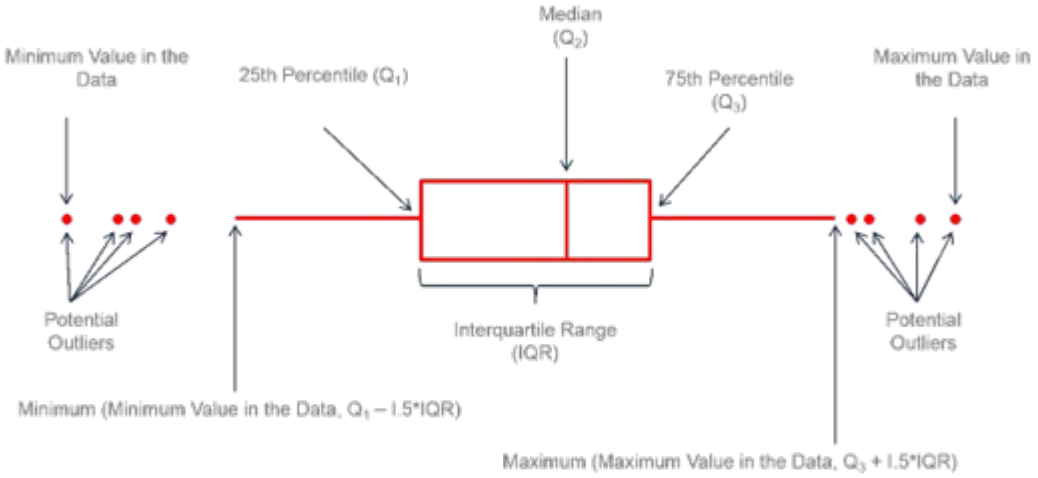
- **Check assumptions about distribution of data** (normality, skew, multimodality)
- **Detect structure in the data** (learning, fatigue, strategy shifts)
- **Identify outliers and data errors** (RT artifacts, missing trials, NaNs)
- **Understand variability** (within-subject vs between-subject)
- **Decide appropriate statistical models** (parametric vs non-parametric)

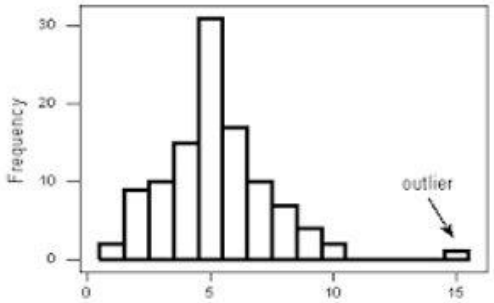


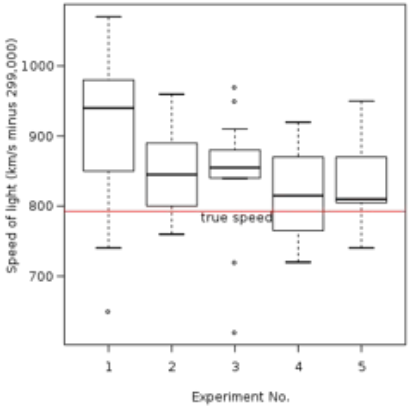


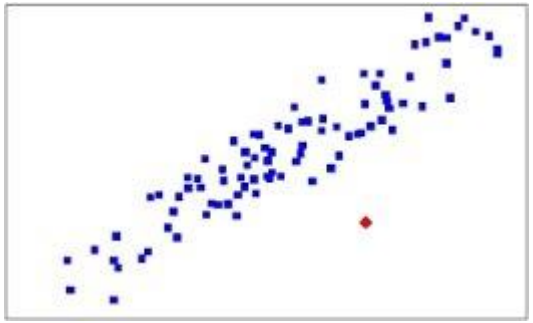


Histogram, BoxPlot, Scatter Plot









Violin Plots

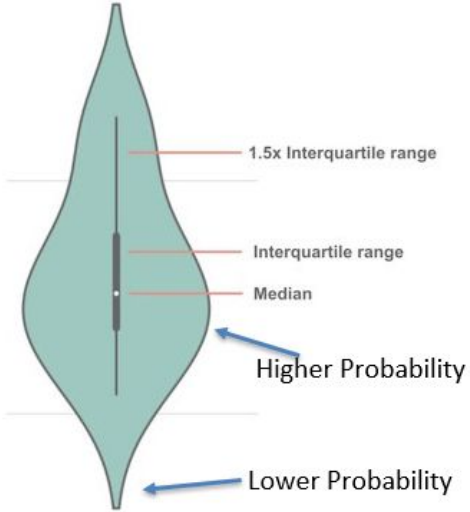
Better than box plots for showing skew and tails.

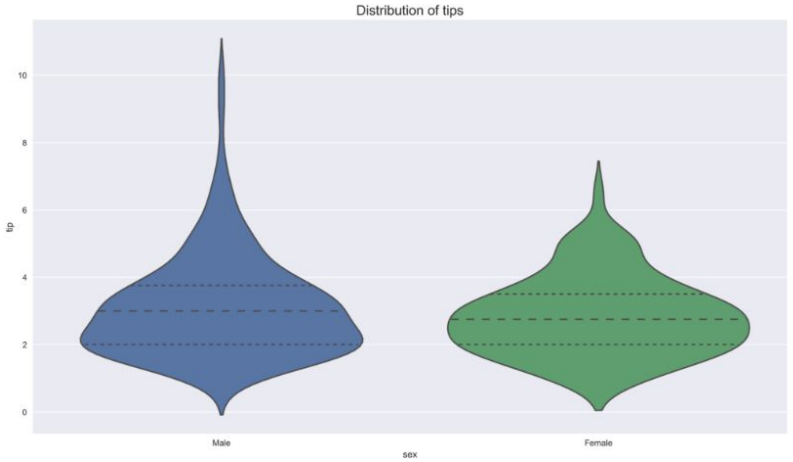
Shows:

- Full distribution
- Median and spread

Use when:

- For comparing distributions across conditions
- RT, confidence, continuous ratings



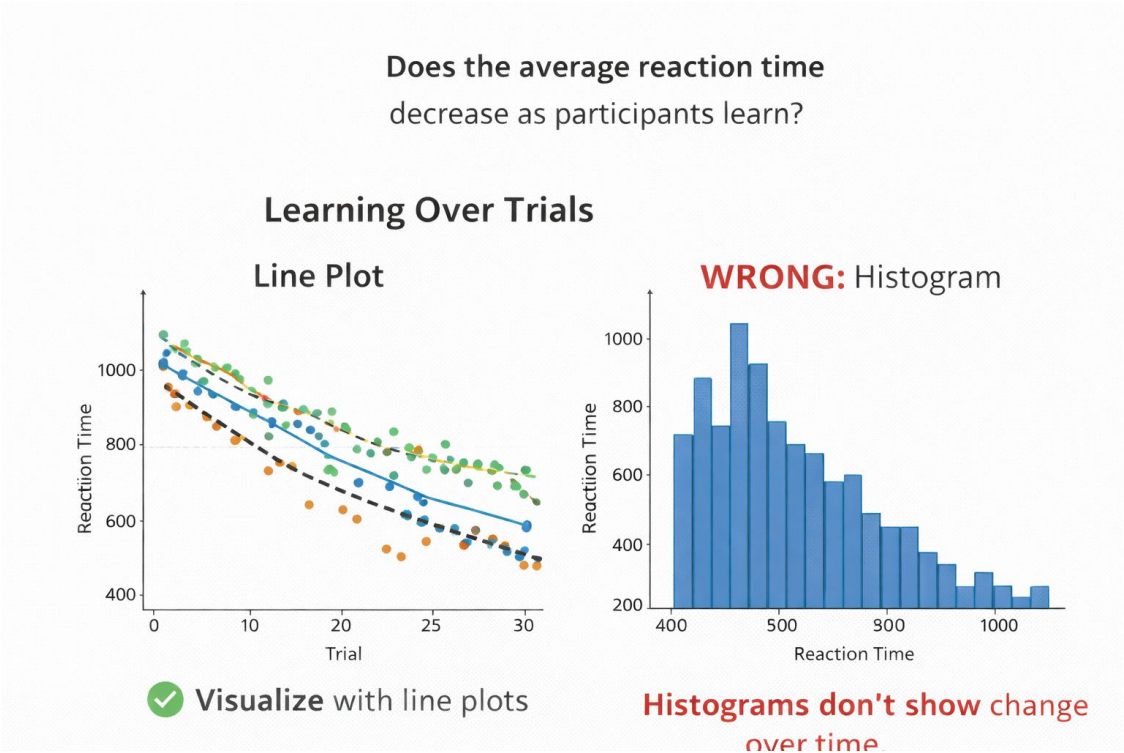


Line Plot (Temporal/Learning Curves)

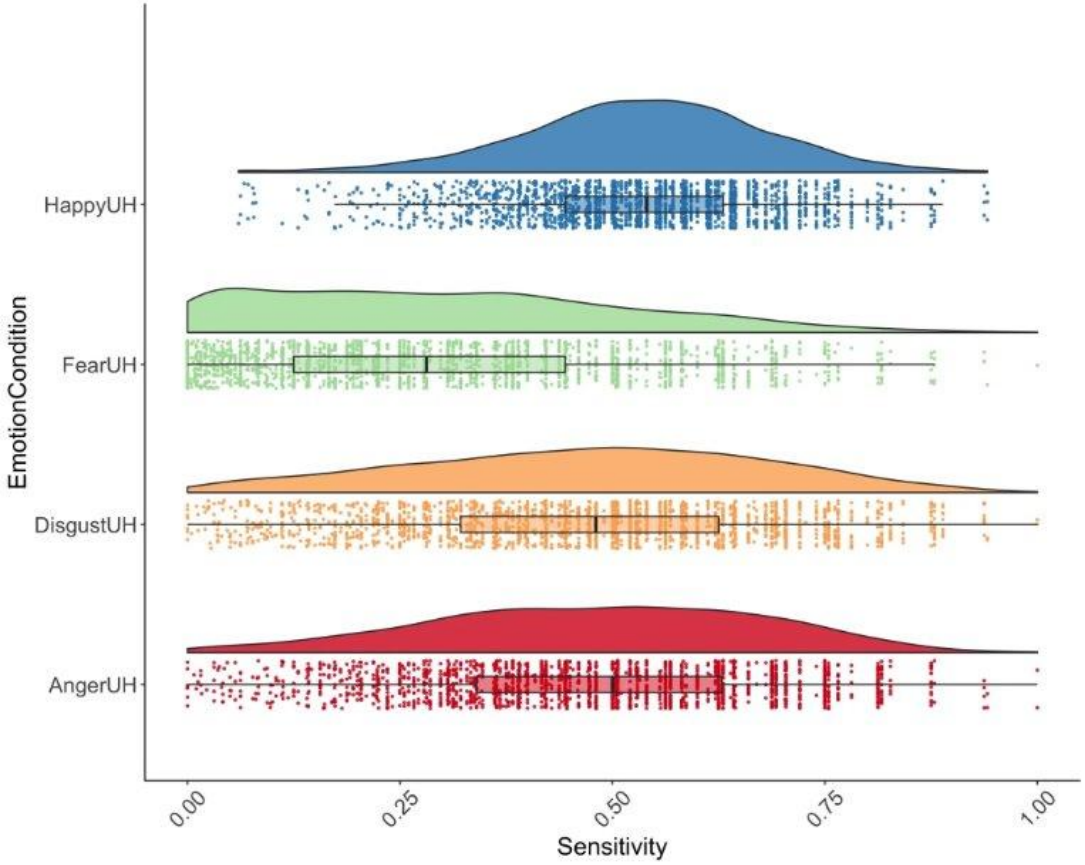
- Shows change over time or trials

Use when:

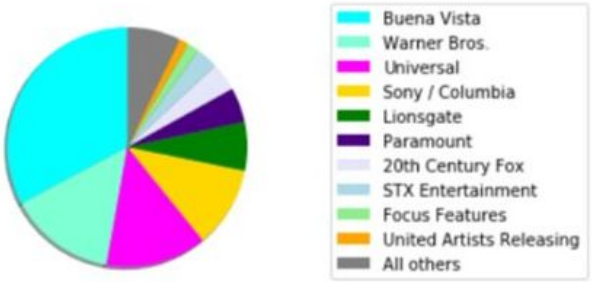
- Learning, fatigue, adaptation
- Neural or behavioral time courses
- Show individual trajectories, not just group averages.

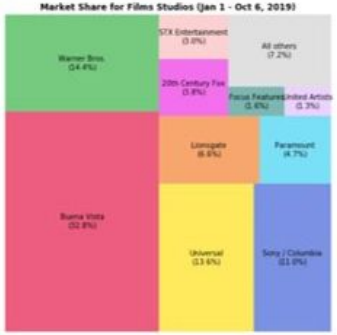


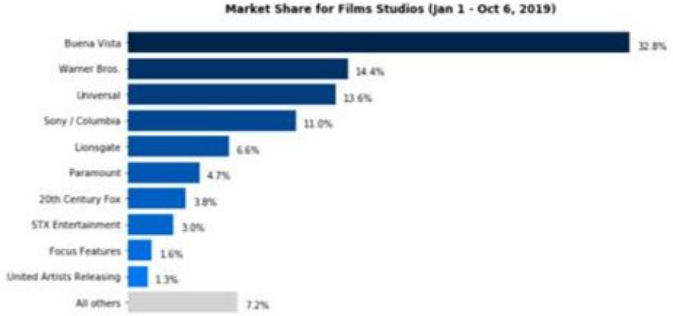
Raincloud/Raindrop Plot



Mosaic Plot and Waffle Plots



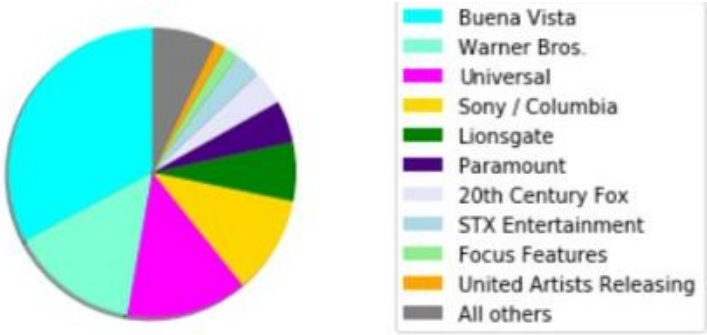




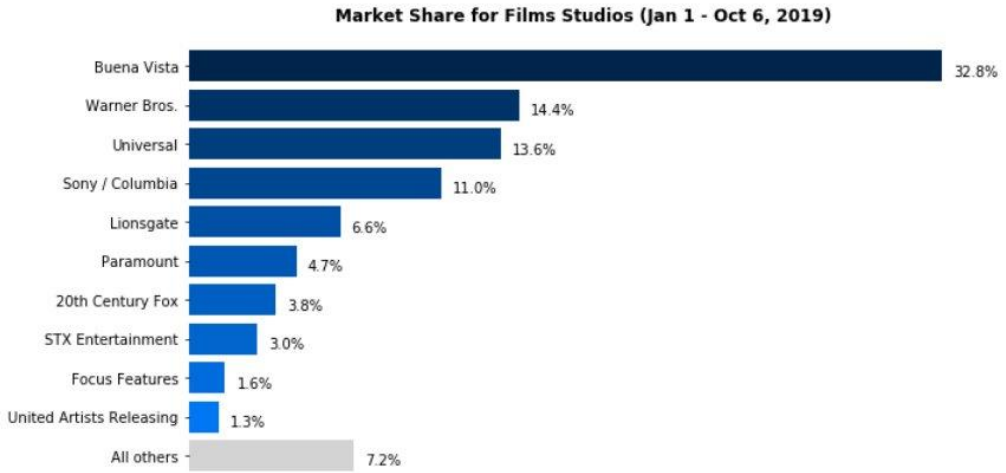
Pie charts

Use pie charts when:

- Smaller no. of categories
- Readers can differentiate slices (unless you are making a point)
- You don’t need to rely on many colors or labels to explain the proportions
- Total adds up to 100%

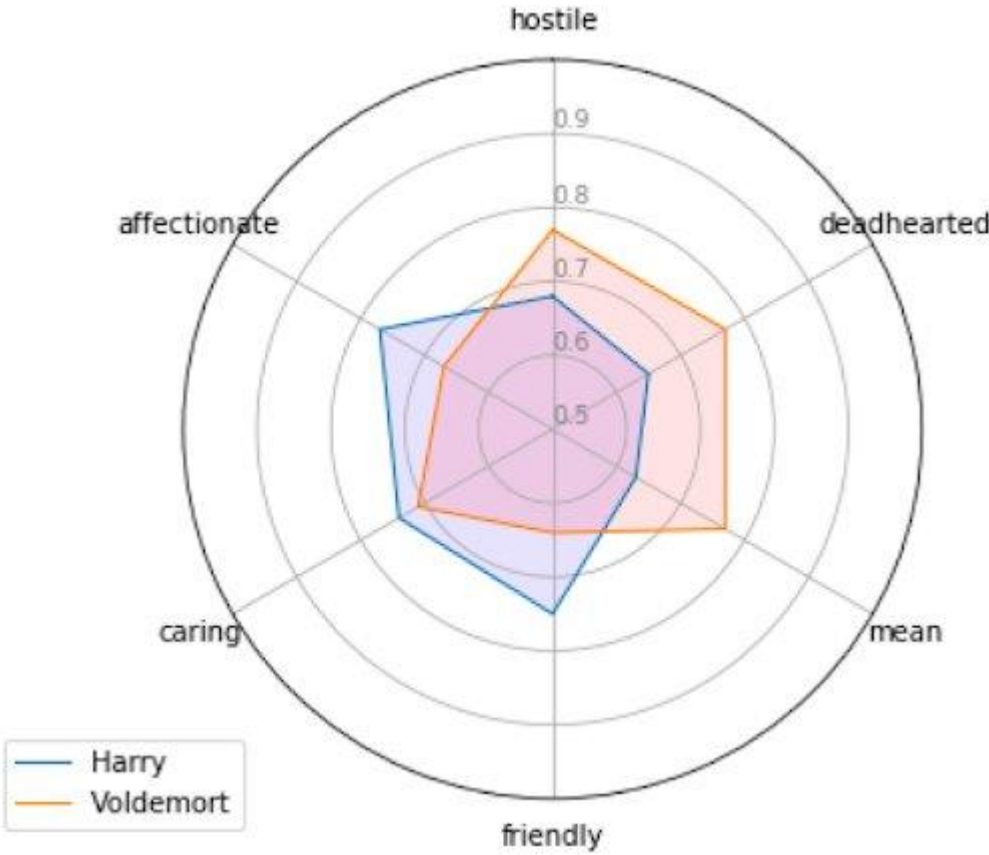


Difficult to comprehend!



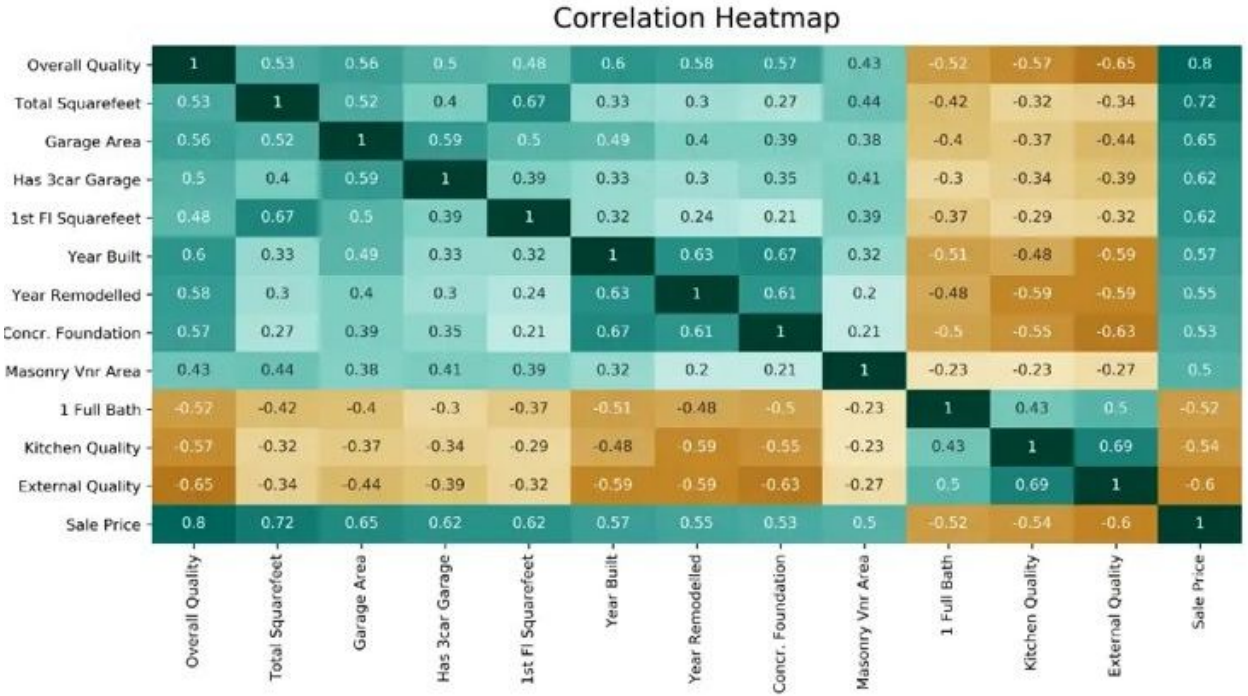
Spider plot/ Radar chart

Looks fancy, often hard to interpret.



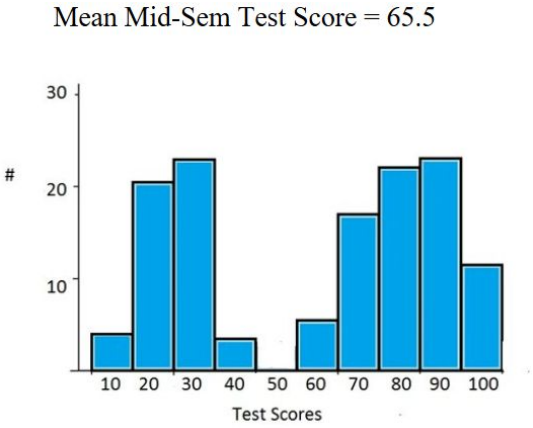
Heat maps

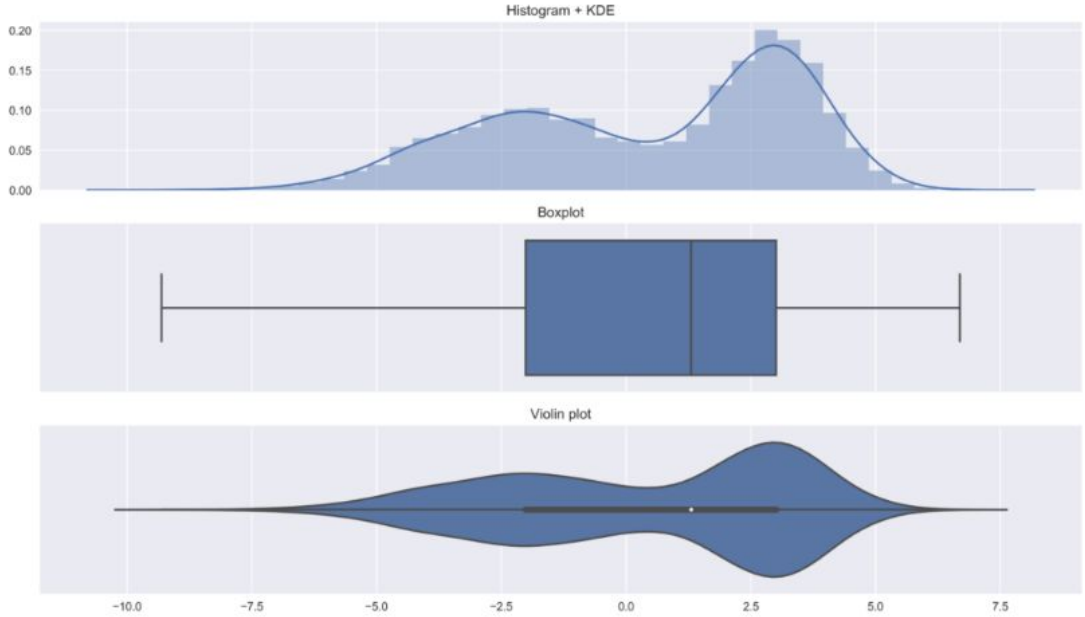
- Depict Correlation



What Can Go Wrong If We Don't Visualize the Data?

- Hidden distribution structure
- Outlier driving effects
- Group mean suggests no effect, individuals show strong but opposite patterns
- Data quality issues



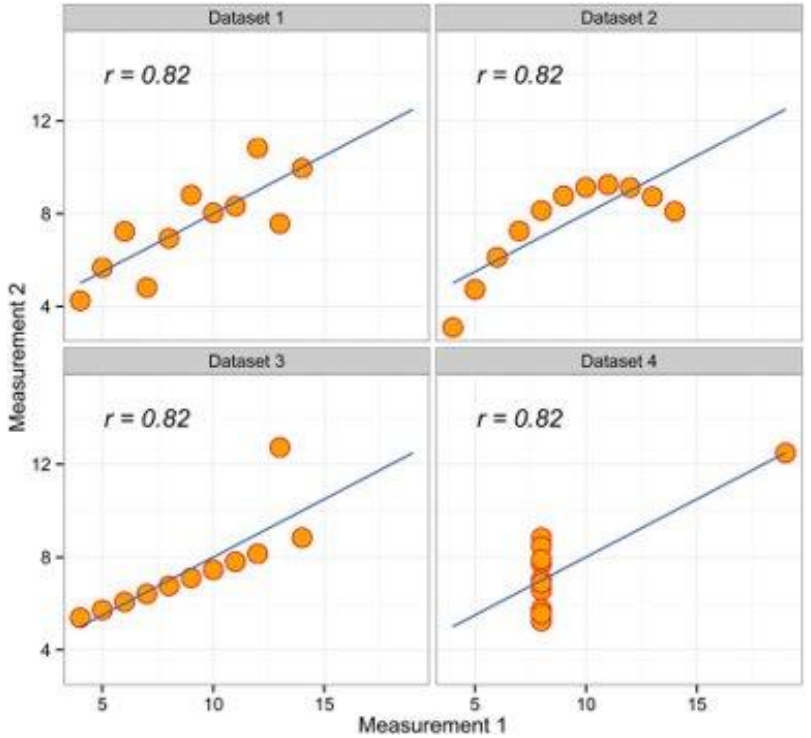


Anscombe's Quartet: Same Statistics, Different Data

All four datasets have the same mean, variance, correlation, and regression line.

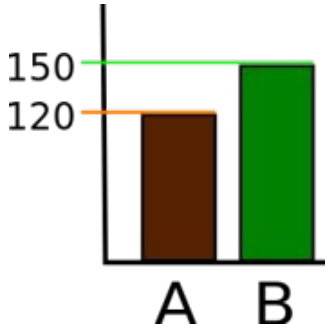
Statistics compress data.
Visualizations reveal structure.

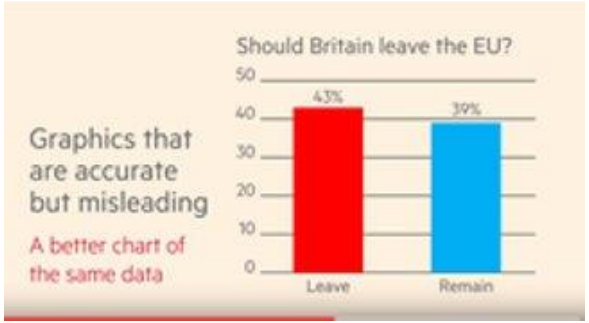
You should visualize *before* choosing a statistical test and not after.

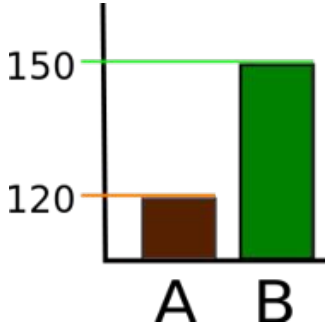


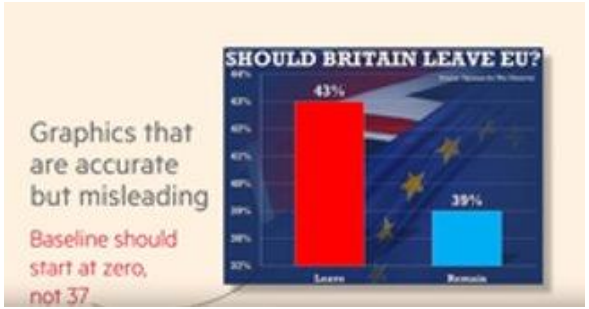
What Makes a Good Visualization?

- Show the data, not just summaries
- Reduce cognitive load
- Match the plot to the question
- Use scales honestly
- Make assumptions visible





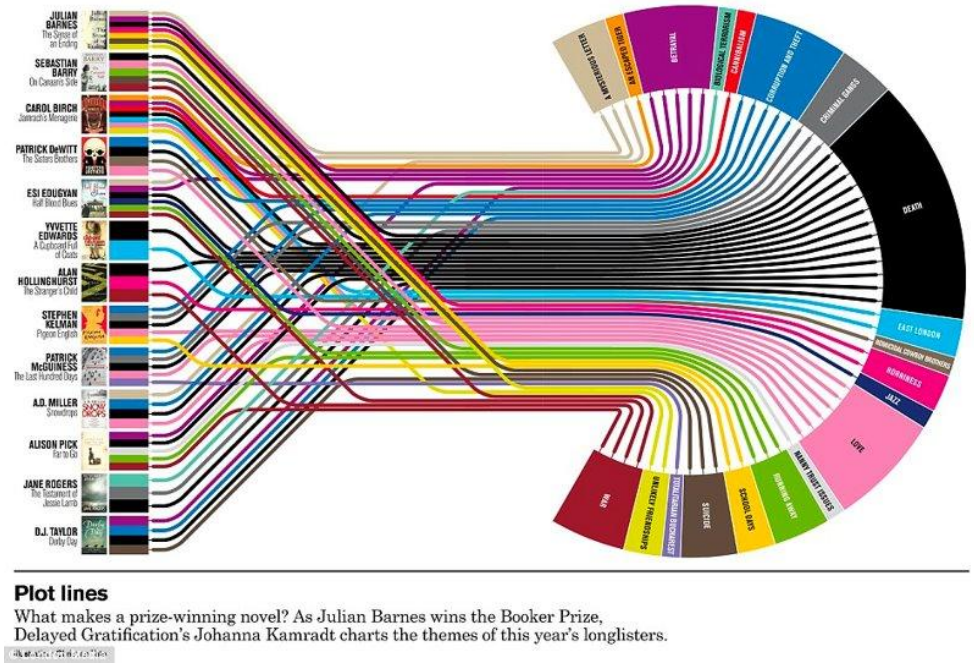




What Makes a ‘Bad’ Visualization?

The wrong visualization can change the scientific conclusion.

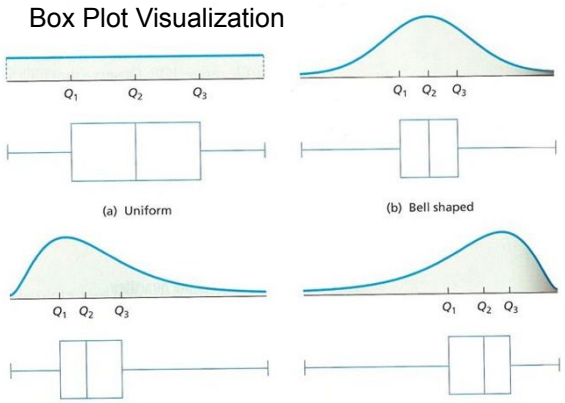
- Bar plots for continuous data (RT): Hides skew and multimodalities
- Truncated or inconsistent axes
- Overplotting without transparency
- Decorative or misleading charts

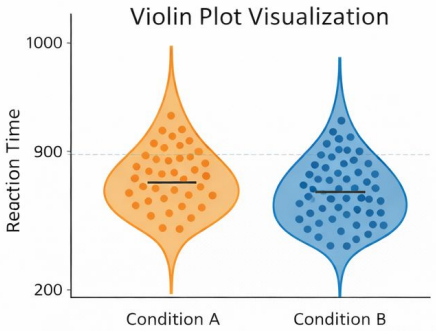


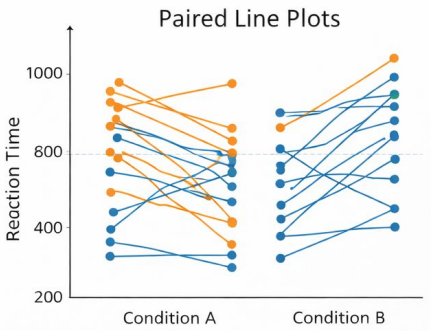
Plot lines

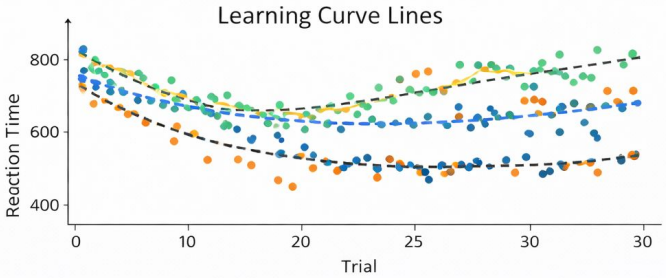
What makes a prize-winning novel? As Julian Barnes wins the Booker Prize, Delayed Gratification's Johanna Kamradt charts the themes of this year's longlisters.

Better Ways to Visualize Behavioral Data





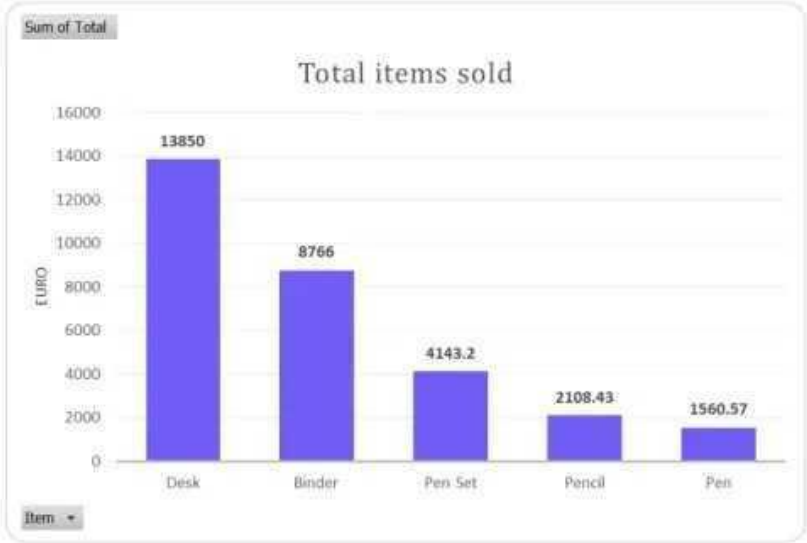




- Box plots encode distributiona shape, not just spread.
- Box plots do **NOT** show bimodality or clusters.

Cluttered chart

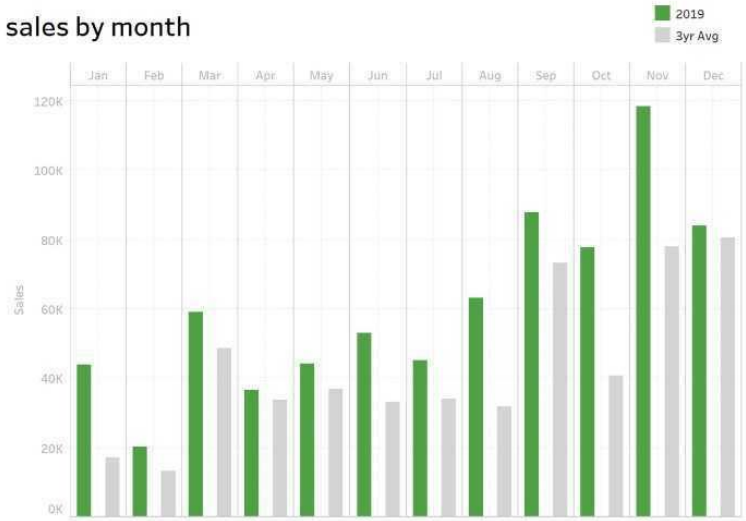
(bad data visualization example)



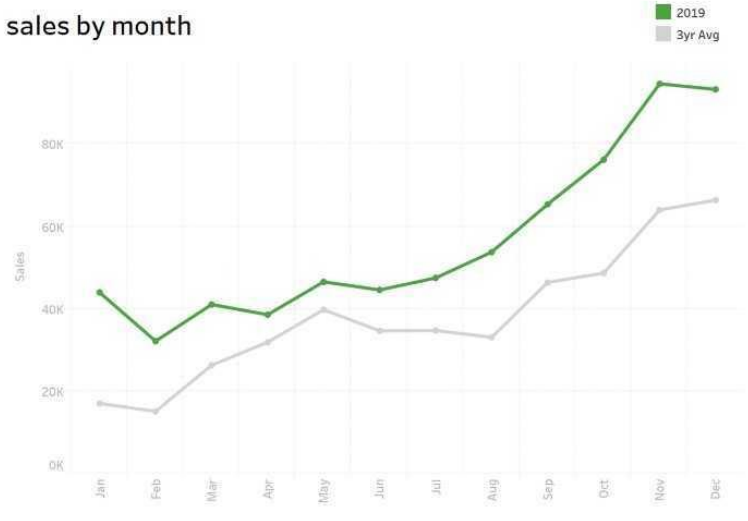
Clean chart

(good data visualization example)

Contributed by: Samarth Srikar, UG4



ineffective

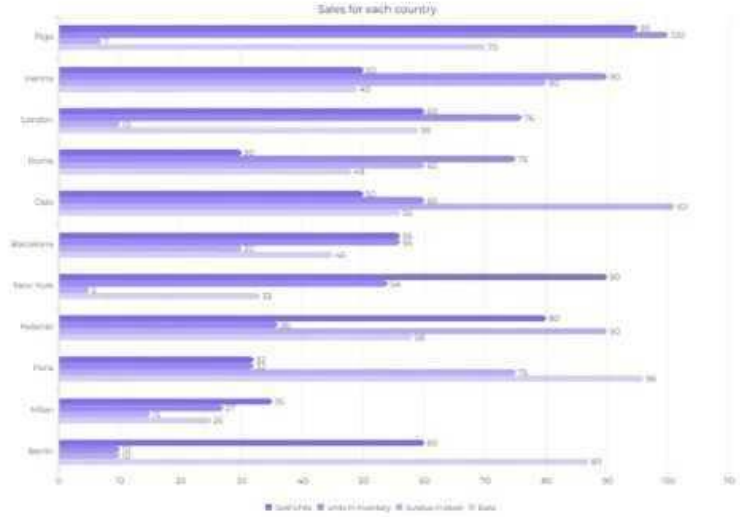


effective

Contributed by: Samarth Srikar, UG4

Bad coloring

(too many colors that distract from the data)



Good coloring

(all colors are one shade and you focus on the data)

Contributed by: Samarth Srikar, UG4

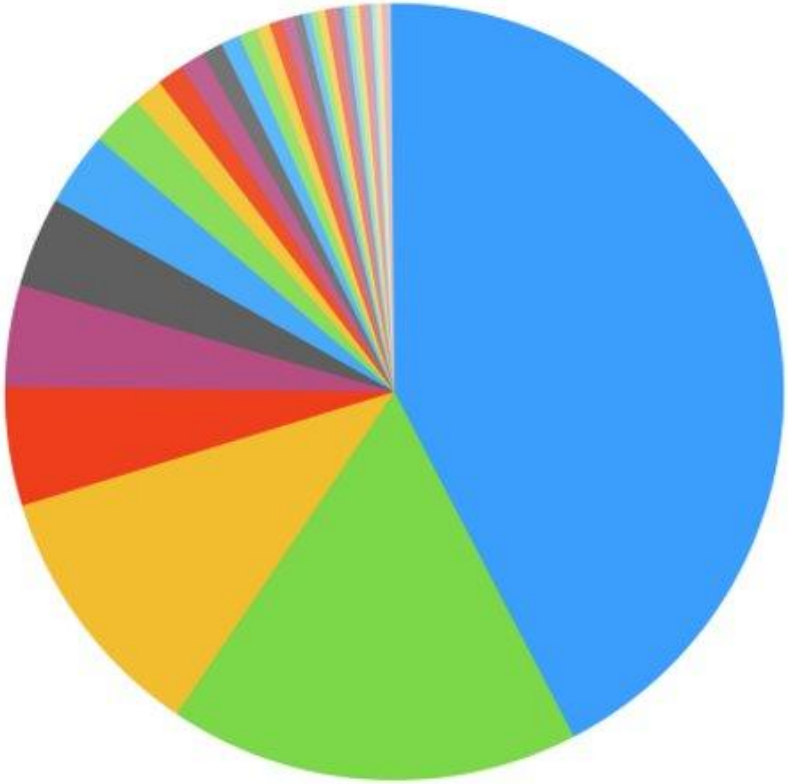
What makes them “good” or “bad”?

Comment on these visualizations

Tim Cook used the particular chart to showcase the rising sale of iPads between the years 2008-2013.

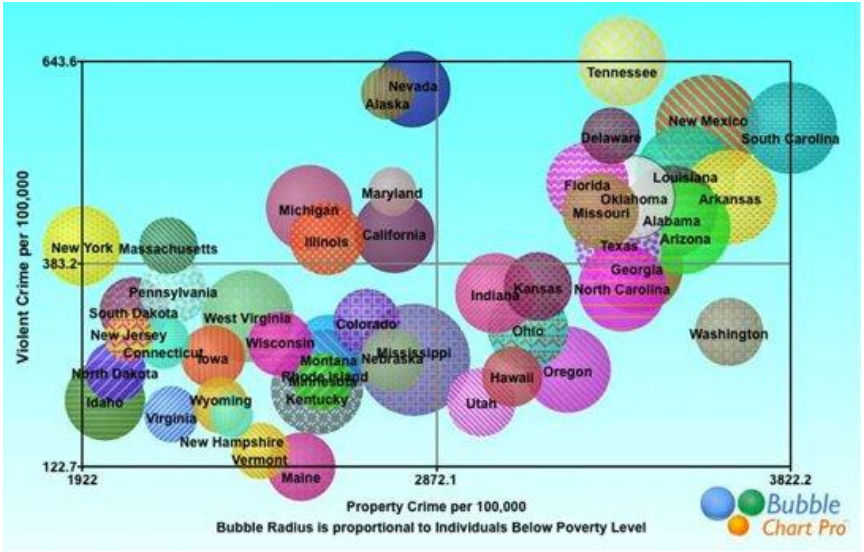
Which game(s) have you played the most?

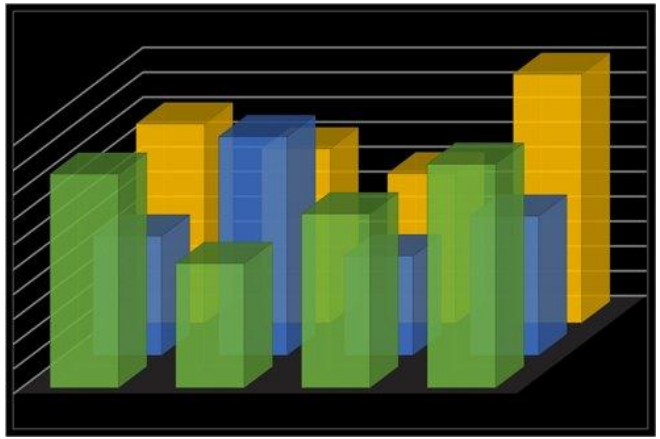
3,994 responses



- Zelda
- The Legend of Zelda: Breath of the Wild
- Breath of the Wild
- BOTW
- Botw
- Breath of the wild
- BotW
- zelda
- Legend of Zelda: Breath of the Wild
- Legend of Zelda
- Zelda BOTW
- BoTW
- botw
- Zelda: Breath of the Wild
- Zelda BotW
- Zelda Breath of the Wild
- The Legend of Zelda
- Breath of The Wild
- The Legend of Zelda Breath of the Wild
- Zelda: BOTW
- Zelda: BotW
- Breath of the Wild
- Zelda breath of the wild
- Breath Of The Wild
- Legend of Zelda Breath of the Wild
- LoZ
- LoZ: BotW
- Zelda botw
- zelda botw
- breath of the wild
- Legend of zelda
- legend of zelda
- LoZ BOTW
- The Legend of Zelda: Breath of The Wild
- The legend of Zelda: breath of the wild
- ZELDA
- Zelda: BoTW

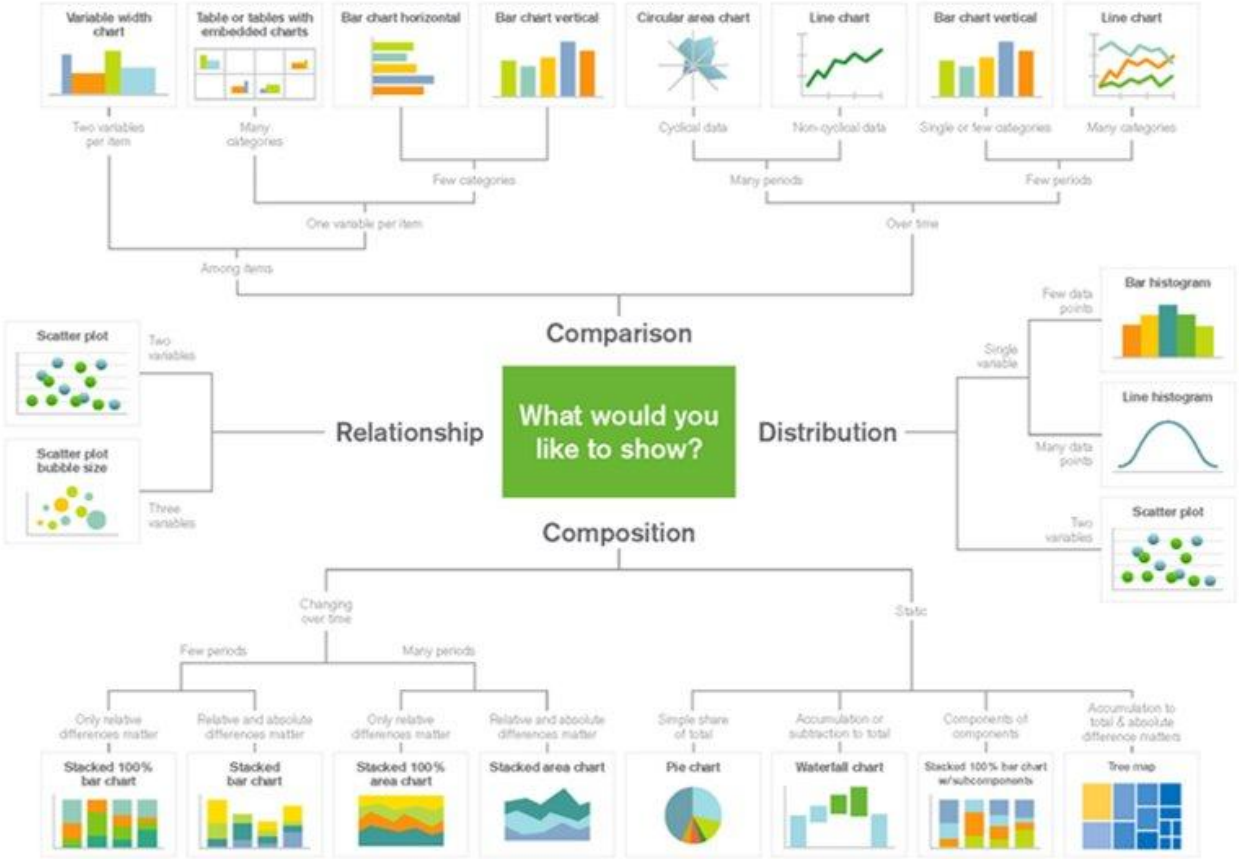

62,500
62,000
61,500
61,000
60,500
60,000
59,500
59,000
58,500
58,000
57,500
57,000
56,500
56,000
55,500
55,000
54,500
11:00:00 GMT+11
BullionVault
click chart to zoom out
08 Mar
22 Mar
05 Apr
19 Apr
03 May
17 May


643.6
383.2
122.7
1922
2872.1
3822.2
Violent Crime per 100,000
Property Crime per 100,000
Bubble Radius is proportional to Individuals Below Poverty Level
Bubble Chart Pro
Nevada
Alaska
Tennessee
New Mexico
South Carolina
Delaware
Florida
Louisiana
Oklahoma
Arkansas
Missouri
Alabama
Arizona
Texas
Georgia
North Carolina
Maryland
Michigan
Illinois
California
New York
Massachusetts
Pennsylvania
South Dakota
West Virginia
New Jersey
Connecticut
Iowa
Wisconsin
Colorado
Indiana
Kansas
Ohio
Washington
Mississippi
Montana
Nebraska
North Dakota
Rhode Island
Minnesota
Kentucky
Hawaii
Oregon
Utah
Idaho
Wyoming
Virginia
New Hampshire
Vermont
Maine



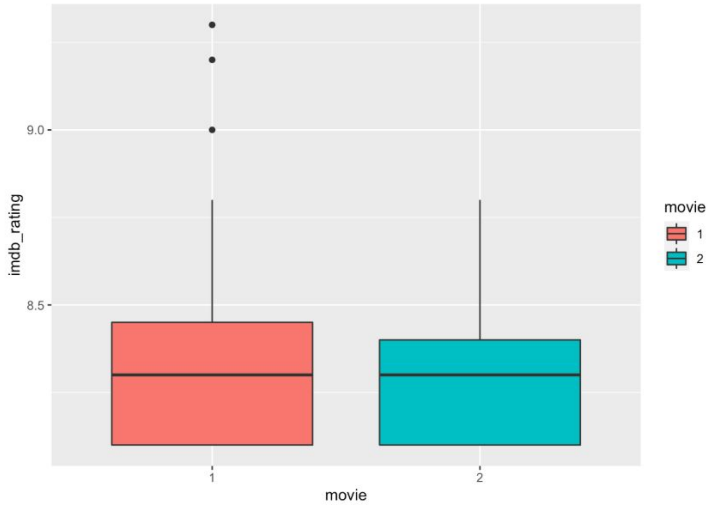
Visualizing Different Research Questions

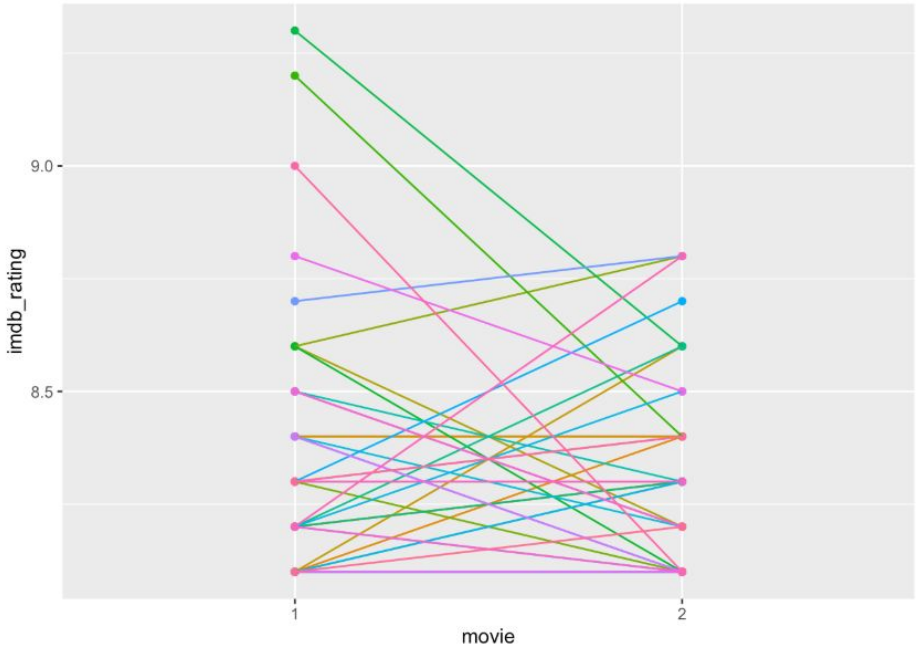
There is no ‘best plot’.
There is only the best plot for your question



Group Differences

- Main effects and interaction plots
- Group differences should be visualized with distributions and individual data—not only means.





To do and **Not to do** during Visualizations

To-do's:

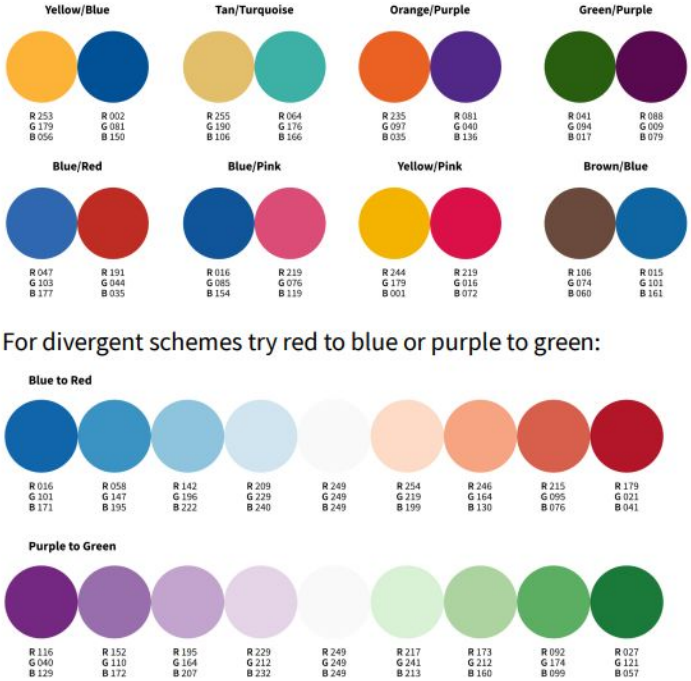
- Provide necessary Context around Visuals
- Ensure Simplicity and Clarity of Information
- Avoid Unnecessary Information
- Use Simple and Easy to Understand Color Palettes
- Pay attention to Graphics that they are Visually Appealing

Not To-do's

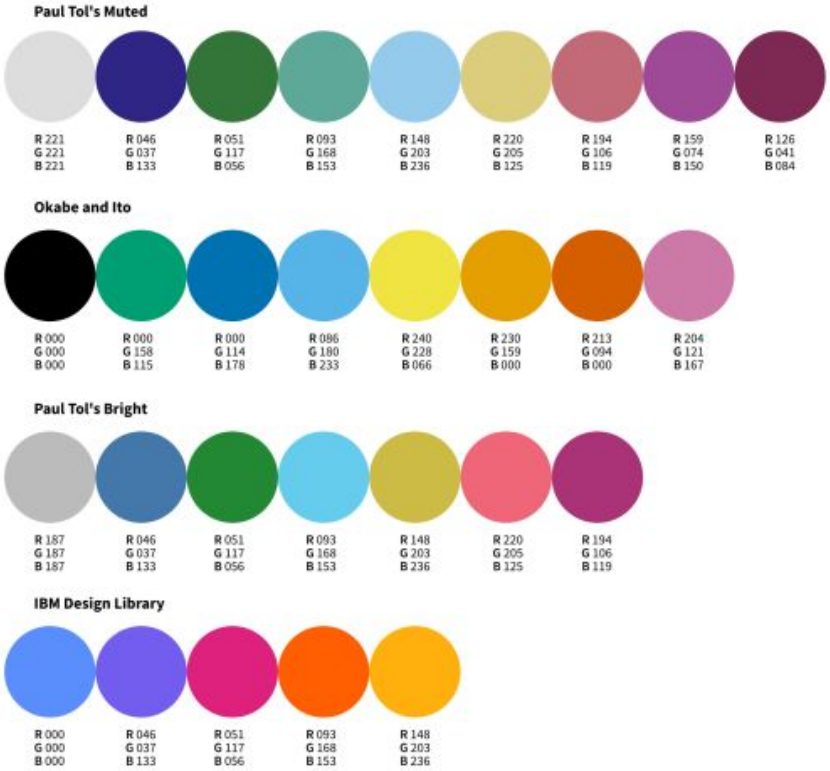
- Avoid using Too Many Variables within a single image
- Visualizing data through an Unsuitable or Incorrect visualization format
- While using Scales in Visualization,it is important to ensure that the scale is consistent
- avoid using colors with negligible contrast
- avoid using too many colors
- avoid using conventional colors to convey opposite meanings
- pay heed to the needs of people who might be colorblind

Colour-blind Friendly Palettes

- You should use color blind friendly schemes for all scientific publications.
- Avoid red, especially with green.
- Avoid rainbow colour maps.
- You can use pre-existing palettes.
- You can use colourblind visualization software like Color oracle that applies a full screen filter to visualize color blindness and grayscale.
- Use symbols and annotations and typography



Pre-existing colour-blind friendly palettes



For reference:
https://www.nceas.ucsb.edu/sites/default/files/2022-06/Colorblind%20Safe%20Color%20Schemes.pdf