

Data Visualization

Outline

0:10	What this workshop is about
0:15	Basics: form or content?
0:20	Types of charts
0:35	Instruments
0:50	Design principles
0:55	Color use
1:50	Visualization software

What problems do you have with visualizations?

1. I **do not know what I want to visualize** and I am not sure where to start
2. I know what I want to show, but **not sure how to do it**
3. My graphs look **bad**, no matter what I do 🙄
4. I have no problems! **I am very good** at visuals

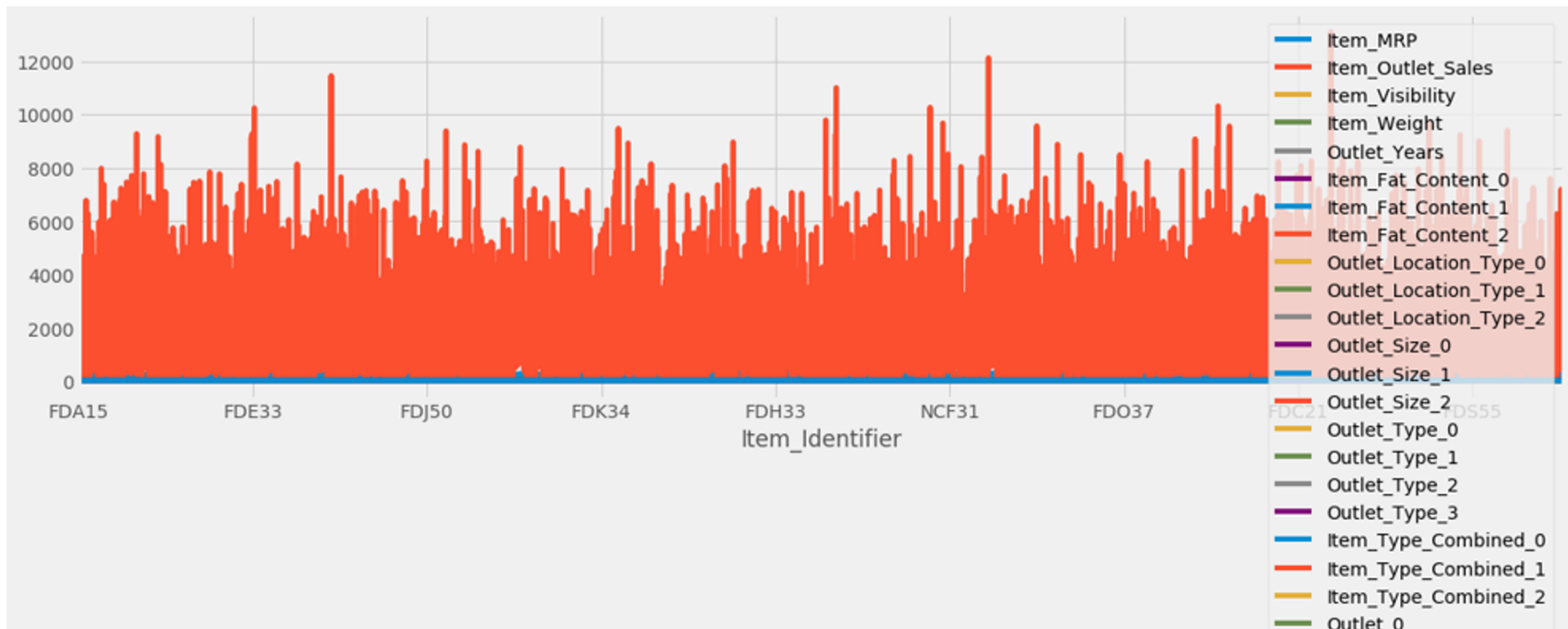
What software do you use for visualization?

1. Python or R
2. Tableau
3. Excel
4. Other (please, share in the chat!)

Learning objectives

1. Understand basic principles behind effective data visualization
2. Learn about graph types and visualization instruments
3. Be familiar with aesthetics features that make graphs clearer and *nicer!*

Do we need to learn how to visualize?



Graph should make a point

1. Hypothesis

What do you want to show?

2. Proof

How to emphasize that?

3. Explaining

What does the chart show?

BASICS

Form

Graph type
Features
Design

Content

Data format
Transformation
Context

Both are important!

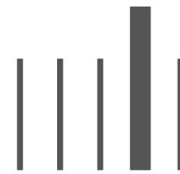
Preattentive Features



Orientation



Length



Width



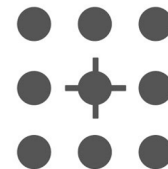
Size



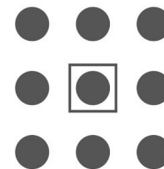
Shape



Curvature



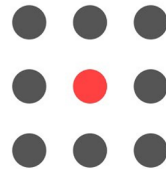
Added Marks



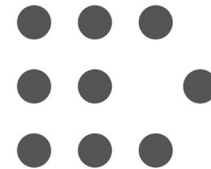
Enclosure



Contrast



Colour

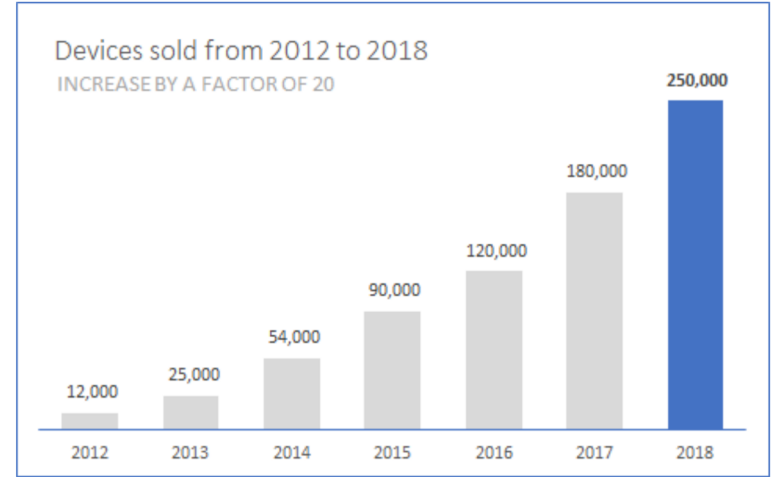
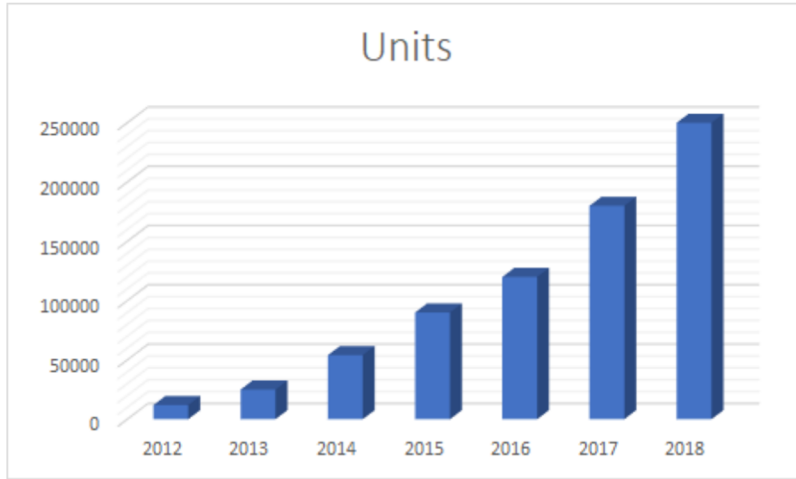


Position



Spatial Grouping

Preattentive features establish the hierarchy of information



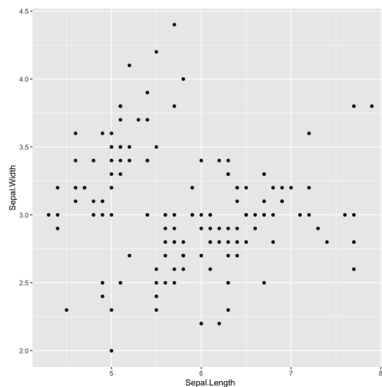
- Remove 3D
- Use direct labeling
- Remove grid (if appropriate)
- Use color meaningfully

If possible, declutter!

Choosing a Graph Type

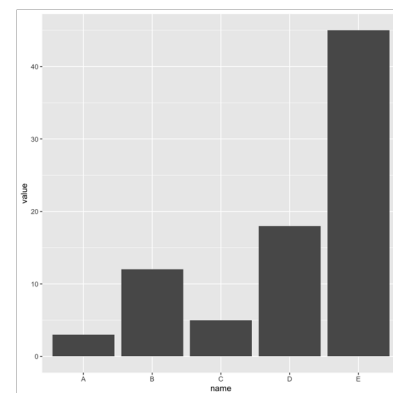
Consider:

- Purpose
- Variable types
- Dimensions

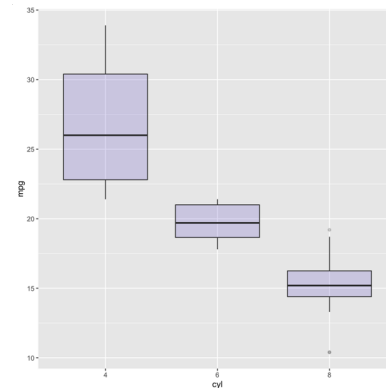


scatterplot

Hall of Fame

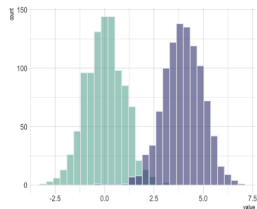


barplot

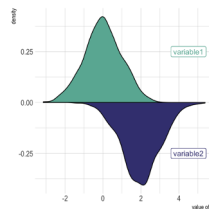


boxplot

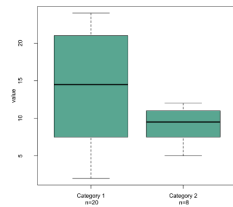
Distribution



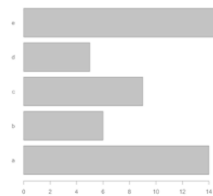
histogram



density plot



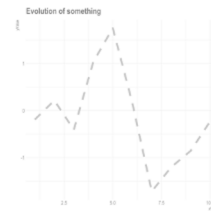
boxplot



barplot



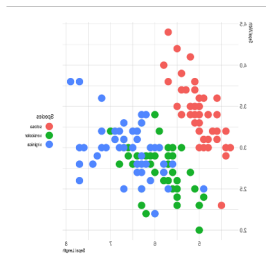
spider plot



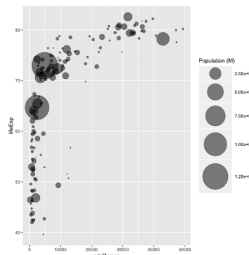
line

Comparison

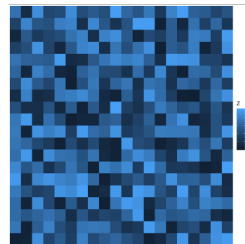
Relationship



scatterplot

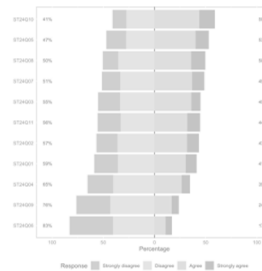


bubble plot

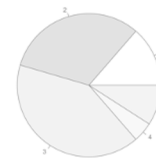


heatmap

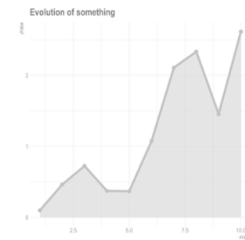
Part of the whole



barplot

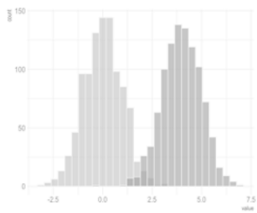


pie chart

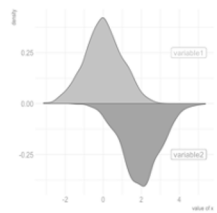


area plot

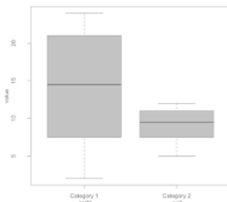
Distribution



histogram

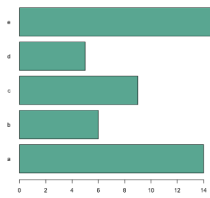


density plot

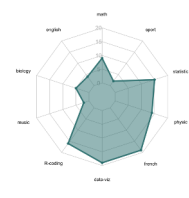


boxplot

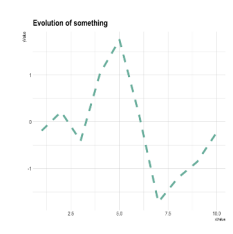
Comparison



barplot

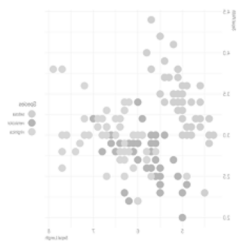


spider plot

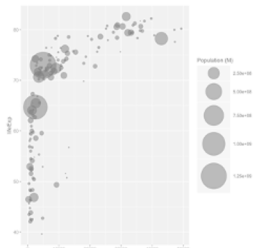


line

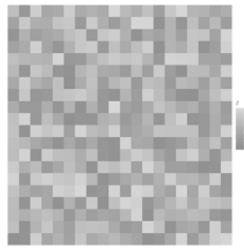
Relationship



scatterplot

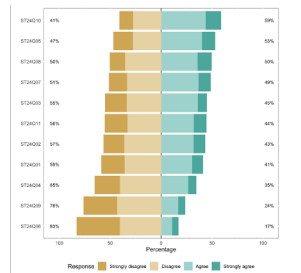


bubble plot

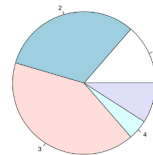


heatmap

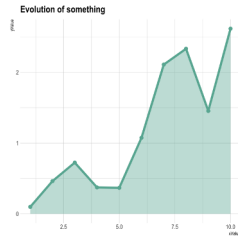
Part of the whole



barplot



pie chart



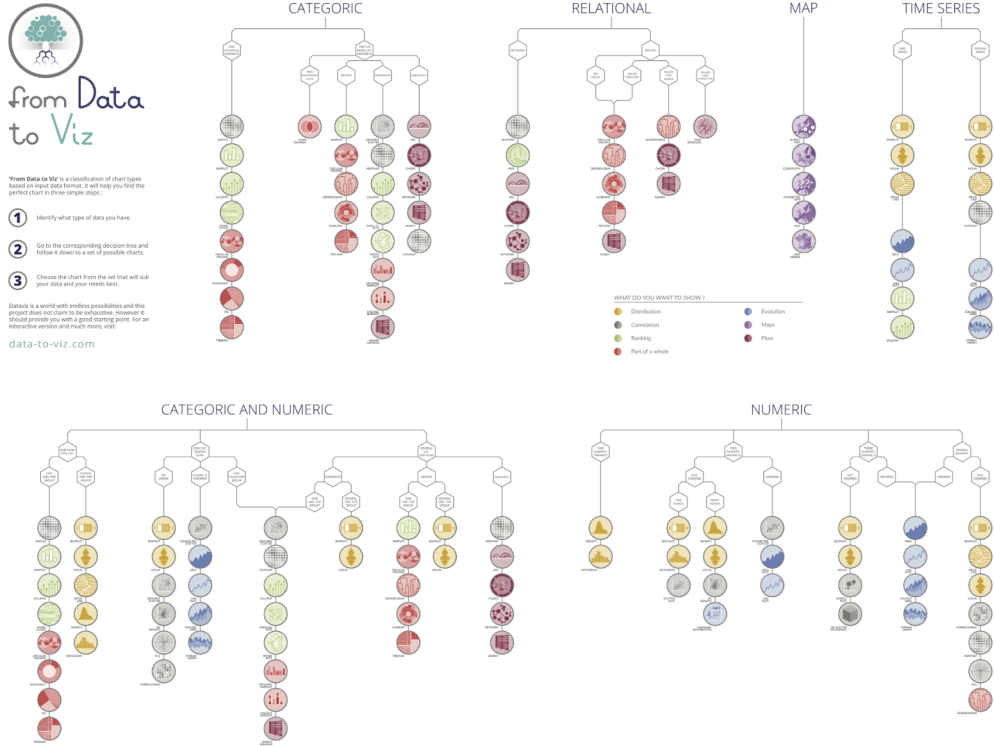
area plot

Graph Types and Data Types

Numeric/Categorical

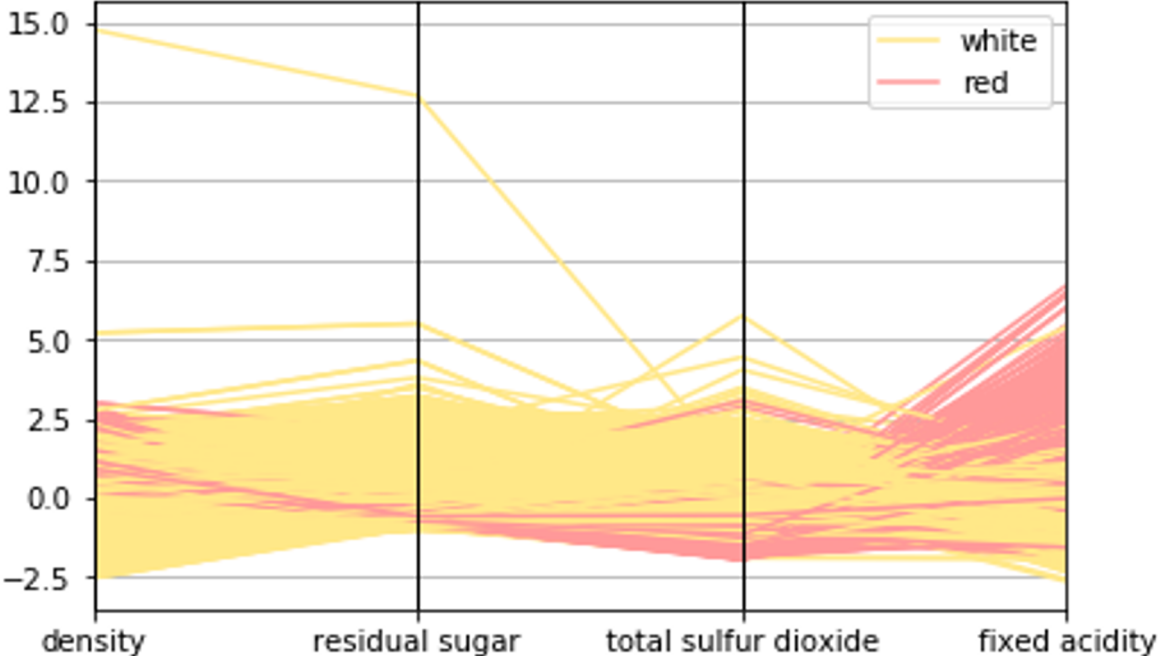
Is it ordered?

How many observations per group?



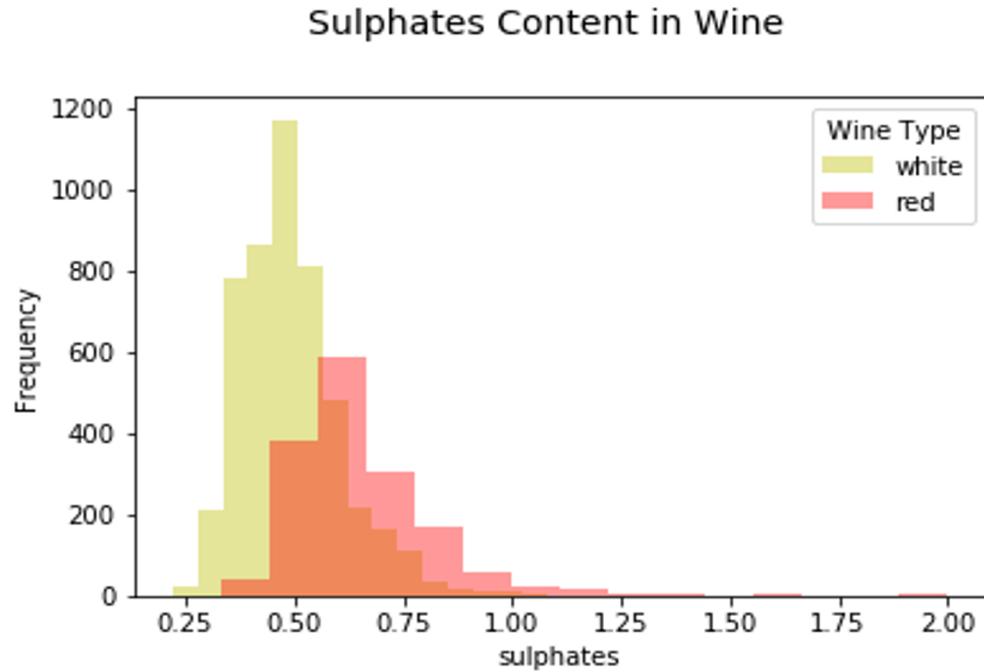
<https://www.data-to-viz.com/>

Multiple dimensions I



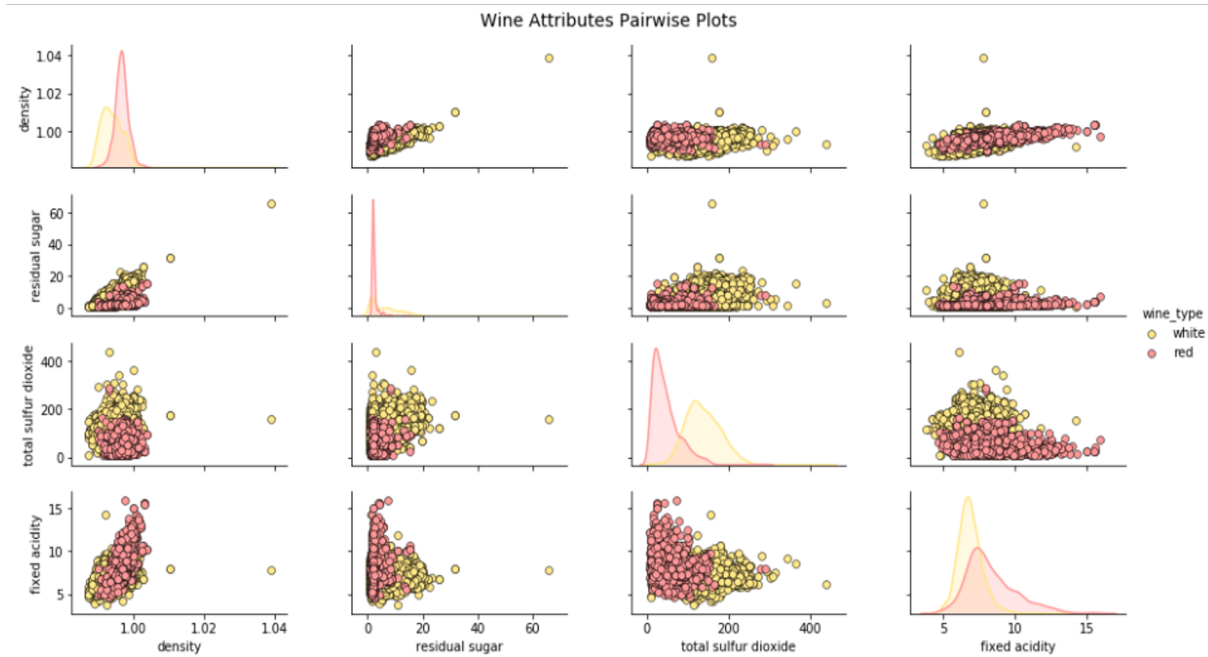
graph types

Multiple dimensions II



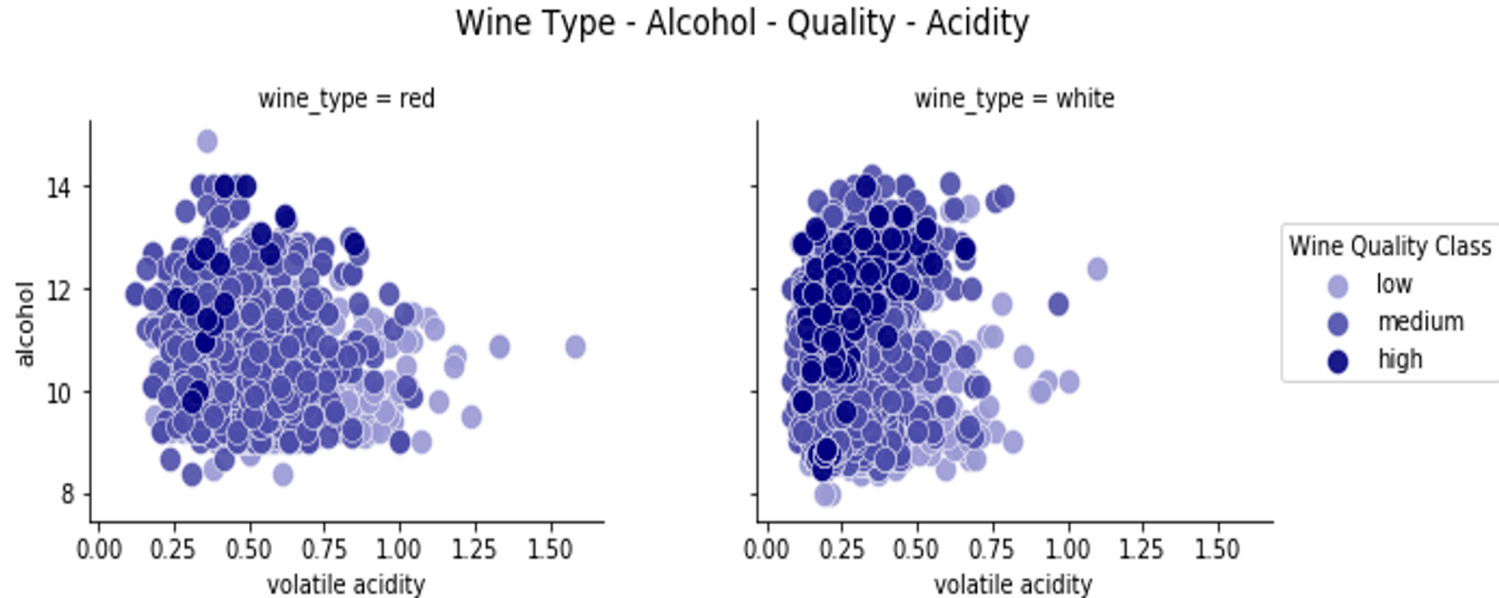
colours

Multiple dimensions III



pairwise correlations

Multiple dimensions IV

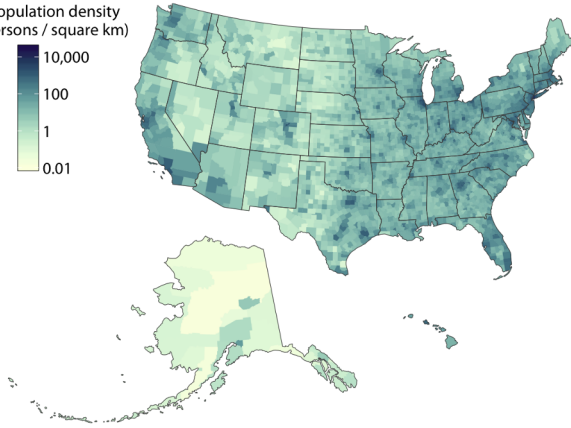
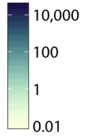


facets

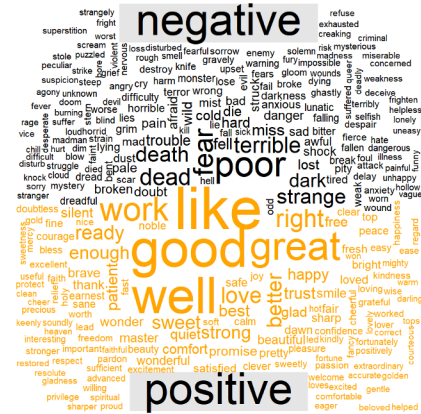
Other Graph Types

geospatial data

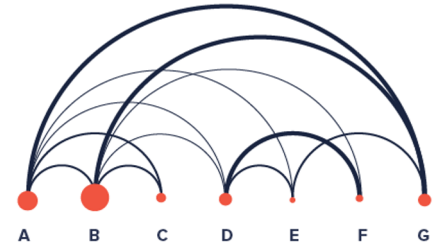
population density
(persons / square km)



wordcloud



connections



Having good data is more important than choosing a right graph type

Ideally, data should be:

- ✓ High volume
- ✓ Historical
- ✓ Consistent
- ✓ Clean
- ✓ Clear
- ✓ Richly segmented

Long vs Wide formats

“Long” format

country	year	metric
x	1960	10
x	1970	13
x	2010	15
y	1960	20
y	1970	23
y	2010	25
z	1960	30
z	1970	33
z	2010	35

“Wide” format

country	yr1960	yr1970	yr2010
x	10	13	15
y	20	23	25
z	30	33	35

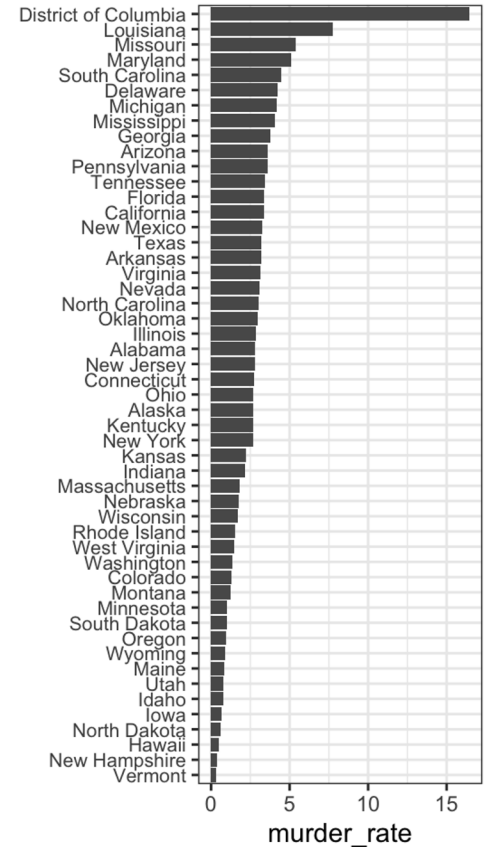
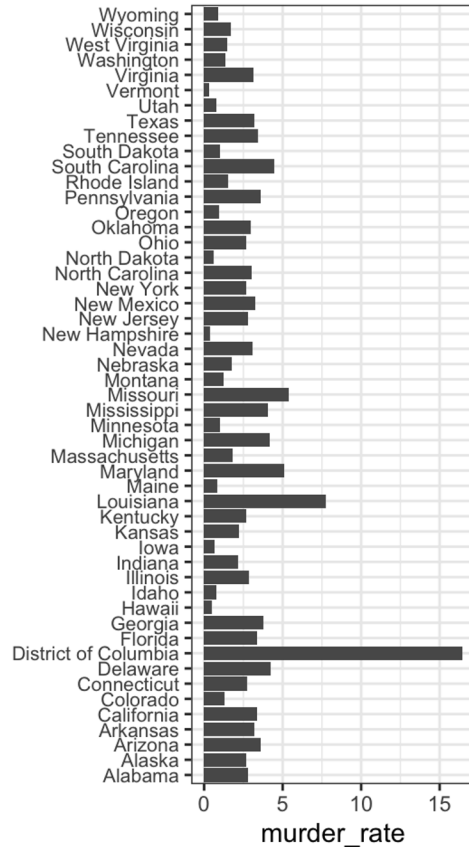
INSTRUMENTS

Most of the data visualization software allows you to:

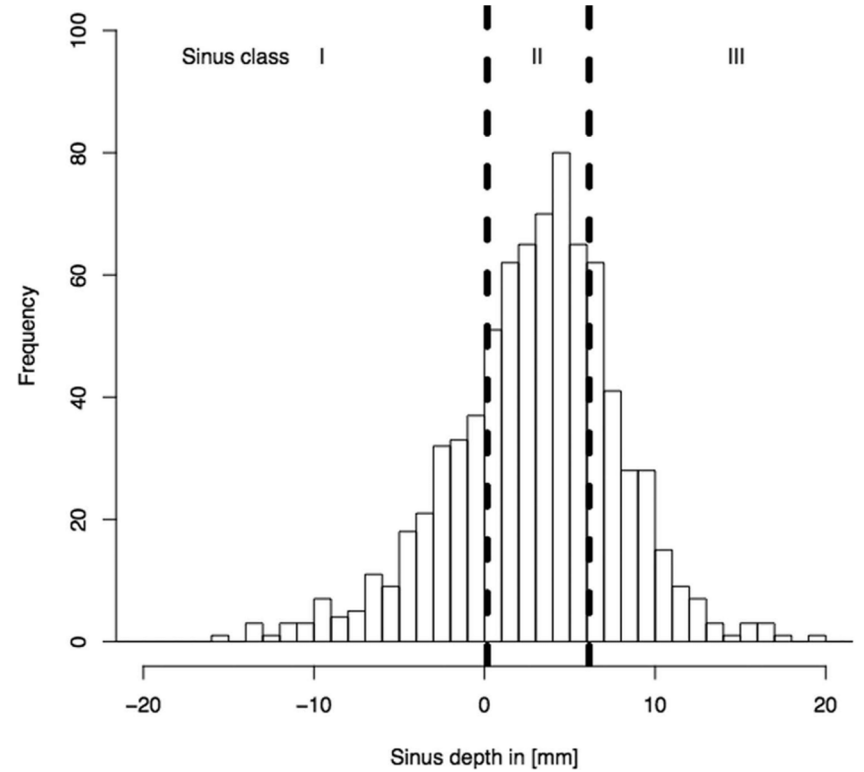
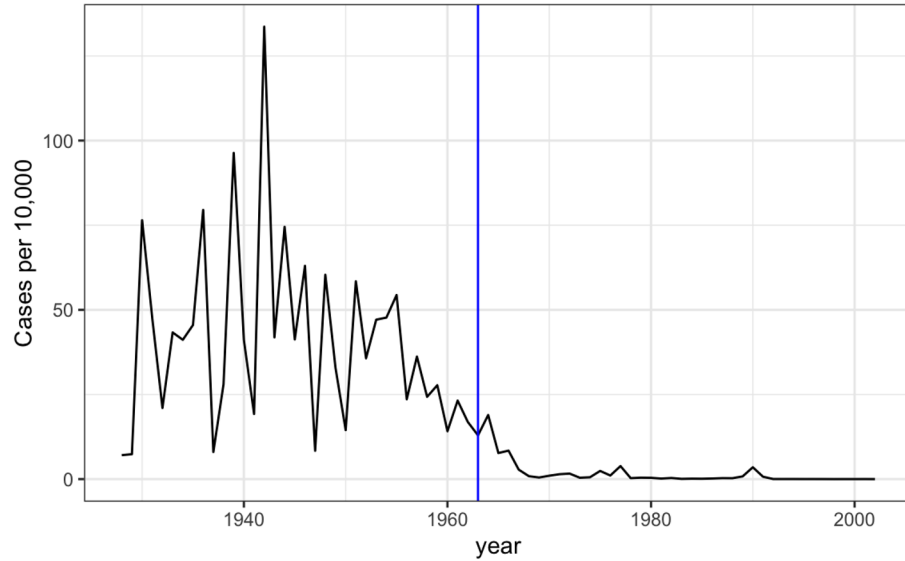
- **Aggregate**
- **Filter**
- **Compare**
- **Annotate**

and much more!

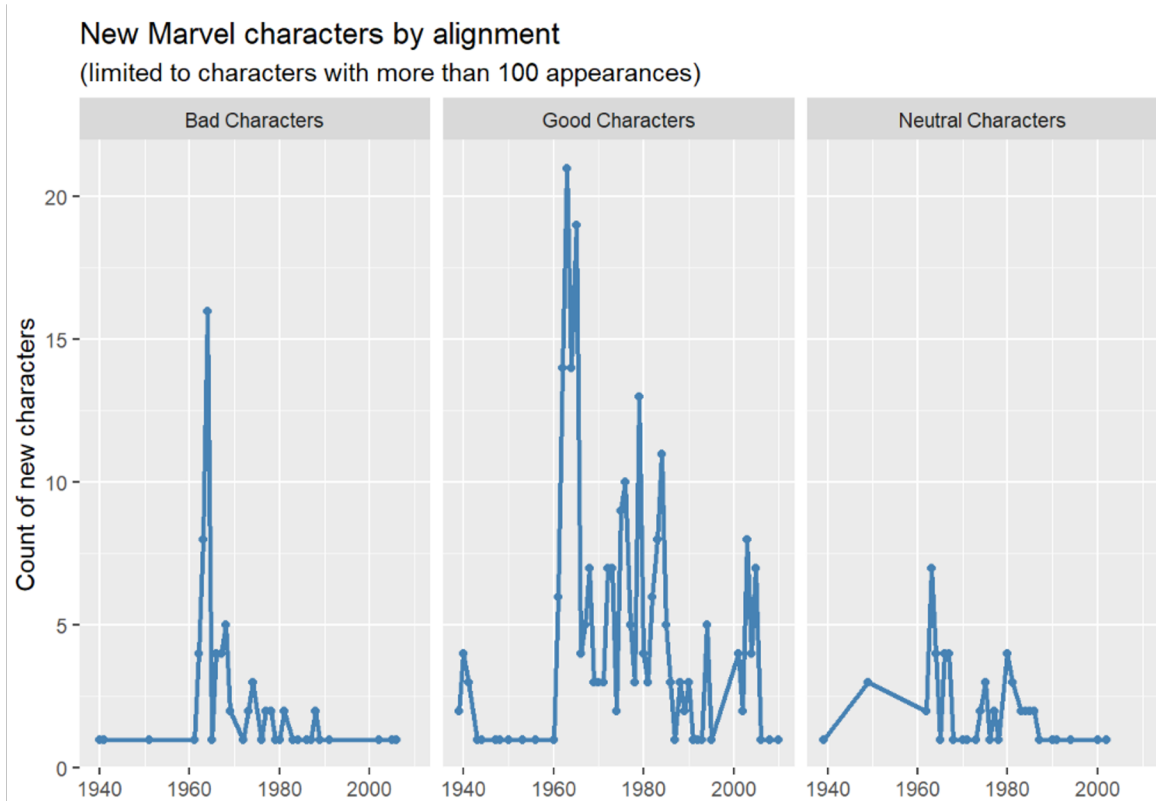
Data ordering



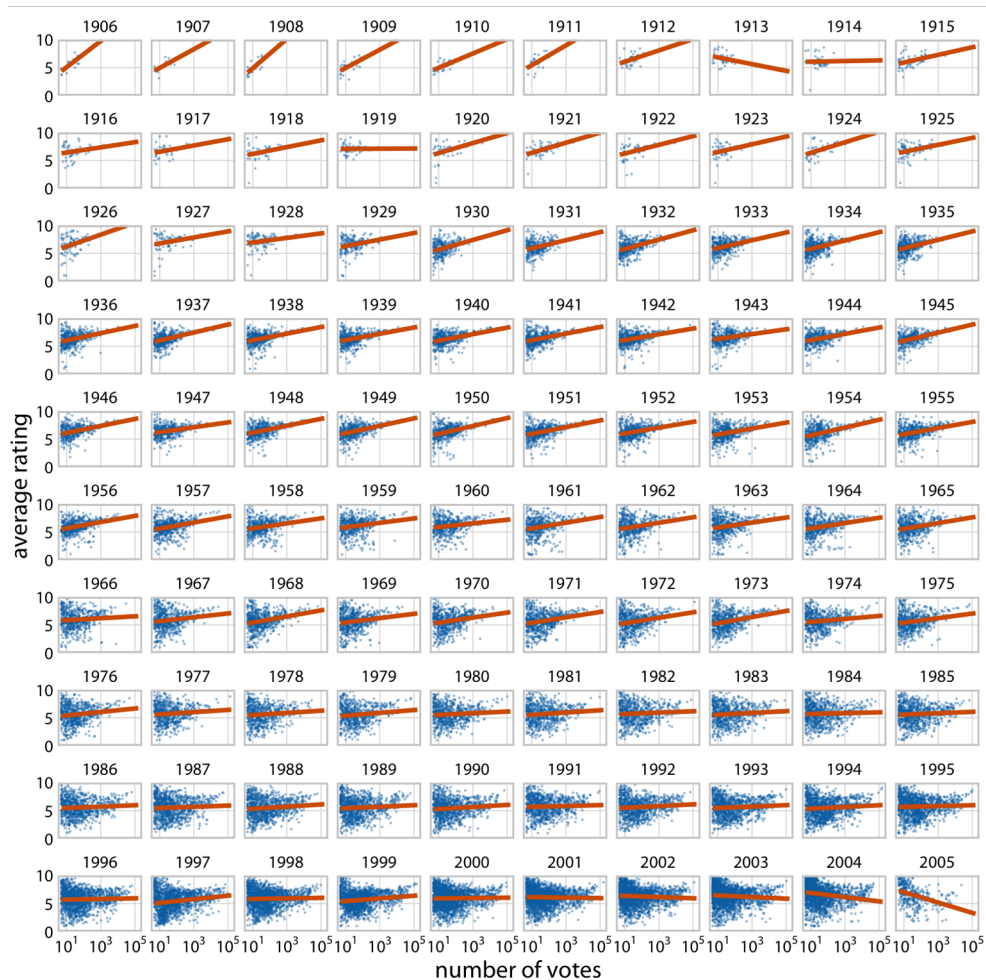
Reference Lines and Regions



Facets

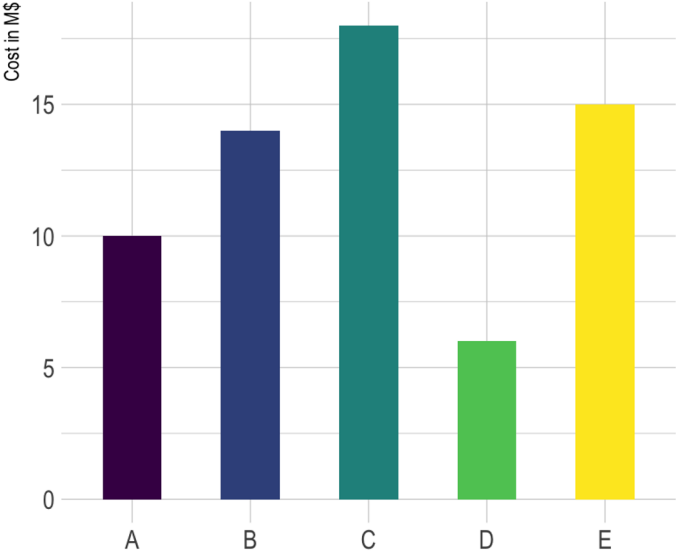


Useful for
multidimensional
or large scale data

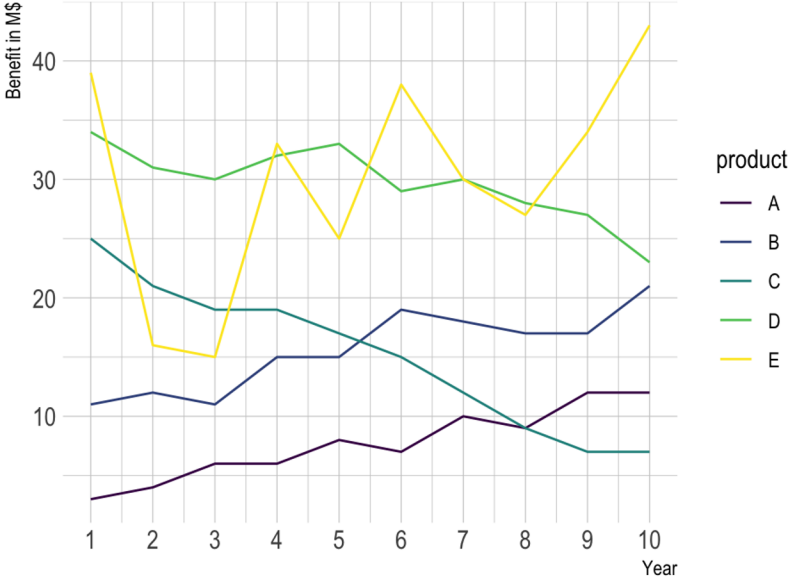


Multiple Views

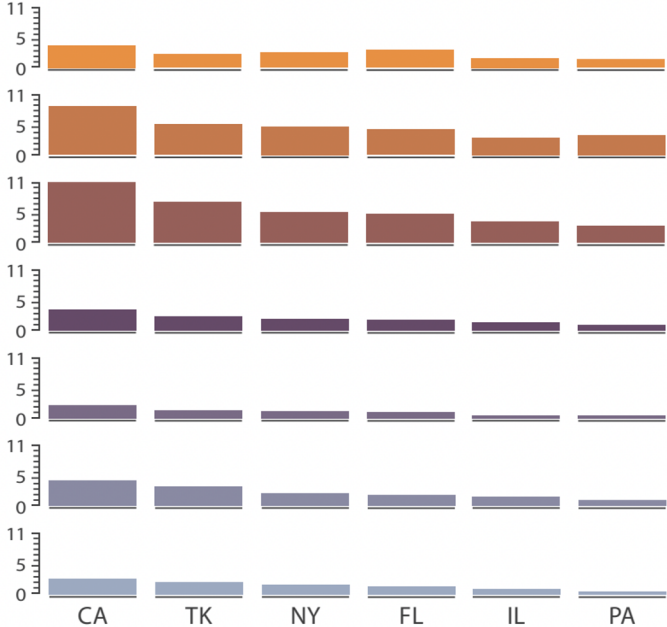
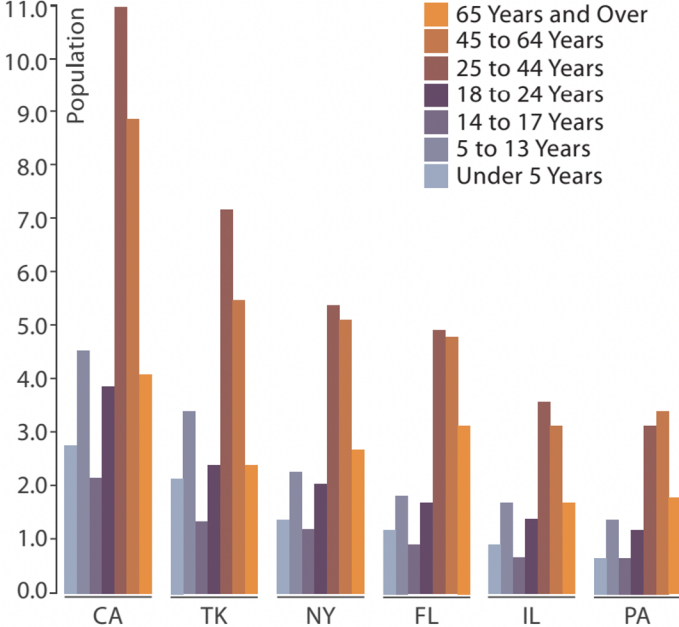
How much the products cost



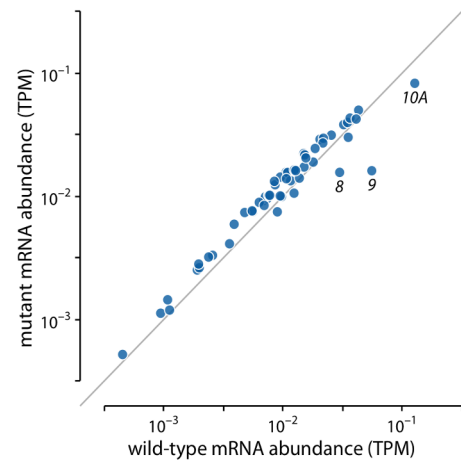
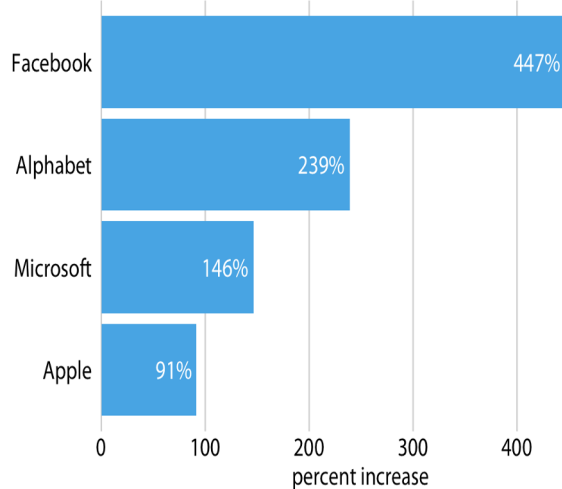
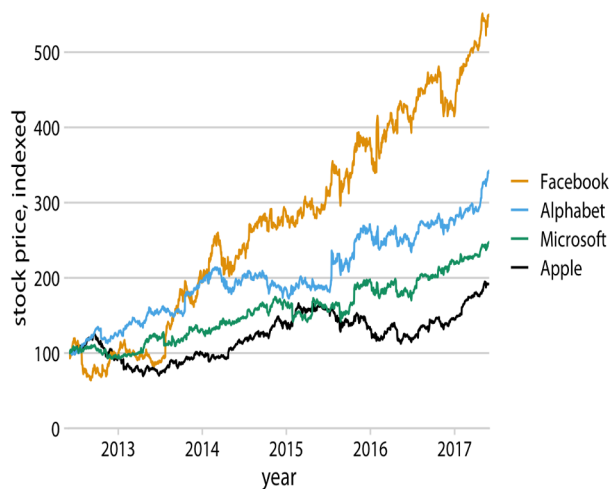
How much the products make



Multiple Views



Context

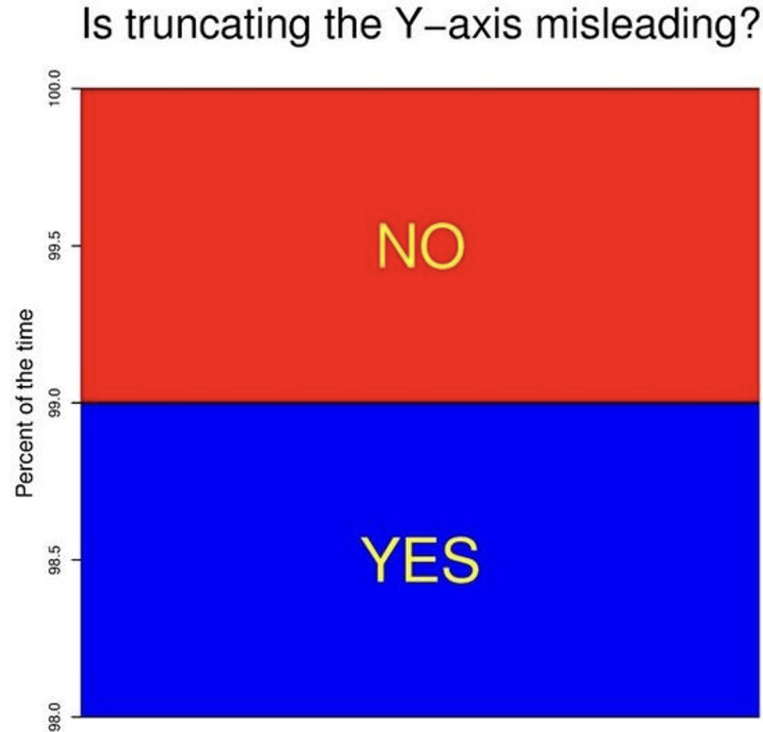


Perpendicular grid lines are
the most useful

For paired data,
use a diagonal line

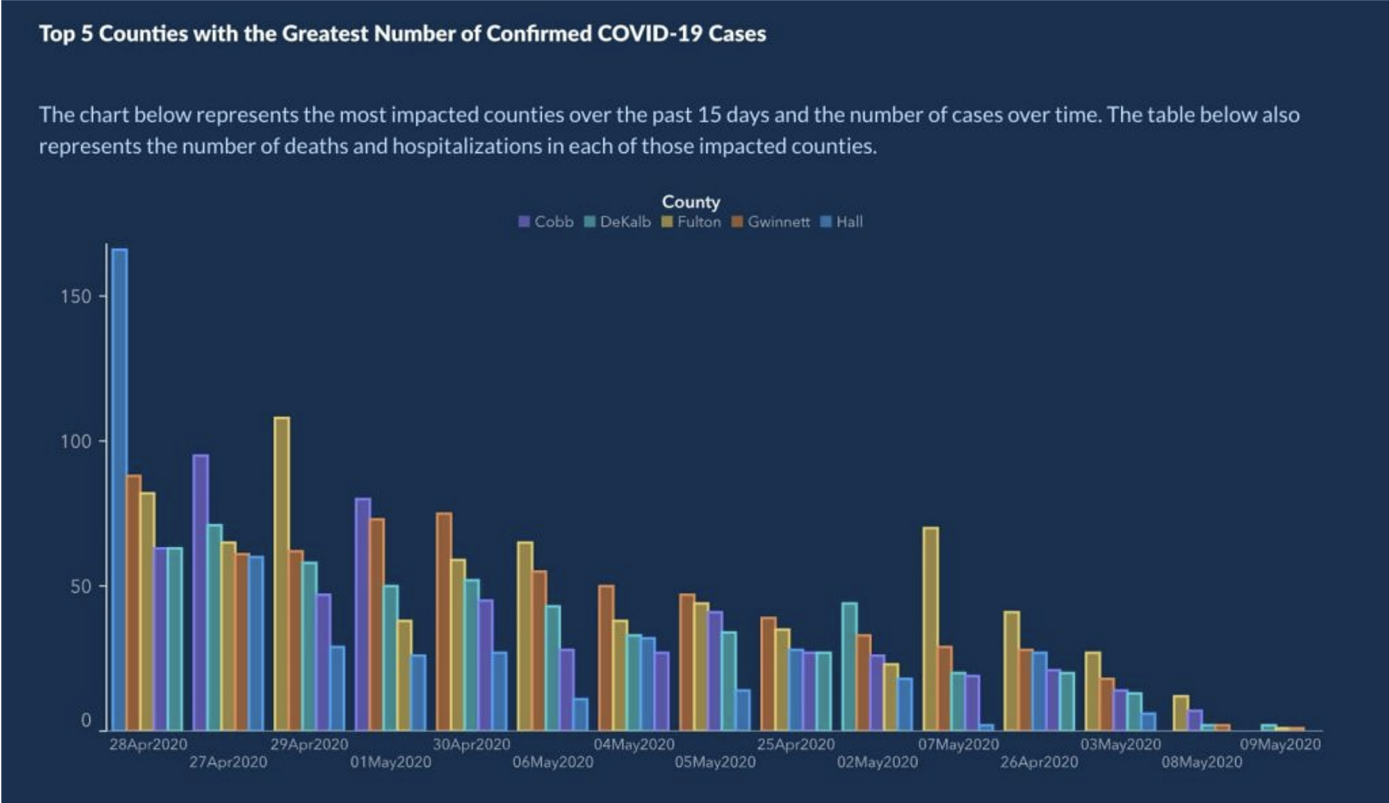
DESIGN PRINCIPLES

Cutting Y-scales



If cutting Y-axis (for example, when using Likert scale data), use **points** instead of bars

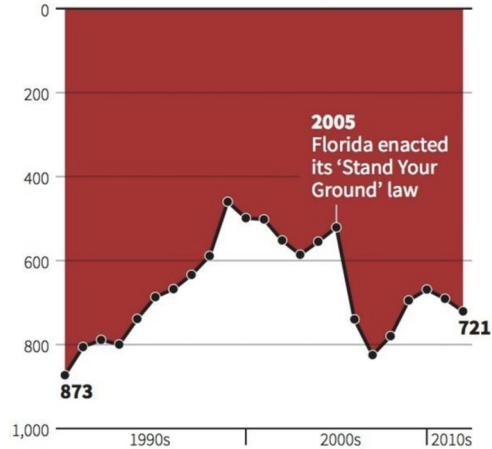
Use reasonable X-axis



flipped

Gun deaths in Florida

Number of murders committed using firearms

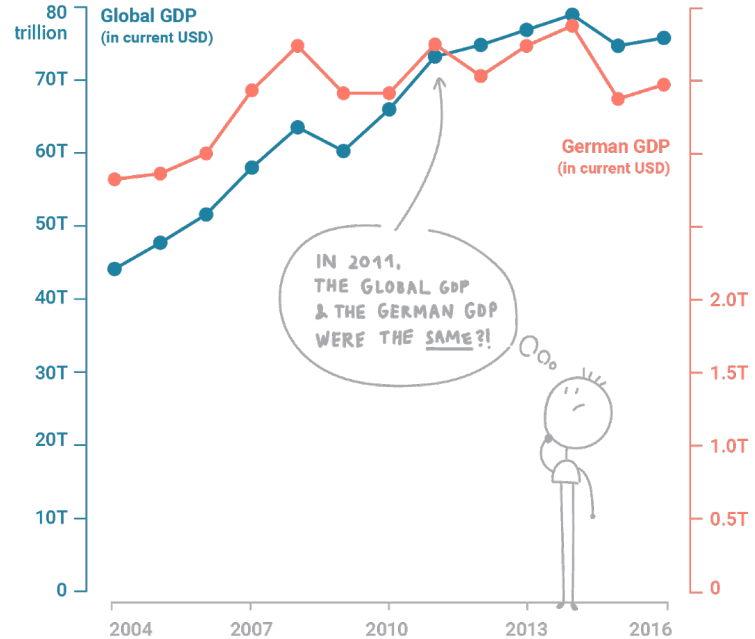


Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS

dual



Avoid misleading axes!

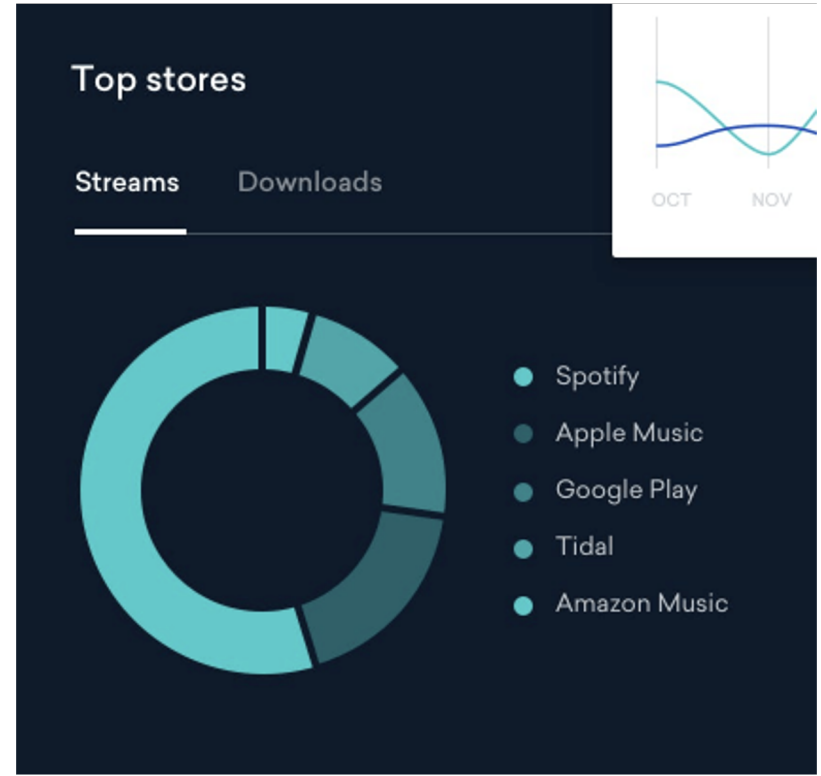
Colour Use

Gradient and diverging palettes

Colorblind-friendly

Do not add more than 8 colours

Use gray for context



Creating Palettes

SANFORD AND SELNICK

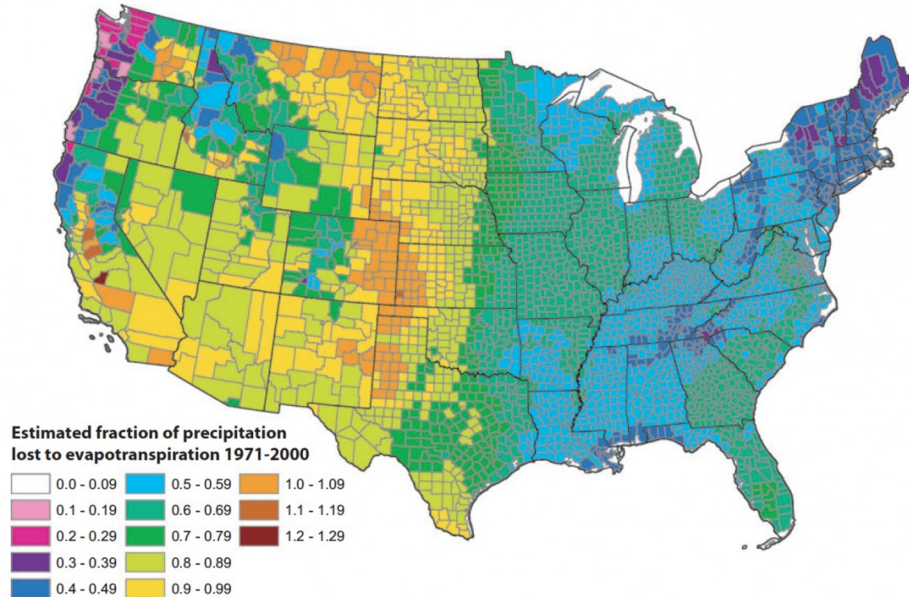


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/ P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

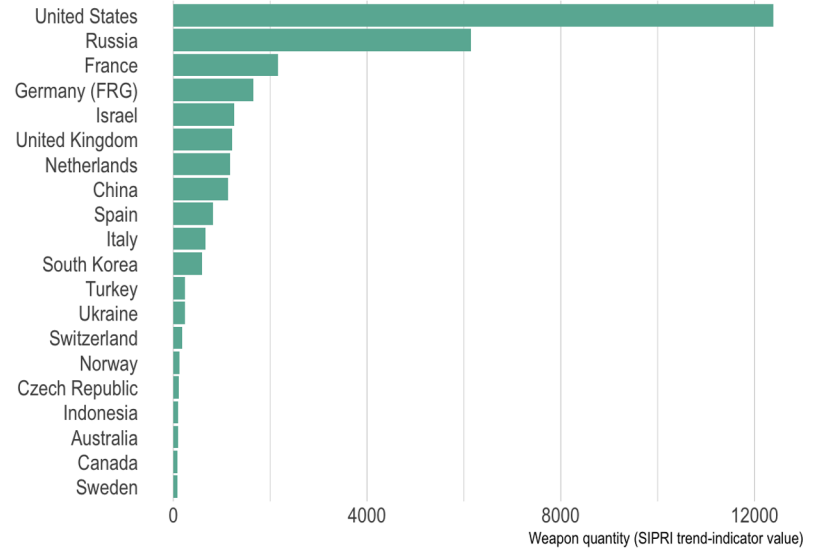
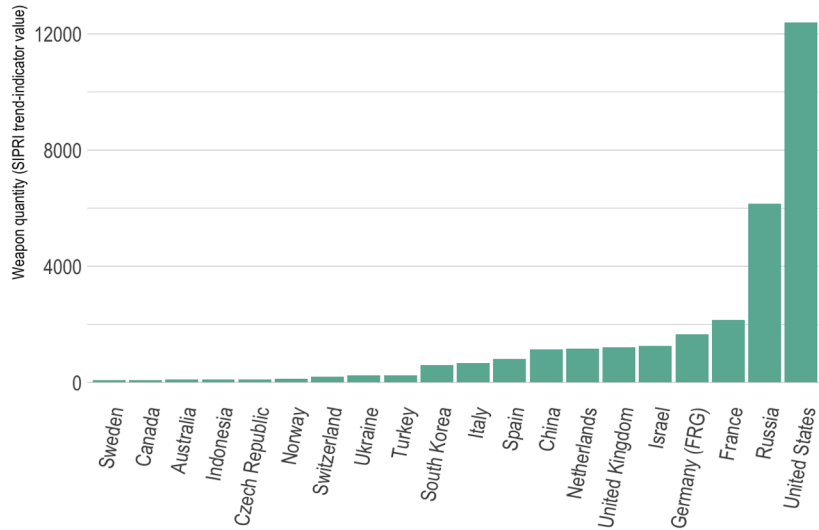
Use meaningful colors

Use color generators

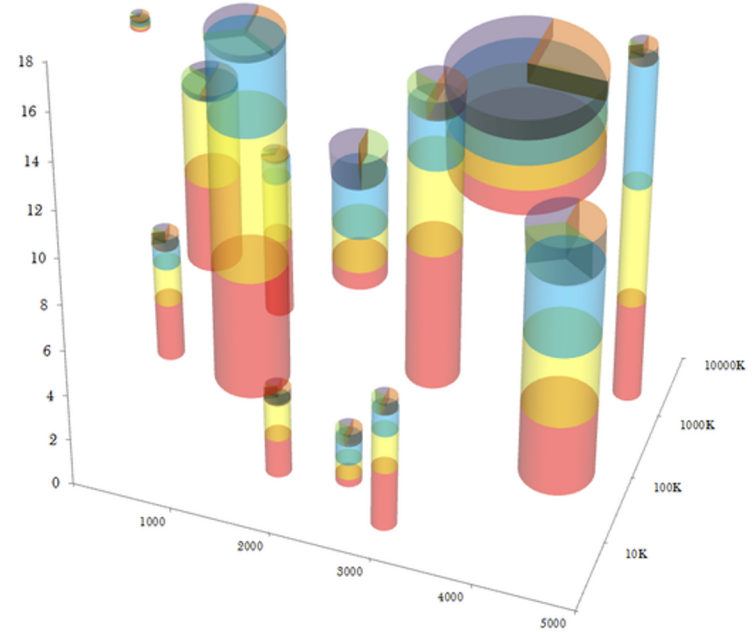
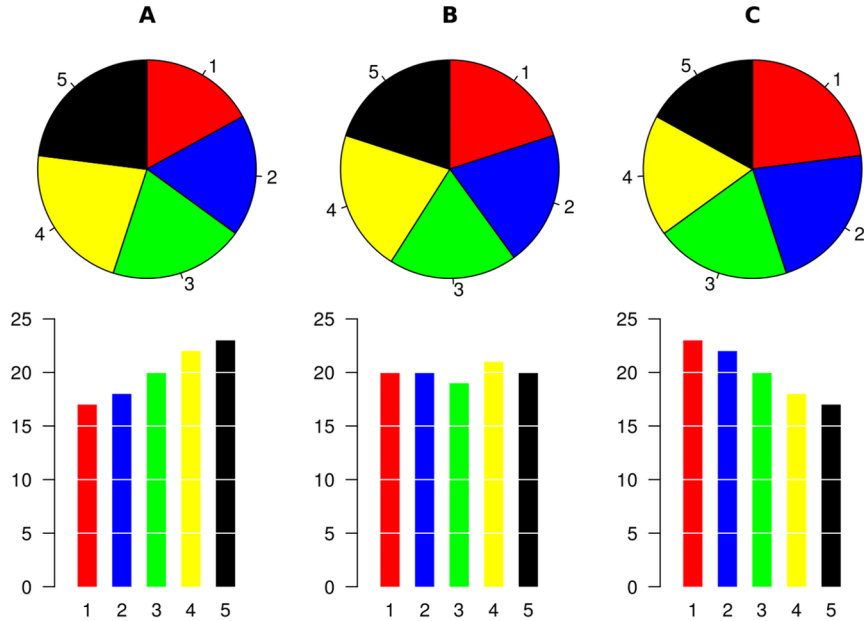
Avoid pure and bright colours

Avoid rainbow palette

Use horizontal bar chart



Avoid pie charts and 3D



Tables

1. Remove style and borders
2. **Do not use vertical lines**, remove all grids if possible
3. **Left align** your text, BUT **right align** your numbers
4. Align headers correspondingly
5. Group and separate

Company	Last Trade	Trade Time	Change	Prev Close
GOOG Google Inc.	597.74	12:12PM	14.81 (2.54%)	582.93
AAPL Apple Inc.	378.94	12:22PM	5.74 (1.54%)	373.20
AMZN Amazon.com Inc.	191.55	12:23PM	3.16 (1.68%)	188.39
ORCL Oracle Corporation	31.15	12:44PM	1.41 (4.72%)	29.74
MSFT Microsoft Corporation	25.50	12:27PM	0.66 (2.67%)	24.84
CSCO Cisco Systems, Inc.	18.65	12:45PM	0.97 (5.49%)	17.68
YHOO Yahoo! Inc.	15.81	12:25PM	0.11 (0.67%)	15.70

SOFTWARE

Data Visualization Tools



Many features
Easy to learn & use
No data cleaning
Data security



Data cleaning
Popular
Confusing
Lacks customization



Free
Data exploration
Requires learning
Debugging