University of Warsaw
MIM
Faculty
of Mathematics, Informatics
and Mechanics

Uniwersytet Warszawski

Data analysis and visualization (DAV)

Lecture 04

Łukasz P. Kozłowski

Warsaw, 2025

l.kozlowski@mimuw.edu.pl

Data analysis and visualization (DAV)

Lecture 04
Plots

Łukasz P. Kozłowski

Warsaw, 2025

l.kozlowski@mimuw.edu.pl

Plots

Plots

COCALC
kaggle
Apache Zeppelin
databricks
Notebooks
jupyter
HEX
Notebooks
Deepnote
Saturn Cloud
Azure Notebooks

Plots

COCALC
kaggle
Apache Zeppelin
databricks
Notebooks
jupyter
X
HEX
Notebooks
Deepnote
Azure Notebooks
Saturn Cloud

Plots

We still do not use Excel

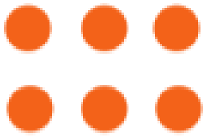
Prefer „one script - one plot” methodology

Build many simple graphs first!

At first you do not care about labels, colors ,etc. (quick & dirty)

Fully Explore Your Data First

KNOW YOUR DATA



QUANTITATIVE
Data that can be counted or measured; all values are numerical.

DISCRETE
Numerical data that has a finite number of possible values. Example: Number of employees in the office.

CONTINUOUS
Data that is measured and has a value within a range. Example: Rainfall in a year.

CATEGORICAL
Data that can be sorted according to group or category. Example: Types of products sold.

BUILD

Static

BEFORE

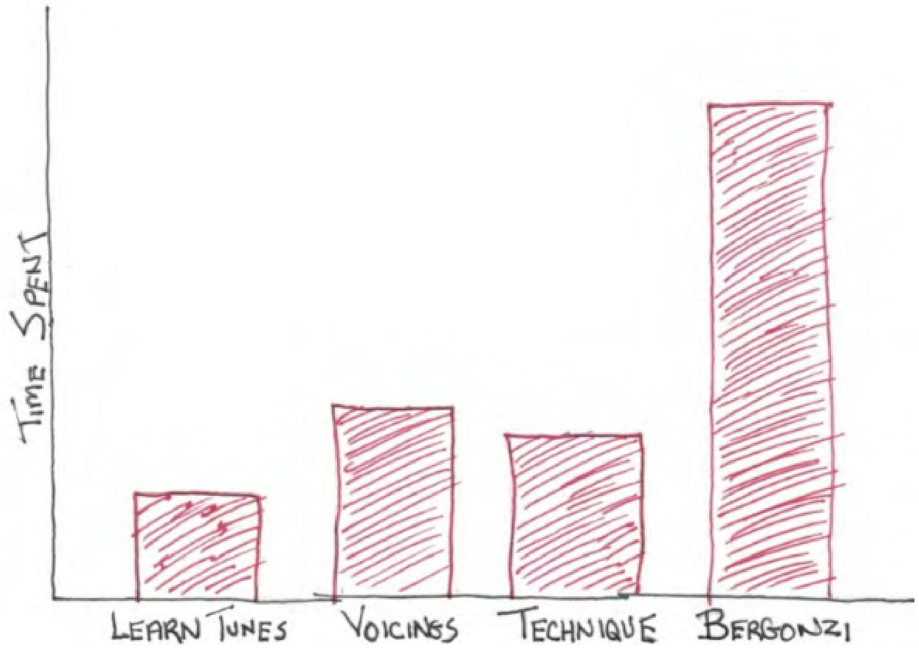
Interactive

BEFORE

Animated

Before you start „coding” the plot it is always good to start the design with …

Before you start „coding” the plot it is always good to start the design with paper and pencil/pen

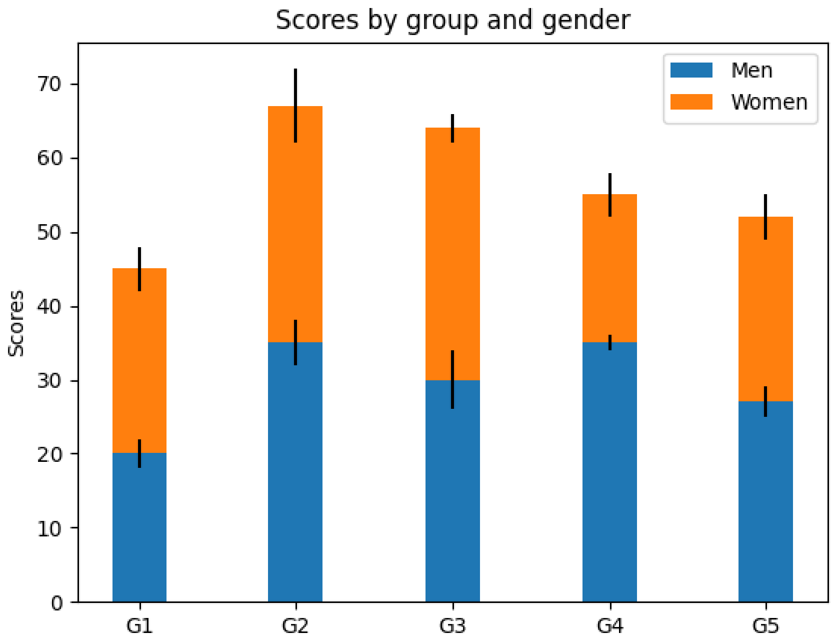


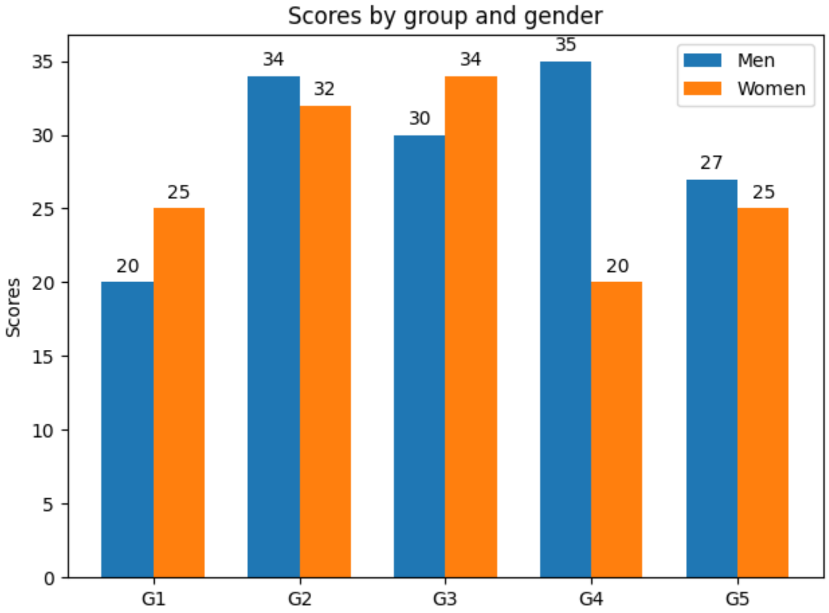
Choose The Right Tool for the Job

Choose The Right Tool for the Job

Welcome in the Visualization ZOO

Bar chart



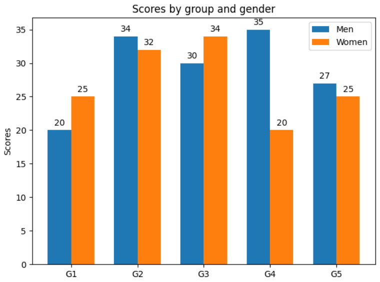


Stacked bar chart

Grouped bar chart with labels

Bar chart

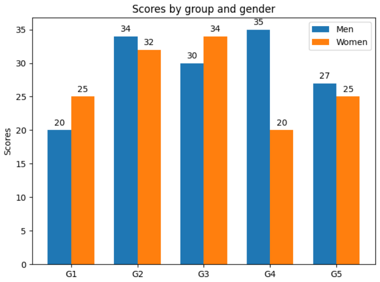
- use if number of categories is quite small (up to 5, not more than 7)



Bar chart

- use if number of categories is quite small (up to 5, not more than 7)

- use if one of your data dimensions is time (years, quarters, months, weeks, days, or hours)

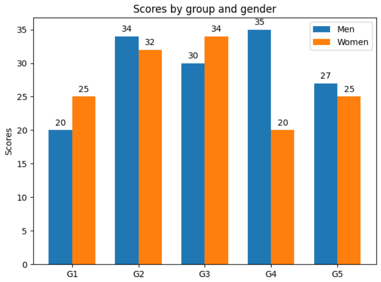


Bar chart

- use if number of categories is quite small (up to 5, not more than 7)

- use if one of your data dimensions is time (years, quarters, months, weeks, days, or hours)

- always set time dimension on the horizontal axis.



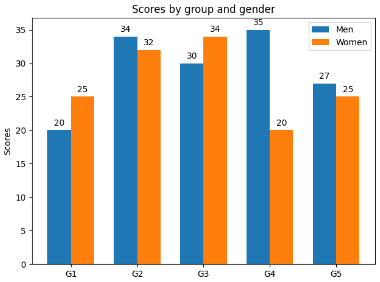
Bar chart

- use if number of categories is quite small (up to 5, not more than 7)

- use if one of your data dimensions is time (years, quarters, months, weeks, days, or hours)

- always set time dimension on the horizontal axis.

- time should always run from left to right (never from top to bottom)



Bar chart

- use if number of categories is quite small (up to 5, not more than 7)

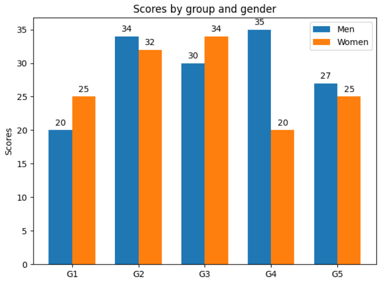
- use if one of your data dimensions is time (years, quarters, months, weeks, days, or hours)

- always set time dimension on the horizontal axis.

- time should always run from left to right (never from top to bottom)

- the numerical axis must start at zero.

Our eyes are very sensitive to the height of columns, and we can draw inaccurate conclusions when those bars are truncated



Bar chart

- use if number of categories is quite small (up to 5, not more than 7)

- use if one of your data dimensions is time (years, quarters, months, weeks, days, or hours)

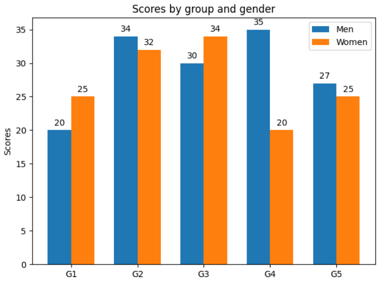
- always set time dimension on the horizontal axis.

- time should always run from left to right (never from top to bottom)

- the numerical axis must start at zero.

Our eyes are very sensitive to the height of columns, and we can draw inaccurate conclusions when those bars are truncated

- avoid using pattern lines or fills



Bar chart

- use if number of categories is quite small (up to 5, not more than 7)

- use if one of your data dimensions is time (years, quarters, months, weeks, days, or hours)

- always set time dimension on the horizontal axis.

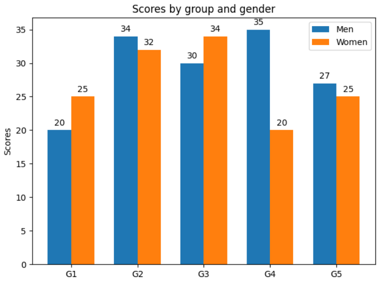
- time should always run from left to right (never from top to bottom)

- the numerical axis must start at zero.

Our eyes are very sensitive to the height of columns, and we can draw inaccurate conclusions when those bars are truncated

- avoid using pattern lines or fills

- use border only for highlights



Bar chart

- use if number of categories is quite small (up to 5, not more than 7)

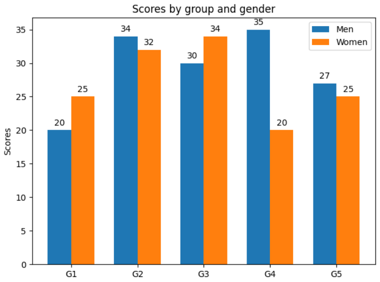
- use if one of your data dimensions is time (years, quarters, months, weeks, days, or hours)

- always set time dimension on the horizontal axis.

- time should always run from left to right (never from top to bottom)

- the numerical axis must start at zero.

Our eyes are very sensitive to the height of columns, and we can draw inaccurate conclusions when those bars are truncated



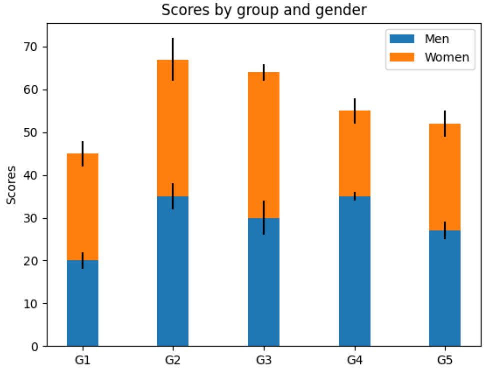
- avoid using pattern lines or fills

- use border only for highlights

- use to show trends only if there are a low number of data points (< 20) and if every data point has a clearly-visible value

Bar chart

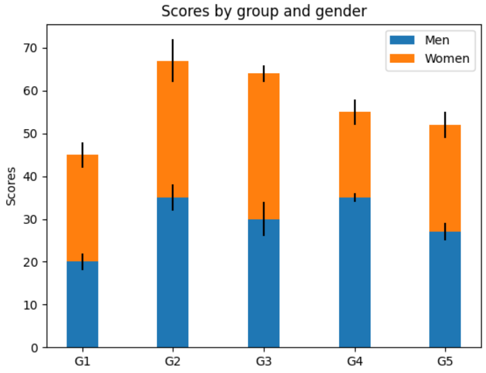
- use stacked column charts to show a composition



Bar chart

- use stacked column charts to show a composition

- do not use too many composition items (max 4)

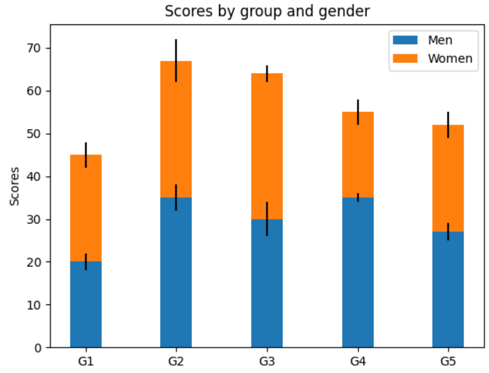


Bar chart

- use stacked column charts to show a composition

- do not use too many composition items (max 4)

- make sure the composing parts are relatively similar in size



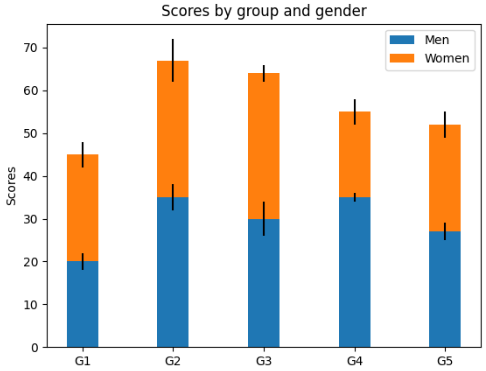
Bar chart

- use stacked column charts to show a composition

- do not use too many composition items (max 4)

- make sure the composing parts are relatively similar in size

- keep axis fixed (or at least test both possibilities)



Bar chart

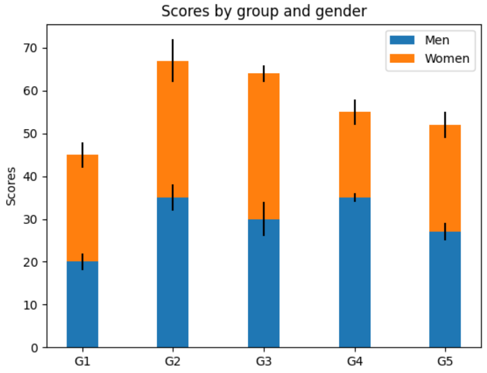
- use stacked column charts to show a composition

- do not use too many composition items (max 4)

- make sure the composing parts are relatively similar in size

- keep axis fixed (or at least test both possibilities)

- stacked bars are not good for comparison or relationship analysis



Bar chart

- use stacked column charts to show a composition

- do not use too many composition items (max 4)

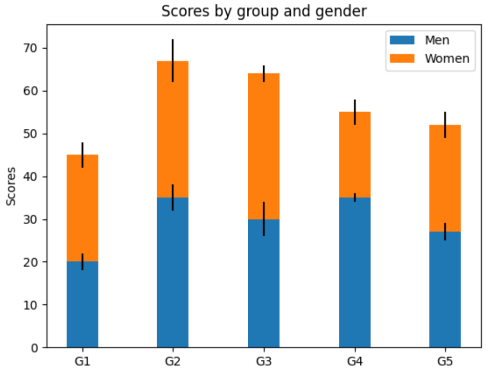
- make sure the composing parts are relatively similar in size

- keep axis fixed (or at least test both possibilities)

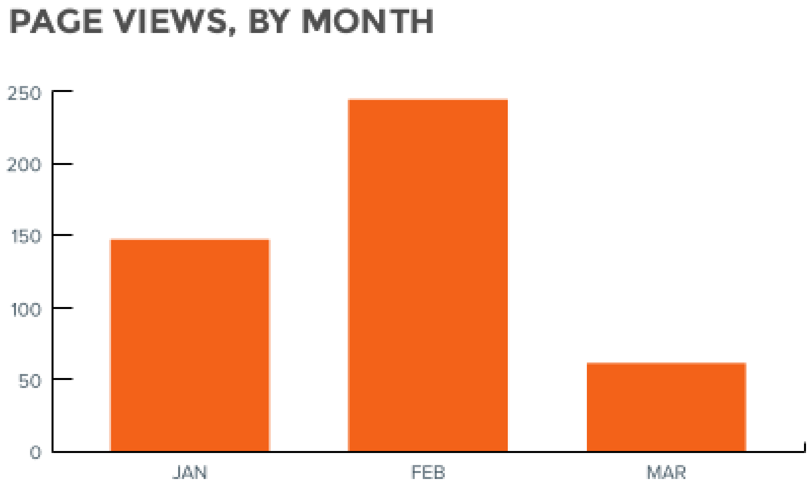
- stacked bars are not good for comparison or relationship analysis

- good for simple comparison of the quantitative values of subcategories

Example: Number of visitors to various websites

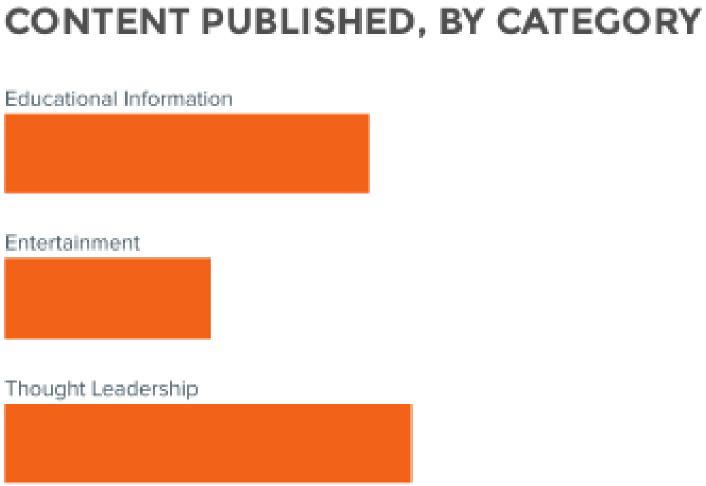


Bar chart



VERTICAL
(COLUMN CHART)

Best used for chronological data (time-series should always run left to right), or when visualizing negative values below the x-axis.

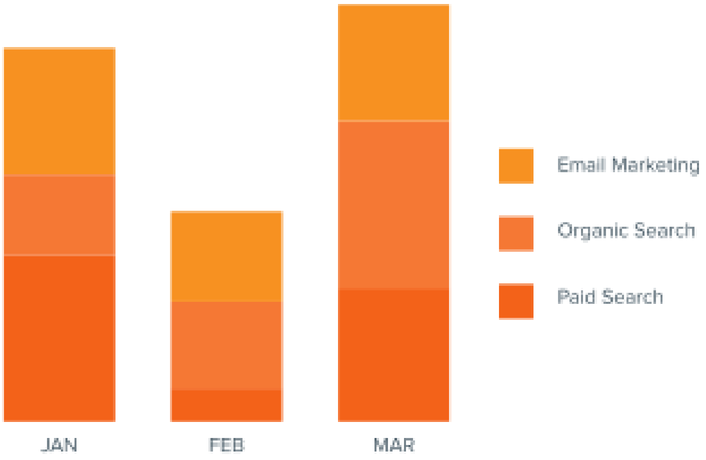


HORIZONTAL

Best used for data with long category labels.

Bar chart

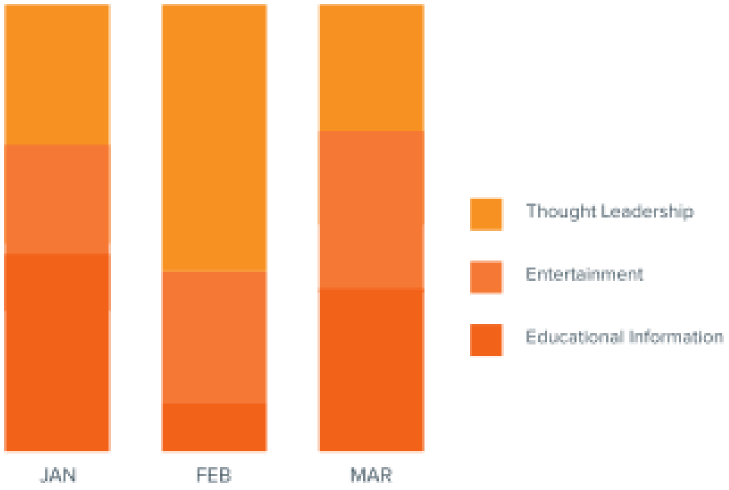
MONTHLY TRAFFIC, BY SOURCE



STACKED

Best used when there is a need to compare multiple part-to-whole relationships. These can use discrete or continuous data, oriented either vertically or horizontally.

PERCENTAGE OF CONTENT PUBLISHED, BY MONTH



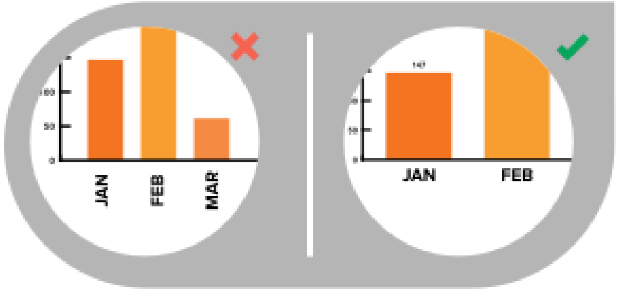
100% STACKED

Best used when the total value of each category is unimportant and percentage distribution of subcategories is the primary message.

Bar chart

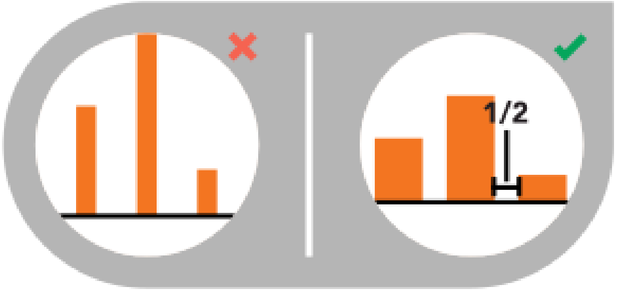
BAR CHART

DESIGN BEST PRACTICES



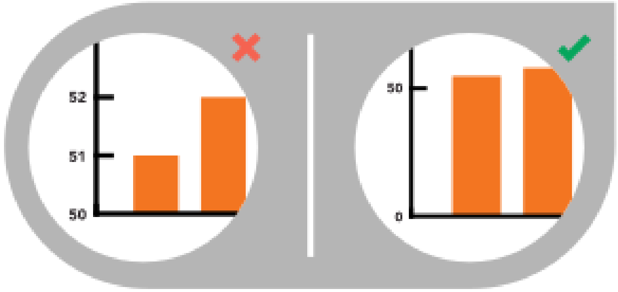
USE HORIZONTAL LABELS

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



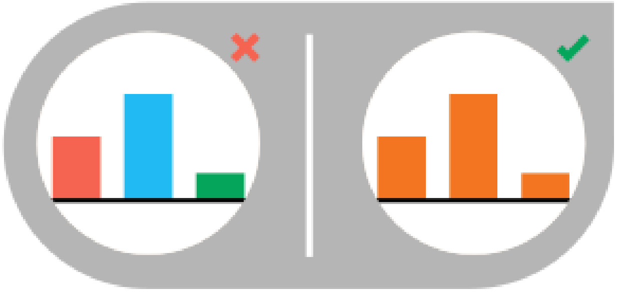
SPACE BARS APPROPRIATELY

Space between bars should be ½ bar width.



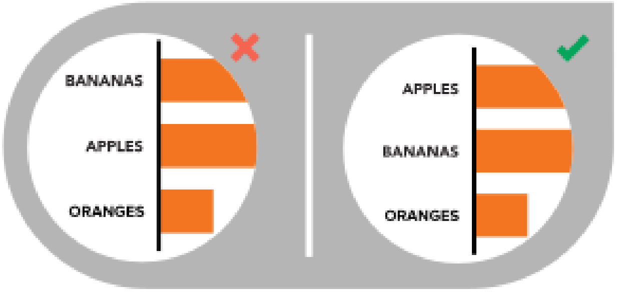
START THE Y-AXIS VALUE AT 0

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.

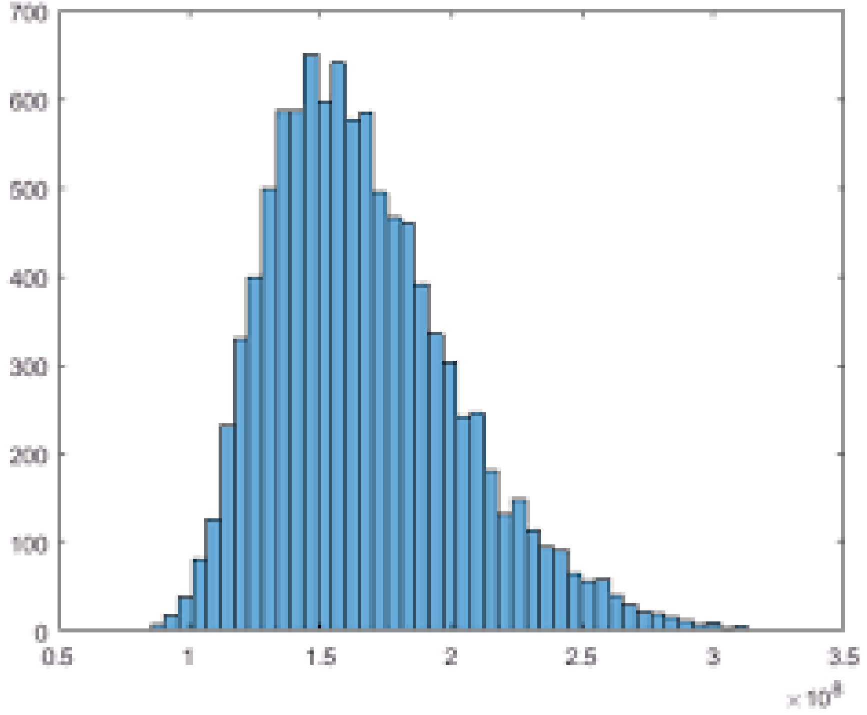


ORDER DATA APPROPRIATELY

Order categories alphabetically, sequentially, or by value.

Histograms

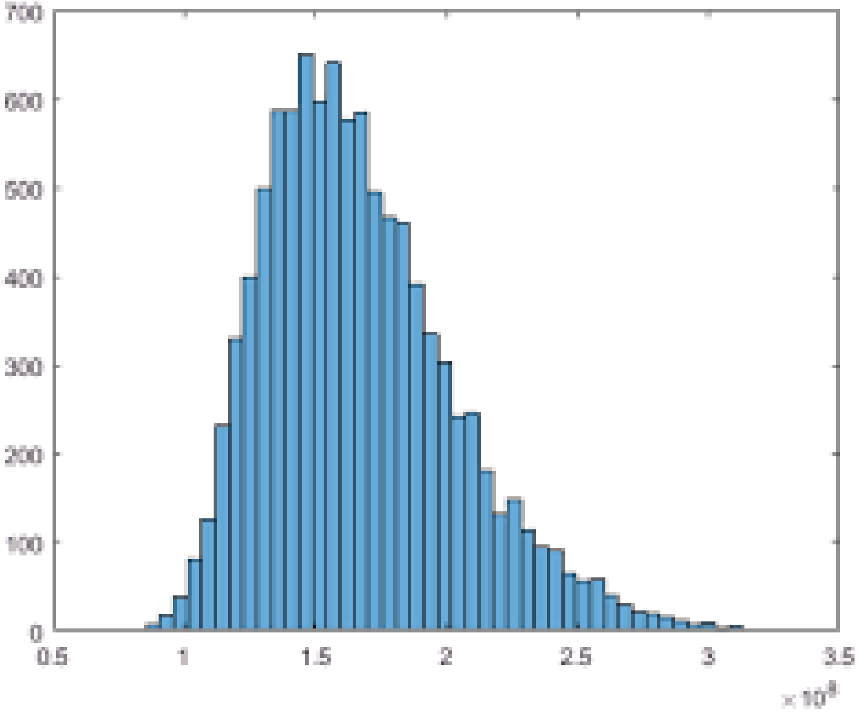
good to show the distribution



Basic example

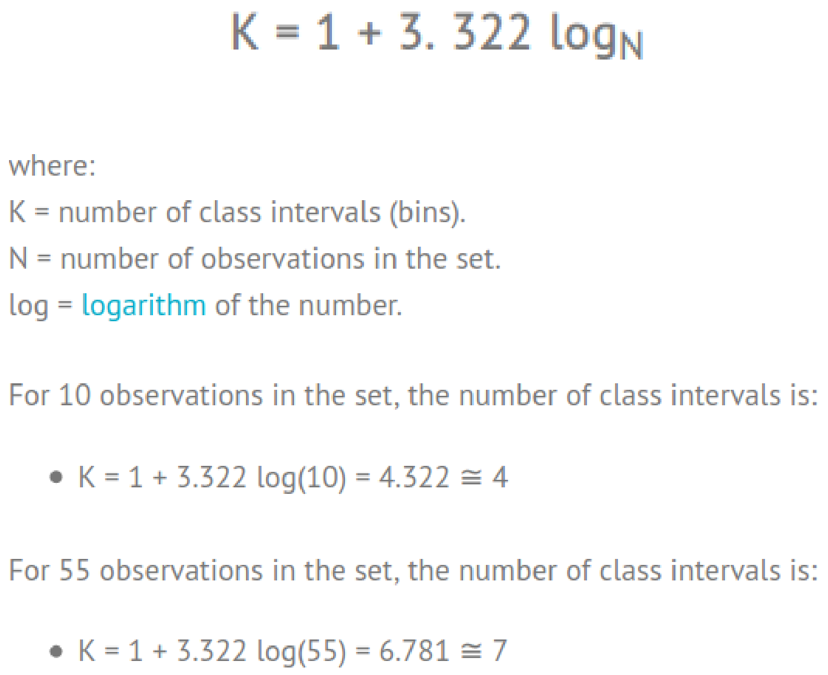
Histograms

good to show the distribution



Basic example

Follow bin sizes with Sturge’s Rule (or similar) if data normally distributed



$$K = 1 + 3.\ 322 \log_N$$

where:

K = number of class intervals (bins).

N = number of observations in the set.

log = logarithm of the number.

For 10 observations in the set, the number of class intervals is:

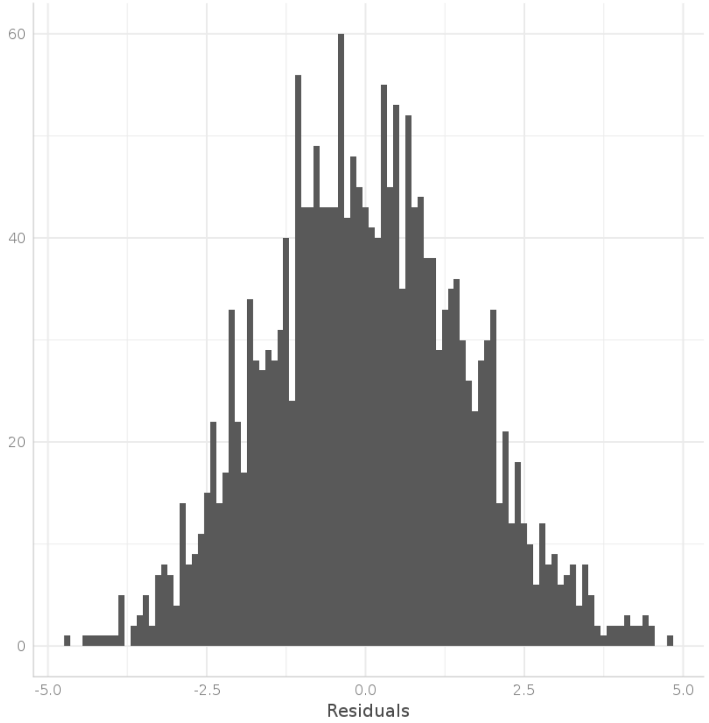
- K = 1 + 3.322 log(10) = 4.322 ≅ 4

For 55 observations in the set, the number of class intervals is:

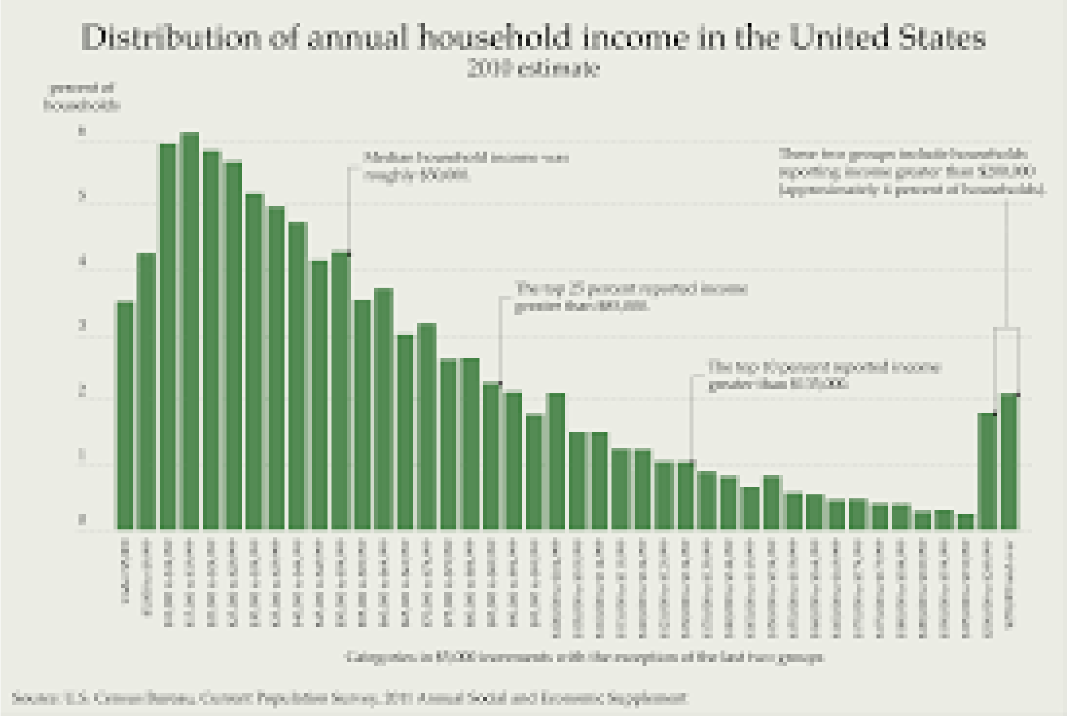
- K = 1 + 3.322 log(55) = 6.781 ≅ 7

Histograms

good to show the distribution



Normal distribution
(above average reading time)



Median household income was roughly $50,000.

These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).

The top 25 percent reported income greater than $85,000.

The top 10 percent reported income greater than $135,000.

Categories in $5,000 increments with the exception of the last two groups

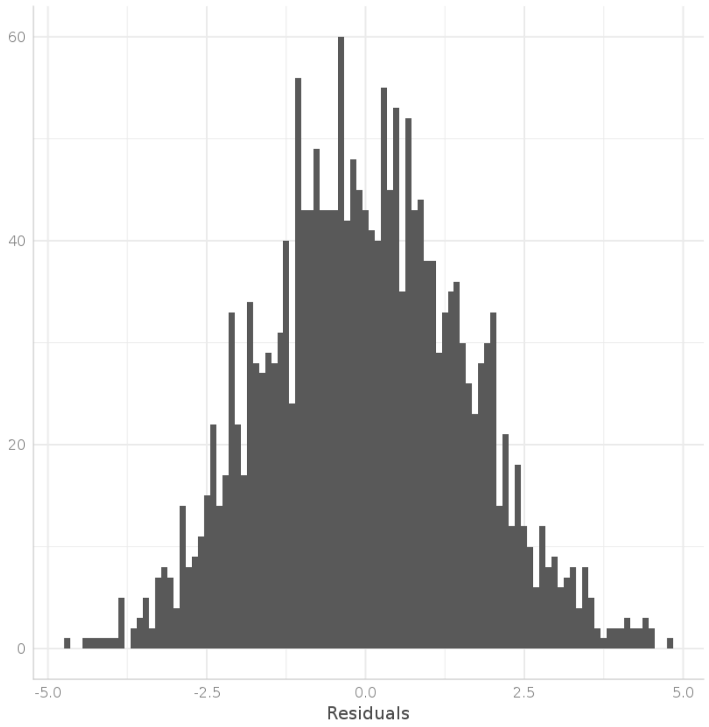
Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

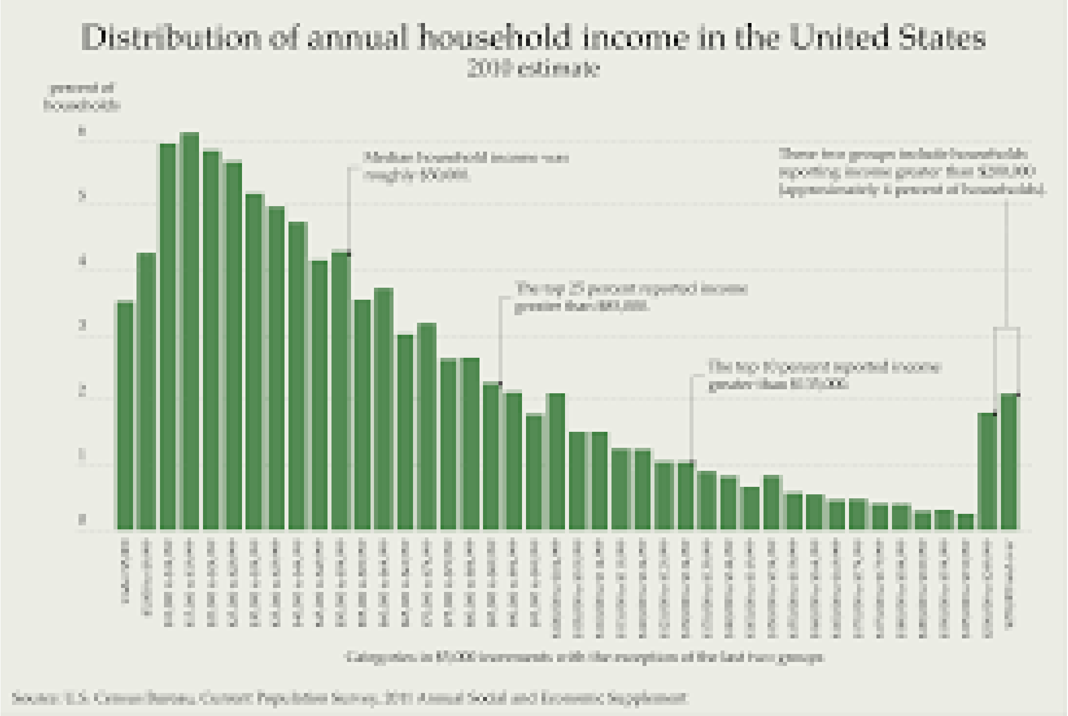
Skewed distribution
(also very popular in nature)

Note the long tail

Histograms

good to show the distribution





Distribution of annual household income in the United States
2010 estimate

Median household income was roughly $50,000.

The top 25 percent reported income greater than $85,000.

The top 10 percent reported income greater than $135,000.

These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).

Categories in $5,000 increments with the exception of the last two groups.

Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

Normal distribution
(above average readding time)

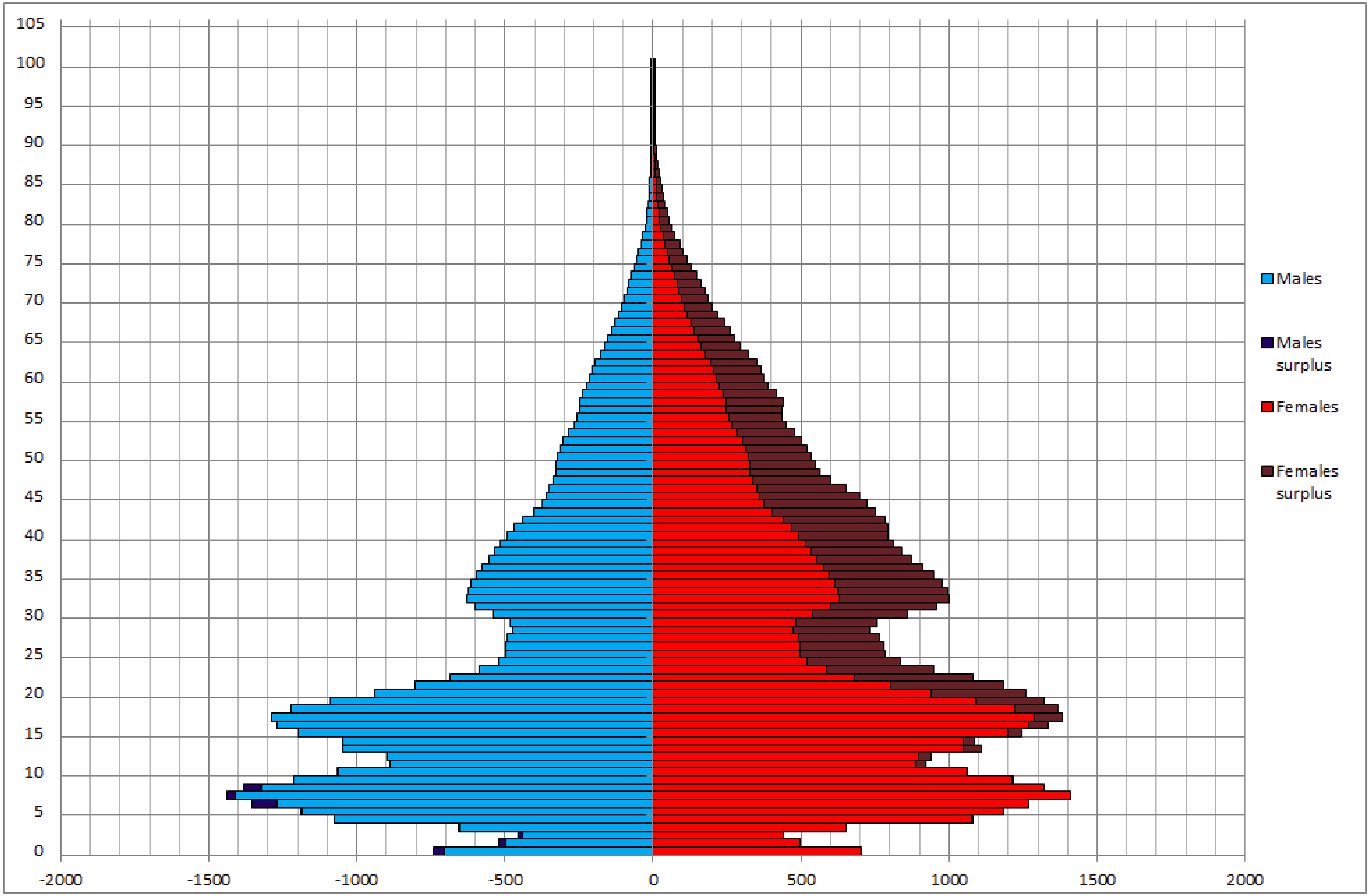
Skewed distribution
(also very popular in nature)

Note the long tail

For more read about:

- Lotka’s law	- Matthew effect
- Zipf's law	- Price's law

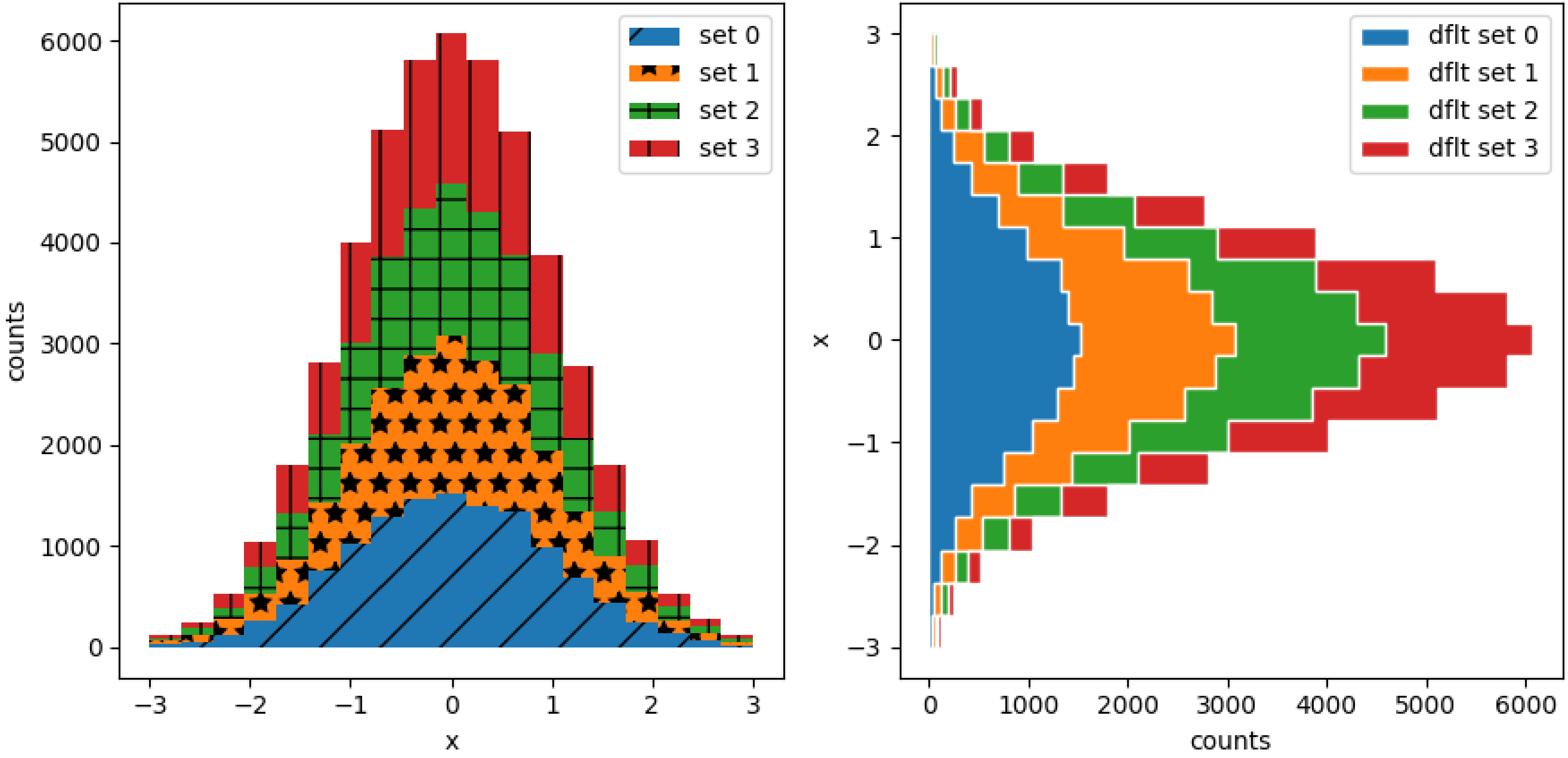
good to show the distribution e.g., population by the age (and sex)



Histograms

good to show the distribution e.g., population by the age (and sex)

can be used with colors and styles (but it is easy to end us with mess)



Histograms cons

Interpreting a histogram can be hard:

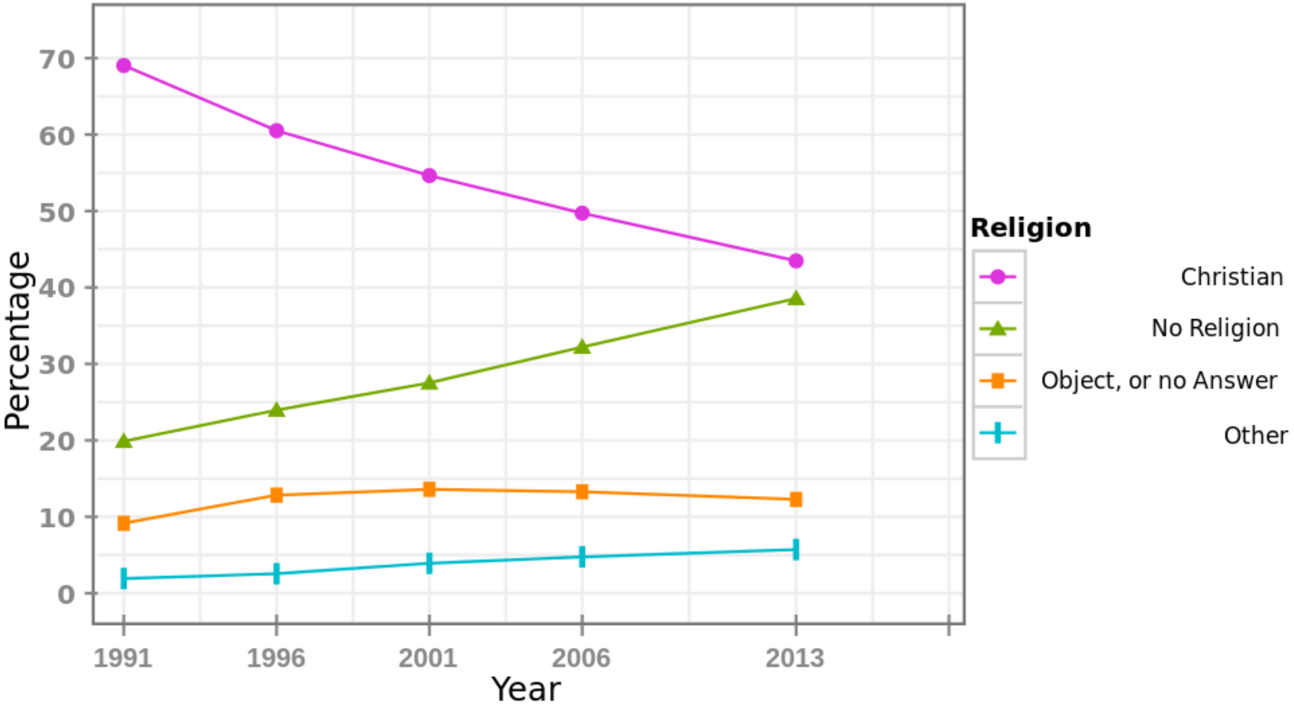
- it is extremely difficult to extract the exact amount of "input" (exception frequency histograms)

For example: if you are given a histogram and asked how many people gave their data in a survey, it would be extremely difficult to pinpoint an exact number

- inconvenient when comparing multiple categories

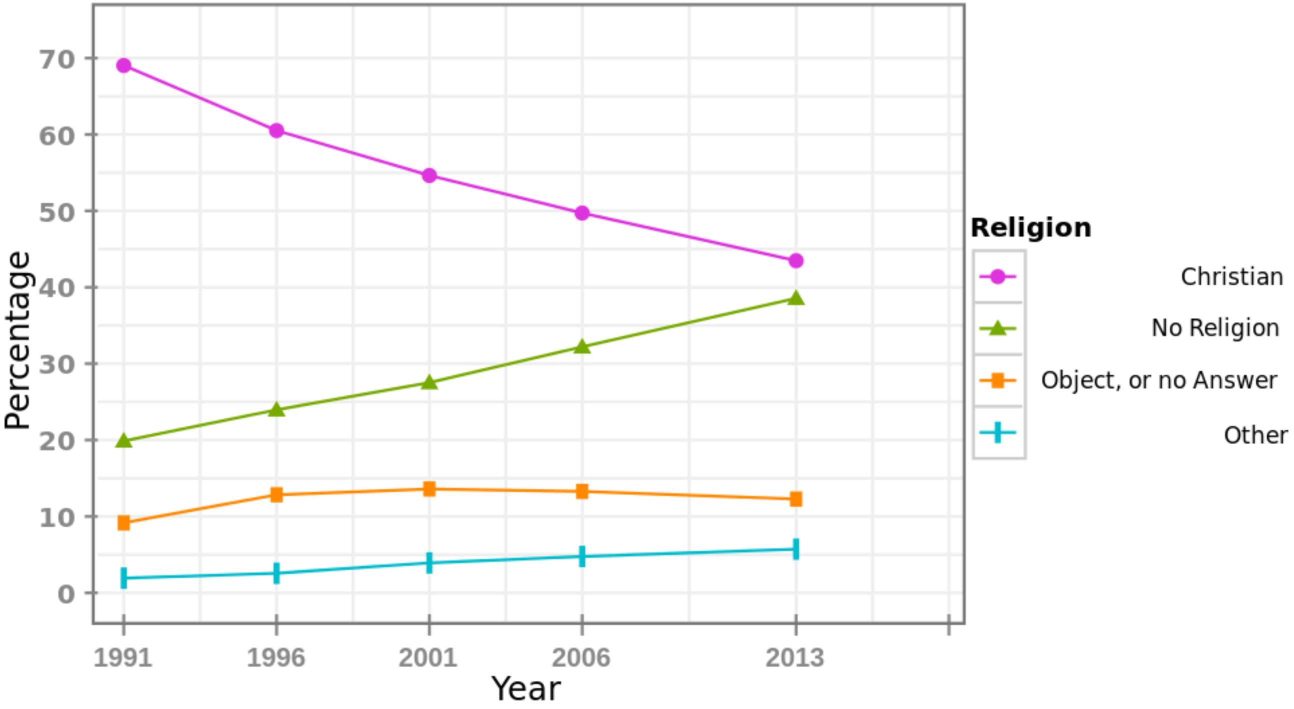
you can compare several histograms side by side, but it doesn't quite create the desired effect

- use to show time-series relationships with continuous data



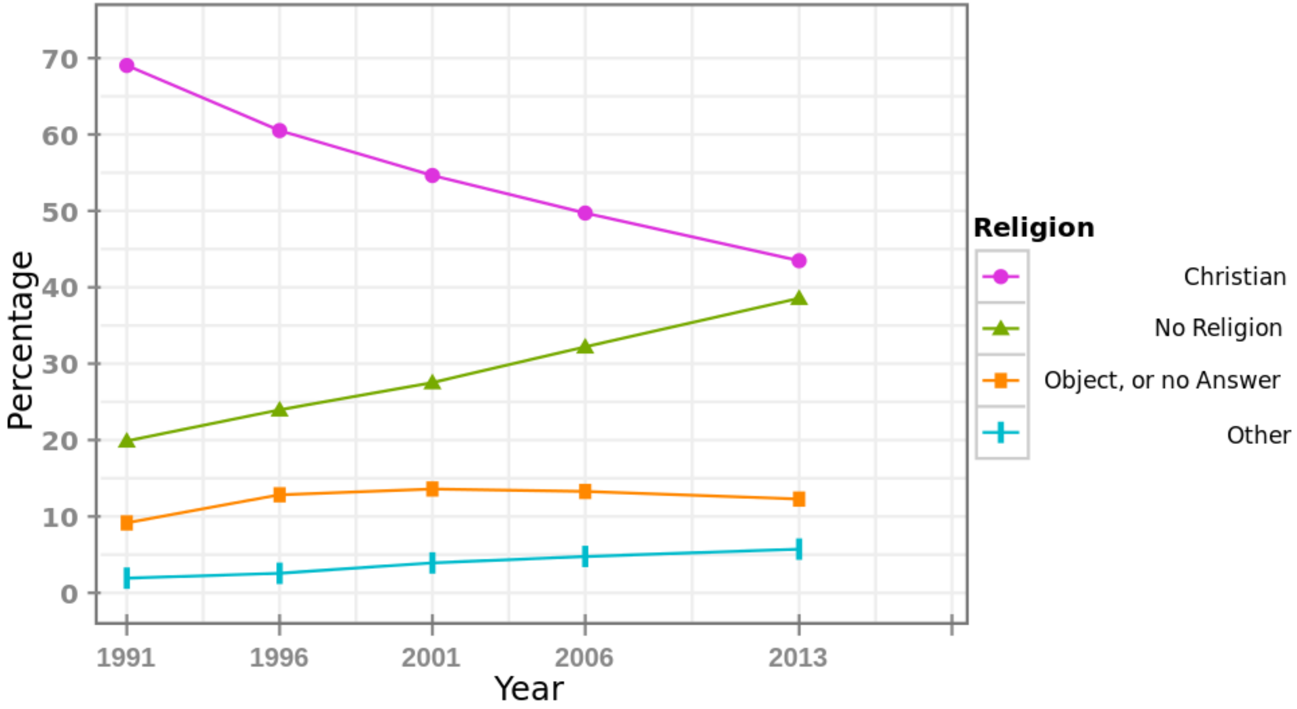
Line chart

- use to show time-series relationships with continuous data
- use to show trend, acceleration, deceleration, and volatility



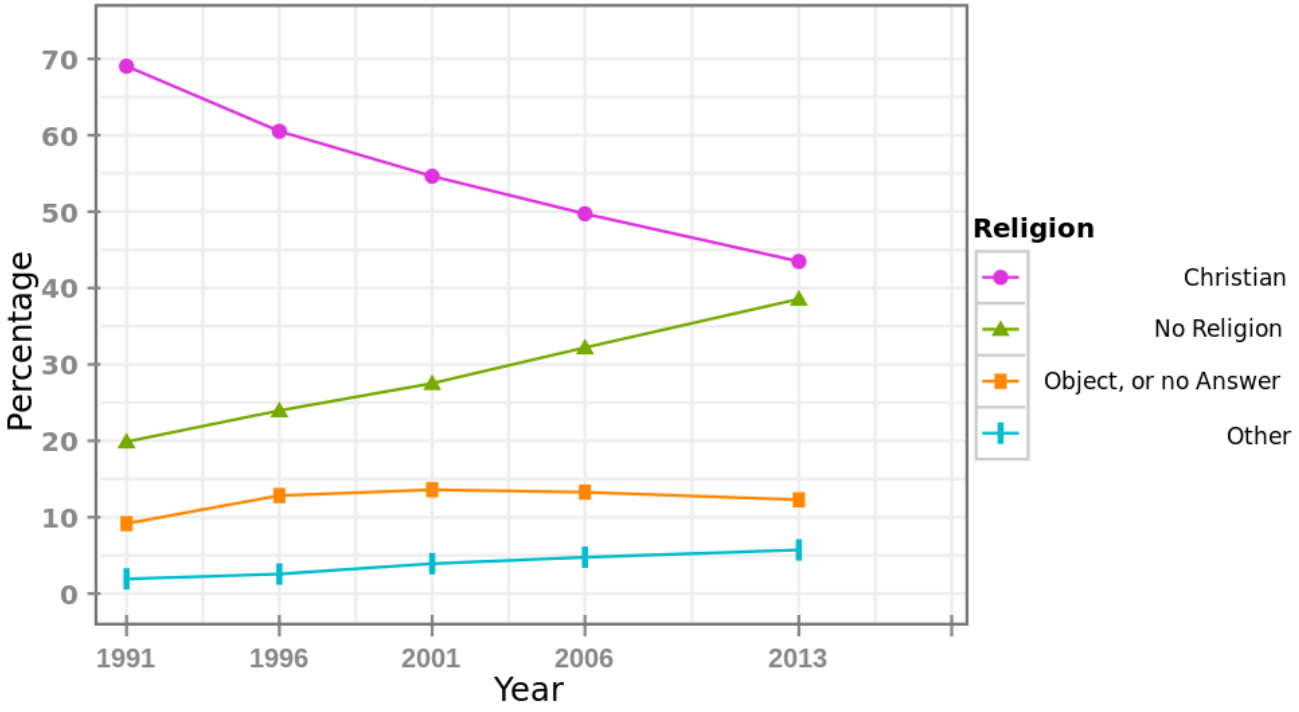
Line chart

- use to show time-series relationships with continuous data
- use to show trend, acceleration, deceleration, and volatility
- number of data points is very high (more than 20)



Line chart

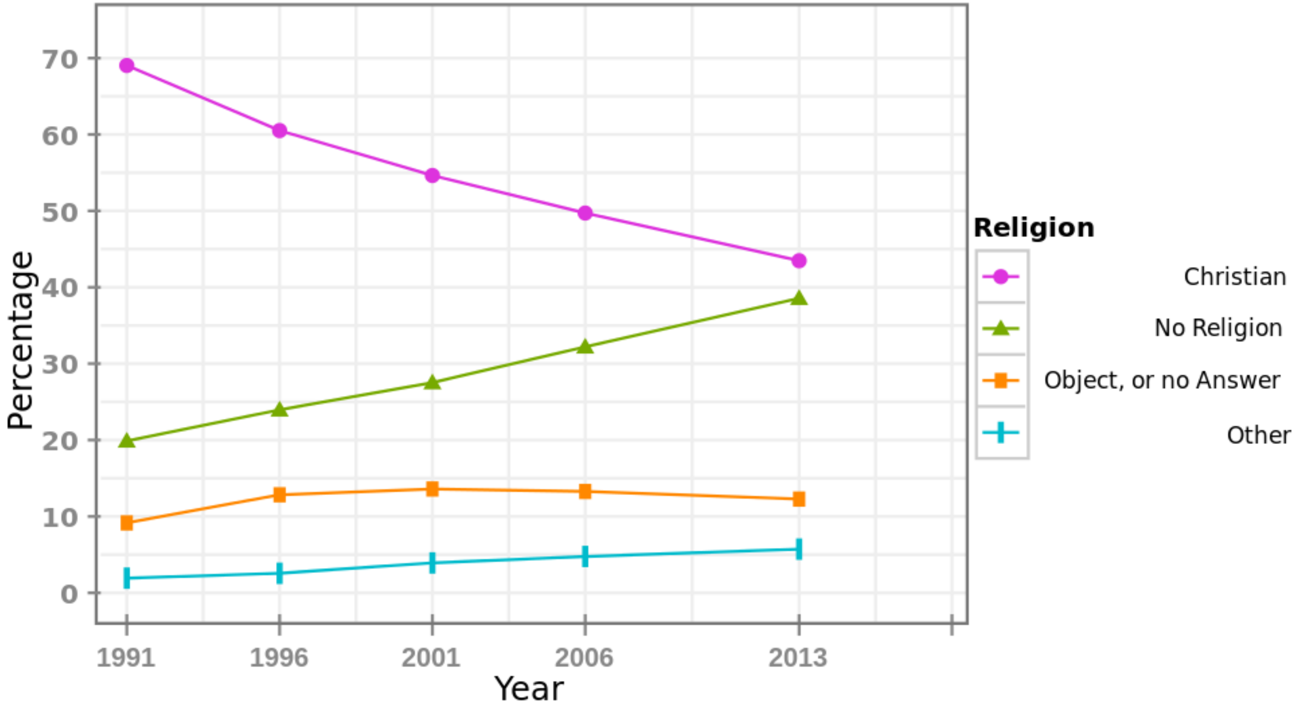
- use to show time-series relationships with continuous data
- use to show trend, acceleration, deceleration, and volatility
- number of data points is very high (more than 20)
- good alternative to column charts when the chart need to be small



Line chart

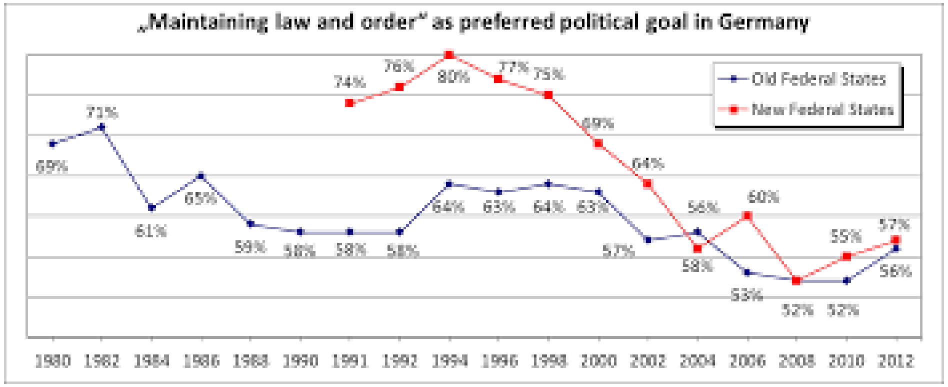
- use to show time-series relationships with continuous data
- use to show trend, acceleration, deceleration, and volatility
- number of data points is very high (more than 20)
- good alternative to column charts when the chart need to be small

Very popular (frequently good starting point)



Line chart

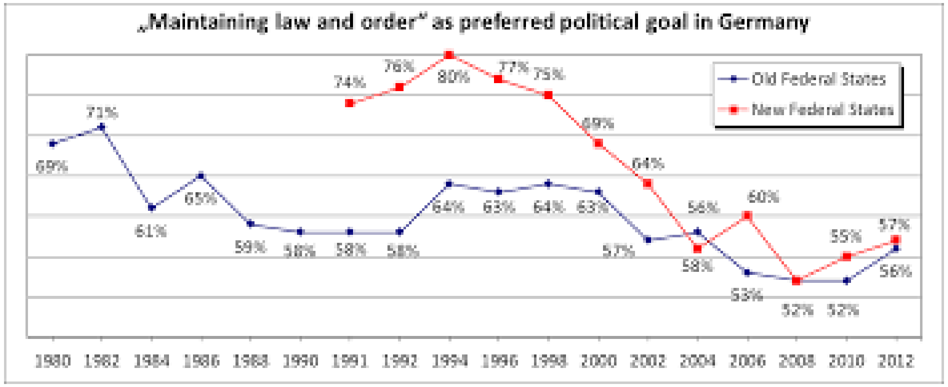
Can be done in 1001 ways, flavours, modifications, ...

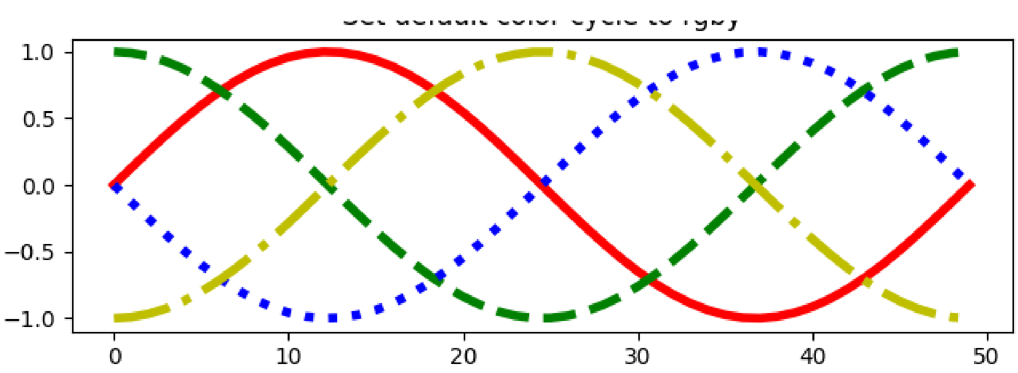


Time + percentage labels

Line chart

Can be done in 1001 ways, flavours, modifications, ...

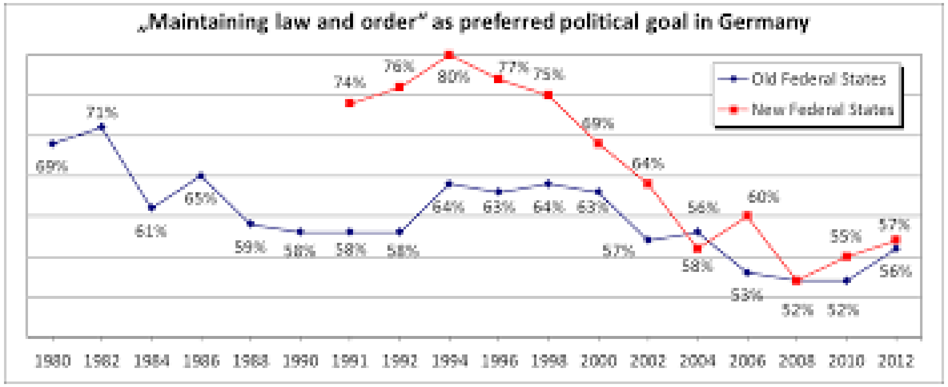


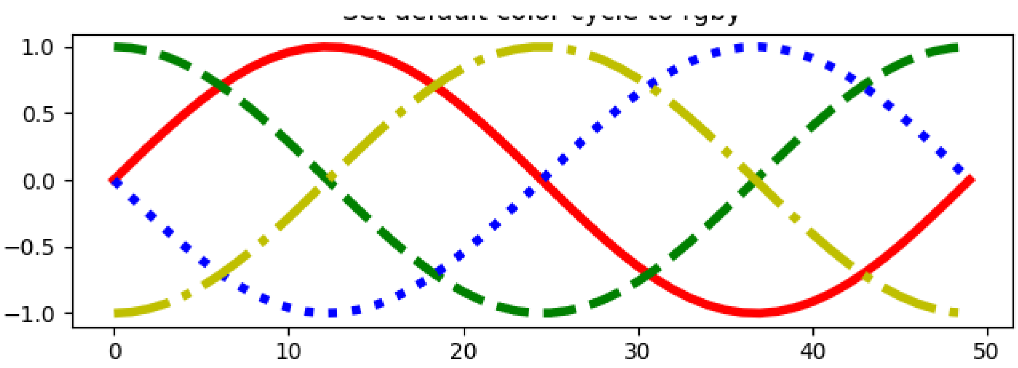


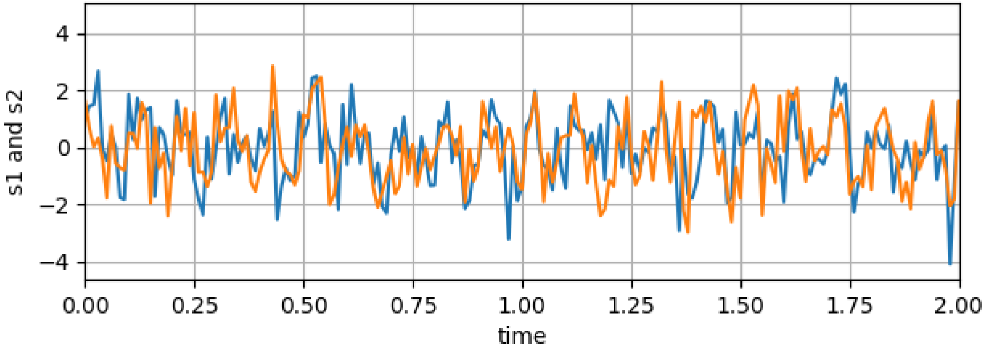
Lines can be distinguished by different color or/and style

Line chart

Can be done in 1001 ways, flavours, modifications, ...

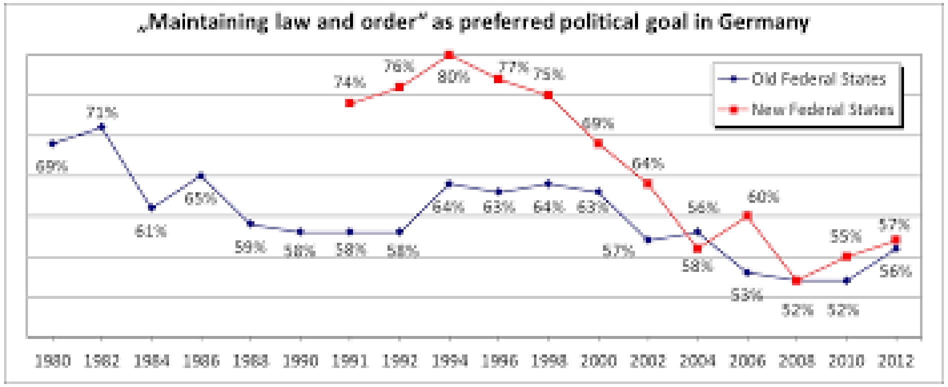


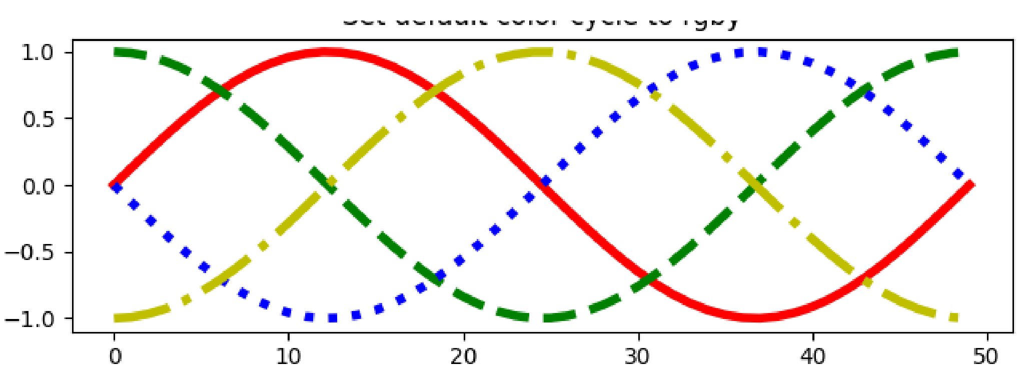


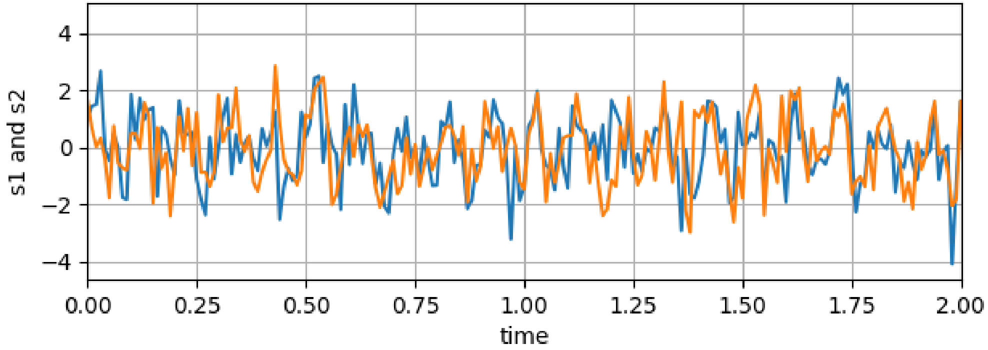


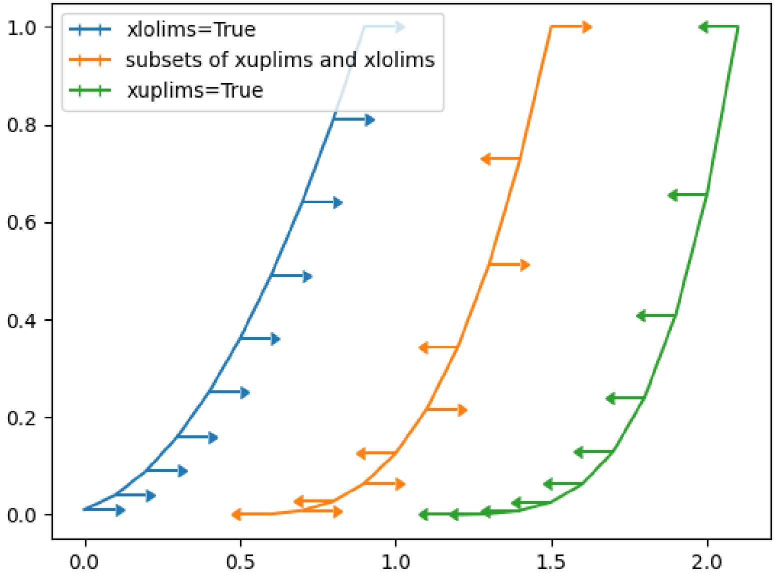
Very good for correlation analysis

Can be done in 1001 ways, flavours, modifications, ...



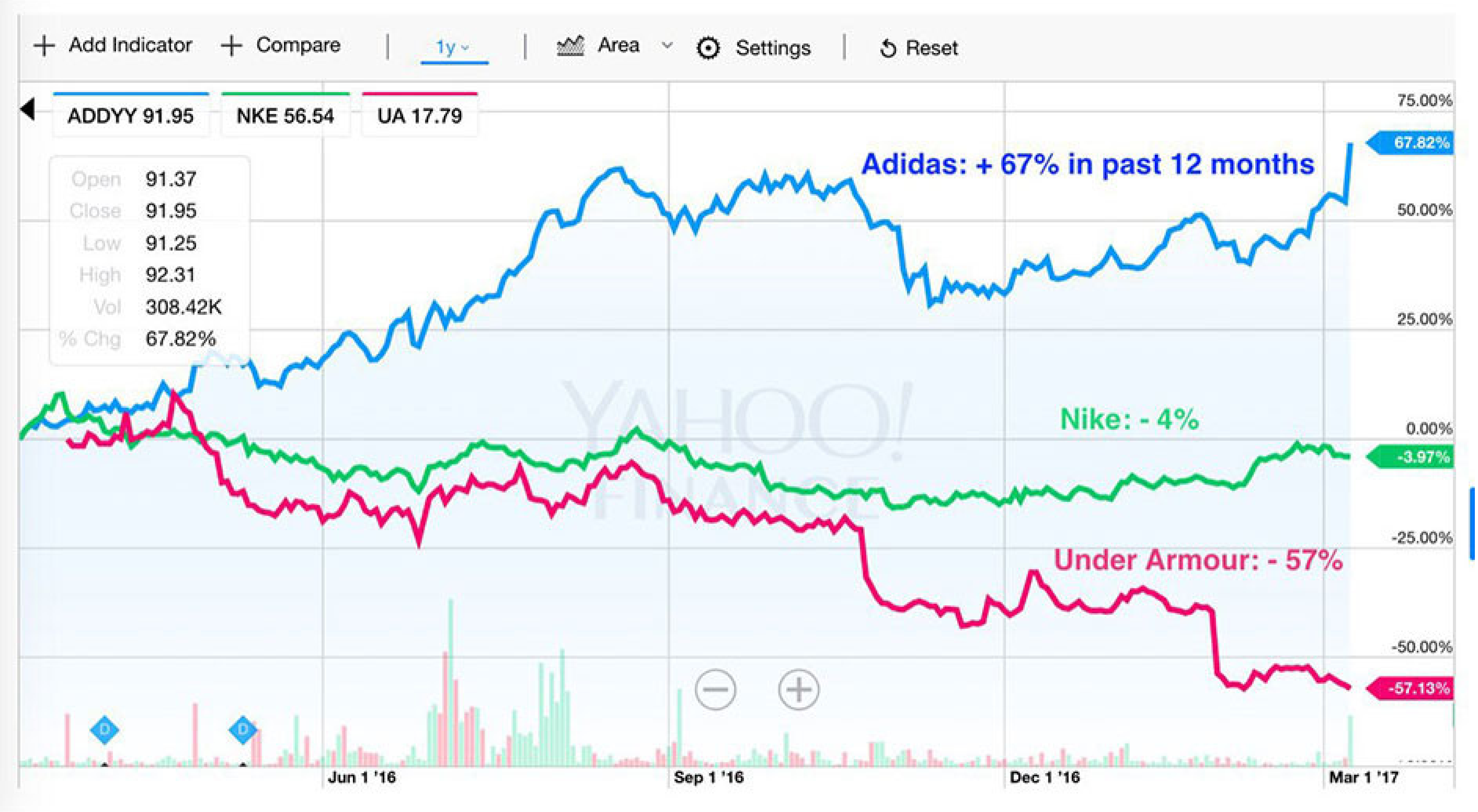






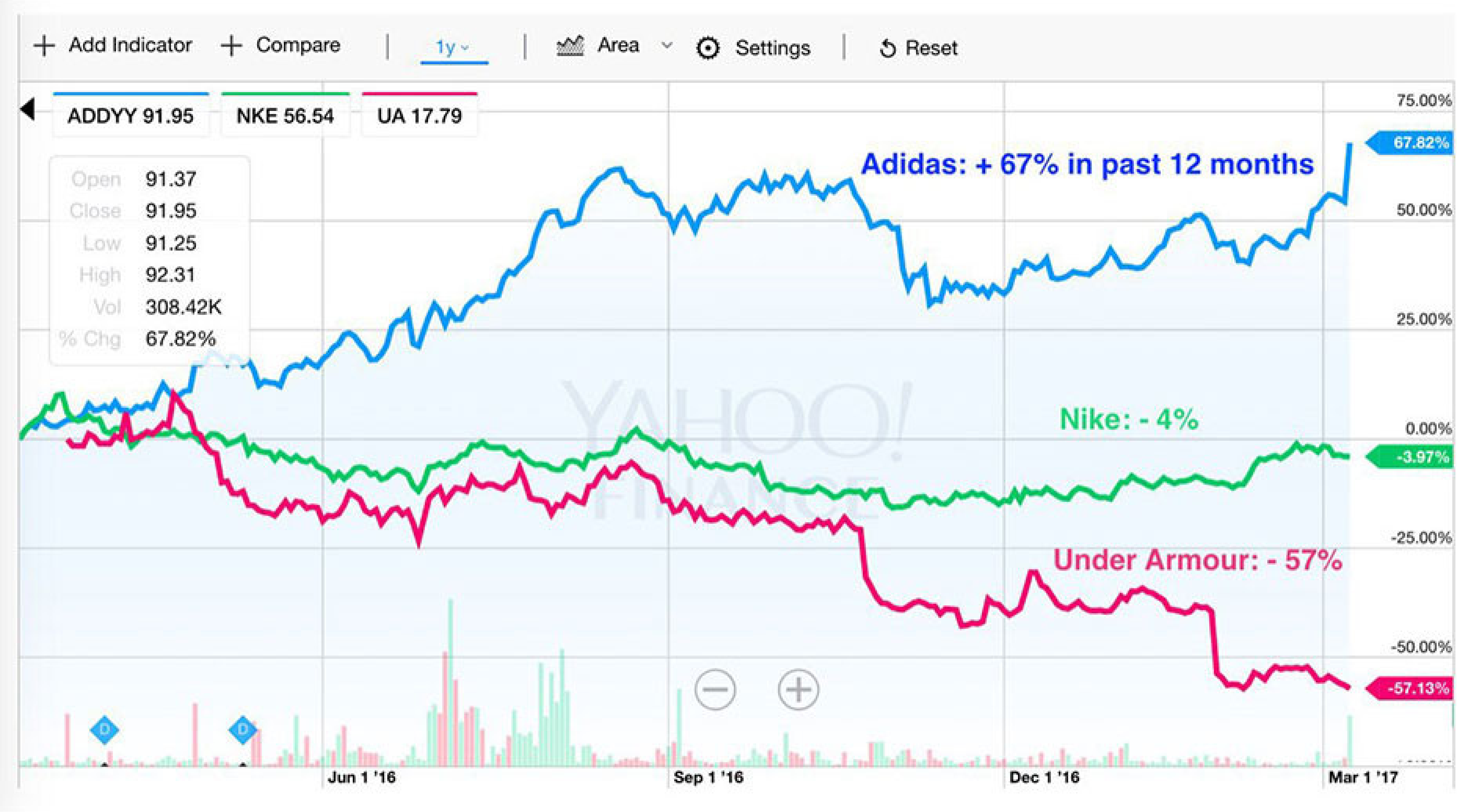
Can show trend or error as extra information

Line chart



Can be easily mixed with other type of charts

Line chart



Open	91.37
Close	91.95
Low	91.25
High	92.31
Vol	308.42K
% Chg	67.82%

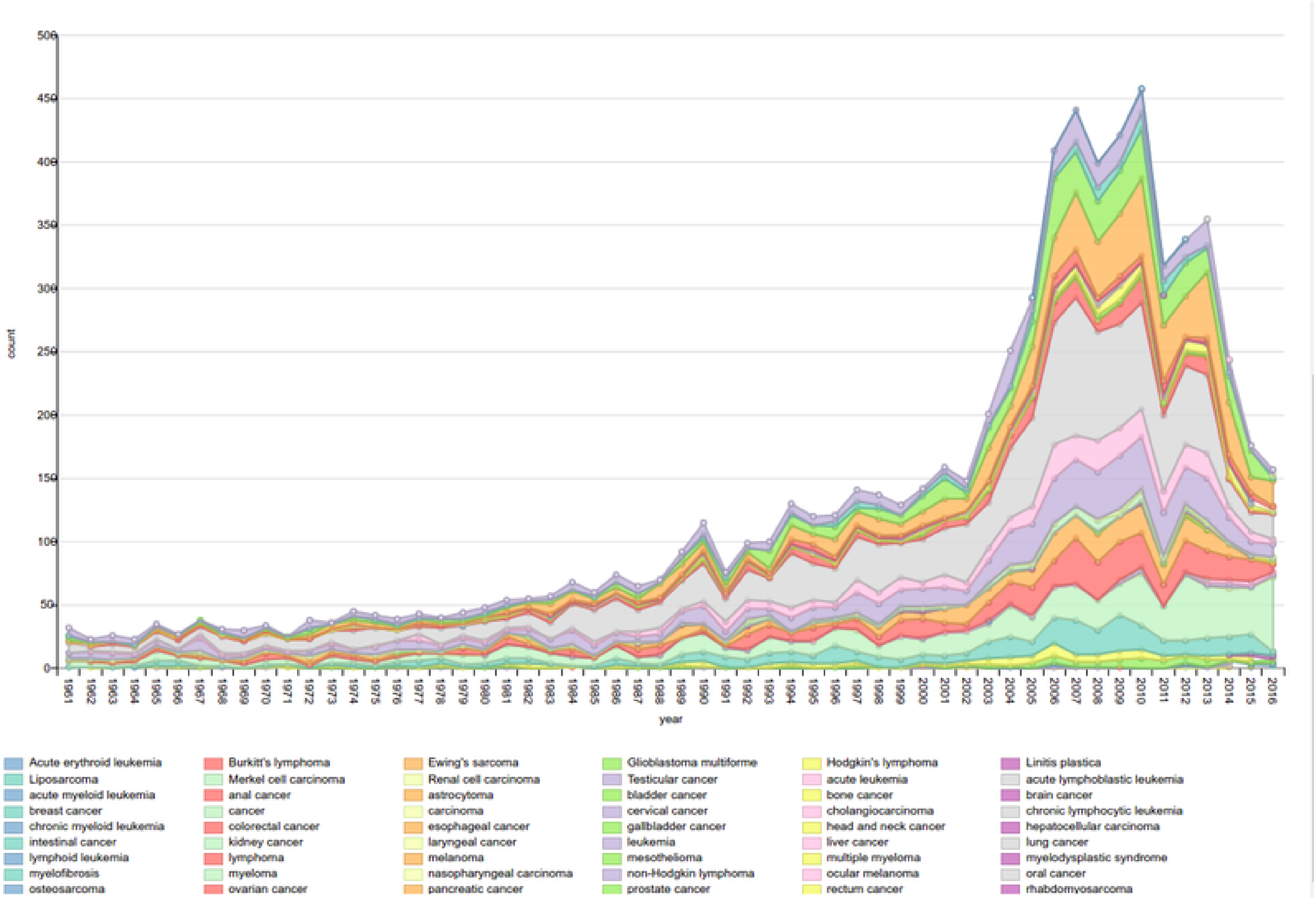
Adidas: + 67% in past 12 months

Nike: - 4%

Under Armour: - 57%

Those related to time are frequently called timeline charts
Let the user zoom in and out and compress or stretch the time axis to see more details or overall trends

Line chart



They can also be used to show area, changes in composition over time, accumulative value changes, over time

Line chart

Use to present continuous data in an interval scale, where intervals are equal in size

The axis may not start from zero if the intended message is to show the rate of change or overall trend, not exact values or comparison. Yet, it is best to start the axis with zero for wide audiences because some people may otherwise interpret the chart incorrectly.

Time should always run from left to right

Do not skip values for consistent data intervals presenting trend information, for example, certain days with zero values

Remove guidelines to emphasize the trend, rate of change, and to reduce distraction

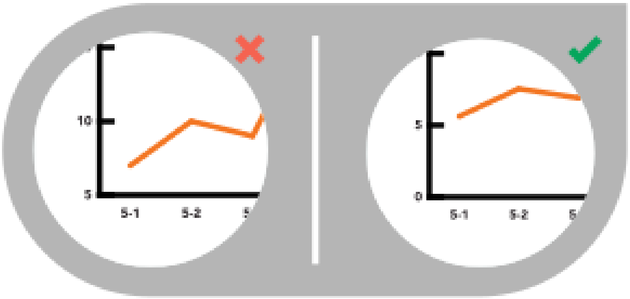
Use a proper aspect ratio to show important information and avoid dramatic slope effects

For the best perception, aim for a 45-degree slope
see https://eagereyes.org/basics/banking-45-degrees

Line chart

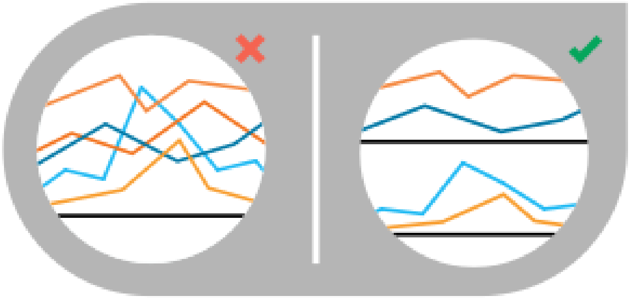
LINE CHART

DESIGN BEST PRACTICES



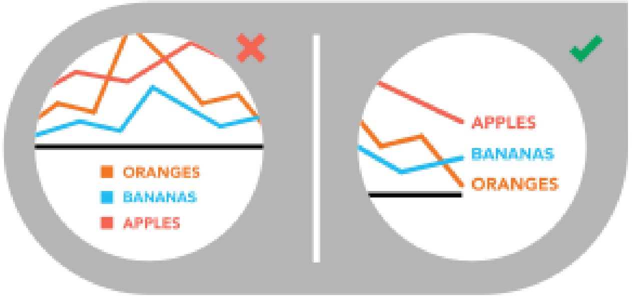
INCLUDE A ZERO BASELINE IF POSSIBLE

Although a line chart does not have to start at a zero baseline, it should be included if possible. If relatively small fluctuations in data are meaningful (e.g., in stock market data), you may truncate the scale to showcase these variances.



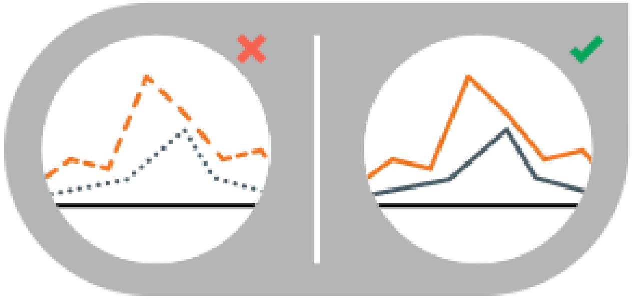
DON'T PLOT MORE THAN 4 LINES

If you need to display more, break them out into separate charts for better comparison.



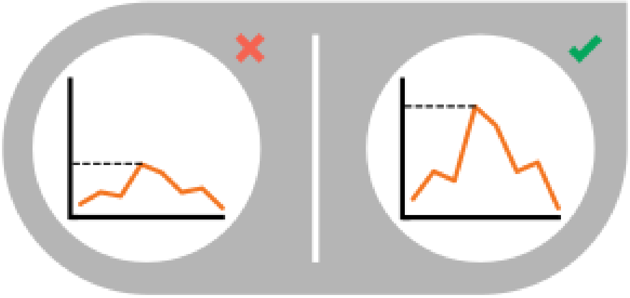
LABEL THE LINES DIRECTLY

This lets readers quickly identify lines and corresponding labels instead of referencing a legend.



USE SOLID LINES ONLY

Dashed and dotted lines can be distracting.



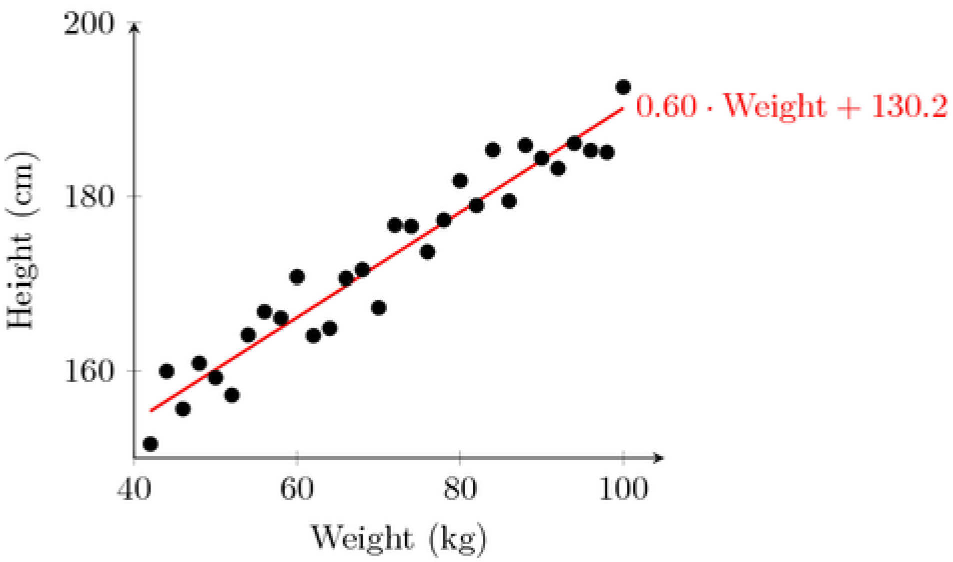
USE THE RIGHT HEIGHT

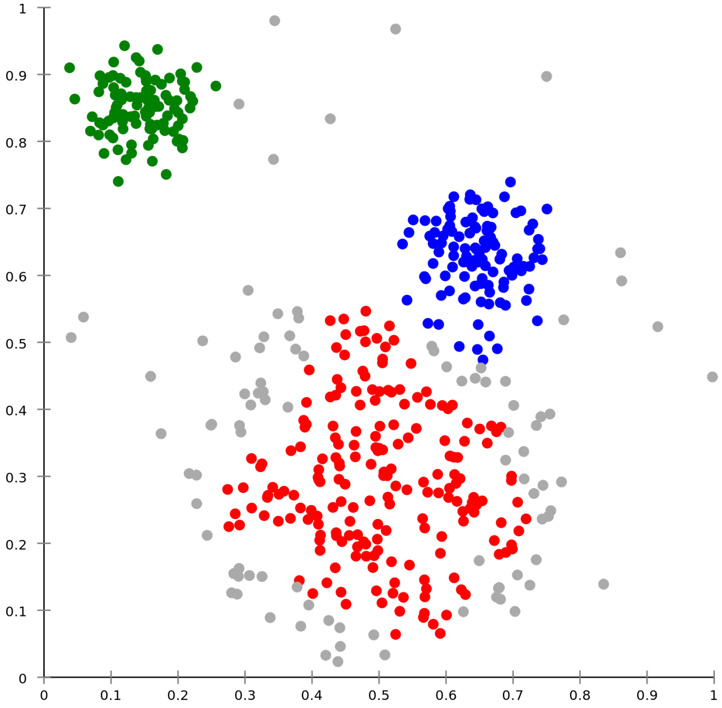
Plot all data points so that the line chart takes up approximately two-thirds of the y-axis' total scale.

Scatter plot

- show the relationship between items based on two sets of variables

- show correlation or clustering effects in a large amount of data

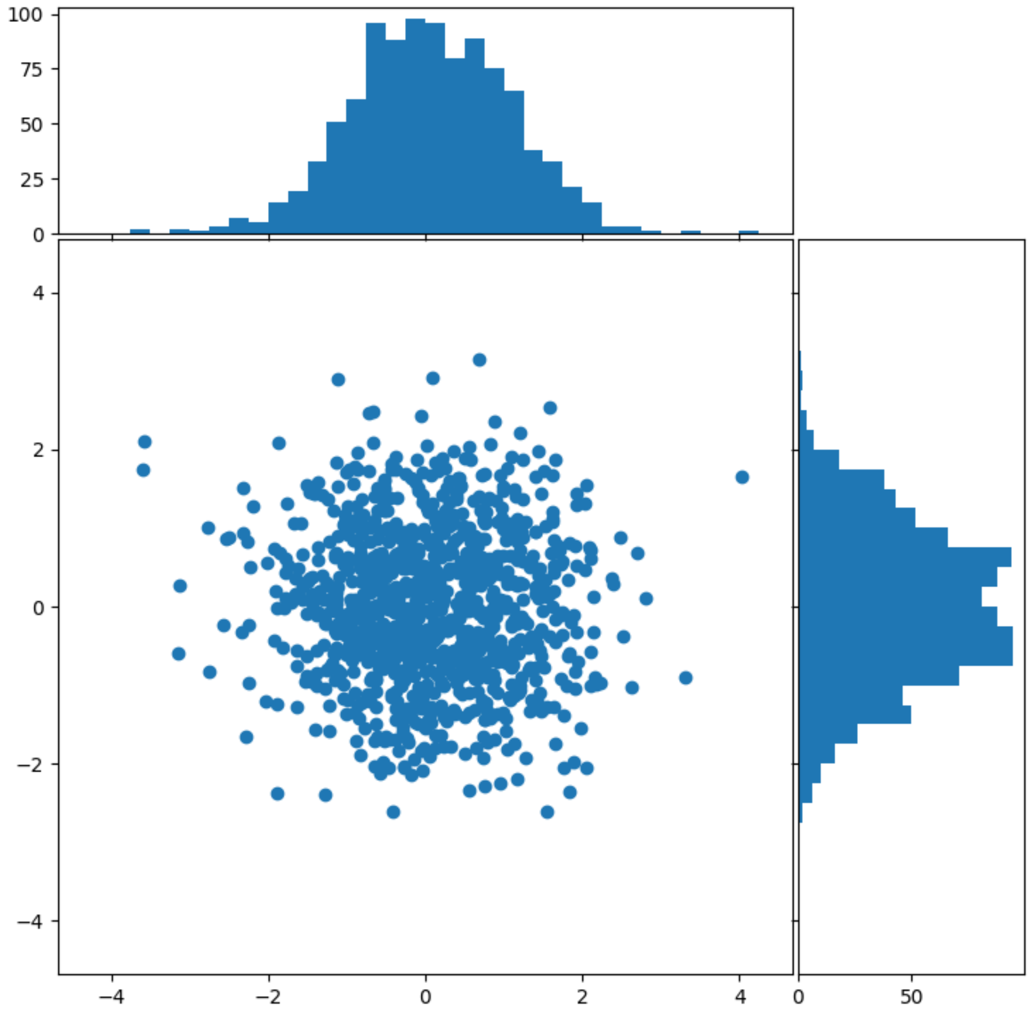




Scatter plot

- show the relationship between items based on two sets of variables

- show correlation or clustering effects in a large amount of data

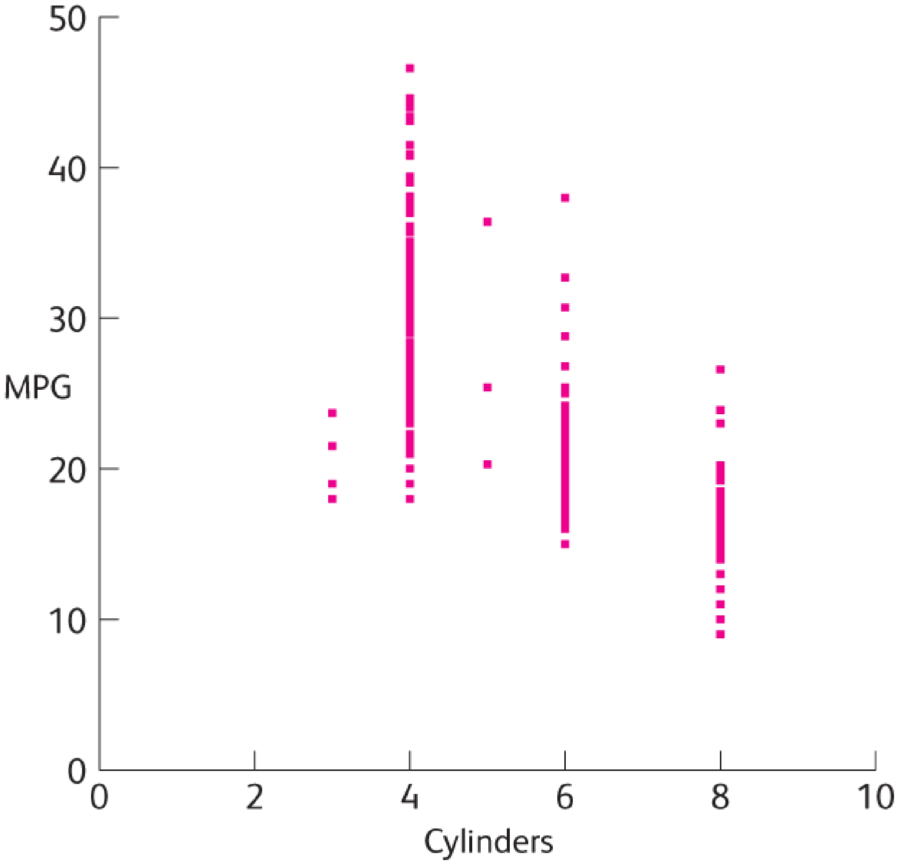


- can be mixed with other (e.g. histograms)

Scatter plot

Not the best option if data are discrete.

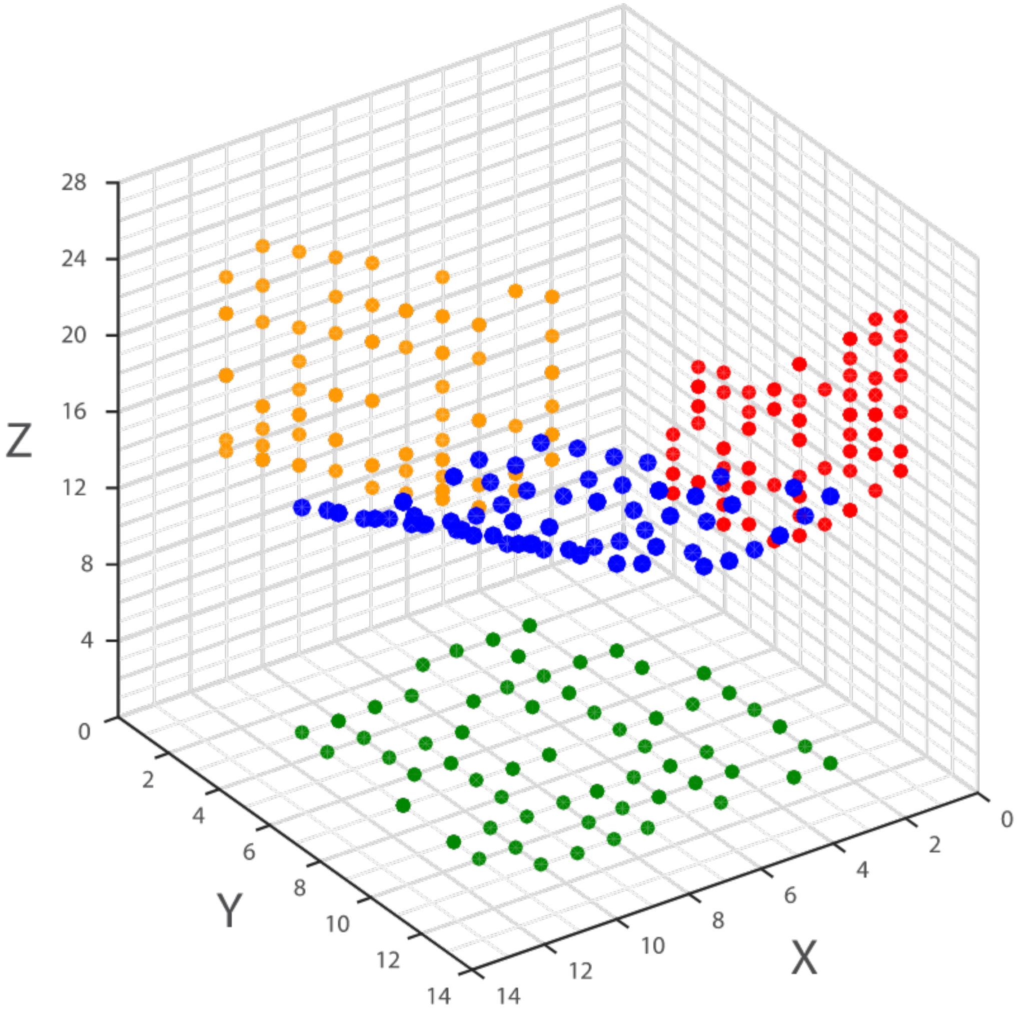
This happens when decimal places are rounded off, measurements are not accurate enough, or a data field is categorical.

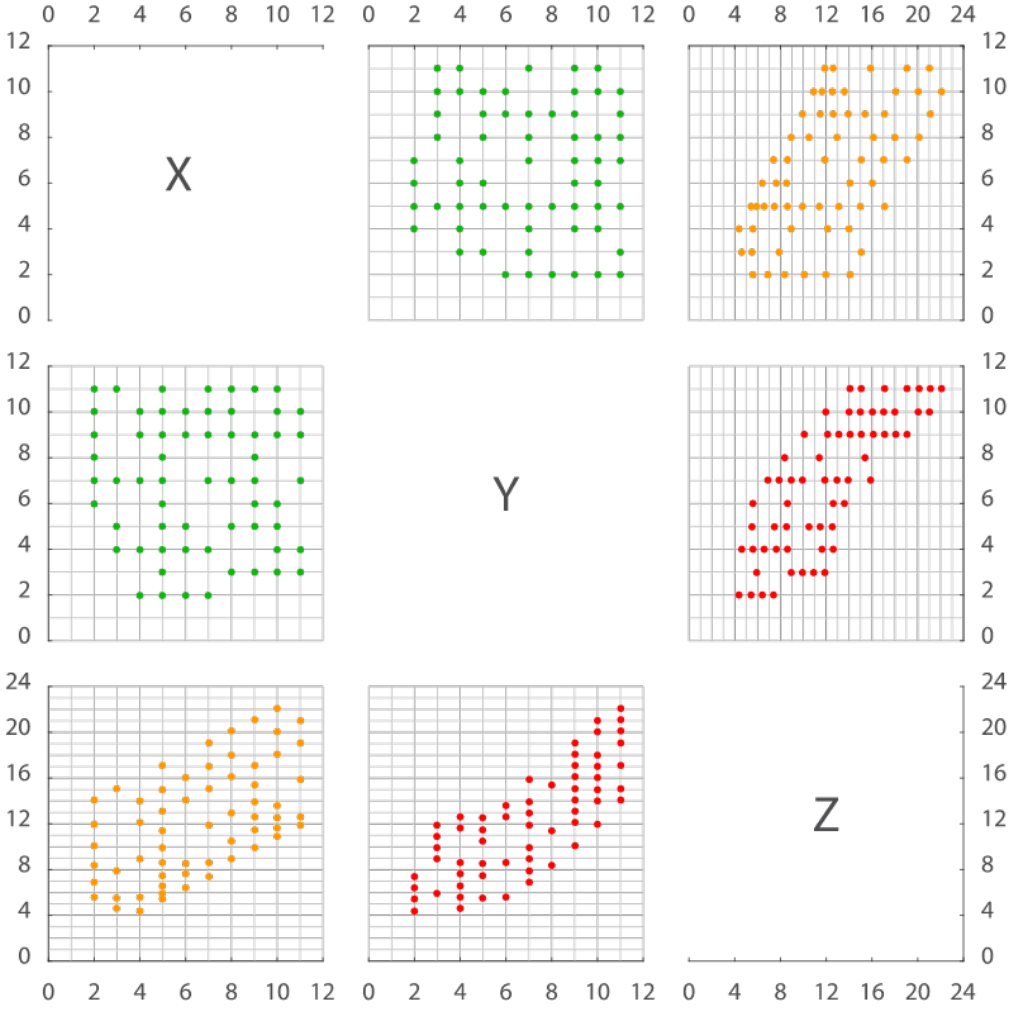


Scatter plot

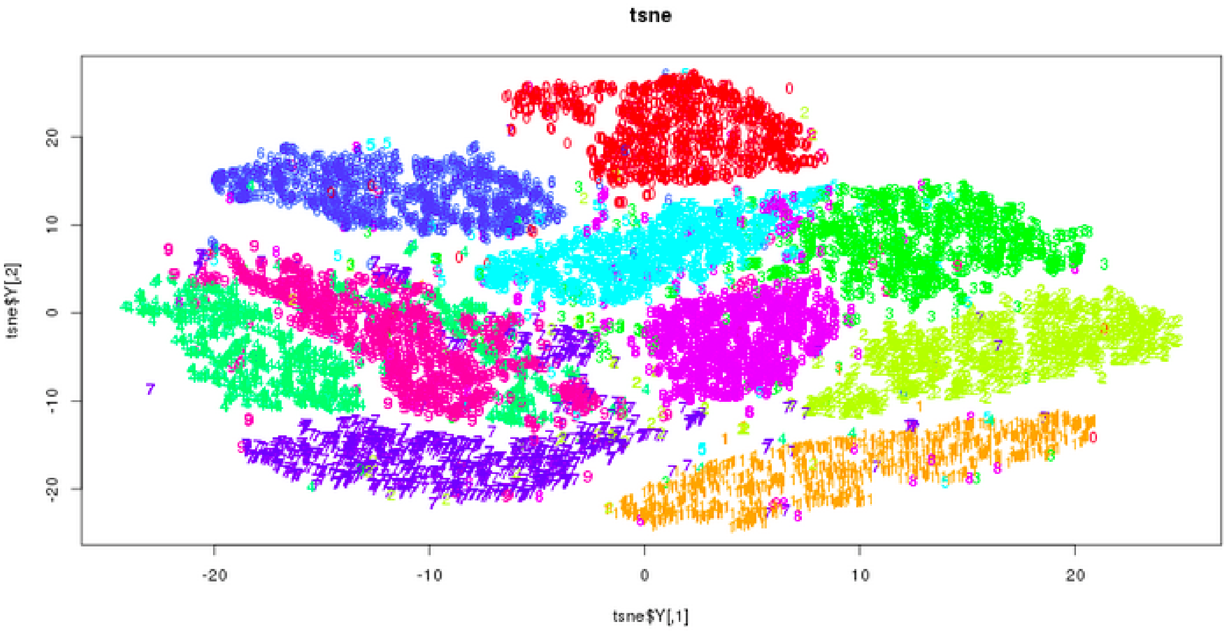
Good for 2D

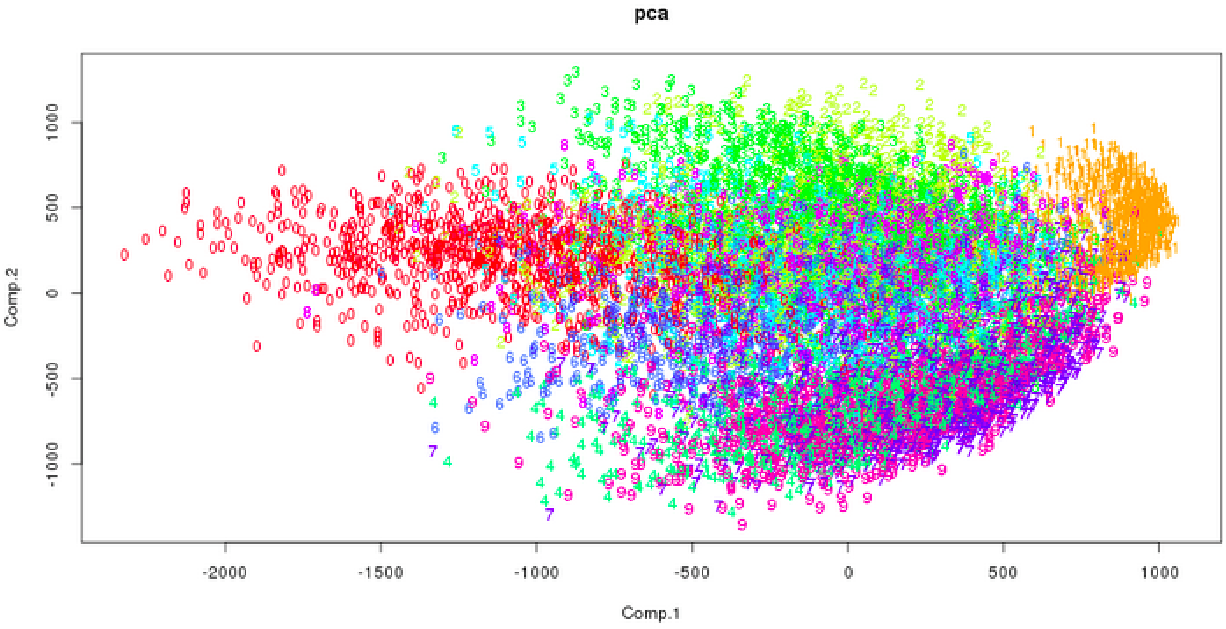
The more dimensions, the worse the effect





Scatter plot





For dimensionality data reduction see:

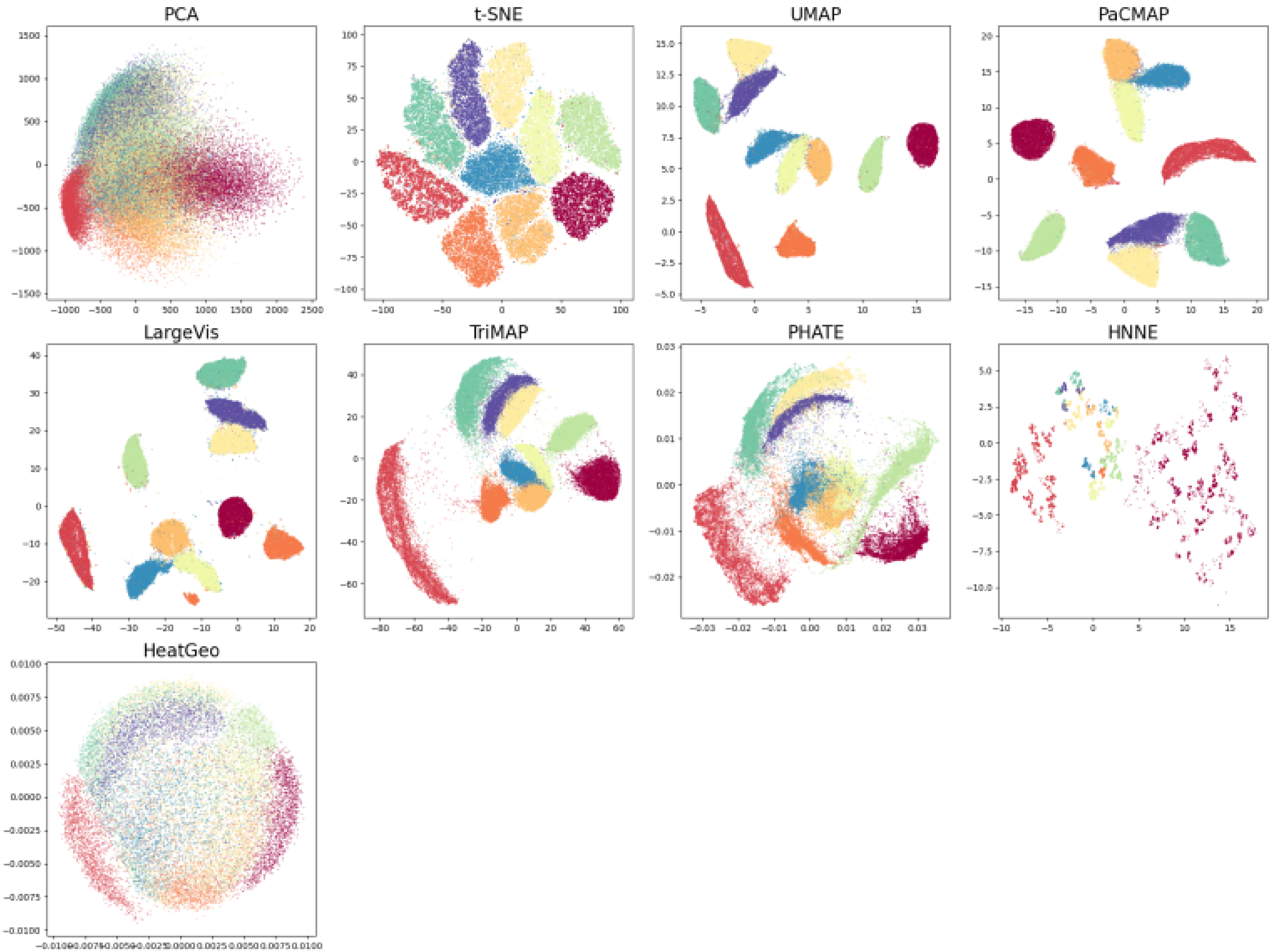
- **t-SNE**

- **PCA**

- **UMAP**

- **...**

Scatter plot



Scatter plot

F

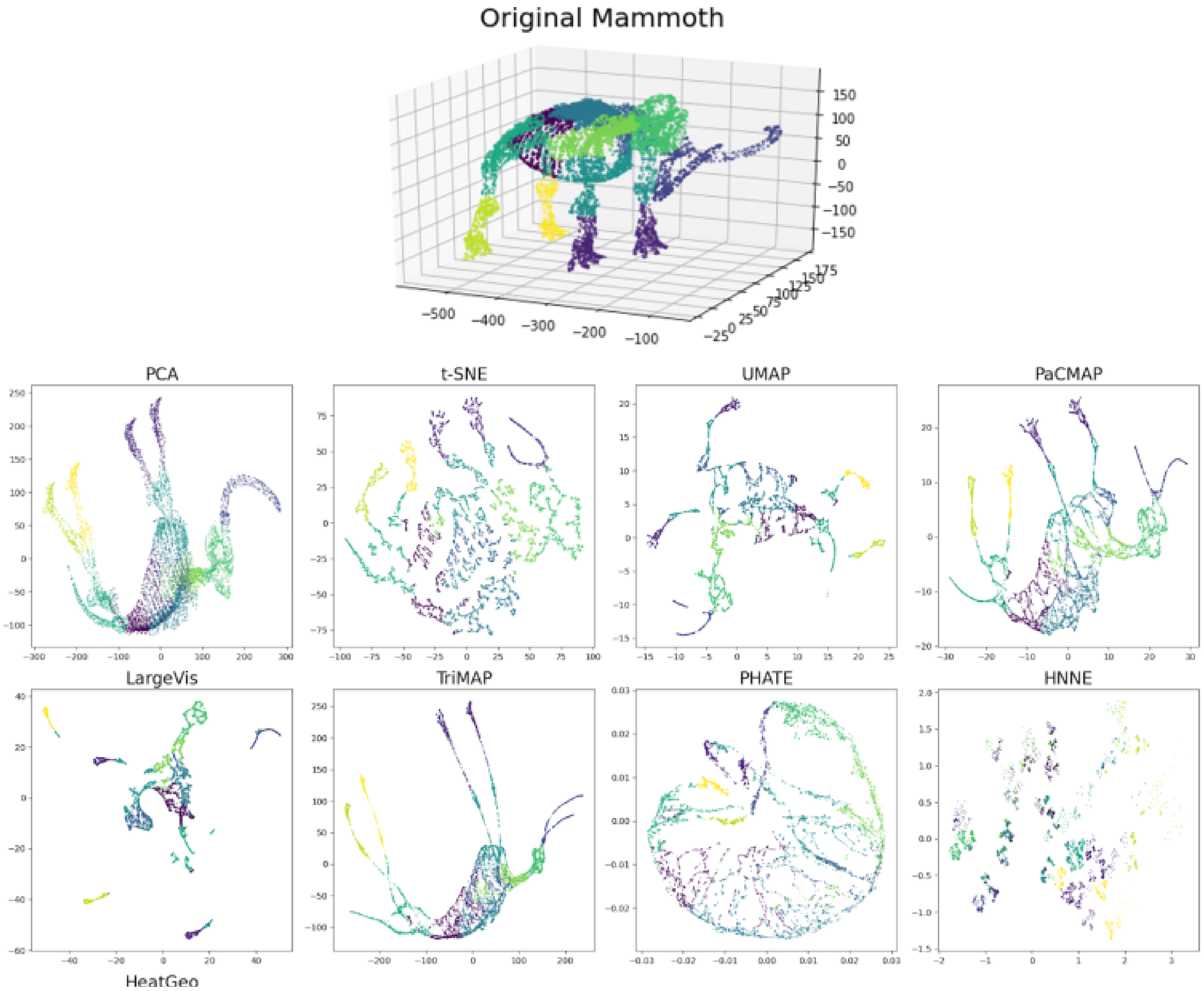
-

-

-

h

h



Scatter plot

For dimensionality data reduction

- t-SNE

- PCA

- UMAP

....

Must read links:

https://www.youtube.com/watch?v=o_cAOa5fMhE

https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/

https://sites.duke.edu/dimensionreduction/

http://jmlr.org/papers/v22/20-1061.html

Plots

To be continued

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

l.kozlowski@mimuw.edu.pl

www: bioinformatic.netmark.pl

Credits:

Infographics: The Power of Visual Storytelling by Ross Crooks, Jason Lankow and Josh Ritchie (Wiley 2012);

The Wall Street Journal Guide to Information Graphics by Dona Wong (Dow Jones & Company 2010)

Visualize This by Nathan Yau (Wiley 2011)

Michael Grossberg (2015) Data Visualization Basics Tools, Principles and Pitfalls

https://www.kaggle.com/timoboz/data-science-cheat-sheets

visage

HubSpot