Data Visualization

Introduction to Data Science

3rd lecture

Prepared by: Assoc. Prof. Alan Jović

Ac. year 2023/2024

Data visualization

Two main purposes:

- **In data analysis**
  - Supports reasoning on present data
- **In communication**
  - Informs and convinces other collaborators

Source: Yvette, W., Pixabay

Data visualization for data analysis

- Goals of visualization for the purpose of data analysis:
  - Discovers relations among variables and among examples
  - Condensed description of variables' values
  - Better understanding of analysis results
- Visualization is very relevant for **understanding the dataset** (types of variables, size of dataset), for discovery of errors in data and for setting feasable analysis goals
  - An important part of **data survey**
  - Some goals are not feasible if the dataset is inadequate (too small, too large, contains some types of variables, doesn't contain other types, etc.)
  - Many errors in data are difficult to discover without visualization

Data visualization for communication

- Goals of visualization for the purpose of communication:
  - Draws attention and involves collaborators a lot better than text/numbers
  - Allows telling stories in a visual way
  - Enables focus on some particular aspects, hiding the details (abstraction)

- Visualization is very relevant for involving other people that did not work on a dataset
  - Problems of a dataset as well as the most important results can easily be shared with others
  - Facilitates discussion and reaching business decisions

Data visualization

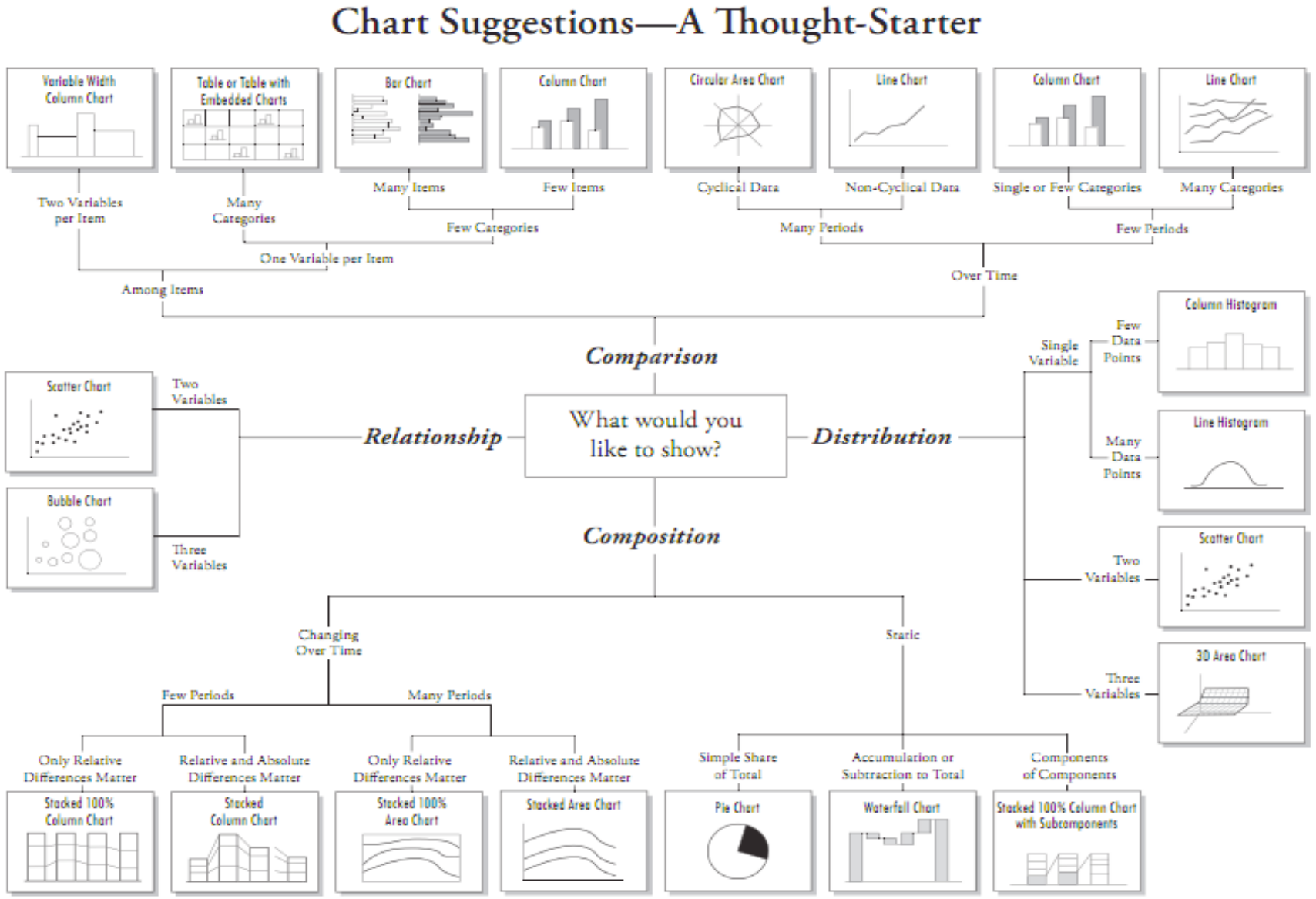
- **Static visualization**
  - Excellent for data exploration
  - In development for the last couple of centuries
  - **Focus of this lecture**
- **Interactive visualization**
  - Enables user's **actions for changing/replacing elements** in a graphical display
  - Becomes very common in displaying results (e.g., dashboard)
  - Based on new web frameworks
  - https://www.heavy.ai/technical-glossary/interactive-data-visualization
  - https://coronavirus.jhu.edu/map.html
  - https://towardsdatascience.com/dashboards-are-dead-b9f12eeb2ad2
  - https://www.xenonstack.com/blog/streaming-data-visualizations

Content

- Navigating through graphs for data visualization
- Principles and best practices of visualization
- Examples of visualization uses
- Tools for visualization

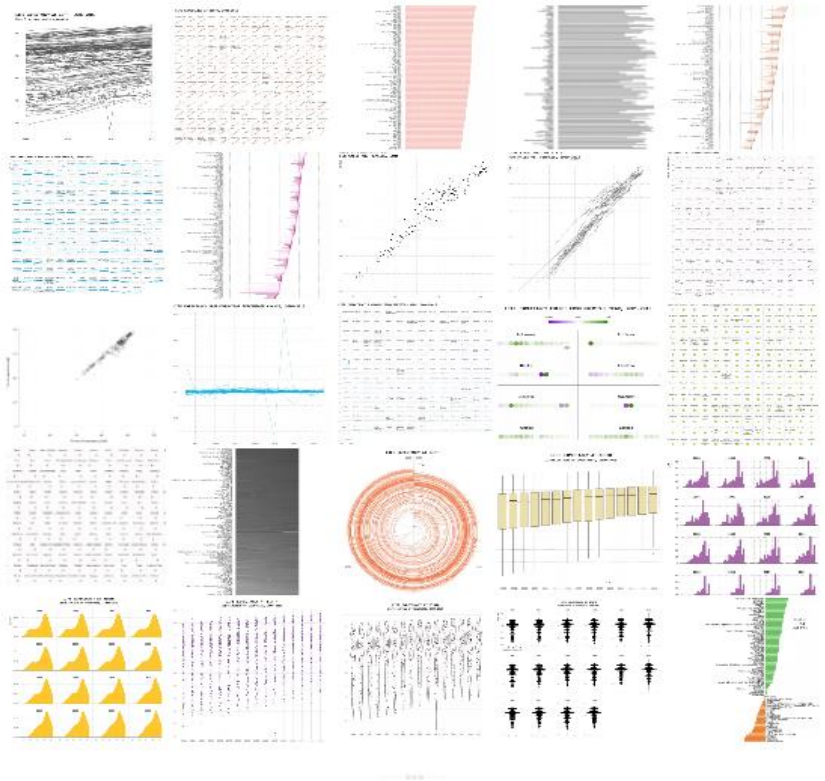
Navigating through graphs for data visualization

Choice of graphs



- Large number of available graphs
- The choice of graph highly depends on:
  - What kind of data do we have?
  - What exactly we want to show from the data?
- Different tools support various numbers and types of graphs

An example of extreme visualizion



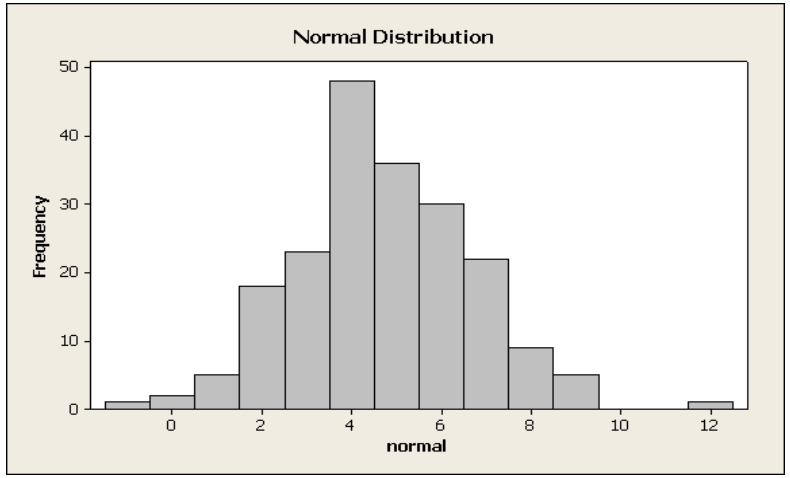
- A single dataset, contains life expectancy data depending on age in world's countries (WHO, 2000 – 2015), visualized in **25** ways
- http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways
- Each graph offers a unique view into the same dataset
- Educational, but mostly unnecessary in practice
- Correct choice of graph for visualization requires **experience** in analysis and displaying data, it is best to start with the basic graphs

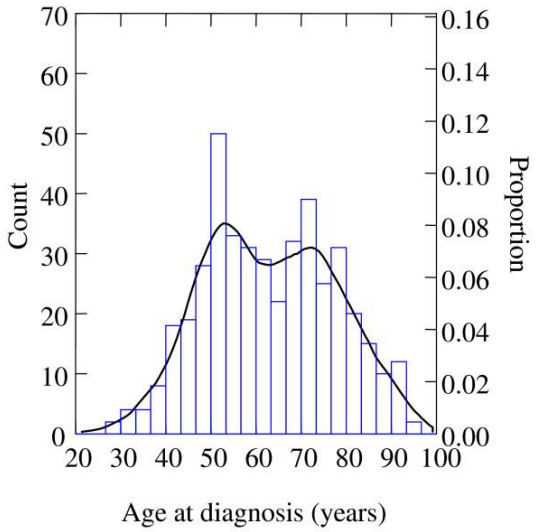
The most common graphs with respect to the number of considered variables

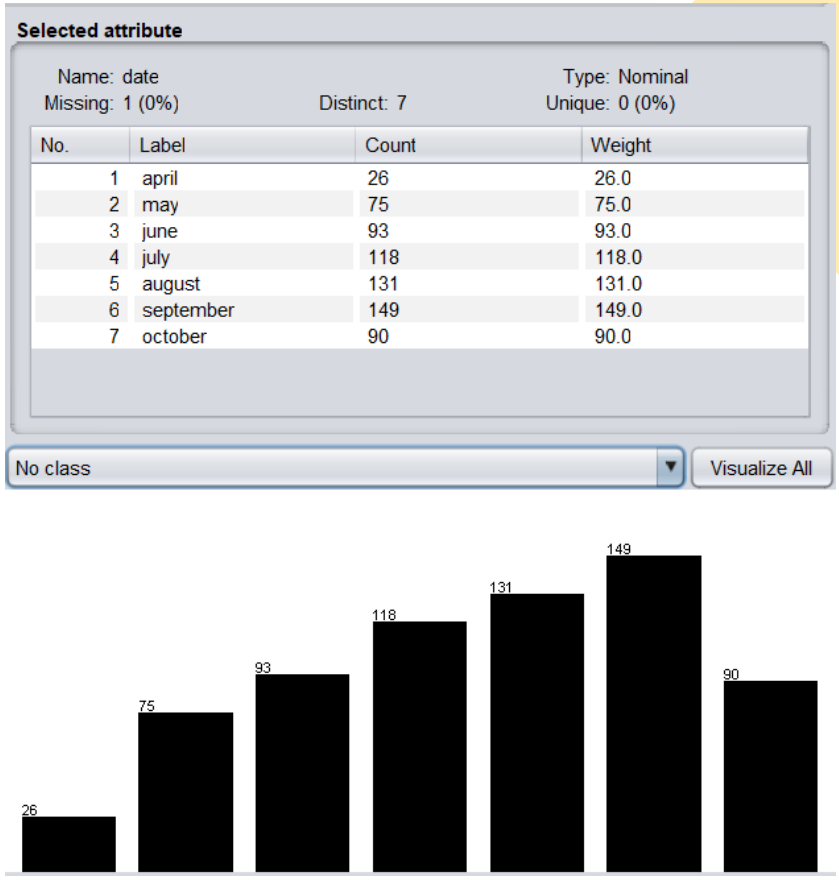
- **One variable**
  - Histogram, box plot (box-and-whisker plot), pie chart, column chart
- **Two variables**
  - Scatter plot, line chart
- **More than two variables**
  - Scatter plot matrix, stacked plot, bubble chart, heatmap, 3D area chart (surface chart), radar chart (spider chart, web chart, star chart)

Histogram

- Shows single variables
- **Categorical** (the usual histogram) and
- **Numerical** (value discretization or line histogram)



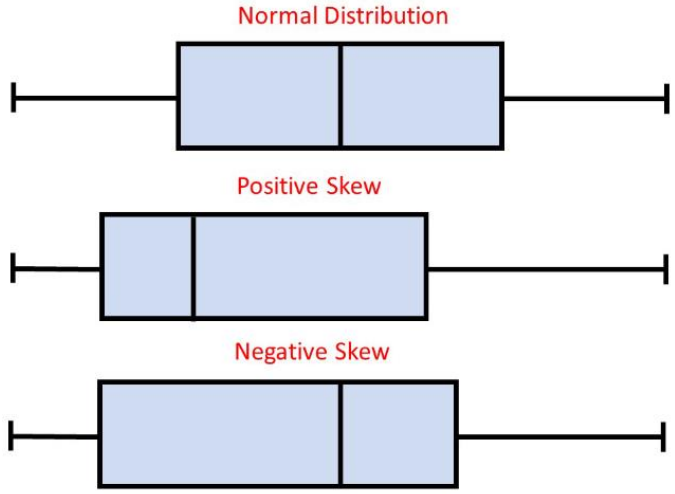


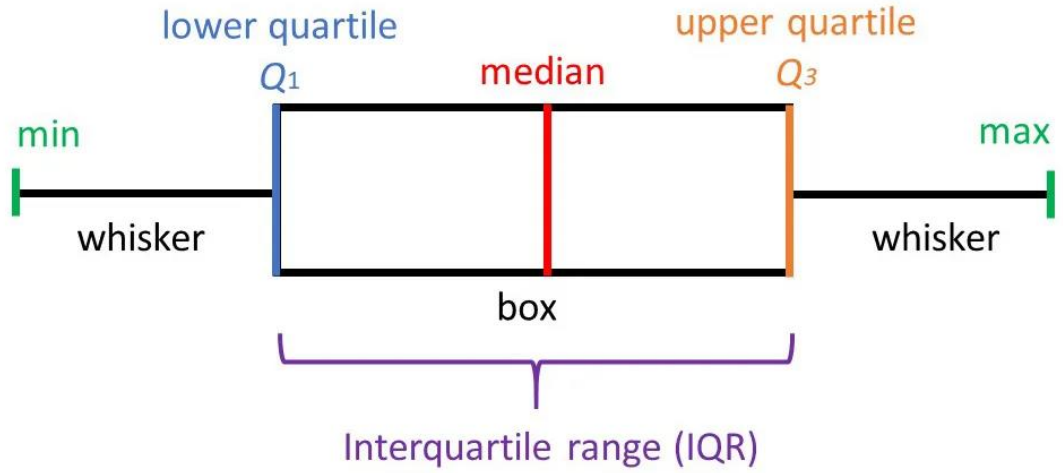


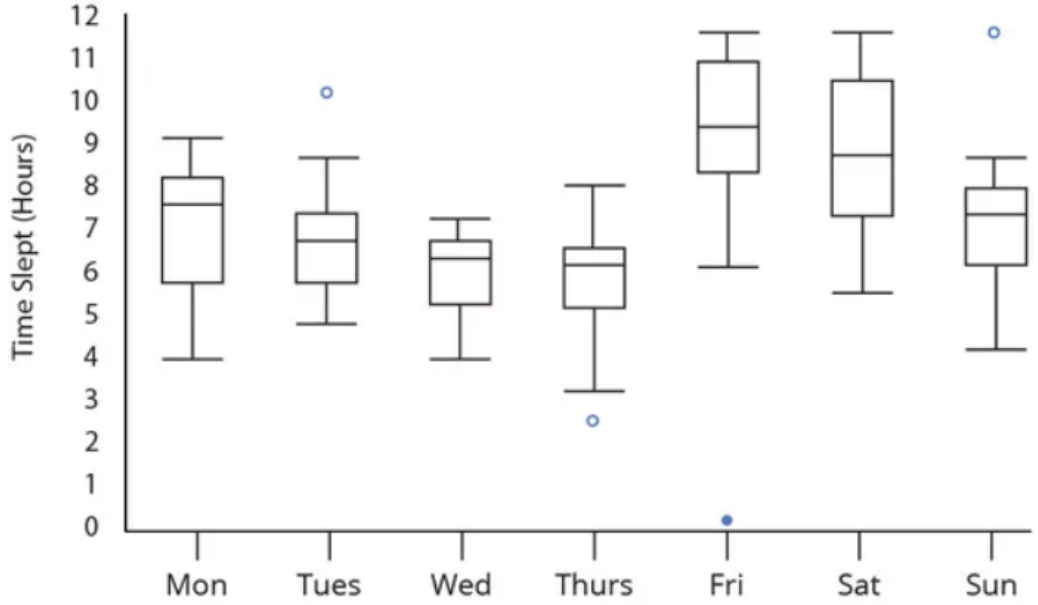
Source: ADA, 3rd lect., EPFL, 2020.

Box plot

- Shows a **numerical** variable, but possibly also multiple variables for comparison
- Suitable for showing the level of skeweness and variability of data value distribution





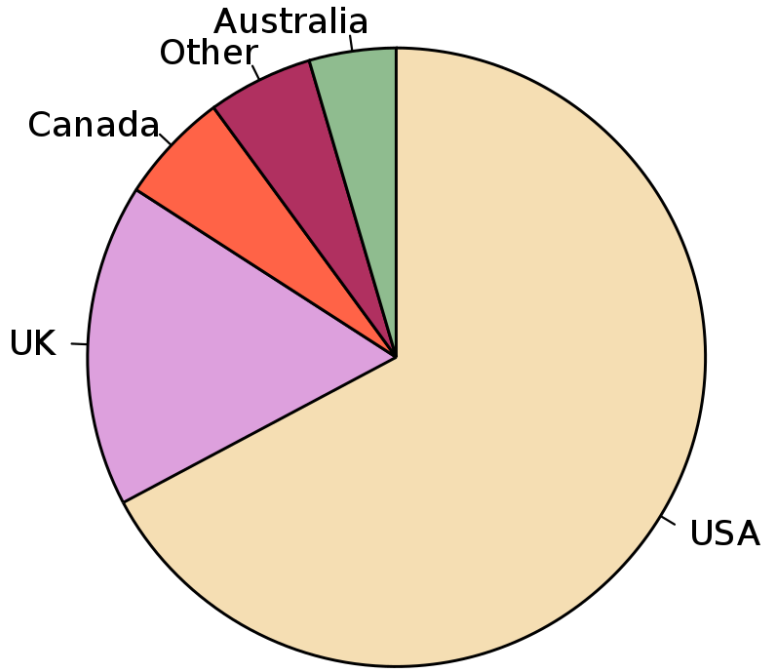


Source: McLeod, S. A. (2019, July 19). What does a box plot tell you? Simply psychology: https://www.simplypsychology.org/boxplots.html

Pie chart

- Showing an individual, **categorical** variable
- Shows a relative relation between the number of examples of individual categories of a specific variable

- In this example: the number of people having the maternal language English, categories are world countries

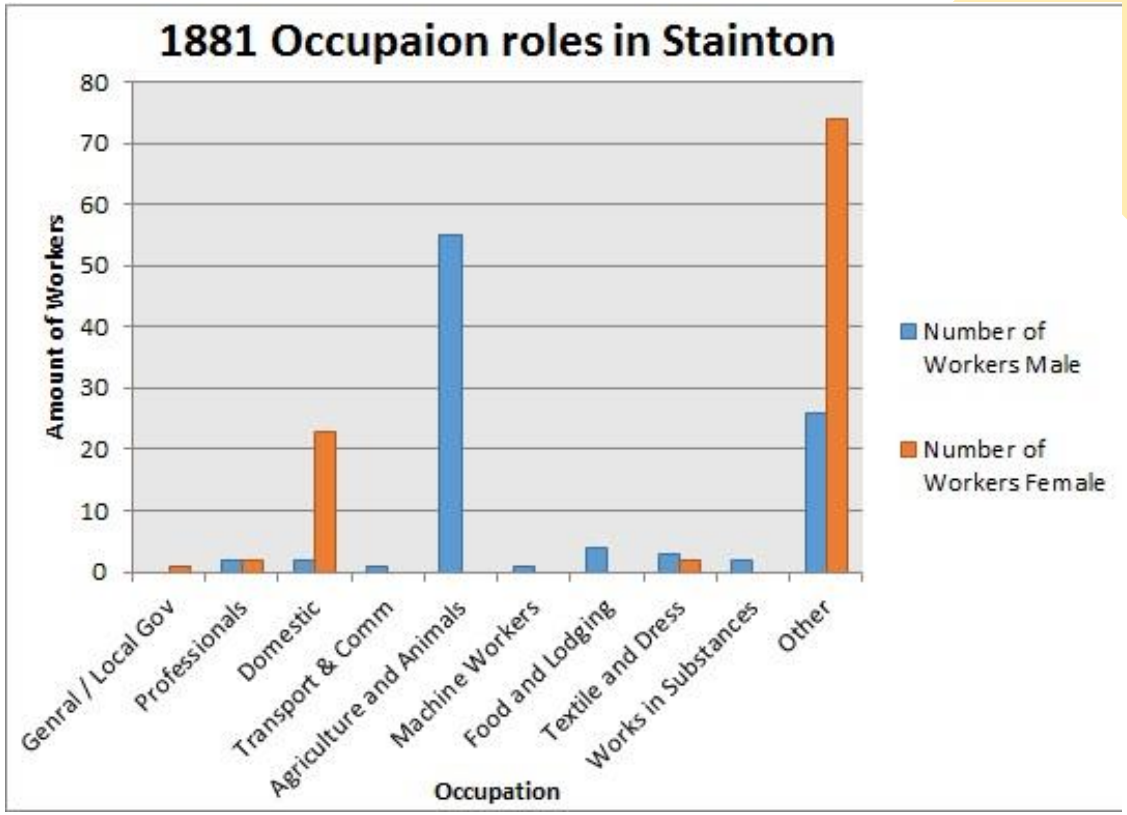


Source: https://en.wikipedia.org/wiki/Pie_chart#/media/File:English_dialects1997.svg

Column chart

- Shows individual **categorical** variables
- Each category has its own name and one or more corresponding columns
- There is more than one column if some other categorical variable's influence on the showed variable is considered in parallel

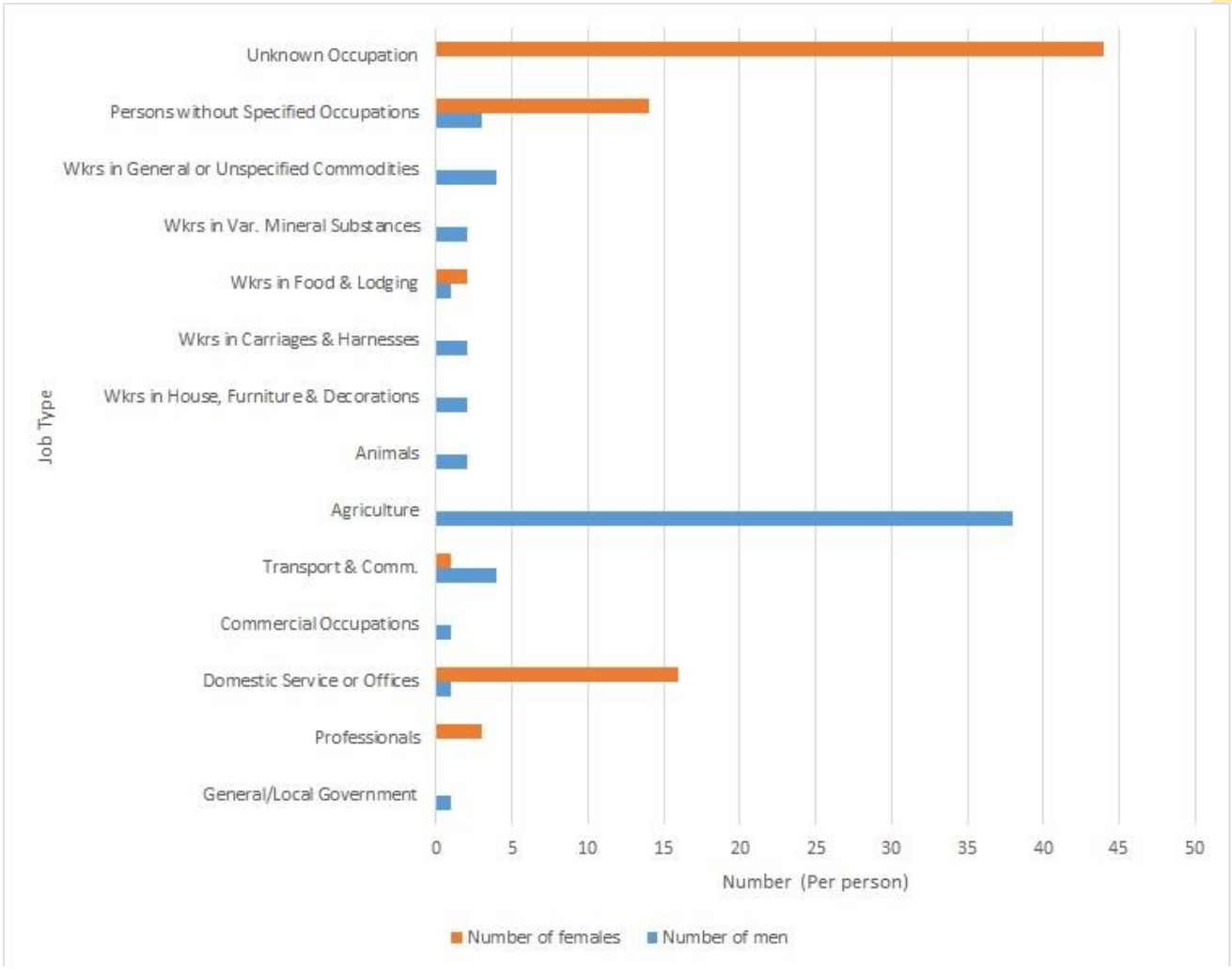
- In this example: variable *Occupation* has 10 categories, each category shows the number of examples for male and female workers (the influence of the second variable – gender)



Source: https://commons.wikimedia.org/wiki/Category:Demographic_bar_charts#/media/File:1881_bar_chart_paint.jpg

Bar chart

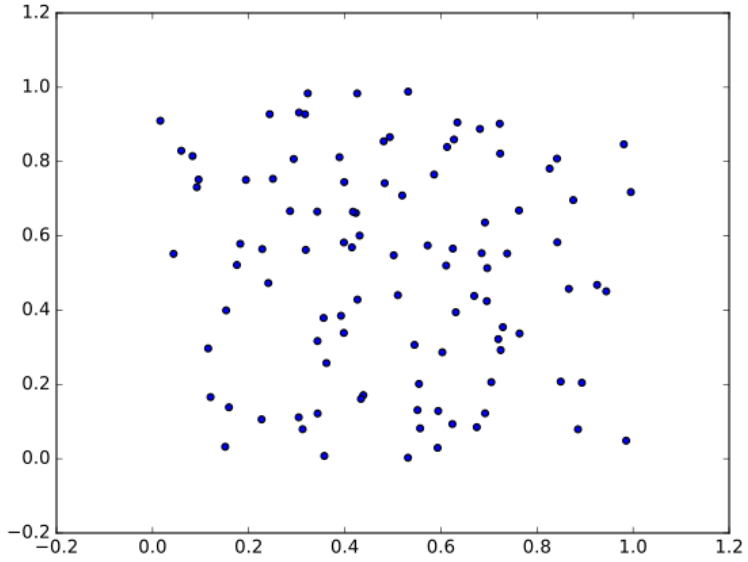
- A variant of column chart – the same as column chart, only the bars are shown horizontaly
- The choice between column chart and bar chart is mostly a matter of taste

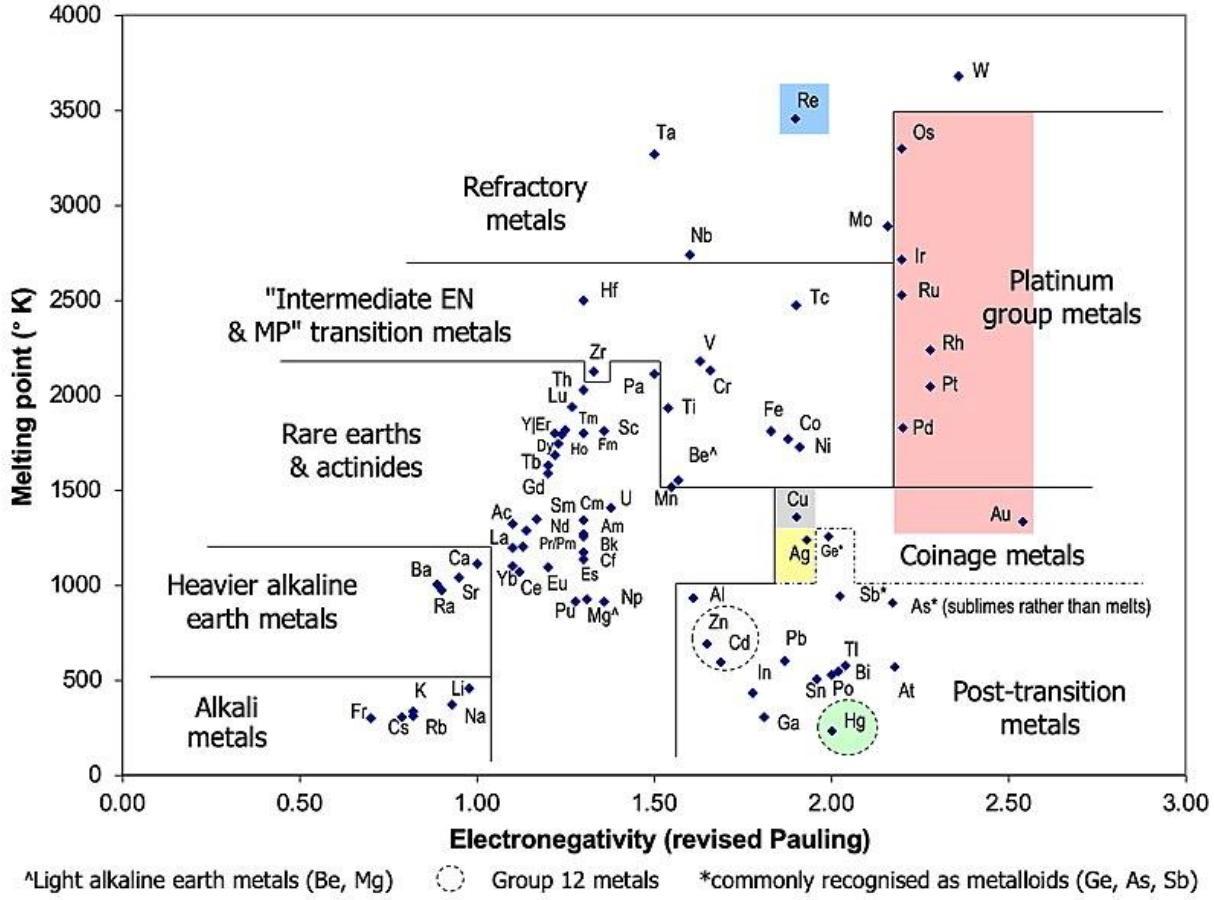


Source: https://commons.wikimedia.org/wiki/Category:Demographic_bar_charts#/media/File:Boyton_occupation_chart_1881.jpg

Scatter plot

- Shows relations **between two numerical** variables
- Can be simplistic or more complex
- Suitable for perceiving correlation

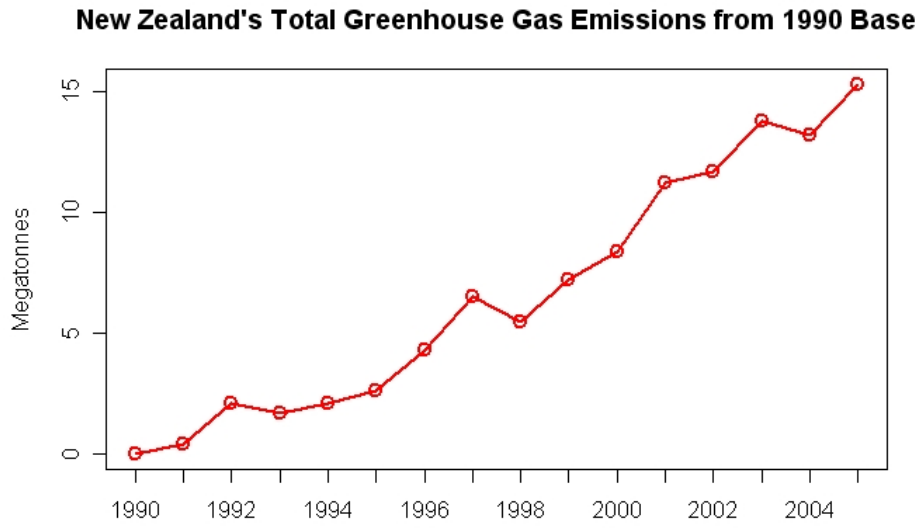


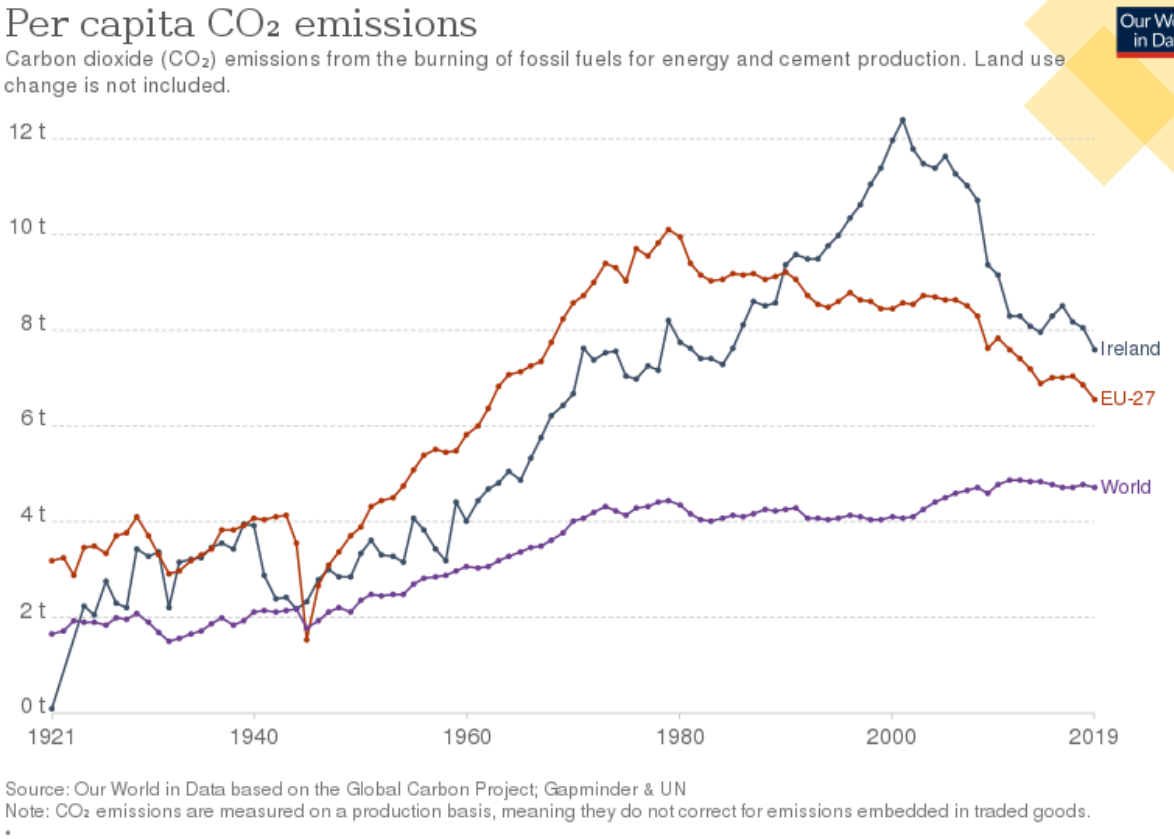


Sources: https://commons.wikimedia.org/wiki/File:Mpl_example_scatter_plot.svg
https://commons.wikimedia.org/wiki/File:Scatter_plot_of_EN_%26_MP_with_NM_shaded.jpg

Line chart

- Shows **one or more numerical** variables, most commonly with respect to the variable of time
- Suitable for consideration and analysis of time series data

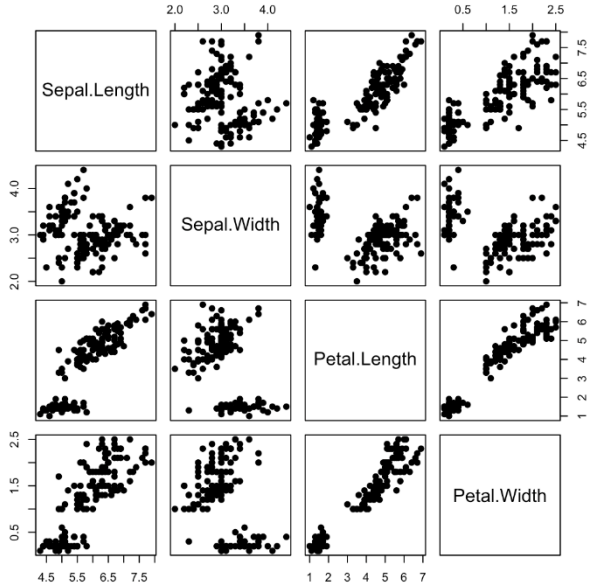


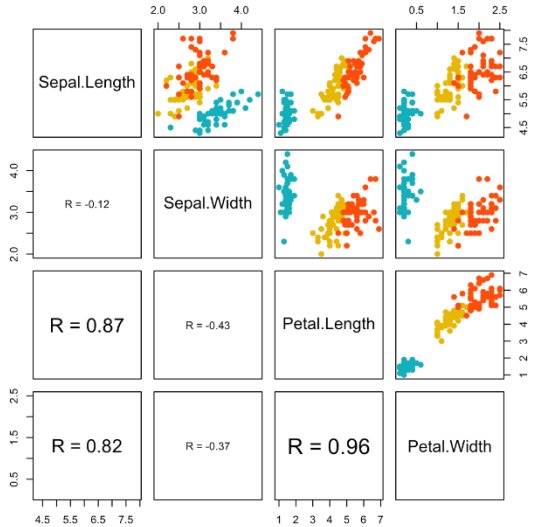


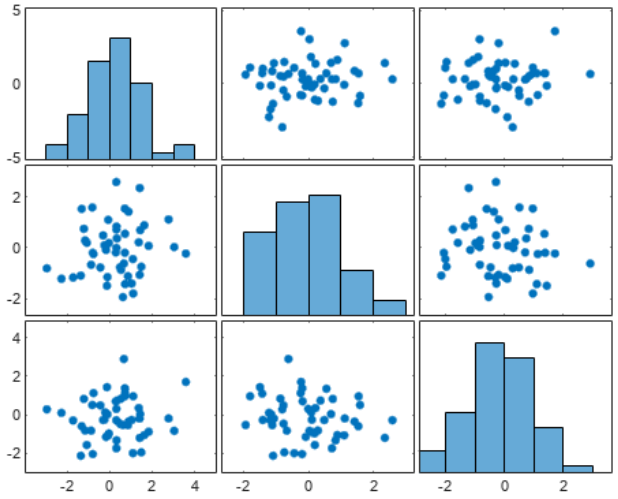
Sources: https://commons.wikimedia.org/wiki/File:New-Zealand-greenhousegases-1990-2005-line-chart.jpeg
https://commons.wikimedia.org/wiki/File:Ireland_v_EU-27_v_World_per_capita_CO2_emissions.svg

Scatter plot matrix

- Shows relation among **multiple paired numerical variables**
- Suitable for quick perception of possible correlations among variables
- Variants of scatter plot matrix are possible that also show variable histograms on the diagonal and correlation coefficient values on one side of the matrix





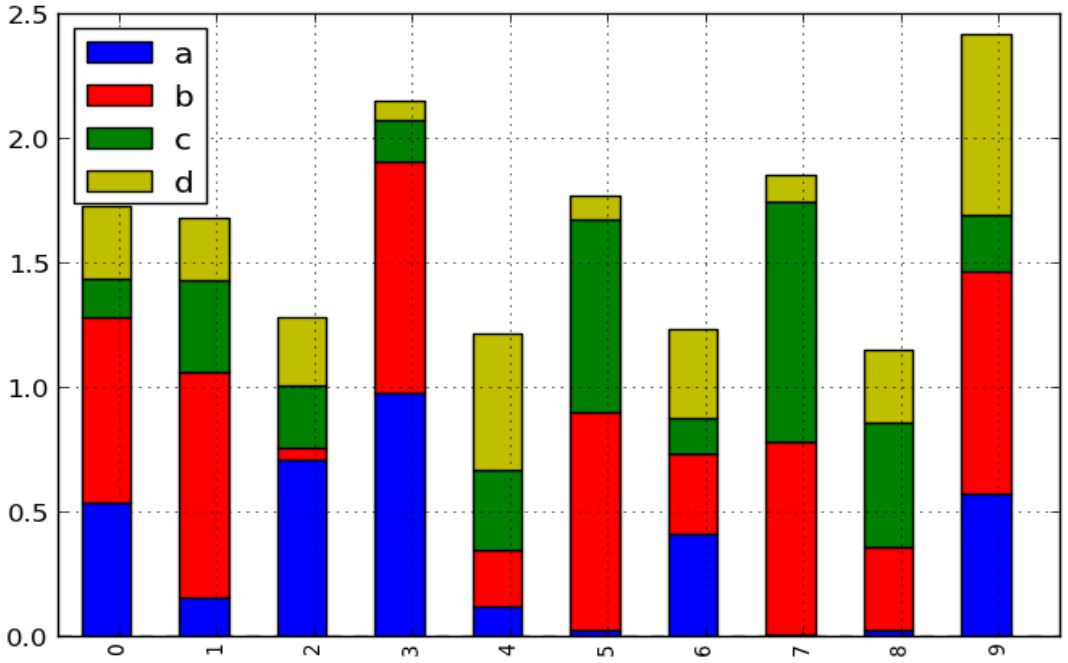


Sources: http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs
http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs
https://www.mathworks.com/help/matlab/ref/plotmatrix.html

Stacked plot

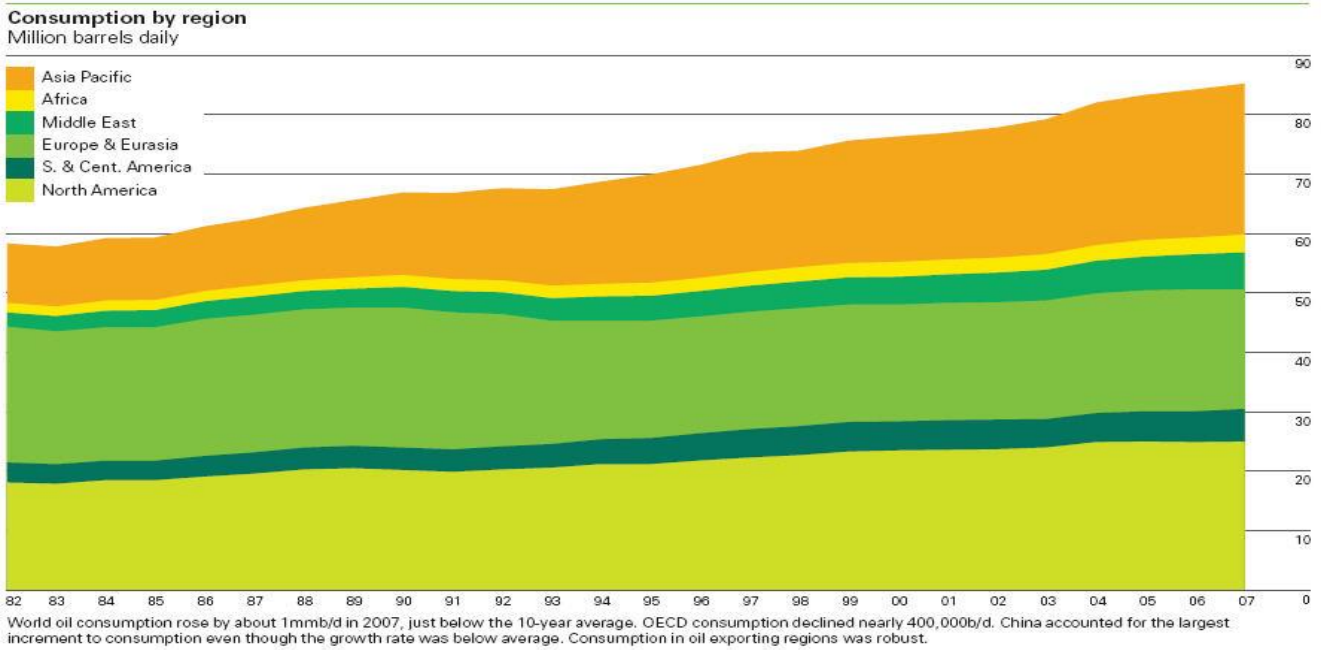
- Shows interrelations among **three or more** variables, comes in several variants

Two categorical (0-9, a-d), one numerical variable (height of column)



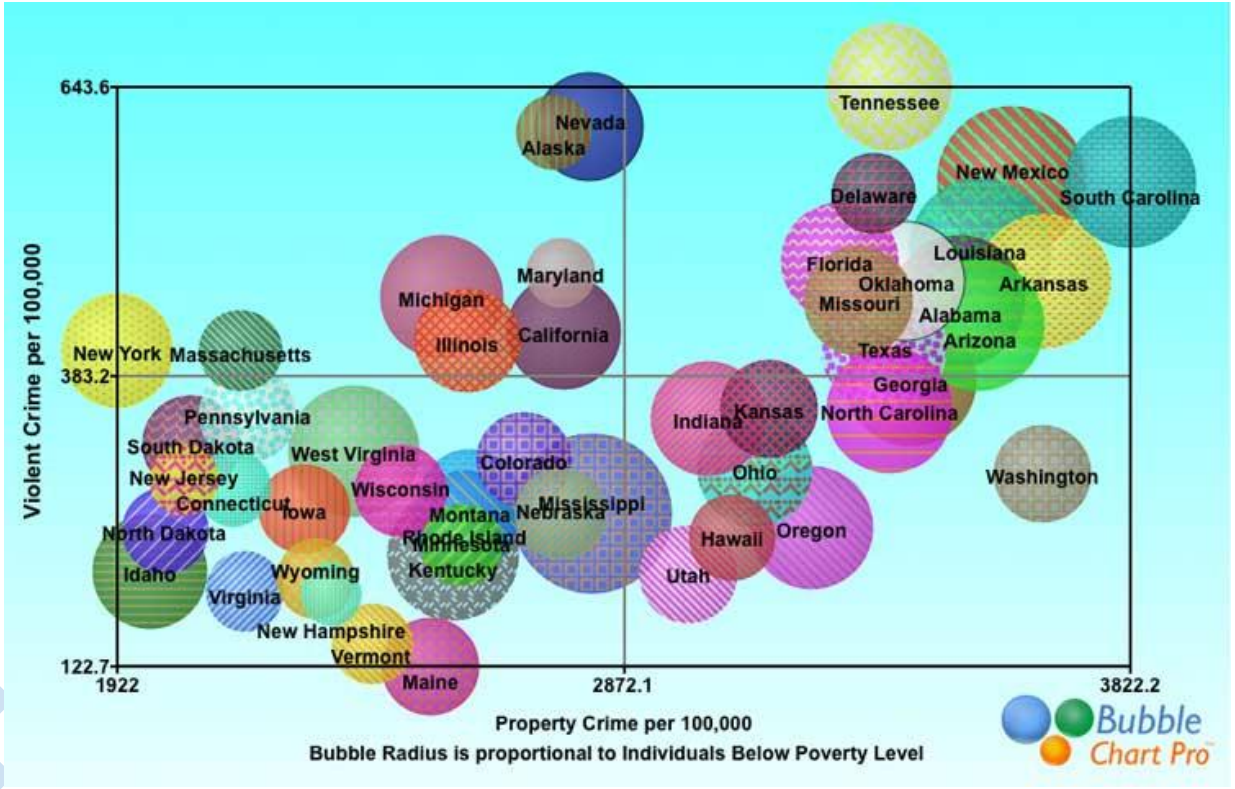
Sources: ADA, 3rd lecture, EPFL, 2020

One categorical (colors of continents), two numerical variables (years from 1982 to 2007, the amount of consumed oil)



Bubble plot

- Show interrelations among **three** (in 2D) or **four** (in 3D) **numerical** variables, the size and/or color of bubbles reflects values of one of the numerical variables



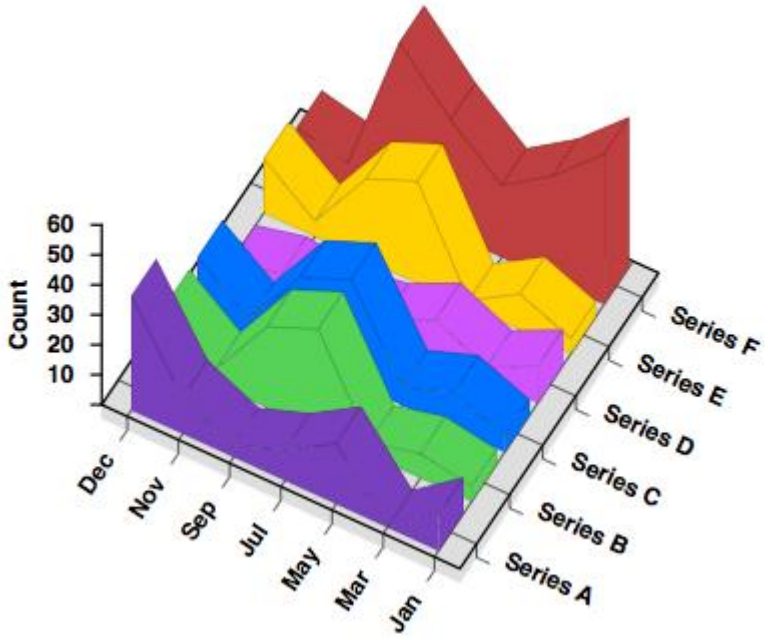
In this example:

- On the x-axis: property crime rate (e.g., house burglary)
- On the y-axis: violent crime rate (e.g., wounding)
- Bubbles reflect citizen percentage below the level of poverty, briefly: the larger the bubbles – the poorer the inhabitants
- A single example = a single US state

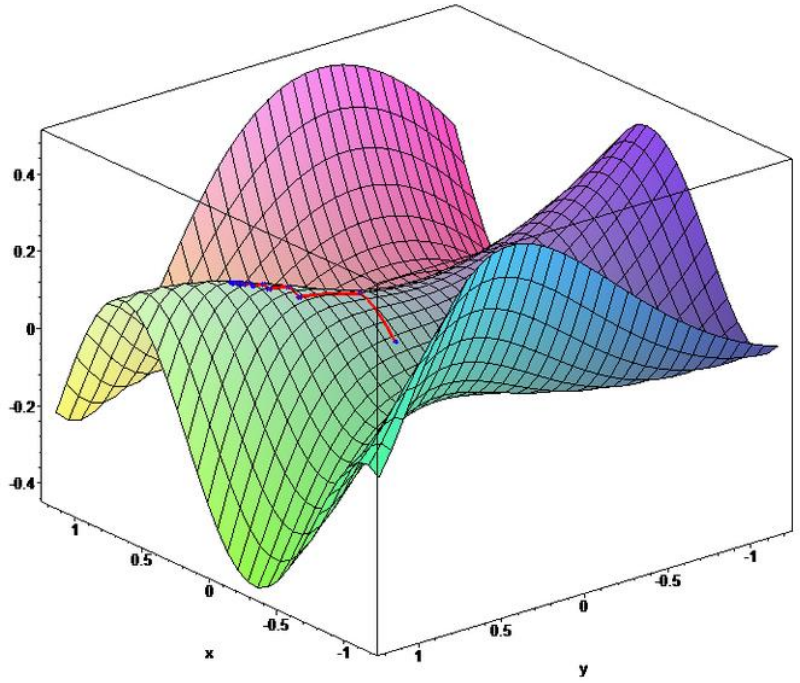
Source: https://commons.wikimedia.org/wiki/Category:Bubble_charts#/media/File:Bubble_Chart_of_Crime_versus_Poverty_in_50_states.jpg

3D area chart

- Shows interrelations among **three** or **four numerical** variables (if some of them are categorical, then we are talking about a column 3D chart)



Source: https://www.gigawiz.com/3d-area.html



Source: https://commons.wikimedia.org/wiki/File:Gradient_ascent_(surface).png

Heatmap

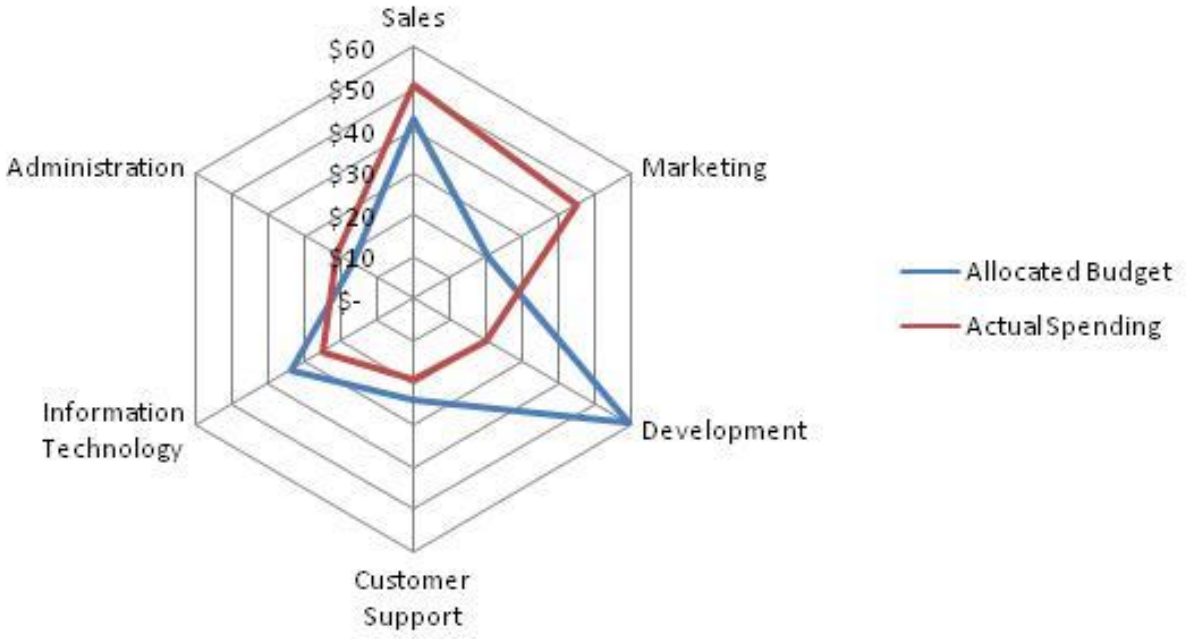
- Shows the relation among three variables (in 2D space) or four variables (in 3D space)
- One of the variables is numerical and is used for color grading
- Other variables that form a coordinate space are most commonly categorical or ordinal, but can also be numerical

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Average Monthly Temperatures at Central Park, New York													
2		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
3	2009	27.9	36.7	42.4	54.5	62.5	67.5	72.7	75.7	66.3	55.0	51.2	35.9	
4	2010	32.5	33.1	48.2	57.9	65.3	74.7	81.3	77.4	71.1	58.1	47.9	32.8	
5	2011	29.7	36.0	42.3	54.3	64.5	72.3	80.2	75.3	70.0	57.1	51.9	43.3	
6	2012	37.3	40.9	50.9	54.8	65.1	71.0	78.8	76.7	68.8	58.0	43.9	41.5	
7	2013	35.1	33.9	40.1	53.0	62.8	72.7	79.8	74.6	67.9	60.2	45.3	38.5	
8	2014	28.6	31.6	37.7	52.3	64.0	72.5	76.1	74.5	69.7	59.6	45.3	40.5	
9	2015	29.9	23.9	38.1	54.3	68.5	71.2	78.8	79.0	74.5	58.0	52.8	50.8	
10	2016	34.5	37.7	48.9	53.3	62.8	72.3	78.7	79.2	71.8	58.8	49.8	38.3	
11	2017	38.0	41.6	39.2	57.2	61.1	72.0	76.8	74.0	70.5	64.1	46.6	33.4	
12														

Sources: https://www.excel-easy.com/examples/heat-map.html

Radar plot

- Shows values of **five or more** variables of a particular example, more than one example is shown for comparison purposes



- In this example: two examples – allocated budget and actual spending are shown, measured by 6 variables (Sales, Administration...)

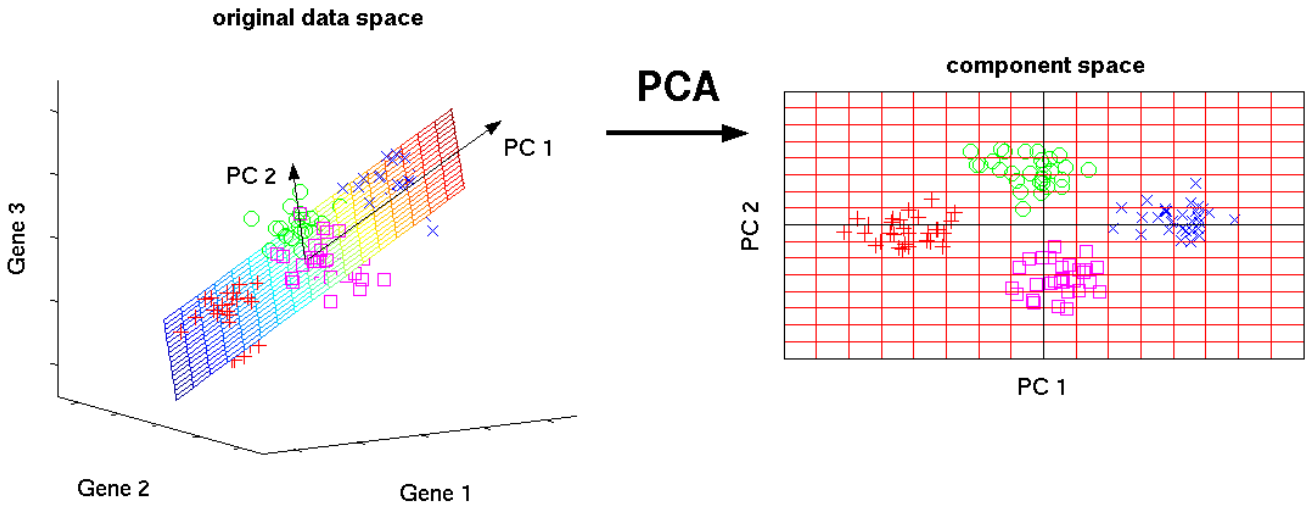
Sources: https://commons.wikimedia.org/wiki/File:Spider_Chart2.jpg

Dimensionality reduction methods with applications in visualization

- Many methods for dimensionality reduction, most of them applicable to visualization
- Assumption: by reduction, we preserve the initial information as much as possible, the initial features are **transformed** (linearly or nonlinearly)
- Here we consider only some of the most common dimensionality reduction methods used for visualization:
  - **Principal Component Analysis, PCA**
  - **t-SNE** (*t-distributed Stochastic Neighbor Embedding*)

Principal component analysis

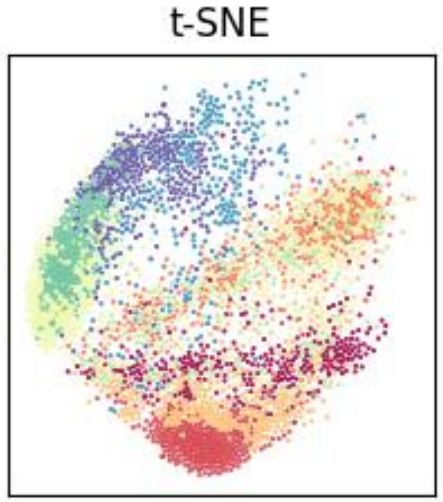
- 1901, Karl Pearson
- Enables visualization of high-dimensional numerical data in a low-dimensional (2D or 3D) space
- **Principal components**
  - Expressed analytically as a **linear combination of starting features** in the order so that they cover the largest variability in data
  - Mutually orthogonal
- Usually, two or three principal components are kept for visualization
- Useful for discovering **clusters** of data



t-SNE

- One of the **manifold learning** techniques, Maaten and Hinton, 2008
- Searches for a low-dimensional structure so that the properties of groups in a higher dimension remain preserved
- The relationship among points in a high-dimensional space is represented by Gaussian mutual probabilities, while in the low-dimensional embedding space it is represented by Student t-distribution (if is more flat)
- Kullback-Leibler divergence between joint probabilities in the original space and the embedded space is minimized by gradient descent method
- The method is computationally demanding but often leads to great results

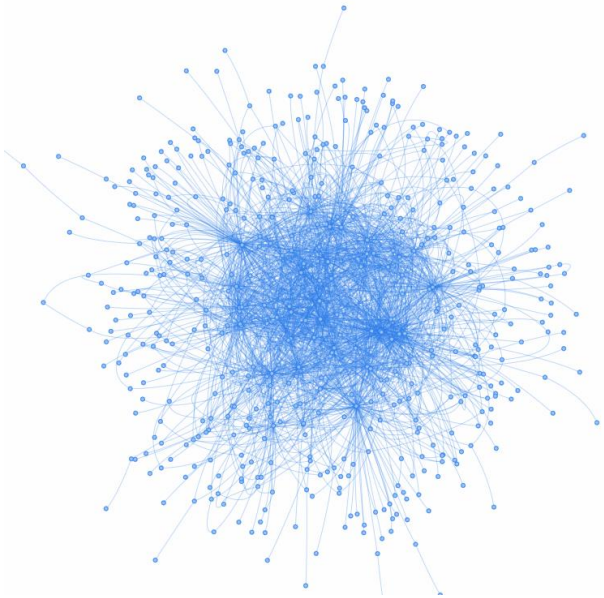
Fashion-MNIST: 28x28, 60000 primjera, 10 klasa



Source: Eugen Vušak, „Manifold learning techniques for increased efficiency of side-channel analysis" diploma thesis, FER, 2020.

Other graphs and visualizations

- Word cloud
  - Application in text visualization (*bag of words*)
  - https://www.wordclouds.com/
- Text network graph
  - Shows connections among words in a text (context)
  - https://infranodus.com/
- …



Principles and best practices of visualization

The choice of colors and tones

- In grayscale tones, one needs to have a **significant difference in intensity** to perceive it
  - Avoid ”slighty noticable difference”, also see Weber-Fechner law: https://en.wikipedia.org/wiki/Weber%E2%80%93Fechner_law
  - Better to ”play it safe” and use a very small number of very different tones

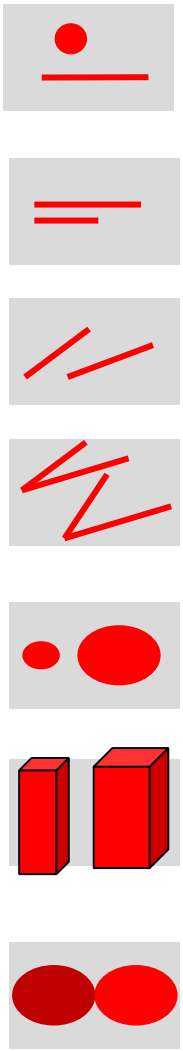
Reality: continuous (almost) spectrum of tone colors

What is easily perceived by the human eye – a discrete (small) number of different tones

Perception of magnitude

Most accurate

Least accurate



Position

Length

Slant

Angle

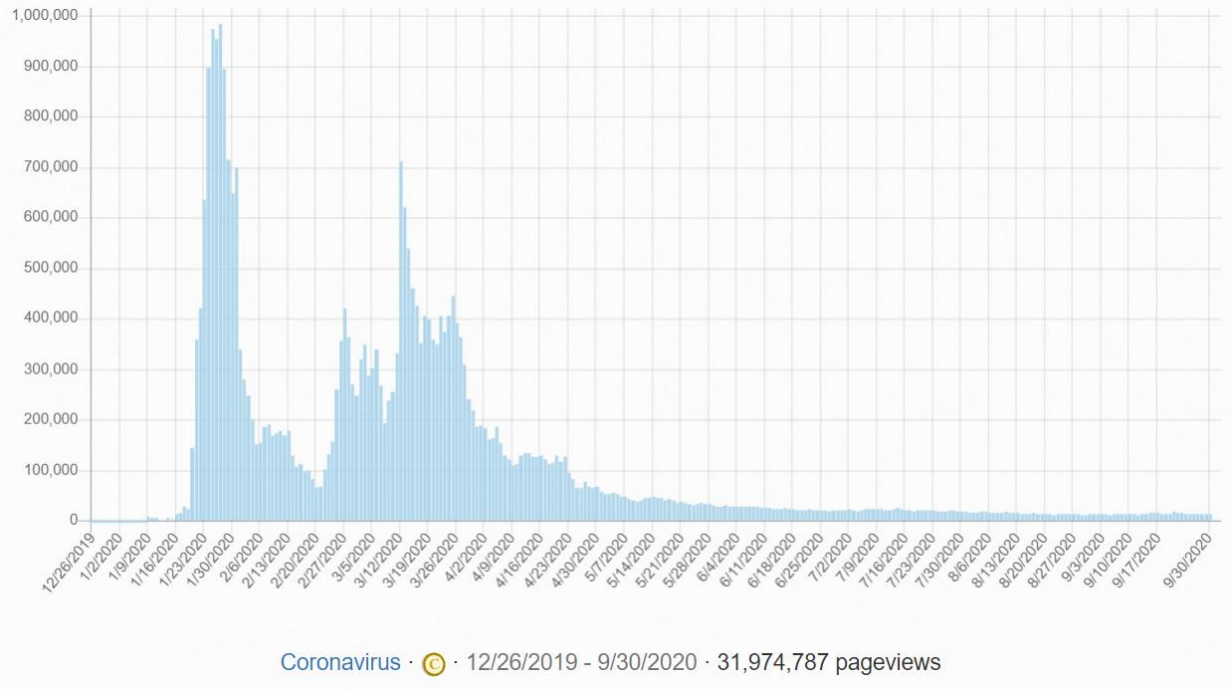
Area

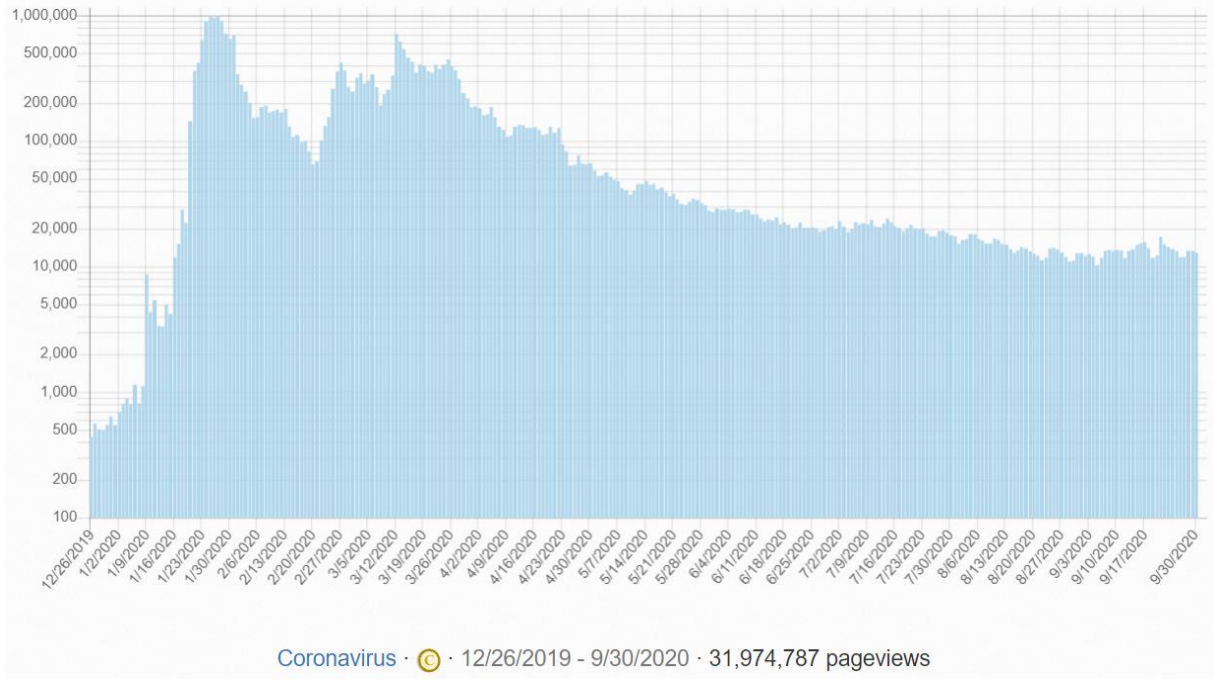
Volume

Tone-saturation-color density

- Focus on the use of picture elements with the most accurate perception of **magnitued and difference**
- Avodi small (and unclear) differences

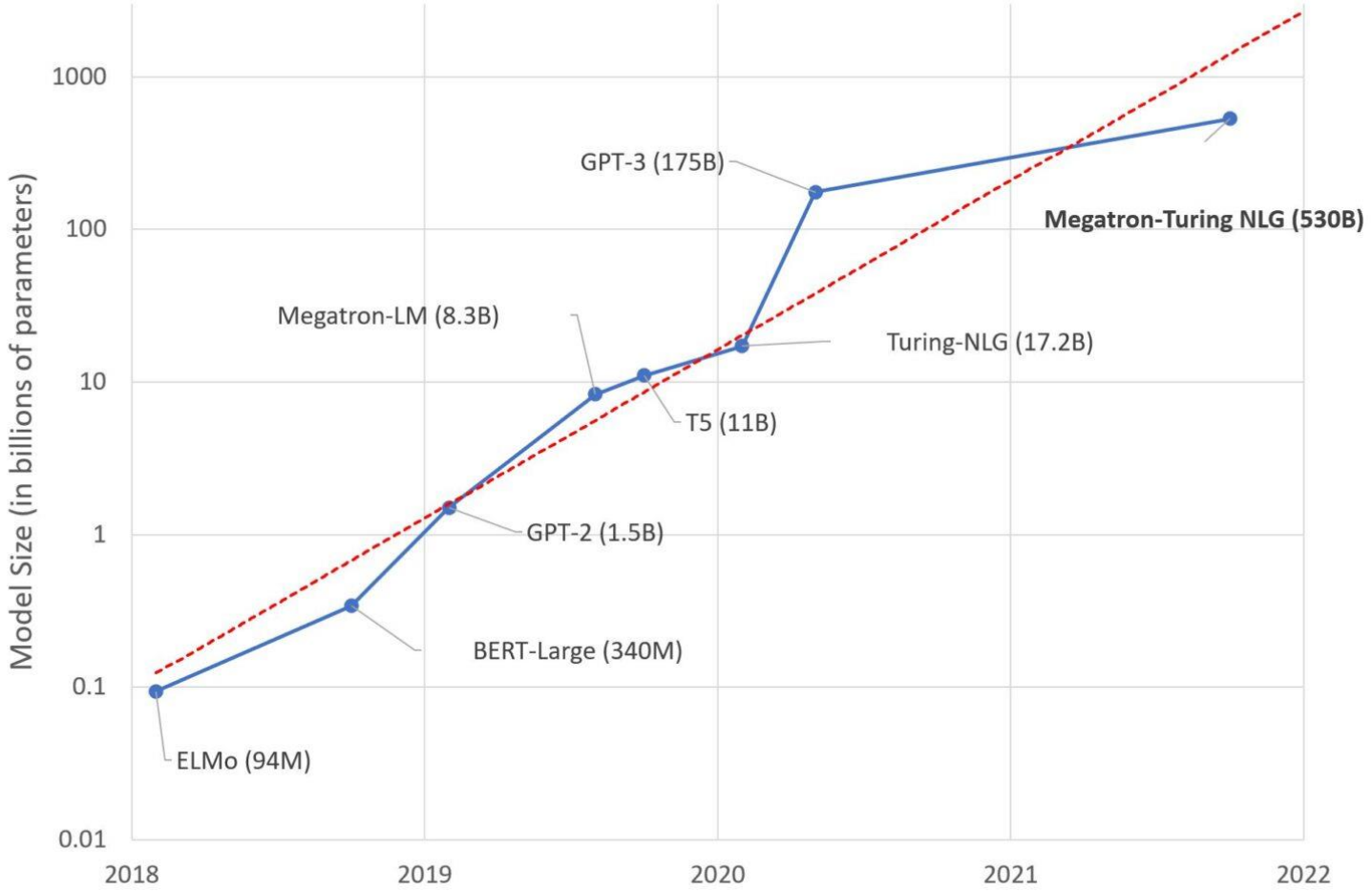
Watch out for the scale of graphs





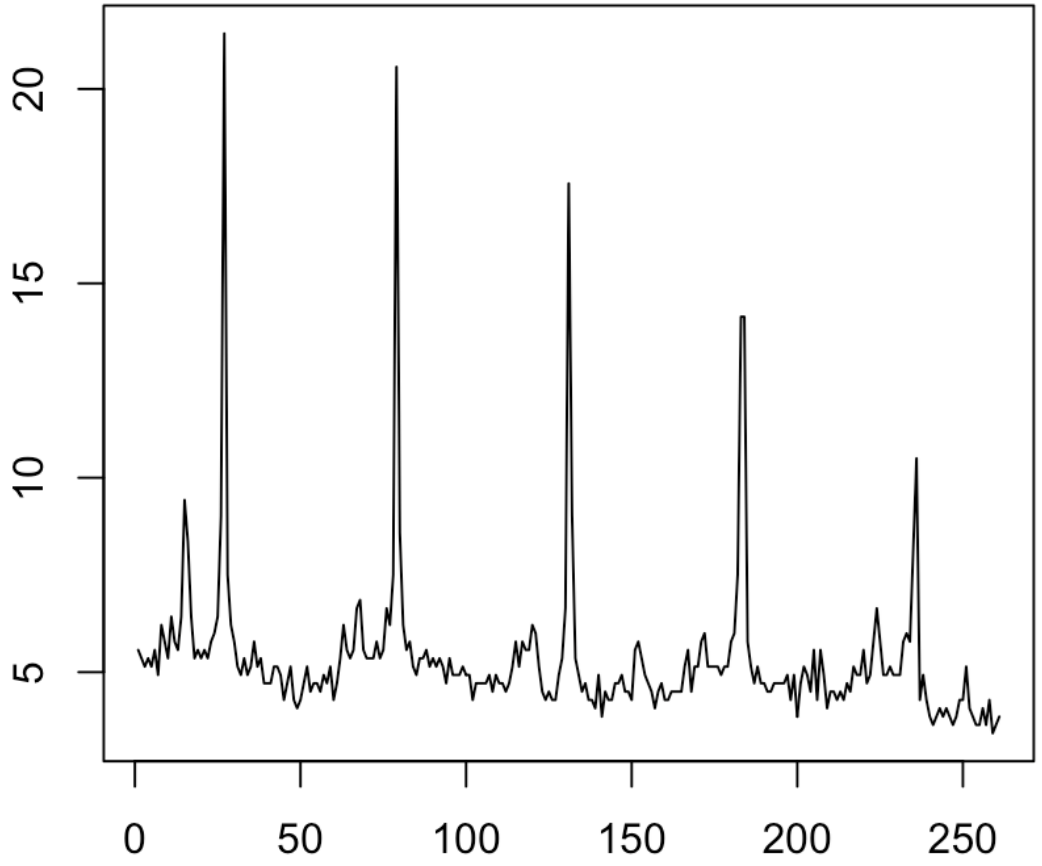
- Sometimes, simple linear scale is better to show information, sometimes it is logarithmic
- It is important to perceive scale on time!

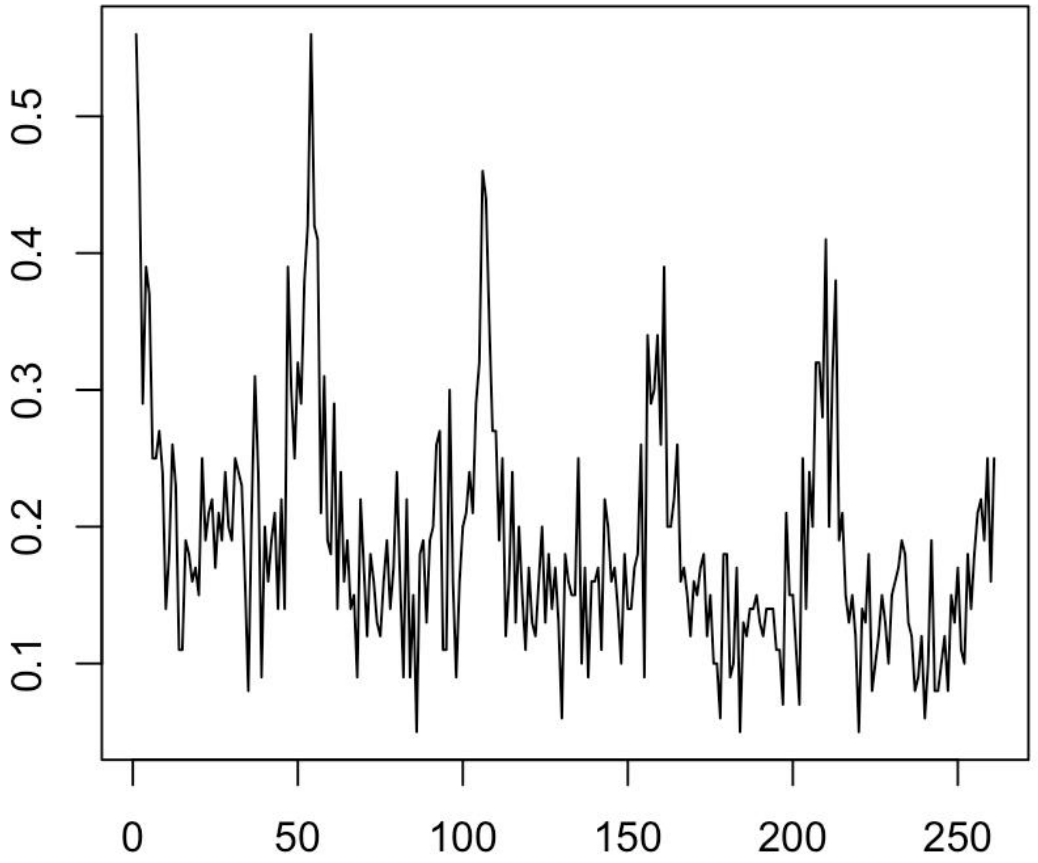
Watch out for the scale of graphs



Source: https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/
https://towardsdatascience.com/counting-no-of-parameters-in-deep-learning-models-by-hand-8f1716241889

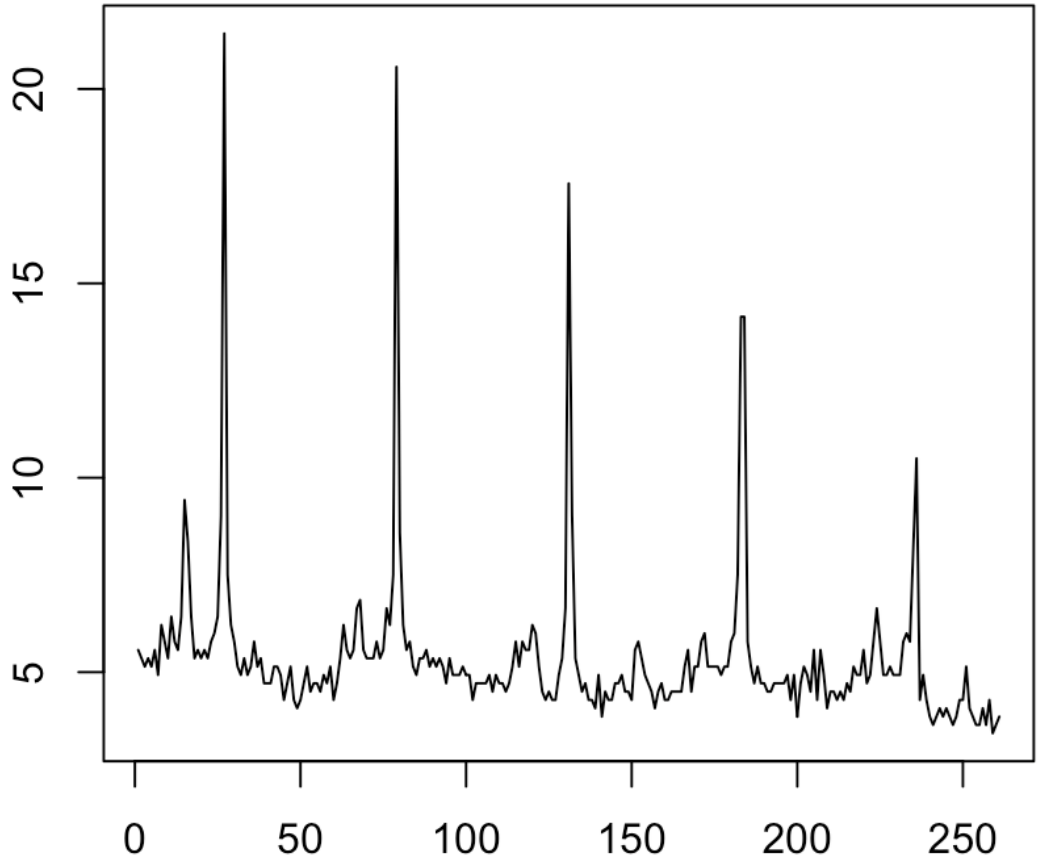
Answer quickly: which time series has a higher average value?

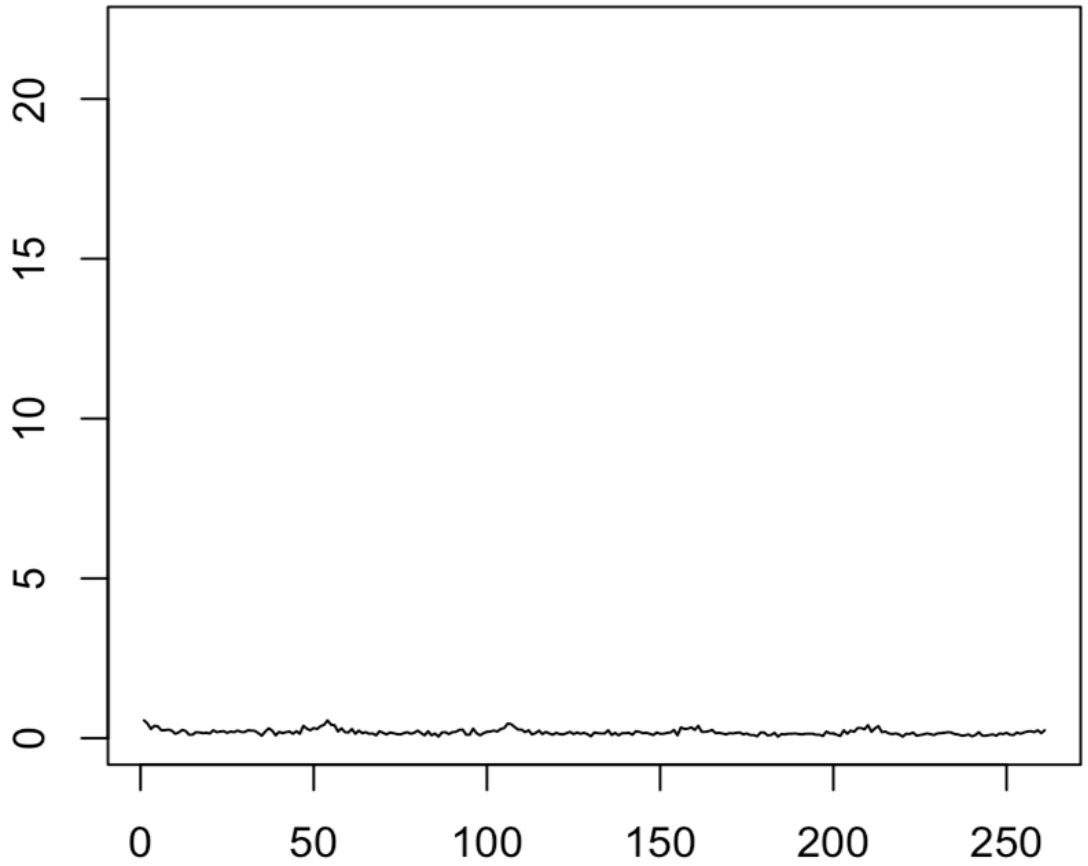




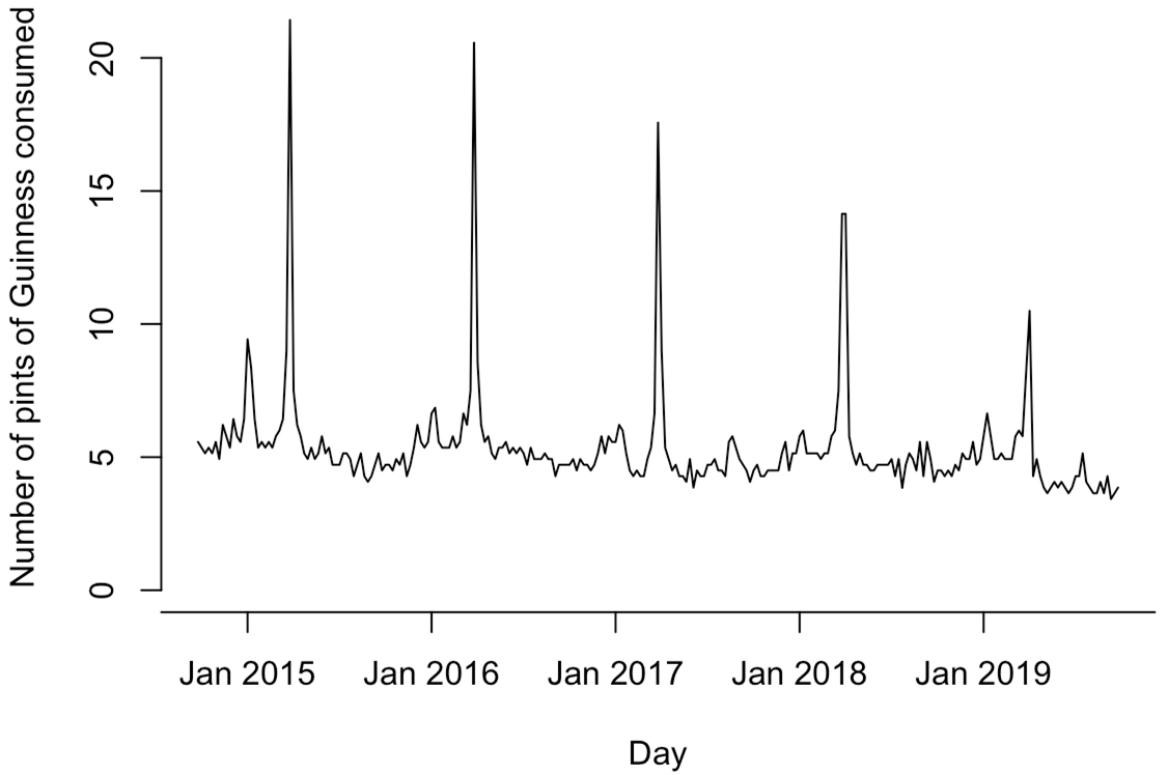
Source: ADA, 3rd lect., EPFL, 2020.

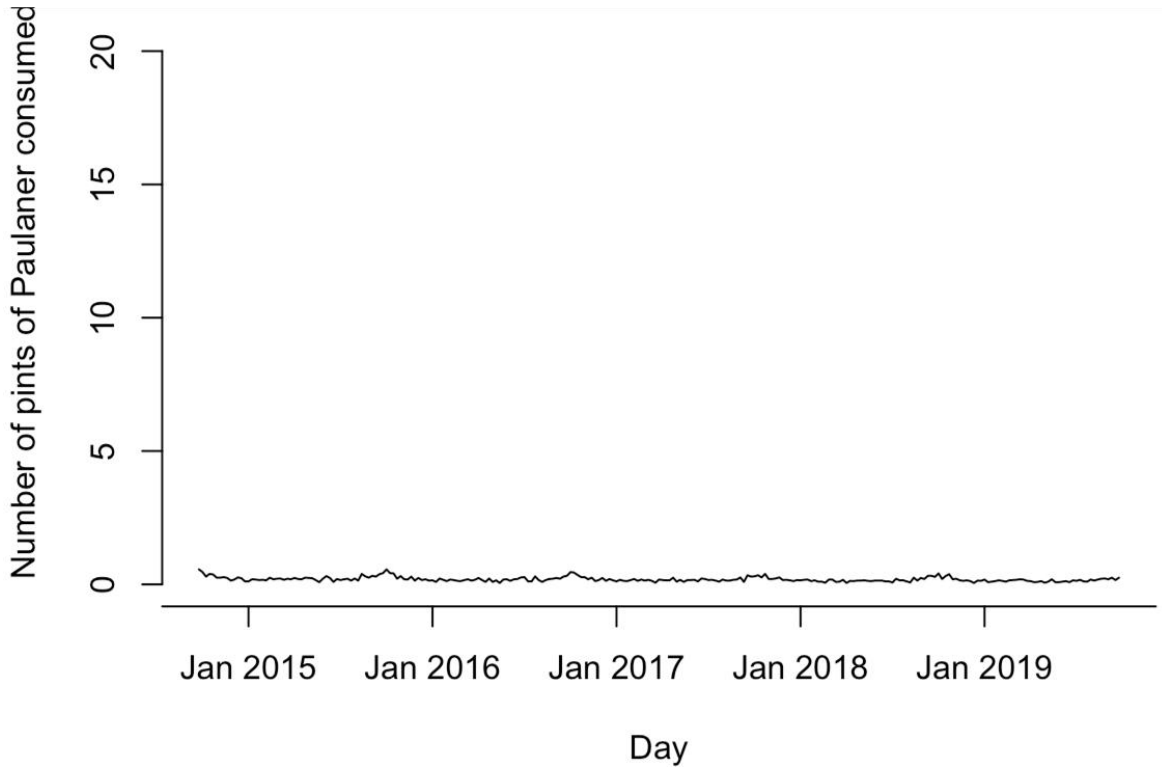
Actual situation



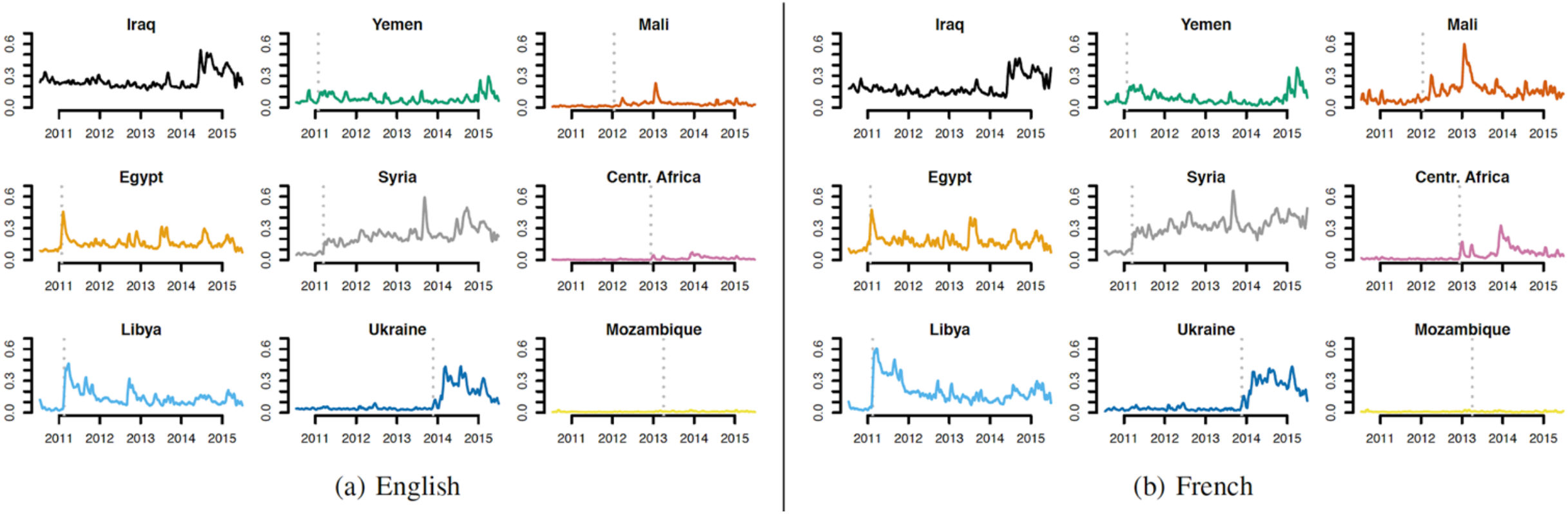


Always label the axes!





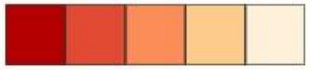
Colors and scales nees to be used consistently between various examples of visualization!



Source: ADA, 3rd lect., EPFL, 2020.

Schemes and color recognition

- Different color schemes
  - **Sequential** – for ordered numerical variables, lower values for lighter tones, higher values for darker tones

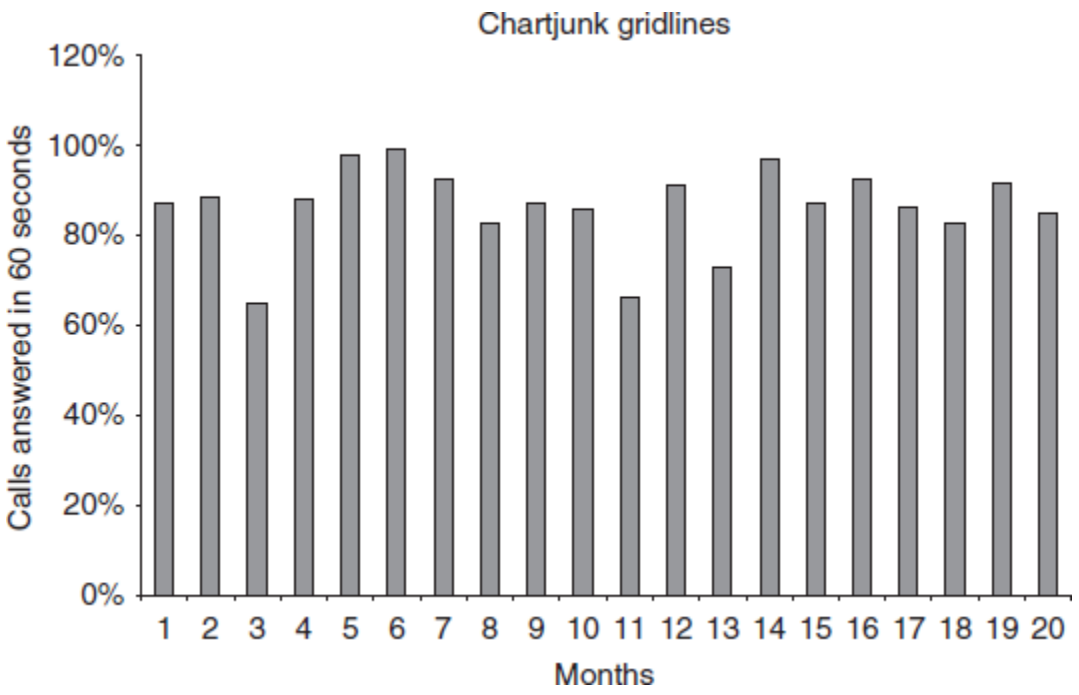


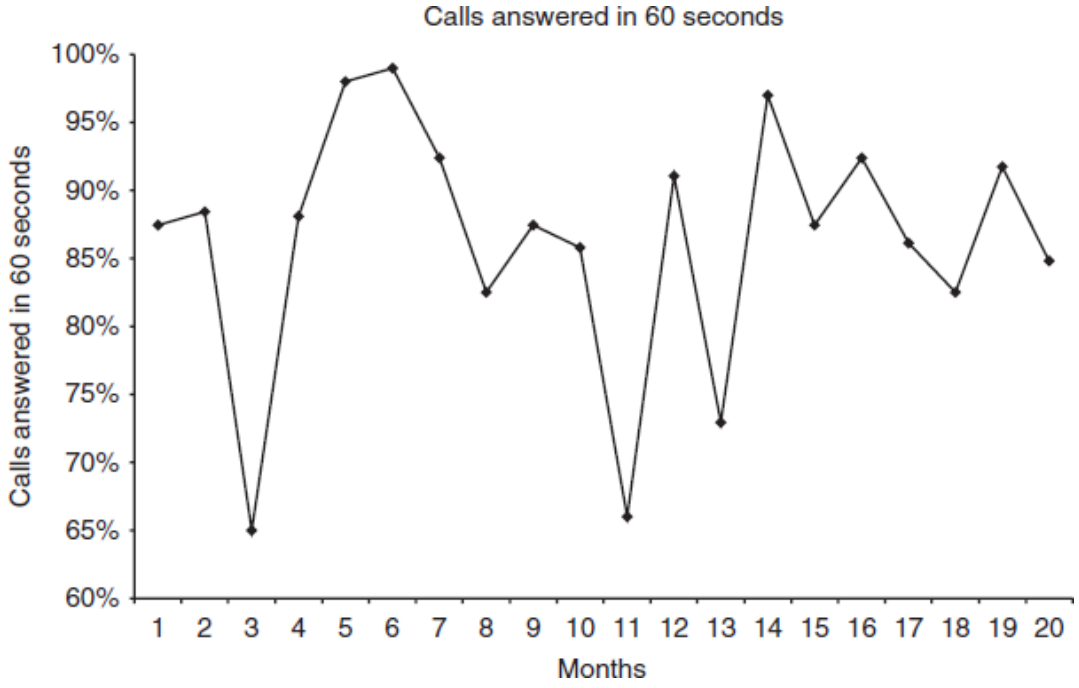
  - **Divergent** – focus is on the mean value with the lightest tone, edge values have darker tones, it is used for numerical variables if one wants to stress the mean value and edge values

  - **Categorical** (qualitative) – used for categorical data, the color are selected independently of class

- Always use **colorblind-safe color palettes**
- About 8% of males has some form of problem with color recognition (only about 0.4% of women)
- Check webpage: https://colorbrewer2.org/

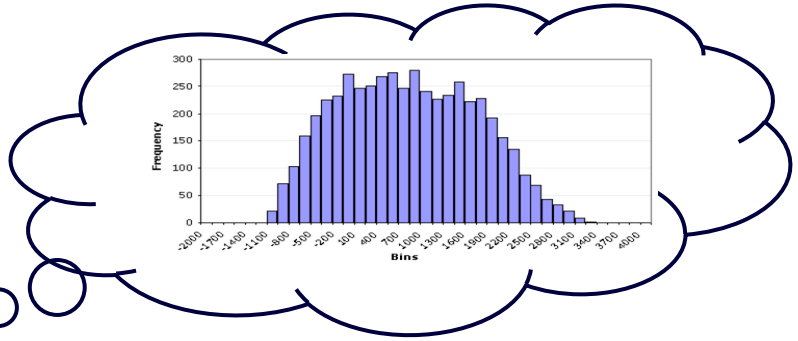
Pay attention to the choice of graph, optimize color ”expenditure”



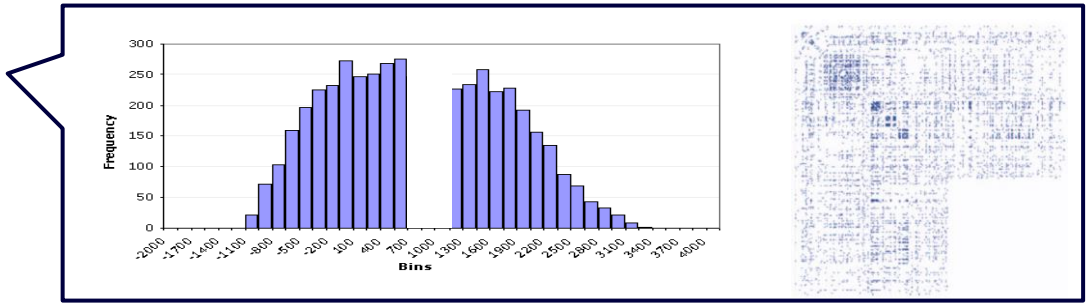


Source: https://www.accessengineeringlibrary.com/content/book/9780071749091/chapter/chapter4

Weird data

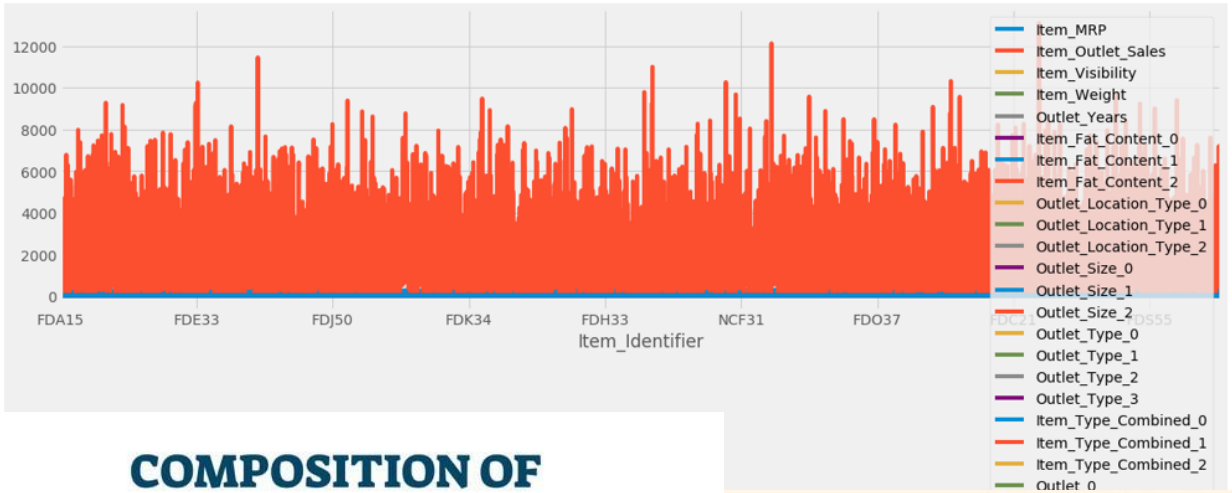


- One needs to have a theory how data should look like
- Some data are very hard to explain
- Never ignore weird pattern in data, always consider them very carefully!

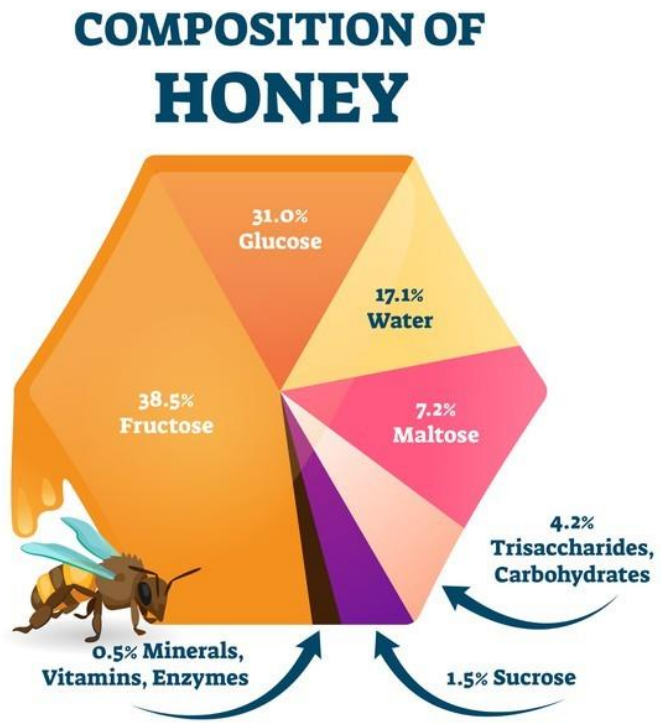


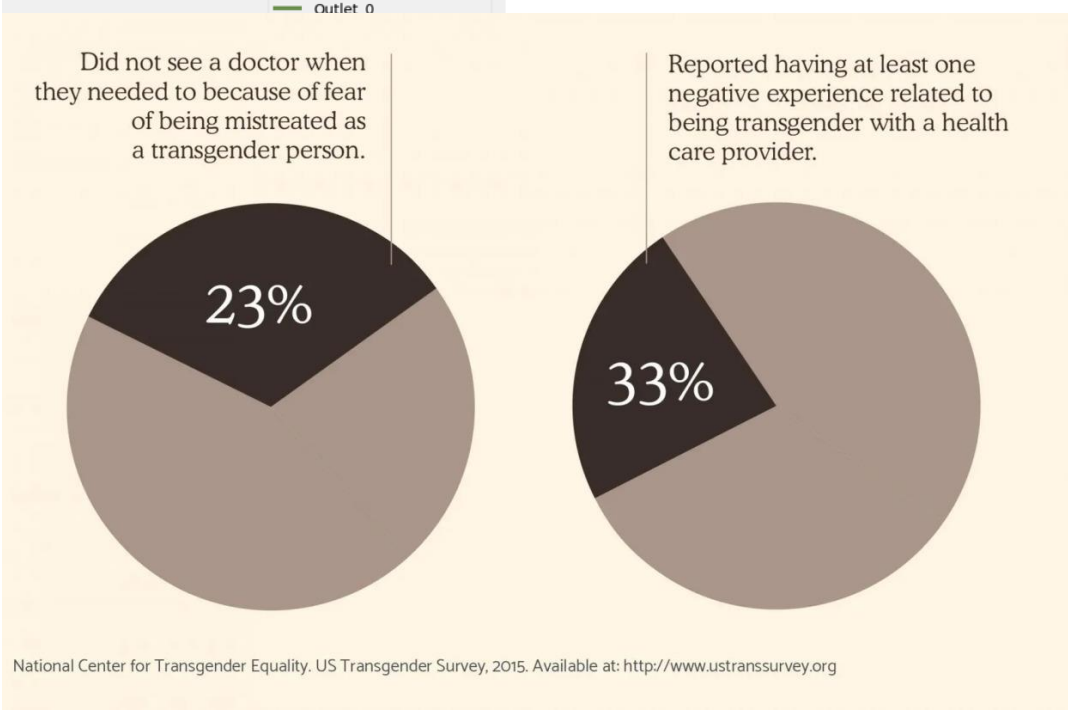
- If visualizations are obtained as a result of a program, first assume that the weird shape is a program error and try to correct it
- If there is no error, it is possible that an interesting discovery was made ☺

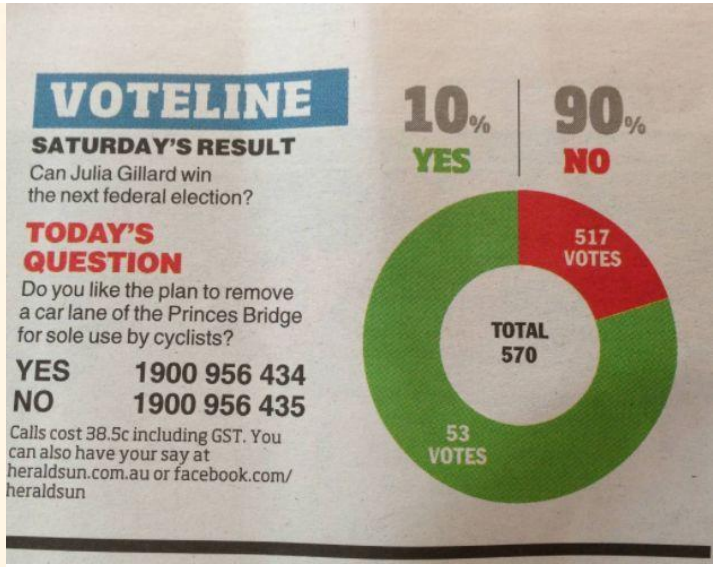
Problematic visualizations in practice: viz.wtf



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Salaried Personel		Jan	Feb	Mar	Apr	may	jun	july	Aug	Sept	Oct	Nov	Dec
2	Utilized 2020	2	24 hrs											
3		1		12hrs										
4		2			36 hrs									
5		4				72 hra								
6		4					76 hrs							
7		3						48hrs						
8		4							72 hrs					
9		2								48 hrs				
10		2									76hrs			
11		4										88 hrs		
12		3											76 hrs	
13		1												48 hrs
14														
15	2021	2	36 hrs											
16		4		84 hrs										



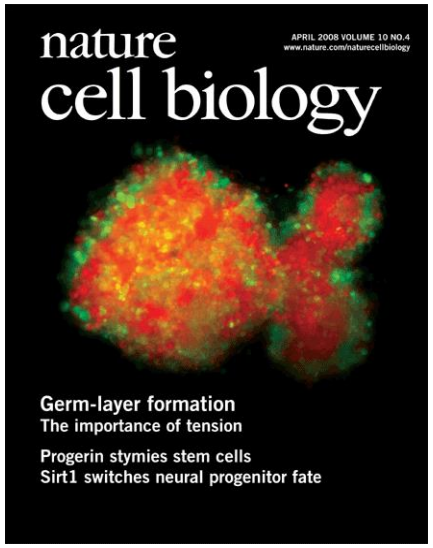


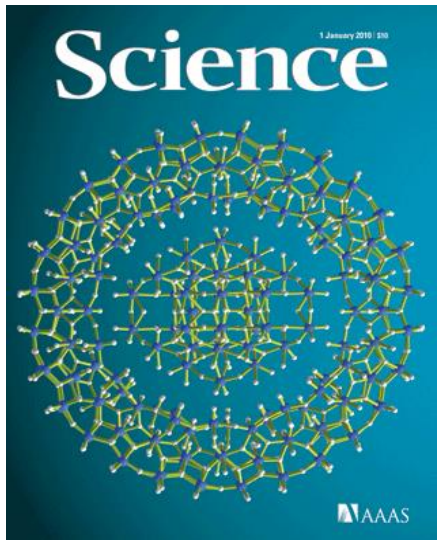


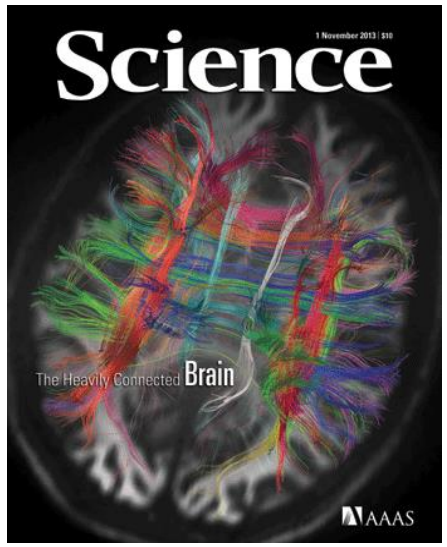
Examples of visualization uses

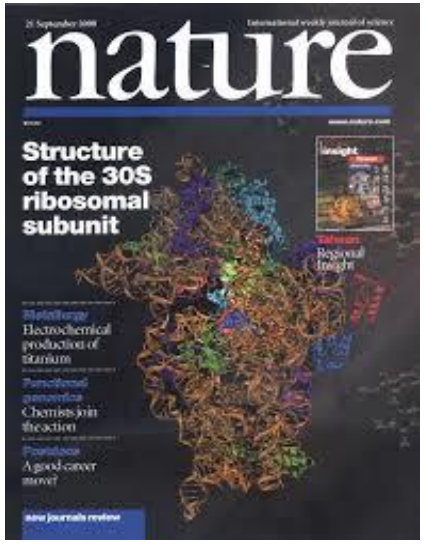
Visualization of scientific results

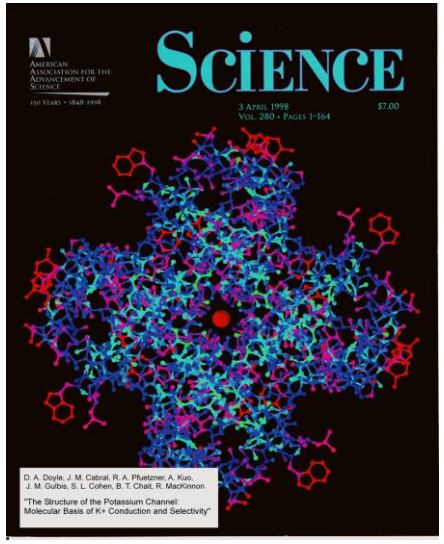
- Often difficult to discern what is science and what is art ☺
- Applies all principles and good practices that we discussed (and some more)
- Quality visualizations enhances readability of scientific articles







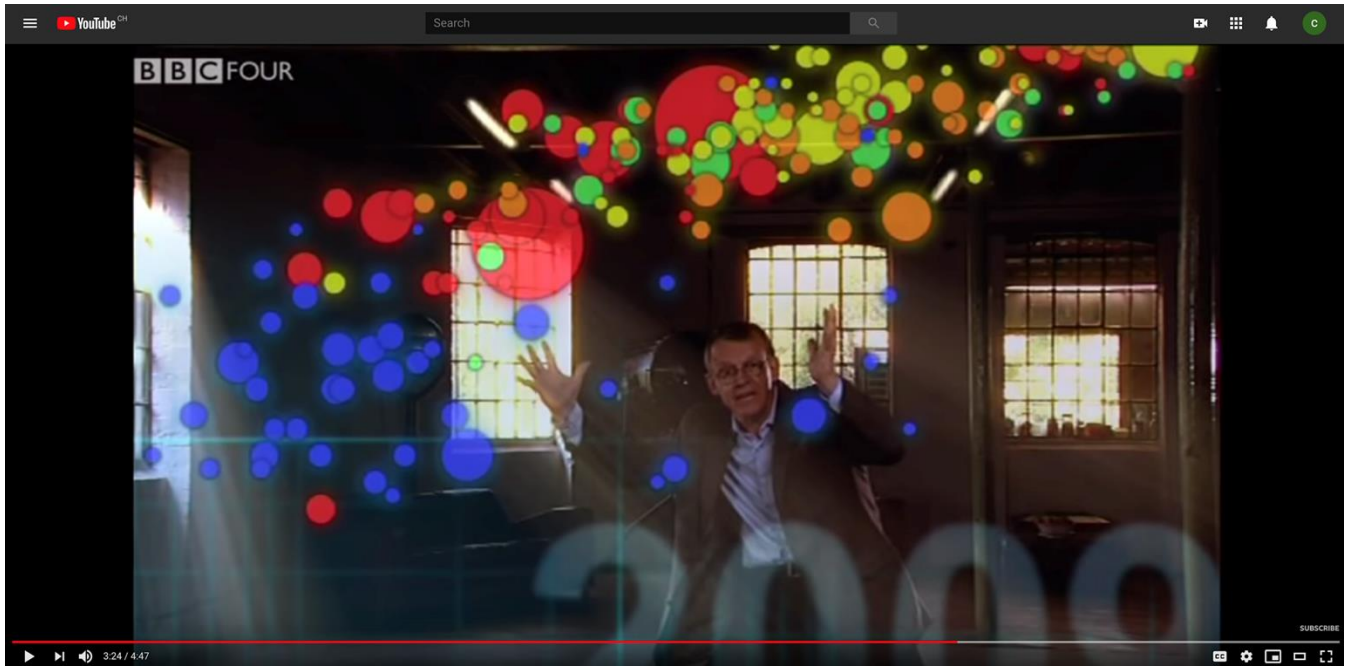




Time progression of a bubble chart

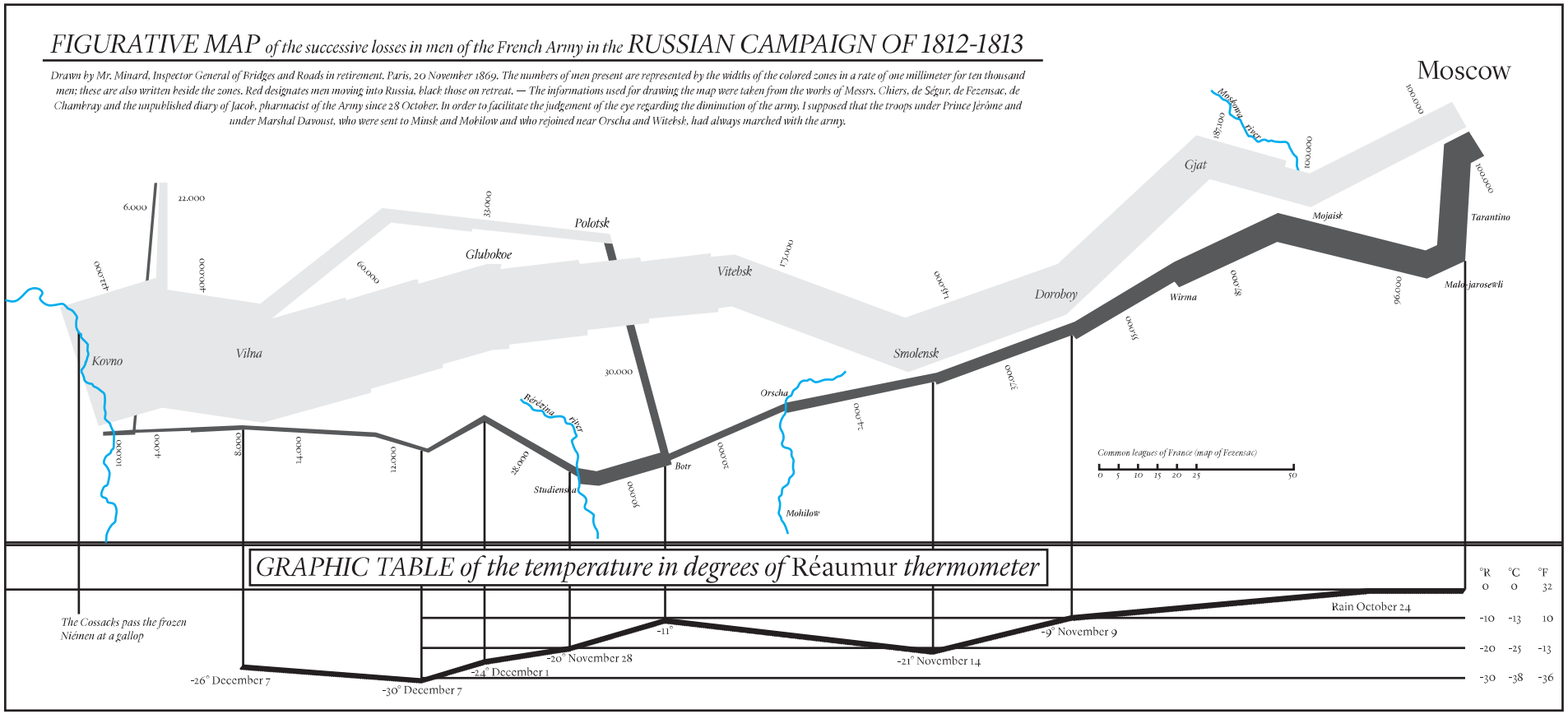
- Hans Rosling: visualization of life expectancy with respect to GDP, **200 world countries, 200 years, 4 minutes**

- https://www.youtube.com/watch?v=jbkSRLYSojo



Complicated visualizations

- Charles Joseph Minard, 1869: Napoleon's march
- https://www.edwardtufte.com/tufte/



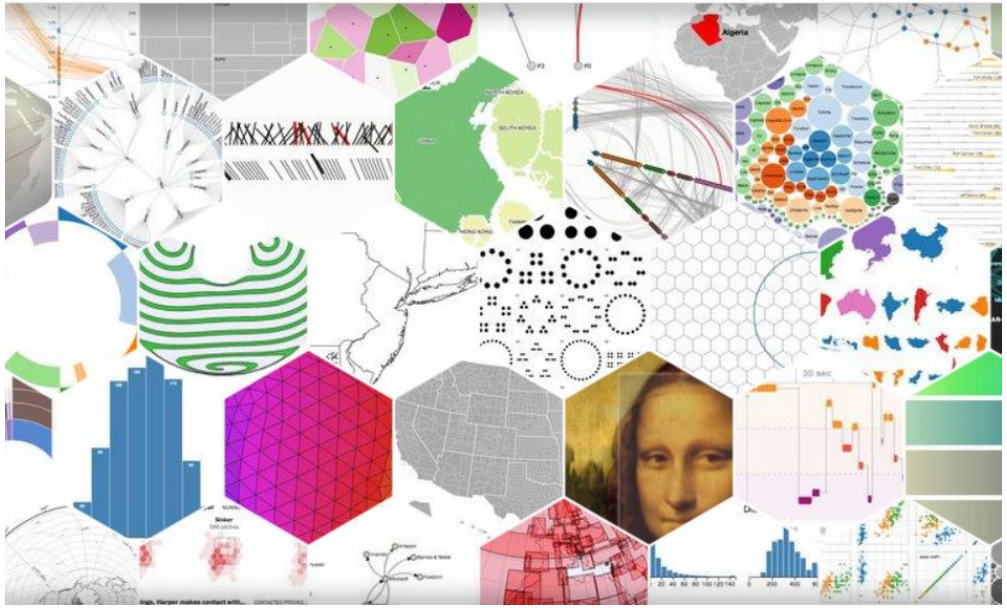
Source: https://en.wikipedia.org/wiki/Charles_Joseph_Minard#/media/File:Redrawing_of_Minard's_Napoleon_map.svg

- According to Edward Tufte: "Possibly the best statistical display ever drawn."
- 5 variables: size of the army, locations, dates, directions, air temperature when retreating

Tools for data visualization

D3

Data-Driven Documents



- Shorter from: *Data Driven Documents*
- One of the most used frameworks for data visualization
- https://d3js.org/
- Intentionally low-level, implemented in JavaScript
- Enables the widest range of options for drawing graphs
- Available for a long time, but still popular

Dash and Plotly

- Dash is a framework for developing analytics applications in Python
- Built over **Plotly.js** and **React.js**, comparable to D3
- Supports a large number of visualization options through the data analytics interfaces
- Enterprise Dash is a commercial version that includes code development, setting up in the environment and integration with the business side
- https://dash.plotly.com/introduction
- Dash and plotly support the principle for developing applications with very little coding – low code data applications

Matplotlib

- Comprehensive Python library for constructing static, animated and interactive visualizations
- Low-level interface
- One of the best choice for data visualizations in Python
- https://matplotlib.org/
- https://github.com/matplotlib/cheatsheets

Seaborn

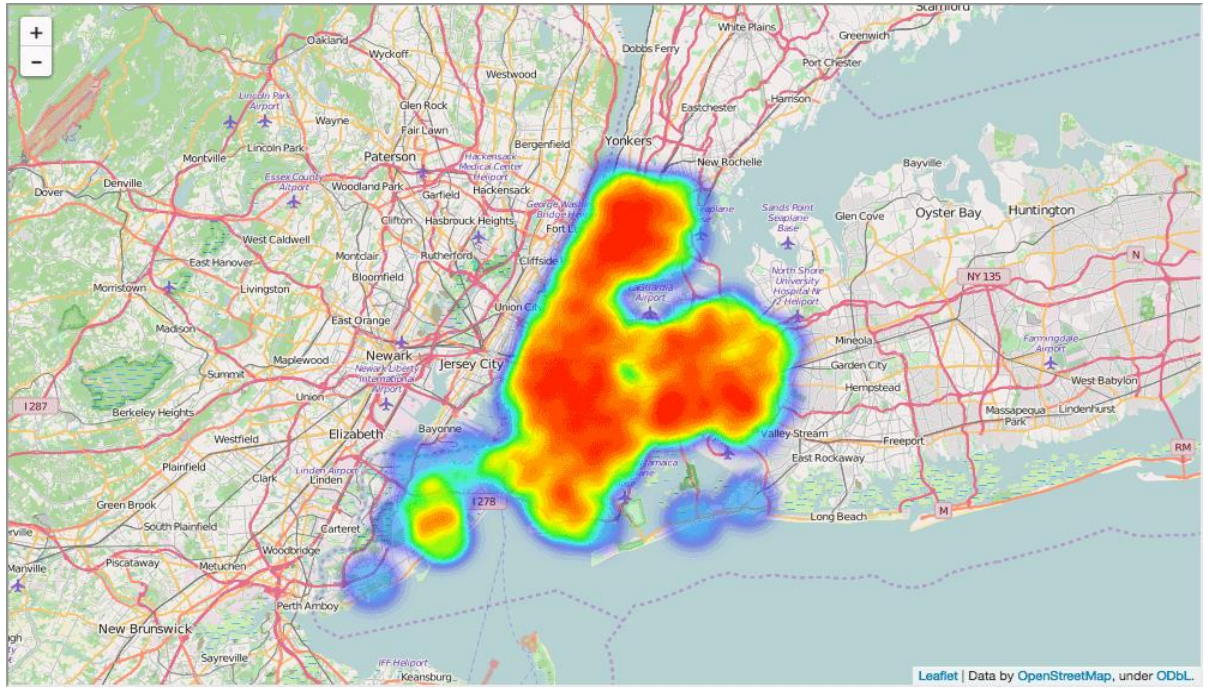
- Python library for statistical visualization contructed on top of Matplotlib
- Relatively high-level interface
- Quite popular tool, excellent for most of the simple visualizations
- https://seaborn.pydata.org/

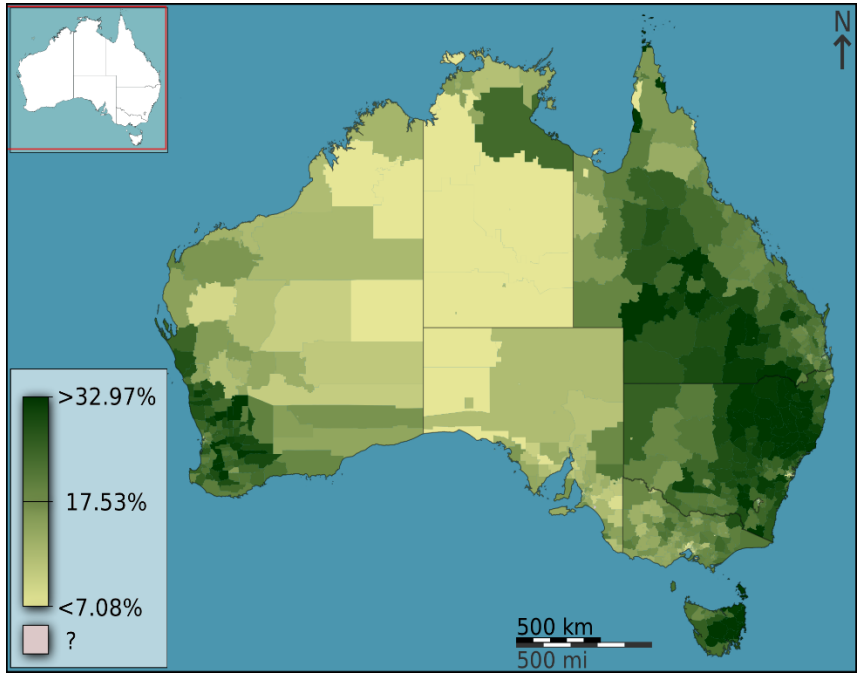
Bokeh

- Independent library for visualizing data
- Focus on visualization for big data and scientific purposes
- https://bokeh.org/
- https://github.com/bokeh/bokeh

Folium

- A tool for visualizing geodata
- https://python-visualization.github.io/folium/





Source: https://en.wikipedia.org/wiki/Choropleth_map#/media/File:Australian_Census_2011_demographic_map_-_Australia_by_SLA_-_BCP_field_2715_Christianity_Anglican_Persons.svg

Recommended literature

- Edward Rolf Tufte (2001), The Visual Display of Quantitative Information, 2nd ed., Graphics Press
- Cole Nussbaumer Knaflic (2019), Storytelling with Data: Let's Practice!, 1st ed., Wiley

Conclusions

- Data visualization enables a better insight into data and results and facilitates communication of collaborators on a project
- There is a large choice of graphs, but one needs to have experience in selecting an appropriate one for a specific purpose
- Use visualizations for discovering various unusual data
- When showing data, pay attention to colors, shapes and scales in order to convey the maximal amount of information in the simplest way
- There is a large number of tools, free and commercial ones, for visualization in Python, mostly good quality