M2 TIW – M2 BIO-INFO

DATA ANALYSIS

INTRODUCTION

WHO AM I

- Antoine Richard (antoine.richard@chu-lyon.fr)
- Research Engineer, CICLY EA3738, HCL – Lyon 1 University
- Service de chirurgie digestive et oncologique – Hôpital Lyon Sud

WHO AM I

- Research topics:
  - AI / ML for Healthcare
    - Natural Language Processing on clinical reports
    - Computer Vision on clinical images
    - Clinical data analysis
  - Data Privacy
  - Ethics of Algorithms
  - eXplainable AI (XAI)
  - AI Environmental Sustainability

- Stages orientés recherche en analyse de données médicales

CLASS OVERVIEW

- Class page: https://a-t-richard.github.io/teaching/tiw-dad
  - All contents: slides, TP, data, corrections…

- Class divided in 2 independent parts:
  - Monday: me, Analyse de de données, manipulation, visualization
  - Wednesday: Fabien De Marchi (frequent patterns and others) (dès ce mercredi)

- My part:
  - How to interpret real data
  - How to explore it using custom interactive visualization
    - Dash

CLASS OVERVIEW

- Courses: (TODO: Maybe rethink order)
  - Course 1: Introduction to Data Analysis
  - Course 2: Unsupervised ML (beyond k-means)
  - Course 3: Network Data
  - Course 4: Projections and other data types

- Exam:
  - 50% of the final note
  - 50% of the exam's note on my part
  - On my part: 60% on known exercices, 40% on unknown exercices

- Project:
  - Building a Data Analysis WebApp with Dash
  - Groups of 2 or 3 students
  - A dataset of your choice
  - 50% of final note

OBJECTIVE

- To learn how to answer these questions in front of a dataset:
  - "What i want to know from my dataset ? "
  - "What are the best ways to extract and visualize these information from my dataset ?"
- To learn how to use code to explore datasets

BEFORE STARTING: CHATGPT

« The brain, as muscles, needs exercices to improve.
And using ChatGPT to learn is like to do pull-ups
assited by freight elevator »

BUT… if you want to use it anyway, I'll ask you to:

- **Be honest and transparent**
  - Tell when, why and how you used it
- **Be critical**
  - Verify the results
  - Justify their use in your work

PART1 - TYPES OF DATA

DATA TYPES

- Data types : What kind of data (feature, variables) can we encounter?
  - People
    - Name, Age, Gender, Revenue, Birth Date, Address, etc.
  - Localisation
    - Surface area, Floor, Address, # of rooms, # of Windows, Elevator, etc.
  - Clinical
    - Weight, blood pressure, molecular biology, CT Scans, histological, etc.

- Types of features?

DATA TYPES

- Nominal:
  - From “names”. No order between possible values
  - Color, Gender, Animal, Brand, etc. (Numbers:Participant ID, class…)
- Ordered:
  - Ordinal
  - Interval
  - Ratio

ORDERED

- Ordinal
  - Order between values, but not numeric
  - Size[small, medium, large], [Satisfied, …, Unsatisfied], Income [0-10k],[10k-15k],[15k-50k]…

- Ratio
  - Numerical values, all operations are valid
  - Height, Duration, Revenue…

- Interval
  - Numeric values, difference is meaningful
  - T°: 30°-20° = 15°-5°, But 30° 2*15°
  - 2022-2020=1789-1787, but 1011 2022/2
  - =>0 is not a meaningful value, is arbitrary

CONTINUOUS OR DISCRETE ?

- Real data is never mathematically continuous (limit of precision in computers…)

- A dataset must be treated as continuous when the number of possible values is at least in the order of the number of observations.

- Think of plotting an histogram distribution…

RELATIVE AND ABSOLUTE

- The world values are divided in two types of things, and two types of interpretation
- Is there a larger difference between two persons:
  - Age 1 / 5? Revenue 1000€/1500€ ?
  - Age 91 / 95? Revenue 10 000 € / 10500 € ?
- Think about it:
  - In country 1, average salary is 100$, p1 salary is 1000$
  - In country 2, average salary is 1000$, p2 salary is 2000$
  - Should you consider that
    - p1 is well paid (10x average salary VS 2x for p2)
    - They are paid the same (1000$ différence)

RELATIVE AND ABSOLUTE

- If your values are expressed in absolute terms, but you think their interpretation should be in relative terms, you can transform them using the ***log scale***: e.g.,
  - In log2, going from x to x+1 means multiplying by 2.
  - In log10, going from x to x+1 means multiplying by 10

- e.g, should you express earthquake strength in:
  - Energy released
  - Richter Magnitude (log scale)
  - Depends if you care about comparing small and large earthquakes (relative) or large from superlarge (raw scale => all those small ones look the same from there)

OTHER TYPES

- Real Data can have many other forms
  - Textual (Unstructured)
  - Relational (networks)
  - Complex objects (picture, video, software…)

DATA QUALITY

- Data coming from the real world is often incorrect
  - Malfunctioning sensors (T°, speed…)
  - Human error or falsification (e.g., entered 100 instead of 1.00)
  - Undocumented change (e.g., Bicycle sharing station was moved…)

- Before applying a method blindly,
  - =>**check your data's quality!**
  - If the data errors are plausible, no simple solutions
  - Common
    - Out-of-range values (e.g., a person's weight is negative or above 1000kg…)
    - Zeros. (Weight of the person is 0. But in many cases, zero is possible too…)
    - Variant: 01/01/1970…

MISSING VALUES

- Real-life datasets are full of missing values
  - Impossible data: ***fur color*** for a sphinx cat
  - More generally, failure to obtain them

- Few methods can deal with missing values
  - =>Imputation
    - Naive: fill with average value
    - Use ML to fill-in missing values (other problems, introduce biases…)
    - Large literature, no good solution

REAL LIFE IS COMPLEX

- You will have to do modeling choices (feature engineering…)
- Possibles values: Blue, Cyan, White, Yellow, Orange, Red.
  - Nominal or Ordinal ?
- Survey: “rate X on a scale from 0 to 5”
  - What if labels are associated ? (“Bad”, “average”, …)
- Always try to justify your choices, even if sometimes, there is no real reason to choose a representation instead of another

PART2 - DESCRIBING A VARIABLE

DESCRIBING VALUES

	adult	belongs_to_collection	budget	genres	homepage	id	imdb_id	original_language	original_title	overview	...	release_date	revenue	runtime	spoken_languages	status	tagline	title	video	vote_average	vote_count
0	False	{'id': 10194, 'name': 'Toy Story Collection', ...	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	http://toystory.disney.com/toy-story	862	tt0114709	en	Toy Story	Led by Woody, Andy's toys live happily in his ...	...	1995-10-30	373554033	81.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	NaN	Toy Story	False	7.7	5415
1	False	NaN	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	NaN	8844	tt0113497	en	Jumanji	When siblings Judy and Peter discover an encha...	...	1995-12-15	262797249	104.0	[{'iso_639_1': 'en', 'name': 'English'}, {'iso...	Released	Roll the dice and unleash the excitement!	Jumanji	False	6.9	2413
2	False	{'id': 119050, 'name': 'Grumpy Old Men Collect...	0	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	NaN	15602	tt0113228	en	Grumpier Old Men	A family wedding reignites the ancient feud be...	...	1995-12-22	0	101.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	Still Yelling. Still Fighting. Still Ready for...	Grumpier Old Men	False	6.5	92
3	False	NaN	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	NaN	31357	tt0114885	en	Waiting to Exhale	Cheated on, mistreated and stepped on, the wom...	...	1995-12-22	81452156	127.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	Friends are the people who let you be yourself...	Waiting to Exhale	False	6.1	34
4	False	{'id': 96871, 'name': 'Father of the Bride Col...	0	[{'id': 35, 'name': 'Comedy'}]	NaN	11862	tt0113041	en	Father of the Bride Part II	Just when George Banks has recovered from his ...	...	1995-02-10	76578911	106.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	Just When His World Is Back To Normal... He's ...	Father of the Bride Part II	False	5.7	173
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16624	False	NaN	37000000	[{'id': 18, 'name': 'Drama'}, {'id': 53, 'name...	http://whatdoyoubelieve.warnerbros.com/	48171	tt1161864	en	The Rite	Seminary student Michael Kovak (Colin O'Donogh...	...	2011-01-28	96047633	114.0	[{'iso_639_1': 'en', 'name': 'English'}, {'iso...	Released	You can only defeat it when you believe.	The Rite	Fa		577
16625	False	NaN	0	[{'id': 18, 'name': 'Drama'}]	http://www.dogpound-lefilm.com/	43920	tt1422020	en	Dog Pound	Three juvenile delinquents arrive at a correct...	...	2010-04-24	430041	91.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	Fighting Back Is The Only Way Out	Dog Pound			104
16626	False	NaN	3500000	[{'id': 16, 'name': 'Animation'}, {'id': 28, '...	NaN	56590	tt1699114	en	All Star Superman	Lex Luthor enacts his plan to rid the world of...	...	2011-02-22	0	76.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	The measure of a man lies not in what he says,...	All Star Superman			121

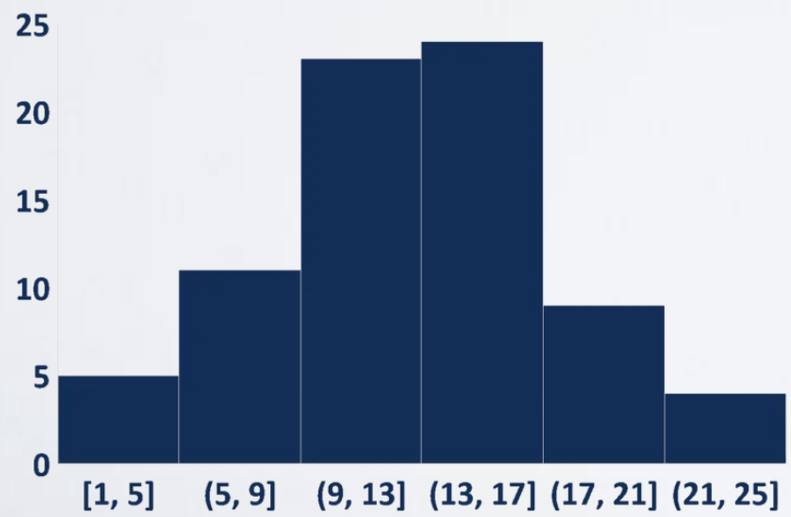
DESCRIBING VALUES

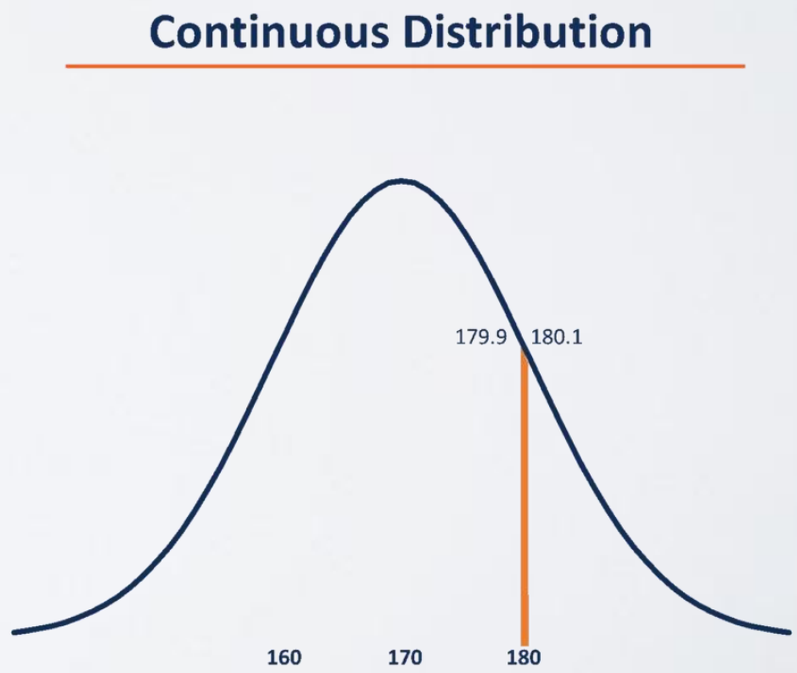
- Mean / Average
  - Be careful, not necessarily representative !
- Median
  - Be careful, not necessarily representative !
- Mode
  - Not representative !
- Min/Max
  - …

None of these metrics tells us about variables' variability !

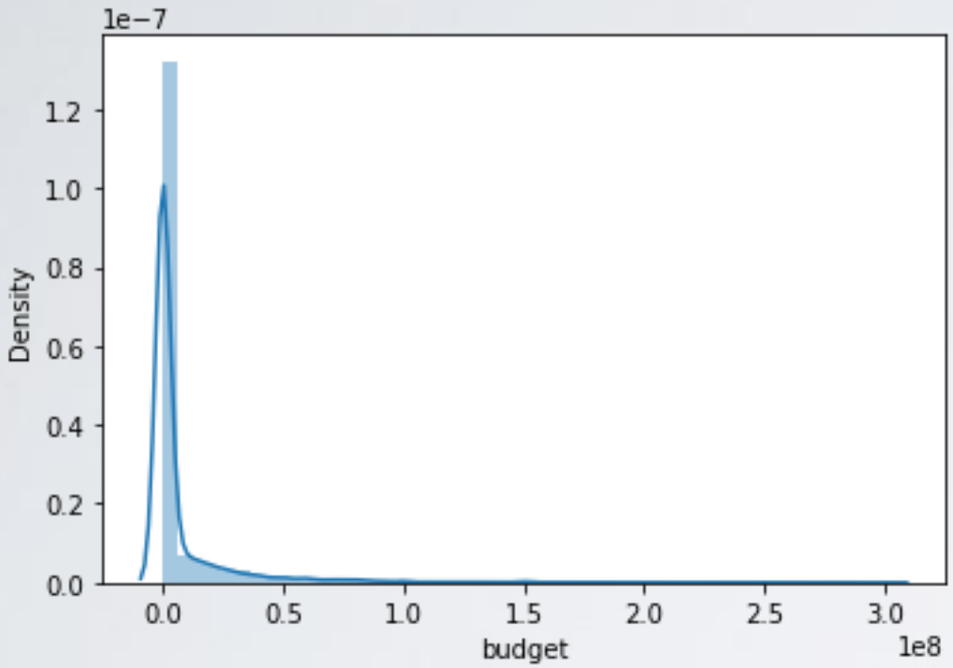
DISTRIBUTION

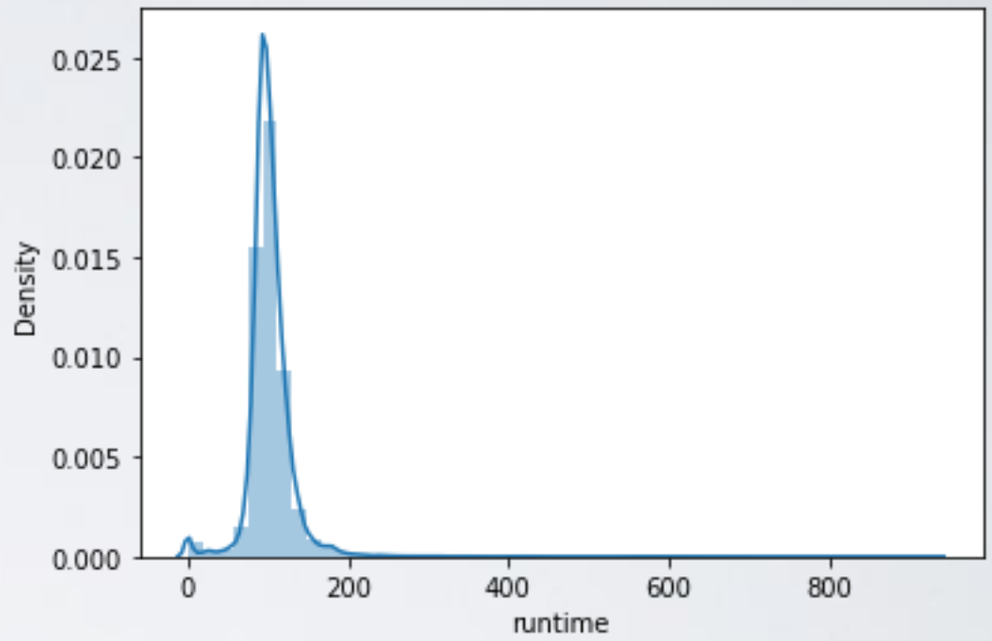
- What is a distribution?
  - A description of the frequency of occurence of items
  - A generative function describing the probability to observe any of the possible events
  - Discrete or continuous

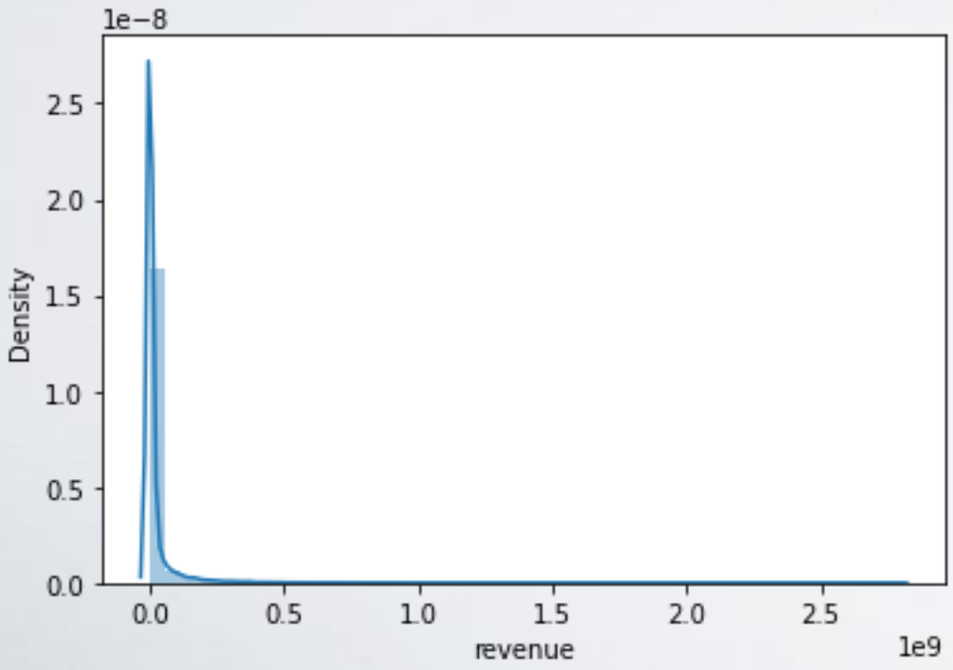


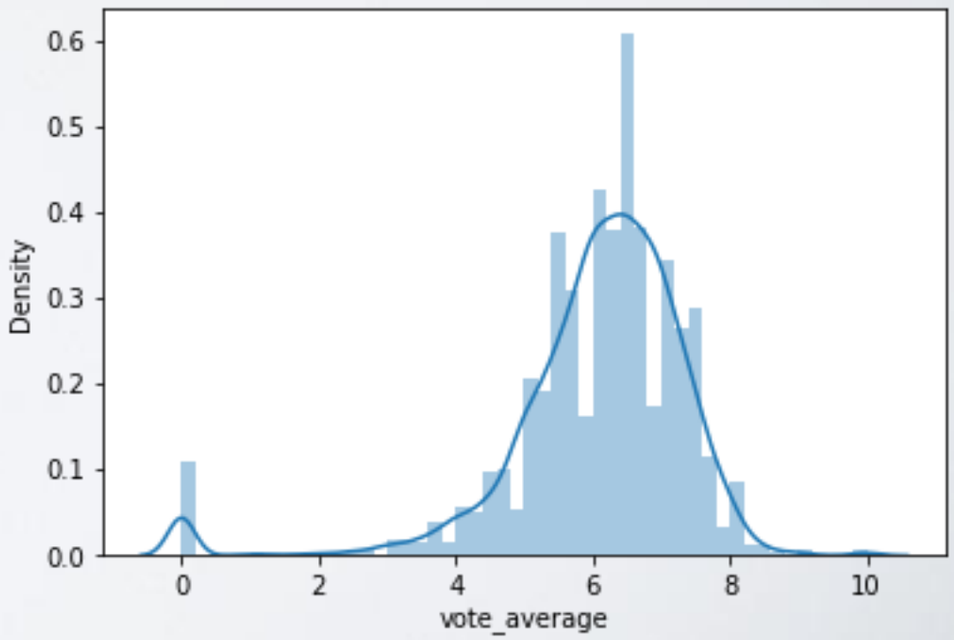


EMPIRICAL DISTRIBUTIONS



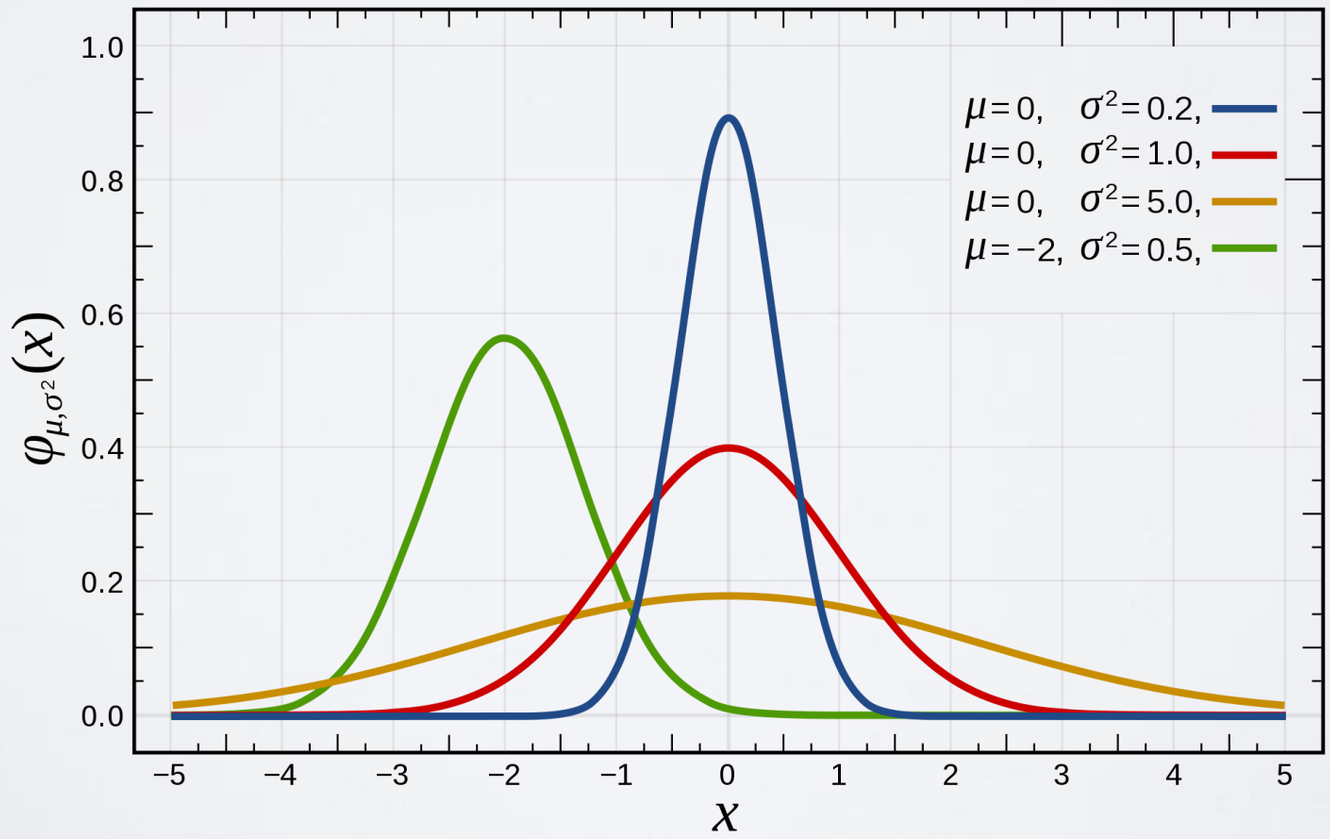




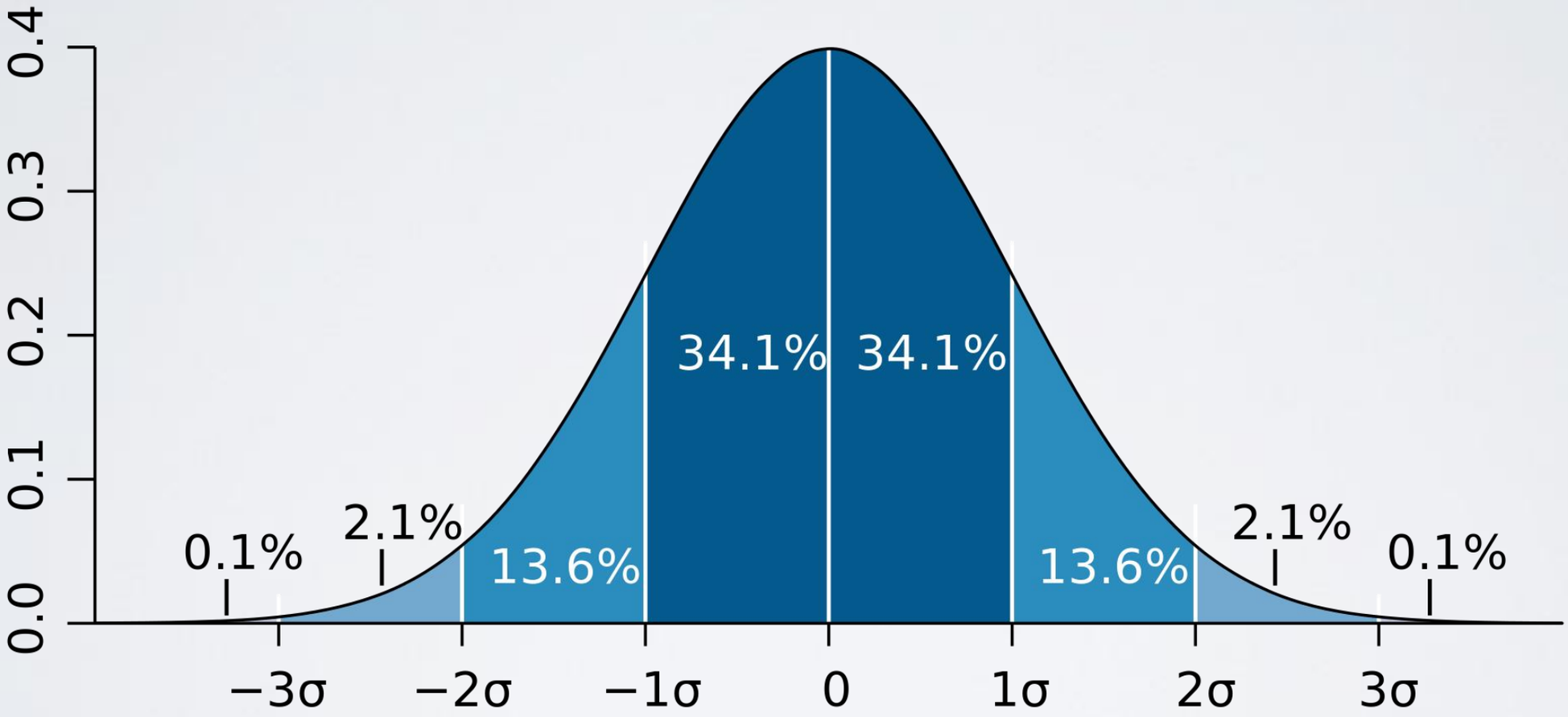


THEORETICAL DISTRIBUTIONS

- Normal distribution
  - Many real variables follow it approximately (height, weight, price of a given product in various locations…
  - Random variations around a well-defined mean
  - Central limit theorem: average of many samples of a random variable converges to a normal distribution



NORMAL DISTRIBUTION



STANDARD DEVIATION

- Squared root of the Variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{E}[(X - \mu)^2]}$$

VARIANCE

- Variance:
  - Expectation of the squared deviation of a random variable from its mean

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}[(X - \mu)^2]$$

Also expressed as average squared distance between all elements

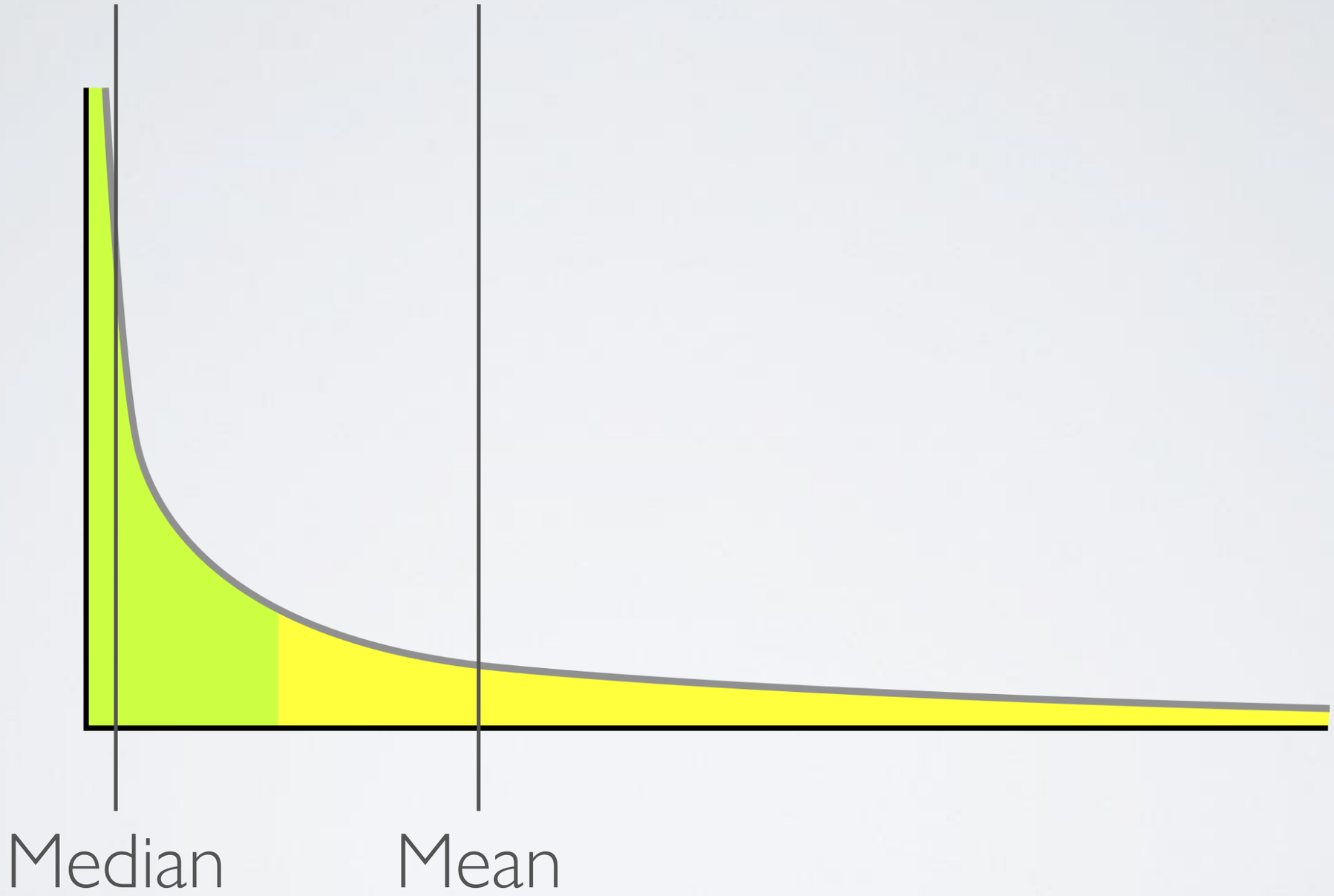
$$\sigma^2 = \frac{1}{N^2} \sum_{i<j} \left(x_i - x_j\right)^2$$

ABSOLUTE DEVIATION

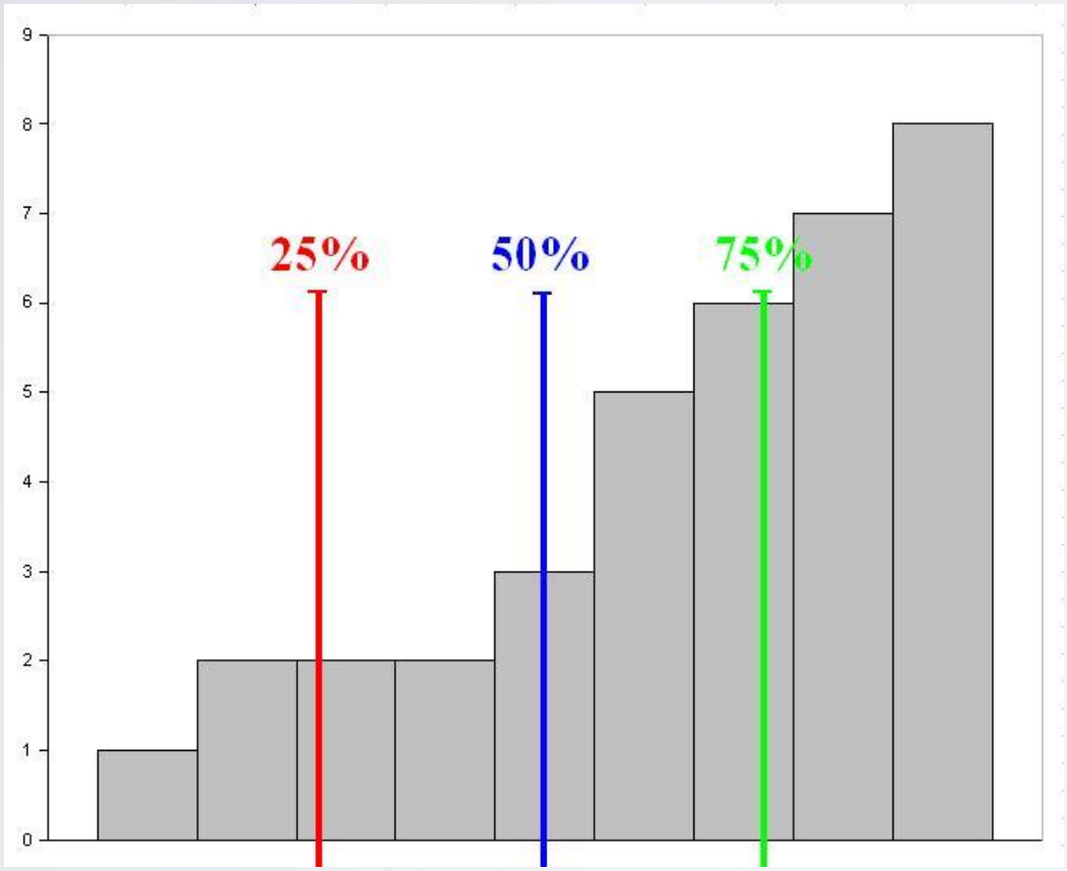
- MAD (Mean Absolute Deviation)
  - Deviation from mean or from median
  - (Variant: Median Absolute Deviation)

- $\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|$

- So why are we using the Standard Deviation again?
  - The mean minimizes the expected squared distance
  - The median minimizes the MAD
  - Leads naturally to least square regression and PCA… see later.

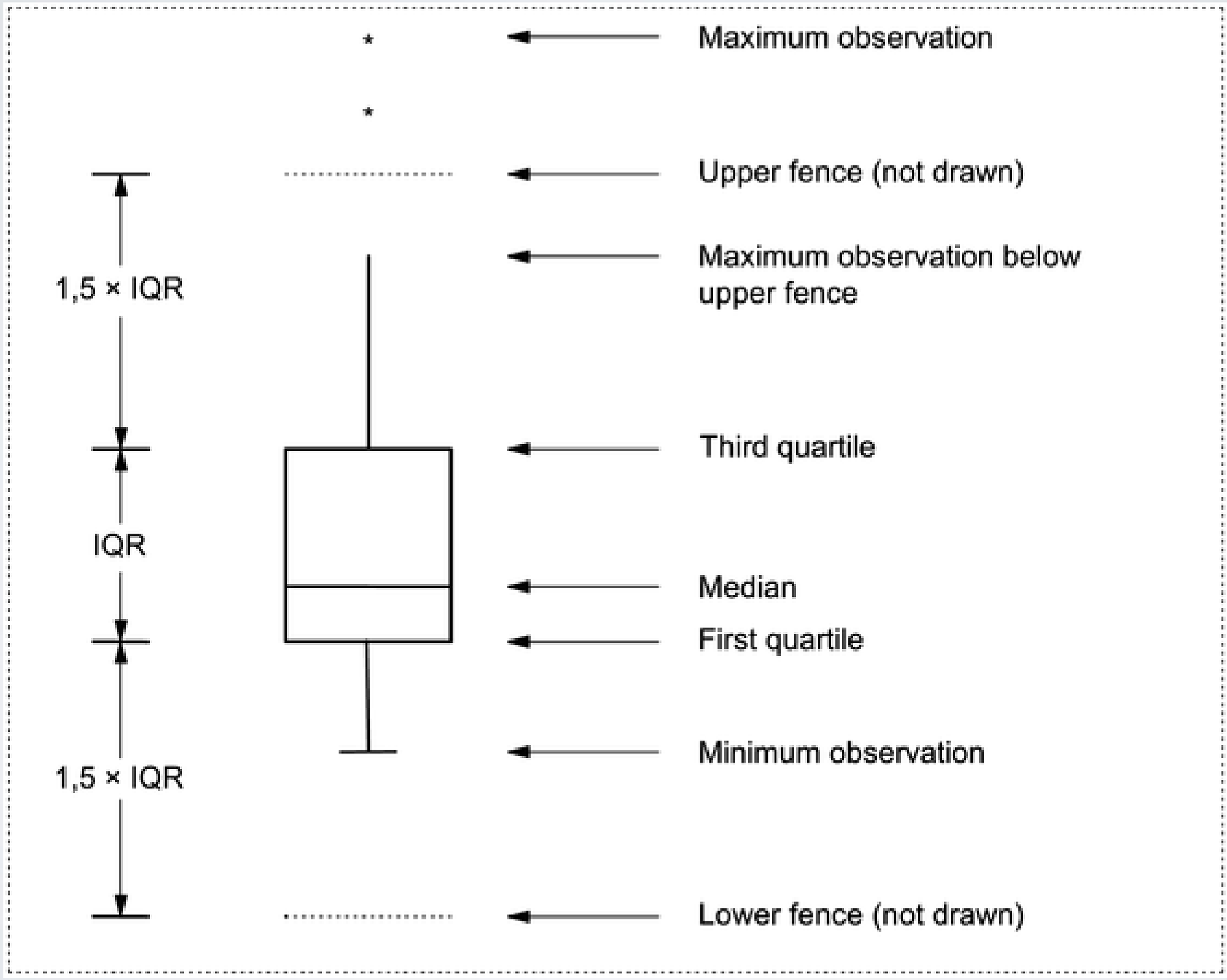

Median
Mean

MEDIAN AND QUARTILES

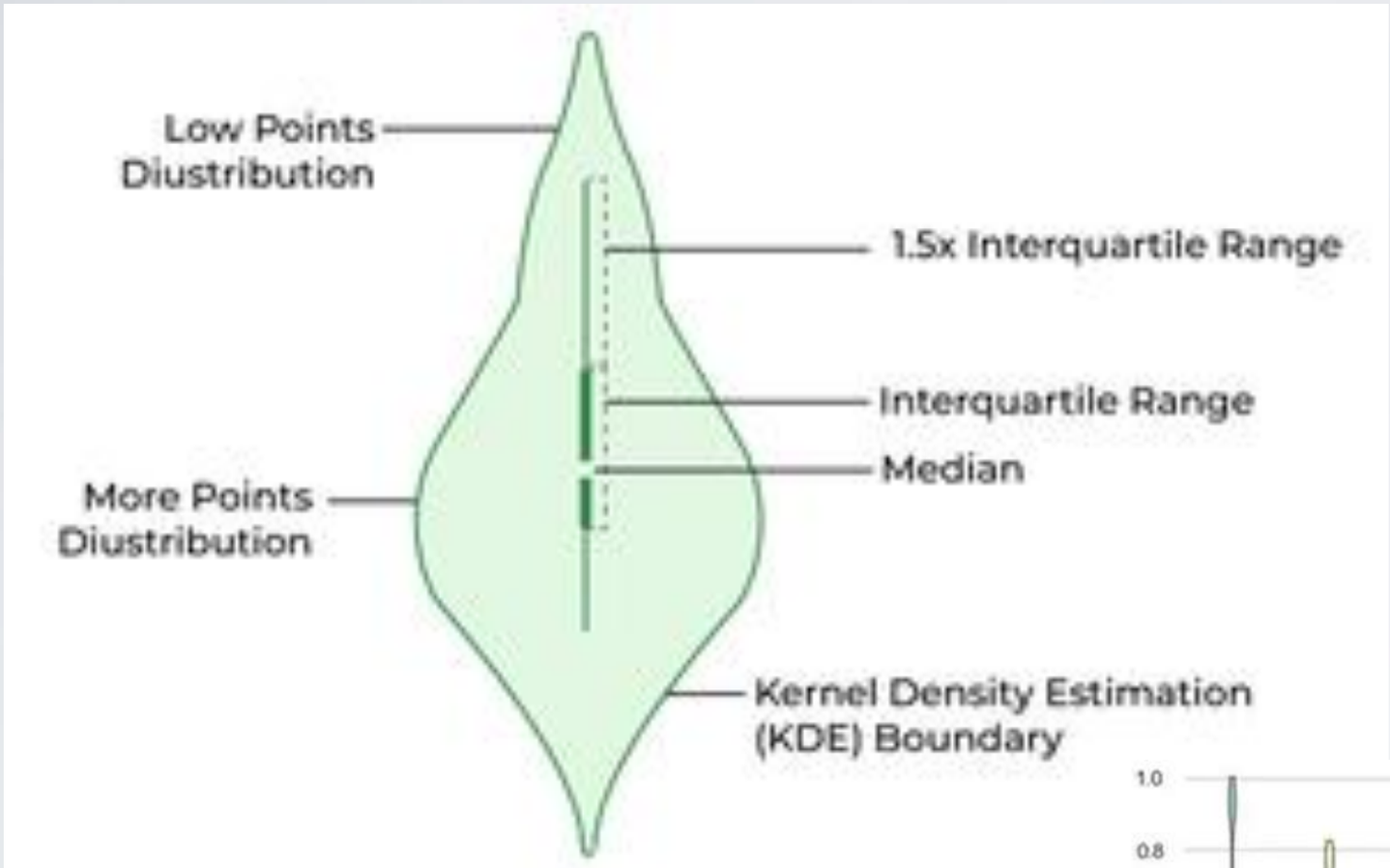


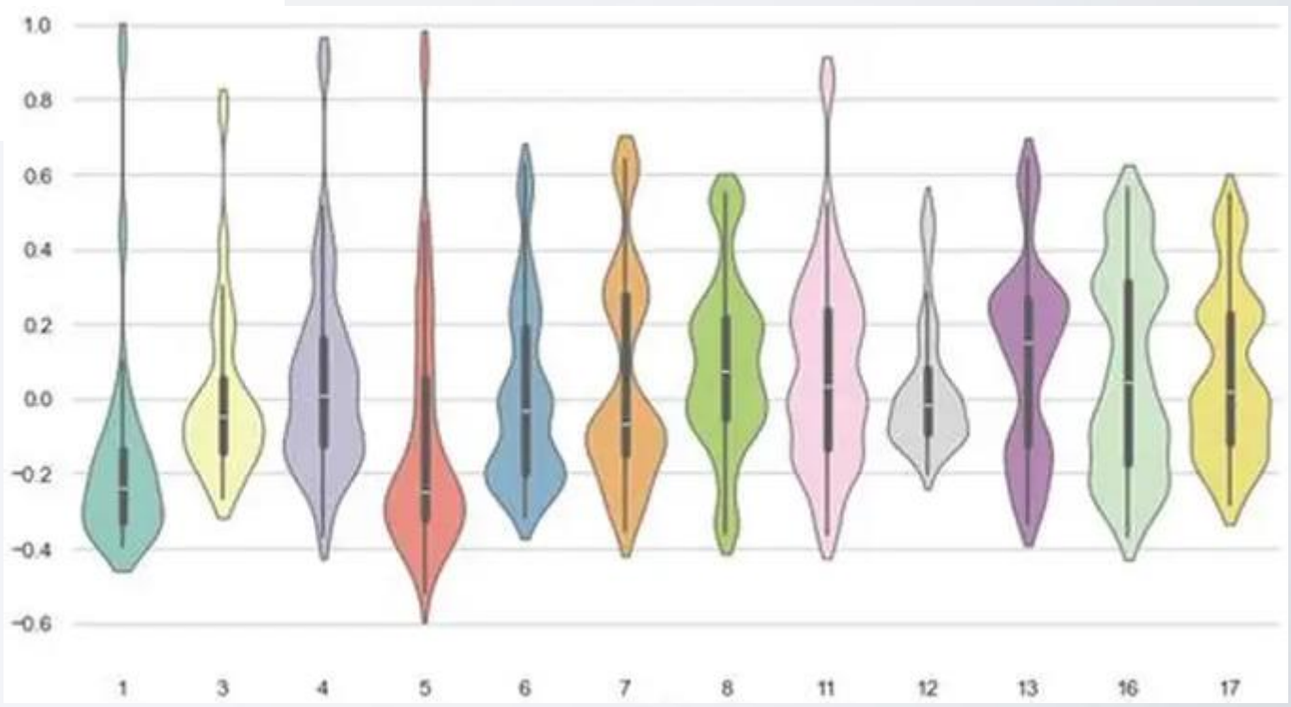
- First Quartile (Q1)
  - Value for which 25% of observed data are below and the remaining 75% above
  - Also called 25th percentile
- Second Quartile (Q1) or Median
  - Value for which 50% of observed data are below and the remaining 50% above
  - Also called 50th percentile
- Third Quartile (Q3)
  - Value for which 75% of observed data are below and the remaining 25% above
  - Also called 75th percentile

THE BOX-PLOT

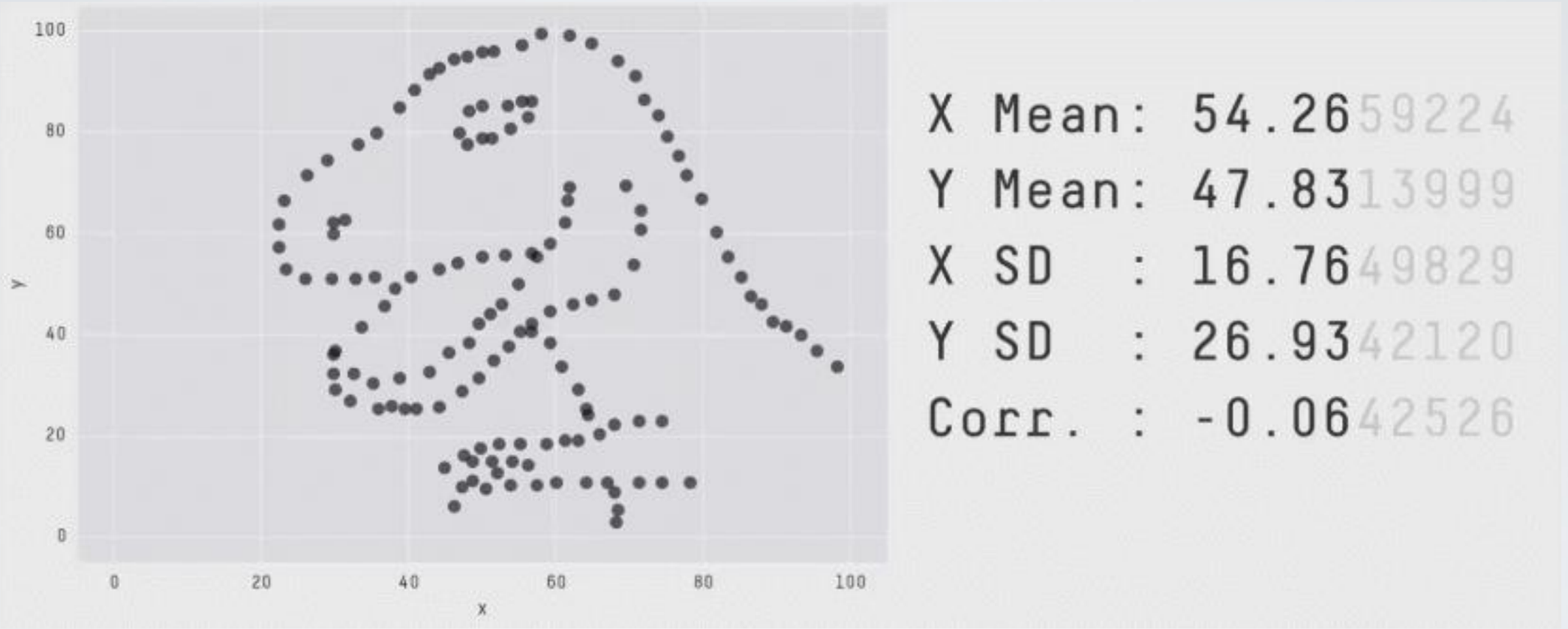

Maximum observation
Upper fence (not drawn)
Maximum observation below upper fence
1,5 × IQR
IQR
Third quartile
Median
First quartile
1,5 × IQR
Minimum observation
Lower fence (not drawn)

THE VIOLON-PLOT





(UN)DESCRIPTIVE STATISTICS



The datasaurus

https://github.com/jumpingrivers/datasauRus

SOME ADVICES

- Always plot the distribution.
- Don’t assume a theoretical distribution
- Don’t believe single-number statistics. Never ever.
  - With Mean (for normally distributed feature):
    - Always add standard deviation at least
    - And plot distribution at best
  - With Median (for not normally distributed feature):
    - Always add quartiles Q1 and Q2 at least (min and max also)
    - Use box plot or violion plot at best

PART 3 - STATISTICAL TESTS

WHAT IS IT?

- Questions such as:
  - Is my data following a normal distribution?
    - I could summarize it by mean and variance…
  - Are two variables coming from the same "population"
    - Is the probability of dying from COVID the same in two countries for 2 "identical" persons?
  - Are two variables independent?
    - "eating chocolate" and "having cancer"?

- You can use statistical tests:
  - Normality: Shapiro-Wik, etc.
  - Categorical variables: Chi-squared χ^2, etc.
  - Comparing distributions: Kolmogorov-Smirnov, t-test if assuming normality, etc.

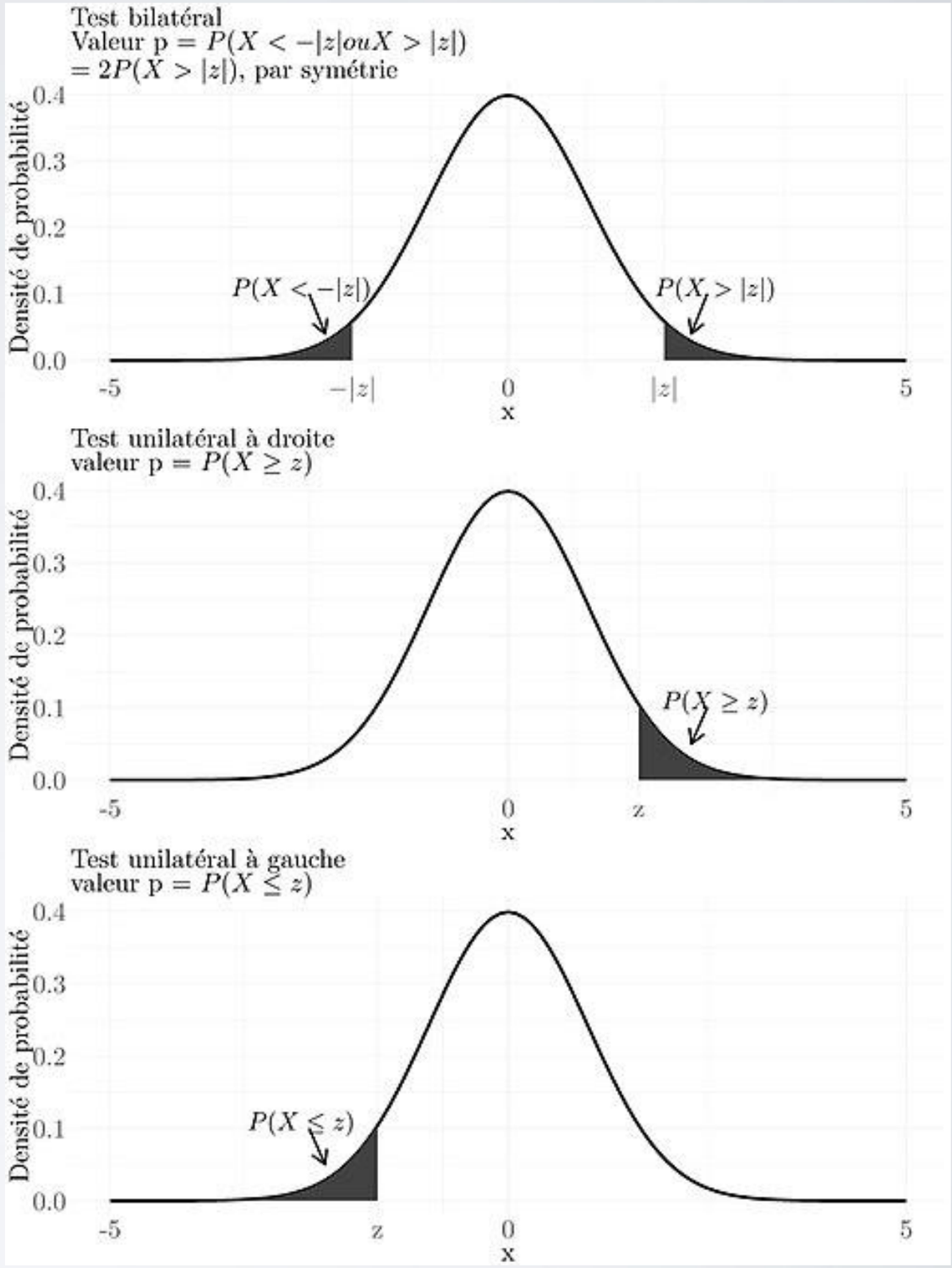
STATISTICAL TESTS

- "Can we reject the null hypothesis?"
  - p-value large => null hypothesis Likely True. (Probability obtain data if hypothesis True)
  - Normality test: Null hypothesis=>distribution is normal.
  - Hypothesis testing: Null hypothesis=>No relation between variables of interest

P-VALUES

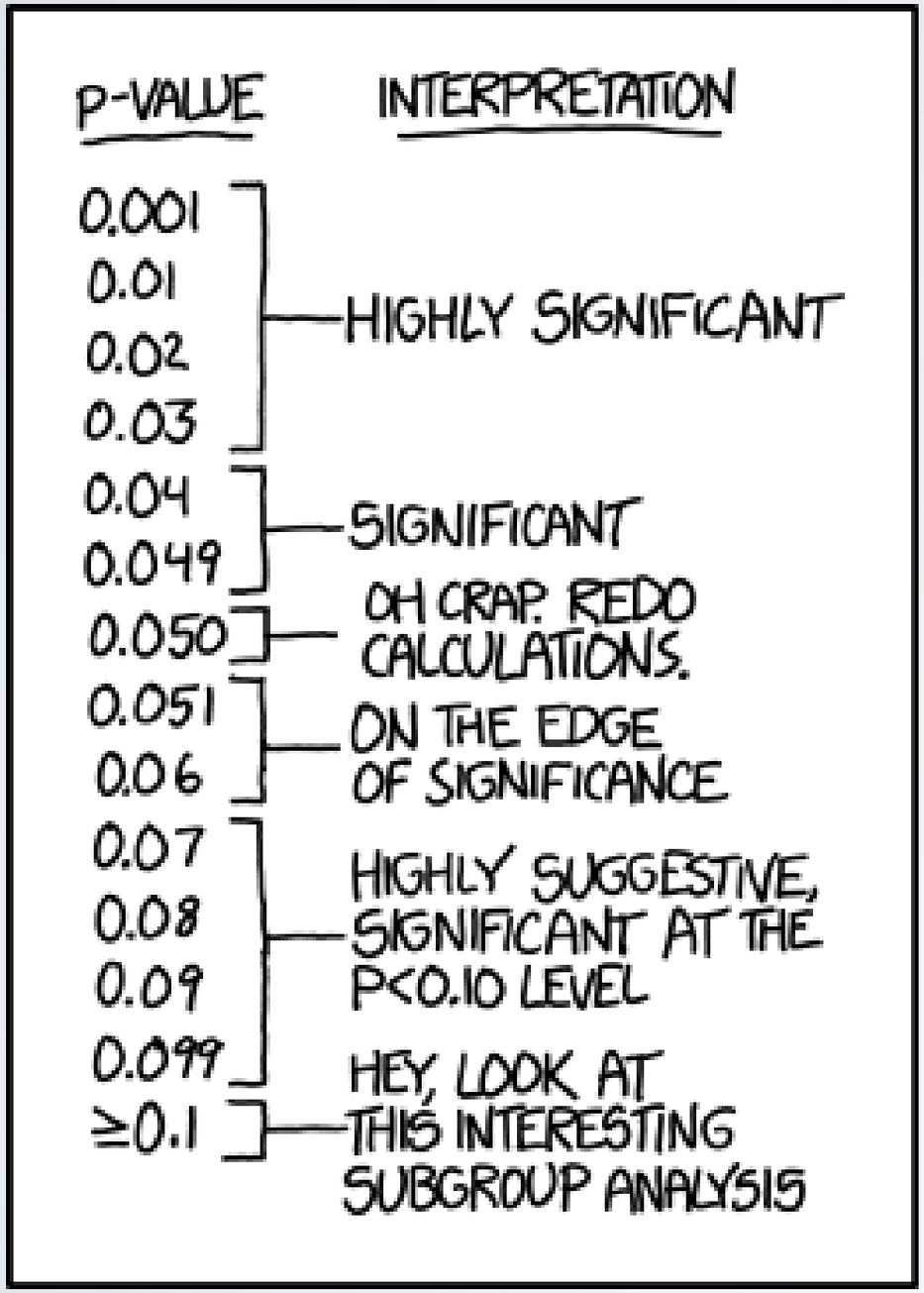
- The **p-value** is the probability of obtaining test results "at least as extreme as" the result actually observed, under the assumption that the null hypothesis is correct.

- The hypothesis is rejected if the p-value is less than or equal to a predefined threshold (generally 0.05)

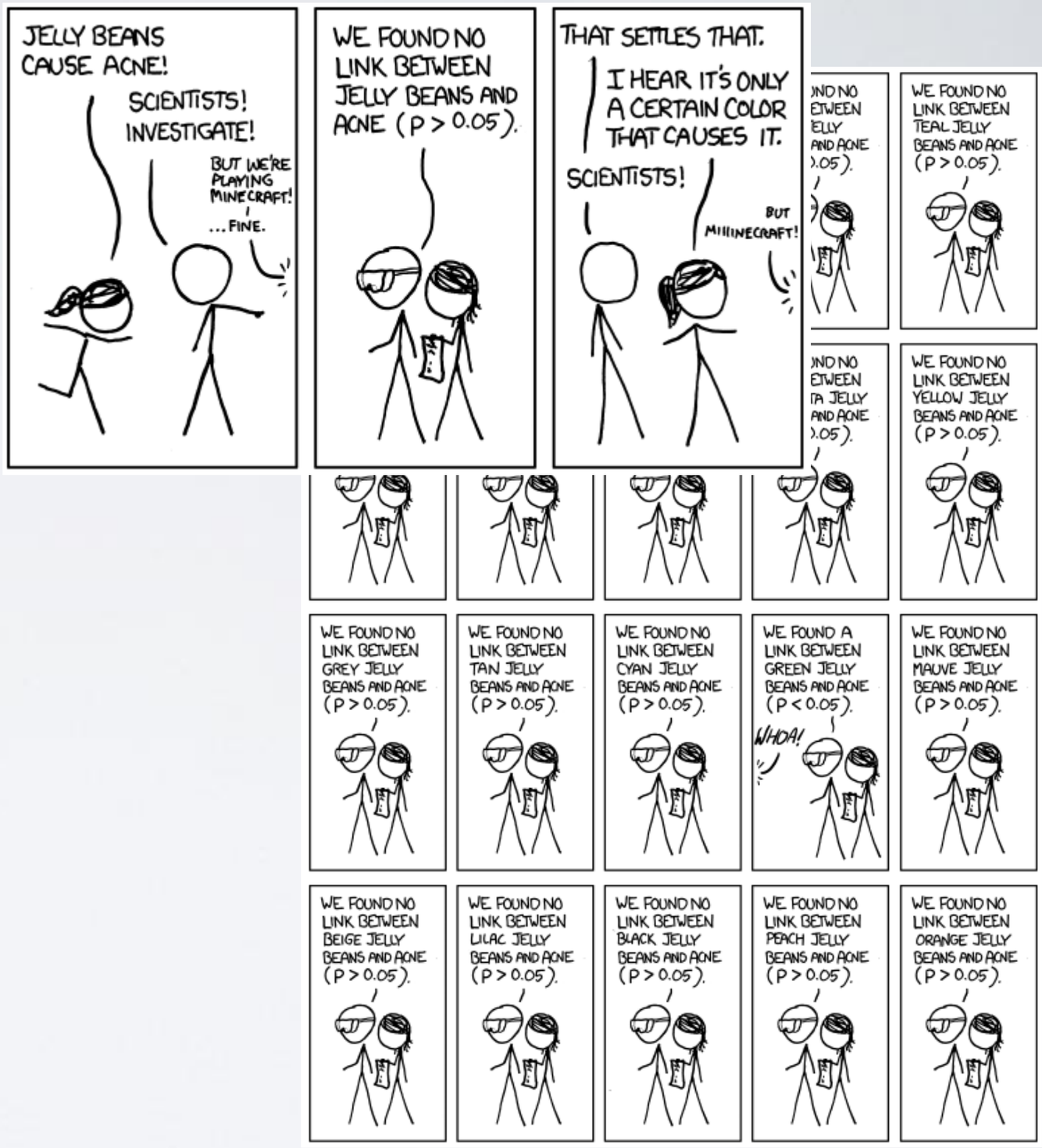


Valeur p — Wikipédia

P-VALUES - INTERPRETATION



xkcd: P-Values



xkcd: Significant

STATISTICAL TESTS

- Useful when you have **very little data** and that you **cannot obtain more** (especially the case in clinical research)

- If you have large datasets, in general, these tests are useless
  - Nothing is exactly normal
  - No pairs of populations are exactly identical
  - No variables are independent
    - Having a cat and owning a SUV? Height of a person and their grades in high school? Etc.

SOME ADVICES

- Plot the data
- If the relation is not so obvious that you have no doubts, don't believe it
- Get more data :)

PART 4 - VARIABLE INTERACTIONS

COVARIANCE MATRIX

Covariance Matrix Formula

cuemath
THE MATH EXPERT

$$\begin{bmatrix} \mathrm{Var}(x_1) & \cdots & \mathrm{Cov}(x_n,x_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n,x_1) & \cdots & \mathrm{Var}(x_n) \end{bmatrix}$$

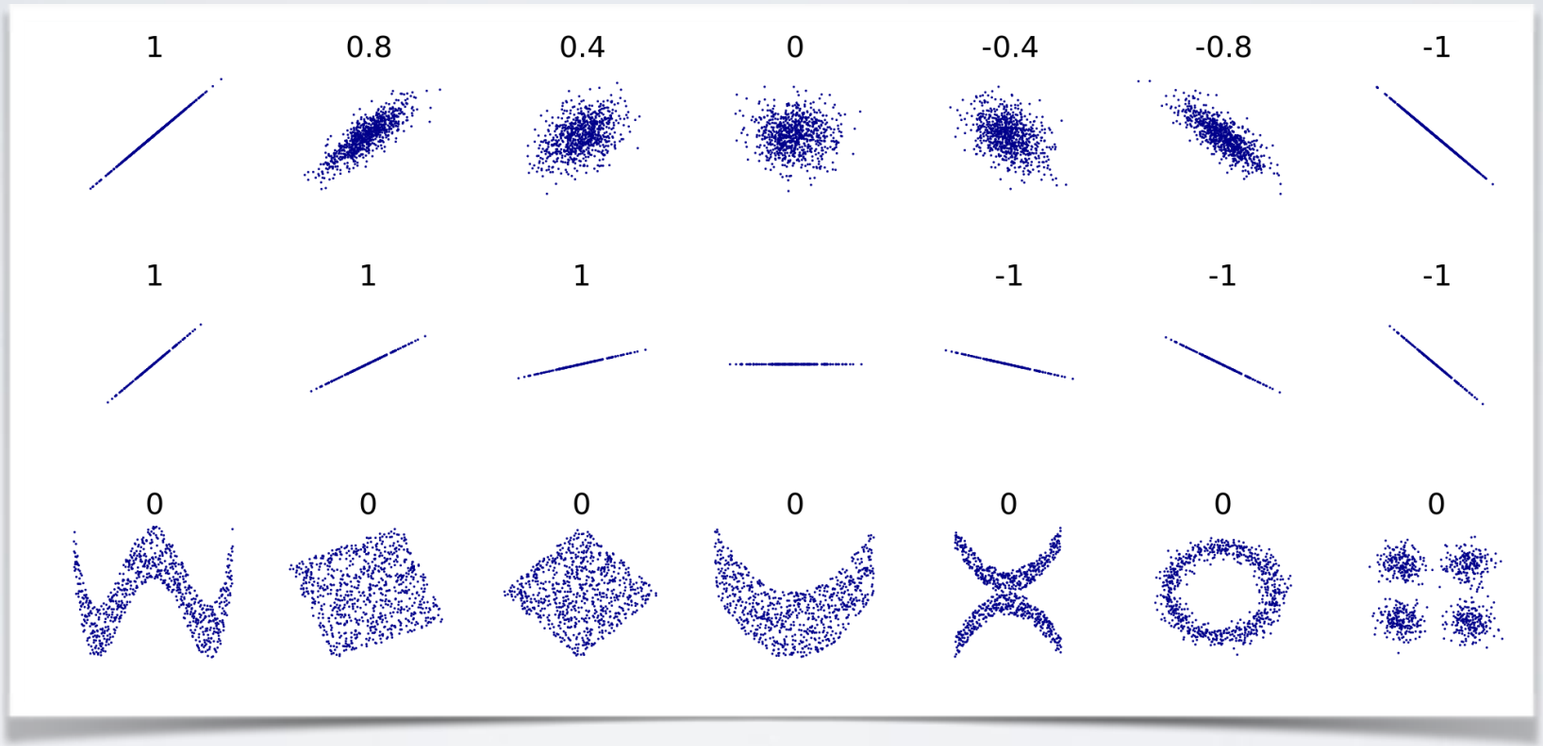
- Covariance matrix $\mathbf{K}$
  - Extension of Variance to multivariate data
  - $\mathbf{Var}(X) = \mathrm{E}[(X-\mu)^2]$
  - $\mathbf{cov}(\mathbf{X},\mathbf{Y}) = \mathrm{K}_{\mathbf{XY}} = \mathrm{E}[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{Y}-\mathrm{E}[\mathbf{Y}])^{\mathrm{T}}]$
    - How much observation X differs from the mean ? And Y ?
    - Multiply the respective divergences of X and of Y for each item
    - Take the average
  - => $\mathbf{cov}(\mathbf{X},\mathbf{X}) = \mathbf{Var}(\mathbf{X})$

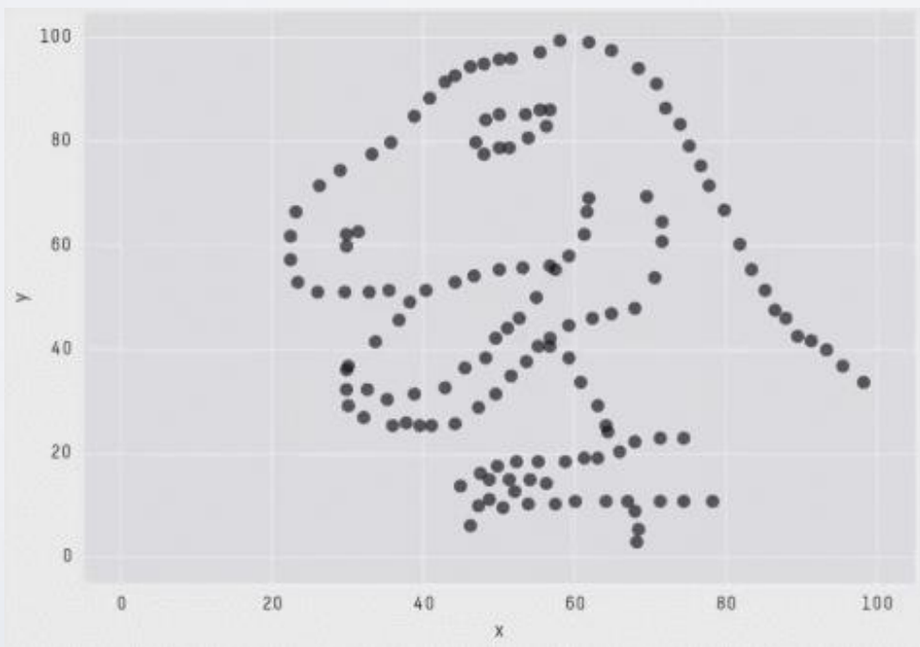
- Covariance is hardly interpretable by itself.
  - If >0, divergences tend to be in the same direction
  - Normalize it to obtain the "correlation coefficient"

CORRELATION COEFFICIENT

- Pearson correlation coefficient : $\rho_{X,Y} = \frac{\mathbf{cov}(X,Y)}{\sigma_X \sigma_Y}$
  - Normalize the Covariance by the Standard deviation.
  - Independent from magnitude, i.e., no need to have normalized data
  - Value in -1, +1.
    - +1 means a perfect positive linear correlation, i.e., X=aY
    - -1 a negative one, i.e., X=-bY
  - 0 can mean many different things

CORRELATION COEFFICIENT





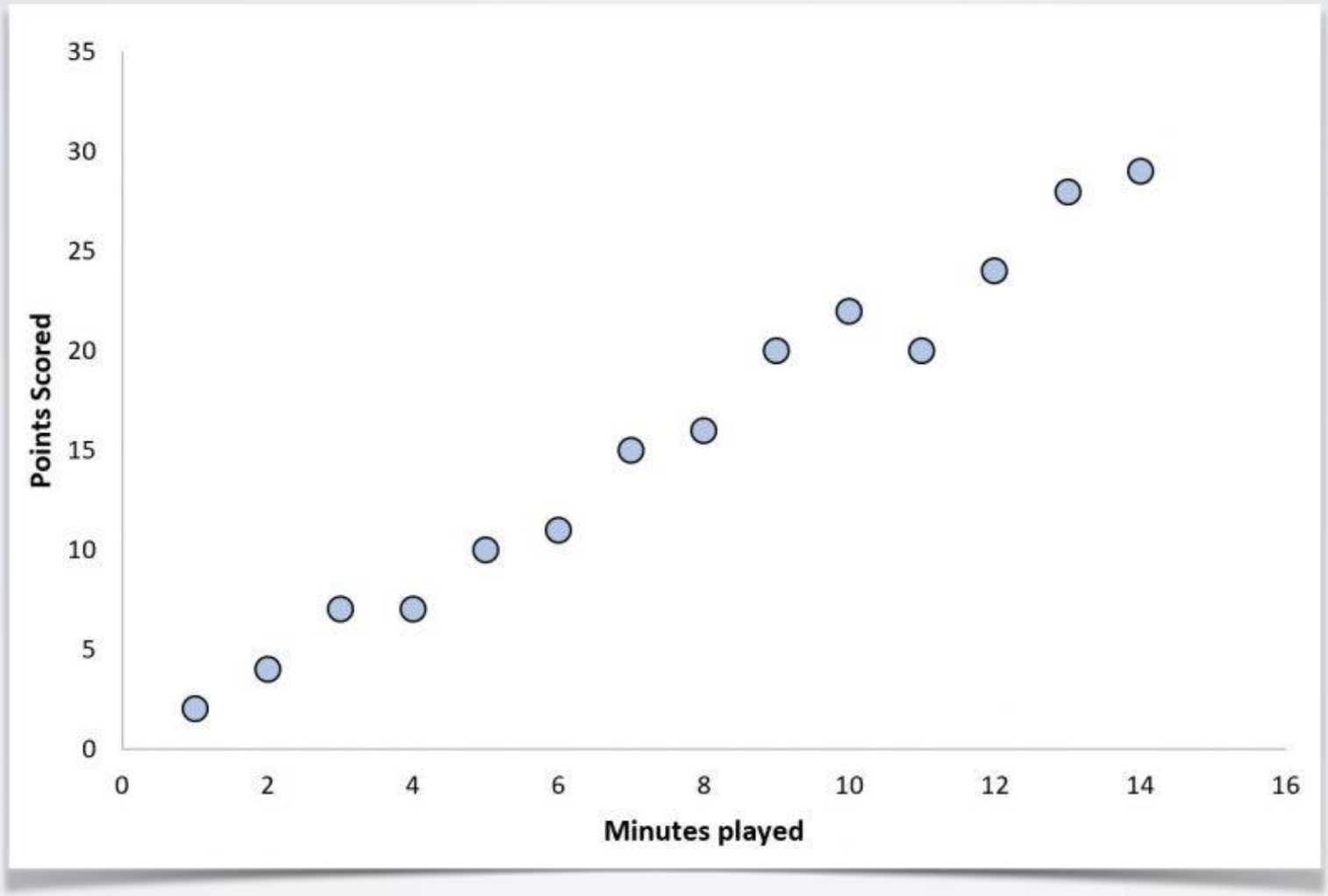
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

CORRELATION COEFFICIENT

- Other possible interpretation, e.g.
  - Cosine similarity of the vectors defined by the observations…

- 0.7 ? Is it a high or low value ?
  - It depends.

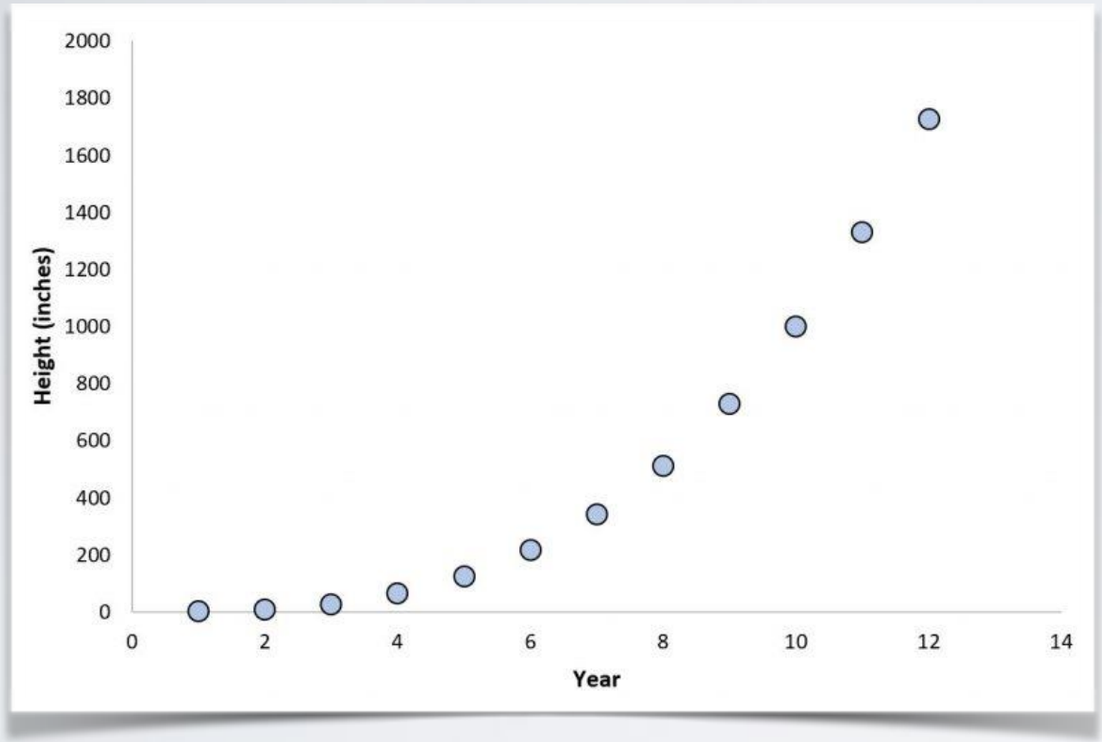
LINEAR RELATIONSHIPS

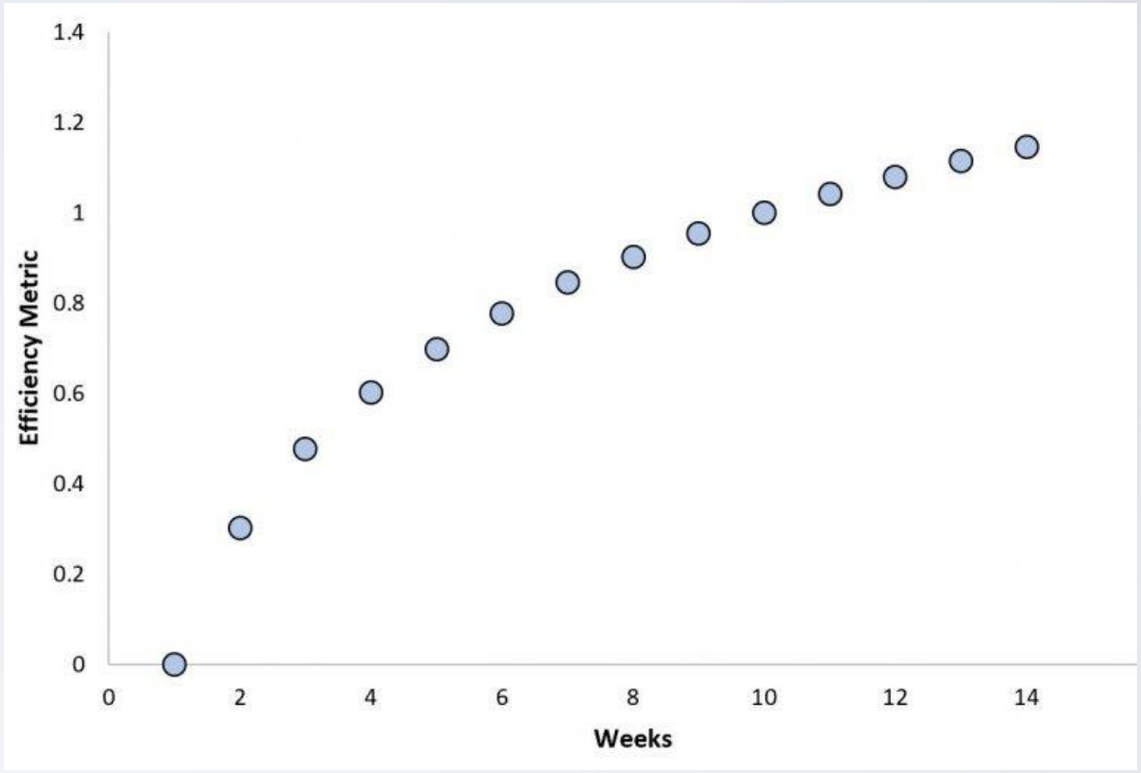


Linear relationship

$Y=a+bX+e$

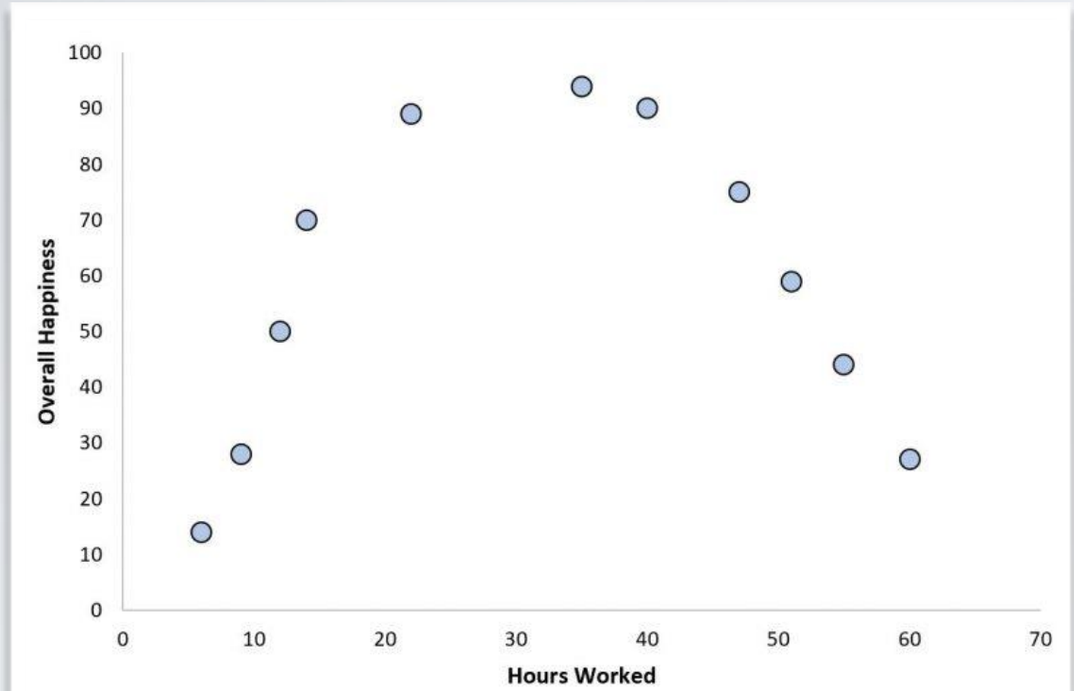
NONLINEAR RELATIONSHIPS

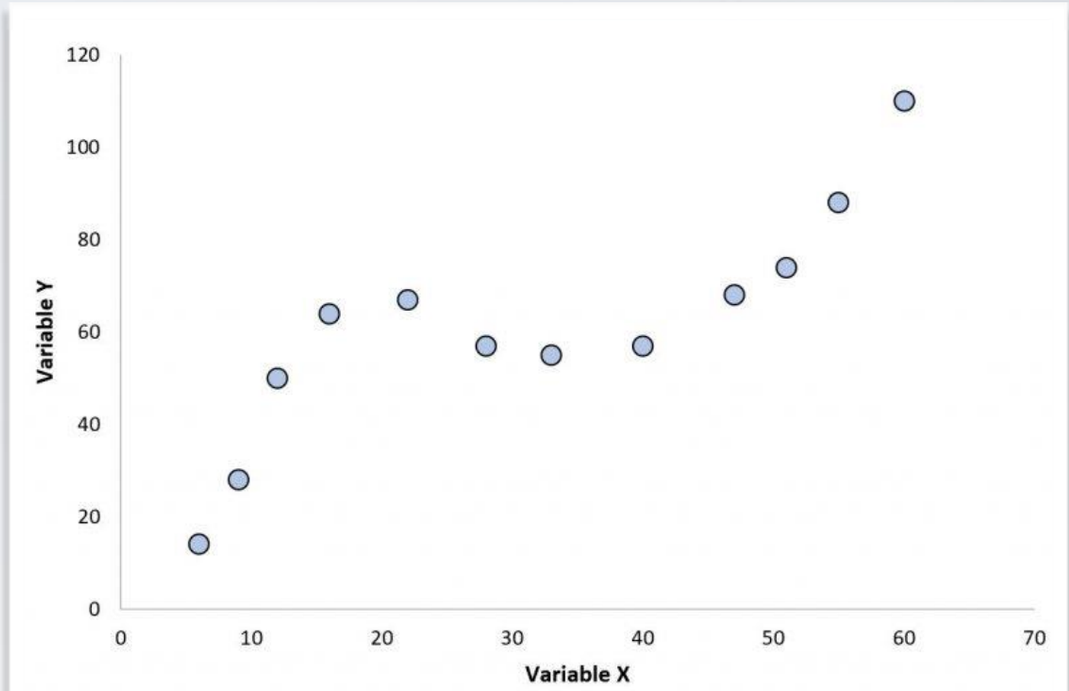




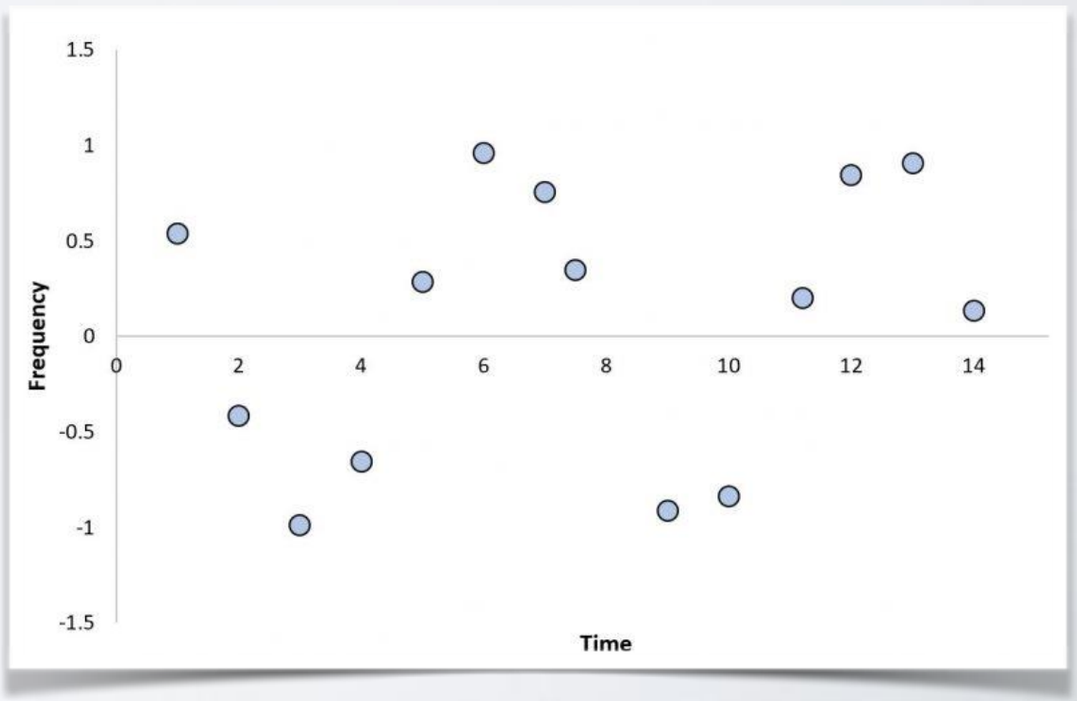
Monotonous, non-linear

NONLINEAR RELATIONSHIPS





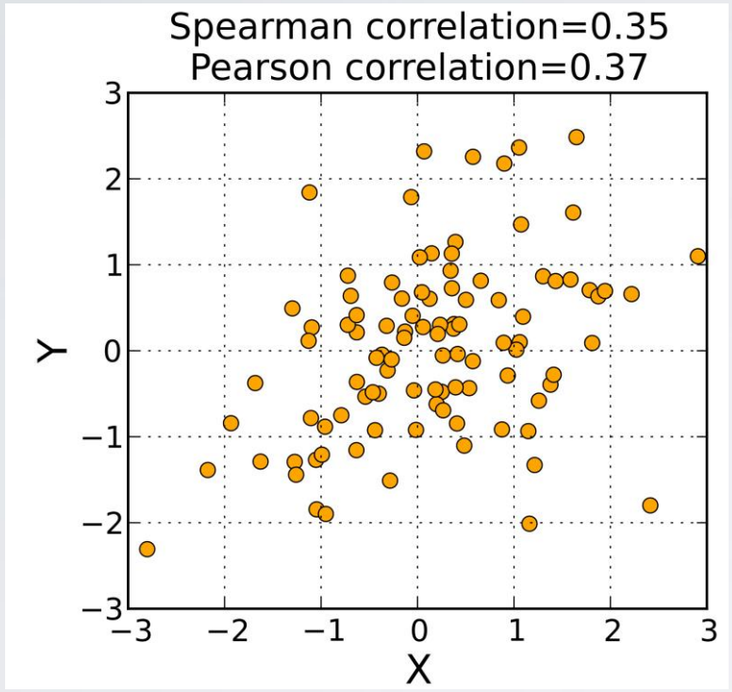
Non-monotonous,
Non-linear

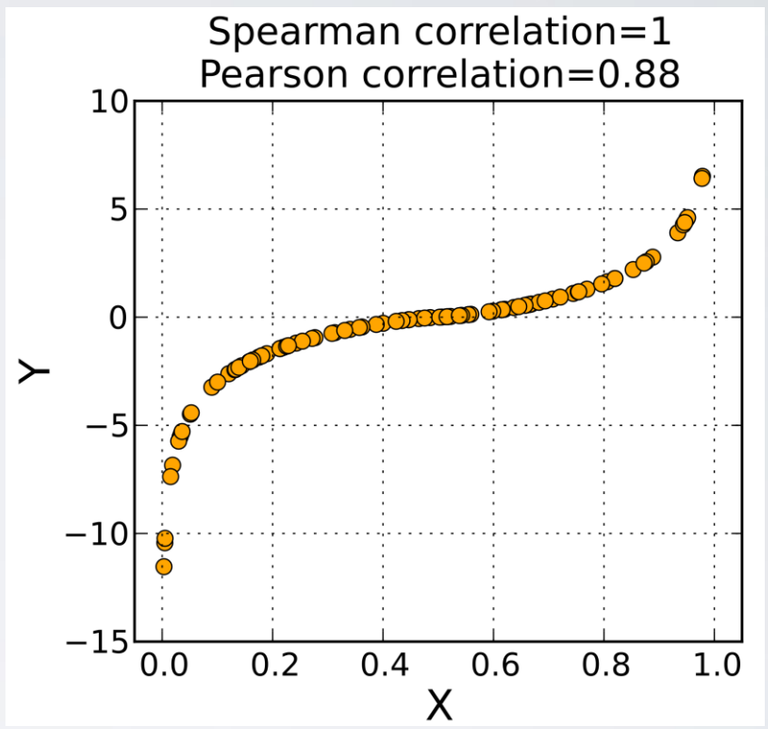


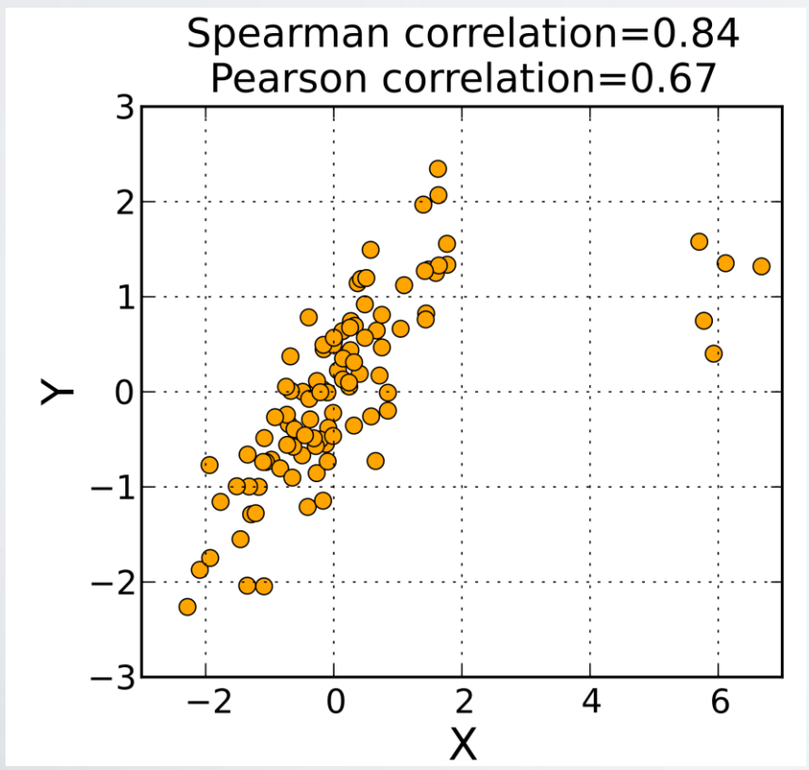
SPEARMAN'S CORRELATION

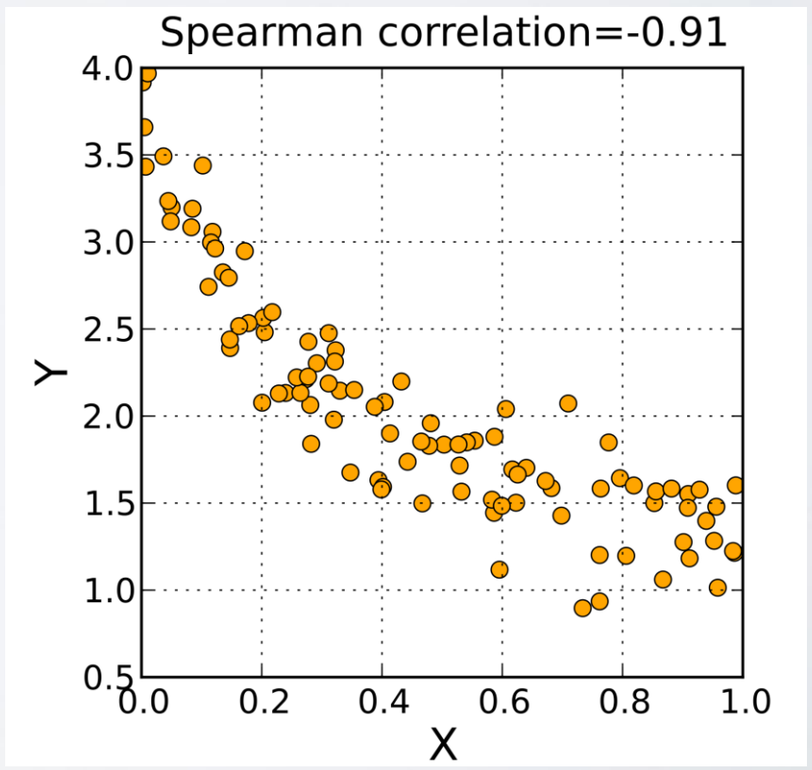
- Spearman's **rank** correlation coefficient
- Assesses how well the relationship between two variables can be described using a monotonic function
  - Not assuming a linear relation
- Pearson correlation coefficient between the rank variables
  - $r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}$

SPEARMAN'S CORRELATION



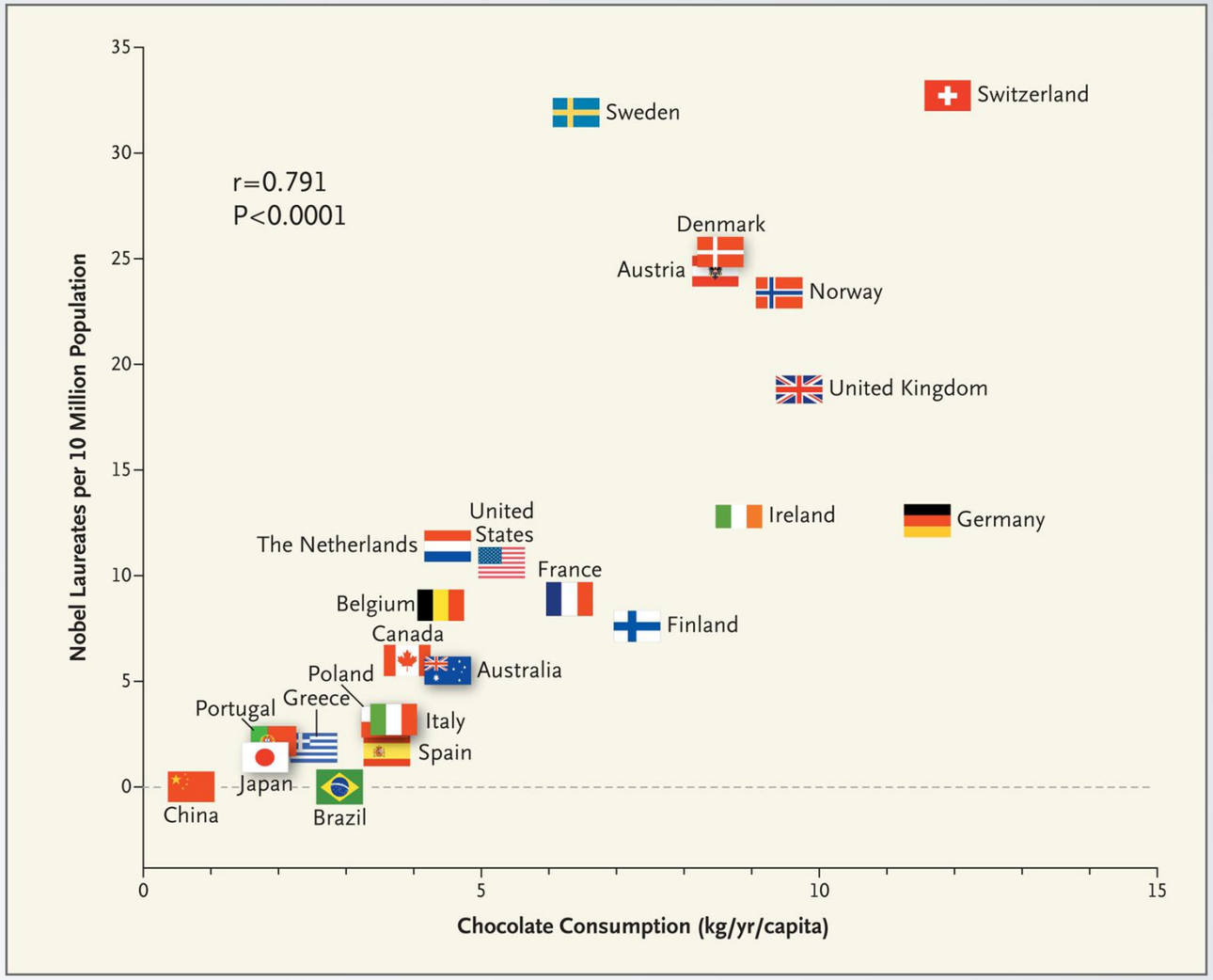






SURPRIZING CORRELATIONS

"There was a close, significant linear correlation **(r=0.791, P<0.0001)** between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries" [1]



Conclusion: "Chocolate consumption enhances cognitive function" [1]

[1] Chocolate Consumption, Cognitive Function, and Nobel Laureates | New England Journal of Medicine

FAKE

WARNING

- Correlation is not causation!!!
  - For more examples: Spurious Correlations

- Confounding variable:
  - an unobserved variable that affects both the cause being studied (Ferrari) and the effect observed (life expectation)
  - =>The main problem of any study. It is impossible (apart from strictly controlled experiments) to avoid this problem.
  - => **Be careful** when drawing conclusions from data

SOME ADVICES

- Always plot the relations between features
- Don’t believe single-number statistics. Never ever.

PART 5 - FEATURE SCALING

FEATURE SCALING: WHY

A

B

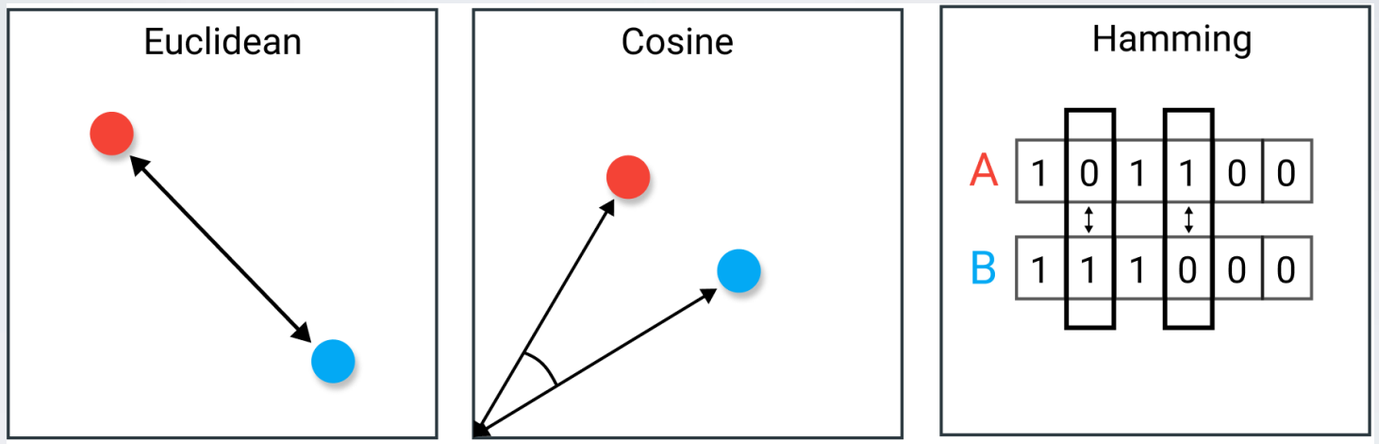
C

Y	Age: 20	Age: 20	Age: 90
m	Height: 1.82	Height: 1.82	Height: 1.50
g	Weight: 80 000	Weight: 81 000	Weight: 81 000

FEATURE SCALING: WHY

- We want to use euclidean distance to compute the "distance" between 2 people
  - a= (y:20,m:1.82,g:80 000), b=(y:20,m:1.82,g:81000), c=(y:90,m:1.50,g:80 020)
    - **d(a,b)=1000.0005**
    - **d(a,c)=72.8**
  - That is not what we expected from our expert knowledge!
    - We should normalize/standardize data

NOTIONS OF DISTANCE



FEATURE SCALING

- Rescaling (Normalization): $x' = \frac{x - min(x)}{max(x) - min(x)}$:[0,1]

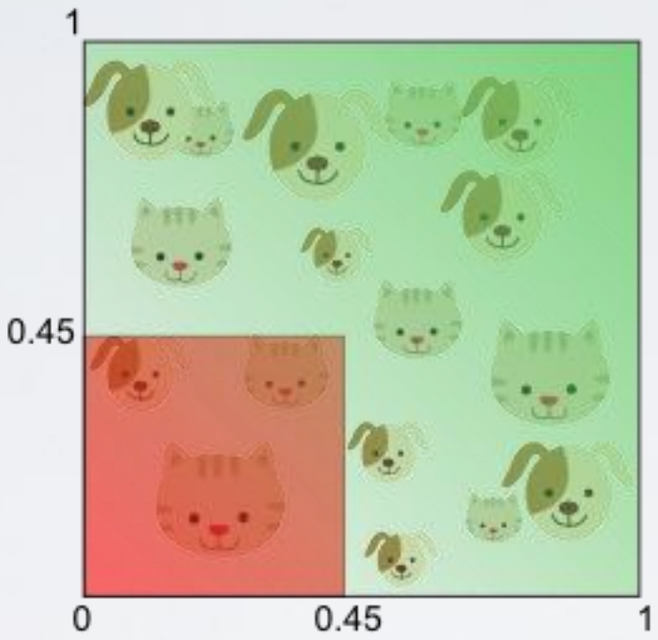
- Mean normalization: $x' = \frac{x - average(x)}{max(x) - min(x)}$: 0=mean

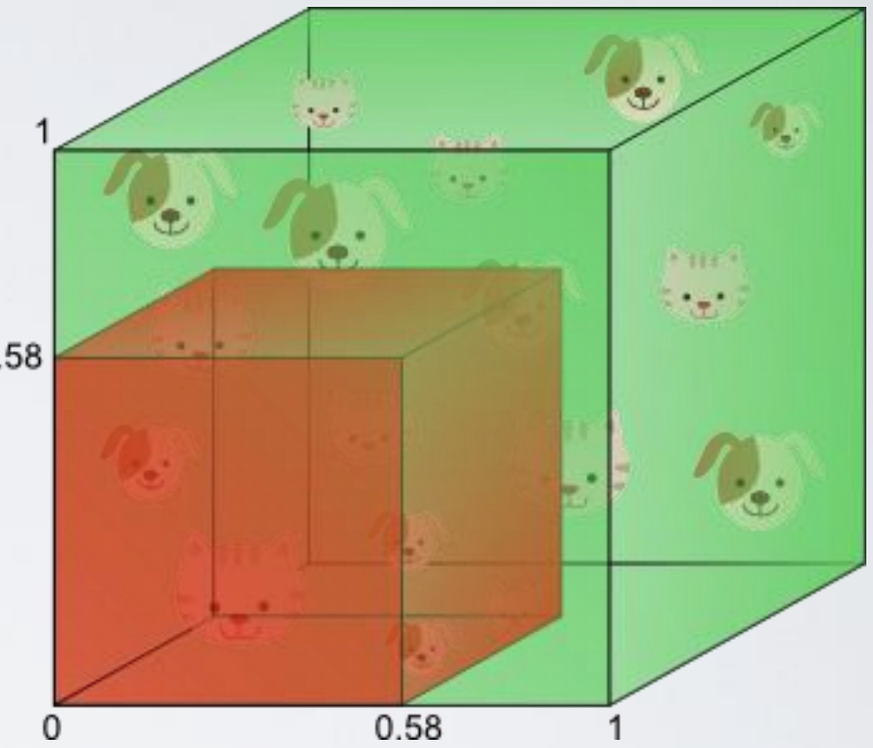
- Standardization (z-score normalization): $x' = \frac{x - \bar{x}}{\sigma}$
  - 0: mean, -1/+1: 1 standard deviation from the mean

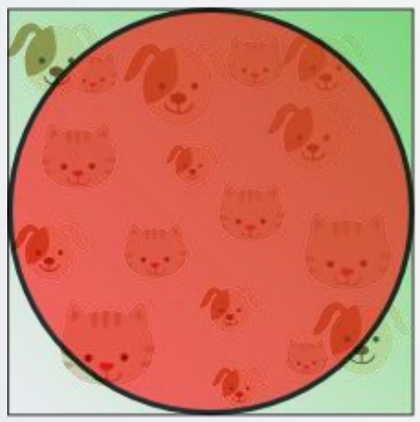
WARNING

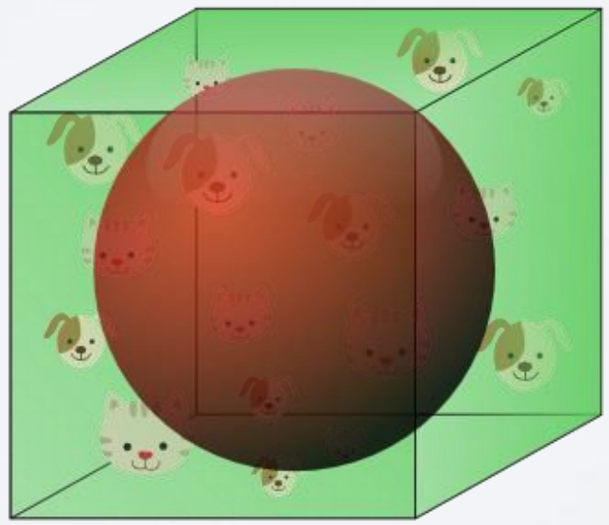
- There is no magic recipe!
- Everything cannot be normalized
  - Percentage scores, grades, scores between 0 and 1…
    - You could make low values big.
  - Binary variables (one hot encoded or not)
  - Careful with variables having an “absolute meaning”
    - Number of observations, duration of time, positions, distances, etc.

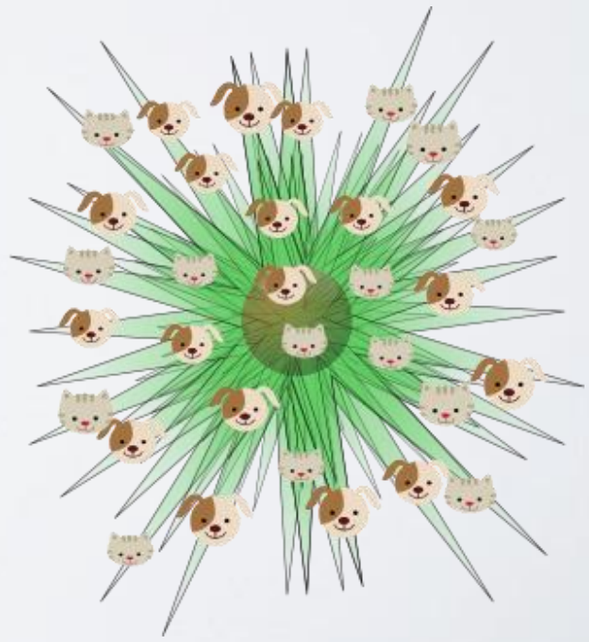
CURSE OF DIMENSIONALITY





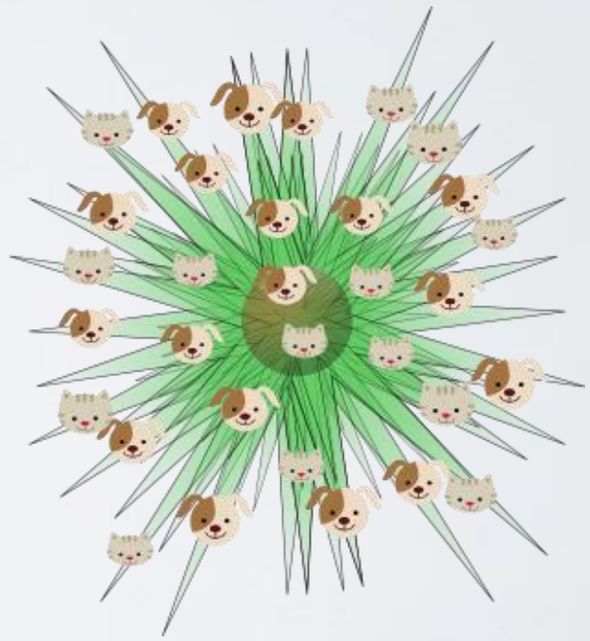






CURSE OF DIMENSIONALITY

- Every observation is "far" from any other observation
- Imagine you focus on the 80% most frequent values for each variable:
  - 1 var: Covers 80% of the population
  - 2 var: 64% of the population
  - 3 var: 51%
  - 10 var: 10%
  - 100 var: 0.00000002%

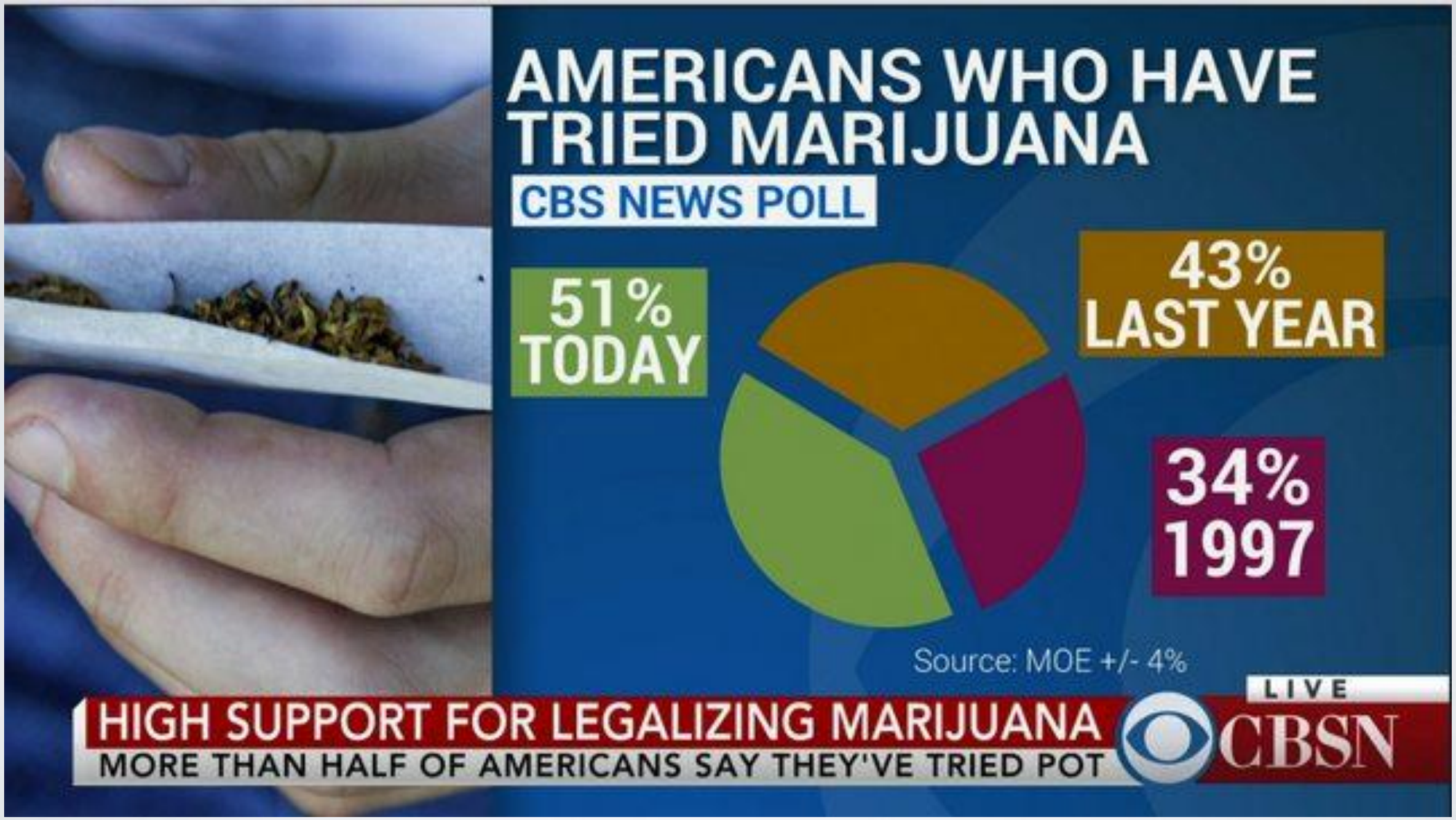


- If you have many variables, you need huge datasets, else you cannot generalize if all observations are completely different

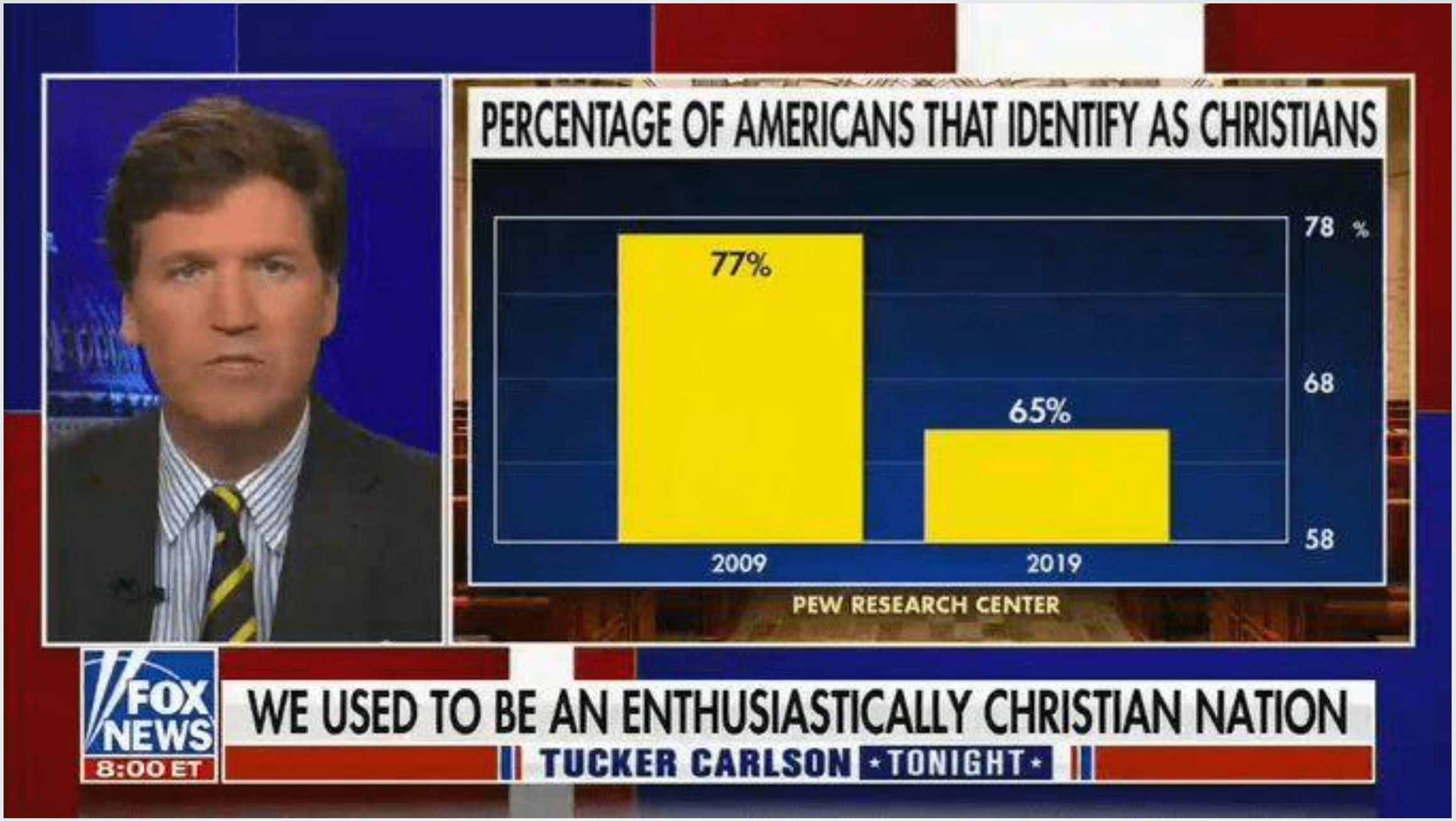
SOME EXAMPLES

of do and don't in data vizualization

SPURIOUS PIE CHARTS



SPURIOUS BAR PLOTS

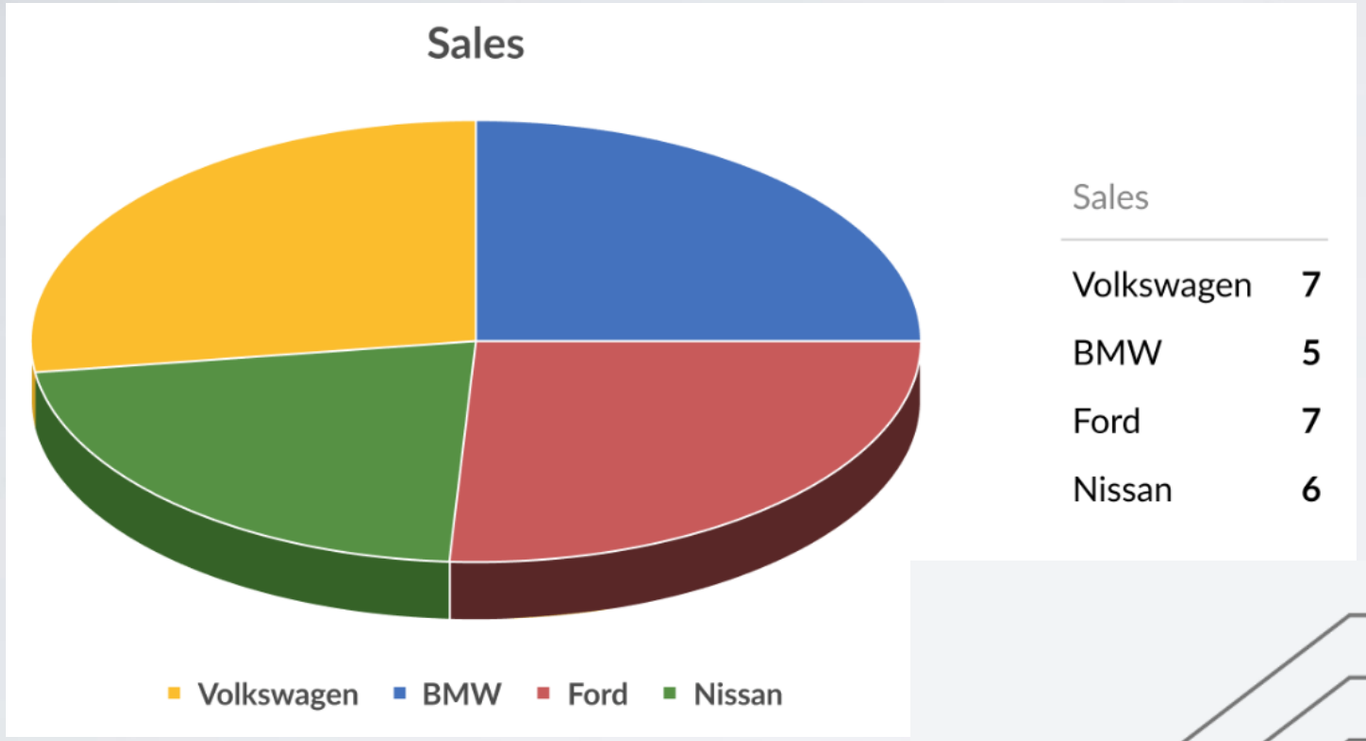


GOING FURTHER

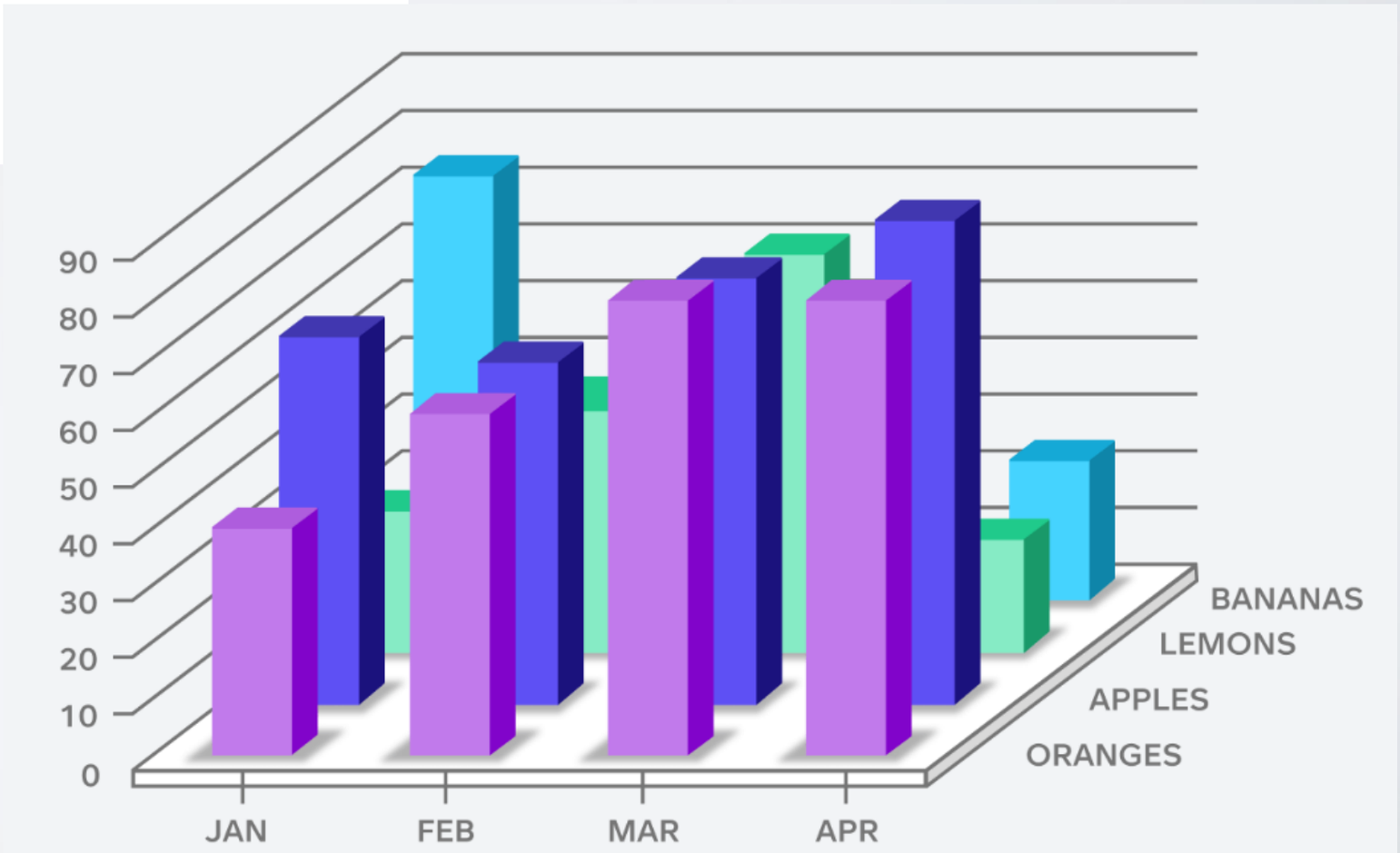
TUTO : Détecter des graphiques trompeurs – DEFAKATOR

TUTO : Survie sur les pics hostiles

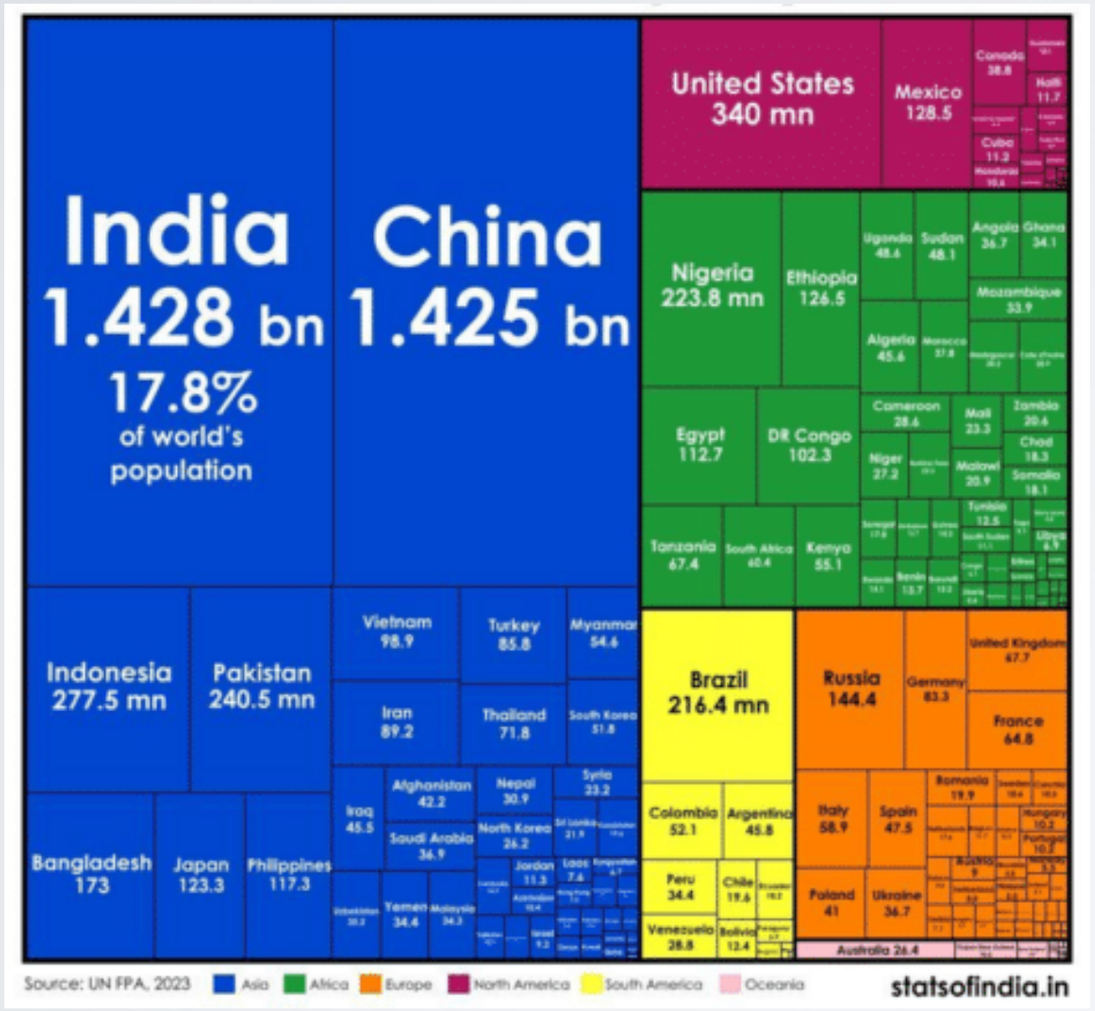
AVOID 3D CHARTS

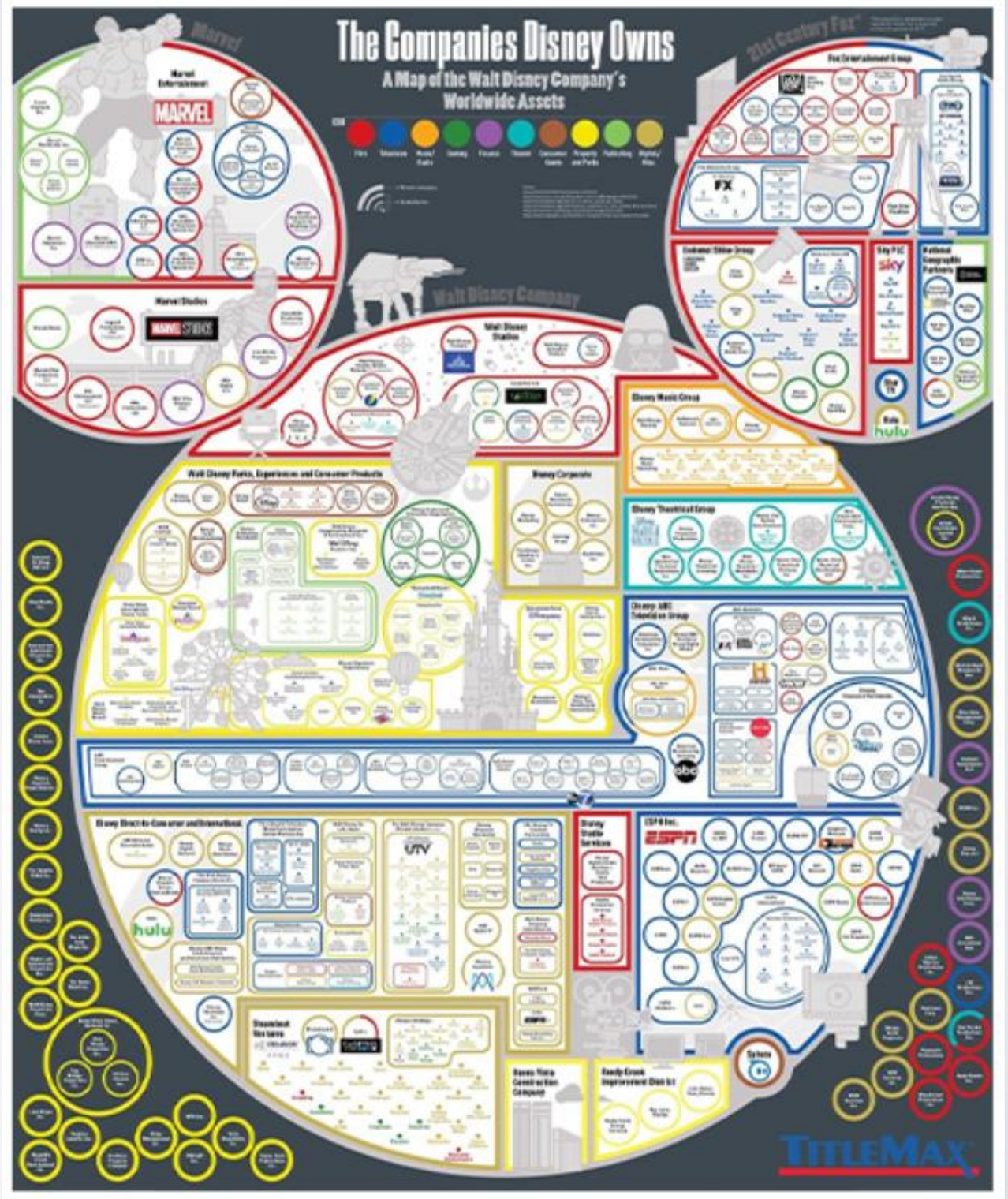


Sales	
Volkswagen	7
BMW	5
Ford	7
Nissan	6

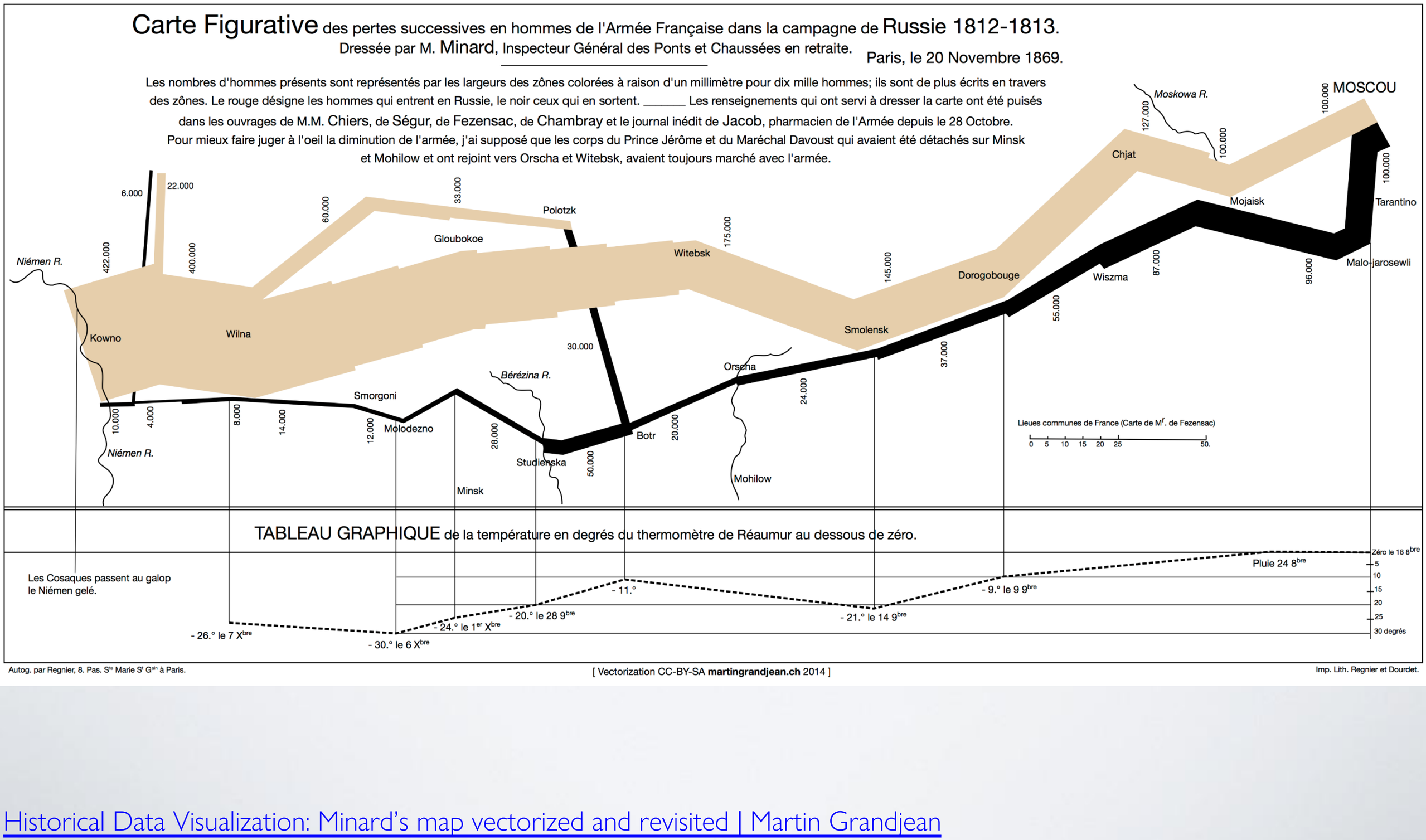


UNREADABLE CHARTS





MINARD'S MAP



[Historical Data Visualization: Minard's map vectorized and revisited | Martin Grandjean]

HOW DATAVIZ CAN BE IMPACTFUL

Quelle est réellement la fortune de Bernard Arnault ?

CONCLUSION

SOME "GOLDEN RULES"

- In real life:
  - Your data does not follow a normal distribution. Nor a power law, nor any other theoretical distribution
  - Your features are always correlated
  - You always have non-linear relationships
- GIGO: Garbage in, Garbage out
  - Real life are always garbage, because they are not produced to be analyzed
- Get to know your data
  - Exploratory Analysis
  - Listen their story and don't try to make them tell the story that you want to tell

IT'S YOUR TURN !

- Go to the webpage of the class and do today's experiments
- The "Advanced" section is not mandatory, you can do it if you have time