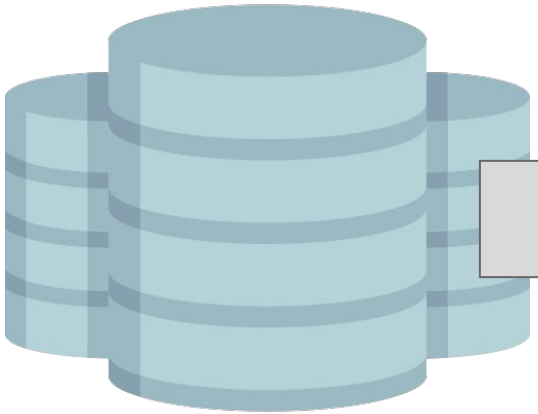
Data Science Survival Skills

Data exploration and visualization



DATA VISUALIZATION

Importance of context

First part

Exploratory

Showing all your data

Explanatory

Second part

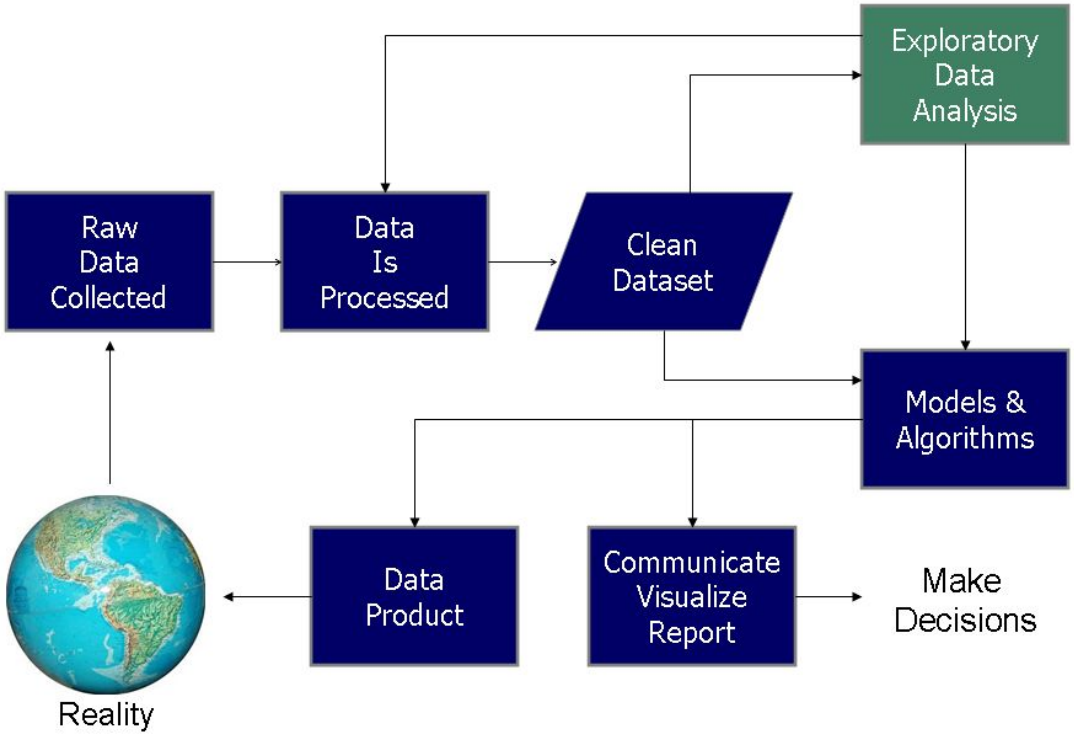
Showing only the relevant data

WHO?

WHAT?

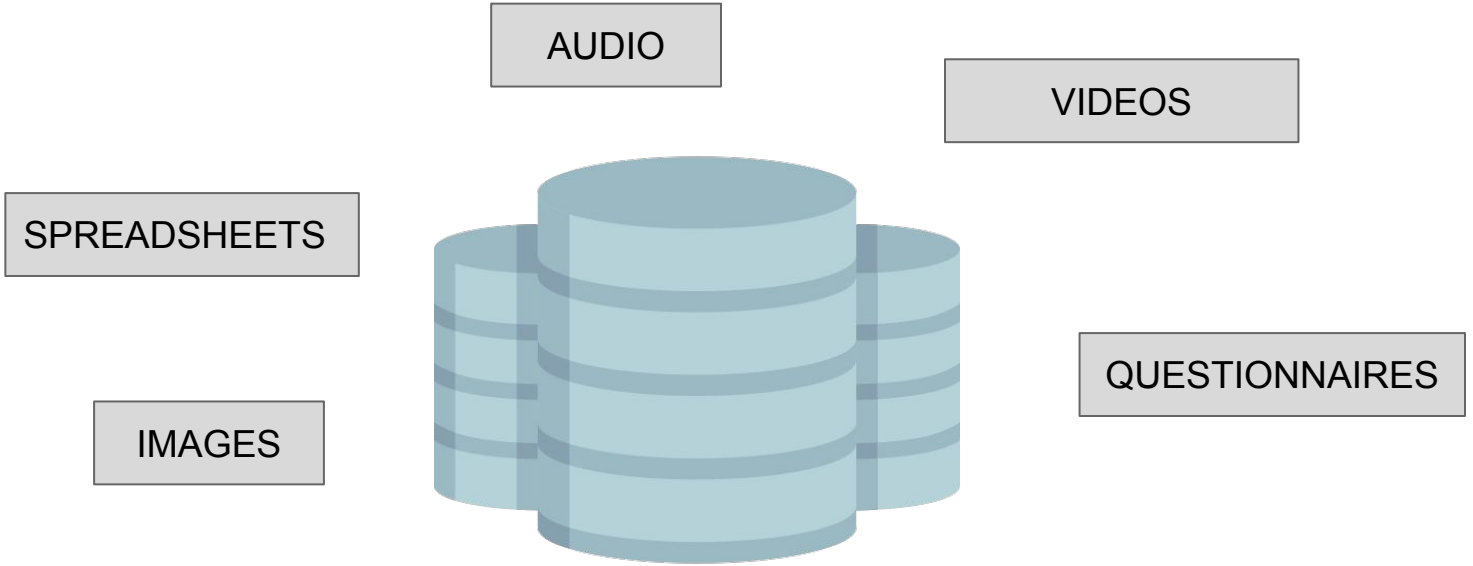
HOW?

Data Science Process

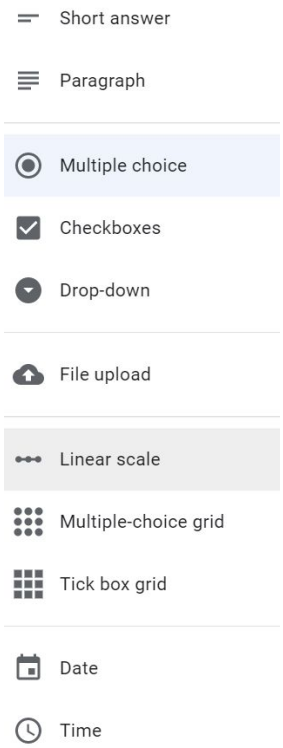


Doing Data Science, by Schutt & O'Neil (2013)

(Raw) Data?!



Data?!



Screenshot Google Forms

Pandas `dtype` mapping

Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

https://pbpython.com/pandas_dtypes.html

Categorical (Nominal)

Definition:
These are data points that represent distinct categories or groups.
There is **no inherent order or ranking** among these categories.
Special case: binary

Examples:
Colors of a shirt (red, blue, green),
types of fruits (apple, banana, orange),
or gender (male, female, non-binary).
Smoker: yes/no

© fruit of the loom

Key Characteristics:

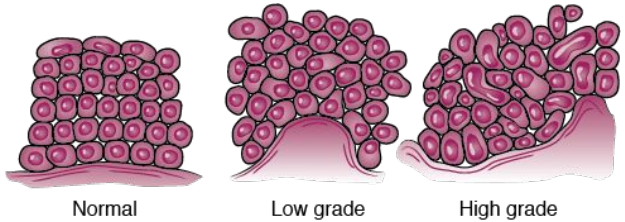
- Categories are **mutually exclusive**.
- **No quantitative relationship** between categories.

Ordinal scales

Definition: Ordinal data represent categories with a meaningful order or ranking, but the distances between the categories are not uniform or defined.

Examples:

- Income (low/middle/high)
- Tumor grading (I, II, III, IV, ...)
- Education level (elementary, high school, college, university)



https://bcan.org/facing-bladder-cancer/bladder-cancer-types-stages-grades/

Key Characteristics:

- There is a clear, **logical order** to the categories.
- Differences between data points do **not have a consistent scale**

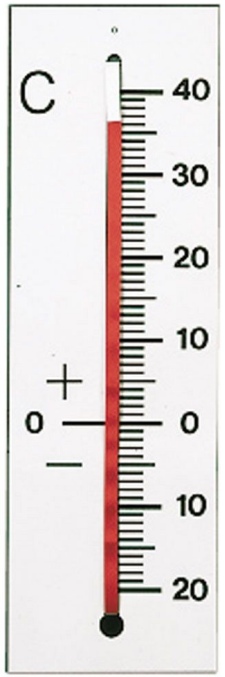
Intervals

Definition: Interval data represent numerical data where the intervals between each value are meaningful and consistent. This type of data allows for the comparison of differences between data points.

Examples: Temperature in Celsius or Fahrenheit (e.g., 10°C, 20°C, 30°C), calendar years (e.g., 1990, 2000, 2010).

Key Characteristics:

- The difference **between any two consecutive values is the same** (e.g., the difference between 20°C and 30°C is the same as between 30°C and 40°C).
- **There is no true zero point**. For example, 0°C does not mean the absence of temperature; it is just an arbitrary point on the scale.
- Arithmetic operations **like addition and subtraction are meaningful** (e.g., you can say that 30°C is 10°C warmer than 20°C).
- Multiplicative operations (e.g., ratios) are **not** meaningful because there is no true zero; saying "20°C is twice as warm as 10°C" would not make sense.



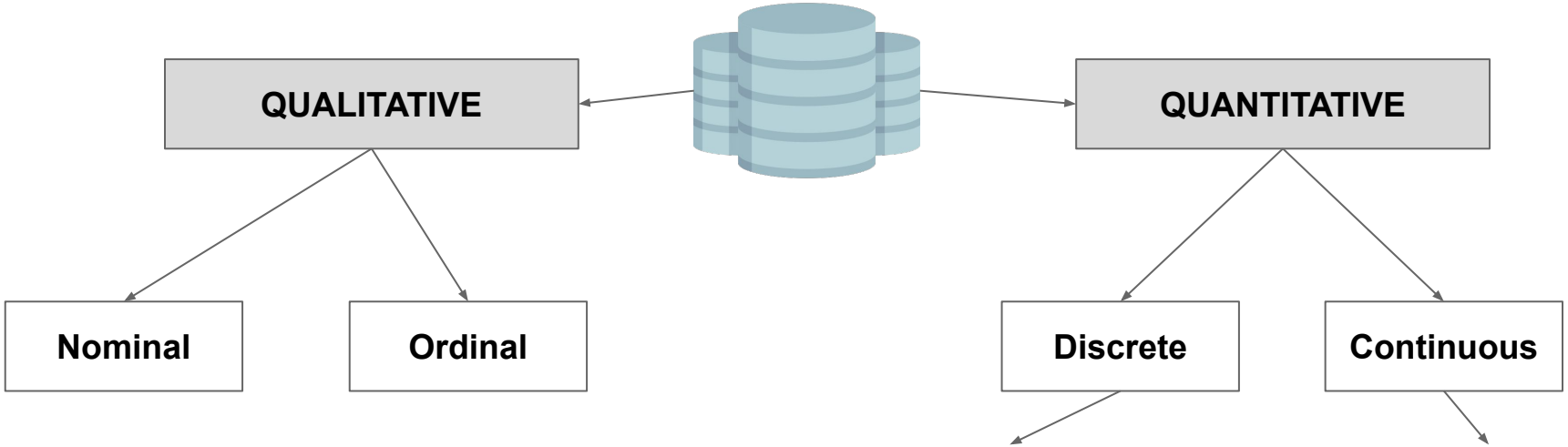
Ratios

Definition: Ratio data represent numerical data where both the order and the exact value between units are meaningful, and a true zero point exists. This true zero indicates the complete absence of the property being measured.

Examples:
Height (e.g., 150 cm, 175 cm),
weight (e.g., 50 kg, 100 kg),
distance (e.g., 5 km, 10 km),
and age (e.g., 20 years, 30 years).

Key Characteristics:

- The difference between values is consistent, just like interval data (e.g., the difference between 5 kg and 10 kg is the same as between 15 kg and 20 kg).
- **True zero point**: 0 means the complete absence of the quantity (e.g., 0 kg means no weight).
- All arithmetic operations, including addition, subtraction, multiplication, and division, are meaningful (e.g., it makes sense to say 20 kg is twice as heavy as 10 kg or that 4 meters is half the length of 8 meters).



Something you can **count**, e.g. damaged items in a shipping or students in the lecture hall

Something you can **measure**, e.g. temperature in the lecture hall, radiation, time until the first student falls asleep

Generating structured data

Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7	2·8	6·7	2·0
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3	2·7		1·8
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7	3·3		
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2		
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8		
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0		
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8		
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

Fisher, 1936 I think

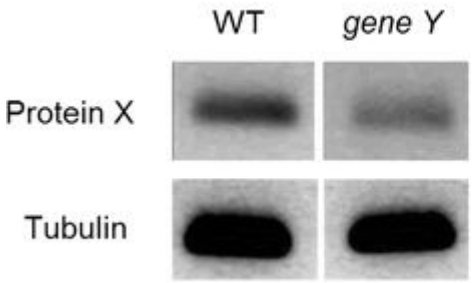
Iris Versicolor **Iris Setosa** **Iris Virginica**

From https://www.datacamp.com/tutorial/machine-learning-in-r

Feature extraction

Quantifying unstructured data

Western blot to quantify protein expression



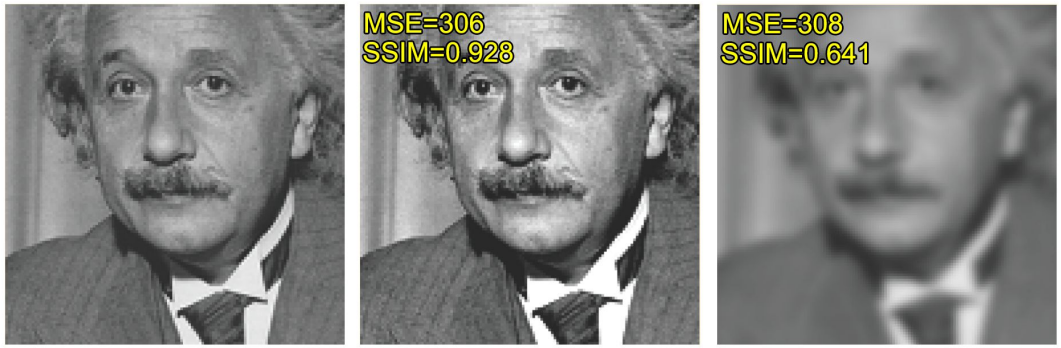
Protein of interest

Reference

Bell, BMC Biology 2016

Ratio: Expression = PoI / Ref

Finding appropriate metrics

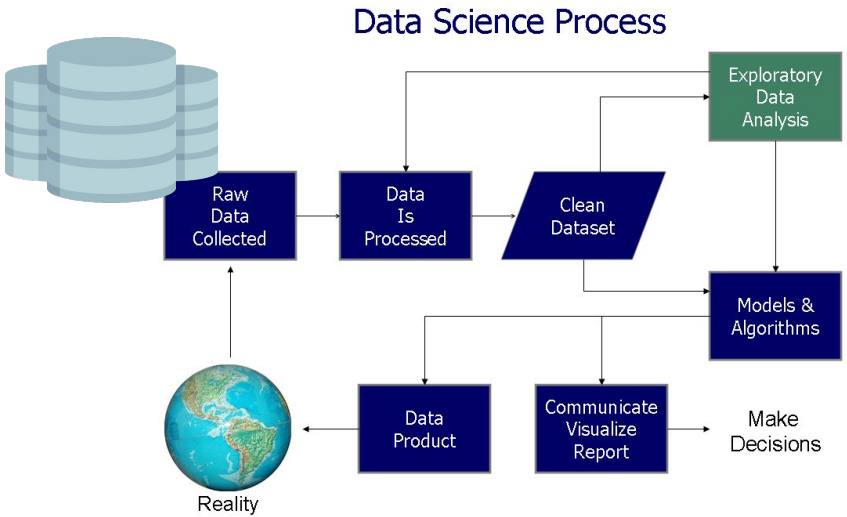


https://videoprocessing.ai/metrics/ways-of-cheating-on-popular-objective-metrics.html

Initial Data Analysis (IDA)

Early data quality checks:

- Is actually the data acquired we are looking for?
- How about outliers or extreme values?
- Is anything clipped?
- ...



Data quality

Descriptive **statistics**

Variable	Mean	SD	Min	P50	Max
DACKO	0.07958	0.08296	0.00041	0.05447	0.52698
AFEE	4.00251	0.35932	3	3.92942	5.27221
LEVERG	33.6211	22.8215	0.39982	30.2928	115.468
SIZE	7.23888	0.63189	5.46952	7.20824	9.08331
GROW	3.4533	31.5135	-8.4	1.02143	500.134
ROA	0.85876	10.6741	-79.328	0.84775	40.3836

Panel B: Descriptive Statistics – Dichotomous Variables

Variable	Frequency of 1's (Yes)	Frequency of 0's (No)	Percentage of 1's (Yes)	Percentage of 0's (No)
AUDSIZE	67	184	26.70%	73.30%

Panel C: Descriptive statistics of continuous variables by audit firm size

Variable	Big Four (N = 69)					Non-Big Four (N = 186)				
	Mean	SD	Min	P50	Max	Mean	SD	Min	P50	Max
DACKO	0.0791	0.0856	0.0007	0.0522	0.4495	0.0787	0.0811	0.0004	0.0543	0.527
AFEE	4.3365	0.4605	3.699	4.0792	5.2722	3.8804	0.2093	3	3.8891	4.2553
LEVERG	41.212	22.865	6.639	39.839	99.815	30.886	22.237	0.3998	27.286	115.46
SIZE	7.5096	0.8645	5.4695	7.3661	9.0833	7.1414	0.4911	5.8608	7.1786	8.0799
GROW	9.1597	60.918	0	1.1066	500.13	1.3978	3.5747	-8.4	0.993	45.429
ROA	-1.055	15.073	-79.33	1.0147	18.659	1.5482	8.5053	-28.37	0.8039	40.383

Mean,
Standard deviation
Min,
Max,
50th percentile (--> median)

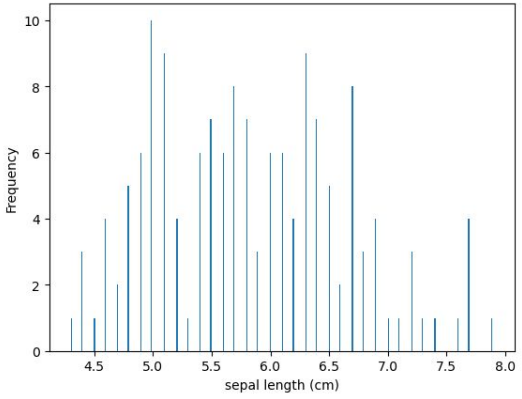
https://www.researchgate.net/figure/Descriptive-Statistics-Table-2-Descriptive-Statistics-Panel-A-Descriptive-statistics-of_tbl2_332371500

Histogram

Table I

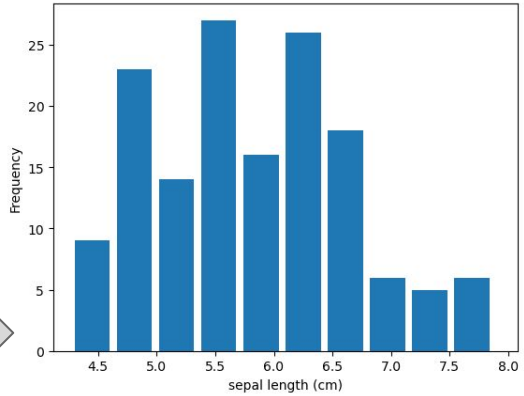
Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7											
5·1											
5·4											
5·1											
4·6											
5·1											
4·8											
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2	6·0	1·8
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8	4·8	1·8
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0	4·9	1·8
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8	5·6	2·1
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

How common are these values?

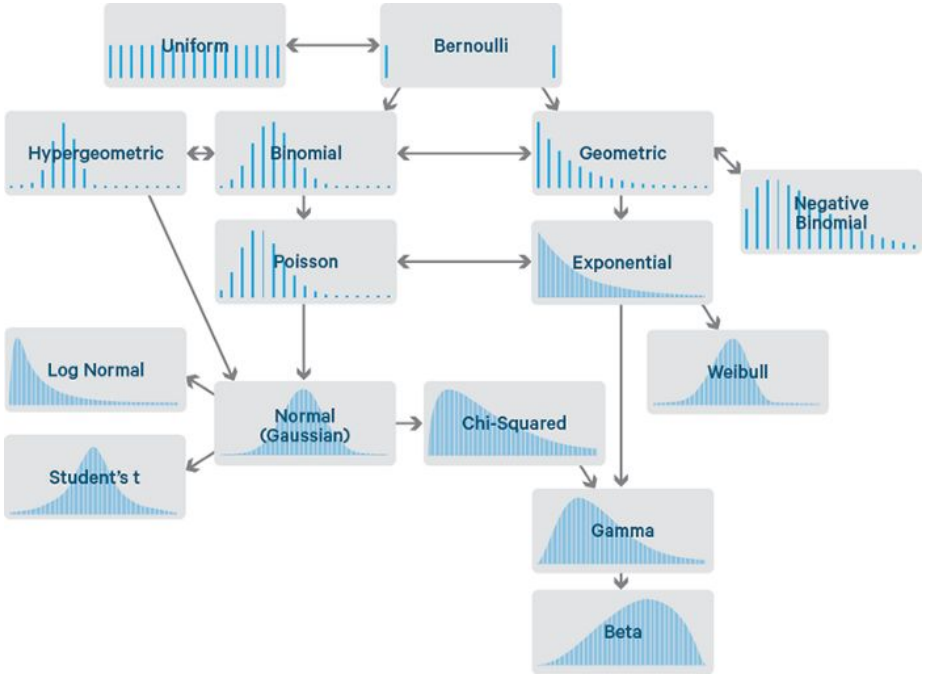


Karl Pearson

Adjust bins to allow intervals



Distributions



https://jagan-singhh.medium.com/types-of-probability-distributions-9333d18ed817

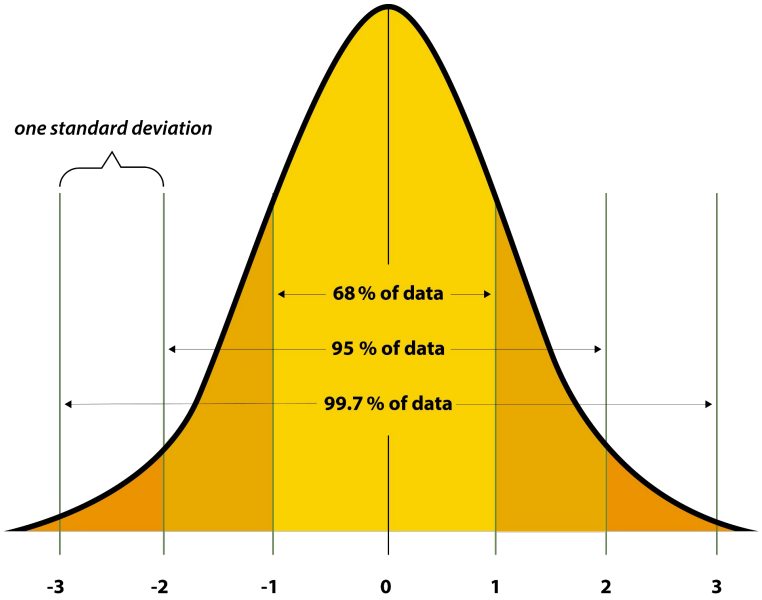
Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

f(x) = probability density function
σ = standard deviation
μ = mean

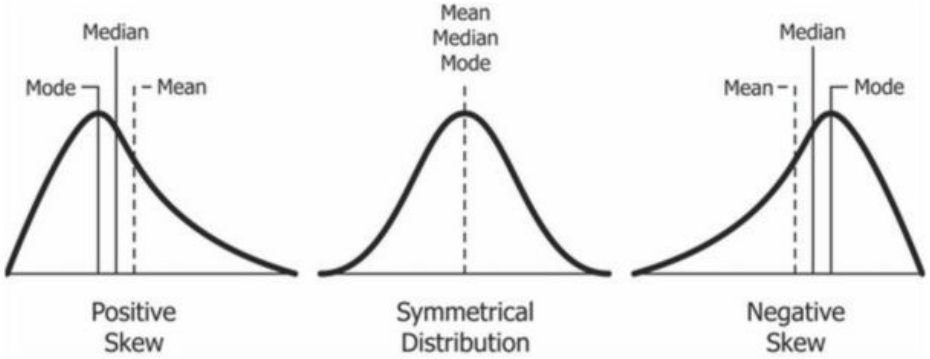
Mean, median and mode is the same value.

Example: heights of adults in a population



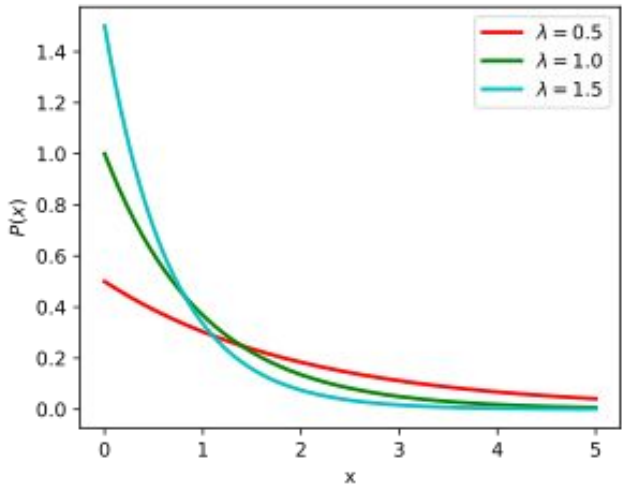
https://www.nlm.nih.gov/oet/ed/stats/02-800.html

Skewness


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Exponential

Definition: A continuous distribution used to model the time between events in a **Poisson process** (events occurring independently and at a constant rate).



Characteristics:

- Defined by the rate parameter (λ).
- The mean is $1/\lambda$, and the variance is $1/\lambda^2$.
- Skewed to the right, with no upper limit.

Example: Time between customer arrivals at a service center, the lifespan of an electronic component.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Poisson distribution

Definition: A discrete distribution that describes the number of events occurring **within a fixed interval of time or space**, assuming the events occur with a known constant mean rate and independently of the time since the last event.

Characteristics:

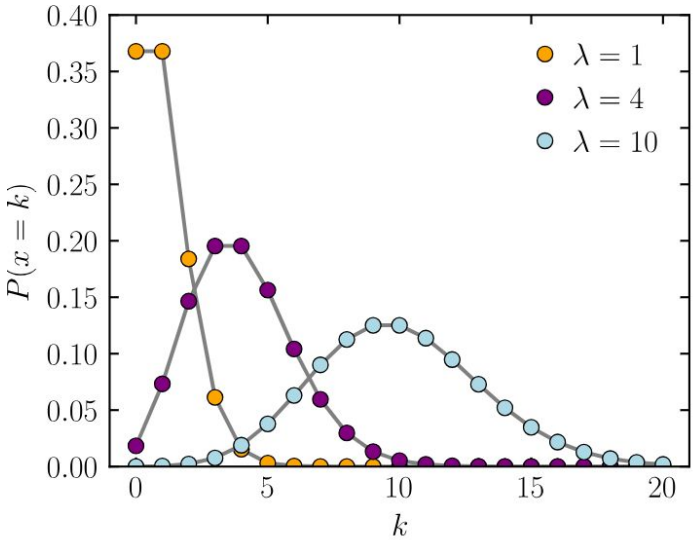
- Defined by the **rate parameter (λ)**, which is both the mean and variance.
- Values are non-negative integers (0, 1, 2, ...).

Example:
Number of emails received per hour,
number of customer arrivals at a store,
Photon shot noise in microscopy

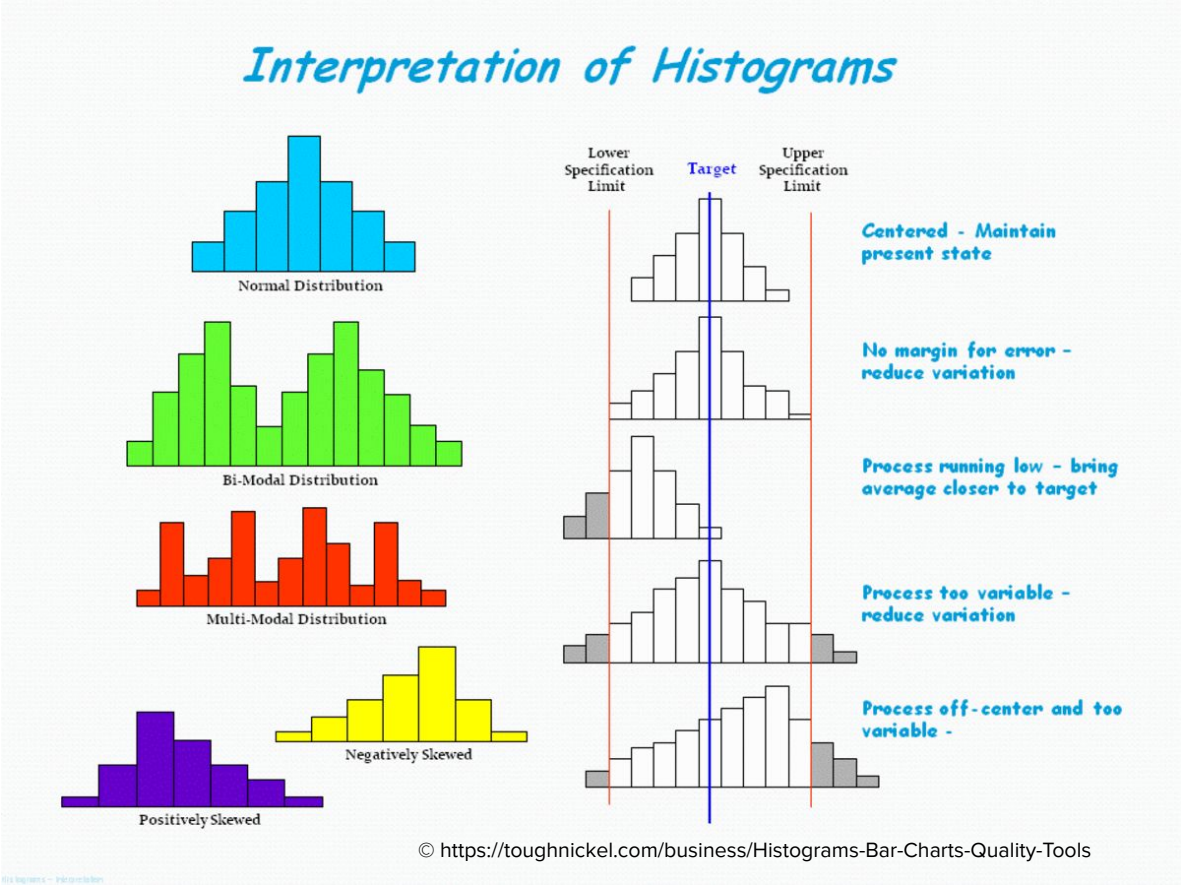
$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where

- k is the number of occurrences ($k = 0, 1, 2, \ldots$)
- e is Euler's number ($e = 2.71828\ldots$)
- $k!$ = $k(k–1) \cdots (3)(2)(1)$ is the factorial.



Histogram interpretations



© https://toughnickel.com/business/Histograms-Bar-Charts-Quality-Tools

Further things to check for...

Common method bias

⇒ Is there more variability due to the measurement than in the effect size?!

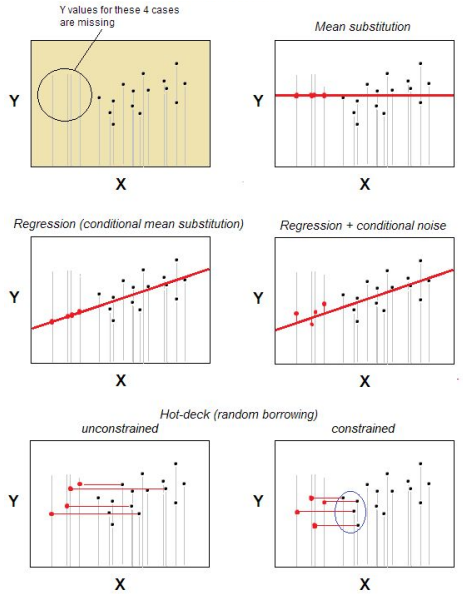
Imputation

⇒ work with missing data, how to treat NaN?

List deletion
➟ removal of complete cases when at least 1 value is missing (most common way)

Hot and Cold deck (within dataset or external, respectively)

About missing values, see next lectures



https://stats.stackexchange.com/questions/307339/hot-deck-imputation-it-preserves-the-distribution-of-the-item-values-how-c

Reliability (internal consistency)

Cronbach's alpha:
Measuring how consistent measurements are.
The higher the alpha, the higher the reliability.

Definition: Cronbach's alpha is a statistic that quantifies the reliability of a scale by assessing the average correlation among items in a test. It essentially shows how well the items in a set correlate with each other.

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Practical Applications

- **Questionnaire and Survey Design**: Researchers and survey designers use Cronbach's alpha to refine questions and ensure that the items are measuring the same concept.
- **Education**: Used to evaluate the reliability of exams and assessments.

Limitations

- **Not a Measure of Validity**: Cronbach's alpha only measures internal consistency, not whether the test actually measures what it is intended to measure (validity).
- **Dependence on Sample Size**: The reliability coefficient can be influenced by the sample size, with larger samples generally yielding more stable estimates.
- **Assumes Equal Contribution**: It assumes that all items contribute equally to the underlying construct, which may not be true in practice.

At the end of IDA

Non-normal distribution

Transformations? Change quantitative/qualitative structure?

Scales

Are the scales correct?
Is everything measured correctly?
Do we need to adapt something?

Missing data

How often does this happen? Which strategy is most applicable?

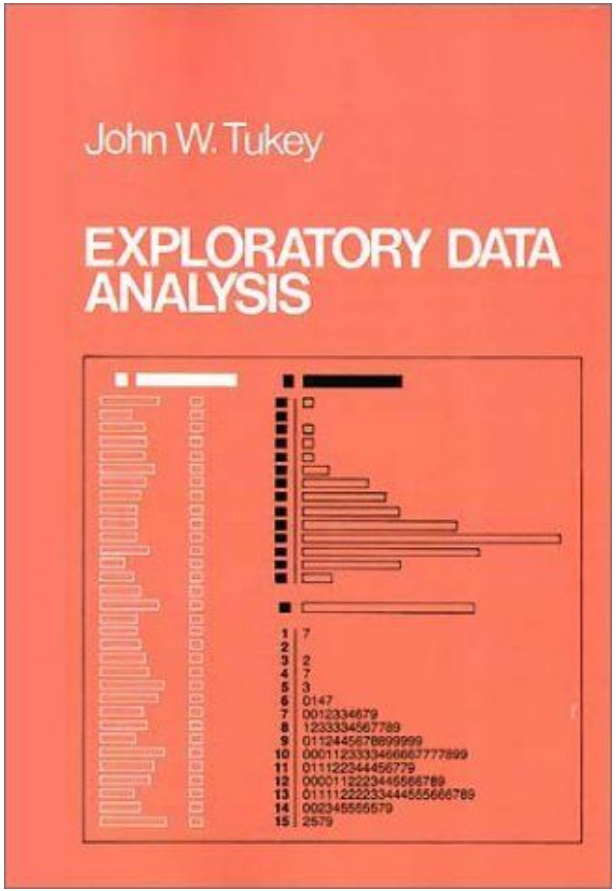
Hypothesis

Is the correct data acquired?
Do we need to refine the hypothesis?

Outliers

Robust analysis techniques? RANSAC methods?

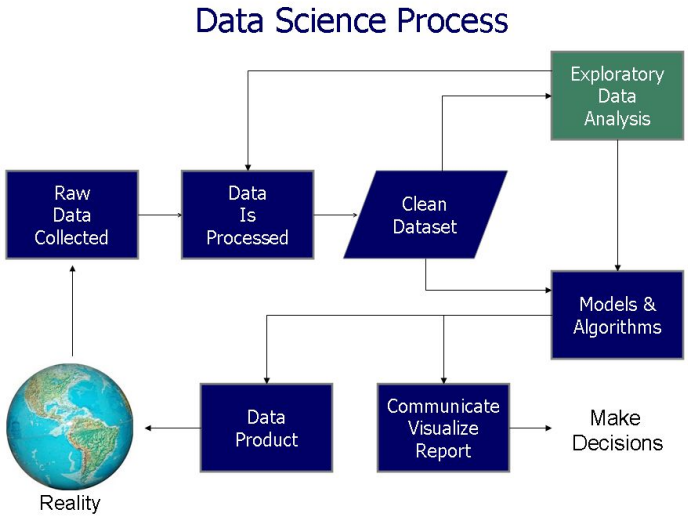
Exploratory Data Analysis (EDA)



John W. Tukey (1915-2000)

- Abolishment of nuclear weapons
- Fast Fourier Transform
- Statistics
- **BOXPLOT**

Process is similar to IDA



Use your **brain** and your knowledge - does this make sense? Is the data replicating reality? (e.g. correct order of magnitude)

Summary statistics

Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7	2·8	6·7	2·0
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3	2·7	4·9	1·8
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7	3·3	5·7	2·1
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2	6·0	1·8
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8	4·8	1·8
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0	4·9	1·8
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8	5·6	2·1
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

⇒ maybe for each target!

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

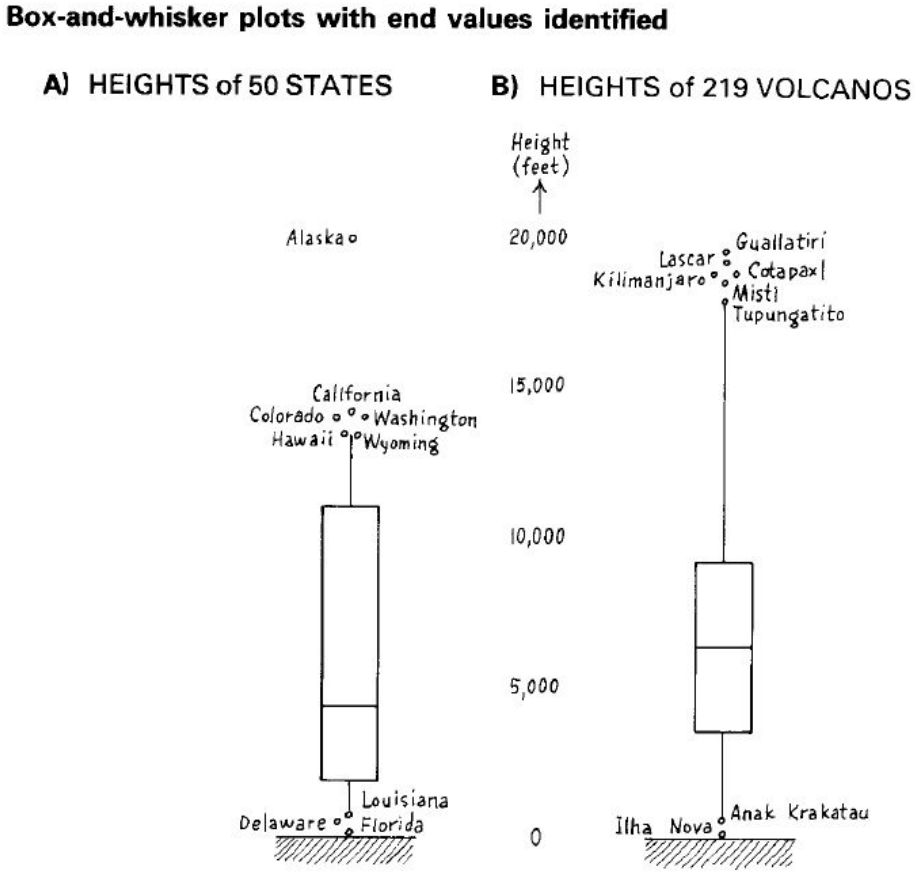
σ = population standard deviation

N = the size of the population

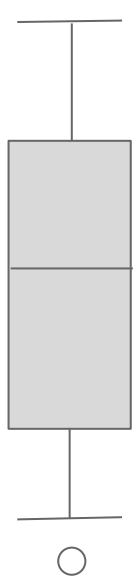
x_i = each value from the population

μ = the population mean

Visualization of quantiles: box-and-whisker



$$\mathrm{IQR} = Q_3 - Q_1 = q_n(0.75) - q_n(0.25)$$



Max, but max. 1.5 * IQR

Q3

Q2

Q1

Min, but max. 1.5 * IQR

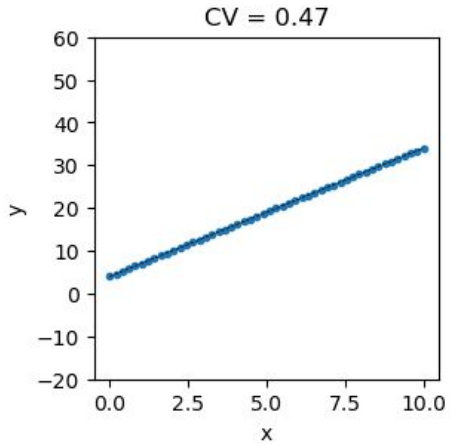
Outlier

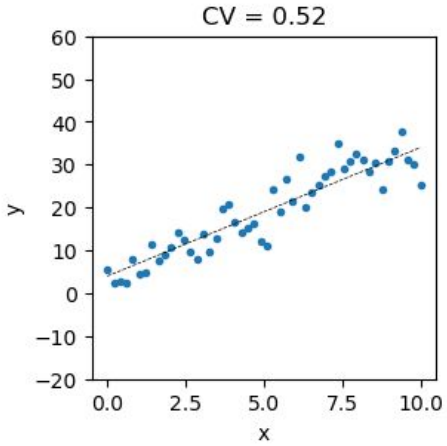
Make a box-and-whisker plot from DataFrame columns, optionally grouped by some other columns. A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of box to show the range of the data. By default, they extend no more than *1.5 * IQR (IQR = Q3 - Q1)* from the edges of the box, ending at the farthest data point within that interval. Outliers are plotted as separate dots.

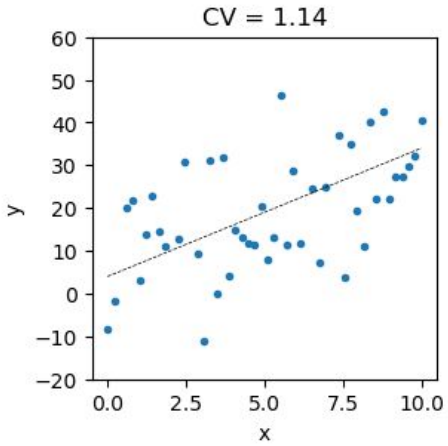
Coefficient of Variation

$$CV = \frac{\sigma}{\mu}$$

How much does the data vary?







Correlation vs. causation

Correlation

Pearson's correlation coefficient

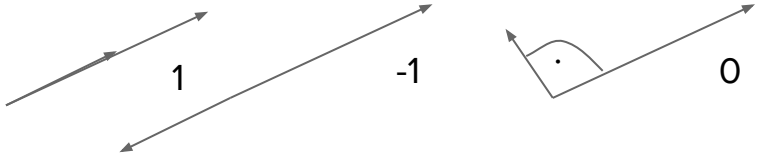
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Correlation:
Pair-wise relationship (how X varies with Y and Y varies with X)

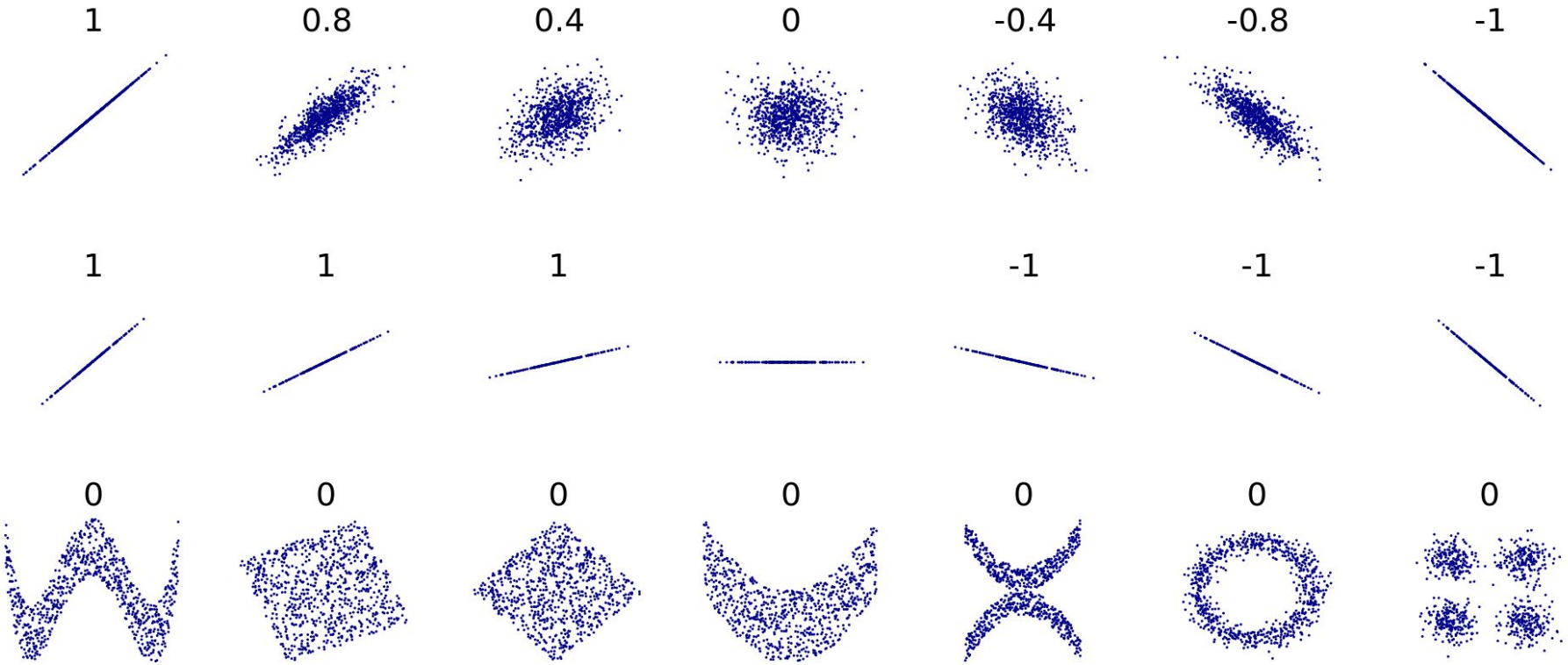
Linear regression:
How Y varies dependent on X
(Y = a + b*X)

Uncentered correlation coefficient

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$



Correlation



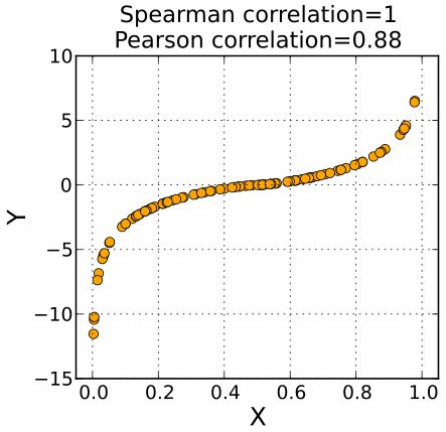
Other correlation metrics

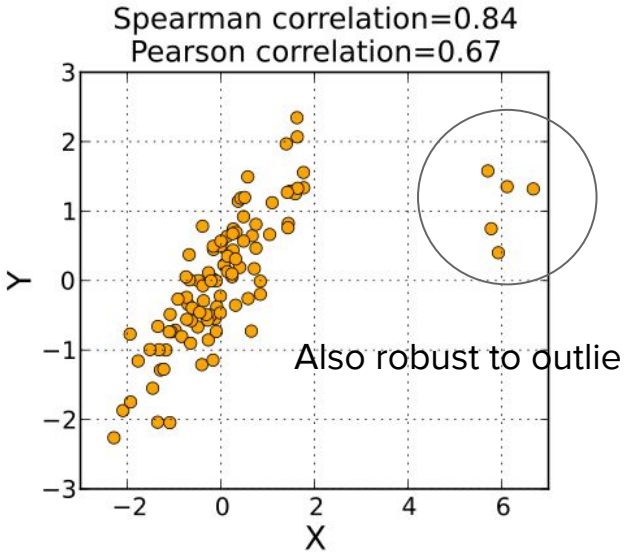
Spearman correlation (Spearman's rho) measures the **monotonic relationship** between two variables, which means it assesses whether as one variable increases, the other tends to increase (or decrease), **without requiring a linear relationship.**

Assumptions:

- The relationship is monotonic (it consistently increases or decreases, but not necessarily at a constant rate).
- The data can be ordinal, continuous, or even ranked.

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}},$$





Also robust to outliers

Phik (ϕk)

A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics

M. Baak[a], R. Koopman[a], H. Snoek[b], S. Klous[a,c]

[a]Advanced Analytics & Big Data, KPMG Advisory N.V., Amstelveen, The Netherlands
[b]Nikhef National Institute for Subatomic Physics / University of Amsterdam, Amsterdam Netherlands
[c]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Sex \ Handedness	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

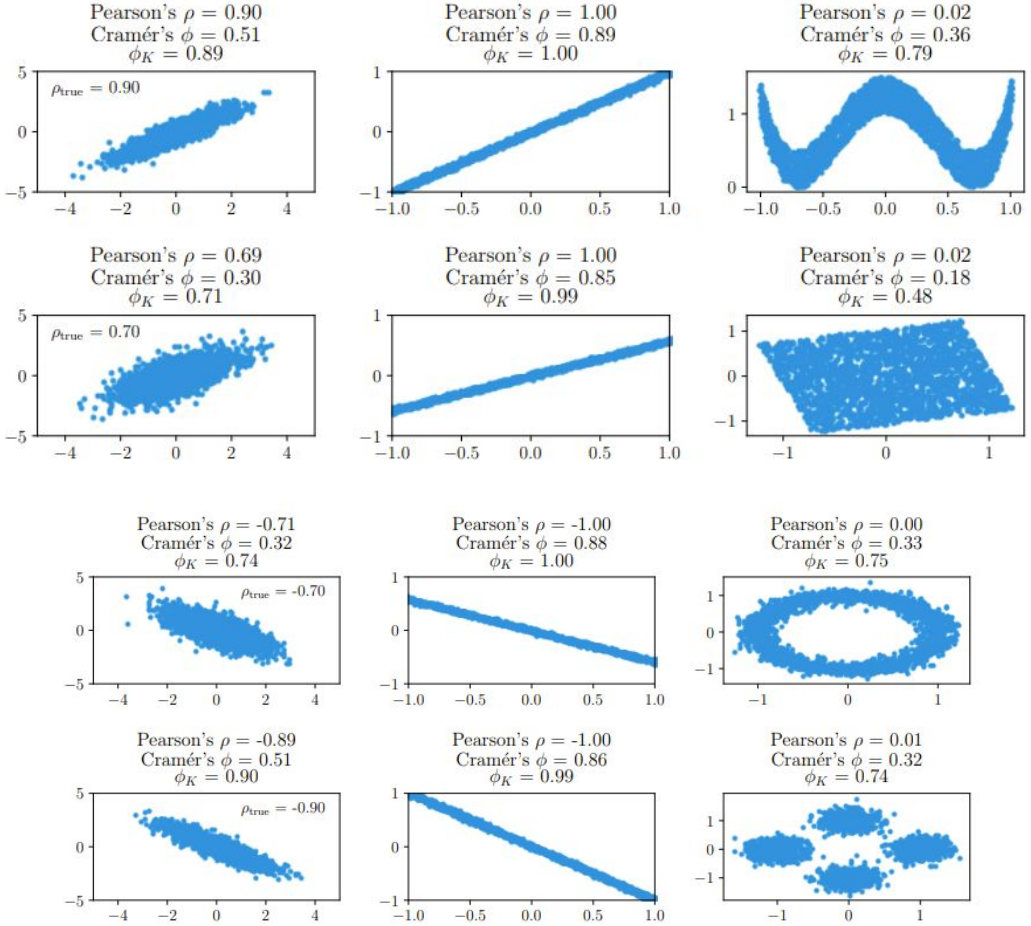
Based on Pearson's $\chi 2$ contingency test, related to Cramer's ϕ

Binning: works with categorical, ordinal and numerical data

Procedure description 1: the calculation of ϕ_K

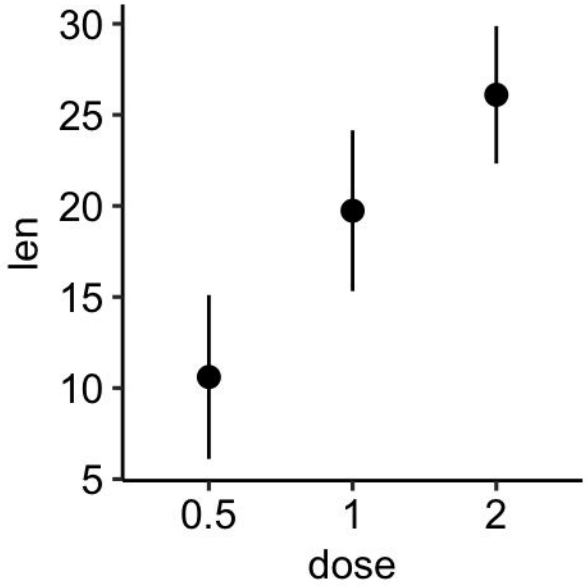
1. In case of unbinned interval variables, apply a binning to each one. A reasonable binning is generally use-case specific, and needs to be chosen such that the bin width is small enough to capture the variations observed in the data. As a default setting we take 10 uniform bins per variable.
2. Fill the contingency table for a chosen variable pair, which contains N records, r rows and k columns.
3. Evaluate the χ^2 contingency test using the Pearson's χ^2 test statistic (Eq. (7)) and the statistically dependent frequency estimates, as detailed in Section 3.1.
4. Interpret the χ^2 value as coming from a bivariate normal distribution without statistical fluctuations, using Eq. (17).
   i. If $\chi^2 < \chi^2_{\text{ped}}$, set ϕ_K to zero.
   ii. Else, with fixed N, r, k, invert the $X^2_{\text{b.n.}}$ function, *e.g.* using Brent's method (Brent, 1973), and numerically solve for ρ in the range [0, 1].
   iii. The solution for ρ defines the correlation coefficient ϕ_K.

Phik (ϕk)



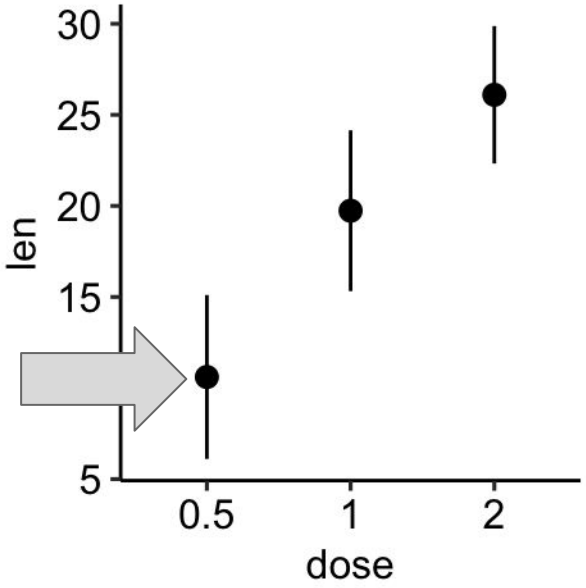
Shows if structure is present, but not its direction!

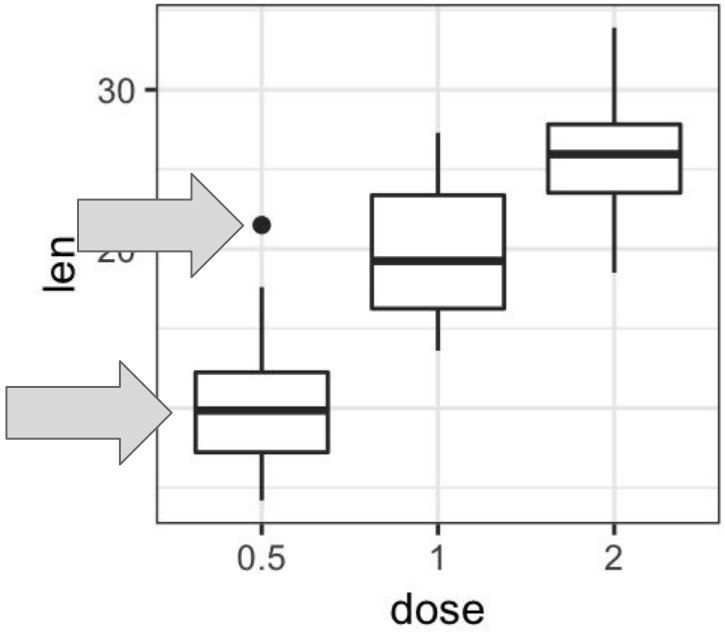
Plotting results



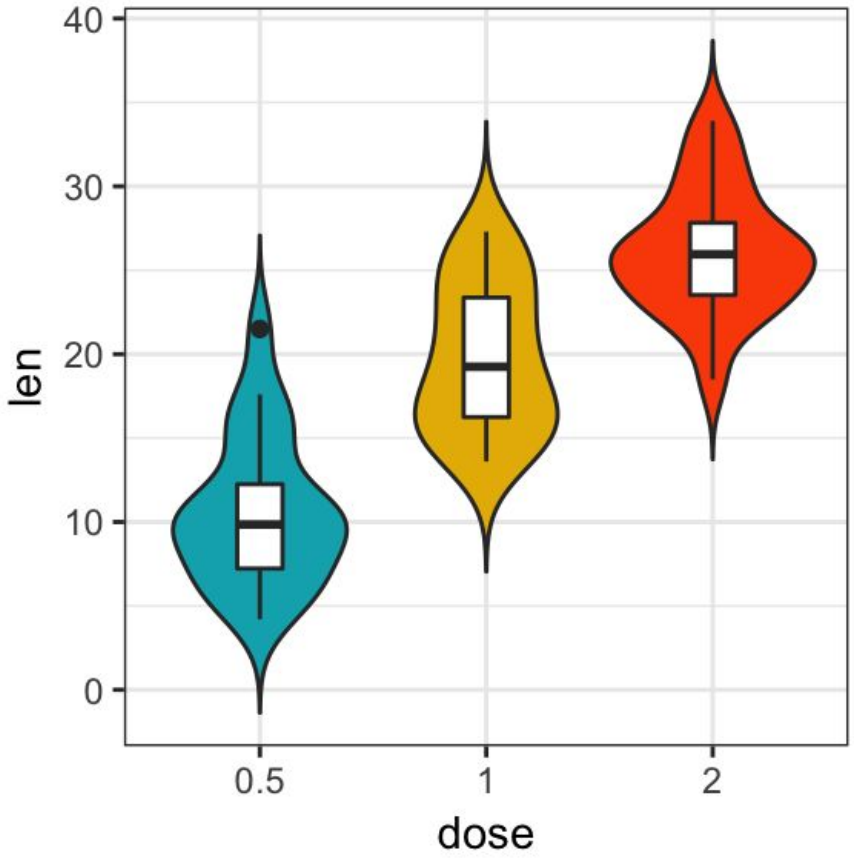
Where is the data?
Is it really well represented as MEAN +- STD?

Plotting results - boxplot?



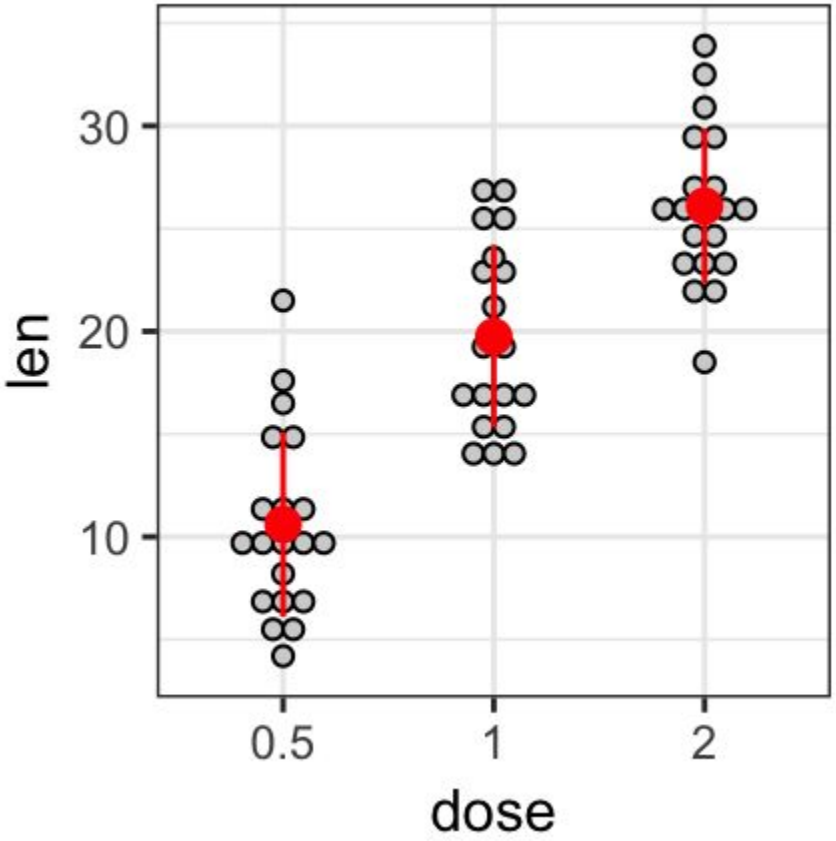


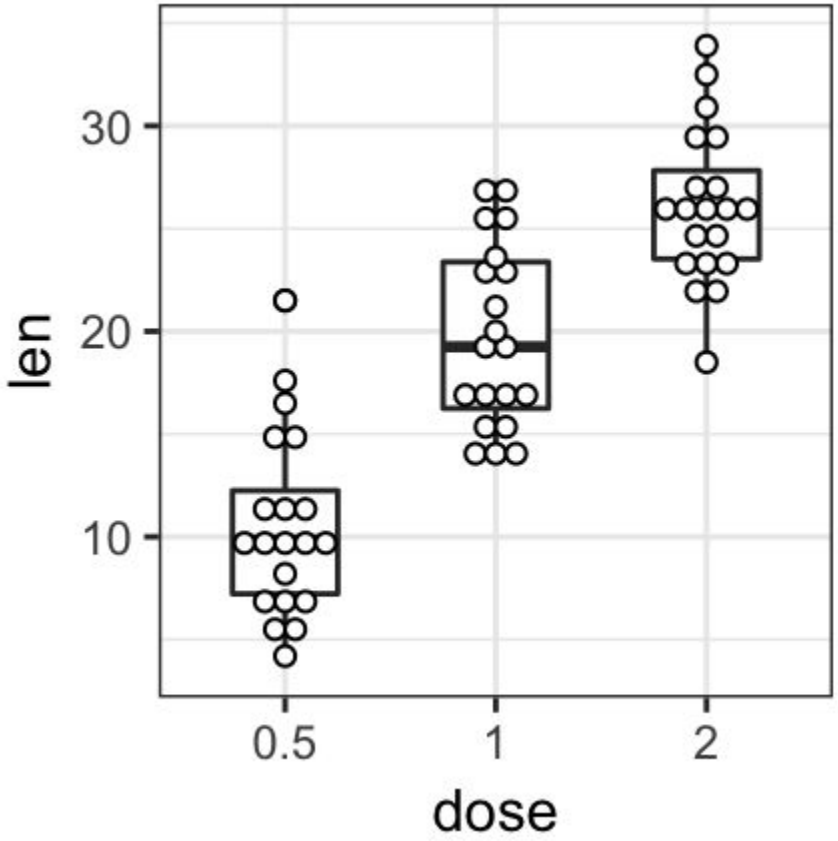
How many points are there?



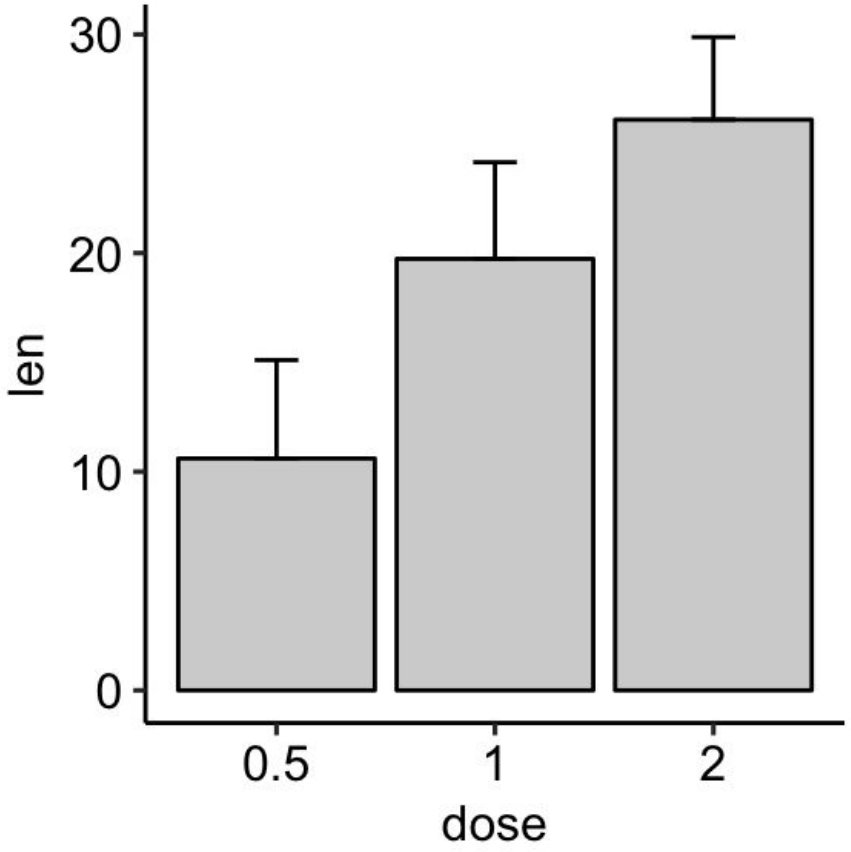
Combined with a violin plot,
Such that I can see the **kernel density**!

Ideally: show me the RAW data





Barplot?



Only if it makes sense!
(e.g. your values could start from 0!)

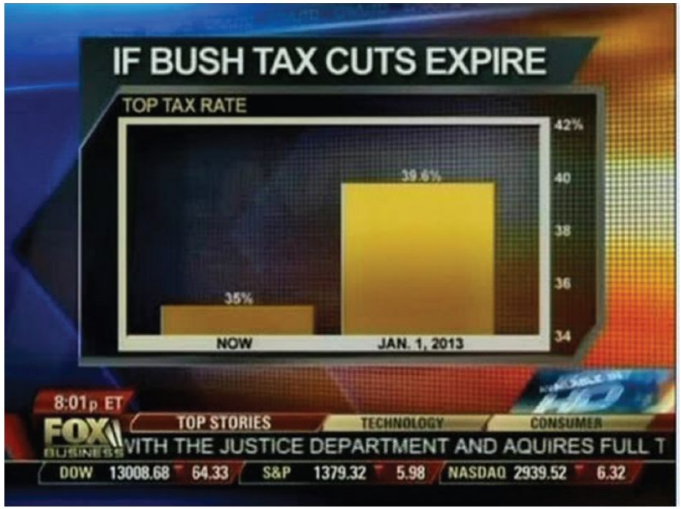


FIGURE 2.12 Fox News bar chart

Let's do this on an example

Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7	2·8	6·7	2·0
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3	2·7	4·9	1·8
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7	3·3	5·7	2·1
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2	6·0	1·8
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8	4·8	1·8
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0	4·9	1·8
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8	5·6	2·1
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

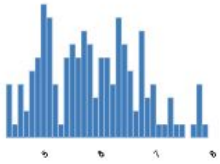
sepal length (cm)

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	35
Distinct (%)	23.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5.843333333

Minimum	4.3
Maximum	7.9
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	1.3 KiB



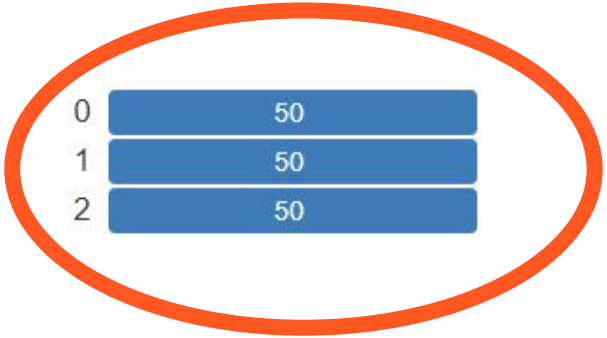
target

Categorical

HIGH CORRELATION

UNIFORM

Distinct	3
Distinct (%)	2.0%
Missing	0
Missing (%)	0.0%
Memory size	1.3 KiB



Classes are balanced!

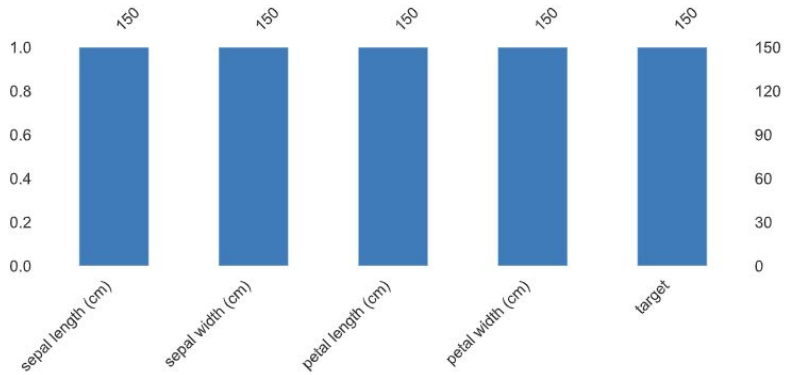
From: Learning from imbalanced data: open challenges and future directions

Application area	Problem description
Activity recognition [19]	Detection of rare or less-frequent activities (multi-class problem)
Behavior analysis [3]	Recognition of dangerous behavior (binary problem)
Cancer malignancy grading [30]	Analyzing the cancer severity (binary and multi-class problem)
Hyperspectral data analysis [50]	Classification of varying areas in multi-dimensional images (multi-class problem)
Industrial systems monitoring [44]	Fault detection in industrial machinery (binary problem)
Sentiment analysis [65]	Emotion and temper recognition in text (binary and multi-class problem)
Software defect prediction [48]	Recognition of errors in code blocks (binary problem)
Target detection [45]	Classification of specified targets appearing with varied frequency (multi-class problem)
Text mining [39]	Detecting relations in literature (binary problem)
Video mining [20]	Recognizing objects and actions in video sequences (binary and multi-class problem)

- Fraud Detection.
- Claim Prediction
- Default Prediction.
- Churn Prediction.
- Spam Detection.
- Anomaly Detection.
- Outlier Detection.
- Intrusion Detection
- Conversion Prediction.

https://machinelearningmastery.com/what-is-imbalanced-classification/

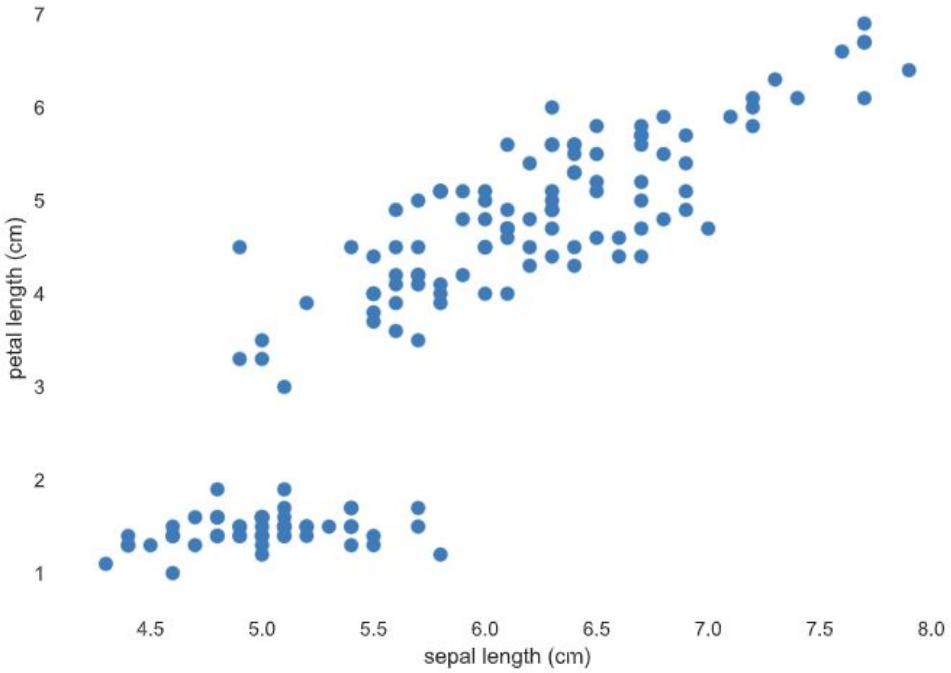
Anything missing or duplicate?



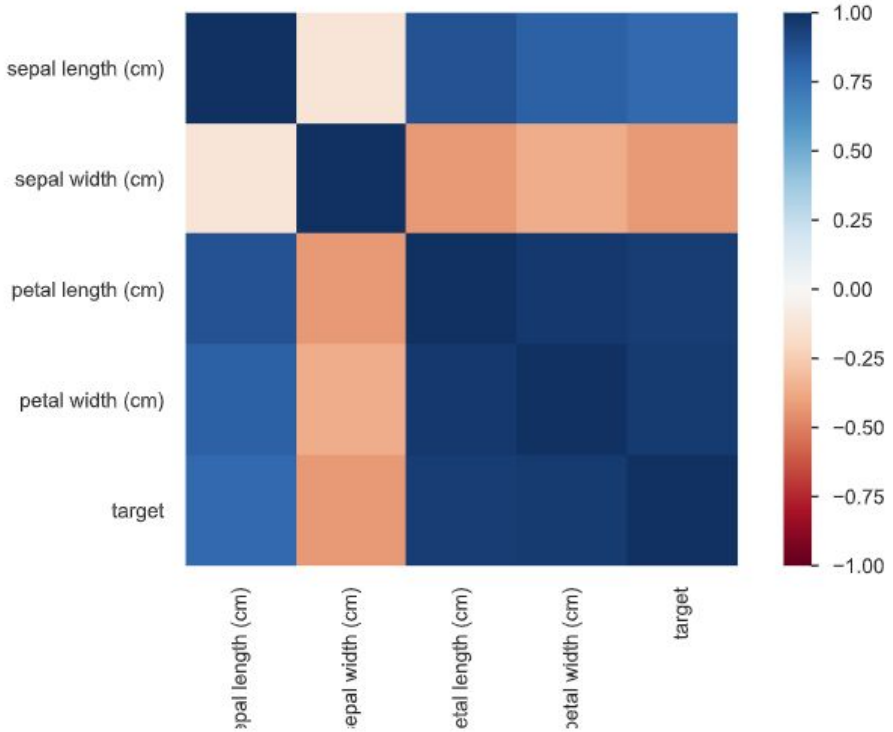
Most frequently occurring

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target	# duplicates
0	5.8	2.7	5.1	1.9	2	2

Correlation?!



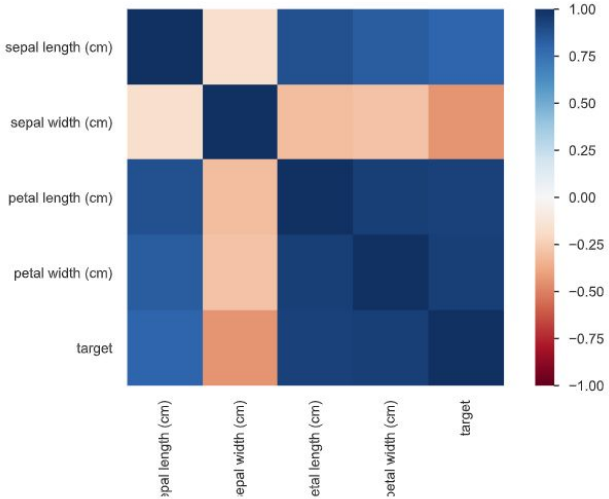
Multiple correlations at one glance



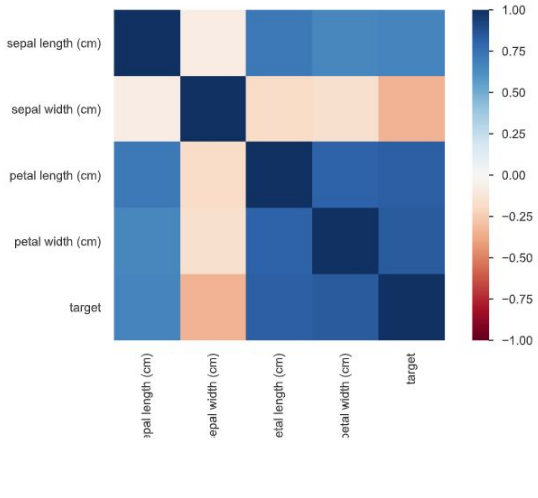
Pearson’s r across samples

All the others

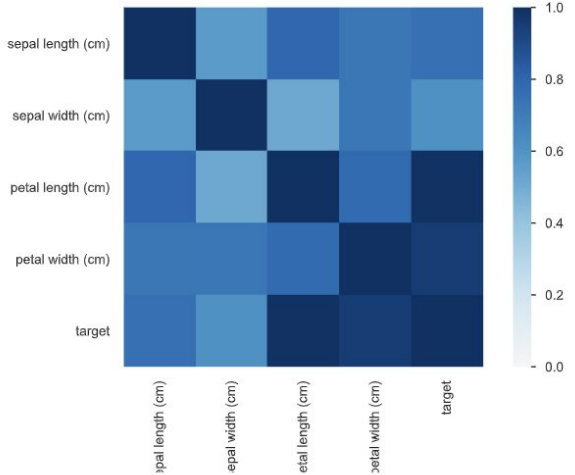
SPEARMAN



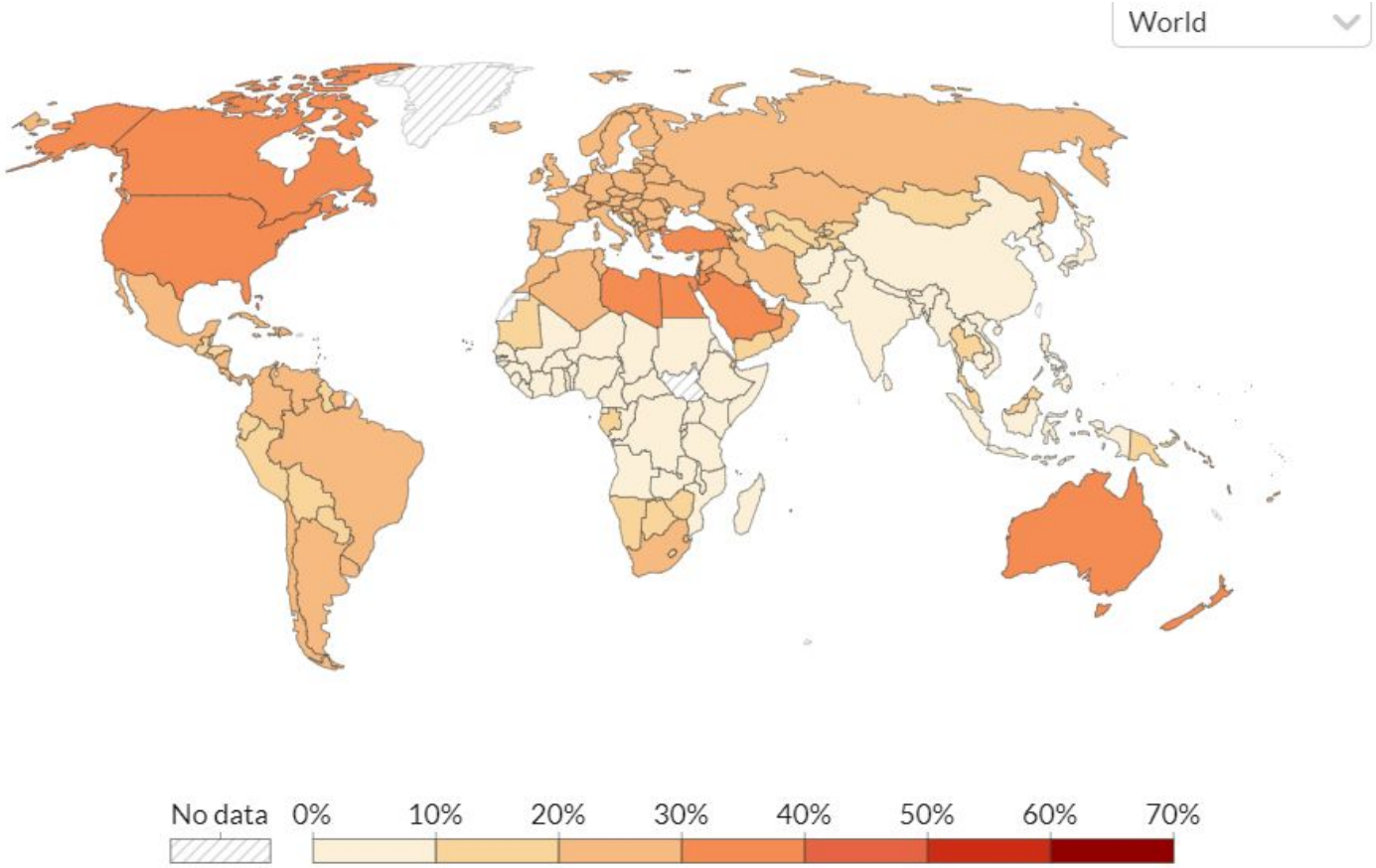
KENDALL



PHI-K



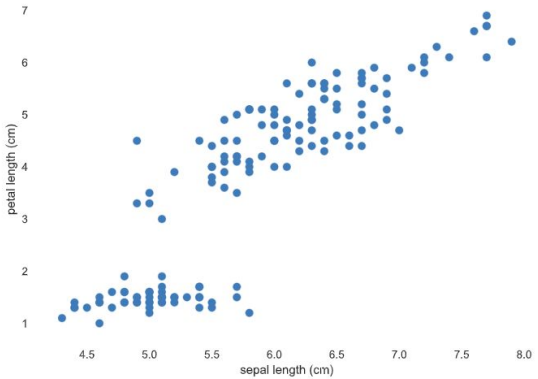
A heatmap

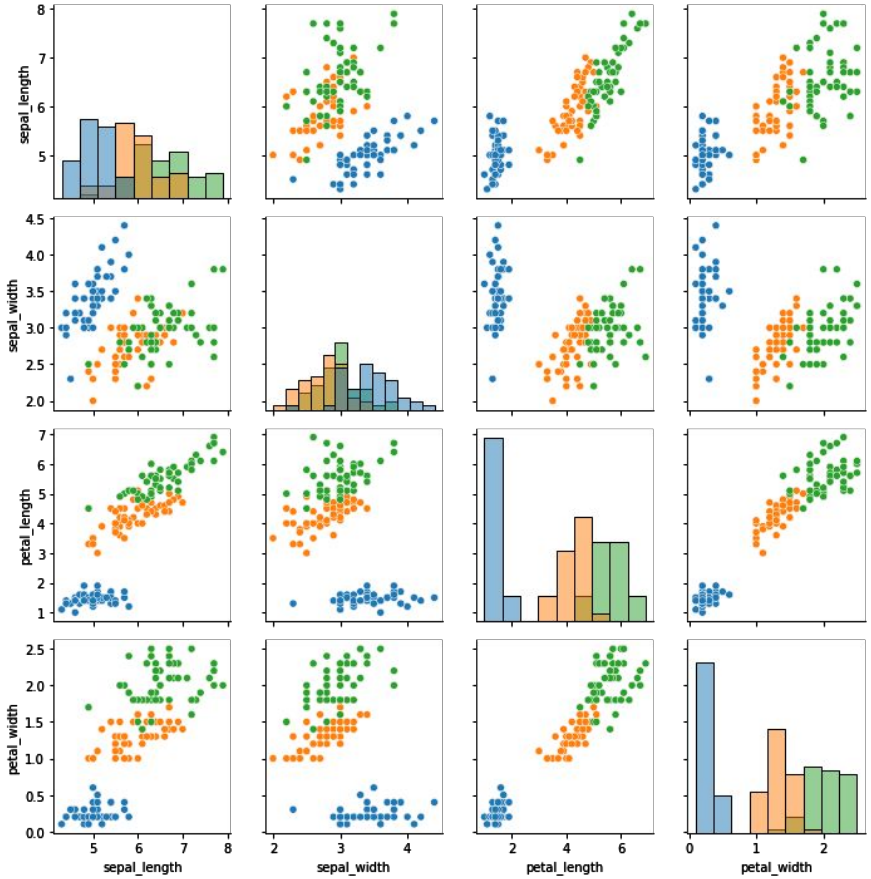


Share adults w/ obesity

Correlation?!

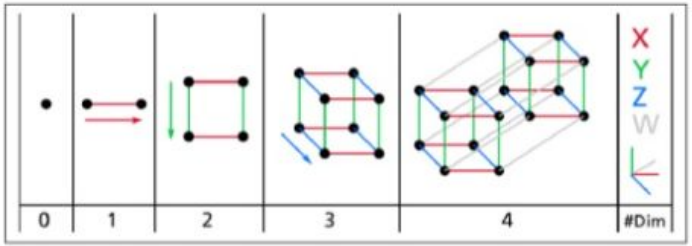
Only two dimensions at a time...





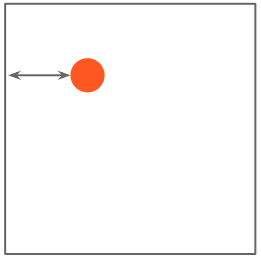
The curse of dimensionality

Many features....
⇒ MANY DIMENSIONS!!



Hard to imagine > 3 Dimensions!

Example: random point in unit square (1x1)



Chance < 0.001 to border? **0.4%**

Example: random point in 10,000 dim. hypercube?

99.99999 %

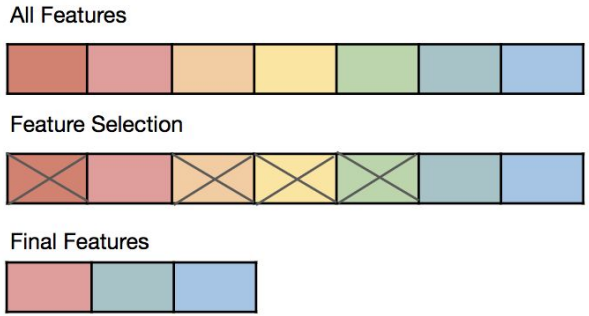
Second example: mean distance of two random points?

Unit square: **0.52**

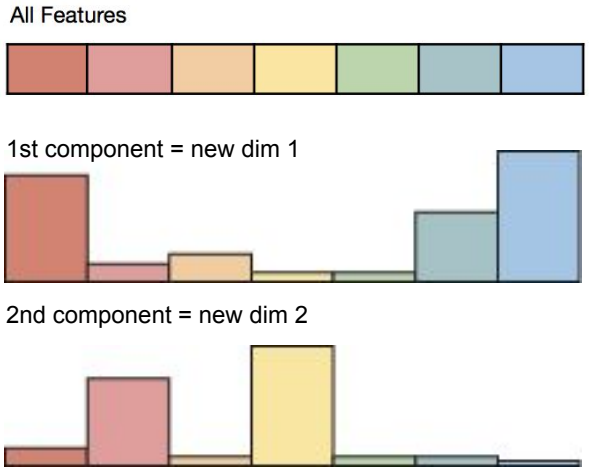
10,000 dim hypercube: **408.25!!!**

Dimensionality reduction

Feature selection



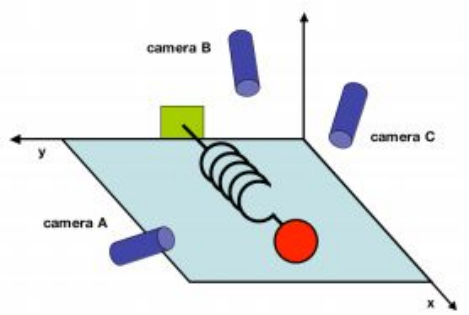
Feature projection



Principal component analysis

A tutorial on principal component analysis 2962 2014
J Shlens
arXiv preprint arXiv:1404.1100

Jon Shlens



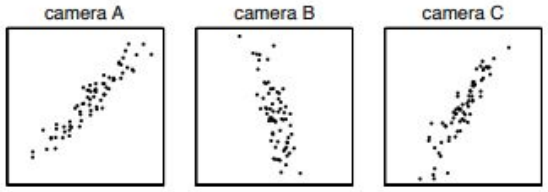
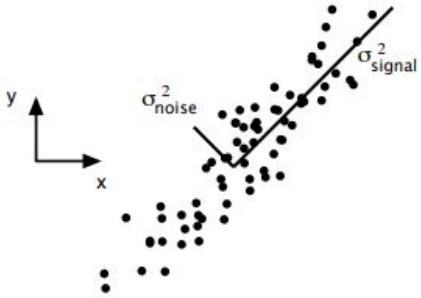


FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

$$\vec{X} = \begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \\ x_C \\ y_C \end{bmatrix}$$

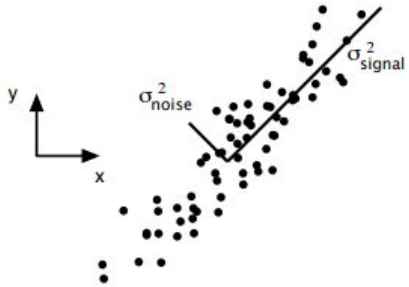
With this rigor we may now state more precisely what PCA asks: *Is there another basis, which is a linear combination of the original basis, that best re-expresses our data set?*



PCA

Via Eigenvectors of covariance matrix

$$\mathbf{C_X} \equiv \frac{1}{n}\mathbf{X}\mathbf{X}^T.$$



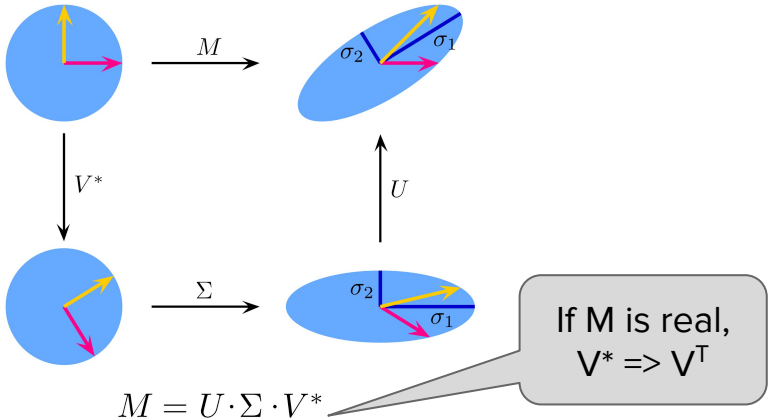
Via Singular Value Decomposition (SVD)

$$\mathbf{XV} = \mathbf{U\Sigma}$$

where each column of $\mathbf{V}$ and $\mathbf{U}$ perform the scalar version of the decomposition (Equation 3). Because $\mathbf{V}$ is orthogonal, we can multiply both sides by $\mathbf{V}^{-1} = \mathbf{V}^T$ to arrive at the final form of the decomposition.

$$\mathbf{X} = \mathbf{U\Sigma V}^T \quad (4)$$

Although derived without motivation, this decomposition is quite powerful. Equation 4 states that *any* arbitrary matrix $\mathbf{X}$ can be converted to an orthogonal matrix, a diagonal matrix and another orthogonal matrix (or a rotation, a stretch and a second rotation). Making sense of Equation 4 is the subject of the next section.



PCA

E. Summary of Assumptions

This section provides a summary of the assumptions behind PCA and hint at when these assumptions might perform poorly.

I. *Linearity*
Linearity frames the problem as a change of basis. Several areas of research have explored how extending these notions to nonlinear regimes (see Discussion).

II. *Large variances have important structure.*
This assumption also encompasses the belief that the data has a high SNR. Hence, principal components with larger associated variances represent interesting structure, while those with lower variances represent noise. Note that this is a strong, and sometimes, incorrect assumption (see Discussion).

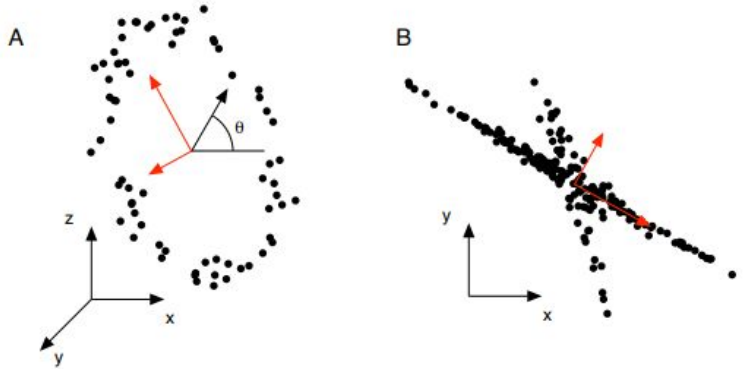
III. *The principal components are orthogonal.*
This assumption provides an intuitive simplification that makes PCA soluble with linear algebra decomposition techniques. These techniques are highlighted in the two following sections.

Quick Summary of PCA

1. Organize data as an $m \times n$ matrix, where m is the number of measurement types and n is the number of samples.
2. Subtract off the mean for each measurement type.
3. Calculate the SVD or the eigenvectors of the covariance.

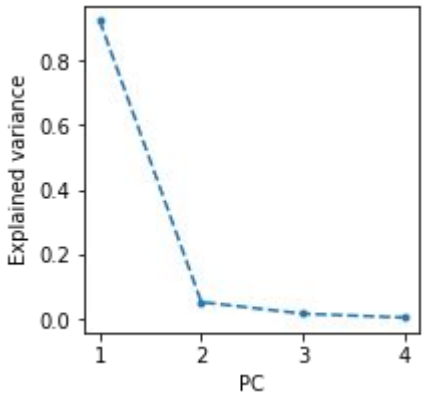
z-scored

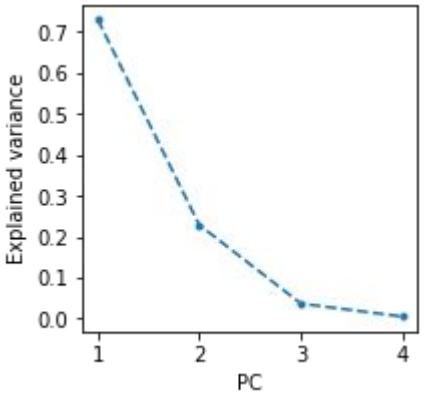
FIG. 5 A step-by-step instruction list on how to perform principal component analysis

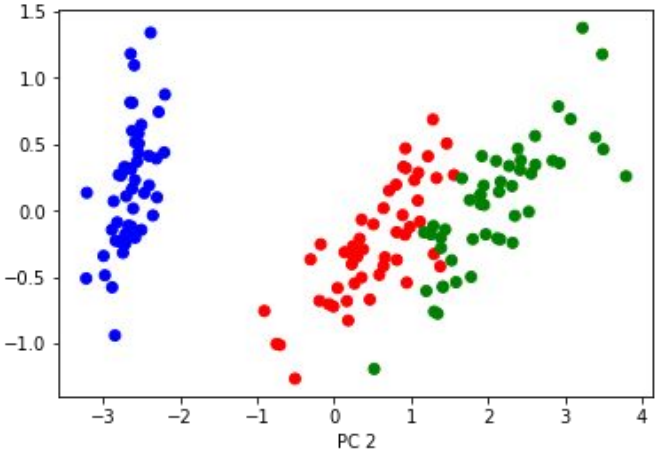


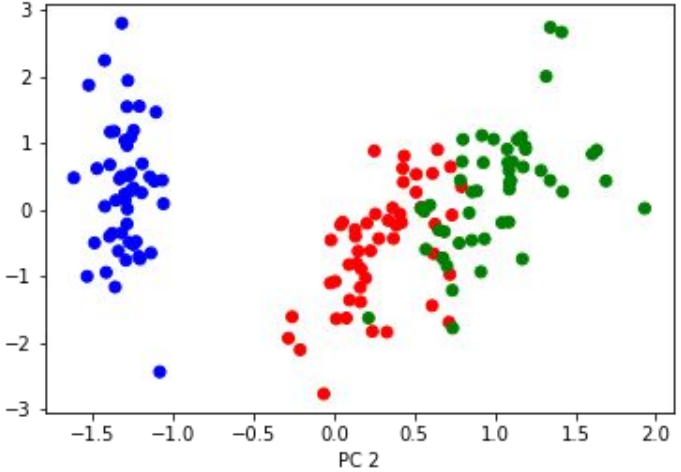
PCA

Z-scoring WHITEN YOUR DATA!



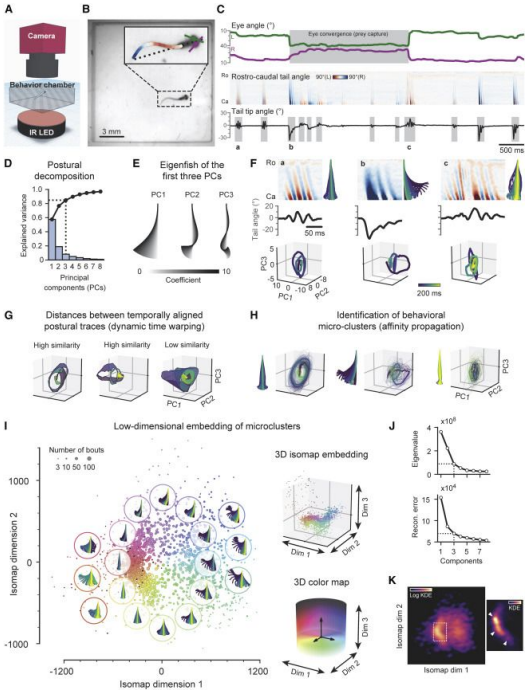




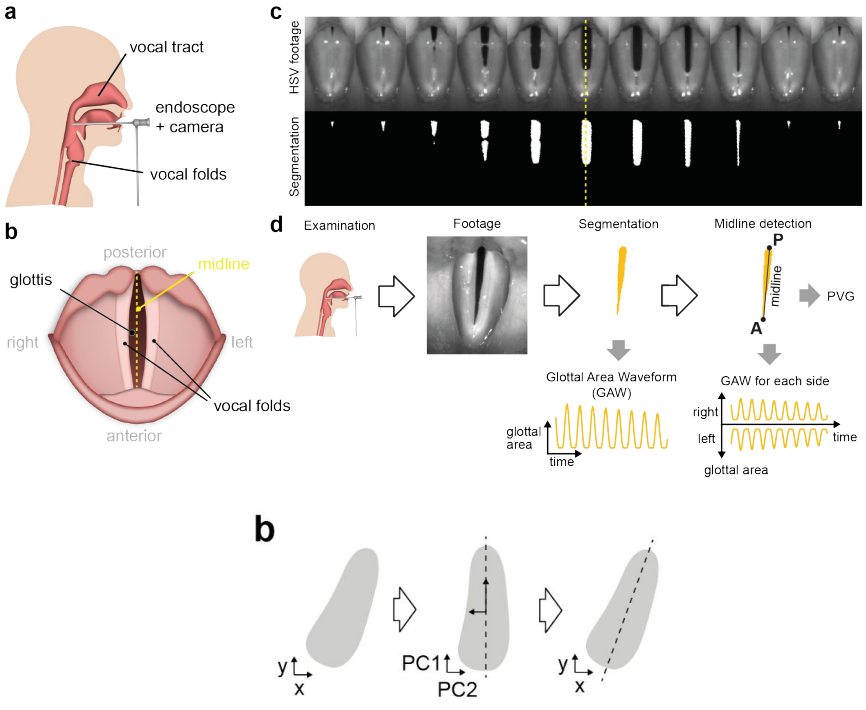


Try exploring this on the WINE dataset!!

Ok, cool, but where do I need it?



Behavioral low-dim embedding
Mearns et al., Curr Biol 2020

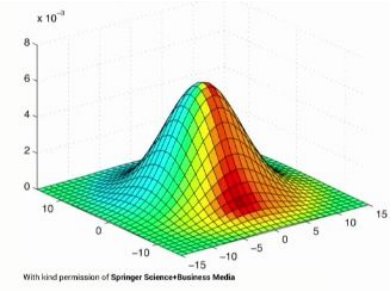


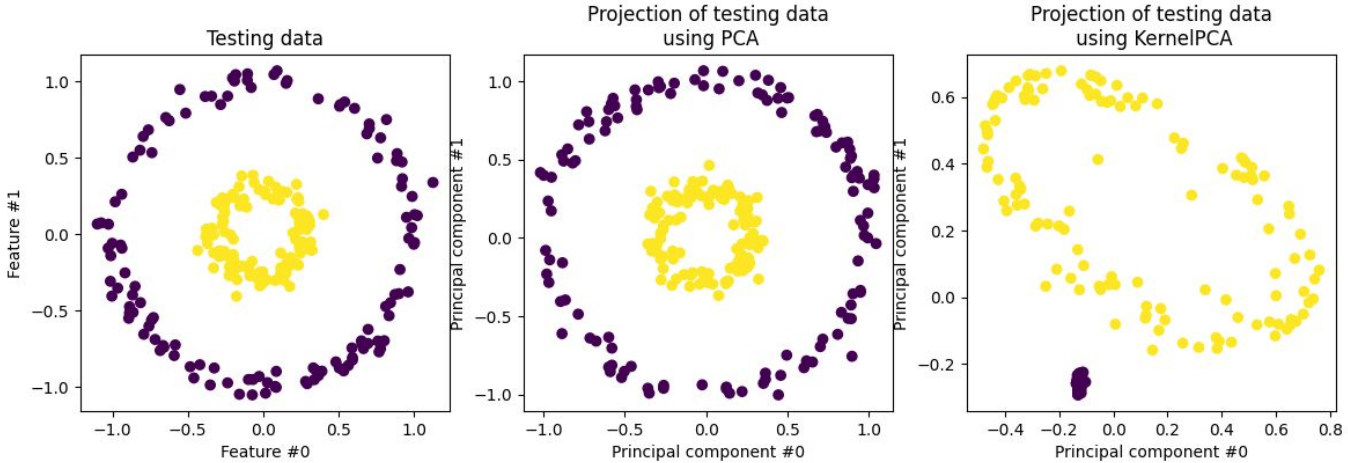
Glottal midline axis computation
Kist et al., Sci Rep 2020

Kernel trick -> Kernel PCA

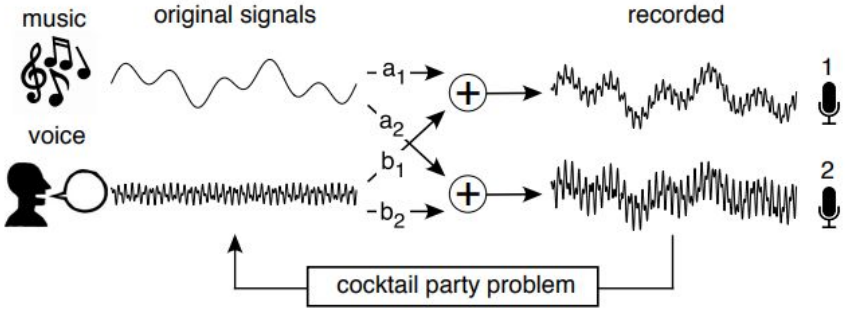
Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models

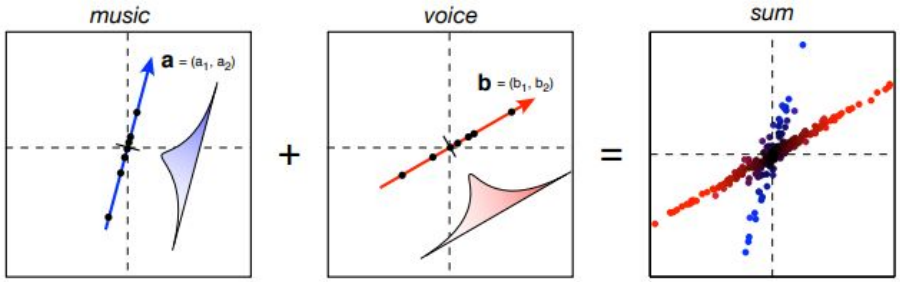
Quan Wang WANGQ10@RPI.EDU
Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180 USA





Independent component analysis (ICA)

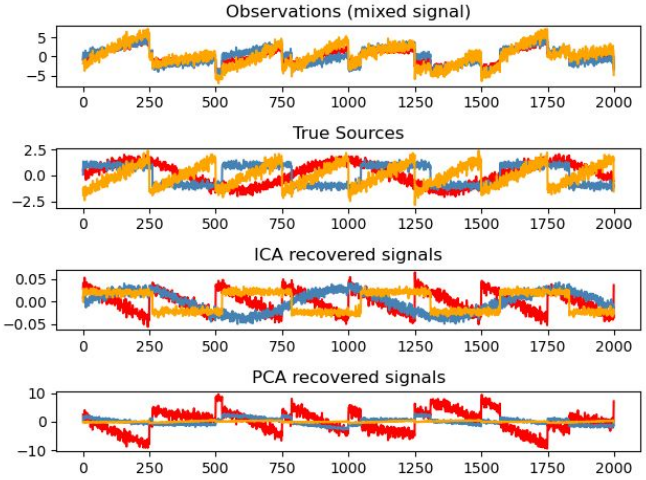




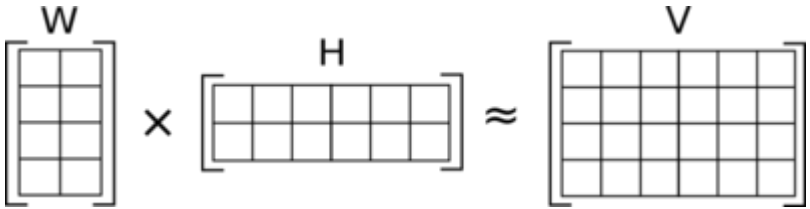
Quick Summary of ICA

1. Subtract off the mean of the data in each dimension.
2. Whiten the data by calculating the eigenvectors of the covariance of the data.
3. Identify final rotation matrix that optimizes statistical independence (Equation 6).

FIG. 9 Summary of steps behind ICA. The first two steps have analytical solutions but the final step must be optimized numerically. See Appendix B for example code.



Non-negative matrix factorization (NMF)



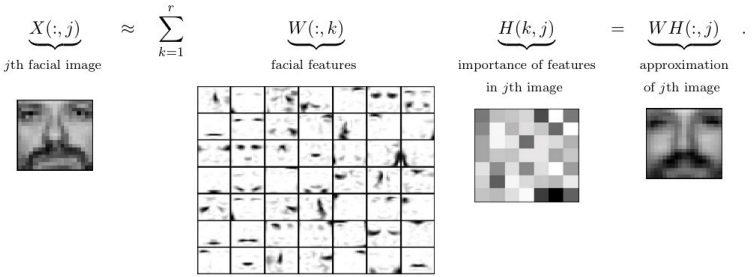


Figure 1: Decomposition of the CBCL face database, MIT Center For Biological and Computation Learning (2429 gray-level 19-by-19 pixels images) using $r = 49$ as in [79].

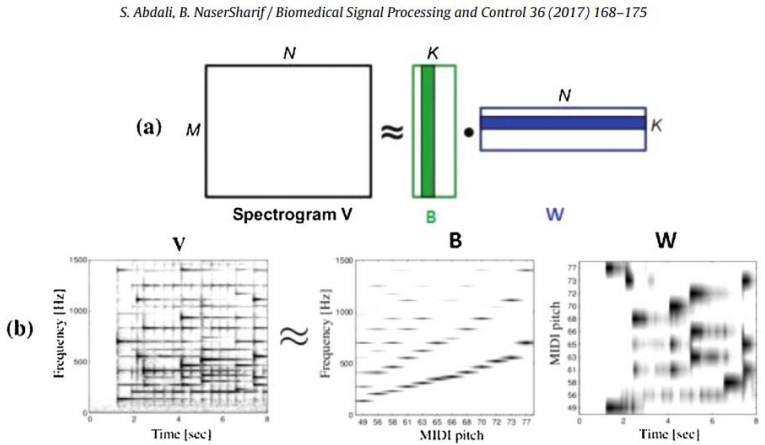


Fig. 1. Non-negative matrix factorization.

HOMEWORK

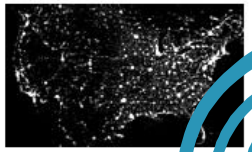
In this lecture, we covered data exploration and visualization. In the homework you will work with the Census Income Dataset.

You will perform data exploration and create adequate plots to visualize the data.

Census Income Data Set

Download: Data Folder, Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Adult" dataset.



Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	722141