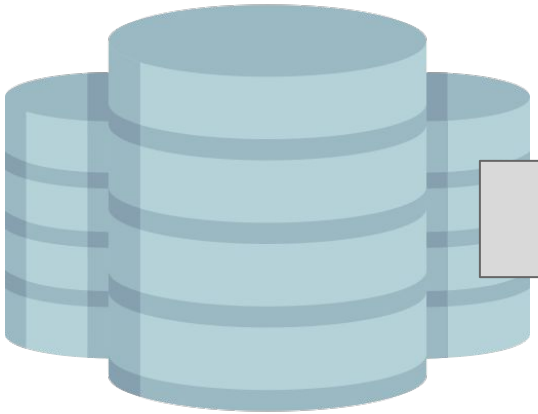
Data Science Survival Skills

Data exploration and visualization



DATA VISUALIZATION

Importance of context

First part

Exploratory

Showing all your data

Explanatory

Second part

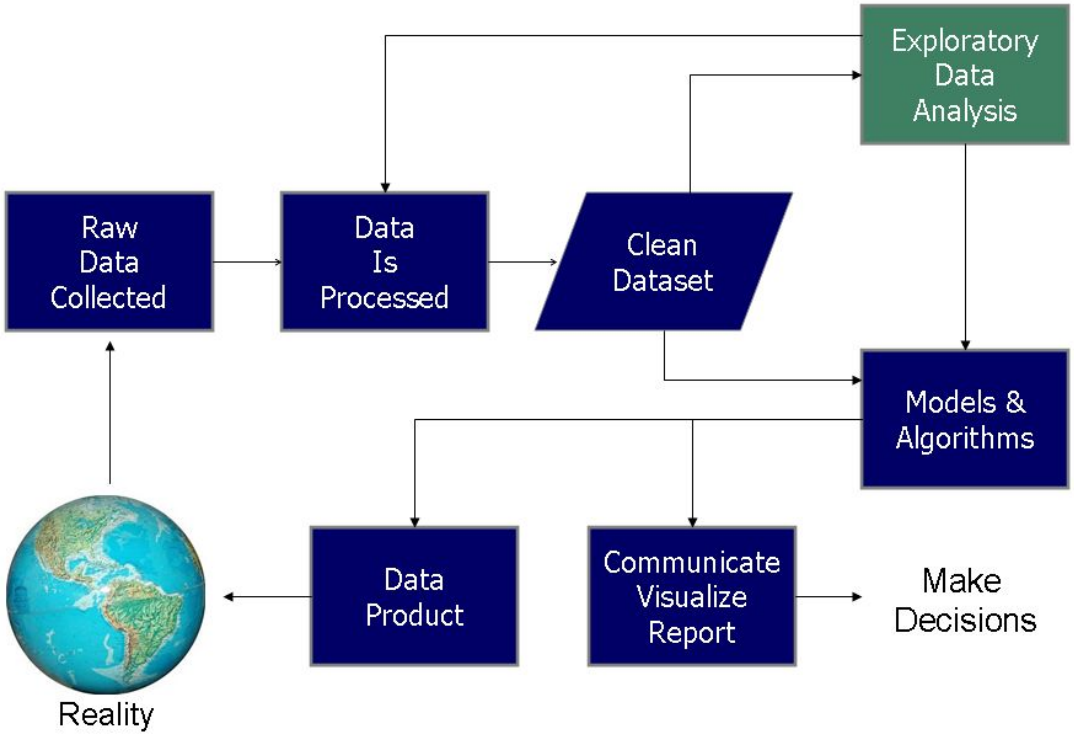
Showing only the relevant data

WHO?

WHAT?

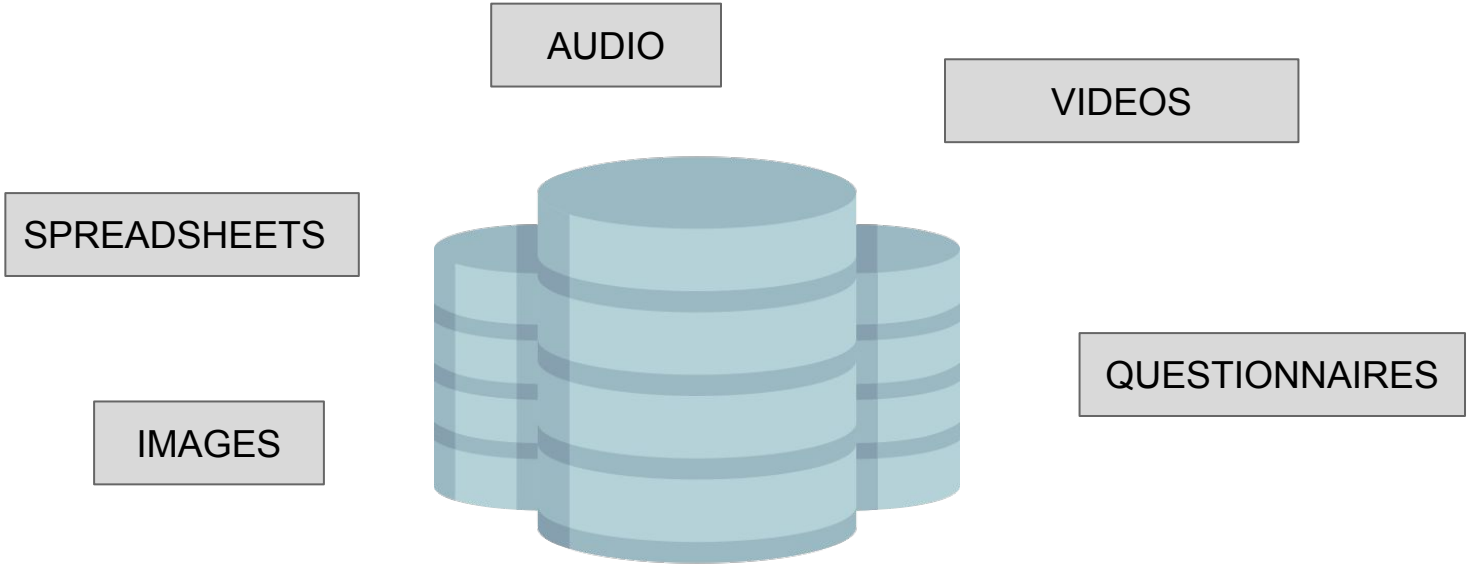
HOW?

Data Science Process

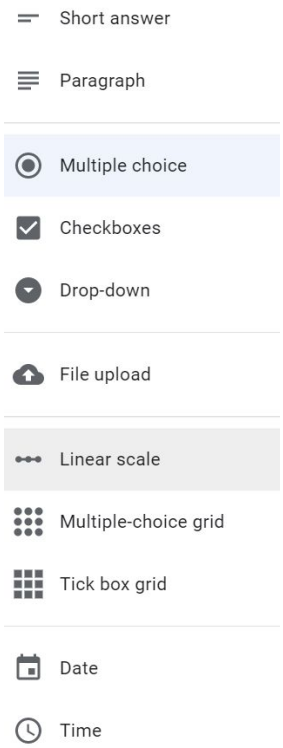


Doing Data Science, by Schutt & O'Neil (2013)

(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

Qualitative values

Categorical (or nominal)

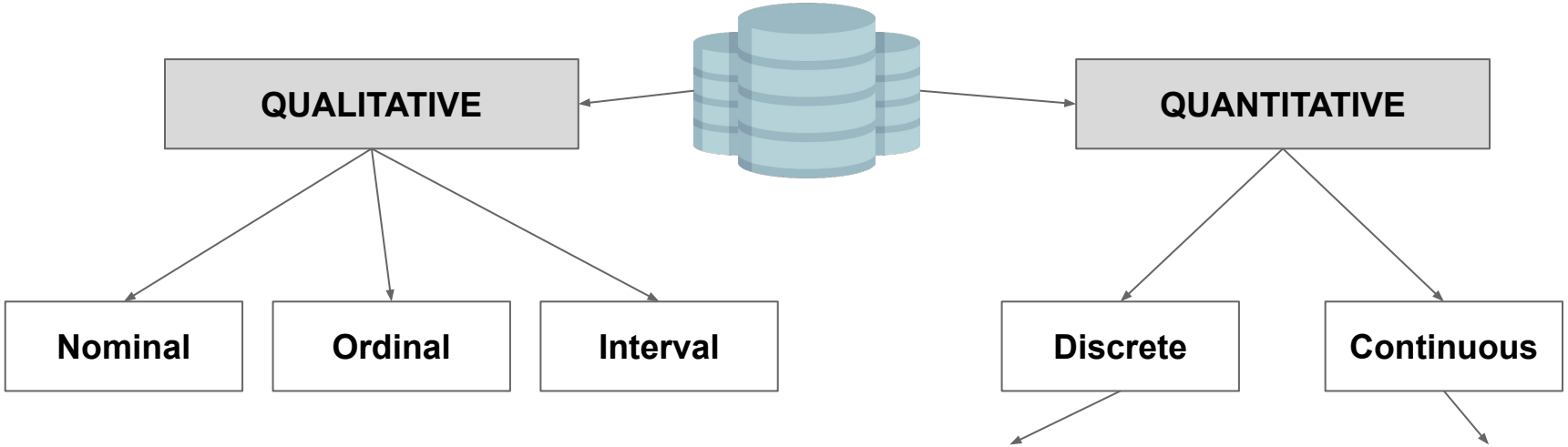
Smoker yes/no Hair color (blonde, brown, black, ...)
At least binary, no order implied

Ordinal

Income (low/middle/high) Tumor grading (I, II, III, IV, ...)
Education level (elementary, high school, college, university)
⇒ rank has order, but spacing might be not consistent

Interval/Numerical

Income 5,000-10,000 USD
10,000-15,000 USD,
⇒ equal spacing!



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 fells 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

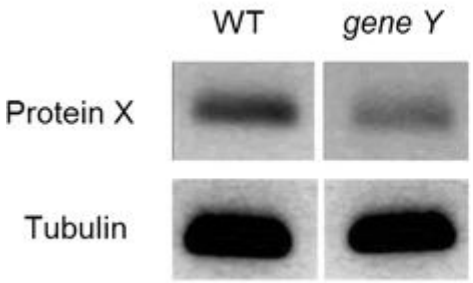
Iris Versicolor **Iris Setosa** **Iris Virginica**

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

Feature extraction

Quantifying unstructured data

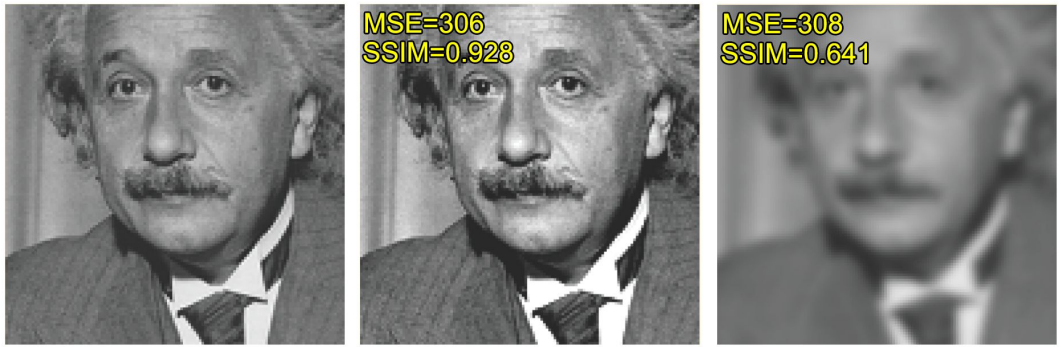
Western blot to quantify protein expression



Protein of interest

Reference

Bell, BMC Biology 2016

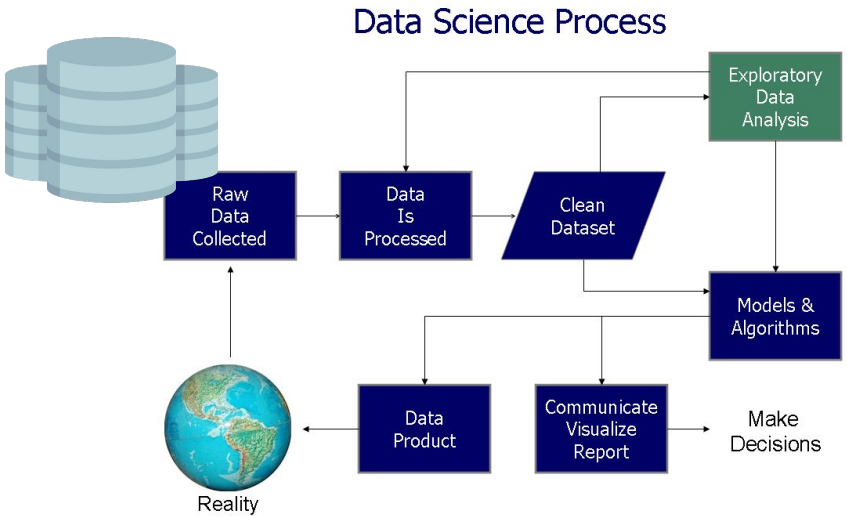


https://videoprocessing.ai/metrics/ways-of-cheating-on-popular-objective-metrics.html

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%

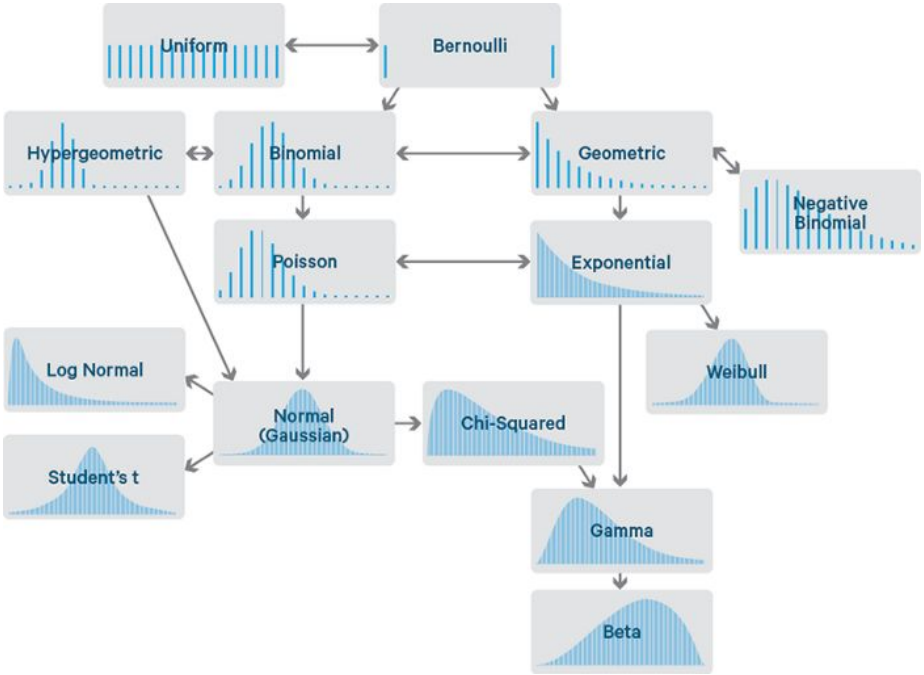
Panel C: Descriptive statistics of continuous variables by audit firm size

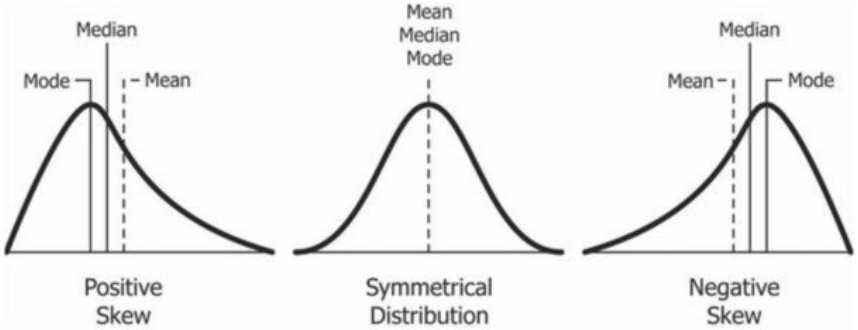
Variable	Big Four (N = 69)					Non-Big Four (N = 186)				
	Mean	SD	Min	P50	Max	Mean	SD	Min	P50	Max
DACKO	0.0791	0.0856	0.0007	0.0522	0.4495	0.0787	0.0811	0.0004	0.0543	0.527
AFEE	4.3365	0.4605	3.699	4.0792	5.2722	3.8804	0.2093	3	3.8891	4.2553
LEVERG	41.212	22.865	6.639	39.839	99.815	30.886	22.237	0.3998	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

Distributions





Determine distributions,
Check transformations (log-scale, etc)

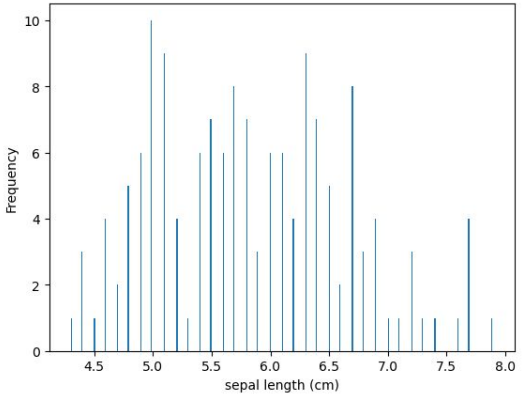
https://jagan-singhh.medium.com/types-of-probability-distributions-9333d18ed817

Histogram

Table I

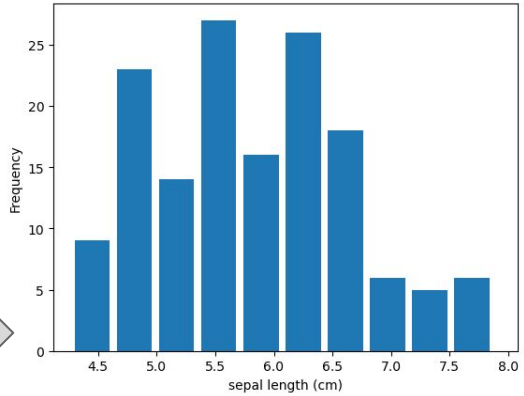
Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9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

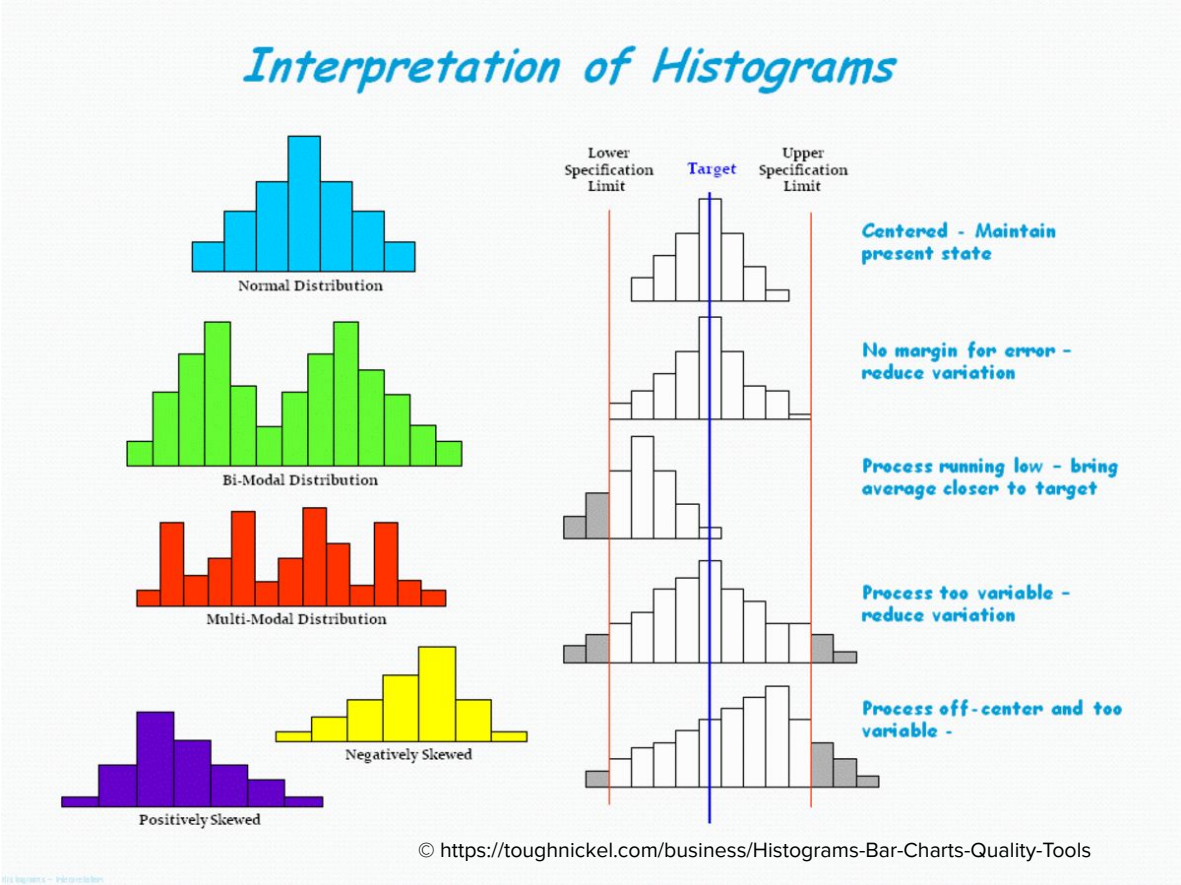


Karl Pearson

Adjust bins to allow intervals



Histogram interpretations



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

Further things to check for...

Common method bias

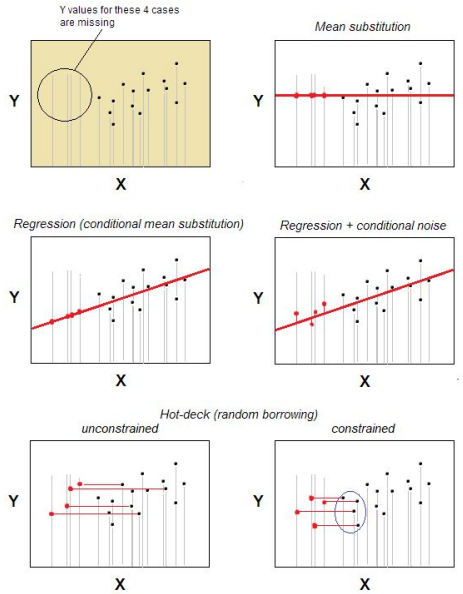
⇒ Is there more variability due to the measurement than in the effect size?!

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 deck
➜ changes with more data is processed
Cold deck
➜ stays constant
(imputation due to neighbour similarity)



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

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

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

Here N is equal to the number of items, $\bar{c}$ is the average inter-item covariance among the items and $\bar{v}$ equals the average variance.

Respondent	Item 1	Item 2	Item 3	Item 4	Item 5
1	3	4	2	5	3
2	4	3	4	4	2
3	2	2	3	3	4
4	5	4	5	5	5
5	3	3	2	4	3
6	4	5	4	4	4
7	3	3	3	3	3
8	2	2	1	2	2
9	4	4	4	4	4
10	3	3	2	3	3

$\bar{c}$

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?

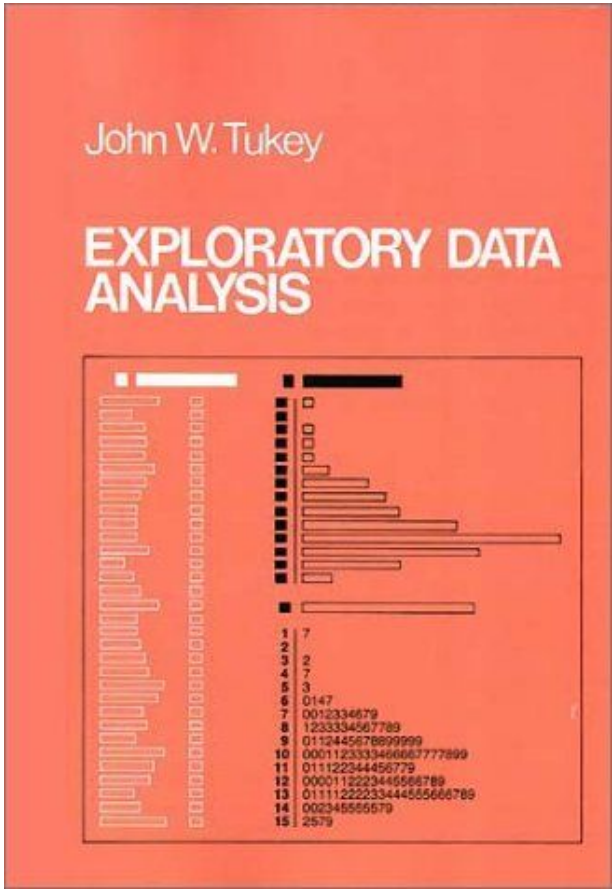
Hypothesis

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

Outliers

Robust analysis techniques? RANSAC methods?

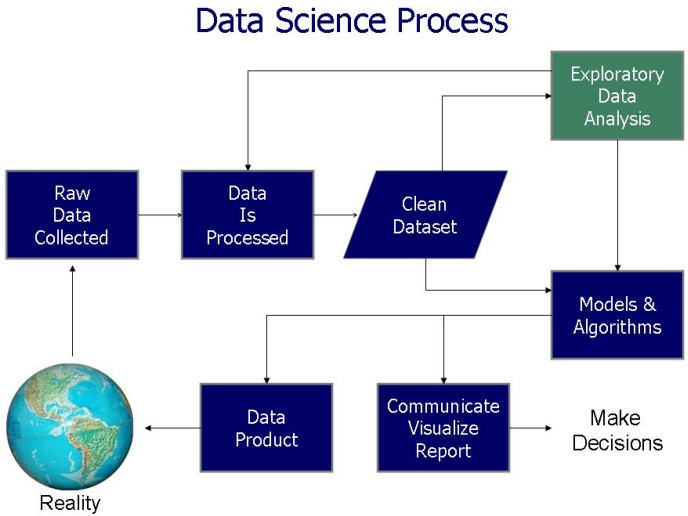
Exploratory Data Analysis (EDA)



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)

Pandas Profiling

- **Type inference**: detect the types of columns in a `DataFrame`
- **Essentials**: type, unique values, indication of missing values
- **Quantile statistics**: minimum value, Q1, median, Q3, maximum, range, interquartile range
- **Descriptive statistics**: mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
- **Most frequent and extreme values**
- **Histograms:** categorical and numerical
- **Correlations**: high correlation warnings, based on different correlation metrics (Spearman, Pearson, Kendall, Cramér's V, Phik, Auto)
- **Missing values**: through counts, matrix, heatmap and dendrograms
- **Duplicate rows**: list of the most common duplicated rows
- **Text analysis**: most common categories (uppercase, lowercase, separator), scripts (Latin, Cyrillic) and blocks (ASCII, Cyrilic)
- **File and Image analysis**: file sizes, creation dates, dimensions, indication of truncated images and existence of EXIF metadata

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

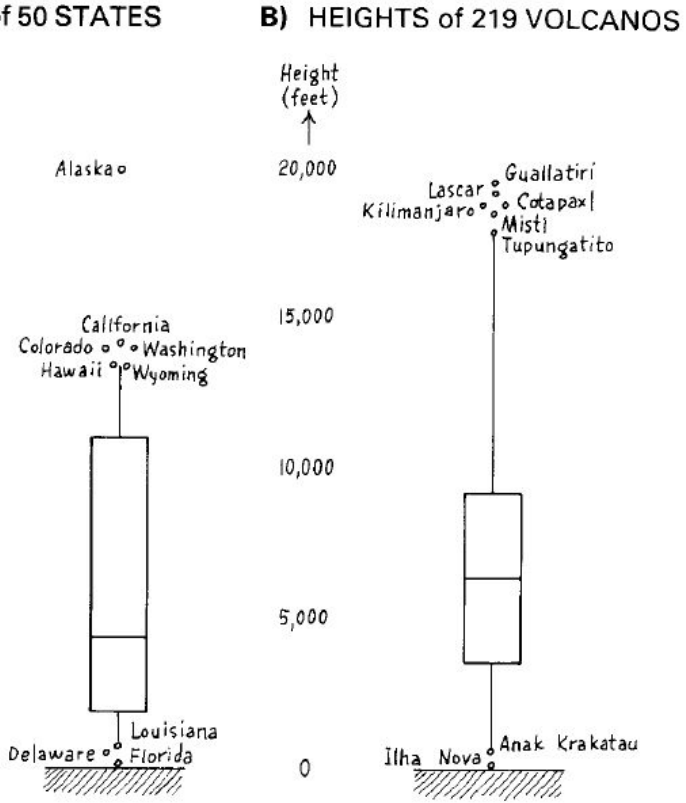
N = the size of the population

x_i = each value from the population

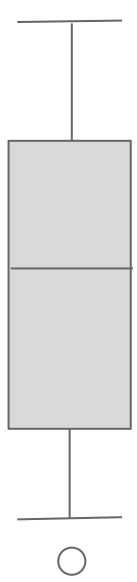
μ = the population mean

Visualization of quantiles: box-and-whisker

Box-and-whisker plots with end values identified



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



Max, but max. 1.5 * IQR

Q3

Q2

Q1

Min, but max. 1.5 * IQR

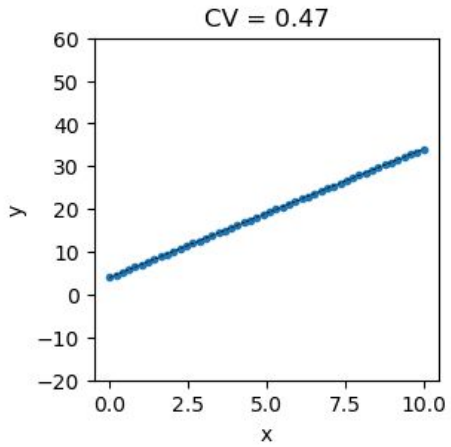
Outlier

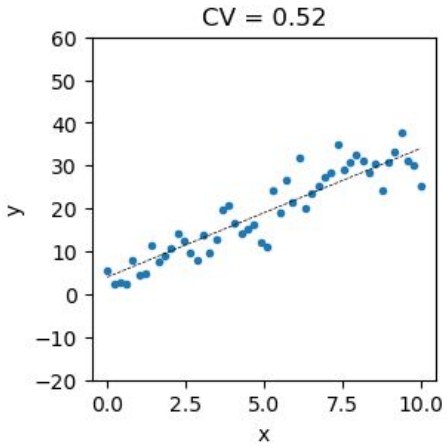
Make a box-and-whisker plot from DataFrame columns, optionally grouped by some other columns. A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers 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.

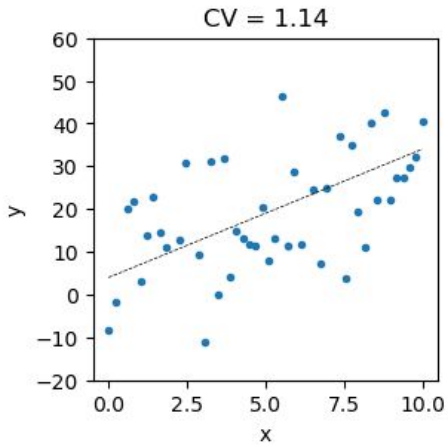
Coefficient of Variation

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

How much does the data vary?







Correlation vs. causation

Correlation

Pearson's correlation coefficient

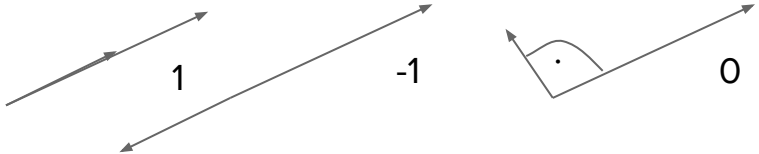
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

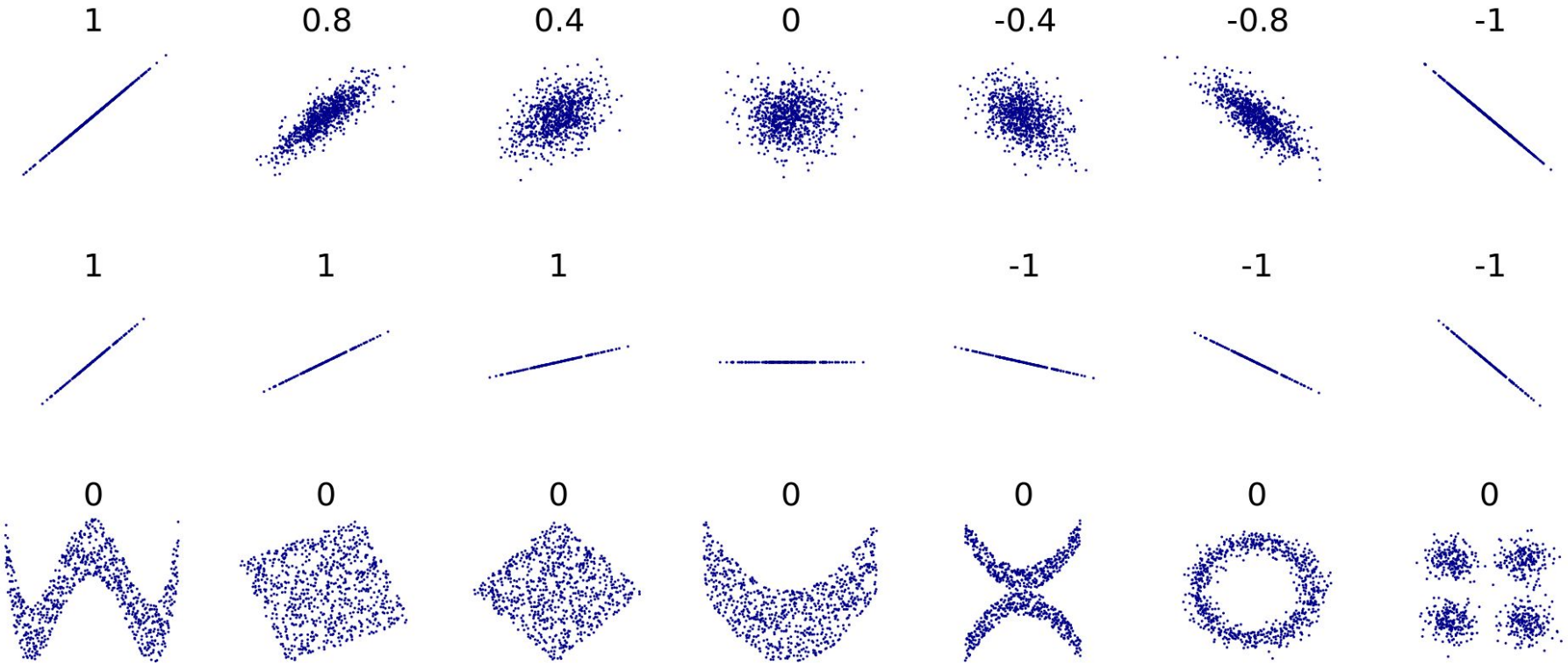
Linear regression:
How Y varies dependent on X
(Y = a + b*X)

Uncentered correlation coefficient

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



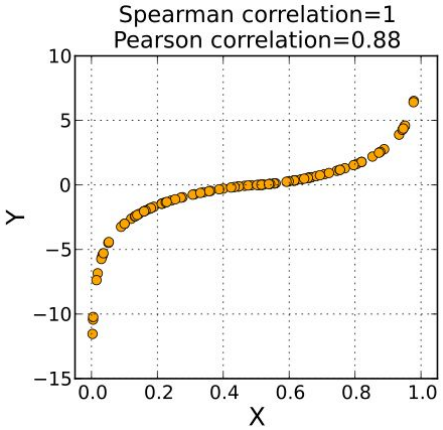
Correlation

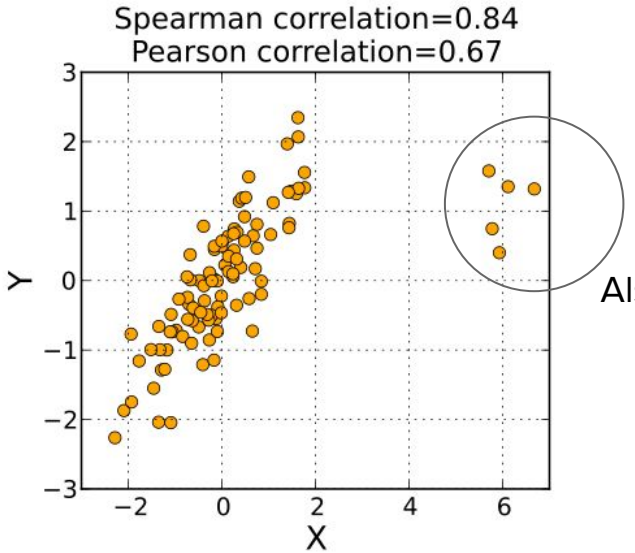


Other correlation metrics

Spearman's correlation coefficient
Looks at the RANK of the data!
➜ also is able to capture NONLINEAR relationships!
Only cares if data follows a MONOTONIC function

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

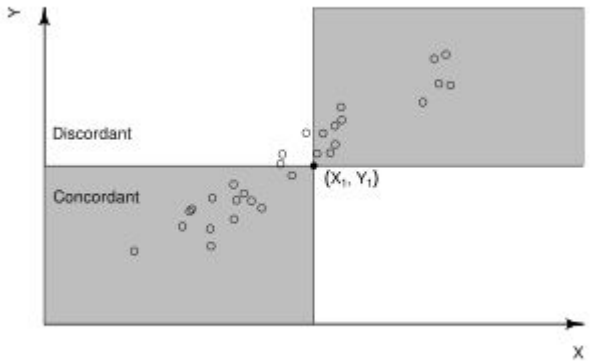




Also robust to outliers

Kendall Rank Correlation

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})} = 1 - \frac{2(\text{number of discordant pairs})}{\binom{n}{2}}.$$



Similar to Spearman's correlation,
But preferable when tied ranks or very small dataset.

Normally used: tau-b

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, The Netherlands
[c]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Pearson ➔ Cramer's phi

CATEGORICAL and BINNED DATA!

Most comparable to this work is Cramér's ϕ [20], a correlation coefficient meant for two categorical variables, denoted as ϕ_C, based on Pearson's χ^2 test statistic, and with values between 0 (no association) and +1 (complete association):

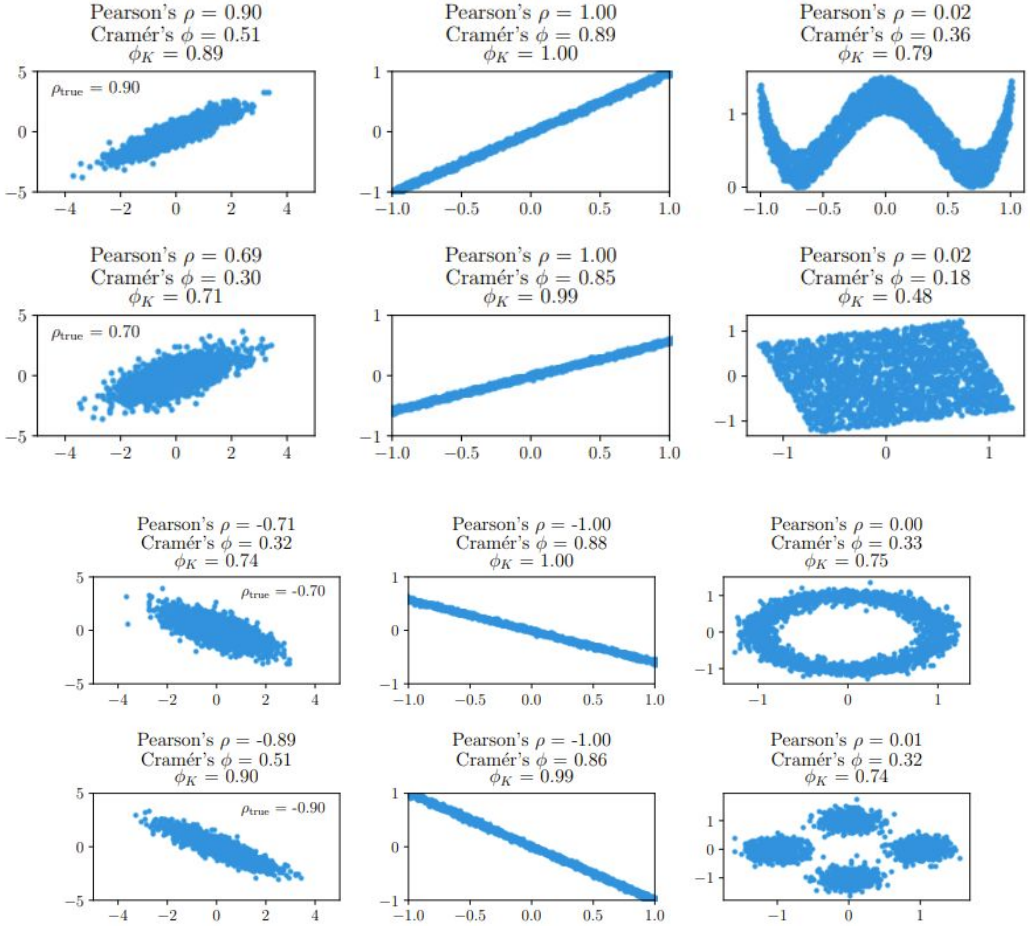
$$\phi_C = \sqrt{\frac{\chi^2}{N \min(r-1, k-1)}}, \tag{5}$$

4 Definition of ϕ_K

The correlation coefficient ϕ_K is obtained by inverting the χ^2 contingency test statistic through the steps outlined below. Although the procedure can be extended to more variables, the method is described with two variables for simplicity.

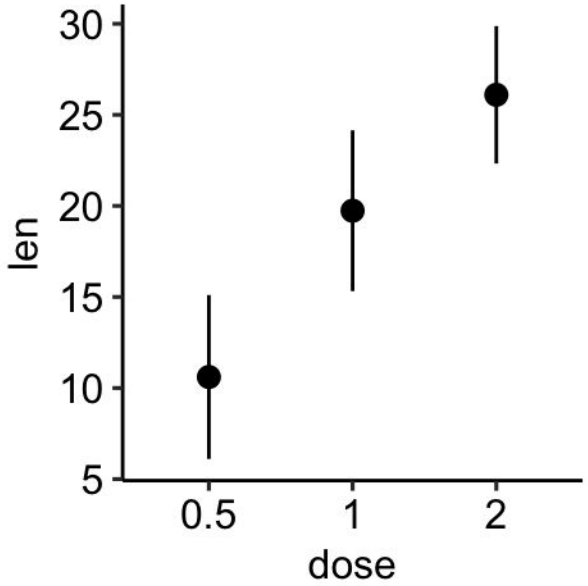
Works on categorical, ordinal, binned and continuous data!

Phik (ϕk)



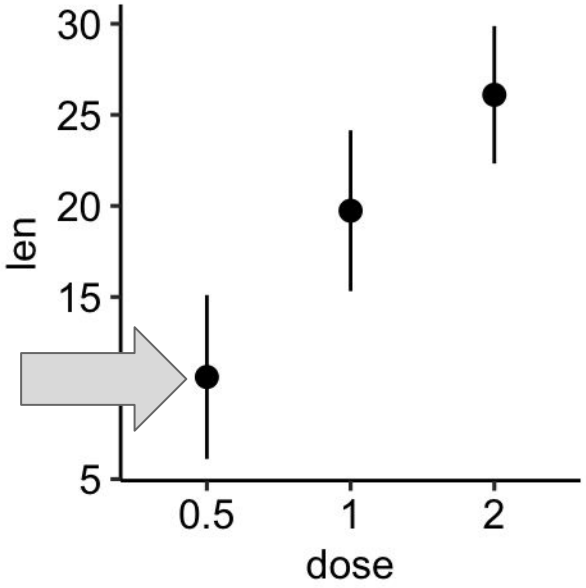
Shows if structure is present, but not its direction!

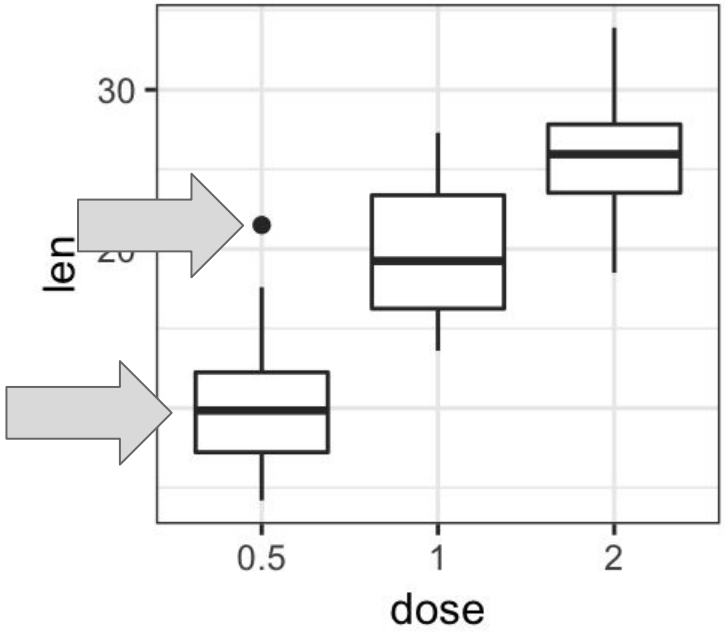
Plotting results



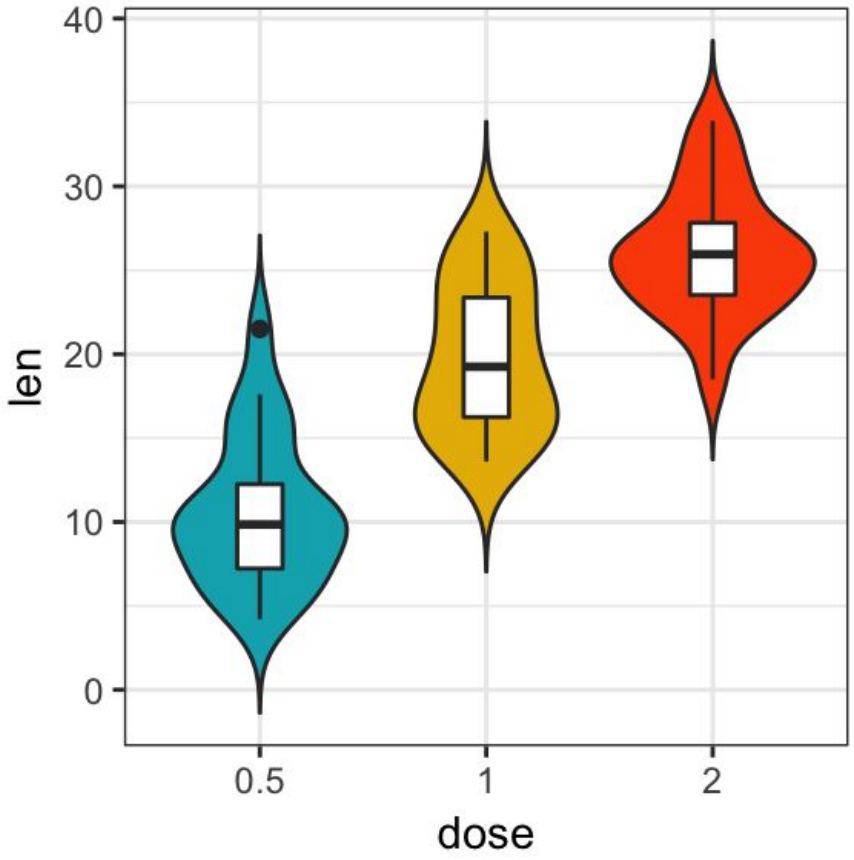
Where is the data?
Is it really well represented as MEAN +- STD?

Plotting results - boxplot?



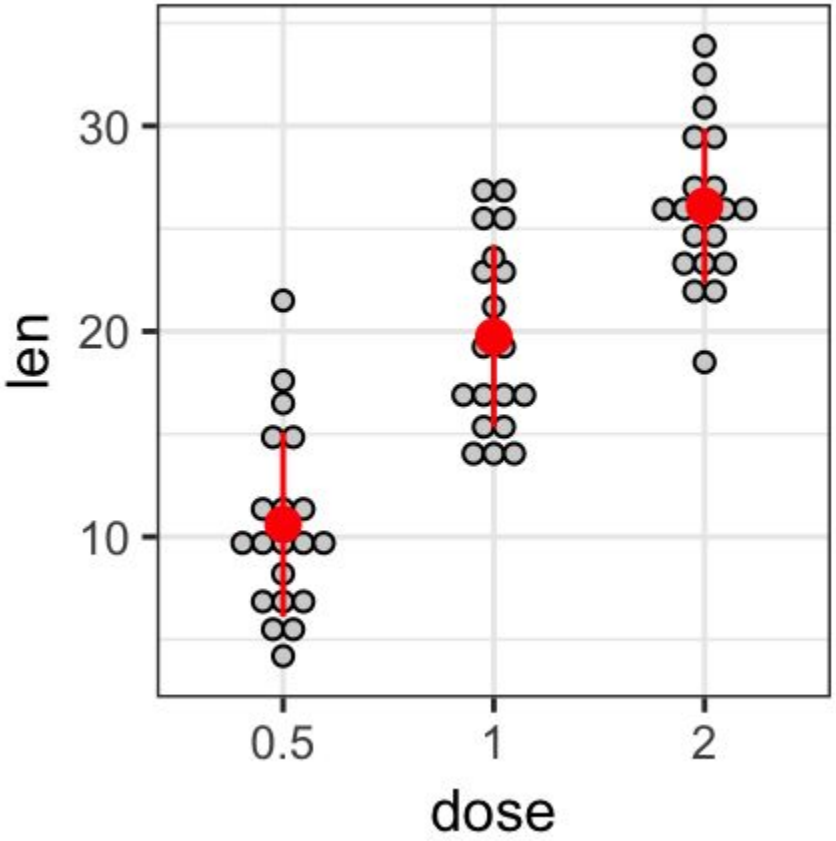


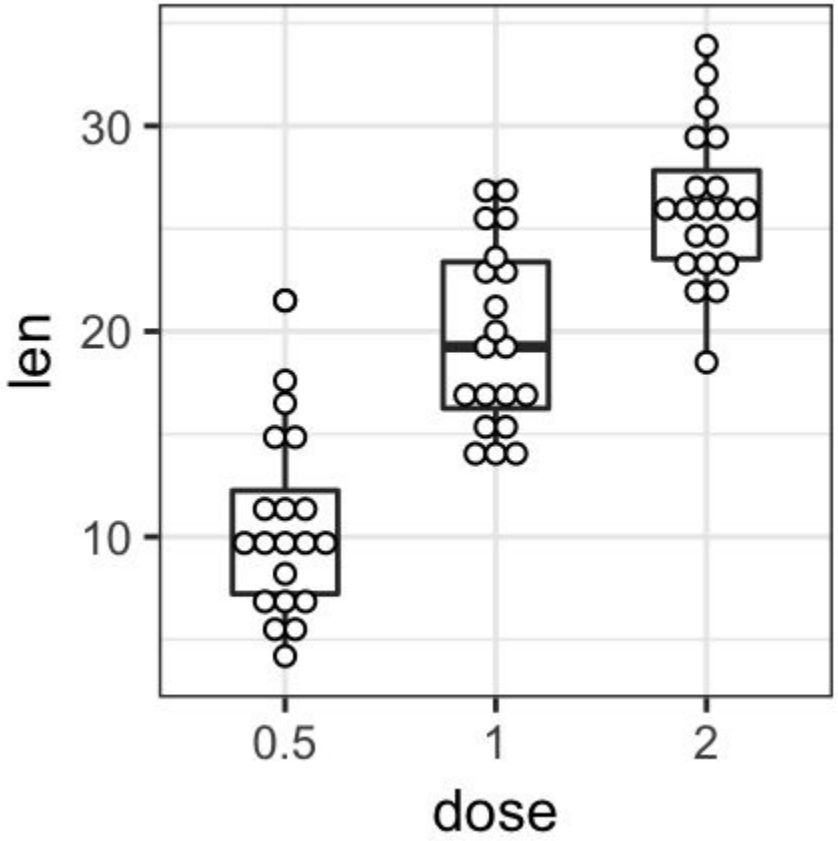
How many points are there?



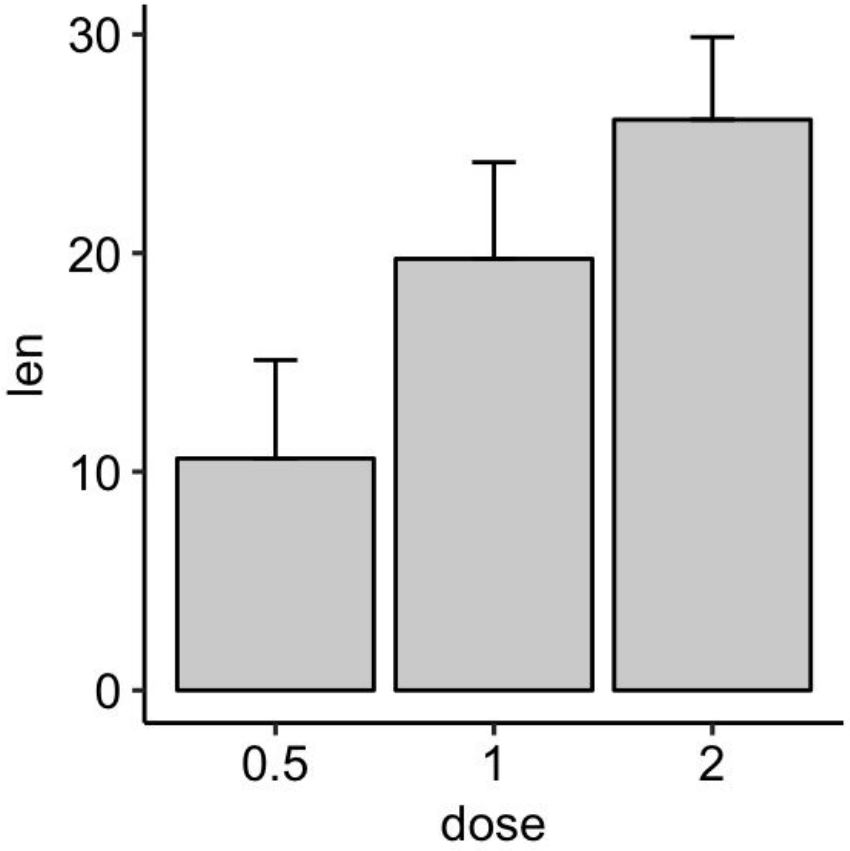
Combined with a violin plot,
Such that I can see the **kernel density**!

Ideally: show me the RAW data





Barplot?



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

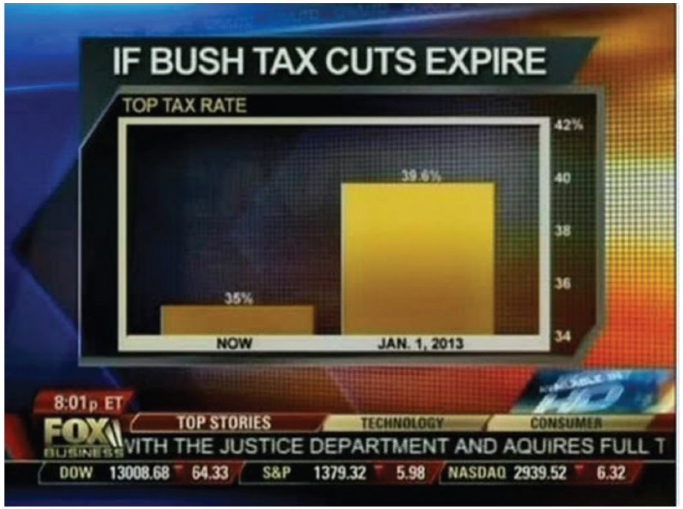
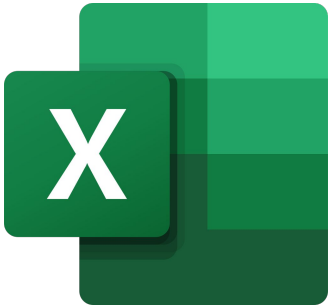


FIGURE 2.12 Fox News bar chart

Nothing is tool specific



matplotlib

bokeh

seaborn

plotly

IBM SPSS

OriginPro 2022
The Ultimate Software for Graphing & Analysis

GraphPad Prism

Prism

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

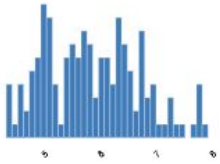
sepal length (cm)

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



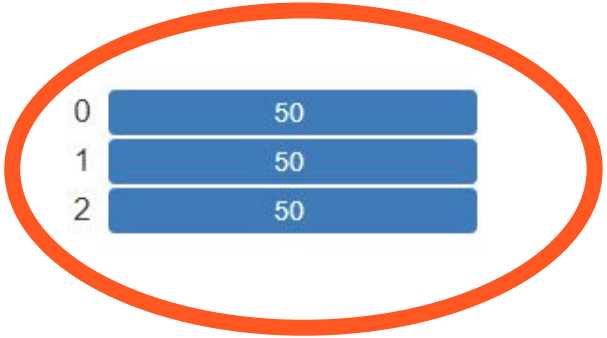
target

Categorical

HIGH CORRELATION

UNIFORM

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



Classes are balanced!

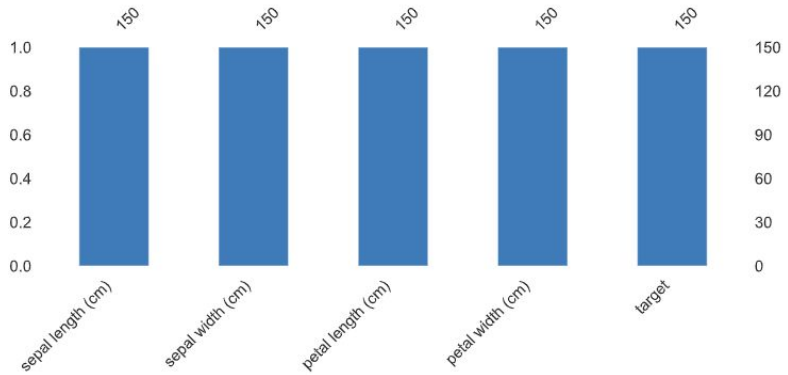
From: Learning from imbalanced data: open challenges and future directions

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/

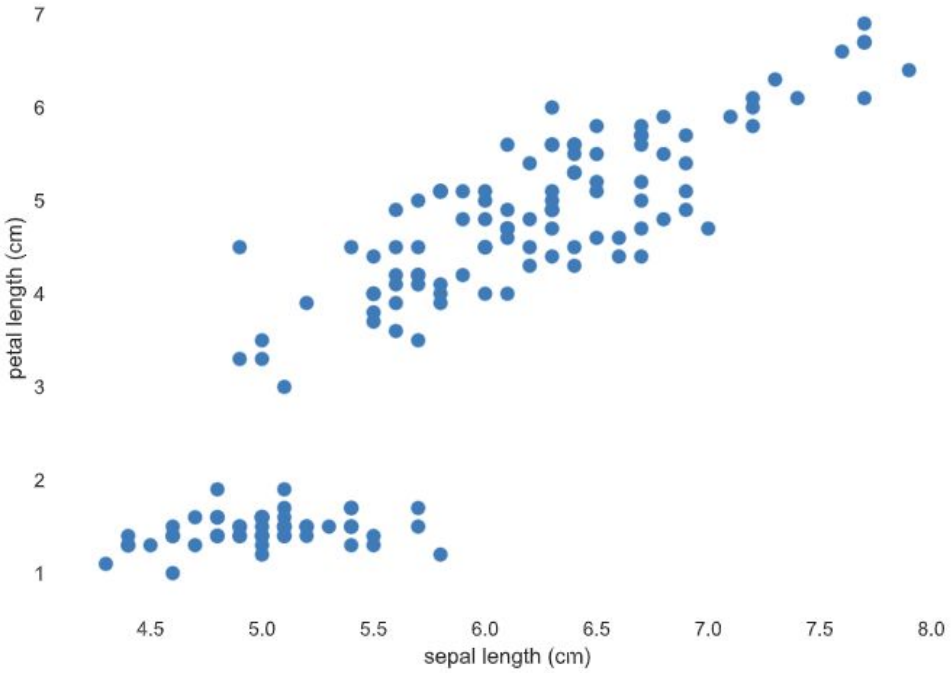
Anything missing or duplicate?



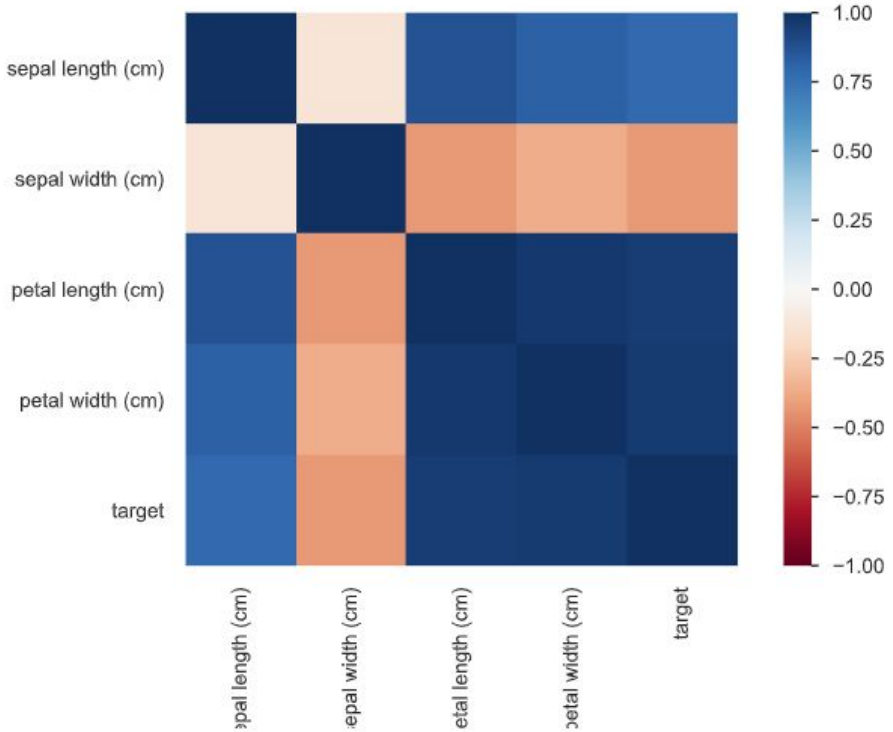
Most frequently occurring

	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?!



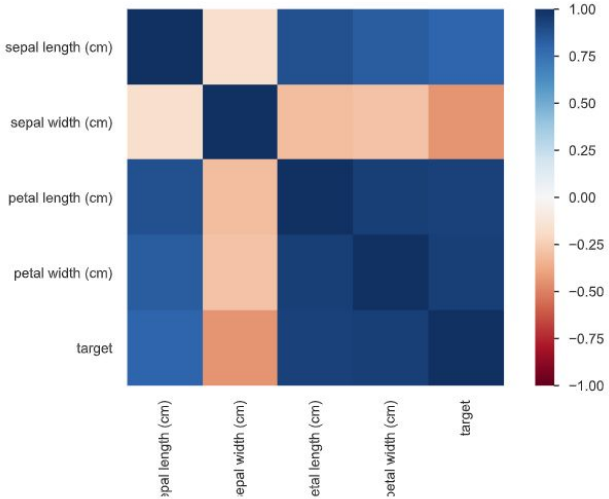
Multiple correlations at one glance



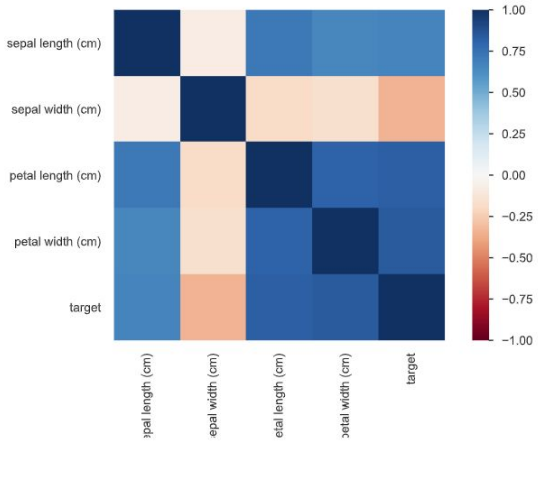
Pearson’s r across samples

All the others

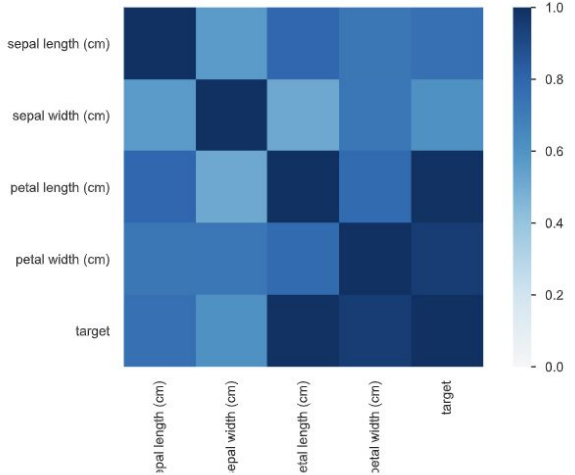
SPEARMAN



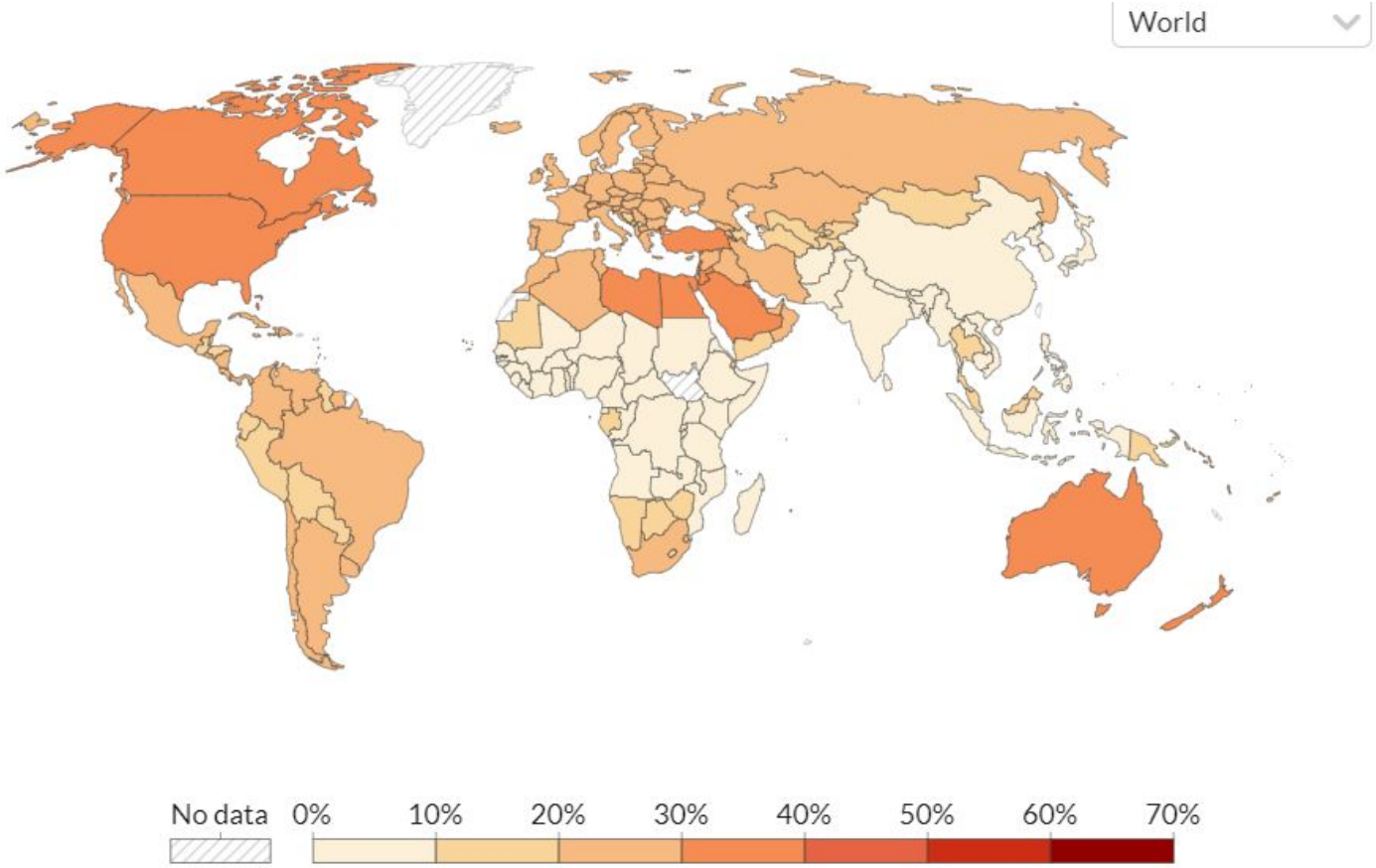
KENDALL



PHIK



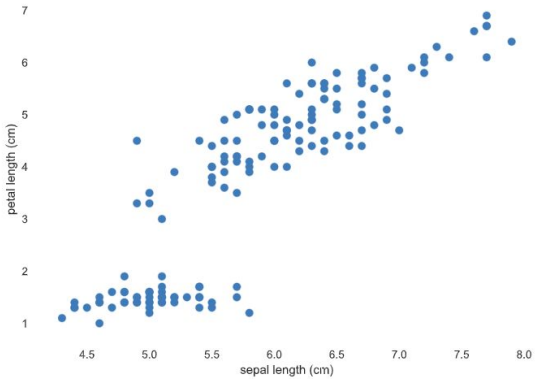
A heatmap

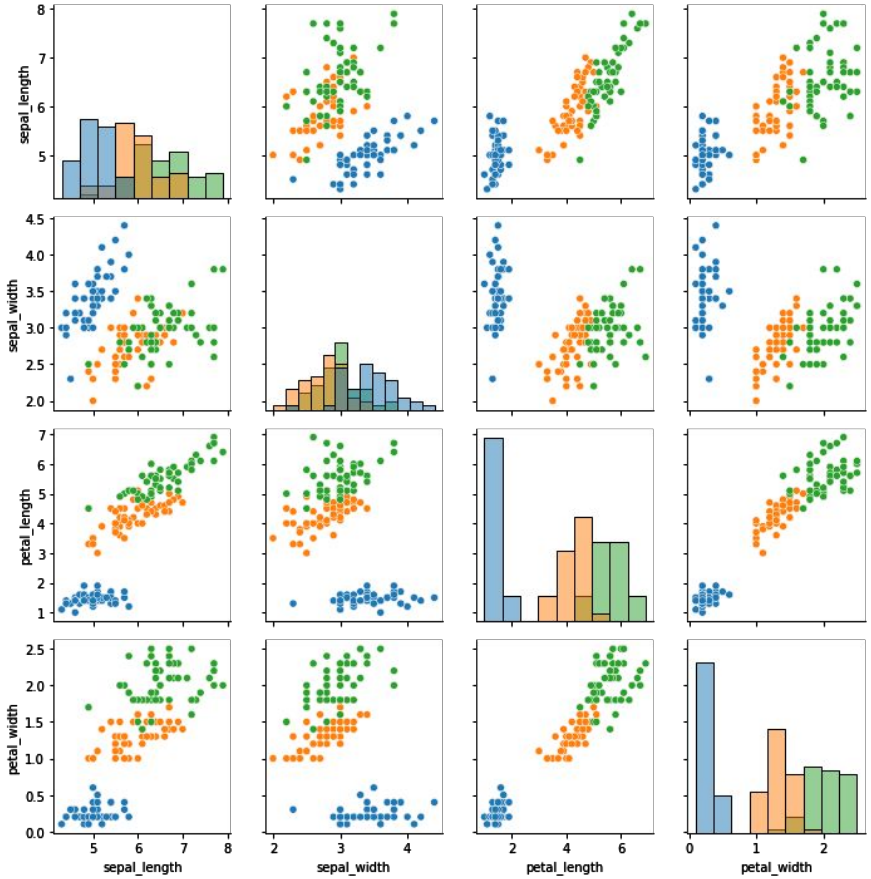


Share adults w/ obesity

Correlation?!

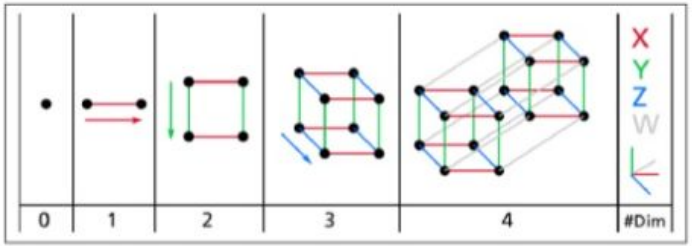
Only two dimensions at a time...





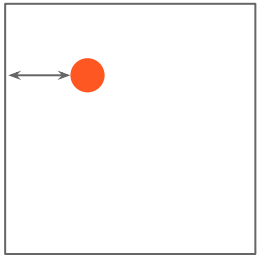
The curse of dimensionality

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 %

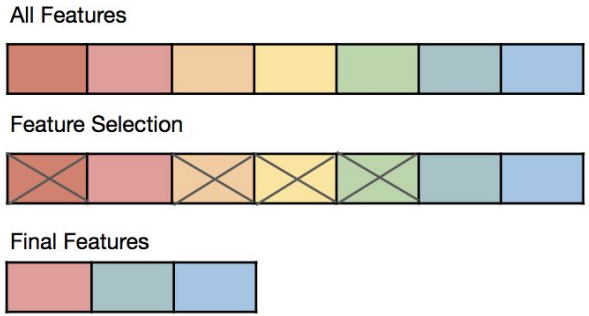
Second example: mean distance of two random points?

Unit square: **0.52**

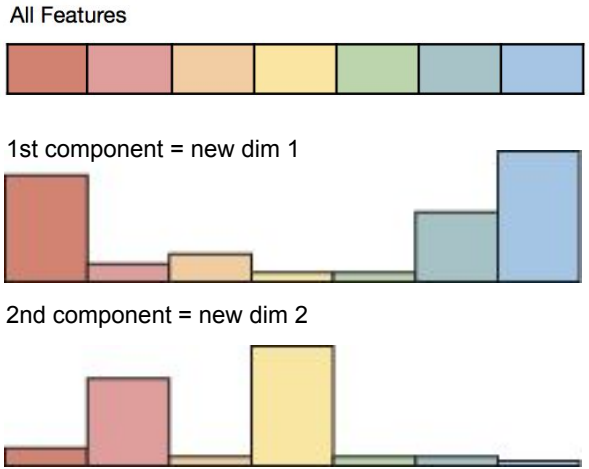
10,000 dim hypercube: **408.25!!!**

Dimensionality reduction

Feature selection



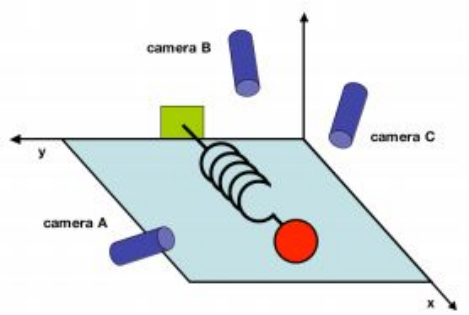
Feature projection



Principal component analysis

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

Jon Shlens



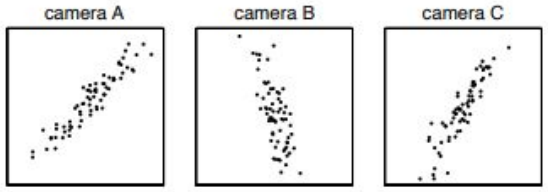
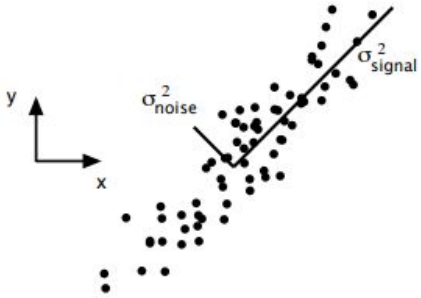


FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

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

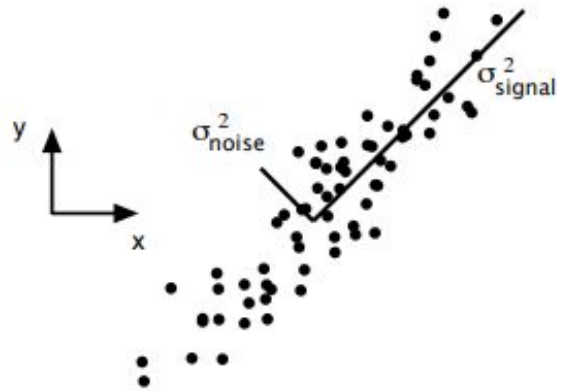
With this rigor we may now state more precisely what PCA asks: *Is there another basis, which is a linear combination of the original basis, that best re-expresses our data set?*



PCA

Via Eigenvectors of covariance matrix

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



Via Singular Value Decomposition (SVD)

$$\mathbf{X}\mathbf{V} = \mathbf{U}\Sigma$$

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.

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

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).

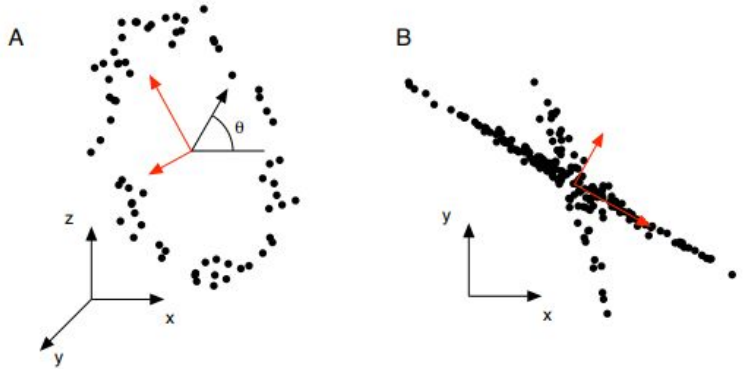
III. *The principal components are orthogonal.*
This assumption provides an intuitive simplification that makes PCA soluble with linear algebra decomposition techniques. These techniques are highlighted in the two following sections.

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.

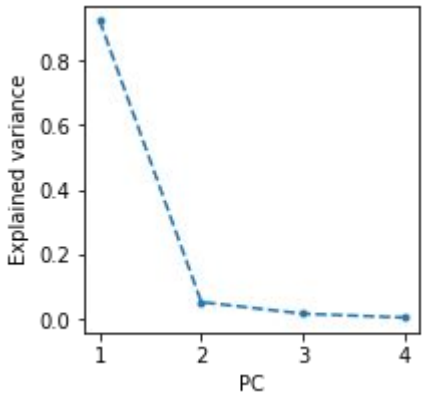
z-scored

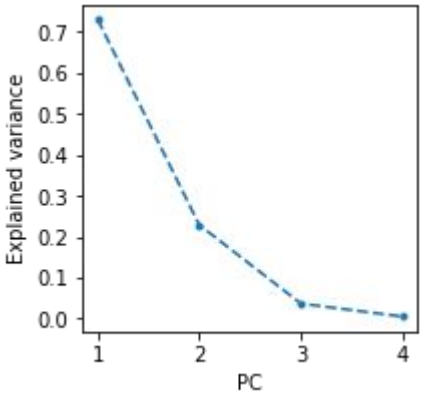
FIG. 5 A step-by-step instruction list on how to perform principal component analysis

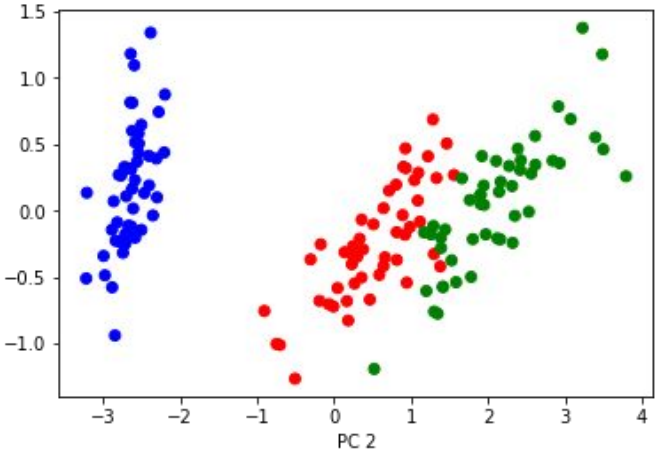


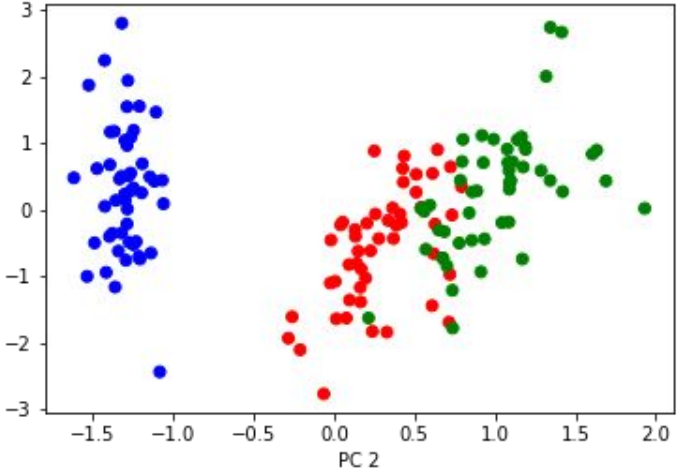
PCA

Z-scoring WHITEN YOUR DATA!



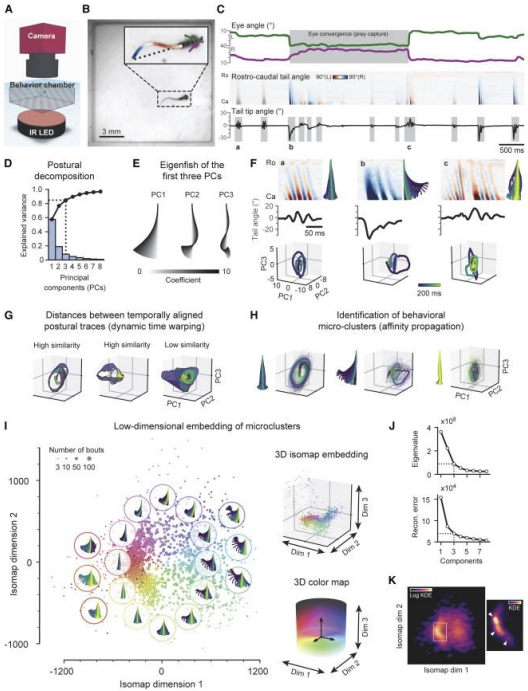




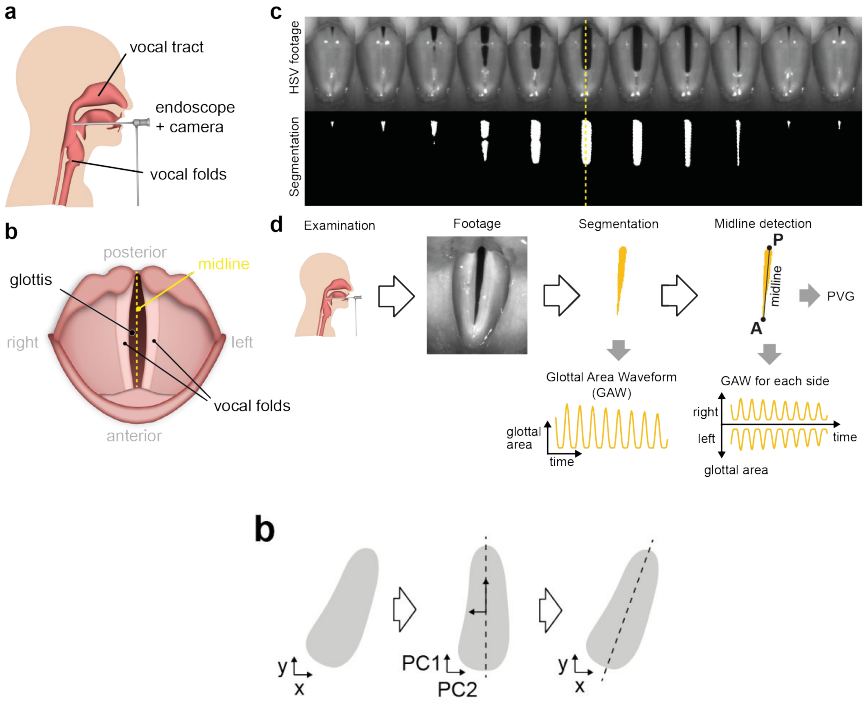


Try exploring this on the WINE dataset!!

Ok, cool, but where do I need it?



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

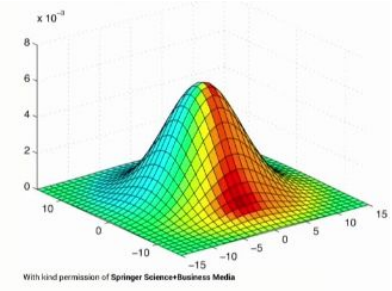


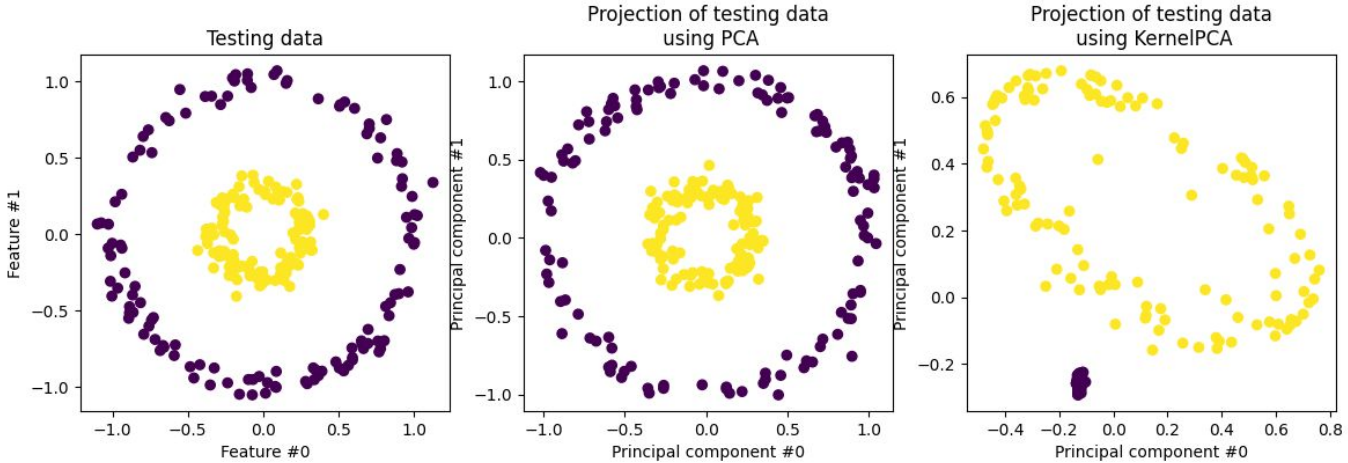
Glottal midline axis computation
Kist et al., Sci Rep 2020

Kernel trick -> Kernel PCA

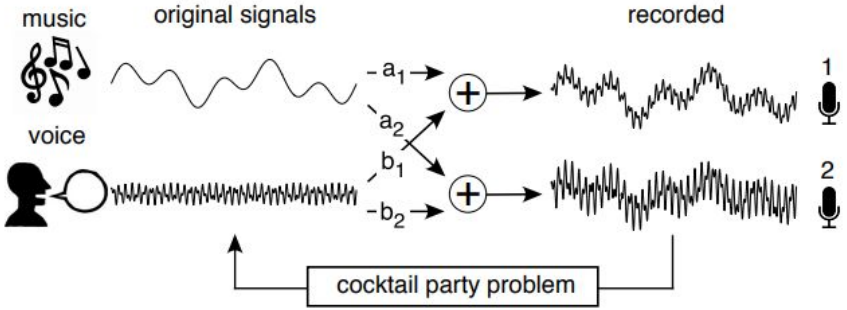
Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models

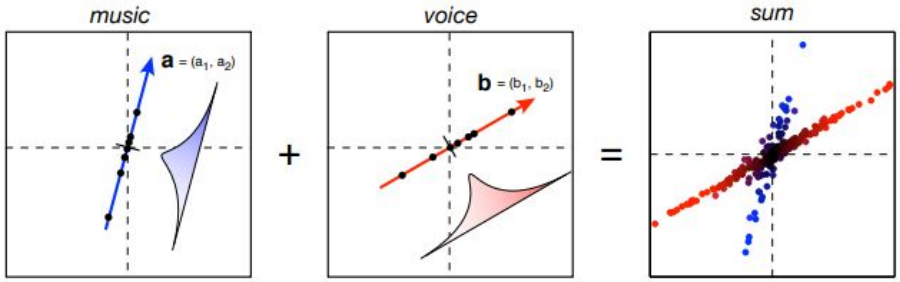
Quan Wang WANGQ10@RPI.EDU
Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180 USA





Independent component analysis (ICA)

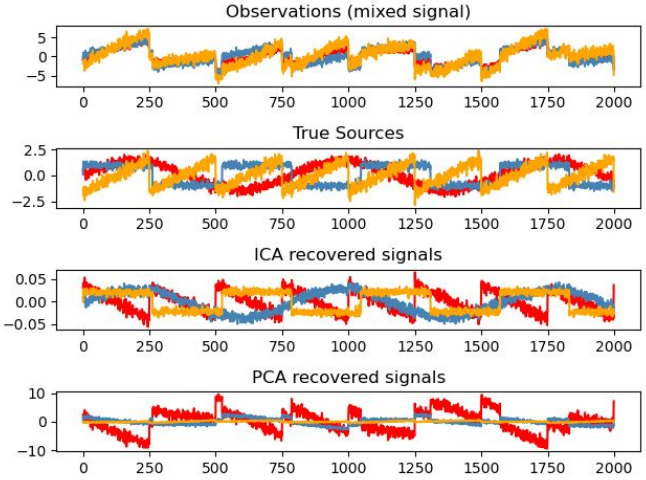




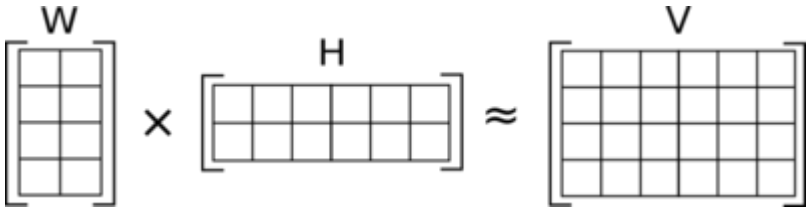
Quick Summary of ICA

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)



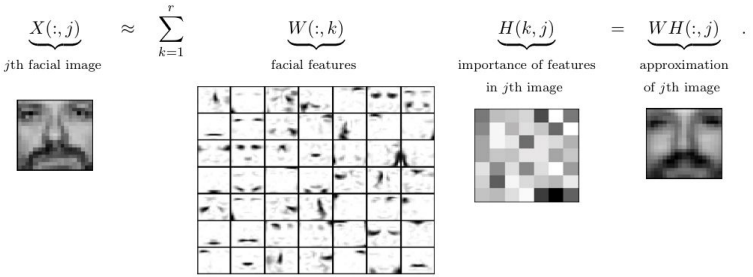


Figure 1: Decomposition of the CBCL face database, MIT Center For Biological and Computation Learning (2429 gray-level 19-by-19 pixels images) using $r = 49$ as in [79].

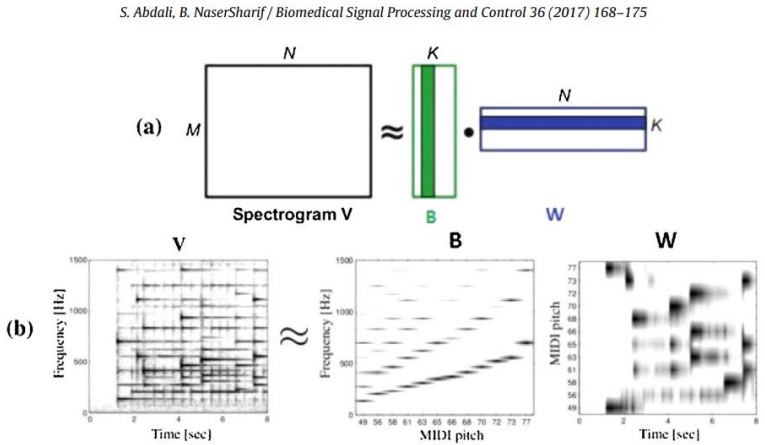
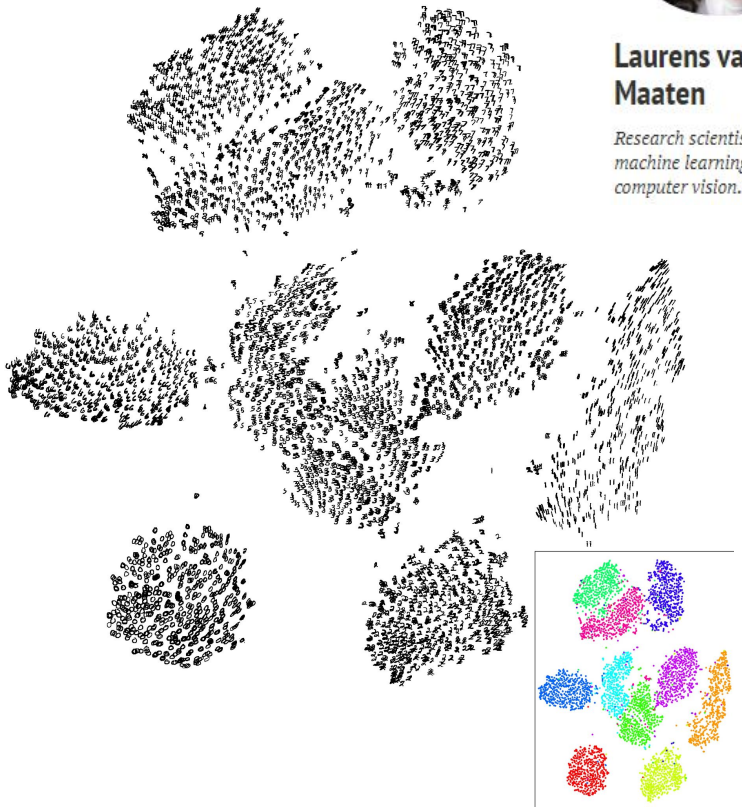


Fig. 1. Non-negative matrix factorization.

t-SNE

Laurens van der Maaten

Research scientist in machine learning and computer vision.



1. Those hyperparameters (e.g. learning rate, perplexity) really matter
2. Cluster sizes in a t-SNE plot mean nothing
3. Distances between clusters might not mean anything
4. Random noise doesn't always look random.
5. You can see some shapes, sometimes
6. For topology, you may need more than one plot

https://distill.pub/2016/misread-tsne/

sklearn.manifold.TSNE

class sklearn.manifold.TSNE(*n_components=2, *, perplexity=30.0, early_exaggeration=12.0, learning_rate='warn', n_iter=1000, n_iter_without_progress=300, min_grad_norm=1e-07, metric='euclidean', init='warn', verbose=0, random_state=None, method='barnes_hut', angle=0.5, n_jobs=None, square_distances='legacy'*) [source

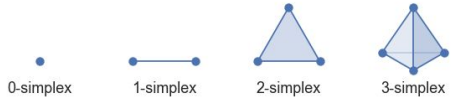
T-distributed Stochastic Neighbor Embedding.

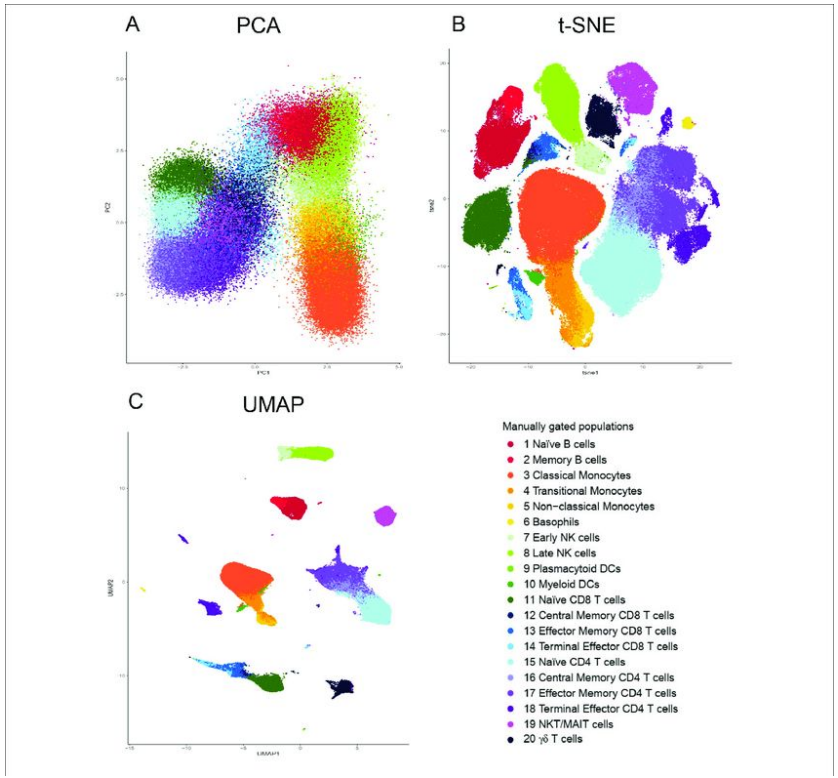
t-SNE [1] is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. t-SNE has a cost function that is not convex, i.e. with different initializations we can get different results.

It is highly recommended to use another dimensionality reduction method (e.g. PCA for dense data or TruncatedSVD for sparse data) to reduce the number of dimensions to a reasonable amount (e.g. 50) if the number of features is very high. This will suppress some noise and speed up the computation of pairwise distances between samples. For more tips see Laurens van der Maaten's FAQ [2].

UMAP

Čech complex → working on simplex





https://pair-code.github.io/understanding-umap/

How to (mis)read UMAP

While UMAP offers a number of advantages over t-SNE, it's by no means a silver bullet - and reading and understanding its results requires some care. It's worth revisiting our previous work on (mis)reading t-SNE, since many of the same takeaways apply to UMAP:

1. Hyperparameters really matter

Choosing good values isn't easy, and depends on both the data and your goals (eg, how tightly packed the projection ought to be). This is where UMAP's speed is a big advantage - By running UMAP multiple times with a variety of hyperparameters, you can get a better sense of how the projection is affected by its parameters.

2. Cluster sizes in a UMAP plot mean nothing

Just as in t-SNE, the size of clusters relative to each other is essentially meaningless. This is because UMAP uses local notions of distance to construct its high-dimensional graph representation.

3. Distances between clusters might not mean anything

Likewise, the distances between clusters is likely to be meaningless. While it's true that the global positions of clusters are better preserved in UMAP, the distances between them are not meaningful. Again, this is due to using local distances when constructing the graph.

4. Random noise doesn't always look random.

Especially at low values of `n_neighbors`, spurious clustering can be observed.

5. You may need more than one plot

Since the UMAP algorithm is stochastic, different runs with the same hyperparameters can yield different results. Additionally, since the choice of hyperparameters is so important, it can be very useful to run the projection multiple times with various hyperparameters.

Importance of context

First part

Exploratory

Showing all your data

Explanatory

Second part

Showing only the relevant data

WHO?

WHAT?

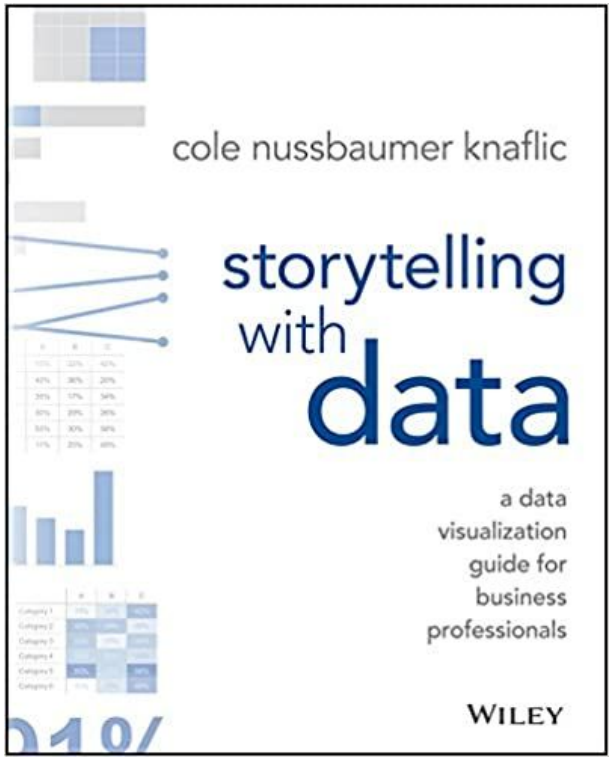
HOW?

What do papers/theses do?

Telling a story.

⇒ People love stories, and storytelling is a key skill to acquire!

What people like or need



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833
Ten Simple Rules for Better Figures

Data needs condensation

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

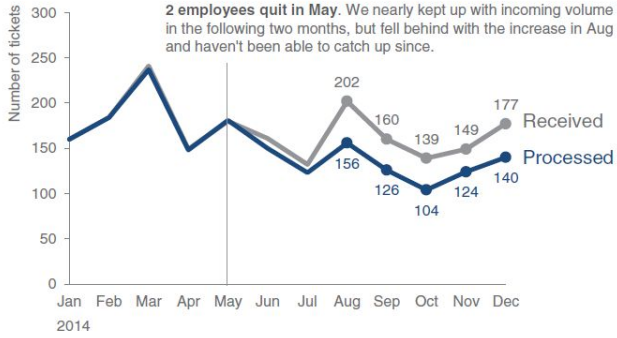
R.A. Fisher, 1936: The Iris Dataset

FIGURE 0.2 Example 1 (before): storytelling with data

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time



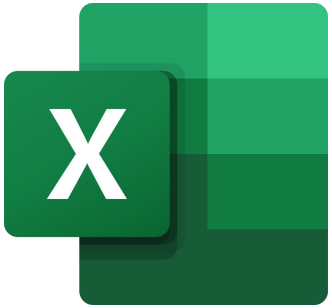
Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

FIGURE 0.3 Example 1 (after): storytelling with data

The golden rules for effective communication

1. Understand the **context**
2. Choose an **appropriate** visual display
3. **Eliminate clutter**
4. Focus **attention** where you want it
5. Think like a **designer**
6. **Tell a story**

Nothing is tool specific



matplotlib

bokeh

seaborn

plotly

IBM SPSS

OriginPro 2022
The Ultimate Software for Graphing & Analysis

GraphPad Prism

Prism

Who, what and how.

Who is your audience? And **who** are you in respect to the audience?

What do you need your audience to know or do?
What action is required?

Prompting action

Here are some action words to help act as thought starters as you determine what you are asking of your audience:

accept | agree | begin | believe | change | collaborate | commence | create | defend | desire | differentiate | do | empathize | empower | encourage | engage | establish | examine | facilitate | familiarize | form | implement | include | influence | invest | invigorate | know | learn | like | persuade | plan | promote | pursue | recommend | receive | remember | report | respond | secure | support | simplify | start | try | understand | validate

The how.

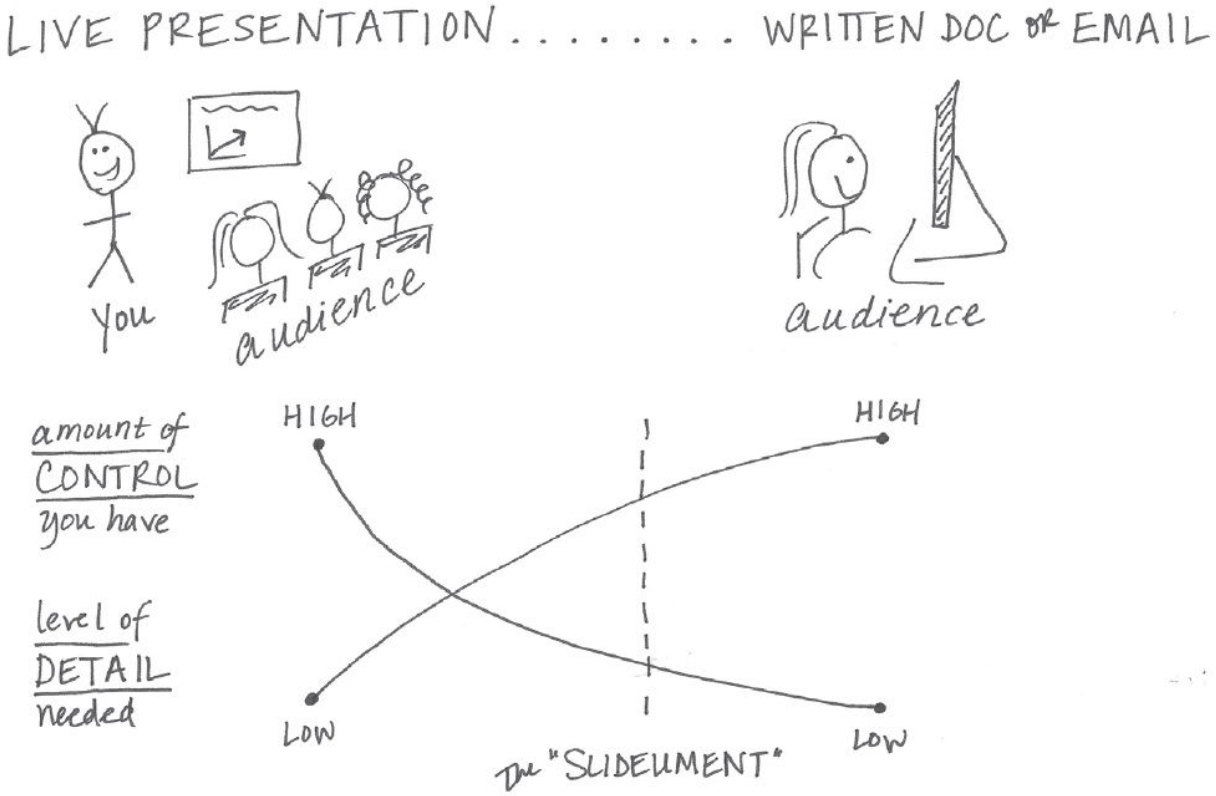


FIGURE 1.1 Communication mechanism continuum

3-Minute story, big idea, elevator pitch

“If you know exactly what it is you want to communicate, you can make it fit the time slot you’re given, even if it isn’t the one for which you are prepared.”

The BIG IDEA components (by Nancy Duarte):

- It must articulate your unique Point of View
- It must convey what’s at stake
- It must be a complete sentence

Example: “The pilot summer learning program was successful at improving students’ perceptions of science and, because of this success, we recommend continuing to offer it going forward; please approve our budget for this program.”

Storyboarding

- Try to avoid PPTX, try to use low tech first (post its, whiteboard, paper, ...)

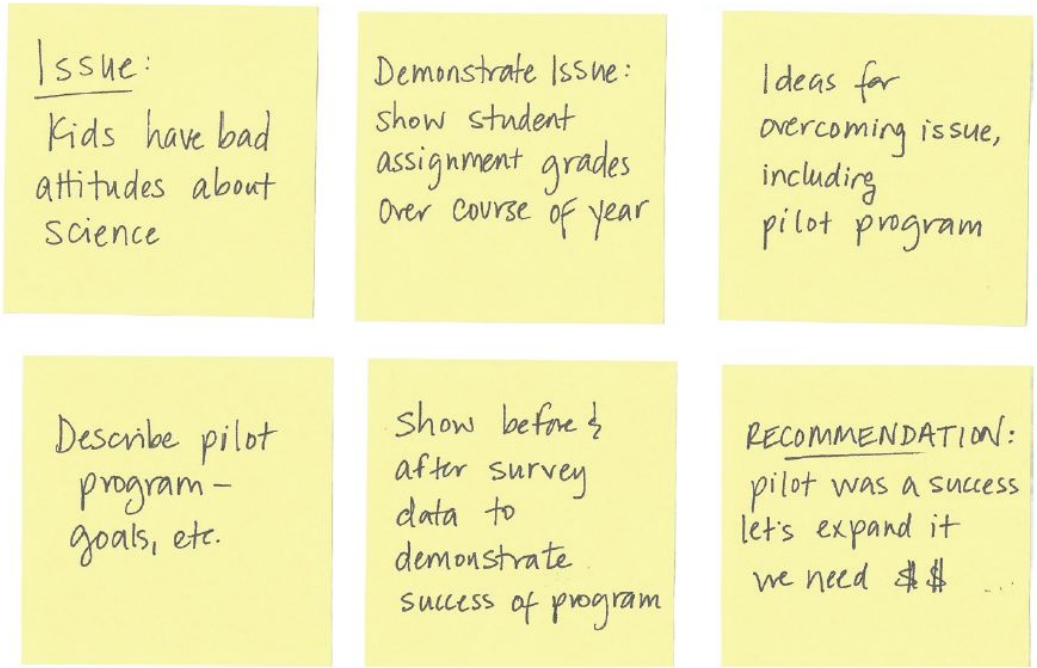
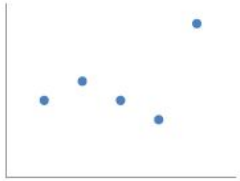


FIGURE 1.2 Example storyboard

Choosing an effective visual

Simple text



Scatterplot

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Table

Line

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

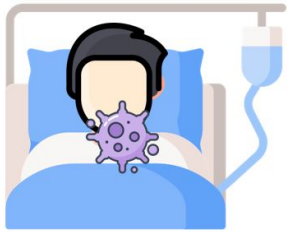
Slopegraph

FIGURE 2.1 The visuals I use most

FAU

Friedrich-Alexander-Universität
Technische Fakultät

Head&Neck cancer is a common and serious threat



Nick

Head&Neck cancer

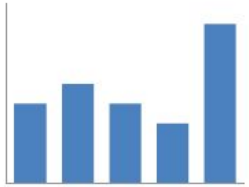
#8 for males, #13 for females

More contrast For gray

43% 5 yr survival rate for males, 55% females

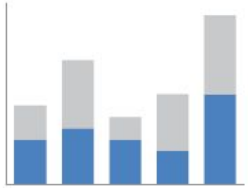
29% 10 yr survival rate for males, 40% females

Choosing an effective visual

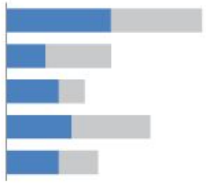


Vertical bar

Horizontal bar

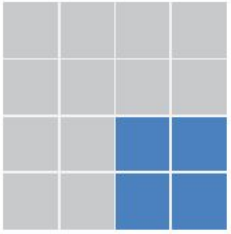


Stacked vertical bar



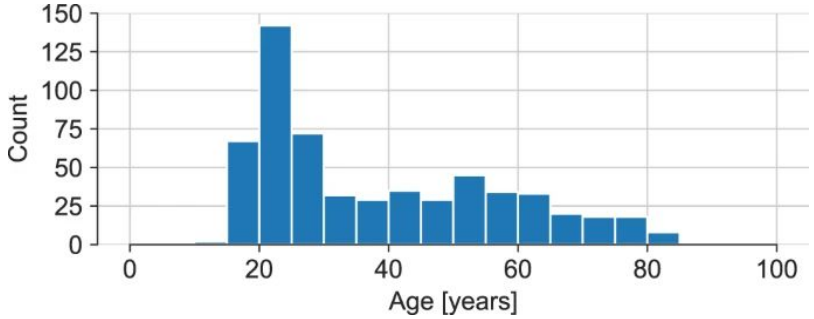
Stacked horizontal bar

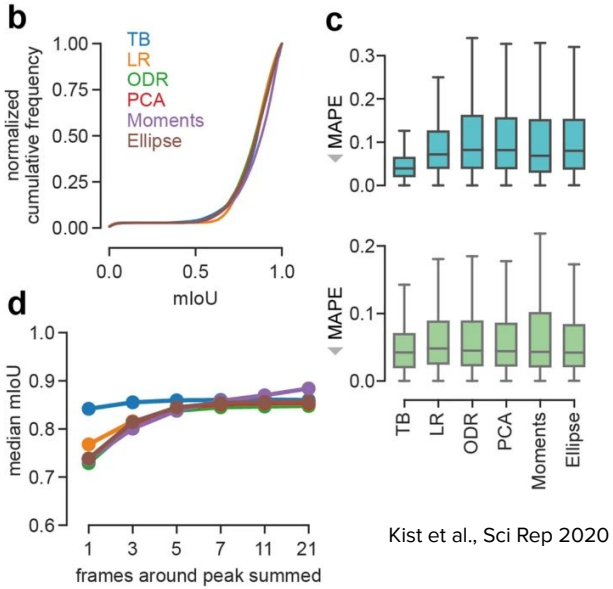
Waterfall



Square area

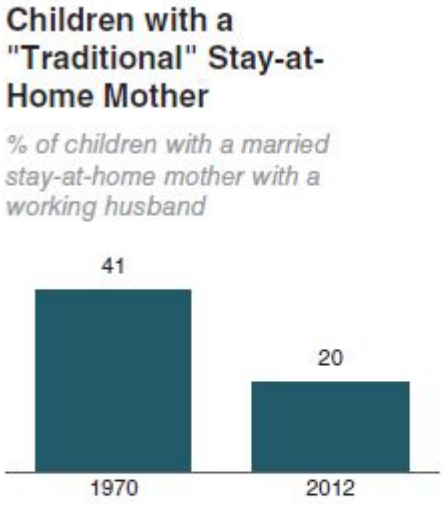
Gómez*, Kist* et al., Sci Data 2020





Kist et al., Sci Rep 2020

Plain text (esp. for talks)



Note: Based on children younger than 18. Their mothers are categorized based on employment status in 1970 and 2012.

Source: Pew Research Center analysis of March Current Population Surveys Integrated Public Use Microdata Series (IPUMS-CPS), 1971 and 2013

Adapted from PEW RESEARCH CENTER

FIGURE 2.2 Stay-at-home moms original graph

20%

of children had a
traditional stay-at-home mom
in 2012, compared to 41% in 1970

FIGURE 2.3 Stay-at-home moms simple text makeover

Tables

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Light borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

FIGURE 2.4 Table borders

➜ Use Gestalt principles

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

FIGURE 2.5 Two views of the same data

Scatter plots

plt.scatter

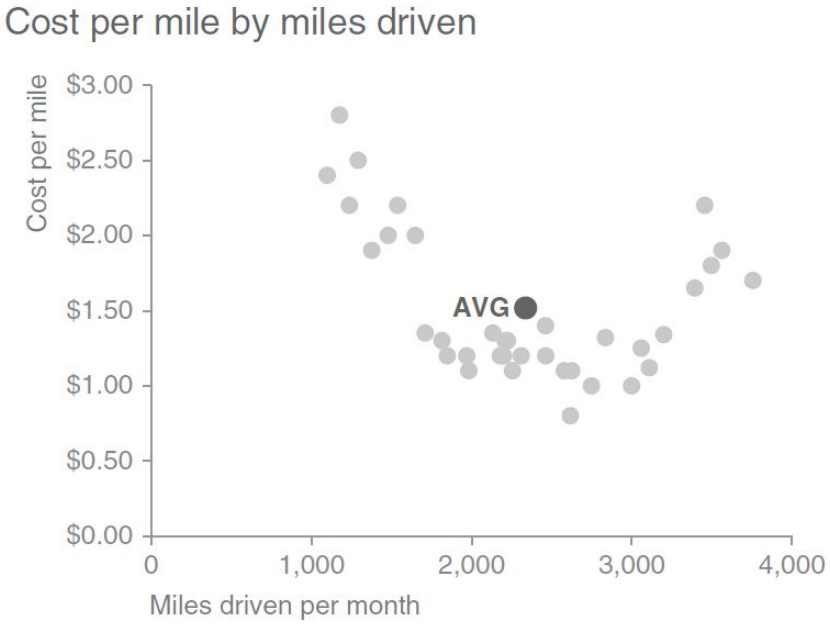


FIGURE 2.6 Scatterplot

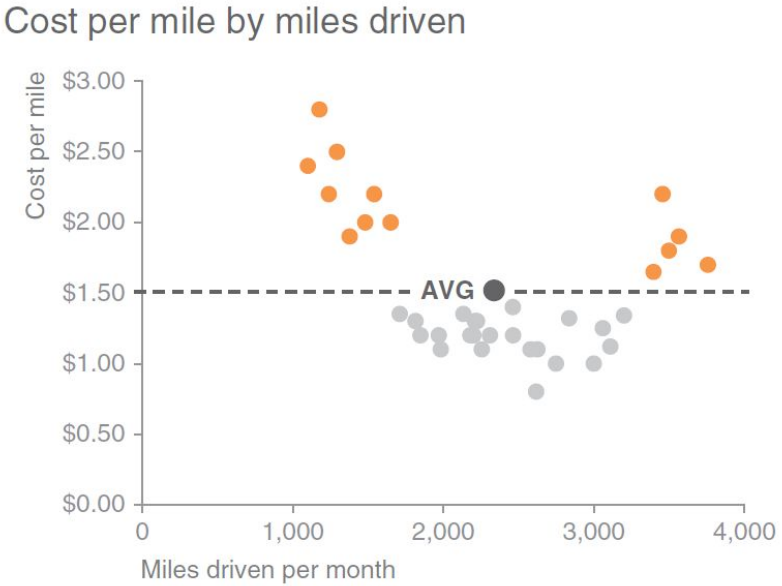


FIGURE 2.7 Modified scatterplot

Line graphs

plt.plot

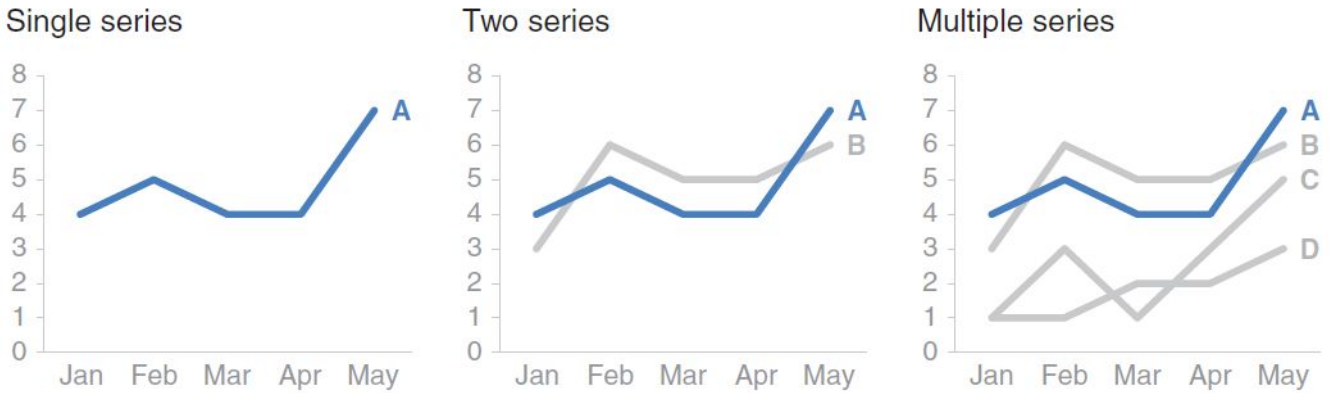


FIGURE 2.8 Line graphs

FIGURE 2.9 Showing average within a range in a line graph

plt.fill_between

Slope graph

plt.plot

FIGURE 2.11 Modified slopegraph

- Good to show connected trends
- Well suited for homogeneity in data (all go up, all go down, or mix of everything)

Bars and how to NOT use them

plt.bar(h)

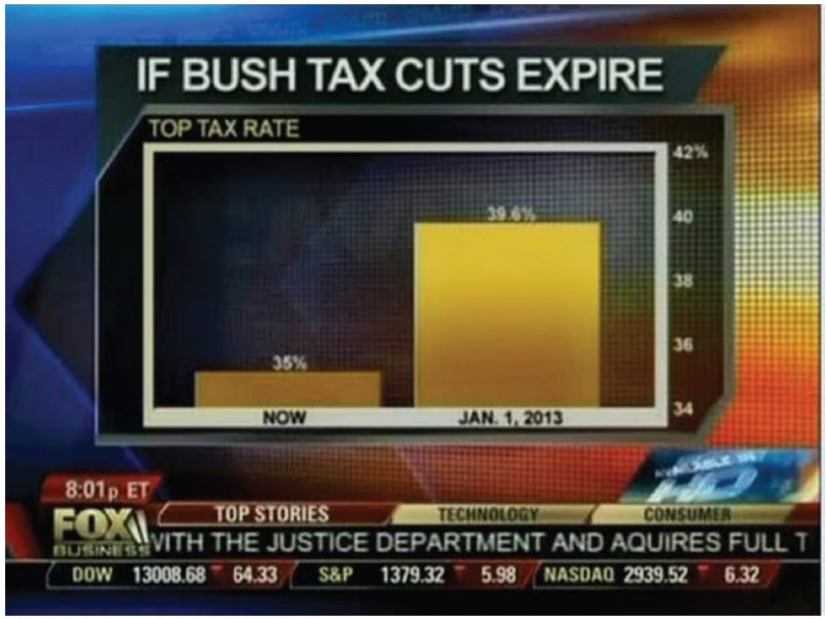
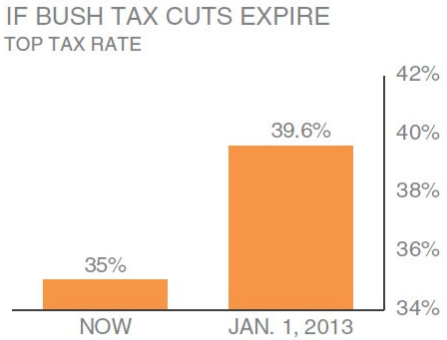


FIGURE 2.12 Fox News bar chart

Non-zero baseline: as originally graphed



Zero baseline: as it should be graphed

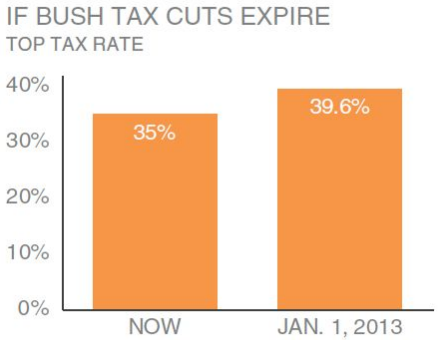


FIGURE 2.13 Bar charts must have a zero baseline

FIGURE 2.14 Bar width

Bar charts

plt.bar(h)

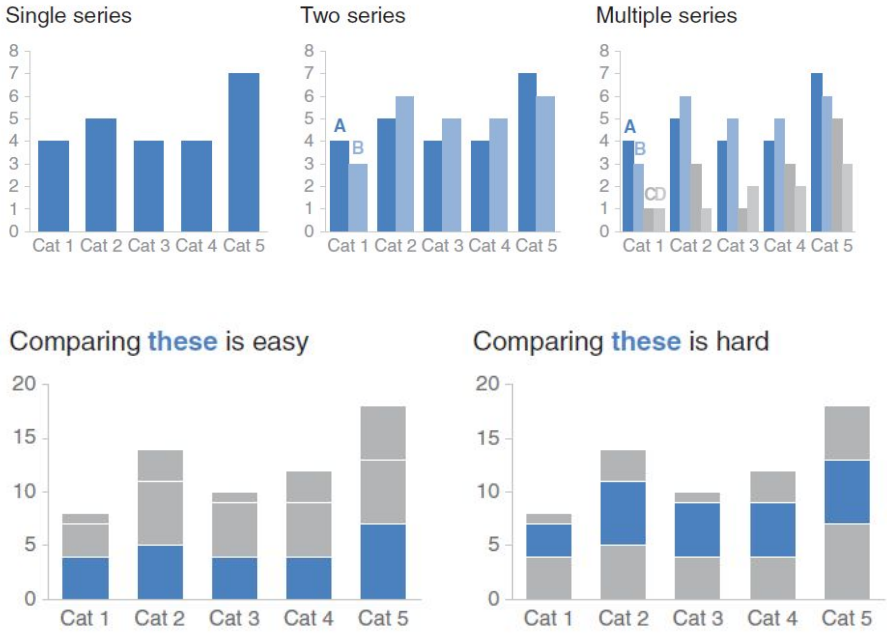


FIGURE 2.16 Comparing series with stacked bar charts

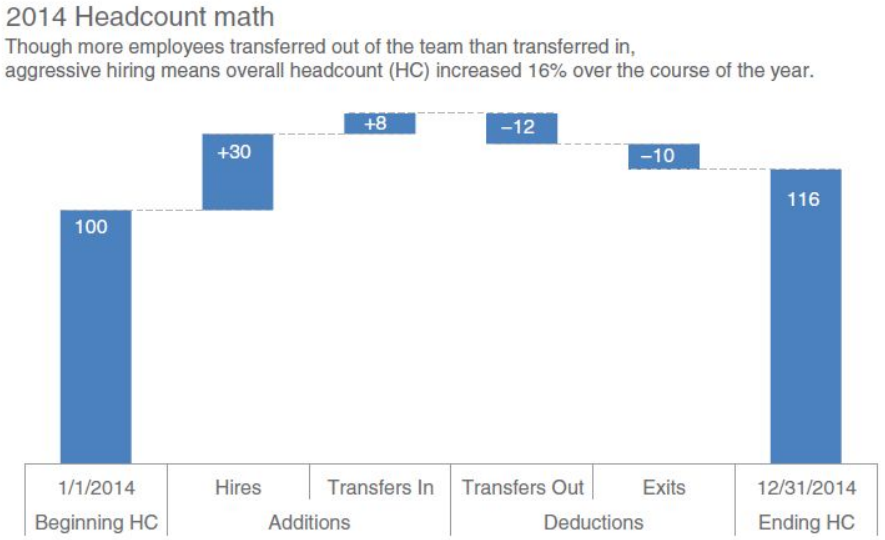


FIGURE 2.17 Waterfall chart

Horizontal bar charts

Especially with long category names!

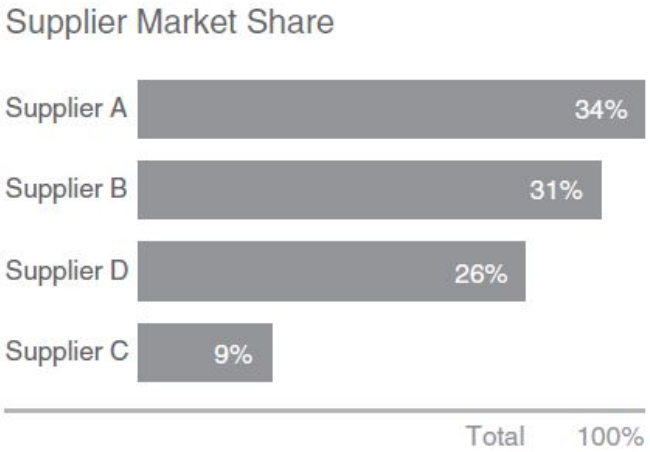


FIGURE 2.23 An alternative to the pie chart

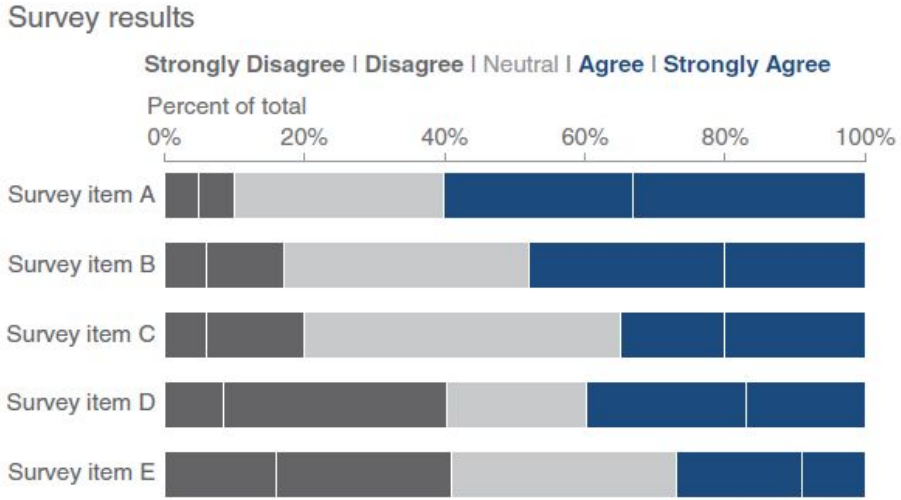


FIGURE 2.19 100% stacked horizontal bar chart

What to avoid in general?

- Pie charts
- 3D effects
- Random color
- Secondary y-axis

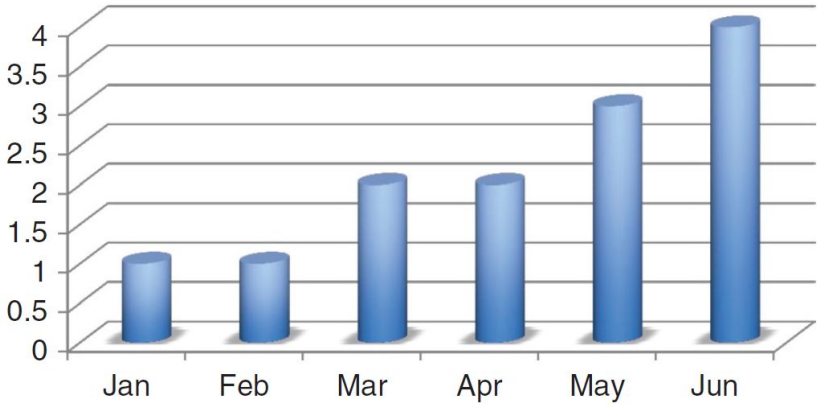


FIGURE 2.25 3D column chart

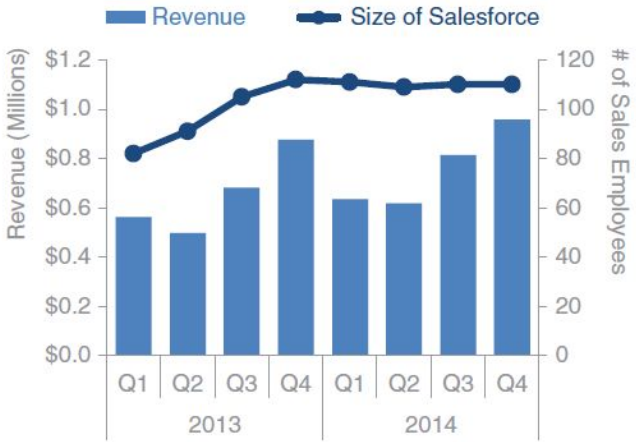


FIGURE 2.26 Secondary y-axis

Alternative 1: **label directly**

Alternative 2: **pull apart vertically**

FIGURE 2.27 Strategies for avoiding a secondary *y*-axis

Remove clutter

proximity, similarity, enclosure, closure, continuity, and connection

Reprise: **Gestalt principles**

FIGURE 3.2 You see columns and rows, simply due to dot spacing

FIGURE 3.7 Gestalt principle of closure

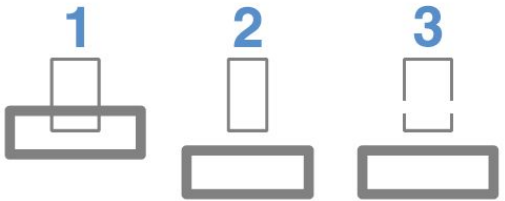


FIGURE 3.9 Gestalt principle of continuity

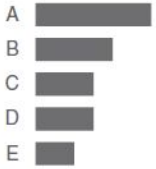


FIGURE 3.10 Graph with *y*-axis line removed

FIGURE 3.4 You see rows due to similarity of color

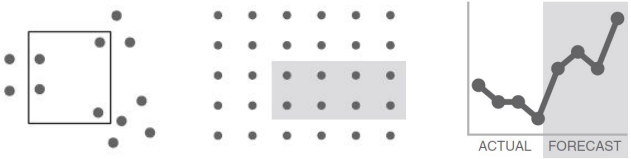


FIGURE 3.5 Gestalt principle of enclosure

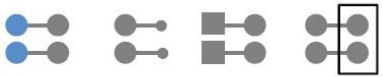


FIGURE 3.11 Gestalt principle of connection

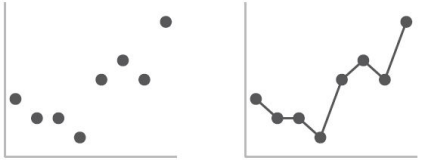


FIGURE 3.12 Lines connect the dots

Lack of visual order

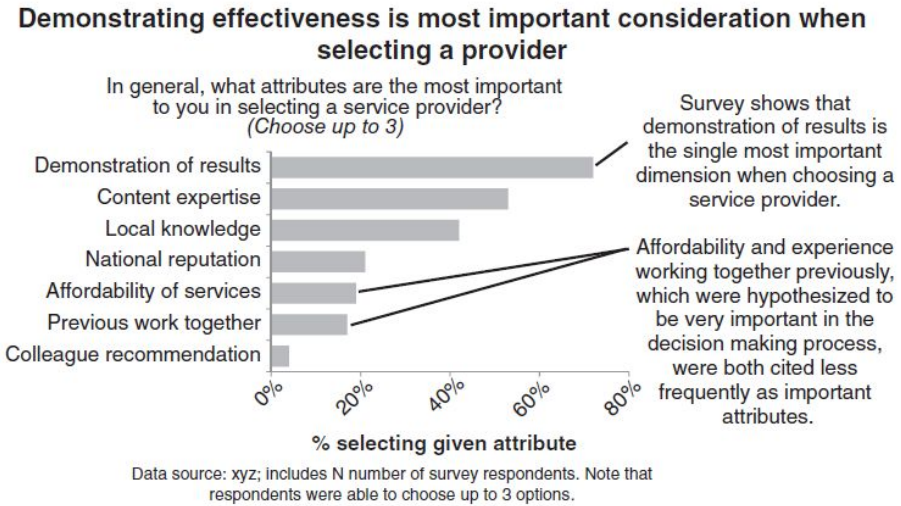


FIGURE 3.13 Summary of survey feedback

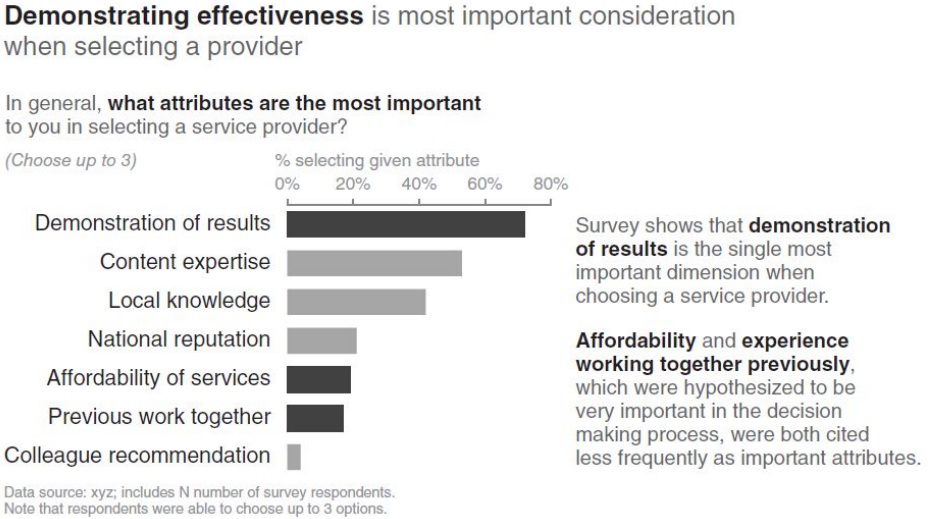
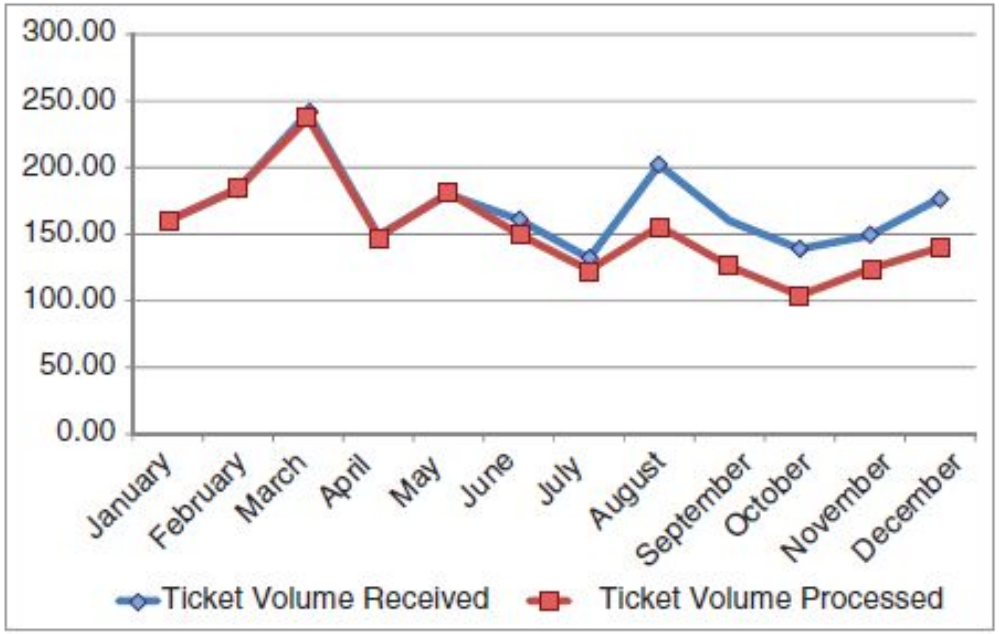


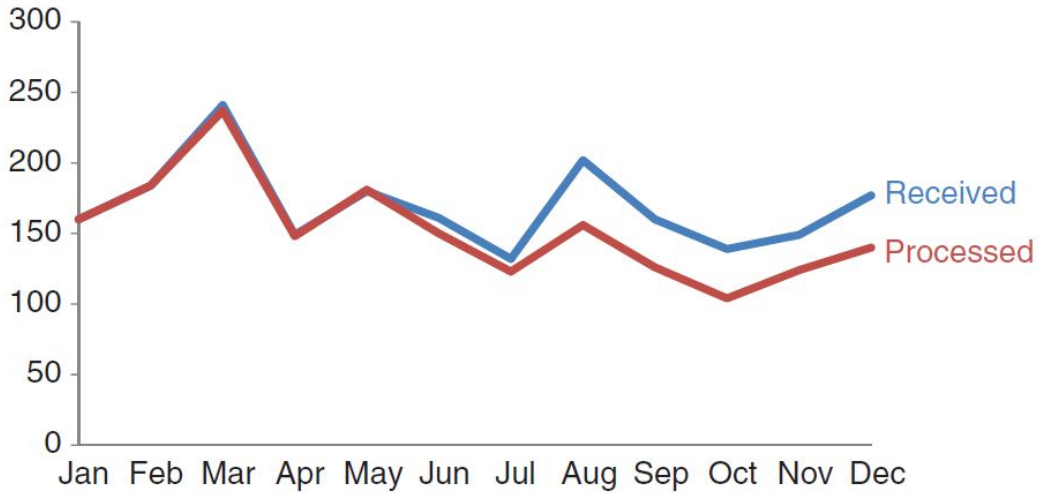
FIGURE 3.14 Revamped summary of survey feedback

How to remove clutter

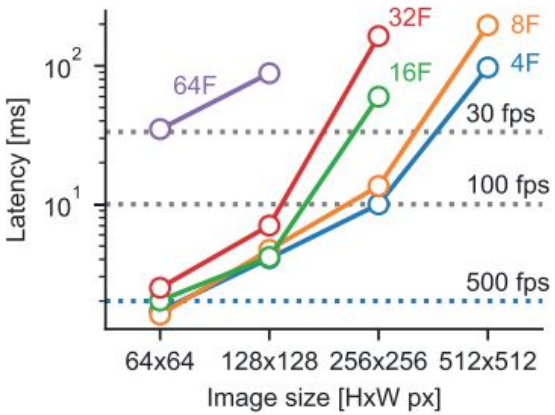


What did we do?

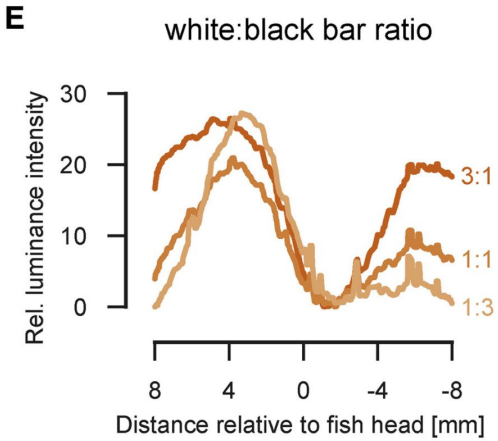
- Remove chart border (Gestalt: closure)
- Remove grid lines (if they don't help)
- Remove data markers
  (personal opinion: add them on purpose!)
- Clean up x-axis
- Label data directly
- Leverage consistent color



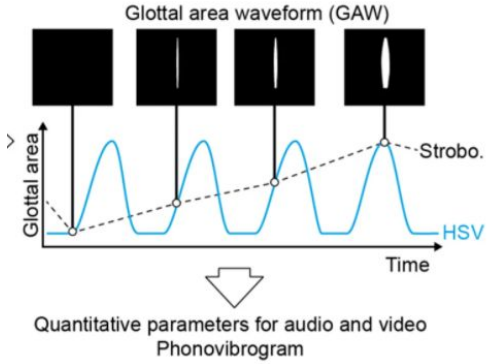
Real life examples



Kist and Döllinger, IEEE Access 2020



Kist and Portugues, Cell Reports 2019



Kist et al., Sci Reports 2021

Drawing attention

FIGURE 4.2 Count the 3s example

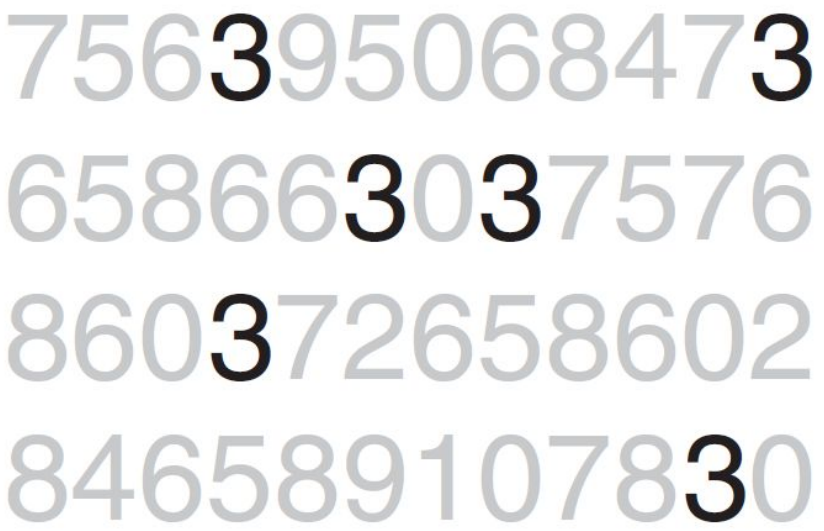


FIGURE 4.3 Count the 3s example with preattentive attribute

Attention/Saliency tricks

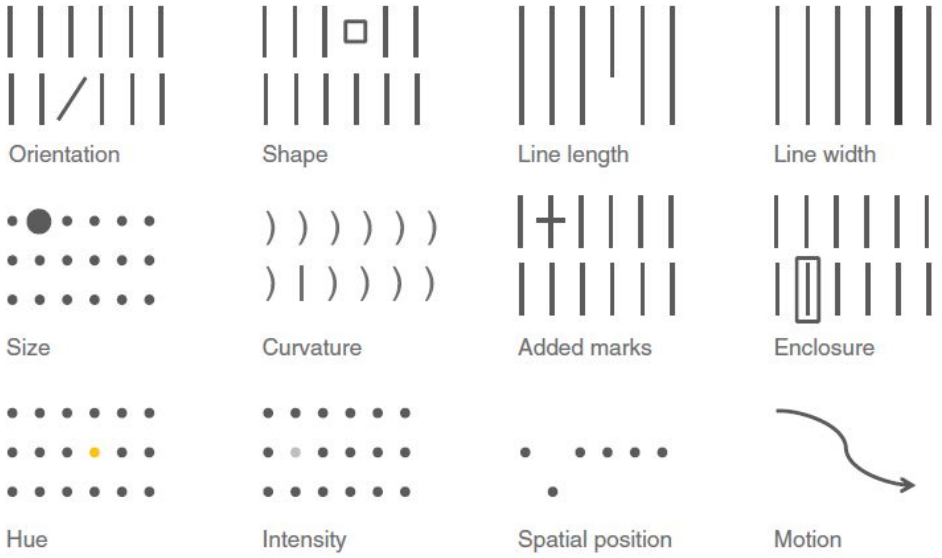


FIGURE 4.4 Preattentive attributes

In text:

- **Bold**
- *Italics*
- Underline
- Color
- Size

- Separate spatially

- Outline (enclosure)

More than 10 per 1,000, 3 are noise-related

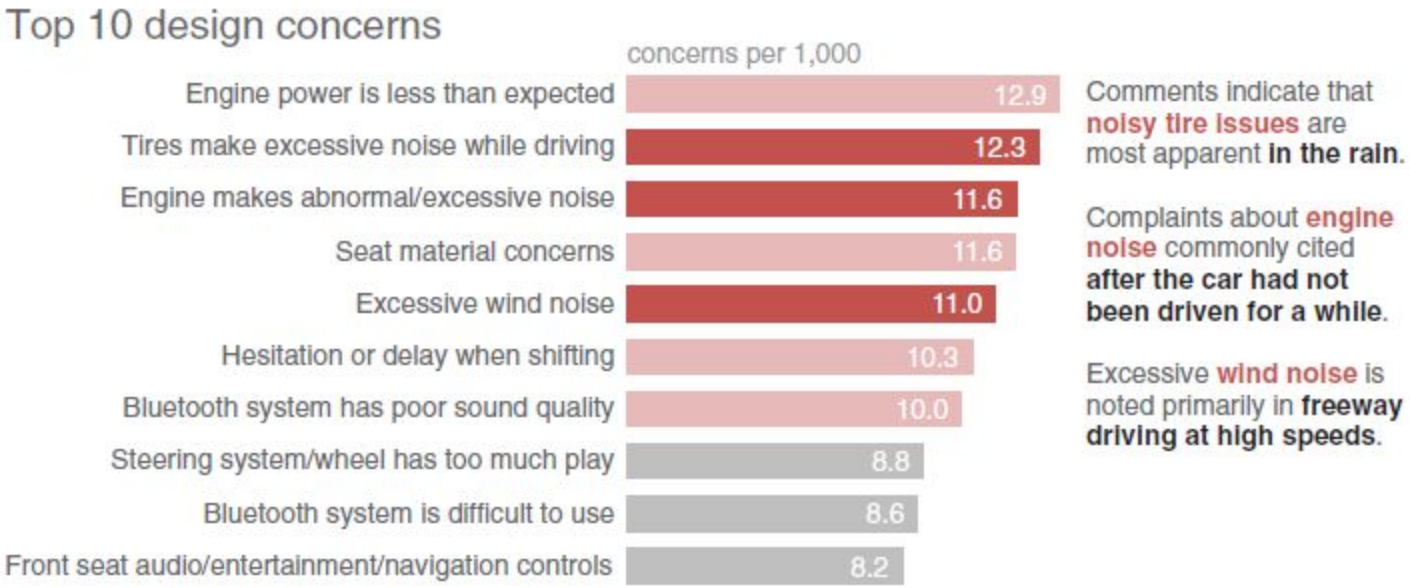


FIGURE 4.9 Create a visual hierarchy of information

Add data directly to labels

FIGURE 4.13 Too many data labels feels cluttered

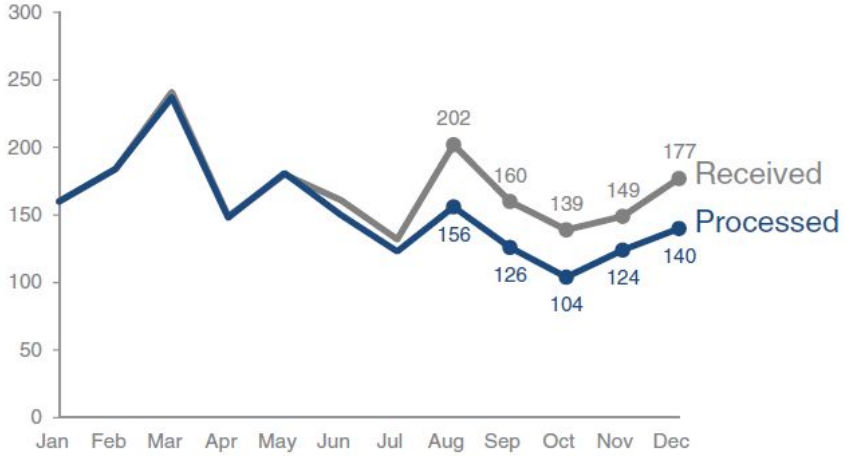


FIGURE 4.14 Data labels used sparingly help draw attention

Use color sparingly

Country Level Sales Rank Top 5 Drugs

Rainbow distribution in color indicates sales rank in given country from #1 (red) to #10 or higher (dark purple)

Country	A	B	C	D	E
AUS	1	2	3	6	7
BRA	1	3	4	5	6
CAN	2	3	6	12	8
CHI	1	2	8	4	7
FRA	3	2	4	8	10
GER	3	1	6	5	4
IND	4	1	8	10	5
ITA	2	4	10	9	8
MEX	1	5	4	6	3
RUS	4	3	7	9	12
SPA	2	3	4	5	11
TUR	7	2	3	4	8
UK	1	2	3	6	7
US	1	2	4	3	5

FIGURE 4.15 Use color sparingly

Top 5 drugs: country-level sales rank

RANK 1 2 3 4 5+

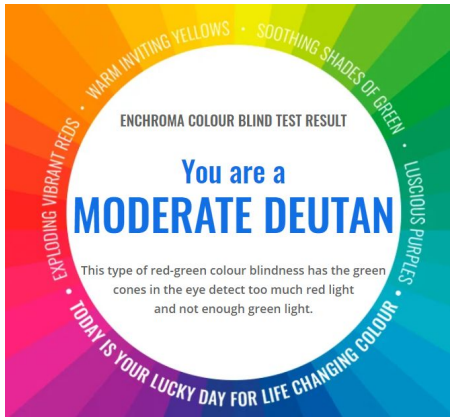
COUNTRY | DRUG

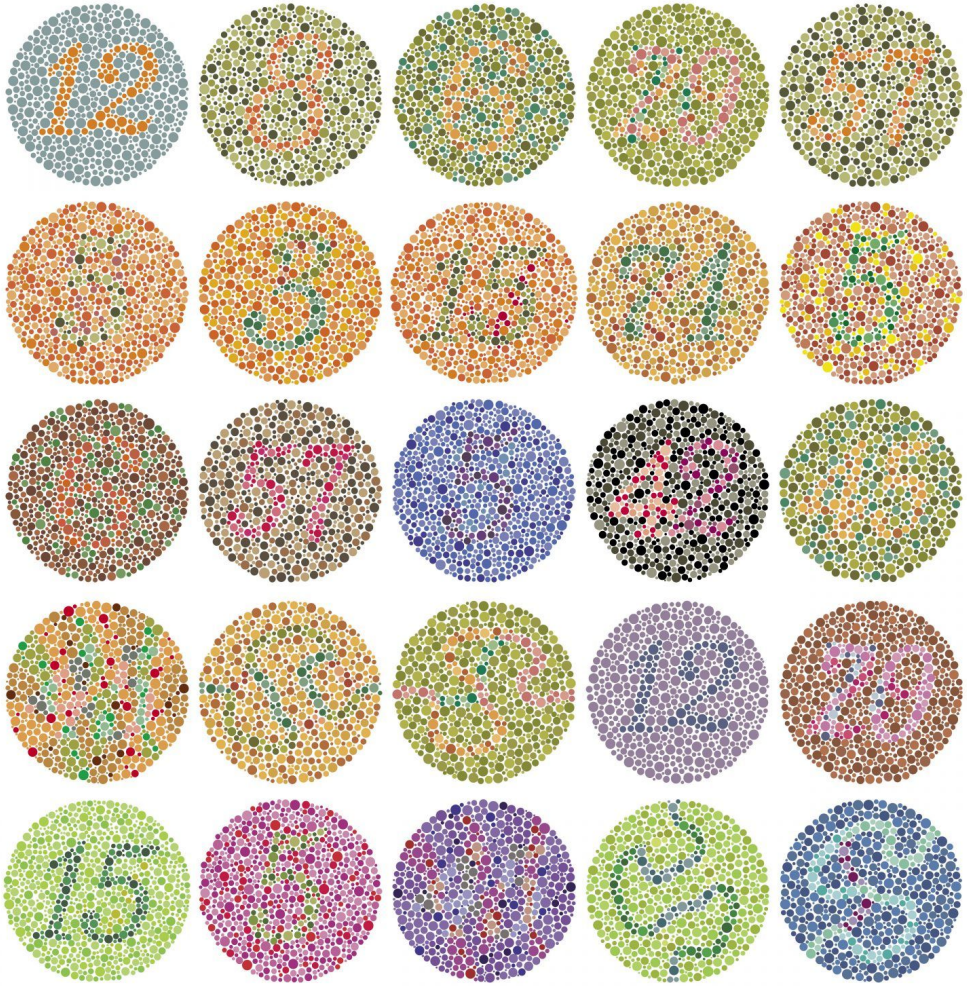
	A	B	C	D	E
Australia	1	2	3	6	7
Brazil	1	3	4	5	6
Canada	2	3	6	12	8
China	1	2	8	4	7
France	3	2	4	8	10
Germany	3	1	6	5	4
India	4	1	8	10	5
Italy	2	4	10	9	8
Mexico	1	5	4	6	3
Russia	4	3	7	9	12
Spain	2	3	4	5	11
Turkey	7	2	3	4	8
United Kingdom	1	2	3	6	7
United States	1	2	4	3	5

Think about colors

Design with colorblind in mind

Roughly 8% of men (including my husband and a former boss) and half a percent of women are colorblind. This most frequently manifests itself as difficulty in distinguishing between shades of red and shades of green. In general, you should avoid using shades of red and shades of green together. Sometimes, though, there is useful connotation that comes with using red and green: red to denote the double-digit loss you want to draw attention to or green to highlight significant growth. You can still leverage this, but make sure to have some additional visual cue to set the important numbers apart so you aren't inadvertently disenfranchising part of your audience. Consider also using bold, varying saturation or brightness, or adding a simple plus or minus sign in front of the numbers to ensure they stand out.





Done for today

HOMEWORK

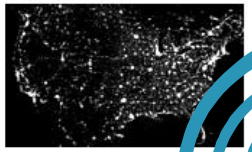
In this lecture, we covered data exploration and visualization. In the exercise you will work with the Census Income Dataset.

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