

# Statistics with Python

Gabriel Torregrosa Cortés

# Index

1. Data and Visualization
2. Association and Correlation
3. Study Design

# Data Exploration and Description

# Index

- Exploratory Data Analysis
  - Data types
  - Plots
  - Summary statistics
  - Outliers I
  - Missing Data
- Learning Outcomes:**

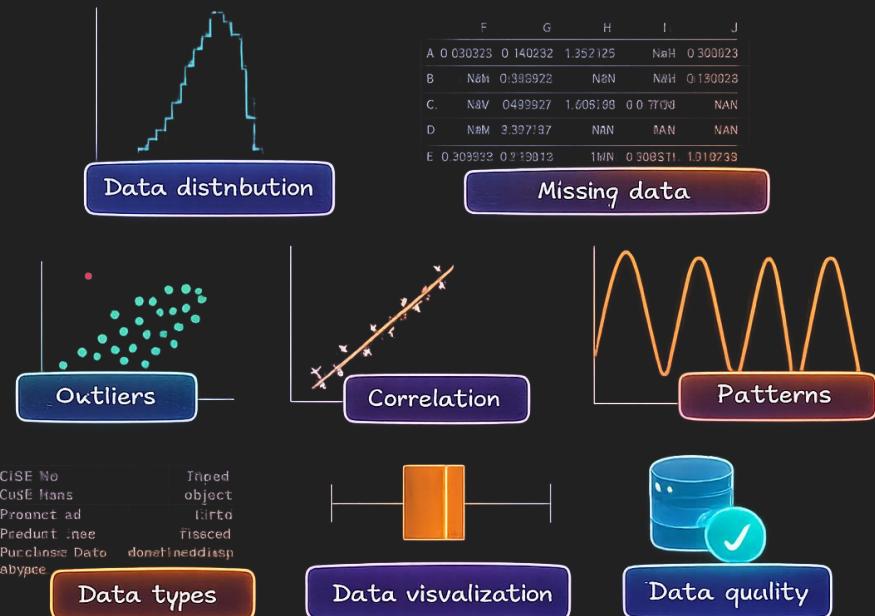
  - EDA Mindset
  - Perform an EDA analysis with pandas and seaborn

# Exploratory Data Analysis (EDA)

- Statistics Start Before Modelling
- Visualization Reveals Patterns EDA is a process, not a checklist
- Goal: Understand data before formal analysis
- Combine:
  - Visualization
  - Summary Statistics
  - Domain knowledge

<https://medium.com/@WenxinZhang98/exploratory-data-analysis-83bfb1f17dc5>

<https://dev.to/yankho817/exploratory-data-analysis-edaultimate-guide-174d>



From <https://medium.com/@WenxinZhang98/exploratory-data-analysis-83bfb1f17dc5>

# EDA Minsdet

- Ask questions before hypothesis
- Expect surprises
- Iterate continuously
- Let the data guide you

**What does a typical observation looks like?**

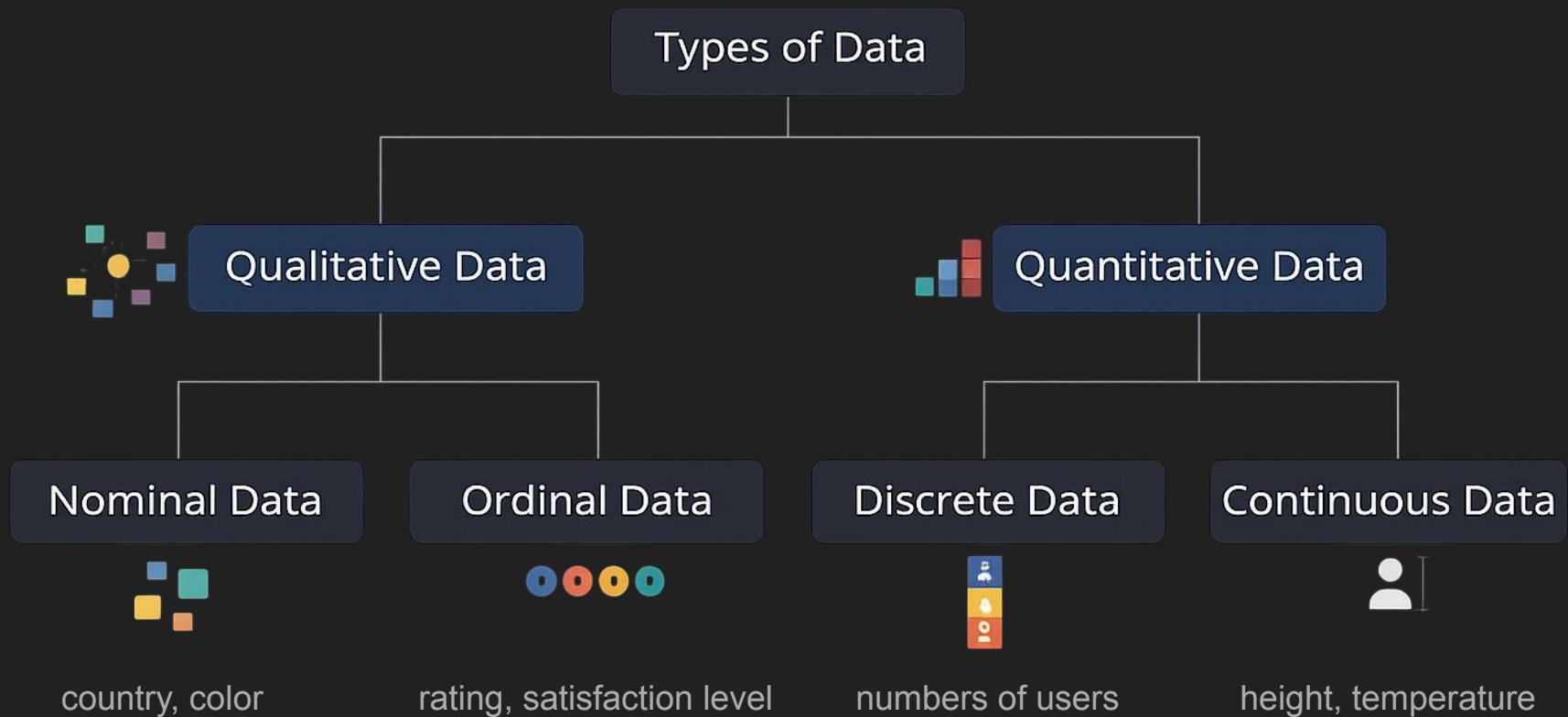
**How variable is the data?**

**Are there unusual values?**

**Are relationships present?**

**Are there data quality issues?**

# Variable Types



# Plots

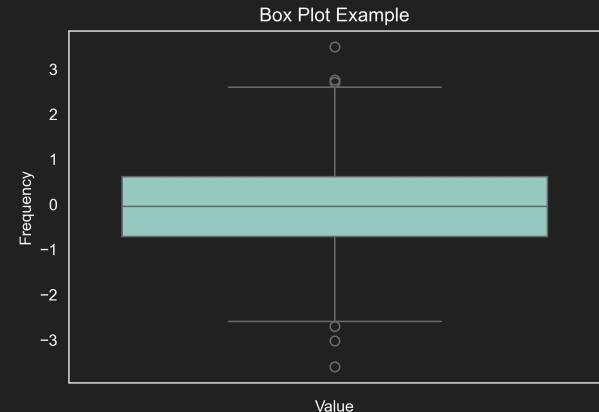
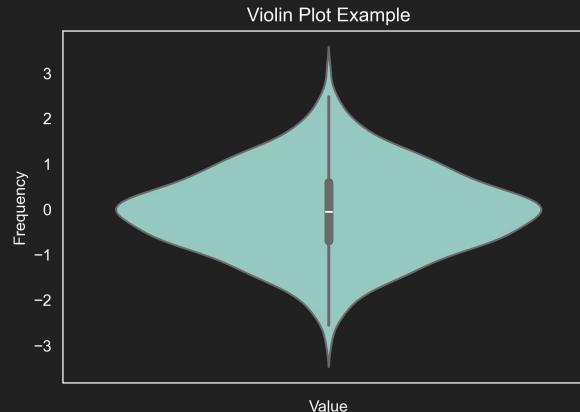
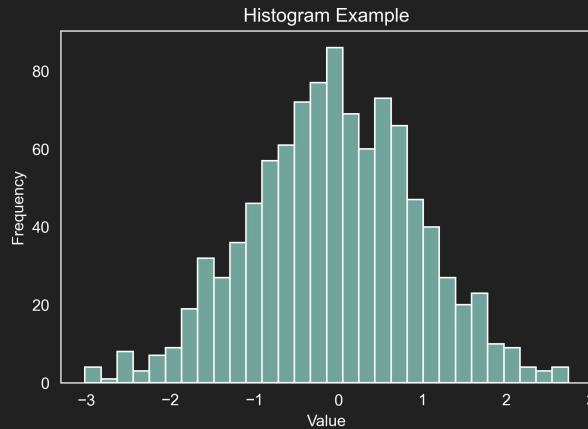
- Visuals reveal patterns numbers can hide
- Help detect:
  - Weird measures
  - Data distributions
  - Relationships between variables
- Essential for communicating the results



# Plot: single quantitative variables

We can use **histograms** to visualize single **quantitative variables**. **Violin plots** or **boxplots** can also be used to visualize **continuous variables**.

Careful note: histogram widths and violin kernel widths can hide fine details in the data

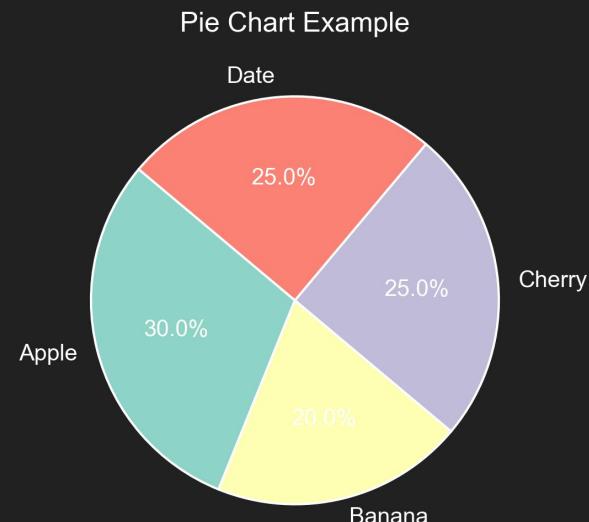
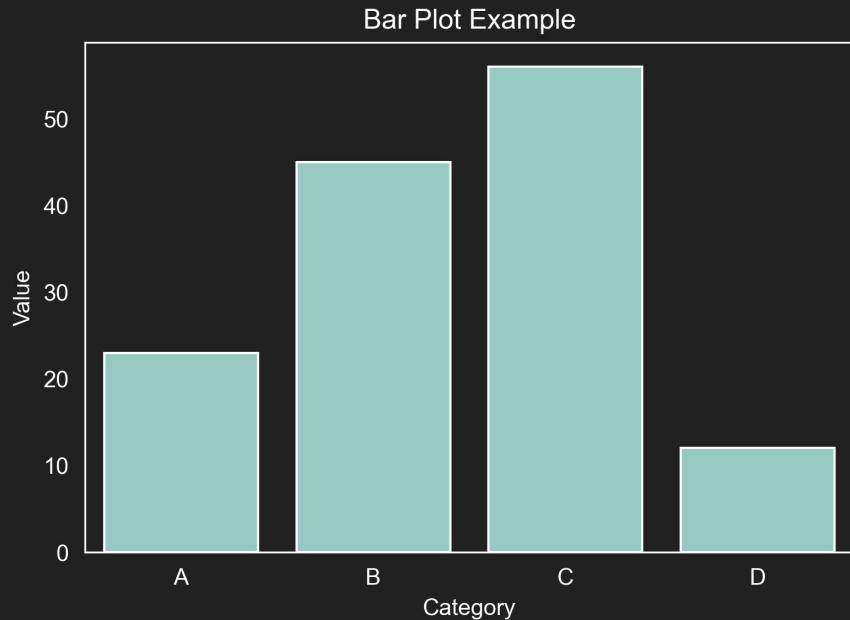


+Raw

+Summarized

# Plot: single qualitative variables

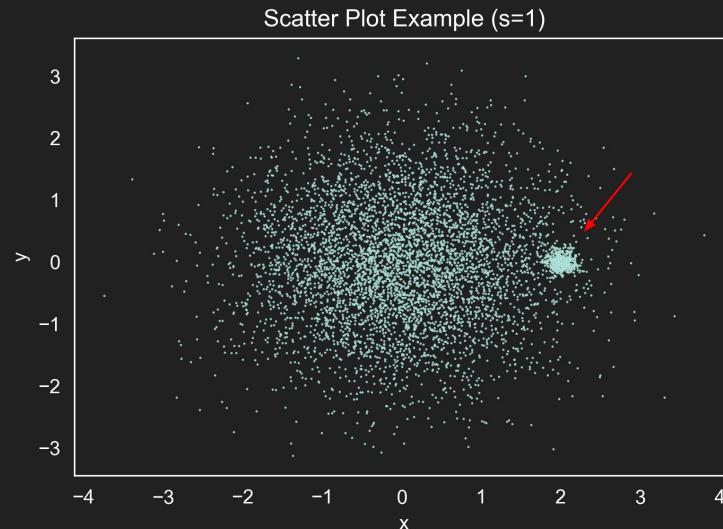
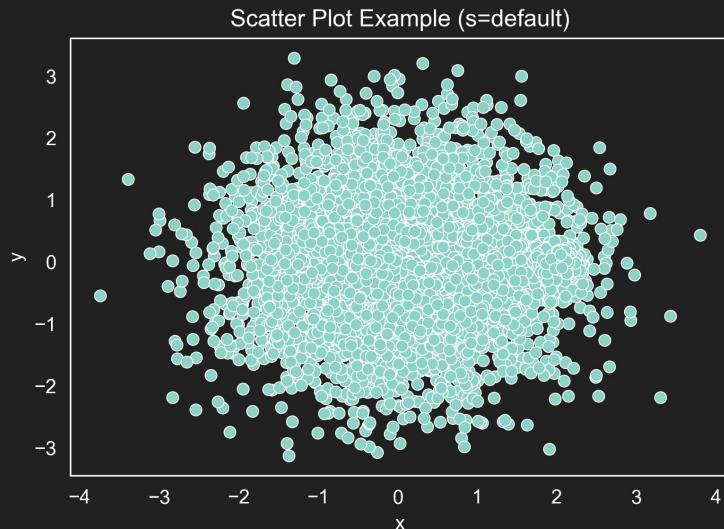
We can use **barplots** or **pie plots** to visualize single qualitative variables.



# Plot: relations quantitative vs. quantitative

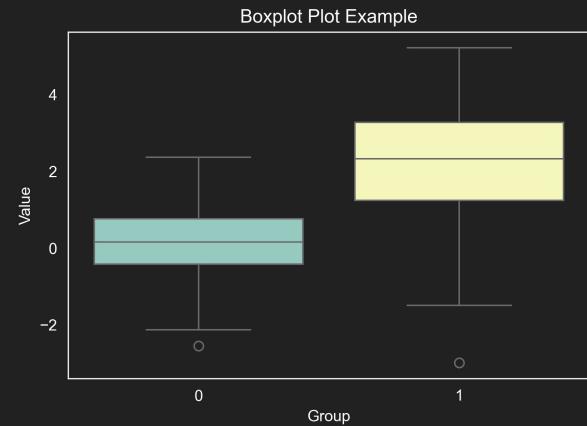
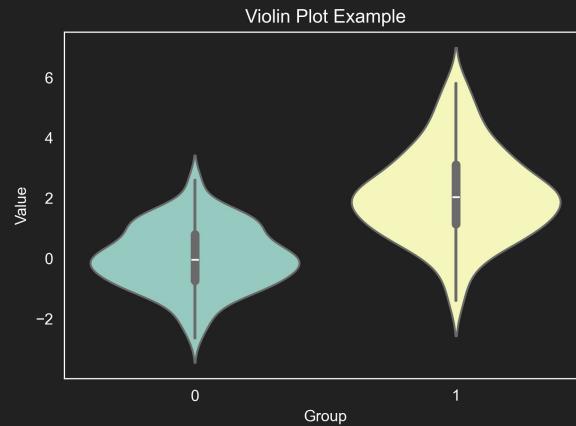
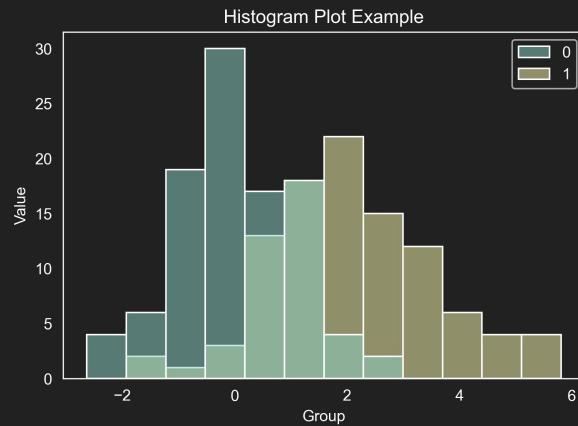
We can use **scatterplots**.

**Careful note:** dot sizes might hide patterns in the data if the size is not adjusted correctly.



# Plot: relations qualitative vs. quantitative

We can use **histograms**, **violin plots** or **boxplots** for the quantitative data separated by qualitative variables.



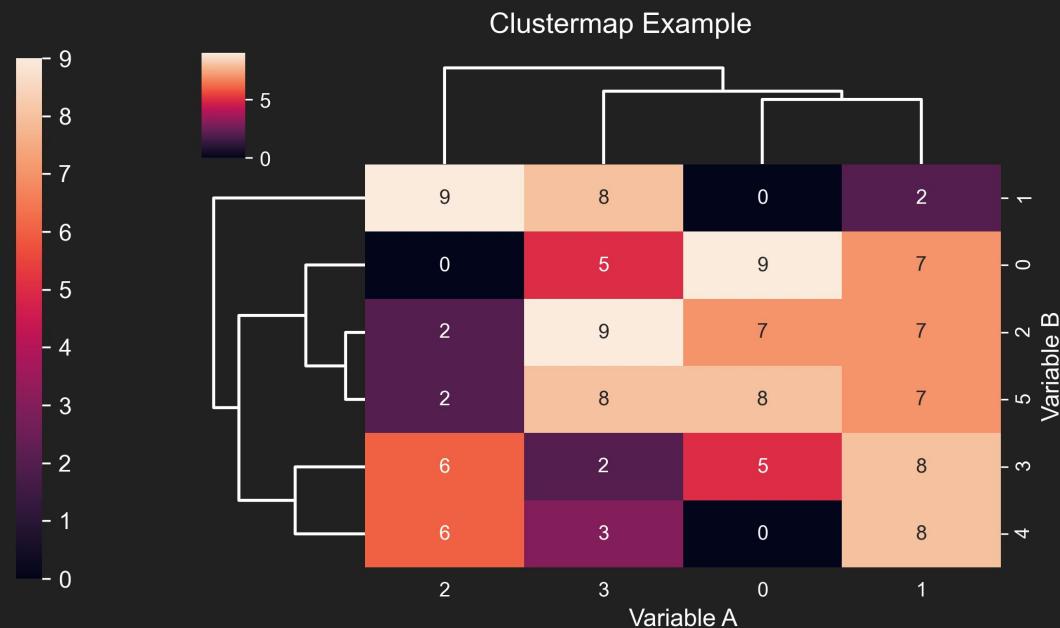
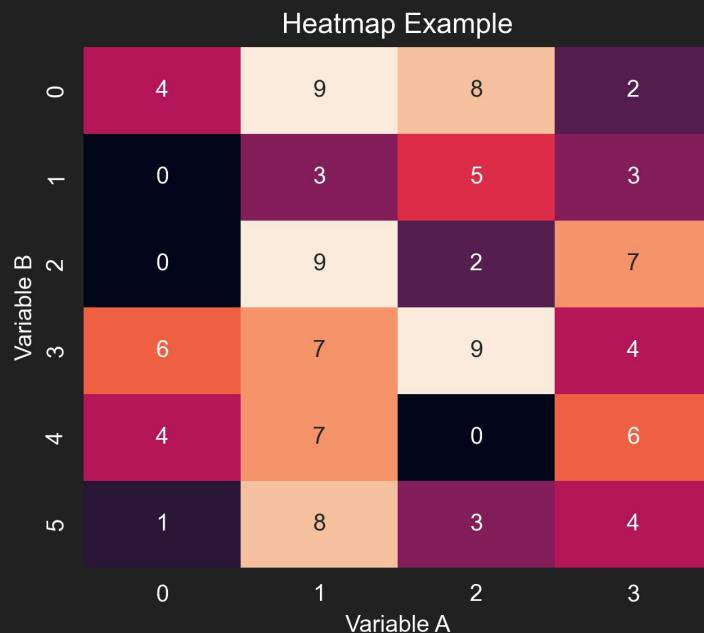
+Raw

+Summarized

# Plot: relations qualitative vs. qualitative

We can use **heatmaps** or **clustermaps**

Careful note: in general, only use clustermap axis with nominal data variables

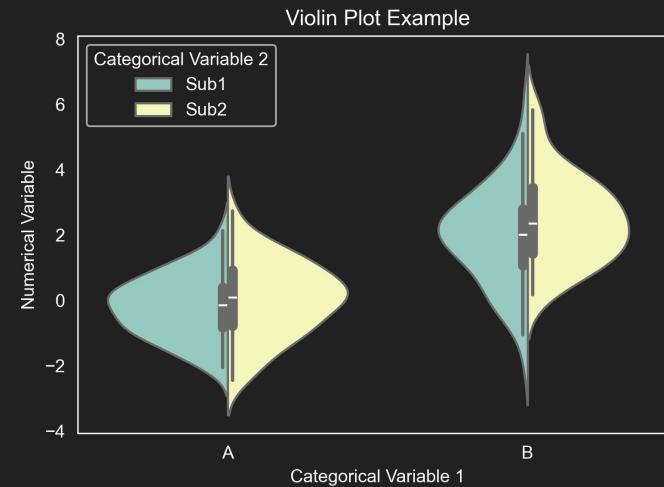
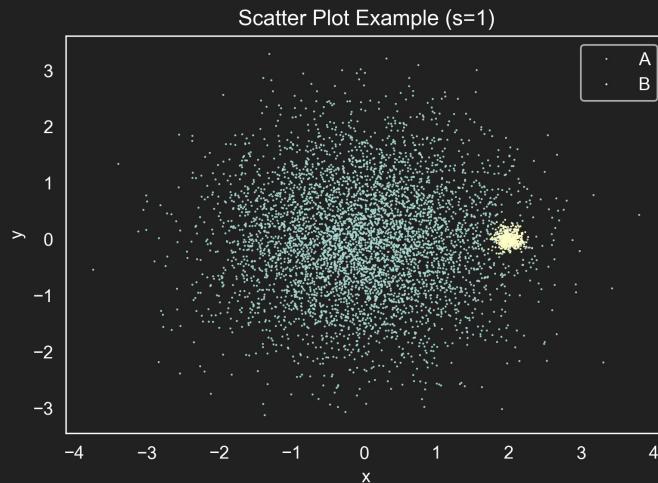


# Plot: 3 variable relationships

We can use **colors** on **scatter plots** for 2 numerical and 1 categorical variable

We can use further splitting in **violin plots** or **bar plots** for 1 numerical and 2 categorical variables.

**Careful note:** try to not overcomplicate plots, always think about the purpose



# Plot: the plotting zoo

# Best practices

- Label axes clearly
  - Use appropriate scales
  - Use color intentionally
  - Always ask: “*What question does this plot answer?*”



<https://seaborn.pydata.org/examples/index.html>

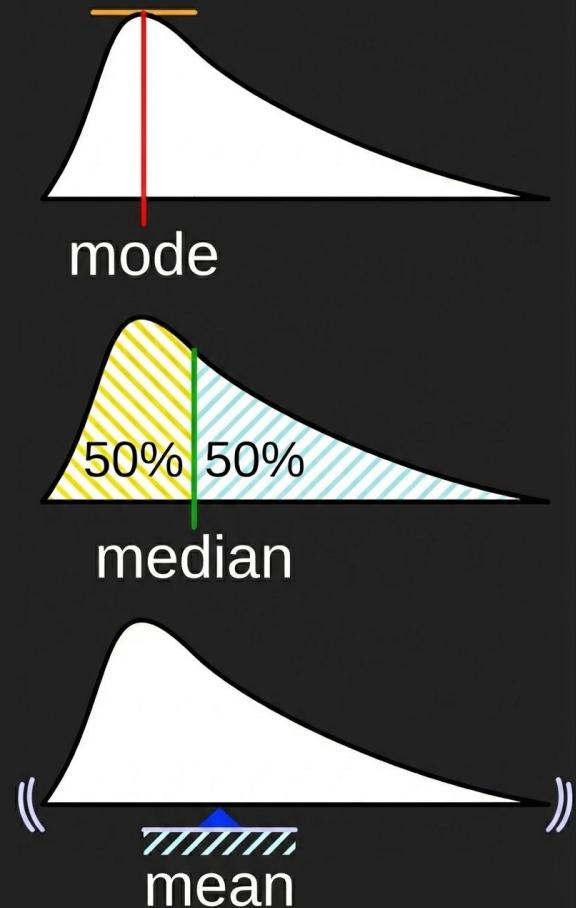
# Summary Statistics

- Connect visual intuition with quantitative interpretation
- Types of summaries:
  - Central tendency
  - Variability
  - Shape
  - Extremeness
- Robustness to outliers
- If you have many variables, easier to see than individual plots

# Summary Statistics: Central tendency

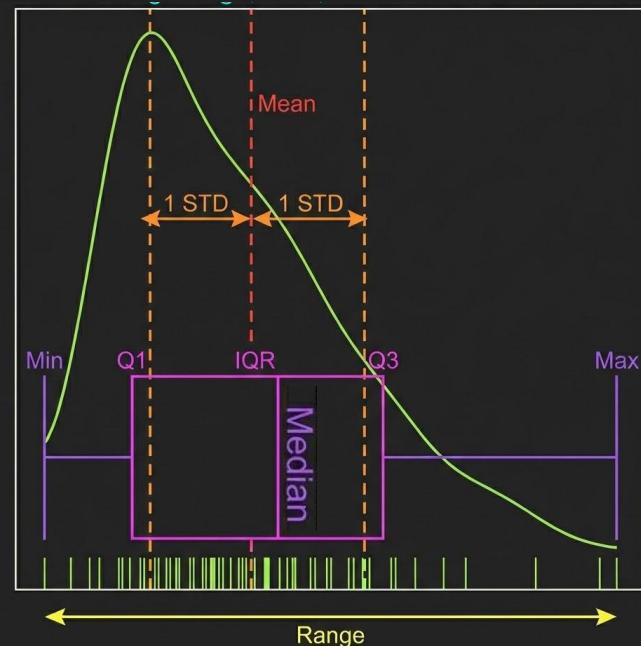
- **Mean:** The average of all observations. Sensitive to outliers.
- **Median:** The middle value when data are ordered. Robust to outliers and skewed distributions.
- **Mode:** The most frequent value in the dataset. Useful for categorical data and multimodal distributions. Not defined for continuous variables\*

\* You can binarize the data (e.g. histogram) and compute the mode in the discrete data



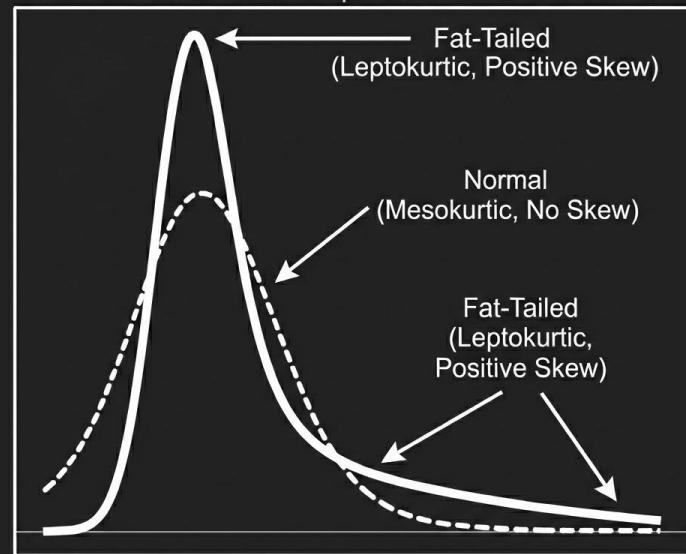
# Summary Statistics: Spread

- **Range:** The difference between the maximum and minimum values. Highly sensitive to outliers.
- **Variance and Standard Deviation (STD):** Variance is the average-squared deviation from the mean. Measures overall variability, but is in squared units. STD is the square root of the variance. Represents typical distance from the mean in original units.
- **Percentile** The value below which a given percentage of observations fall.
- **Quartiles:** Values that divide the data into four equal parts. ( $Q_1 = 25\%$ ,  $Q_2 = \text{median}$ ,  $Q_3 = 75\%$ ).
- **Interquartile Range (IQR):** The difference between the 75th and 25th percentiles. Robust measure of spread.



# Summary Statistics: Shape

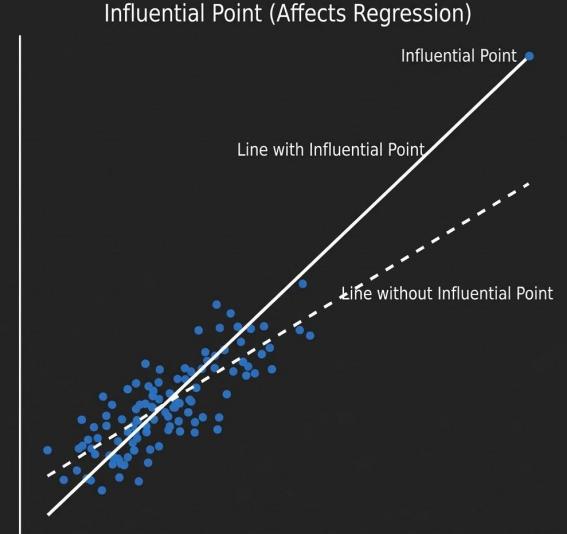
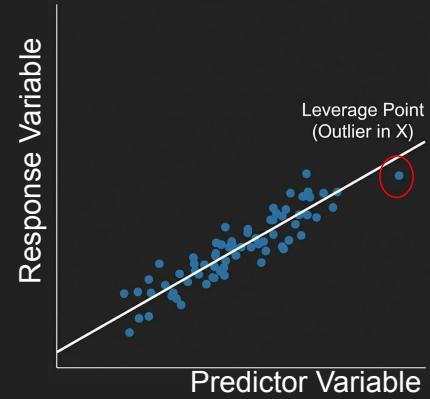
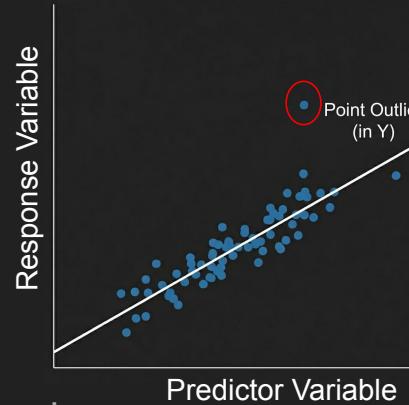
- **Skewness:** Quantifies asymmetry of a distribution. Positive skew: long right tail. Negative skew: long left tail
- **Kurtosis:** Describes tail heaviness and peak sharpness relative to a normal distribution. High kurtosis → heavy tails and extreme values.



Normal:	Kurtosis=0, Skewness=0
Fat-Tailed:	Kurtosis > 0, Skewness > 0

# Outliers I

- **Definition:** An observation that deviates markedly from the rest of the data.
- **Types of outliers:**
  - **Point Outlier:** An unusually large or small value in the response variable.
  - **Leverage Point:** An observation with an unusual value in a predictor variable.
- **Influential Point:** An observation that substantially affects statistical results or model estimates.
- **Always make the question:** Is an error or rare-but-real?



# Missing data

- **Definition:** A missing value for a measure and some variables.
- **Questions to ask:**
  - Why we have missing data?
  - Can we work with that missing information?

ID	Date	Value_A	Value_B	Category
1	2022-02-17	NaN		NaN
2	2022-02-18	1.5	0.7	Chang
3	2023-03-19	NaN	NaN	Flying
4	2023-04-20	NaN	8.5	Category
5	2023-04-21	2.0	NaN	Cheet
6	2023-05-30	3.2	NaN	Cheet

# Association and Correlation

# Index

- Types of relations between variables
- Correlation metrics as summaries
- Correlation is not a relationship
- Partial Correlation
- Simpson's Paradox
- Correlation vs. Causation
- Why to bother?

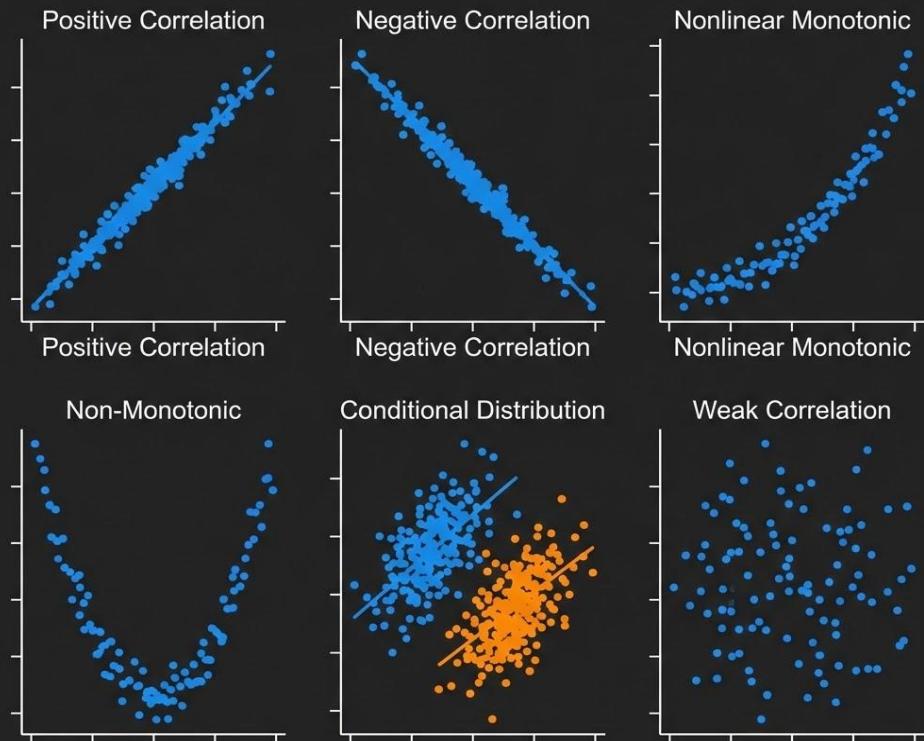
## Learning Outcomes:

- Choose appropriate correlation measures
- Pitfalls of correlation measures
- What correlation measure
- Understand how conditioning changes relationships

# Types of relations between variables

Two variables are related if knowing one gives information about the other.

- Linear or non-linear
- Monotonic or non-monotonic
- Positive, negative or conditional
- Strong or weak



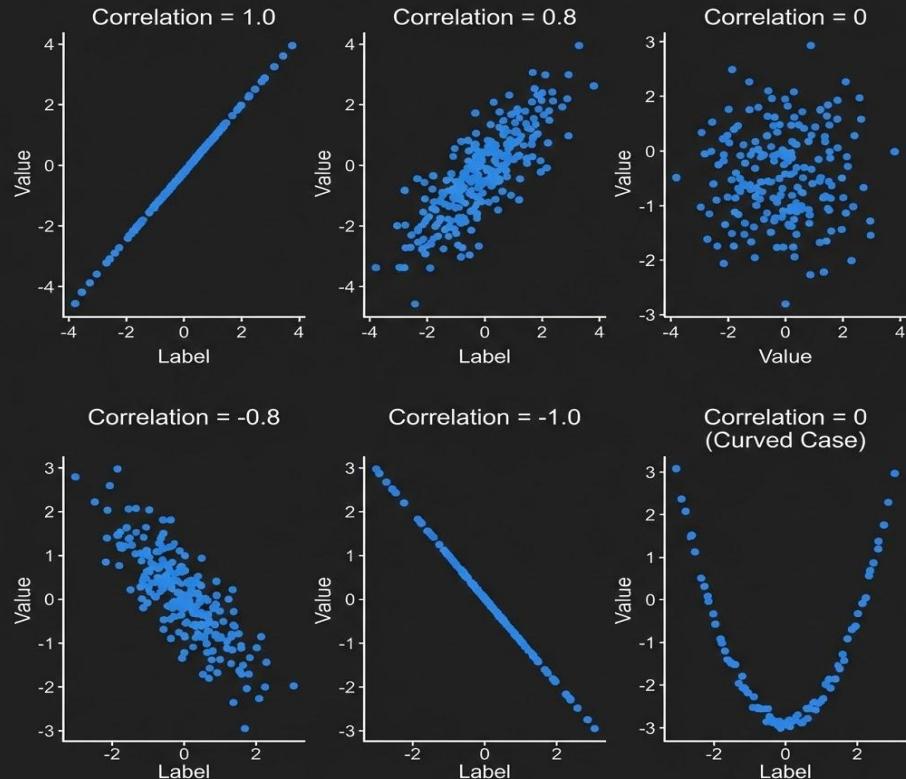
# Correlation metrics: Pearson

Pearson correlation is a descriptive measure of **linear association** between **two continuous variables**. It answers the question:

**Do these two variables move together linearly?**

Properties:

- Measures **linear relationships**
- Ranges from
  - 1: Positive linear relationship
  - 0: No linear relationship
  - -1: Negative linear relationship
- Sensitive to outliers



# Correlation metrics: Spearman

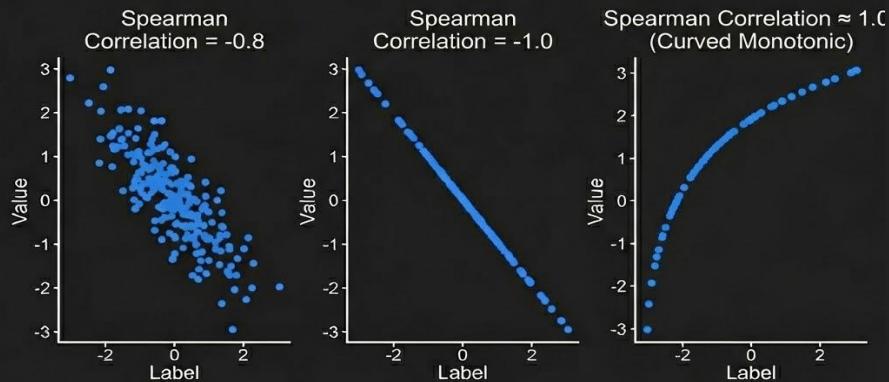
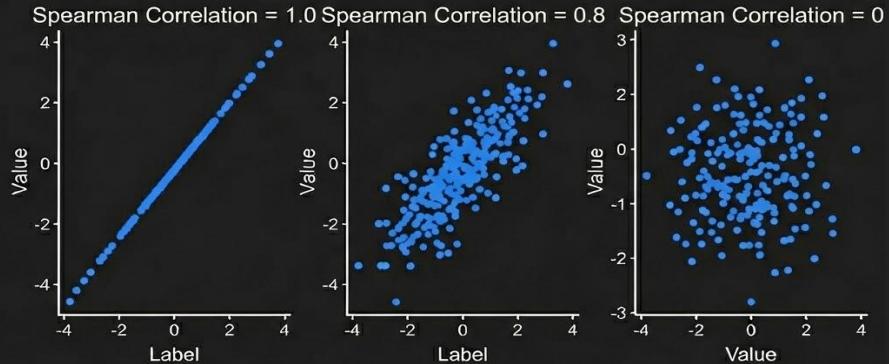
Spearman correlation measures **monotonic association** between two variables.

It answers a different question than Pearson:

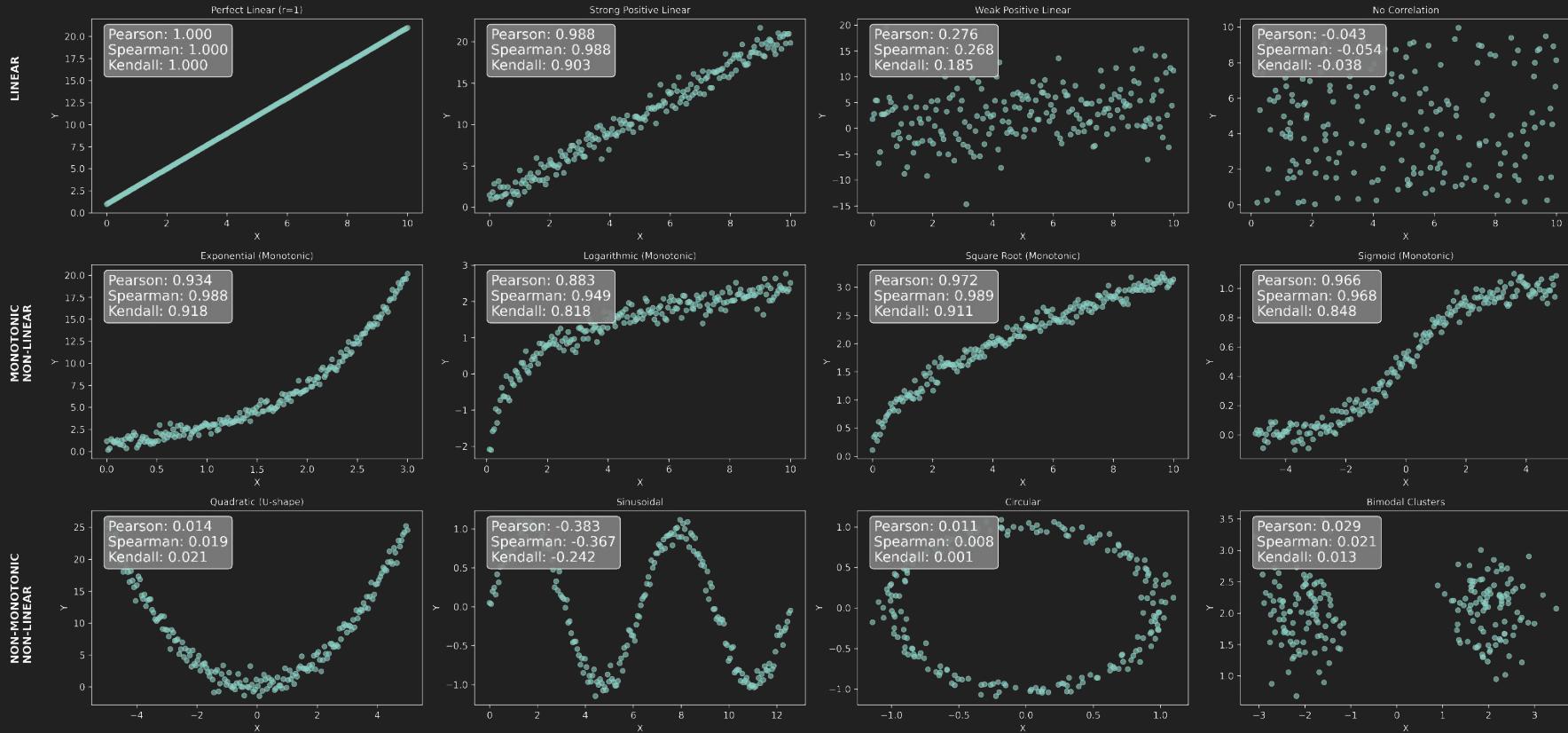
**As one variable increases, does the other tend to increase or decrease — consistently, even if not linearly?**

Properties:

- Measures **linear relationships**
- Ranges from
  - 1: Positive linear relationship
  - 0: No linear relationship
  - -1: Negative linear relationship
- Insensitive to outliers
- Invariant under invariant transformations



# Correlation metrics: Comparison



# Correlation metrics: Contingency Tables

A contingency table summarizes **how often combinations of categorical variables occur**

They are not a single number but a relation between categories

Should look at:

- Rows with different distribution across columns
- Rows with different distribution across rows
- Cells that are unexpectedly large or small

Uniform colors indicate small gains



# Correlation metrics: Cramér's V

Cramér's V is a **measure of association strength** between **two categorical variables**.

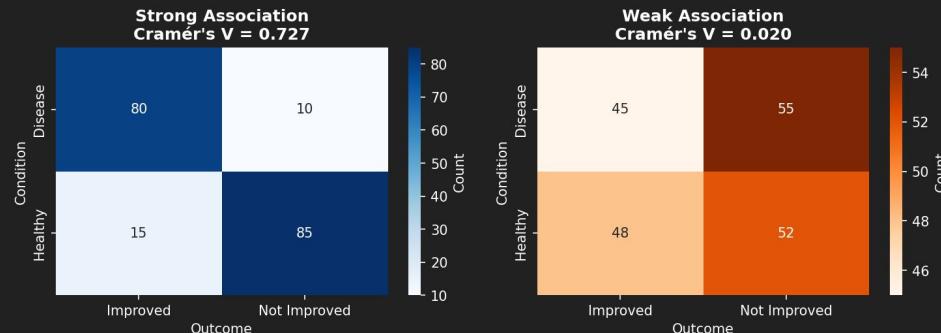
It answers one question:

**How strongly are these two categorical variables associated in the observed data?**

Cramér's V is computed from a **contingency table**. So it is a way of summarizing how uneven a table is compared to what uneven the table would look like.

Properties:

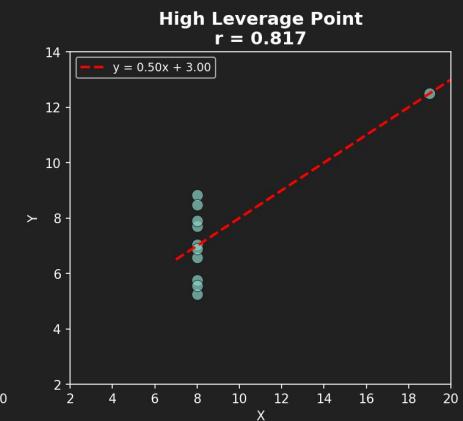
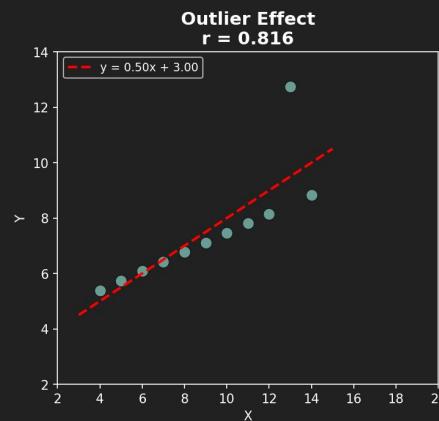
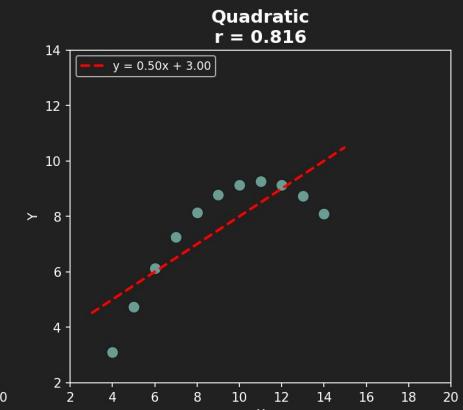
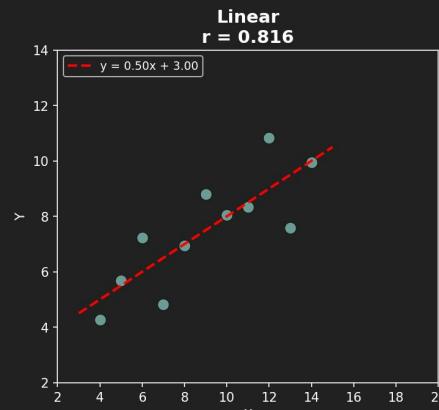
- Scale between 0 (no association) and 1 (perfect association)



# Correlation is not the same as Relationship

- Correlation values do not uniquely identify structure.
- Correlation *is* useful — just not as a verdict.
- It is good for:
  - Detecting potential linear or monotone structure
  - Comparing strength across many variable pairs
  - Flagging candidates for deeper analysis
  - Summarizing patterns *after visualization*
- **If you report a correlation without a plot, you haven't described the relationship.** Correlation is a shadow of the relationship — useful, but never the object itself.

Same Pearson coefficient,  
different associations

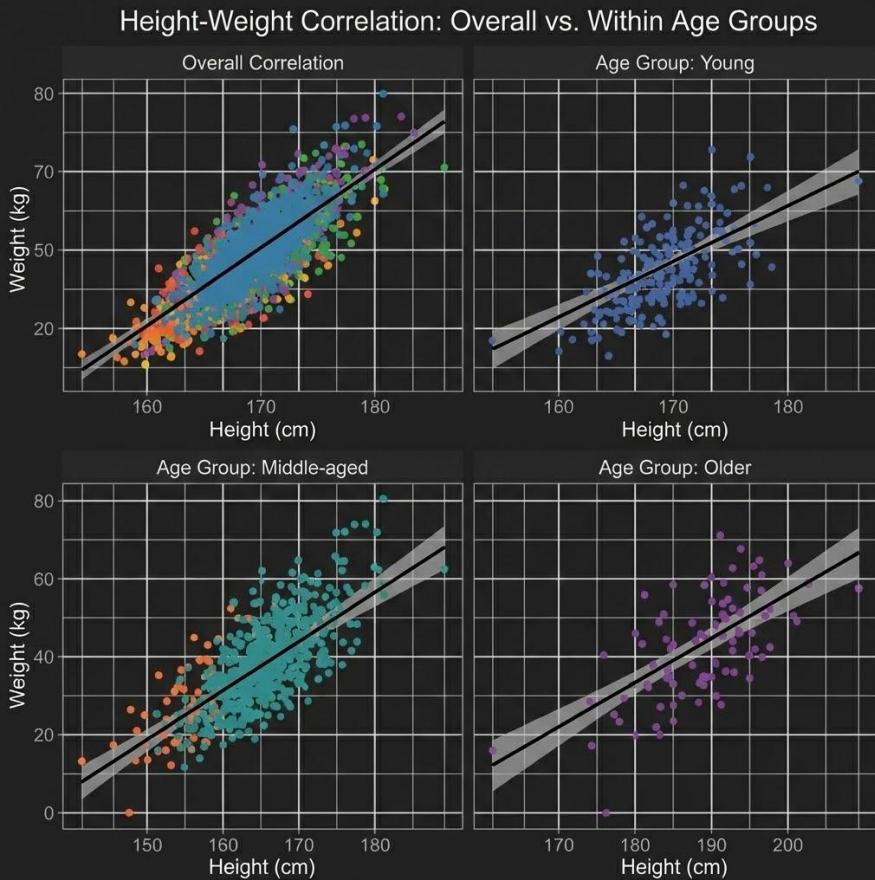


# Conditional Correlation

**Conditional correlation** asks whether two variables still are correlated once we look only at situations where a **third variable is the same**.

e.g. weight and height are correlated overall — are they also correlated when we compare people of the same age?

- **Aggregation:** pooling all data together to compute a single summary measure. Assumes:
  - All observations are comparable
  - Differences between groups do not matter
  - Mixing doesn't distort meaning.
- **Stratification:** separating the data by condition and computing measures within each group. It compares like with like, instead of mixing different situations.



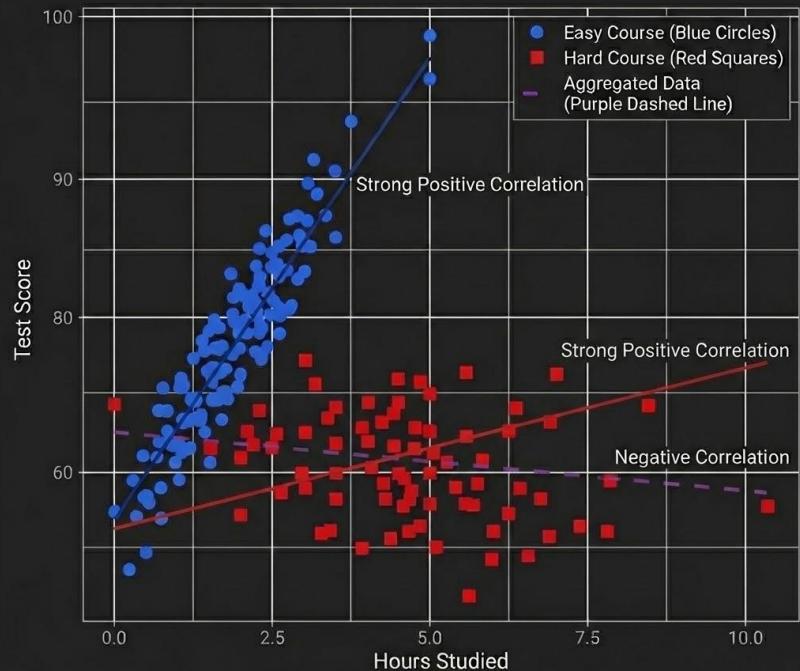
# Simpson's paradox

**Simpson's paradox happens when aggregation hides patterns that stratification reveals.**

e.g. **Overall**, test score and hours of study appear negatively related. **Within each difficulty level**, more study is associated with higher scores.

Simpson's paradox shows that aggregation can mislead — **not that stratification is always correct.**

**Always ask: am I mixing different situations?**



# Correlation is not Causality

Simpson's paradox happens when aggregation hides patterns that stratification reveals.

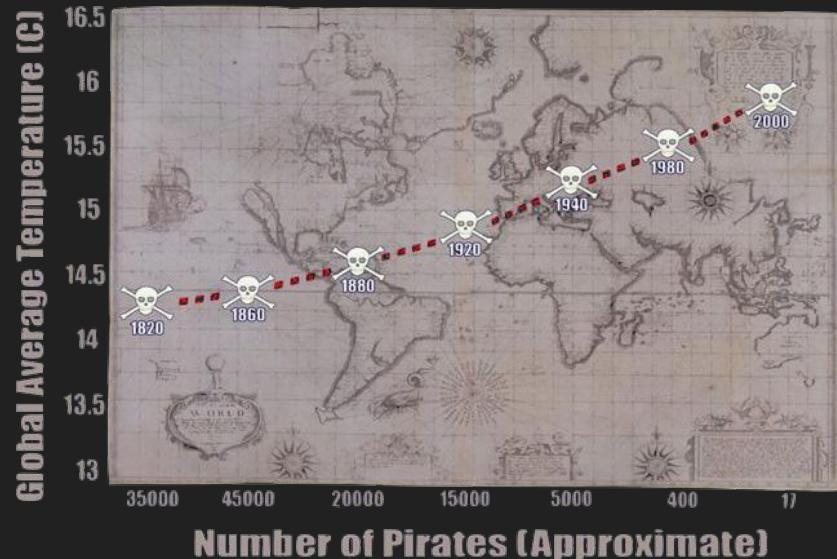
e.g. There is an inverse monotonic correlation in the temperature of the sea and the number of pirates in the ocean.

What if we do conditioning on time as the third variable?  
We would be answering:

"Do pirates and temperature have a trend *within the same year*?"

The causal question is:

"If I changed the number of pirates, would temperature change?"



"Correlation needs context,  
design and humility"

# So why we bother with correlations at all?

Correlation is useful when you want to:

- compare many relationships
- track changes
- summarize consistently
- communicate succinctly
- feed later analysis

Correlation is *not* useful when:

- you haven't plotted
- the relationship is clearly nonlinear
- the number will be misread as causal

Humans are great at: spotting patterns, seeing direction, noticing outliers

Humans are *terrible* at: comparing subtle strength differences, being consistent across plots, remembering magnitudes

Two plots can *look* similar but differ meaningfully in strength. Correlation gives you:

- a stable scale
- a reference point
- a way to say “this is stronger than that”

Not truth — **comparability**

# Study Design

# Index

- Experimental vs. Observational Studies
- Bias
- Randomization and Blocking
- Confounders
- Outliers II: Trimming vs. Winsorizing
- Pre-registering Analysis
- Audit Trails and Data Provenance

## Learning Outcomes:

- Understand the sources of bias
- How to prevent some of them with randomization
- What is a pre-registering analysis
- Audit and analysis

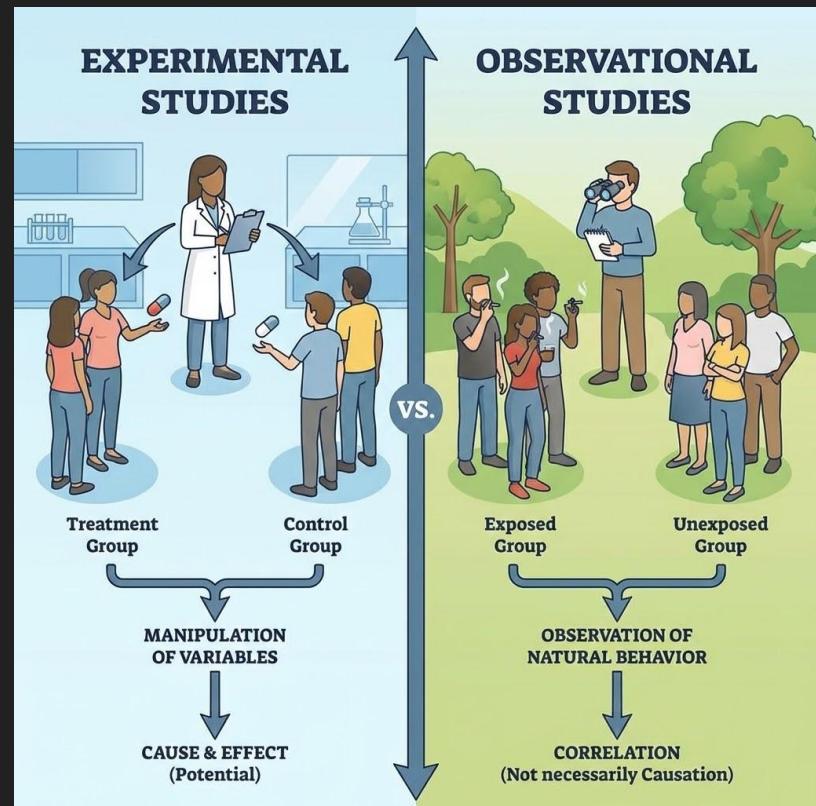
# Experimental vs. Observational Studies

## Experimental studies

- Assignment is under control
- Groups are comparable *by design*
- Randomization breaks many spurious associations

## Observational Studies

- No control over assignment
- Groups may differ in many hidden ways
- Correlation and conditioning must do more work

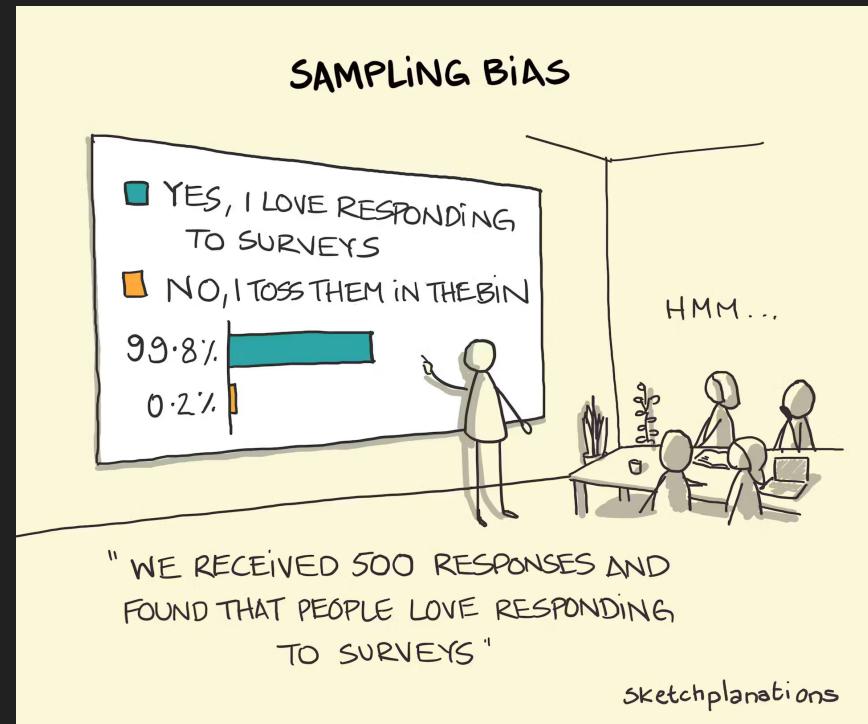


# Bias

**Bias** is **systematic distortion**, not random noise. It does not average out with more data.

Common types:

- Selection bias (who appears in the data)
- Measurement bias (how variables are recorded)
- Survivorship bias (who remains observable)
- Reporting bias (what is missing or censored)



# Randomization and Blocking

**Randomization** and **Blocking** are techniques to balance out **some preconditioning** known and unknown biasing factors.

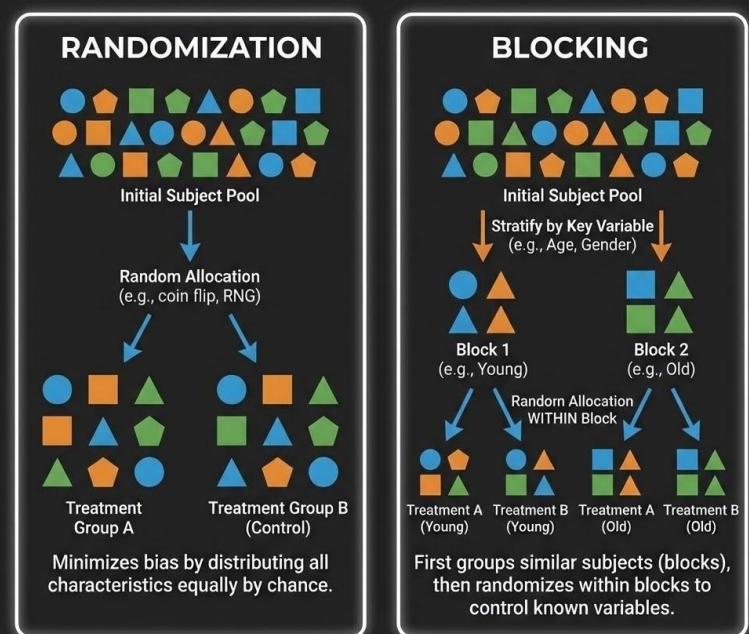
**Randomization** assigns conditions by chance, not choice.

**Blocking** groups similar units **before randomization**.

**It does not solve:**

- measurement bias (bad sensors, bad labels)
- attrition bias (who drops out)
- noncompliance (people not following assignment)

## EXPERIMENTAL DESIGN: RANDOMIZATION VS. BLOCKING



# Randomization and Blocking

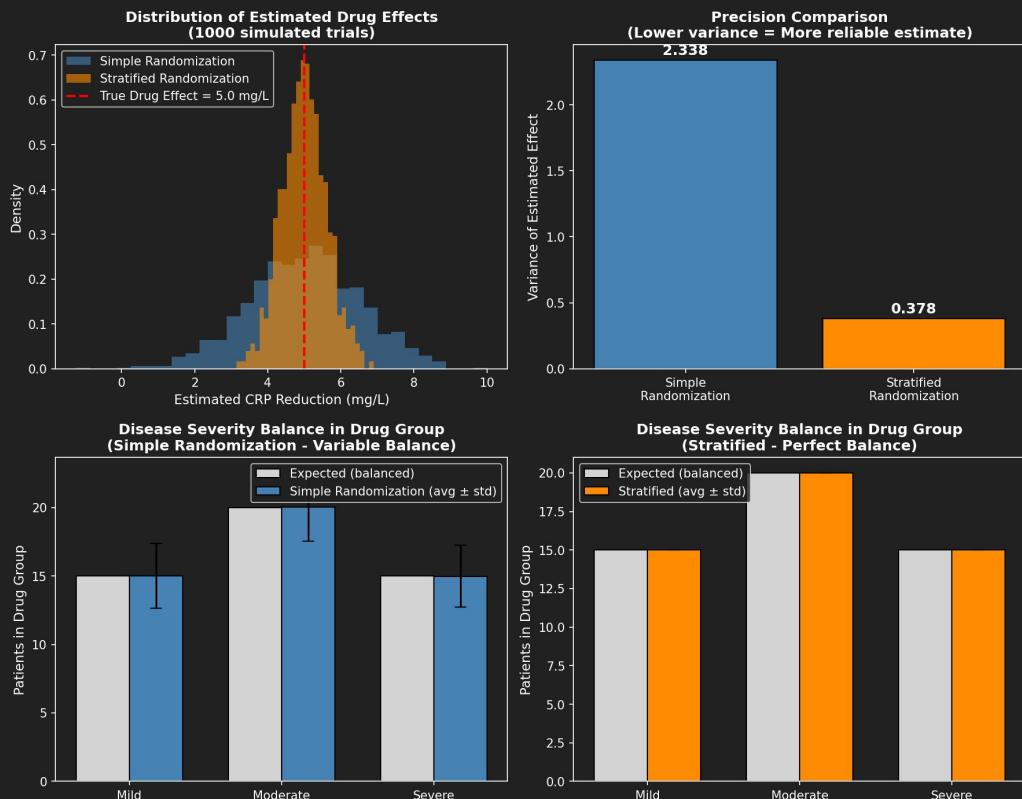
e.g. Scenario: Testing a new anti-inflammatory drug's effect on CRP (C-reactive protein) levels

- **Parameter A (confounding):** Disease severity (Mild/Moderate/Severe) - affects baseline CRP
- **Treatment:** Drug vs Placebo
- **Outcome:** CRP reduction (mg/L)

## Comparisons:

- Simple Randomization: ignores disease severity, leading to potential imbalance
- Stratified Randomization: ensures equal severity distribution in drug/placebo groups

**Key insight:** Stratified randomization (blocking) reduce variance and require smaller sample sizes.



# Confounders

A **confounder** is a type of bias coming by a variable that is related to multiple variables of interest that mix distinct situations together.

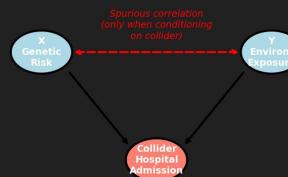
- Produces the Simpson's paradox
- Misleading correlations
- Unstable conditional relationships

Confounding can be mitigated by randomization/blocking by conditioning and for other statistical techniques (we will see them later).

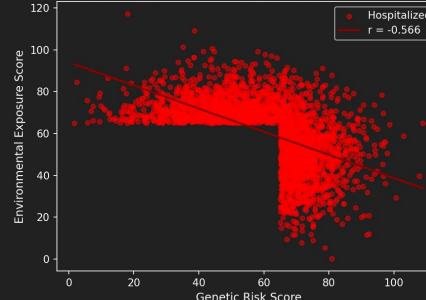
**Care note:** Conditioning on confounders that depend on the tested variables can generate spurious correlations.

Berkson's Paradox: Conditioning on a Collider Creates Spurious Correlations  
X and Y are independent  $\rightarrow$  but appear correlated when we select on their common effect

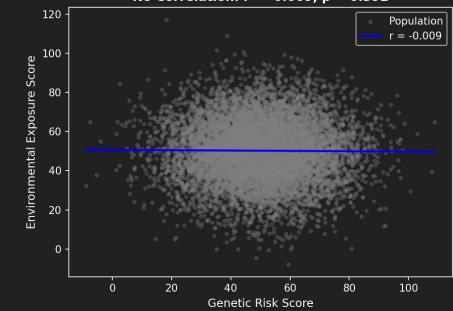
Causal Diagram: Collider Bias



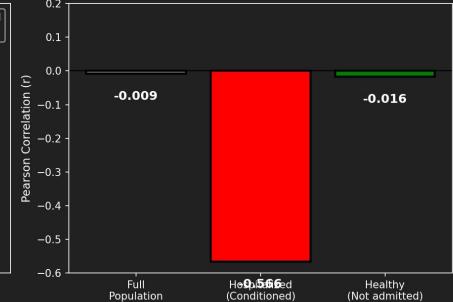
Hospitalized Patients Only (n=2961)  
SPURIOUS Correlation:  $r = -0.566$ ,  $p = 2.32e-250$



Full Population (n=10000)  
No Correlation:  $r = -0.009$ ,  $p = 0.392$



Correlation Between Genetic Risk and Environmental Exposure by Subgroup



# Outliers II: Trimming vs Winsorizing

**Trimming removes extreme observations entirely.**

e.g. drop top and bottom 1% of values

## Pros

- simple
- limits extreme influence
- transparent if documented

## Cons

- reduces sample size
- can bias estimates
- throws away real data
- silently changes the population

**Winsorizing keeps all observations but caps extreme values.**

e.g. values above 99th percentile → set to 99th percentile

## Pros

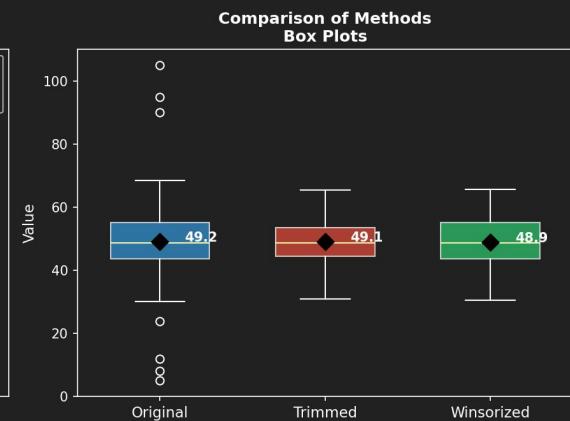
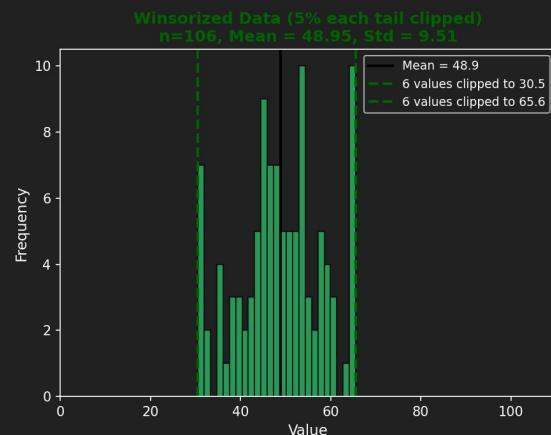
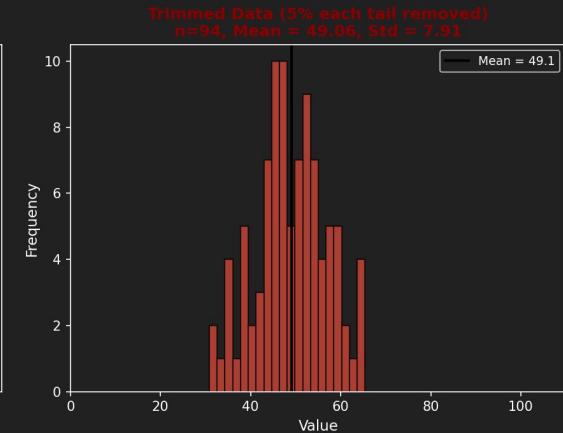
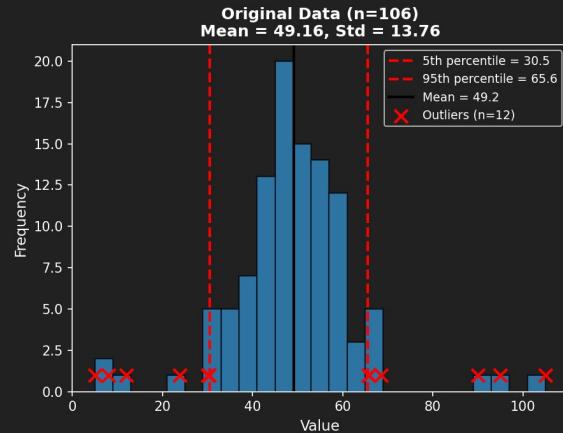
- keeps sample size
- limits leverage of extremes
- often more stable than trimming

## Cons

- alters real values
- hides true extremes
- still arbitrary if unjustified

# Outliers II: Trimming vs Winsorizing

Handling Outliers: Trimming vs Winsorizing  
Trimming removes extremes | Winsorizing clips extremes to boundary values



# Pre-registering Analysis Plans and Cleaning

Pre-registration means deciding what you will do *before* seeing the answers.

That includes:

- which variables matter
- which comparisons you will report
- how you will clean the data
- how you will handle outliers
- which analyses are **confirmatory vs exploratory**

Patterns discovered in exploratory analysis must be validated on data that did not influence their discovery.



# Audit Trails and Data Provenance

**Data provenance:** “Where did this data come from?”

Data provenance means knowing:

- how the data were collected
- by whom
- under what conditions
- with which instruments
- and with which known limitations

If you don't know this:

- you don't know what the data can support
- no amount of statistics can rescue it

**Audit trails:** “What happened to the data?”

An **audit trail** is a complete record of:

- filtering
- exclusions
- transformations
- recording
- imputations
- modeling decisions

This should answer:

**How did raw data become results?**