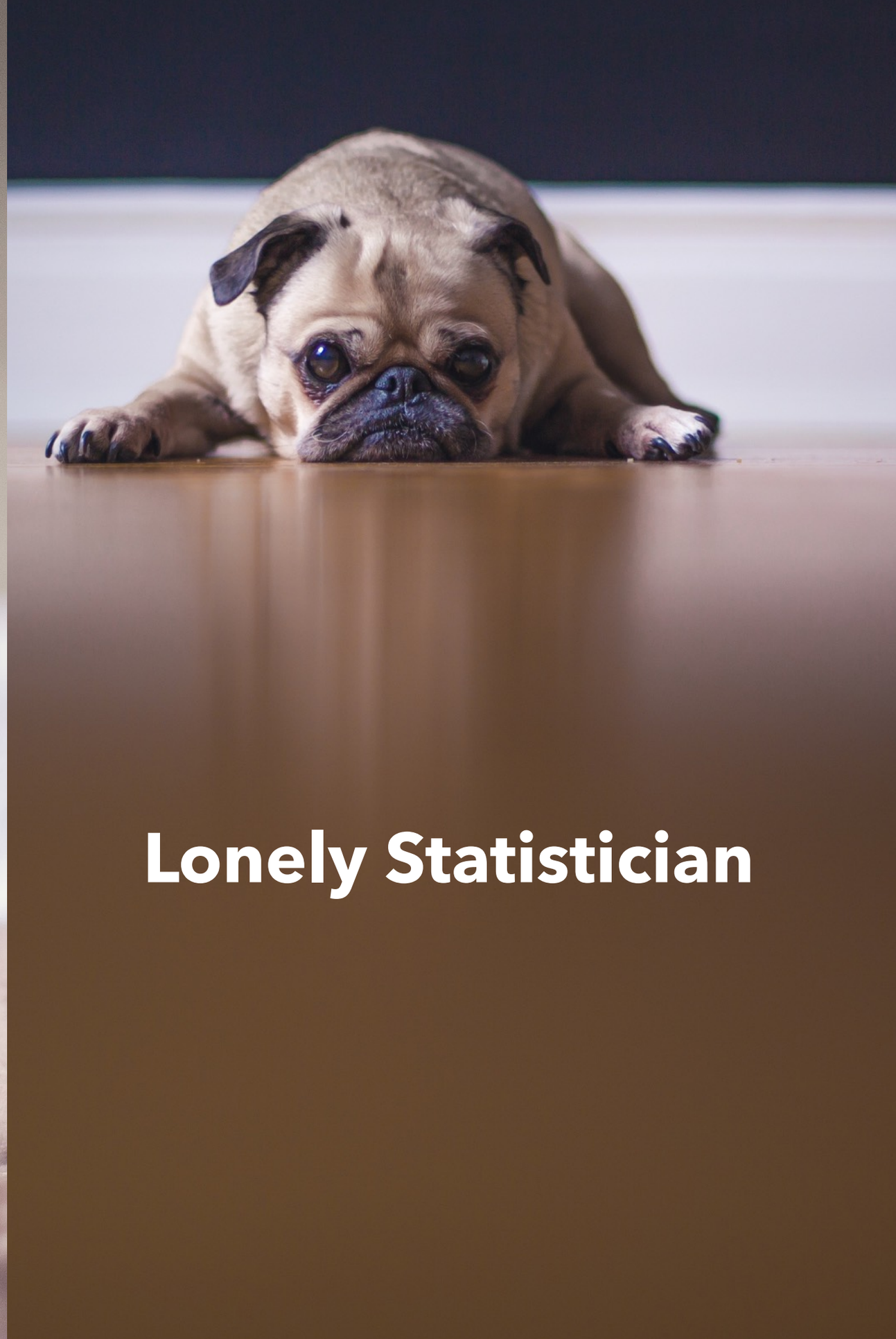
Statistics for Data Science: what you should know and why

Gabriela de Queiroz

Data Scientist and Founder of R-Ladies

Lonely Data Scientist



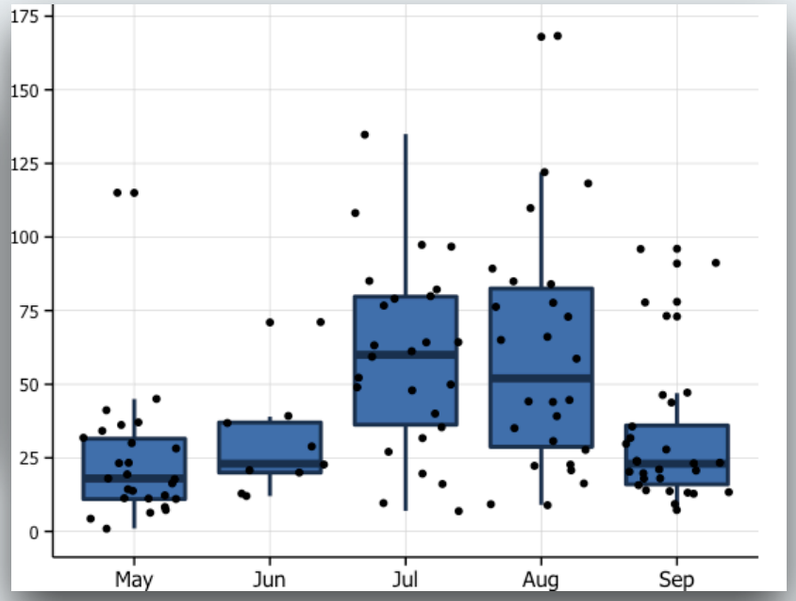
Lonely Statistician

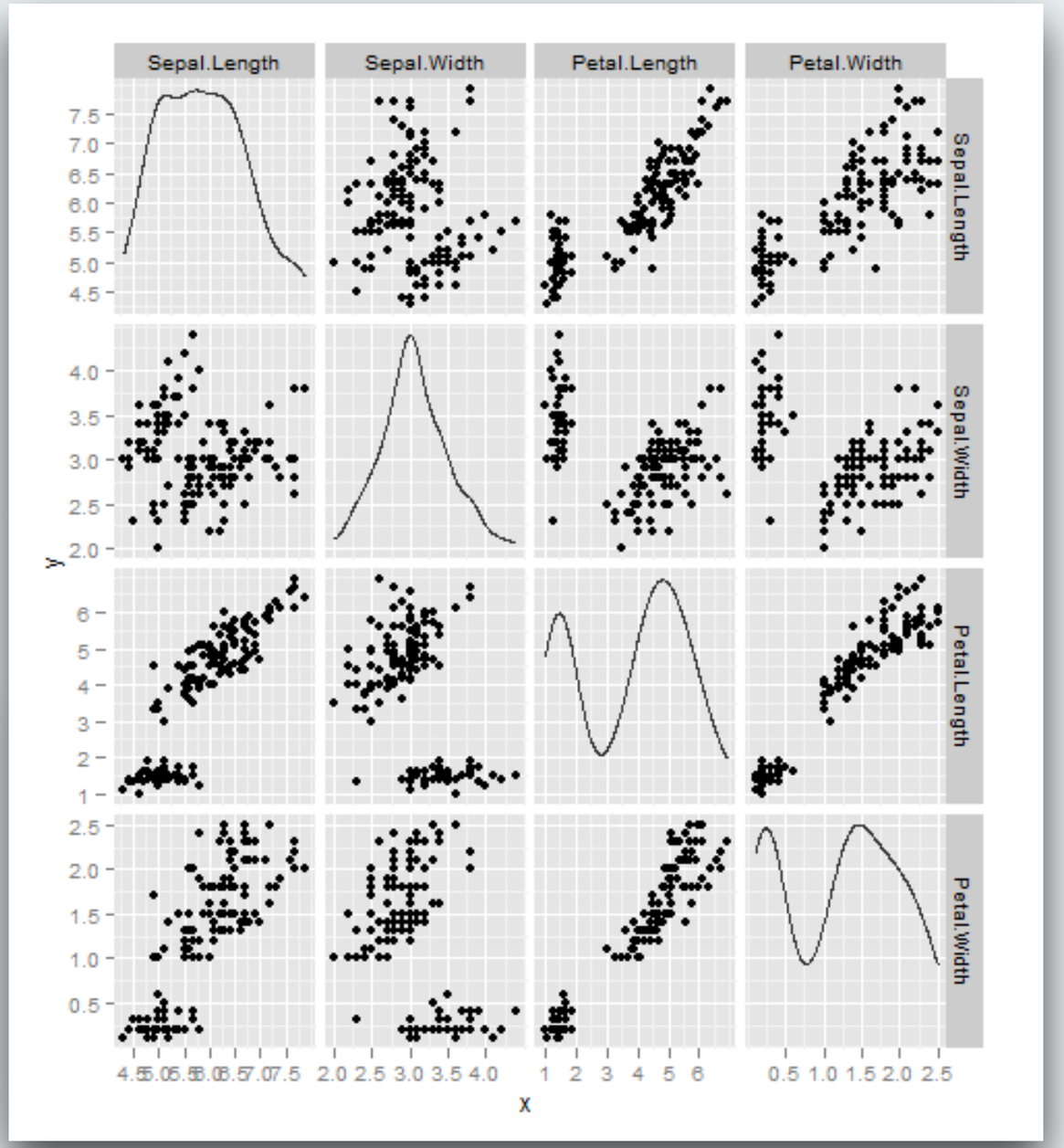
TOP 5 STATISTICAL CONCEPTS

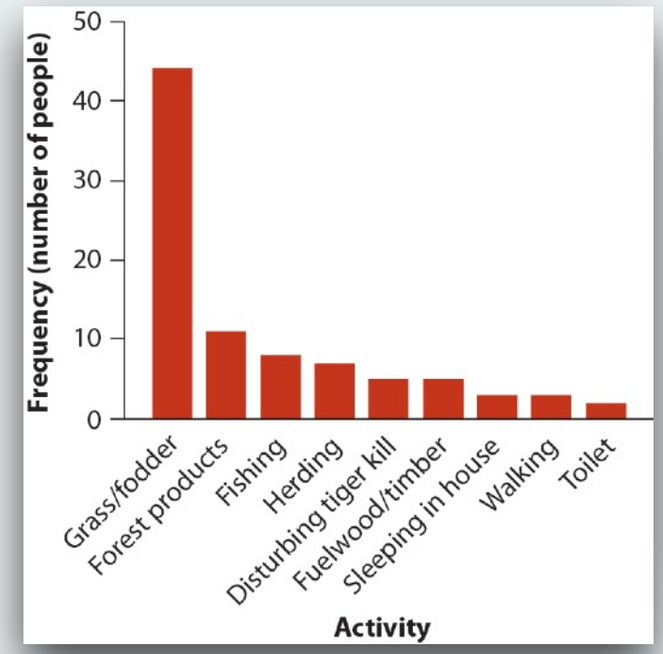
1. Know your data

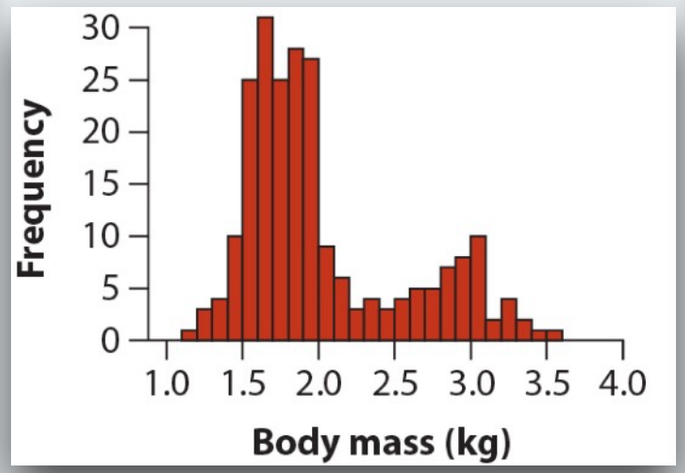
Some ways to know your data

```
Min.   : 1.00
1st Qu.: 3.00
Median : 8.00
Mean   :11.88
3rd Qu.:17.00
Max.   :62.00
```



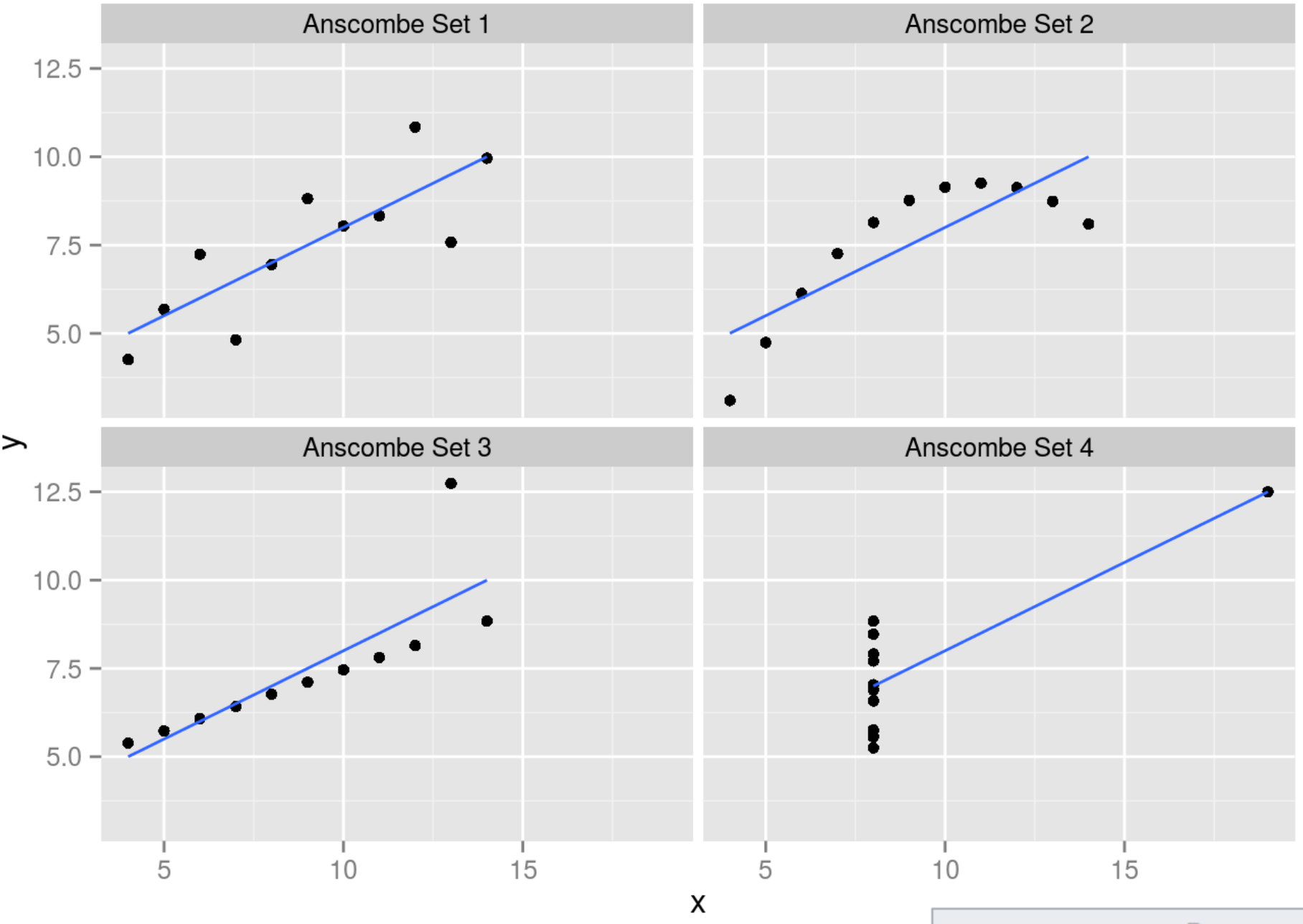






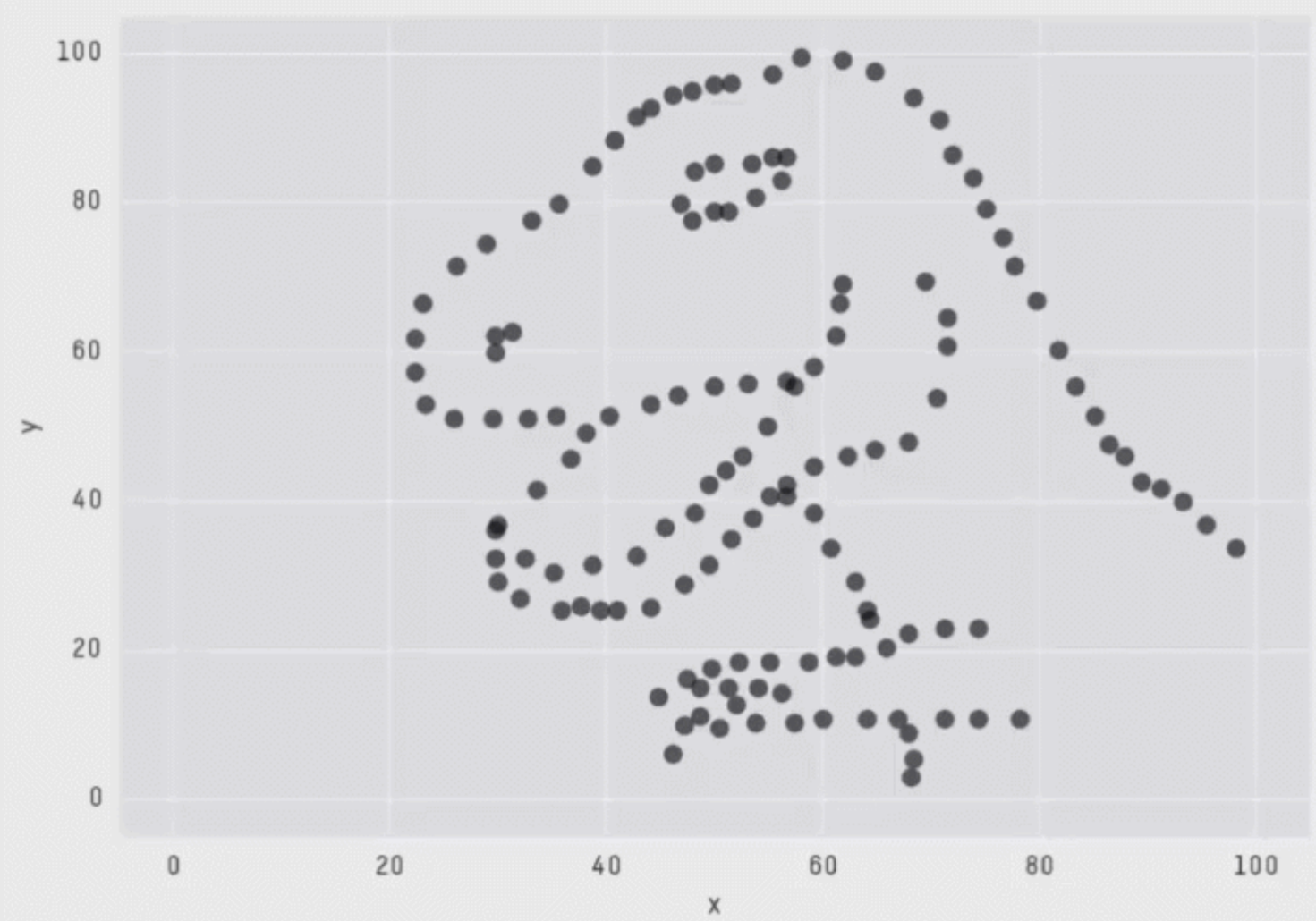
```
   Min.    : 1.00
 1st Qu.:  3.00
 Median  :  8.00
 Mean    : 11.88
 3rd Qu.: 17.00
 Max.    : 62.00
```

Summary Statistics



Anscombe's quartet

Property	Value
Mean of *x*	9
Sample variance of *x*	11
Mean of *y*	7.50
Sample variance of *y*	4.125
Correlation between *x* and *y*	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression	0.67



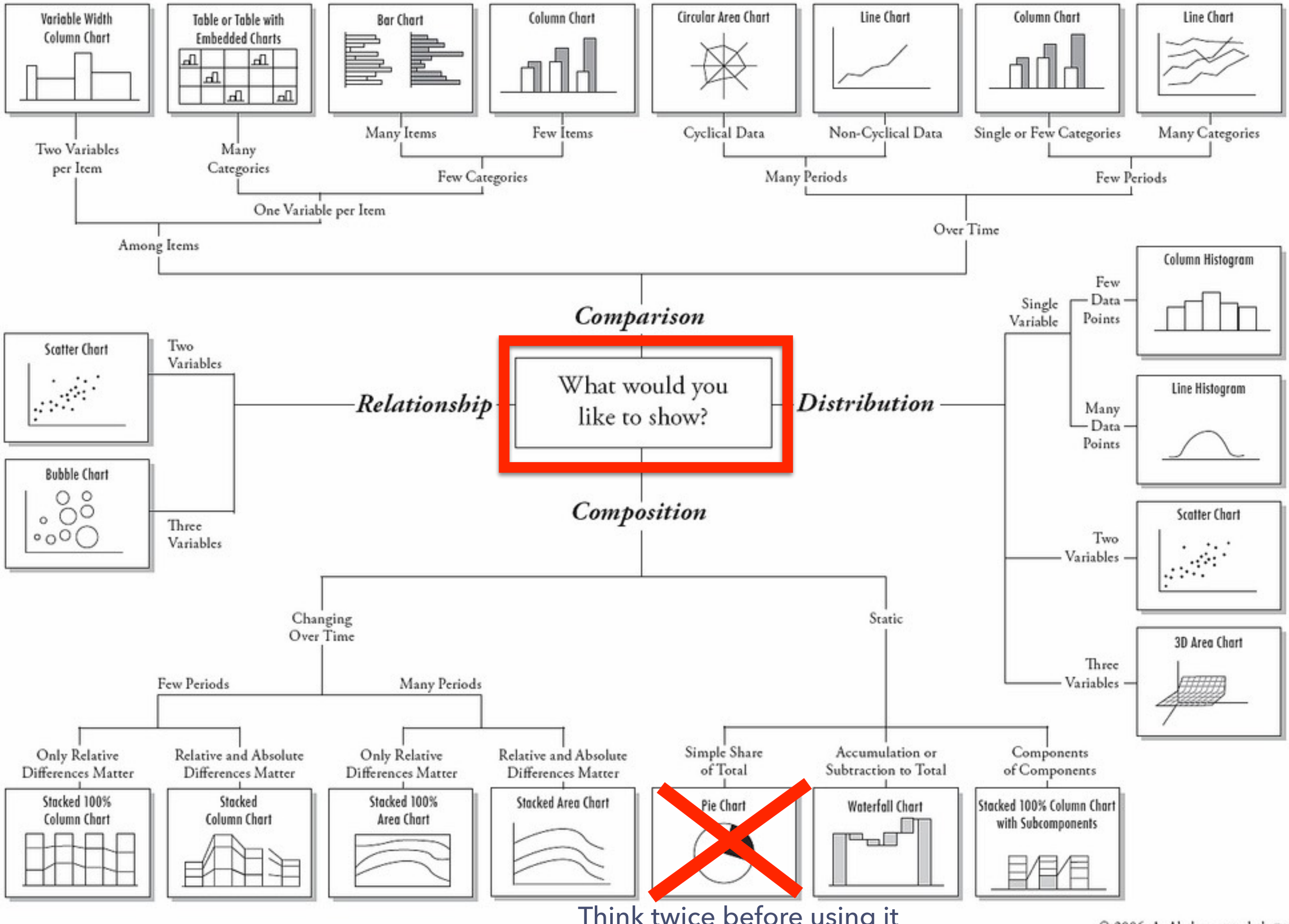
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

The Datasaurus Dozen

AutoDesk Research: https://www.autodeskresearch.com/publications/samestats

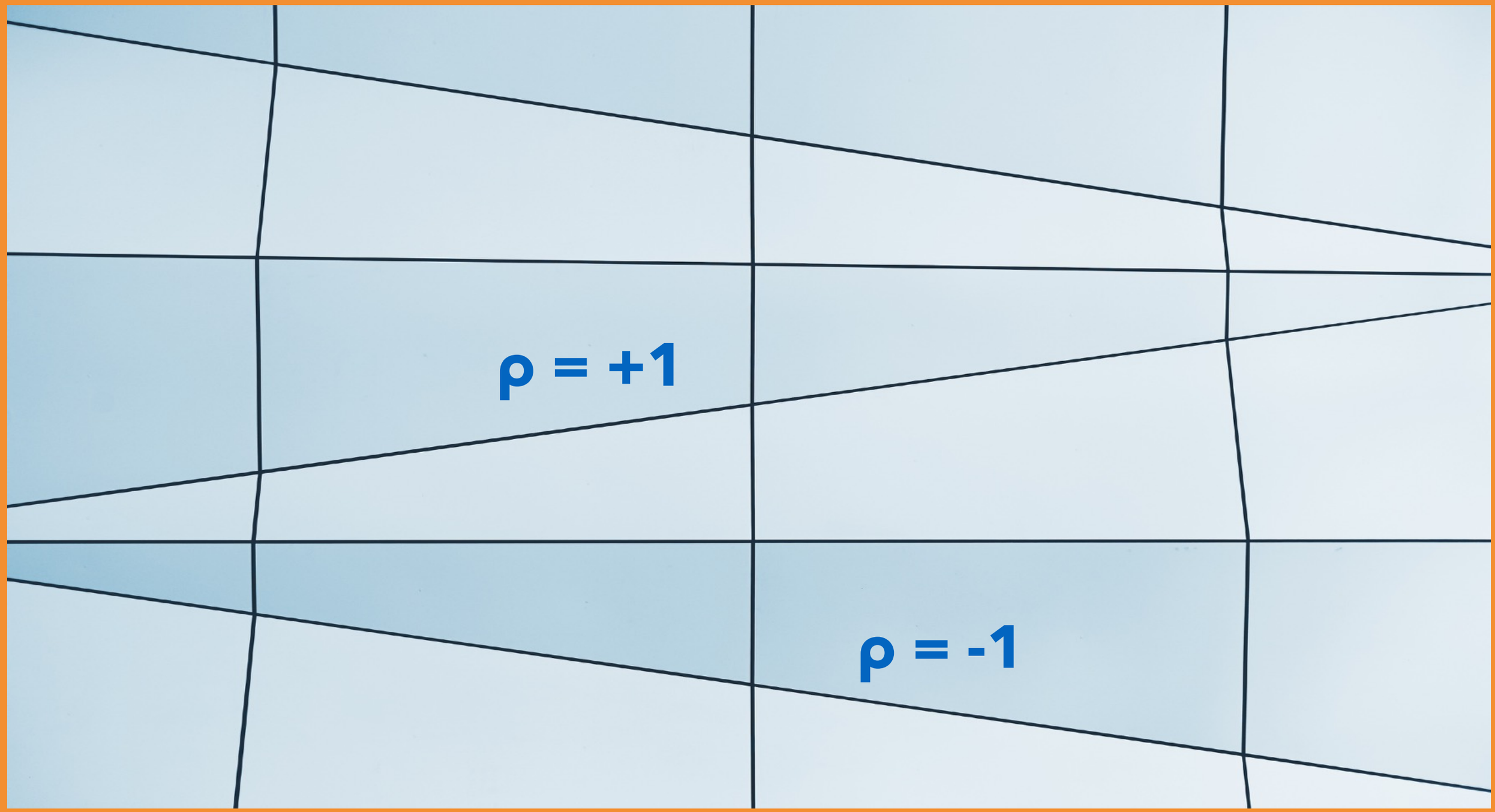
R-package: https://github.com/stephlocke/datasauRus

Chart Suggestions—A Thought-Starter

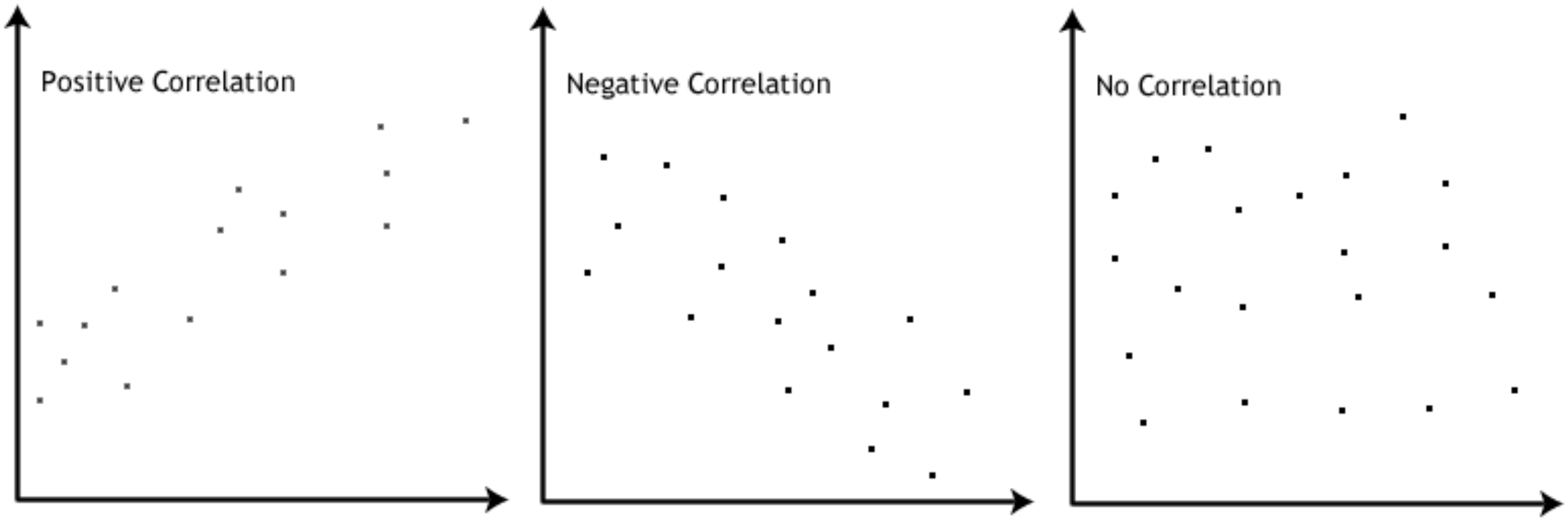


Think twice before using it

2. Correlation*

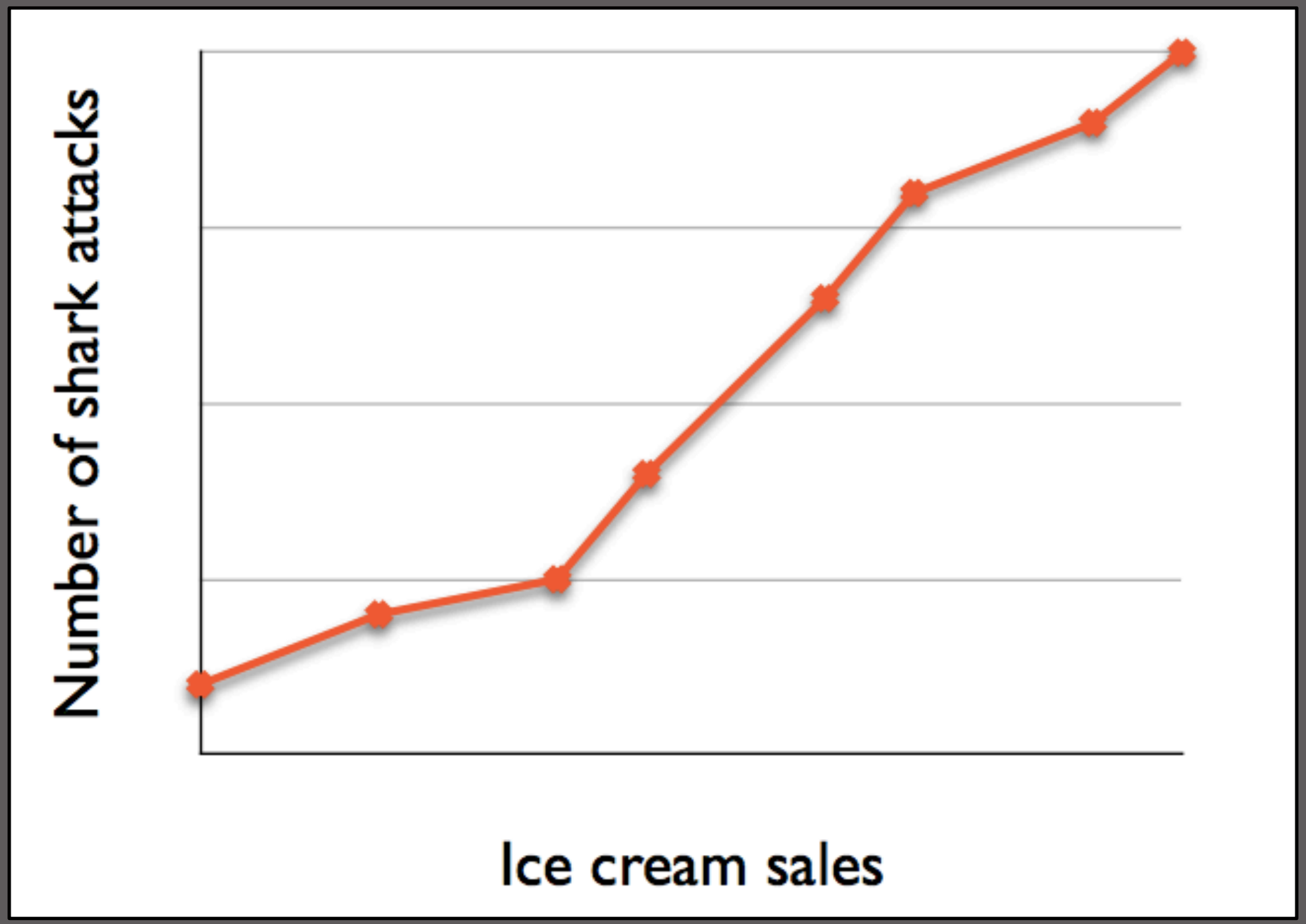


*** Pearson correlation**



Correlation describes the strength of the **linear relationship** between two variables.

What can we say about this chart?



Credits: http://www2.stat.duke.edu/~mc301/ARTSCI101_Su16/post/slides/w2_d2_smoking_research.pdf

ICE CREAM SALES

CAUSE

SHARK ATTACKS?

ICE CREAM SA

CAUSE

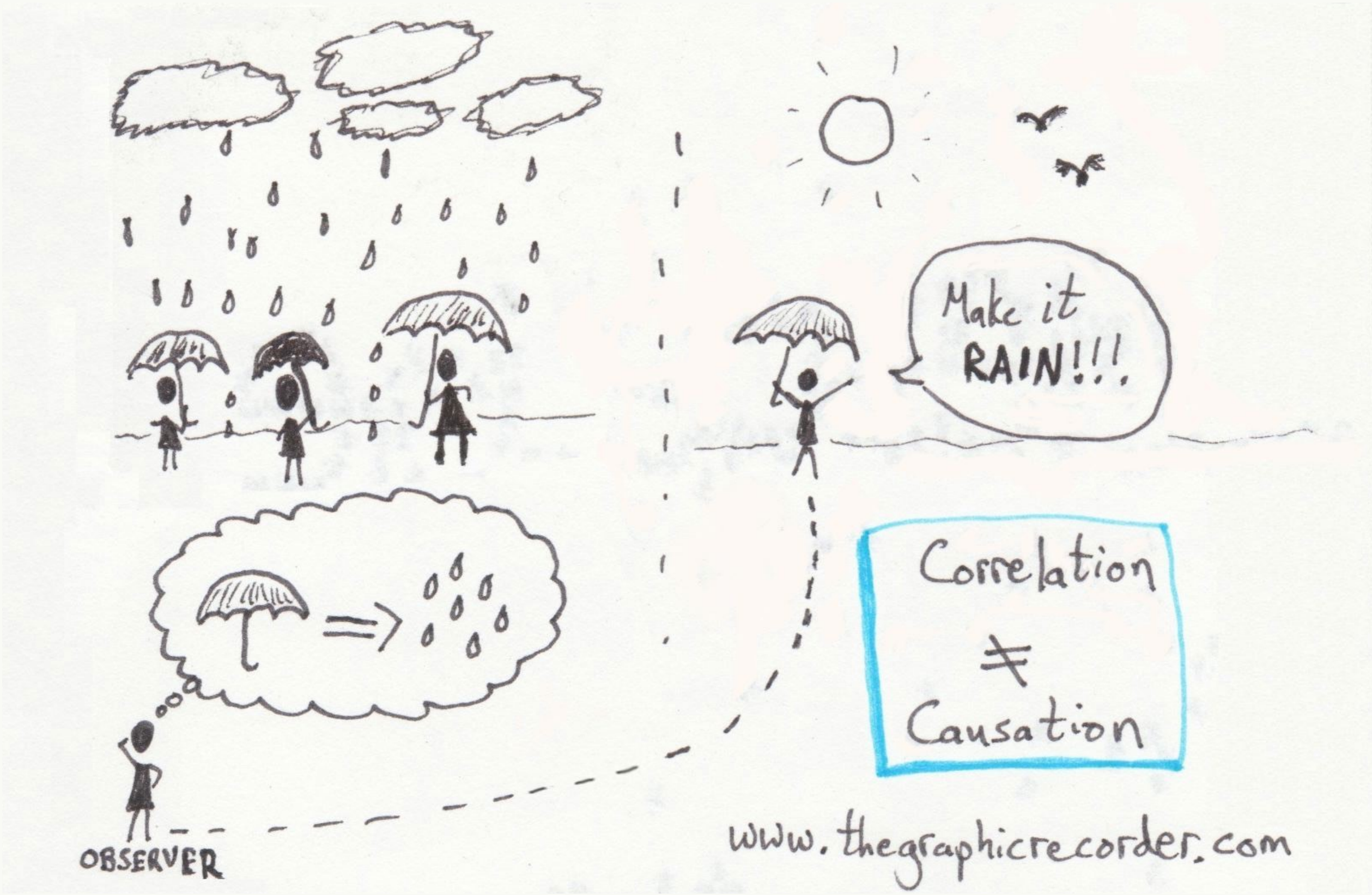
RK ATTACKS?

SUMMER?

observer

umbrella => rain

Where is the rain???



Correlation doesn't imply causation

Causation vs Correlation

- **Causality** indicates that one event is the result of the occurrence of the other event.
- **Correlation** between two things can be caused by a third factor (confounder) that affects both of them.

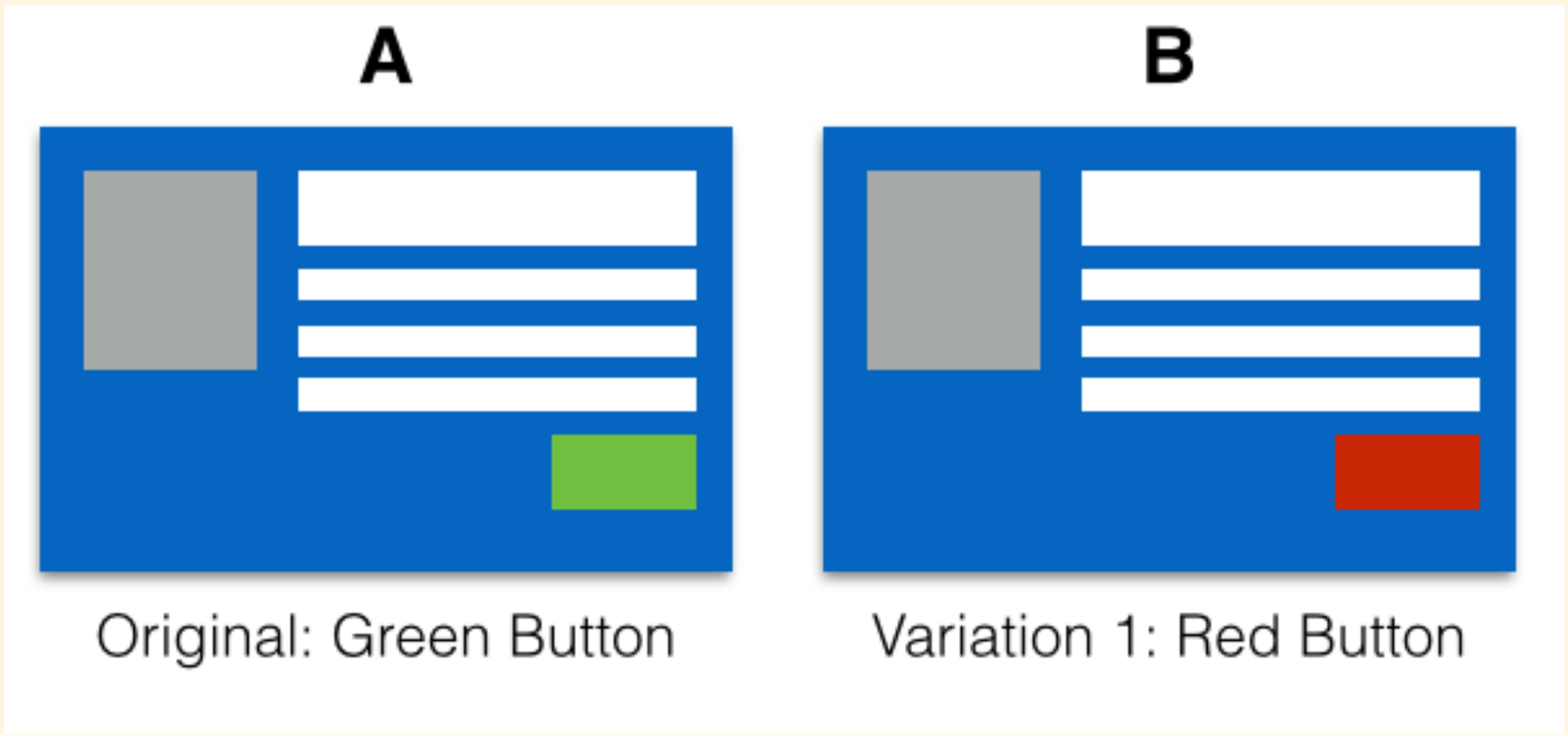
Is there any time where correlation implies causation?

The gold standard for establishing **cause** and **effect** is a controlled trial (aka A/B test).

3. A/B Testing

A/B Testing

Online experiments are used to test a new design, a machine learning model, or any new feature.



A/B Testing - Hypothesis Tests

A hypothesis test is a way to decide whether the data strongly support one point of view or another.

How do you set up an experiment?

DEFINE THE GOAL

AND

FORM THE HYPOTHESIS

DEFINE THE GOAL

AND

FORM THE HYPOTHESIS

From stats:
hypothesis tests
significance level

IDENTIFY THE CONTROL AND THE TREATMENT GROUP

IDENTIFY KEY METRICS

AND

DESIRED IMPROVEMENT

From stats:
effect size

DETERMINE THE FRACTION

IN BOTH GROUPS

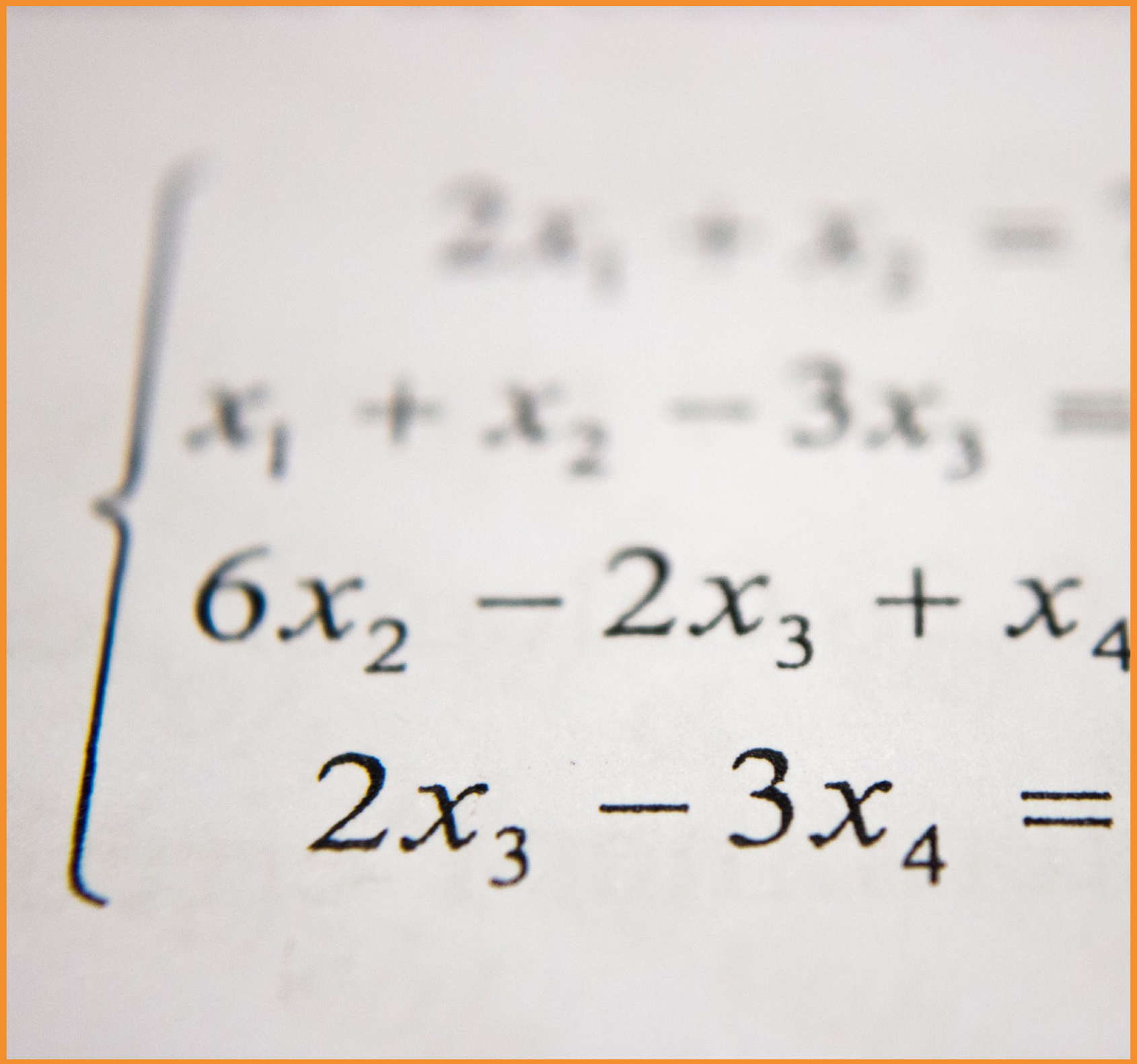
RUN THE TEST FOR A

CERTAIN AMOUNT OF TIME

From stats:
sample size

ANALYZE THE RESULTS

4. Statistical Models



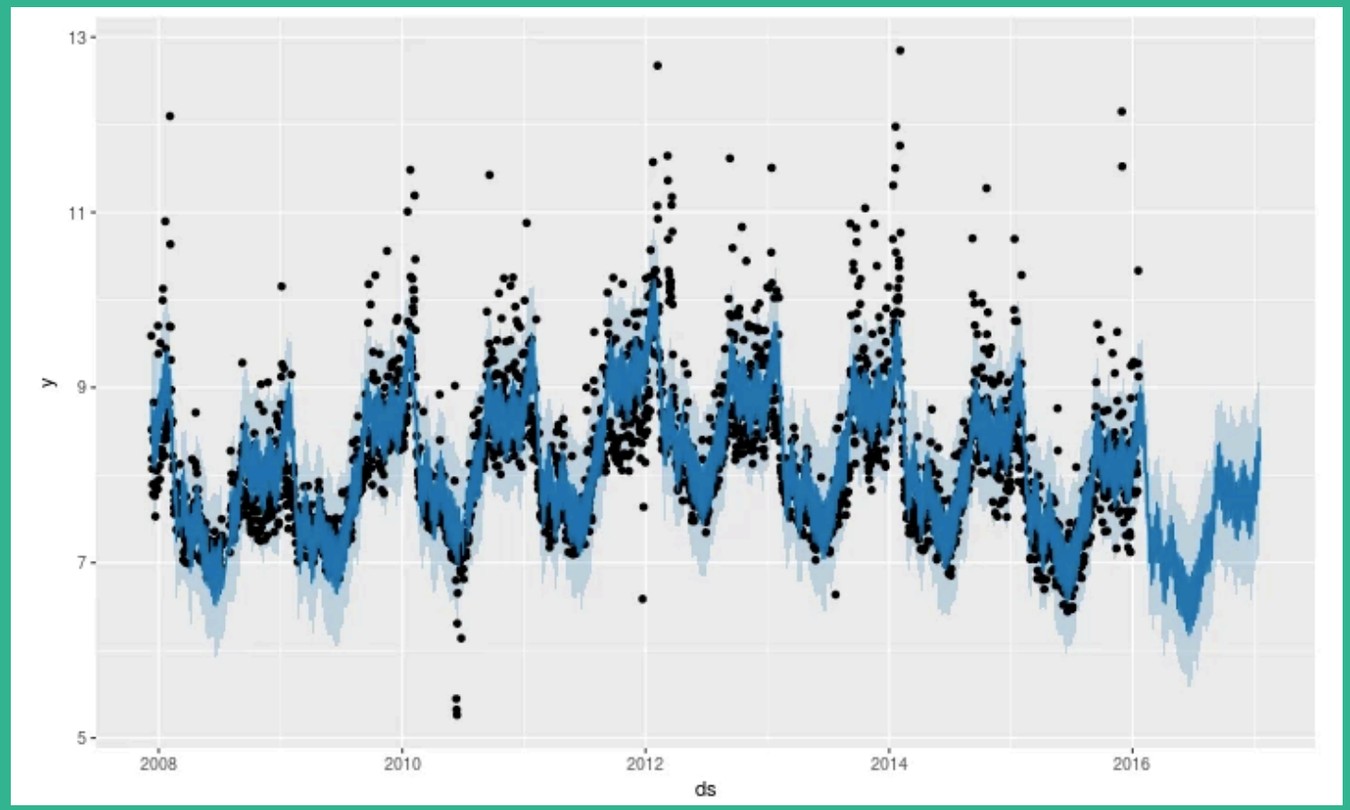
The **response** is the one whose content we are trying to model with other variables (**explanatory variables**)

In any given model:

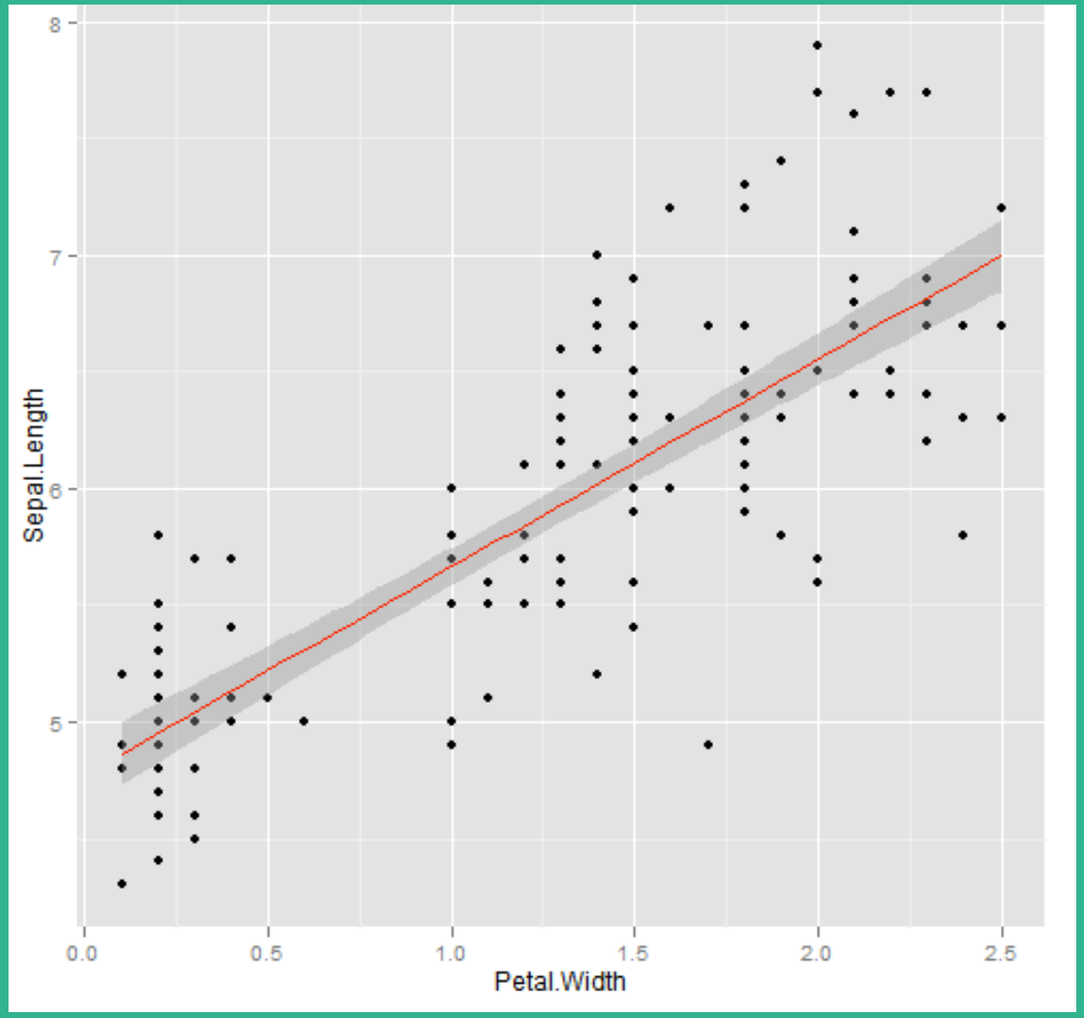
- response variable (Y)
- explanatory variables (X1, ...Xn)

Examples of models

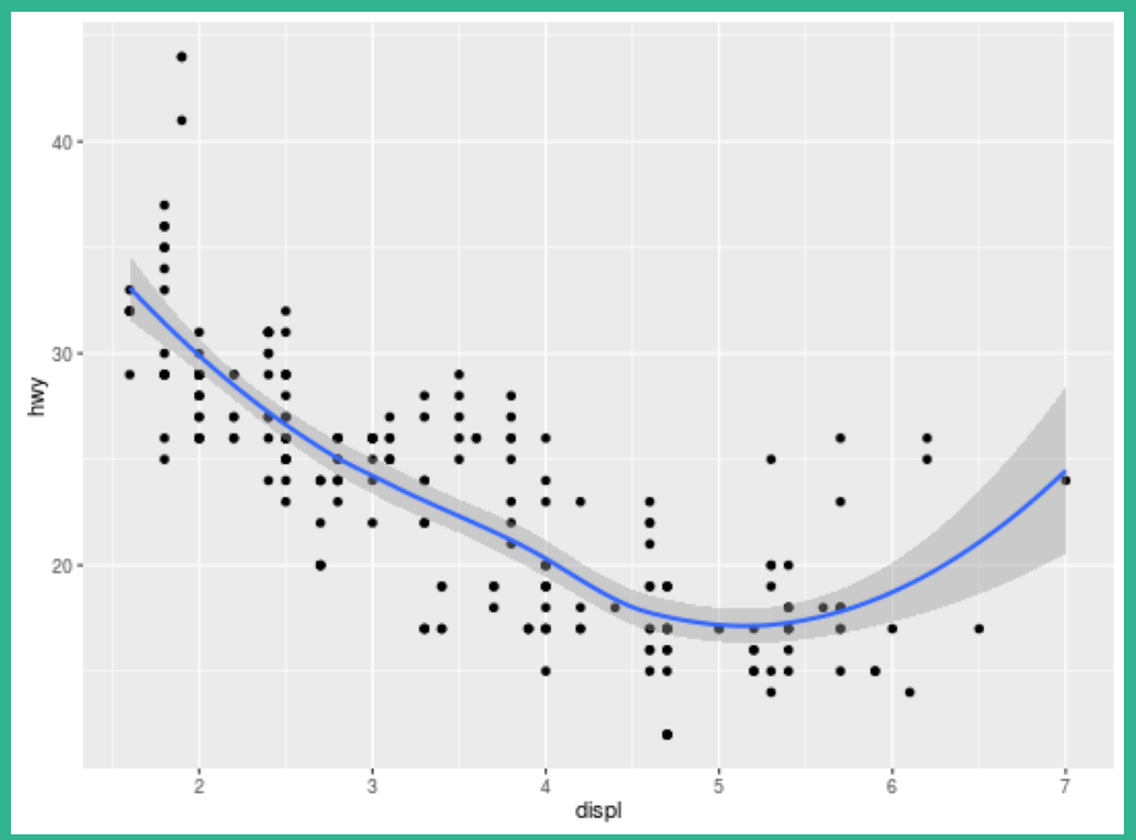
Time Series



Linear Regression



Non-Linear Regression



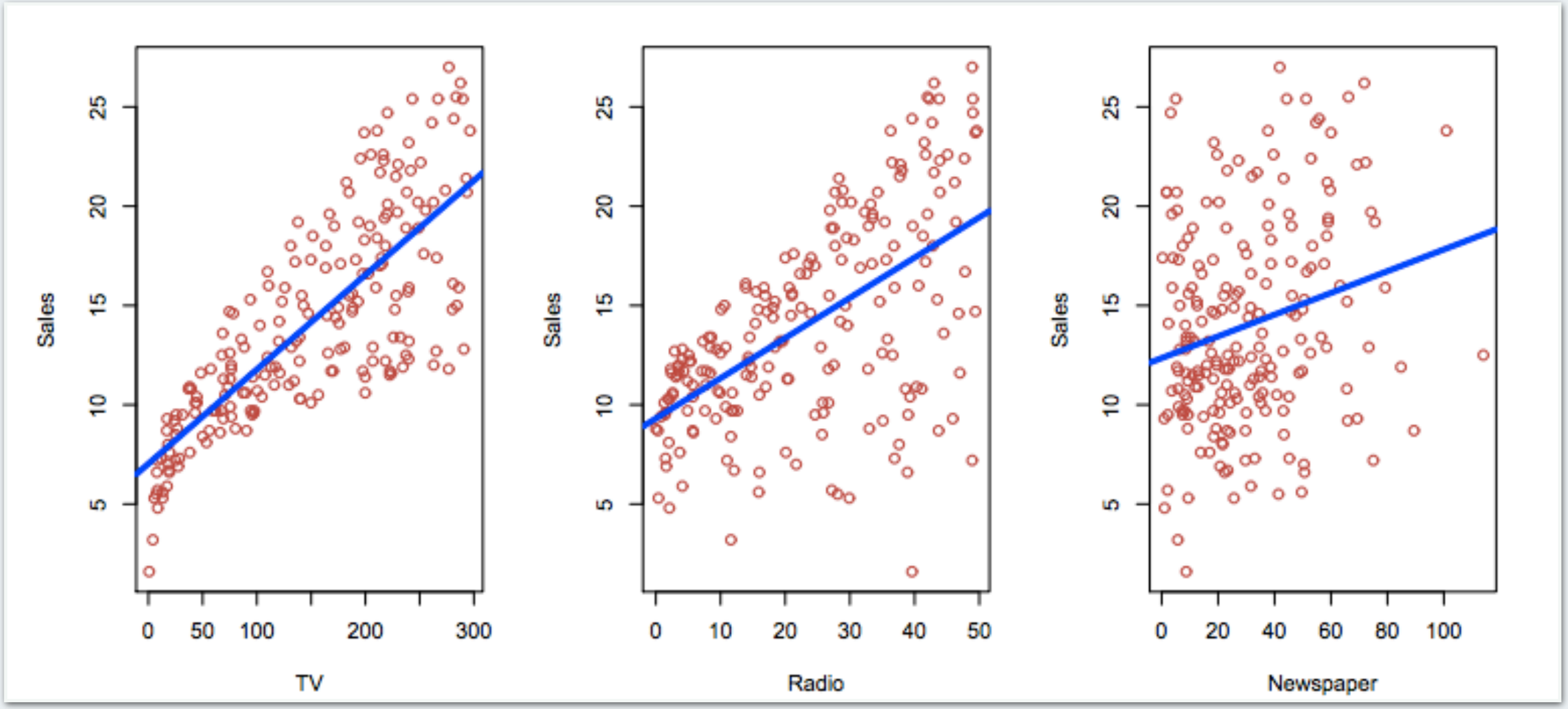
Use Case: Improve Sales of a product

- Let’s say we were hired to provide advice on how to improve sales of a particular product.

- Our goal is to develop an accurate **model** that can be used to **predict sales** based on these 3 media budgets.

Example extracted from the book "An Introduction to Statistical Learning with Applications in R"

The data consists of the **sales of the product** in 200 different markets, along with **advertising budgets** for the product in each of those markets for three different media: TV, radio, and newspaper.



output variable: sales (in thousands of units)

input variables: advertising budgets (in thousands of dollars)

The **sales** for a particular product is a *function* of **advertising budgets**.

Suppose we are asked to suggest a marketing plan for next year that will result in **high product sales**.

WHAT INFORMATION WOULD BE USEFUL TO PROVIDE?

1. Is there a relationship between advertising budget and sales?

Our first goal should be to determine whether the data provide evidence of an **association** between advertising spend and sales.

2. How strong is the relationship between advertising budget and sales?

3. Which media contribute to sales?

Do all three media contribute to sales,

or do just one or two?

4. How accurately can we estimate the effect of each media on sales?

For every dollar spent on advertising in a particular media, by what amount will sales increase?

5. How accurately can we predict future sales?

For any given advertising, what is our prediction for sales, and what is the accuracy of this prediction?

6. Is the relationship linear?

If the **relationship** between advertising spend in the various media and sales is approximately a **straight-line** then linear regression is an appropriate tool.

If not, then it may still be possible to **transform** the predictor or the response so that linear regression can be used.

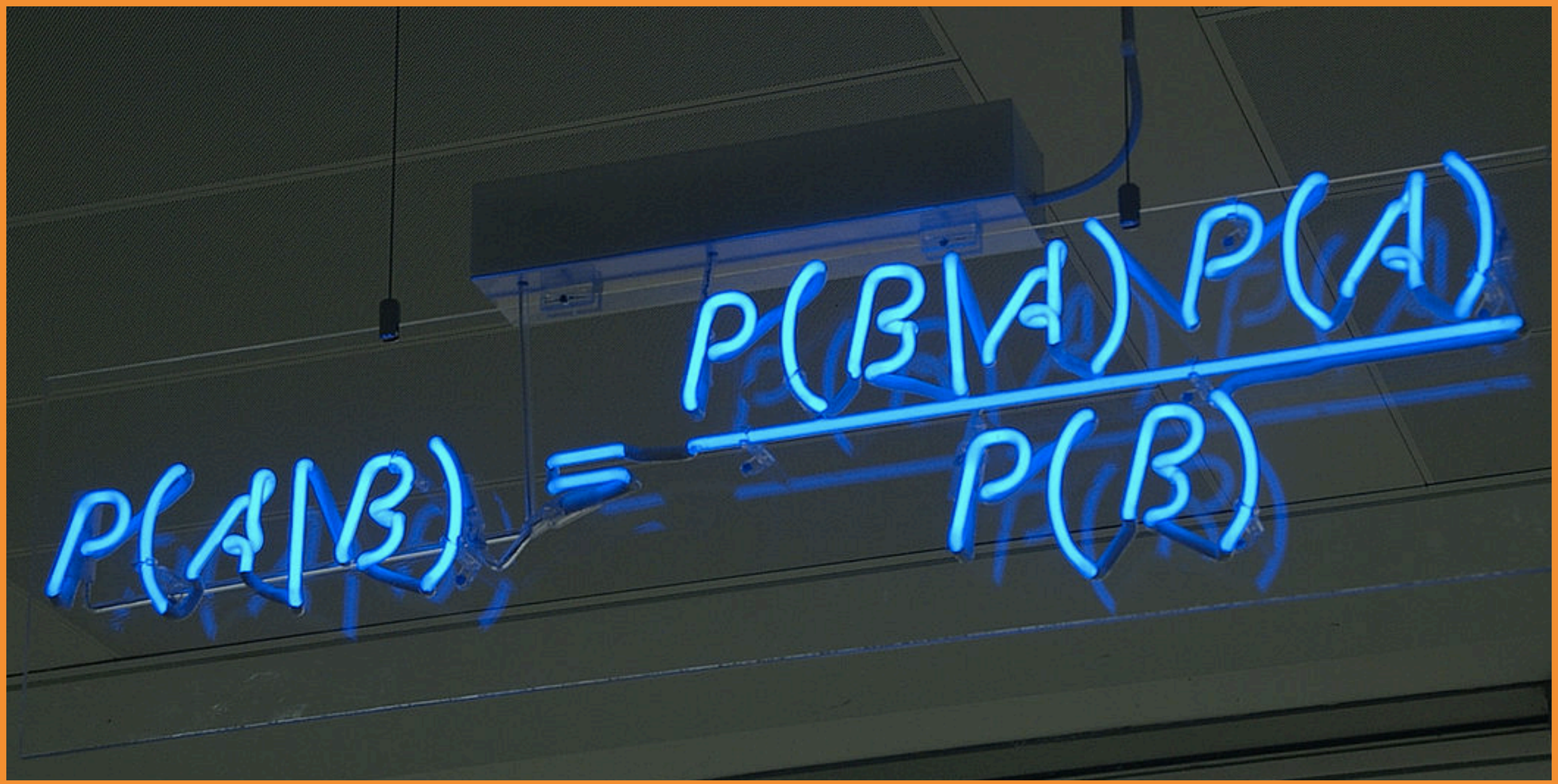
We could answer all those questions by setting up a multiple linear regression:

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon$$

Why can’t we throw all these in a black box algorithm?

INTERPRETABILITY

5. Probability



- Naive Bayes
- Logistic Regression
- *k*-NN
- Latent Dirichlet Allocation
- Decision Trees
- Association Rules (ex: Basket Analysis)
- ...

It doesn’t matter what technique you choose, the most important skill is critical thinking.

THANK YOU!

@gdequeiroz

k-roz.com

@RLadiesGlobal

www.rladies.org