ESTATÍSTICA DESCRITIVA

Estatística Descritiva

Nesse módulo

- Medidas de Centralidade
- Medidas de Dispersão
- Correlação

MEDIDAS DE CENTRALIDADE

Importância

Ao usar uma medidas de centralidade, resumimos nossa informação em um único dado – um valor central, do meio.

Mesmo quem não tem conhecimento matemático, já faz isso inconscientemente.

- “A temperatura na minha cidade é por volta de 30°C na época do carnaval.”
- “Quando eu trabalhava com vendas, eu ganhava uns 200 reais por dia.”

Importância

Além disso, é uma forma de comparar dois valores com número de observações diferentes.

Ana vendeu um total de 1000 reais em 10 dias e Bruna 800 reais em 5 dias.

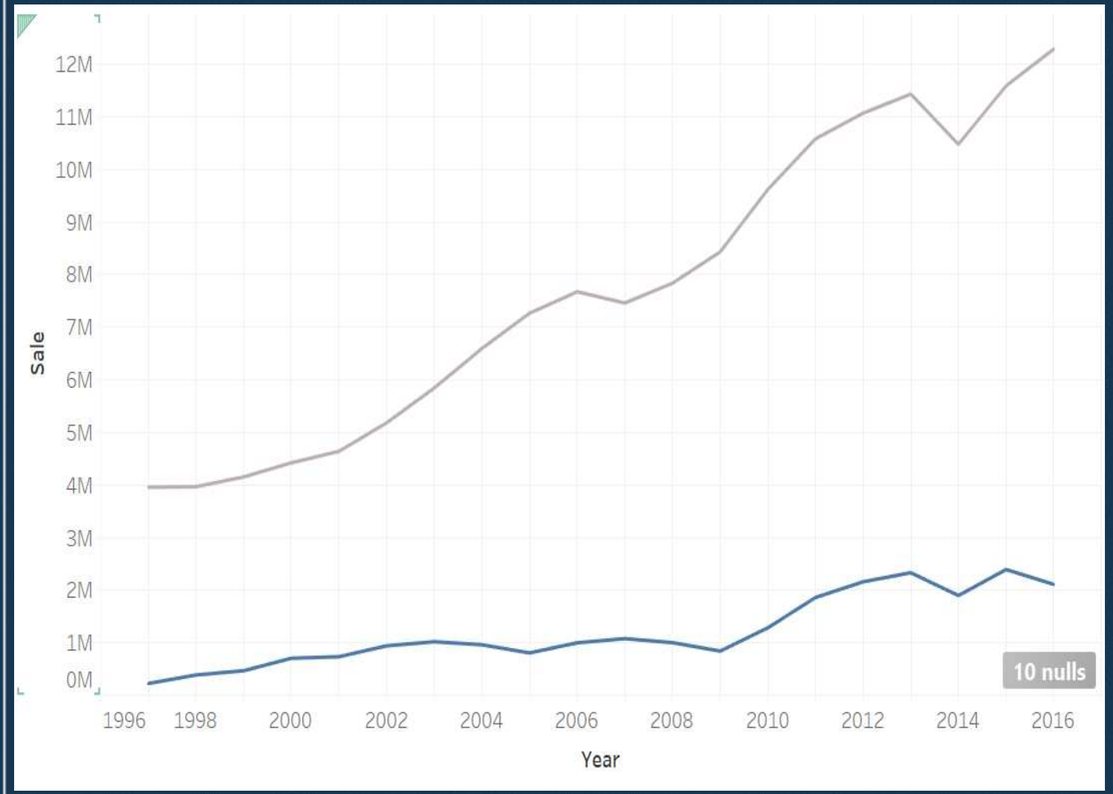
- Por mais que Ana tenha vendido mais no total, em média, ela vendeu 100 reais por dia, enquanto Bruna vendeu 160 reais por dia.

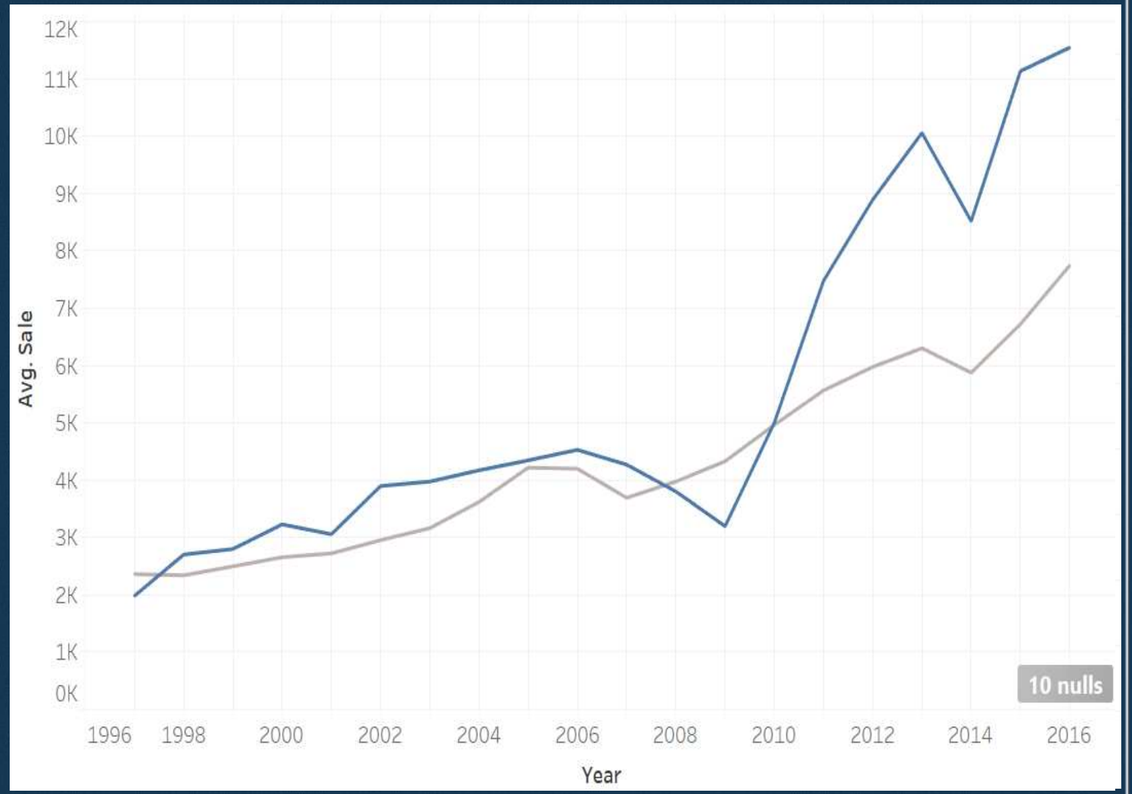
Exemplo

Você precisa realizar uma análise que descreva as vendas de empresas que analisam dados ao longo do tempo.

Para conduzir esta análise, você precisará identificar o desempenho em vendas de empresas que analisam seus dados com empresas que não analisam dados, ao longo dos anos.

Exemplo


12M
11M
10M
9M
8M
7M
6M
5M
4M
3M
2M
1M
0M
Sale
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
10 nulls


12K
11K
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Avg. Sale
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
10 nulls

Média

A média é a medida primária de centralidade.

Como é calculada: a soma de todas as observações dividida pelo número de observações.

$$\bar{x} = \frac{\sum x_i}{n}$$

Média

Em uma startup de tecnologia, foram coletados os salários dos funcionários.

Cargo	Salário (R$)
Desenvolvedor Júnior	4000
Desenvolvedor Júnior	4200
Desenvolvedor Sênior	6000
Desenvolvedor Sênior	6200
Analista de Dados	5500
Analista de Dados	5800
CEO	25000

$$\text{Média: } \frac{\text{Soma dos salários}}{\text{número de pessoas}} = \frac{4000 + 4200 + 6000 + 6200 + 5500 + 5800 + 25000}{7}$$

$$= \frac{54200}{7} = 7742{,}86$$

Média

Possível problema: a média é sensível aos outliers.
[Outliers: observações que são mais extremas que os outros valores]

No exemplo anterior, a média não reflete o salário típico.

Cargo	Salário (R$)
Desenvolvedor Júnior	4000
Desenvolvedor Júnior	4200
Desenvolvedor Sênior	6000
Desenvolvedor Sênior	6200
Analista de Dados	5500
Analista de Dados	5800
CEO	25000

Média: R$ 7742,86

Mediana

A mediana é outra medida de tendência central, mas não é afetada por outliers.

Como calcular: coloque os dados em ordem crescente e selecione o que está no meio.

Se dois valores estiverem no meio (número par de observações), tire a média entre eles.

Mediana

Exemplo da startup:

- Primeiro, ordenamos os salários em ordem crescente:
- 4000, 4200, 5500, 5800, 6000, 6200, 25000
- Como temos um número ímpar de observações (7), a mediana é o valor central da lista ordenada.
- Mediana = 5800

Portanto, a média dos salários é aproximadamente R$ 7742,86 e a mediana é R$ 5800.

A mediana é não é impactada por valores extremos e pode fornecer uma melhor representação do centro da distribuição.

Mediana

Se tivéssemos um número par de observações:
4000, 4200, 5500, 5800, 6000, 6200, 6200, 25000

Neste caso, a mediana é a média dos valores centrais da lista ordenada.

$$\text{Mediana} = \frac{5800 + 6000}{2} = 5900$$

Média ou Mediana?

Precisamos olhar a distribuição dos dados:

- Simétrica: a média é a medida ideal, já que os dados estão concentrados no centro.

- Assimétrica: a mediana é a medida ideal, já que pega o valor do meio.

Caso tenha muitos outliers, a mediana também é a mais recomendada.

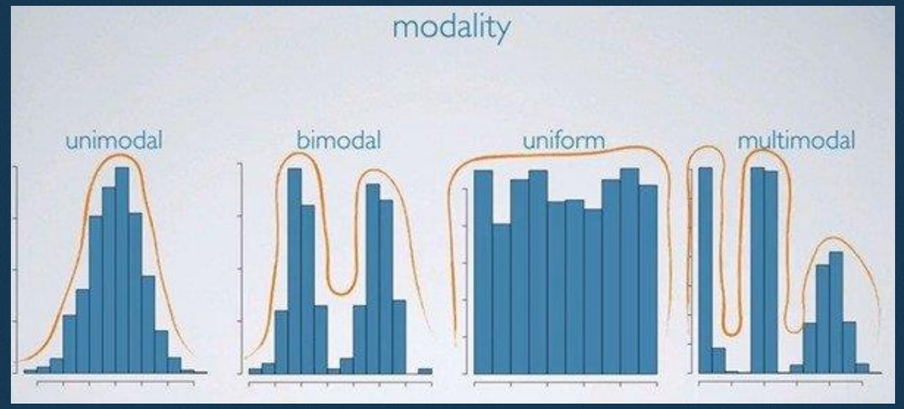
Apresente as duas medidas juntas, mas entenda o que cada uma está representando.

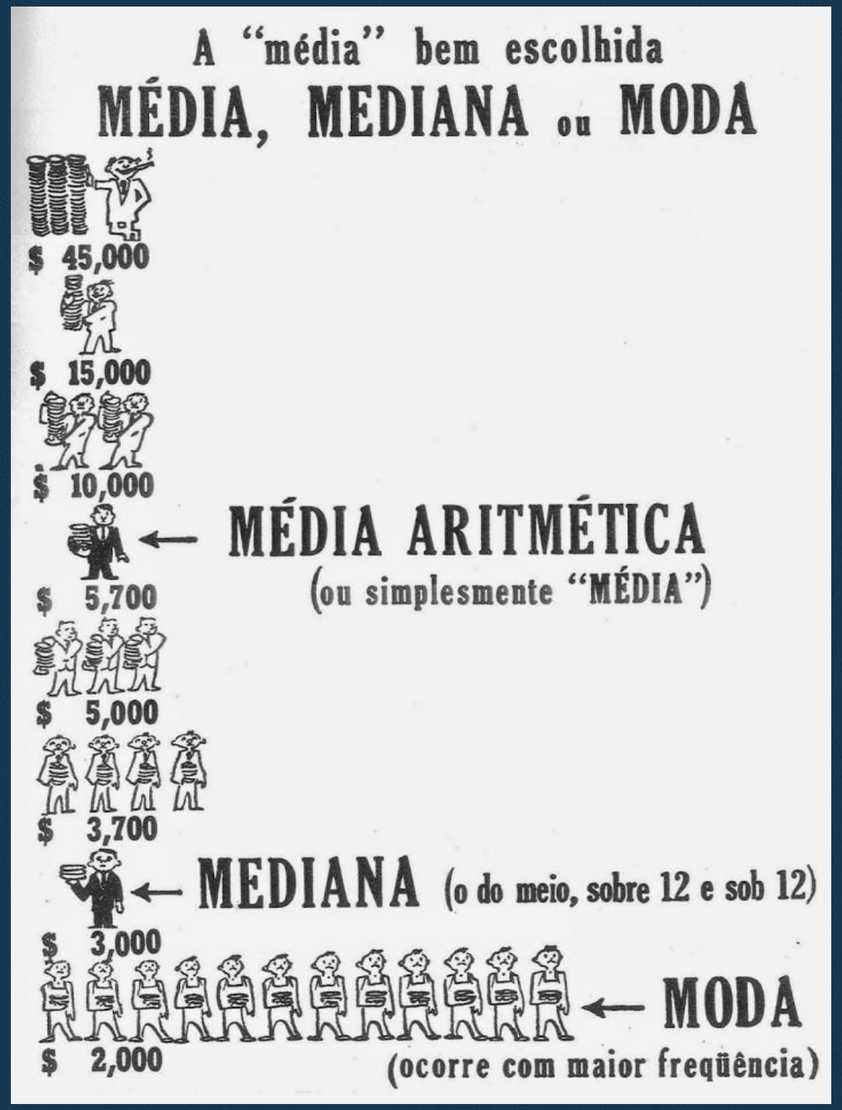
Moda

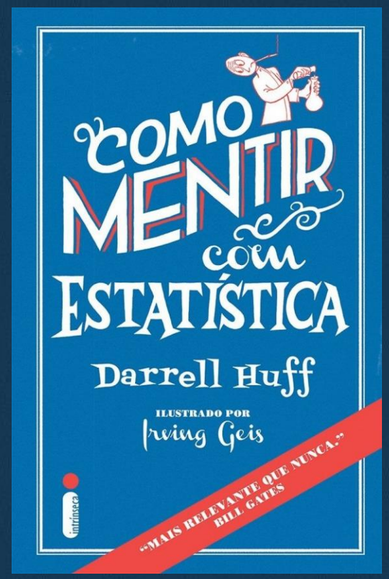
A moda é o valor mais frequente em um conjunto de dados.

No exemplo da startup, não temos um valor que se repete, portanto, não temos uma moda.

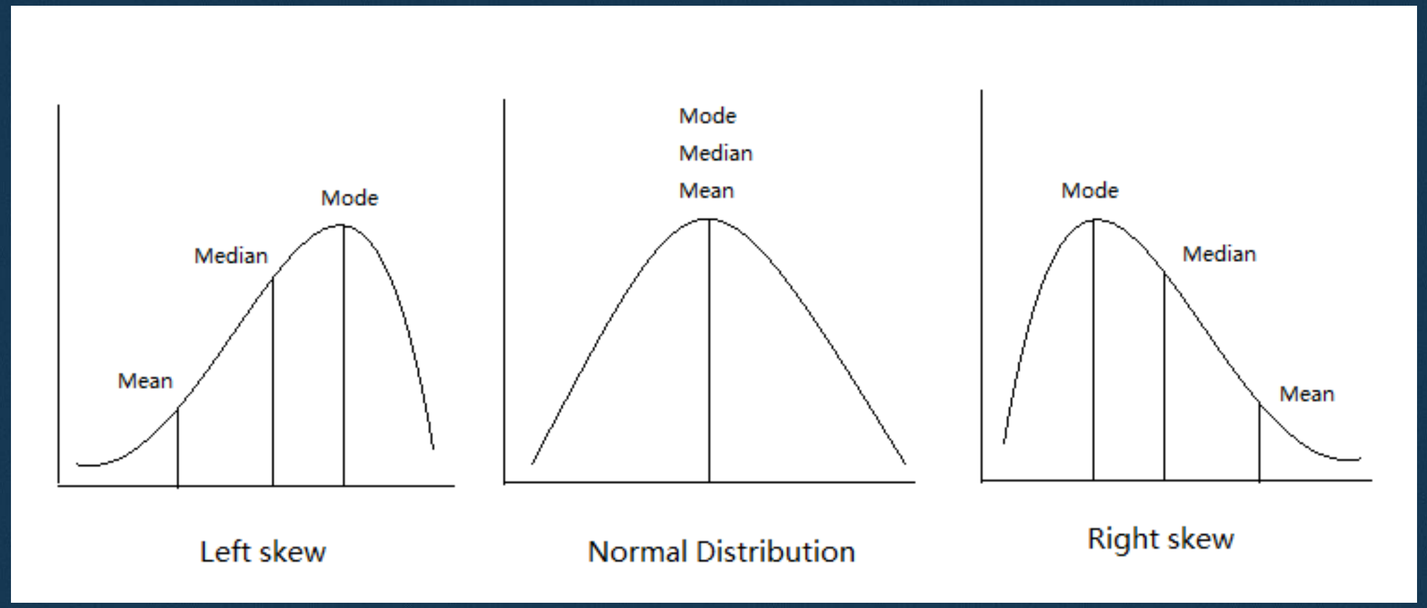
A moda é mais usada com dados qualitativos, assim observamos a categoria com maior frequência.



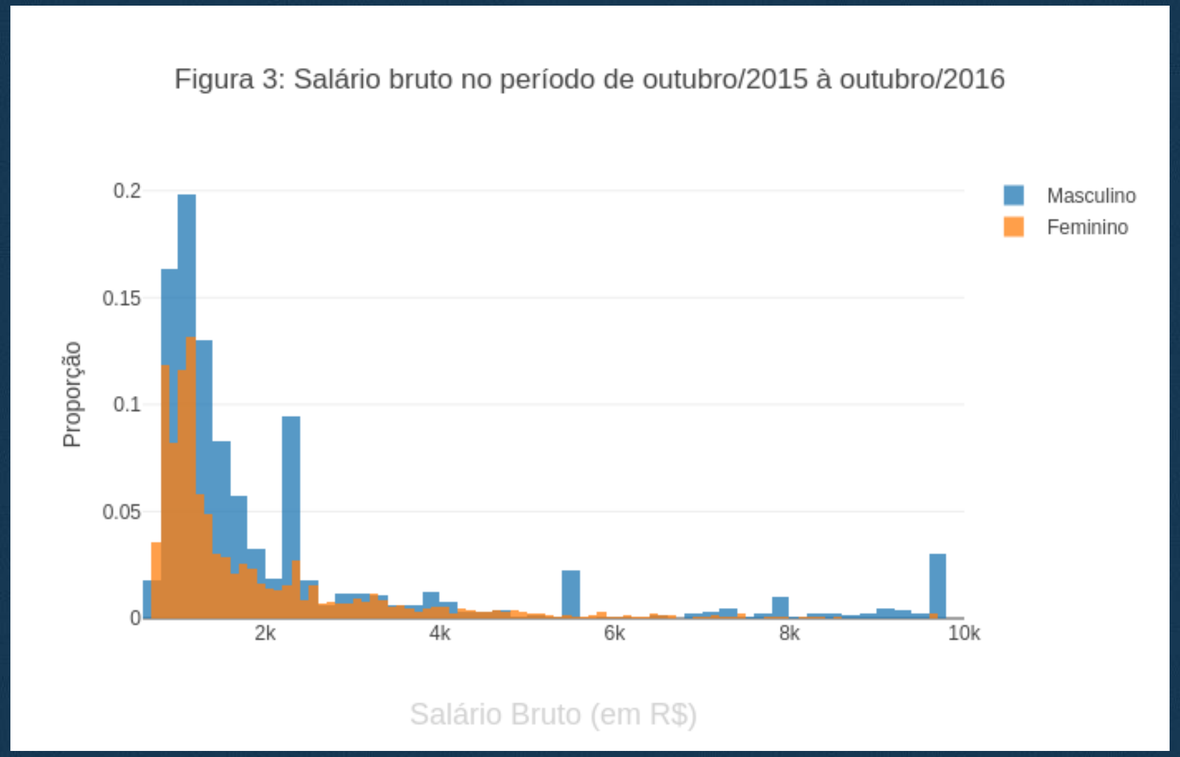

A "média" bem escolhida
MÉDIA, MEDIANA ou MODA
$ 45,000
$ 15,000
$ 10,000
MÉDIA ARITMÉTICA
(ou simplesmente "MÉDIA")
$ 5,700
$ 5,000
$ 3,700
MEDIANA (o do meio, sobre 12 e sob 12)
$ 3,000
MODA
$ 2,000
(ocorre com maior freqüência)


COMO MENTIR com ESTATÍSTICA
Darrell Huff
ILUSTRADO POR
Irving Geis
"MAIS RELEVANTE QUE NUNCA."
BILL GATES

Centralidade e simetria



Distribuição dos dados



Média ponderada

Utilizamos a media ponderada quando os dados já estão agrupados por frequência.

No exemplo, temos 73 pessoas que deram 5 estrelas, 2 pessoas que deram 4, nenhuma deu 3 ou 2 e 1 pessoa deu 1 estrela.

Média ponderada: $\frac{73x5 + 2x4 + 0x3 + 0x2 + 1x1}{73+2+1} = \frac{374}{76} = 4{,}92$

MEDIDAS DE DISPERSÃO

Medidas de dispersão

Medidas de dispersão

Mostram a extensão ou variabilidade de um conjunto de dados.

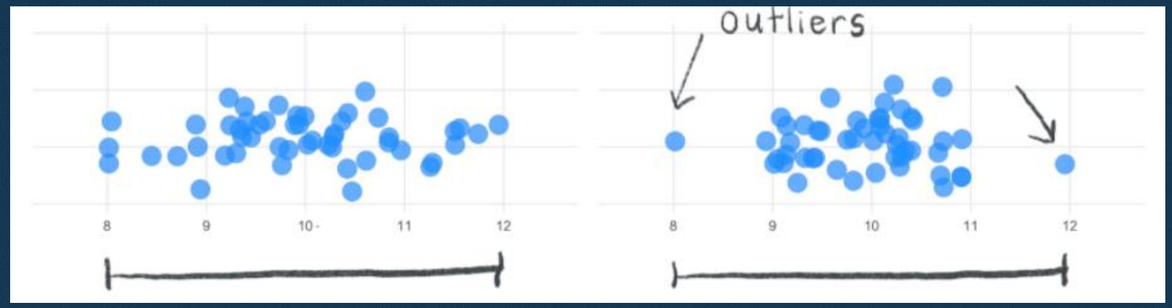
Algumas medidas de dispersão são:

- Amplitude
- Variância
- Desvio padrão
- Coeficiente de variação

Amplitude

Amplitude = Valor máximo – valor mínimo

A amplitude é a medida mais simples.
Não é uma boa medida quando usada sozinha, já que é computada usando apenas dois valores.

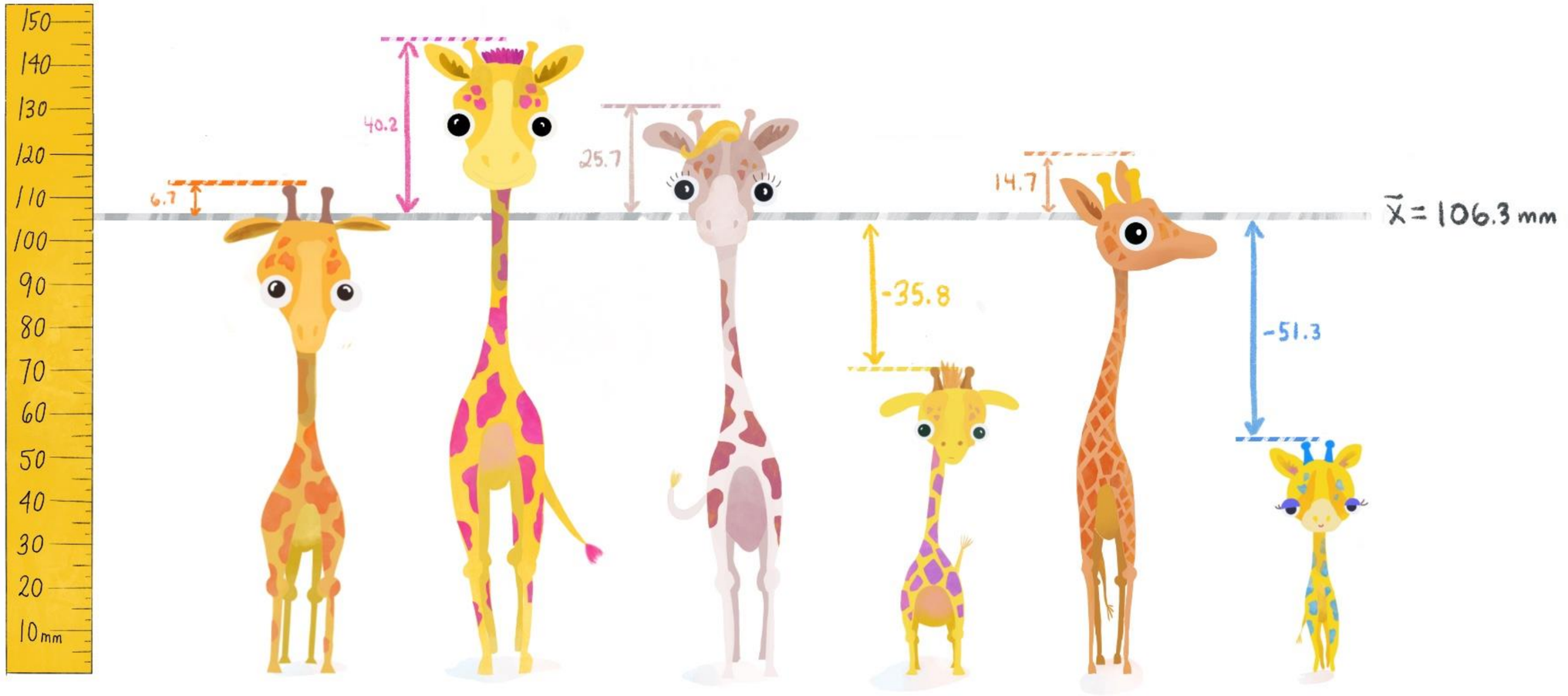


Variância

É uma medida de variação baseada na média.

A distância que um valor está da média é seu desvio; a variância é a média de desvios ao quadrado.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$


150
140
130
120
110
100
90
80
70
60
50
40
30
20
10 mm
6.7
40.2
25.7
-35.8
14.7
-51.3
$\bar{x}$ = 106.3 mm

Desvio Padrão

É a raiz quadrada da variância.

$$s = \sqrt{s^2}$$

É uma medida de variabilidade na unidade de medida original dos dados (já que a variância resulta em unidades ao quadrado).

Obs: o desvio padrão e a variância são as medidas mais usadas de dispersão. São geralmente reportadas com a média para descrever o centro e a dispersão dos dados.

Desvio Padrão

Não é interpretado, já que o que ele mede é a dispersão em volta da média.

Mas você poderia utilizar para comparar a variação entre dois conjuntos de dados.

Exemplo: um conjunto de dados tem o desvio padrão de 4,5, com as observações mais dispersas do que um conjunto de dados, que tem o desvio padrão de 2,3.

Usos:

- Regra empírica
- Inferência Estatística

Erro padrão

O erro padrão é uma medida usada quando trabalhamos com inferência estatística.

Vamos abordar esse assunto no módulo referente.

Coeficiente de variação

É o desvio padrão dividido pela média.

$$CV = \frac{s}{\bar{x}}$$

É a medida da variação dos dados em relação ao seu valor médio.

Utilizada para comparar a dispersão entre conjuntos de dados com unidades de medidas diferentes.

Coeficiente de variação

Exemplo: em um grupo, coletamos a altura e peso das pessoas. Para altura, a média é de 170 cm e o desvio padrão é de 15 cm. Para o peso, a média é 75 kg e o desvio padrão é de 10 kg. Qual medida tem a maior dispersão?

$$CV = \frac{s}{\bar{x}}$$

CV (altura) = $\frac{15}{170}$ = 0,09 ou 9%

CV (peso) = $\frac{10}{75}$ = 0,13 ou 13%

CORRELAÇÃO

Correlação

A correlação descreve uma associação linear entre duas variáveis.

A correlação pode ser:

- Positiva: quando uma aumenta, a outra aumenta (quando uma diminui e outra diminui)
- Negativa: quando uma aumenta, a outra diminui

Exemplo: eu tenho gastos com eletricidade e gás (água quente).

- Qual a relação entre a temperatura e a conta de energia?
- Qual a relação entre a temperatura e a conta de gás?

Tipos de Correlação

Tipos: Pearson, Spearman, Kendall

Vamos focar no Pearson que é o tipo mais comum e o padrão nos programas estatísticos.

Suposições:

1) As duas variáveis precisam ser contínuas;
2) Devem ter uma relação linear (verificamos pelo gráfico de dispersão);
3) Não devemos ter muitos outliers afetando a correlação.

O que observamos?

Ao analisar a correlação, observamos:

- Gráfico de dispersão
- Coeficiente de correlação r

Correlação Positiva

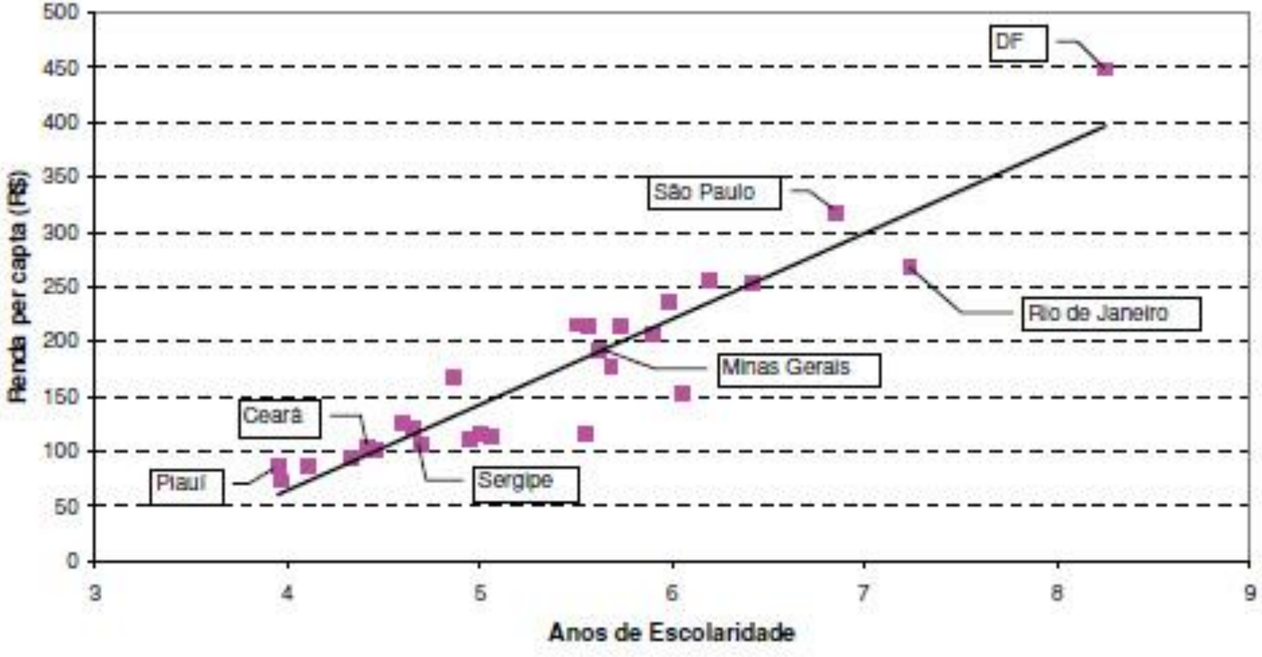
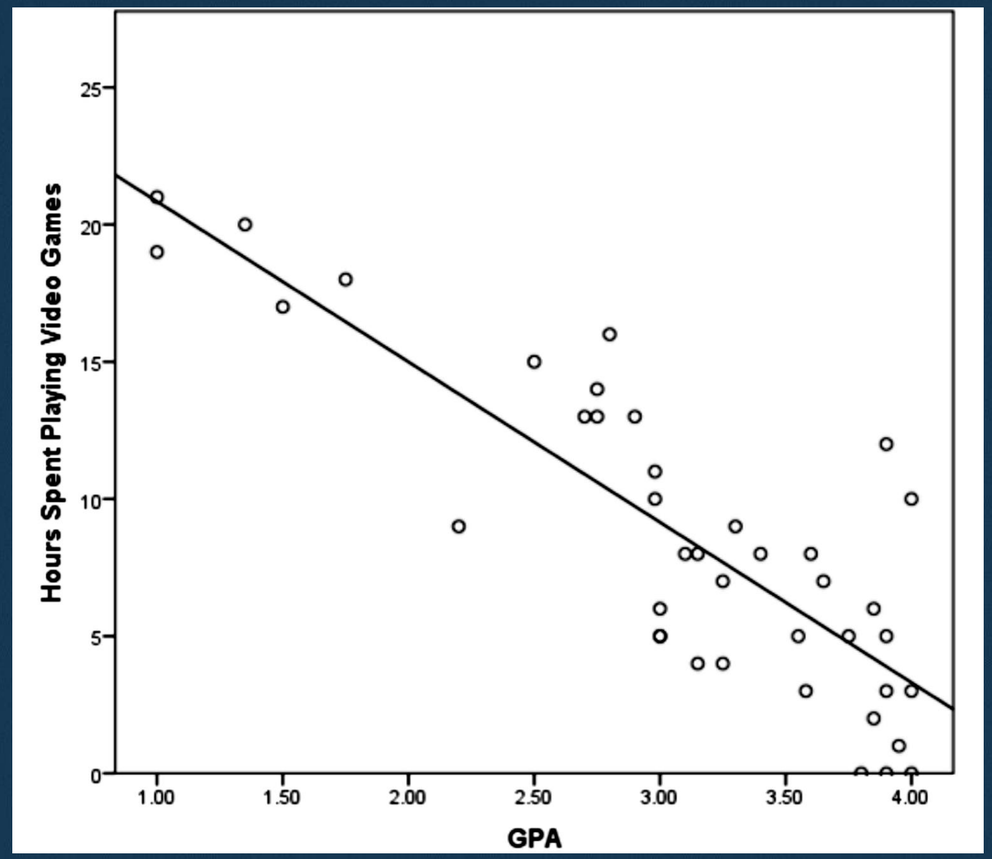


Gráfico 2 – Renda *Per Capita* x Escolaridade Média

Fonte: IBGE.

Correlação Negativa

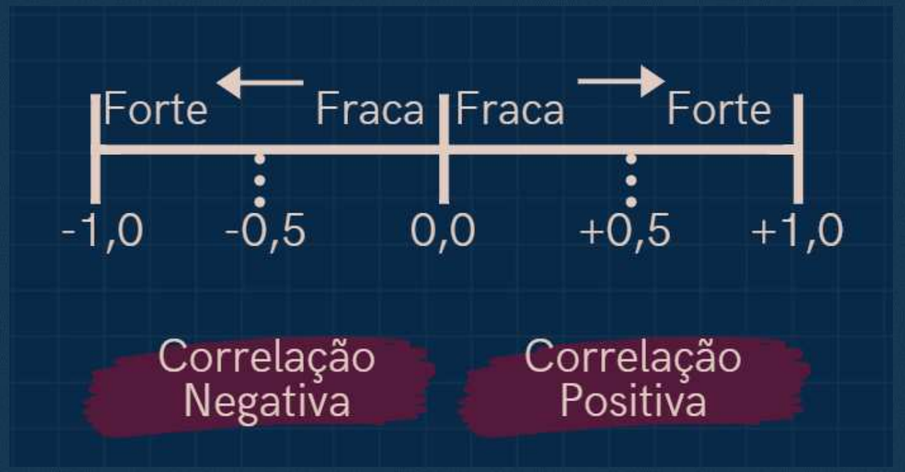


Coeficiente de Correlação

O coeficiente de correlação é um número entre -1 e 1 que representa a direção e o grau da relação.

Direção:
Correlação>0: relação linear positiva
Correlação<0: relação linear negativa
Correlação=0: sem relação linear

Grau:
Correlação=1: relação linear positiva forte
Correlação=-1: relação linear negativa forte

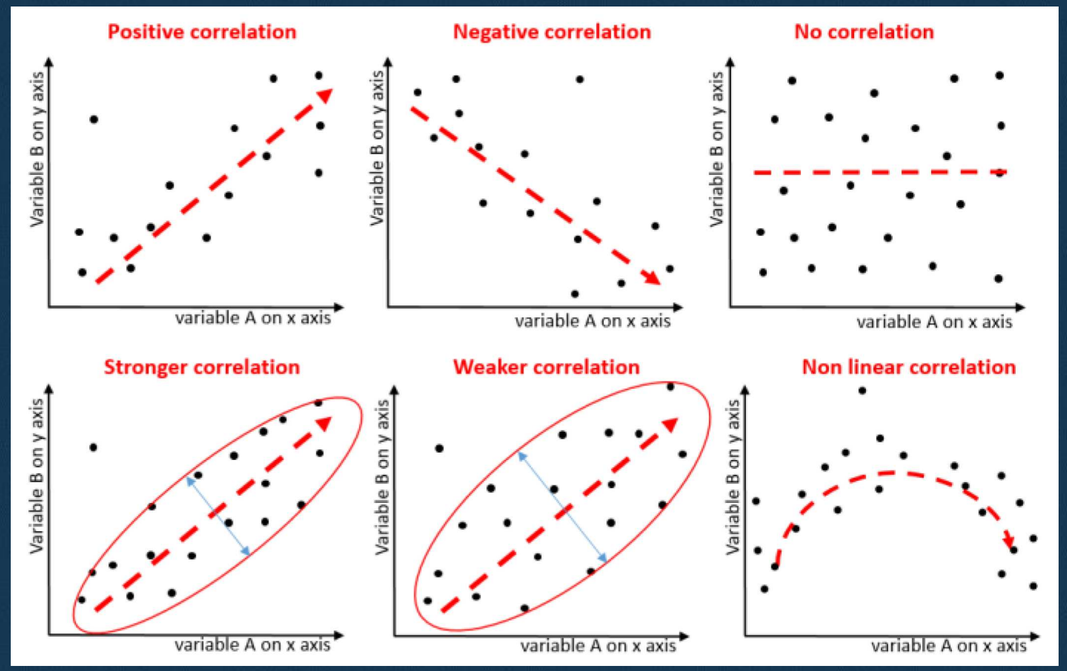


Padrões da correlação

Direção

Coeficiente de correlação > 0:

- Relação positiva
- Temperatura & conta de energia

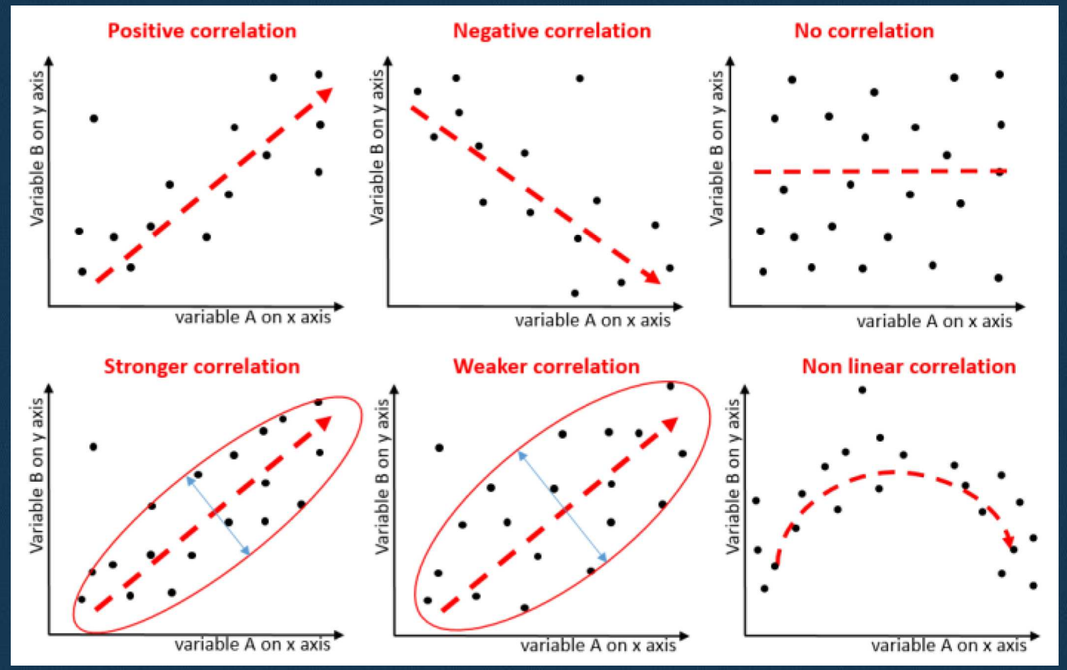


Padrões da correlação

Direção

Coeficiente de correlação < 0:

- Relação negativa
- Temperatura & conta de gás

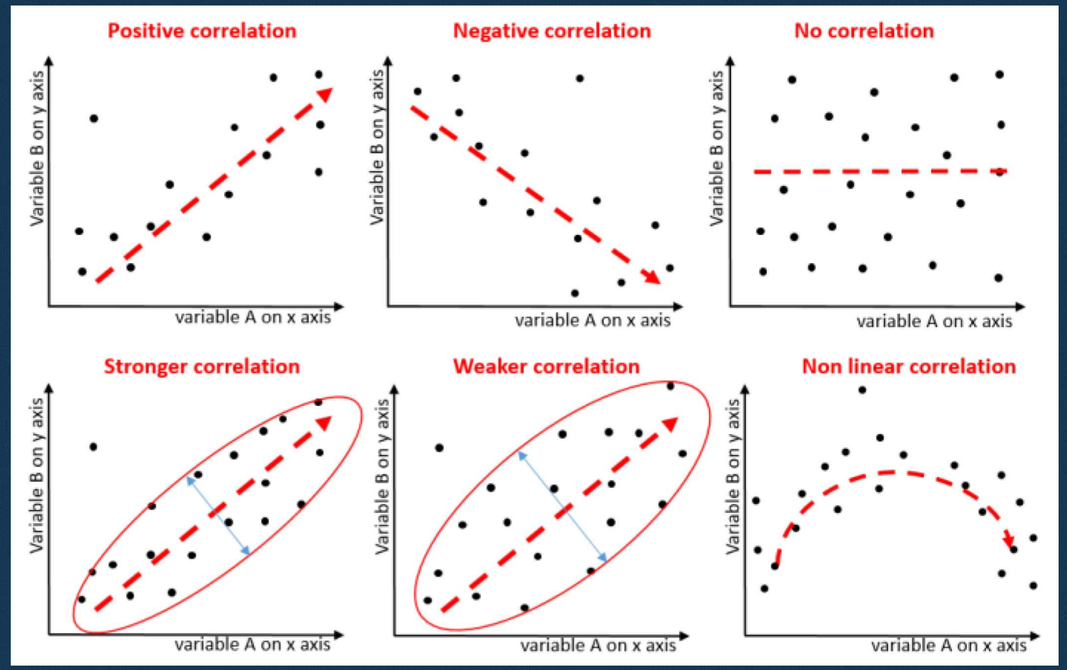


Padrões da correlação

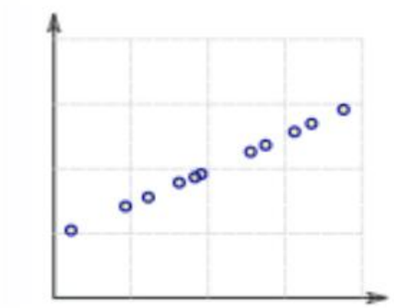
Direção

Coeficiente de correlação = 0:

- Relação inexistente
- Temperatura & vendas de jeans

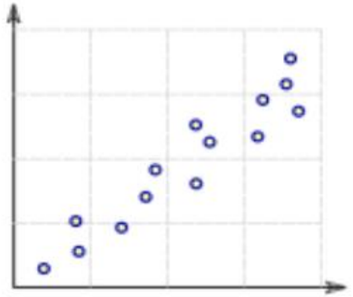


Correlação positiva perfeita



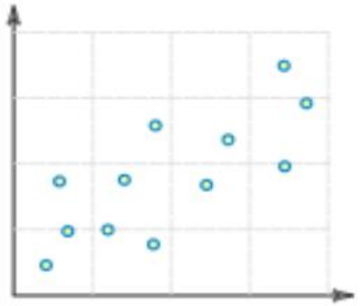
r = 1

Correlação positiva alta



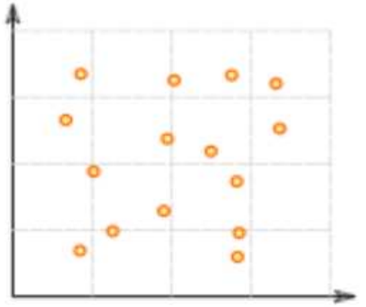
r = 0,9

Correlação positiva moderada



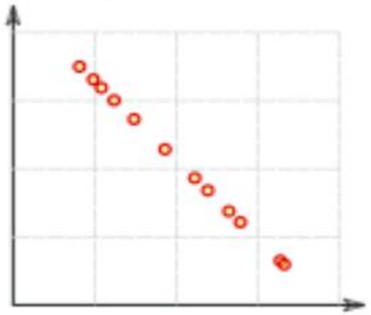
r = 0,5

Sem correlação



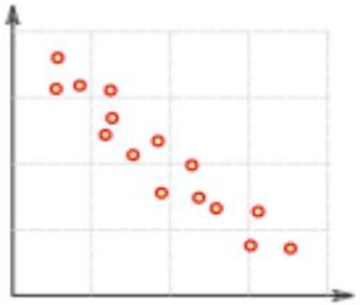
r = 0

Correlação negativa perfeita



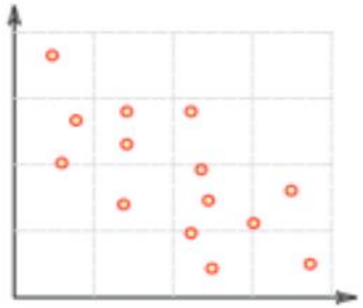
r = -1

Correlação negativa alta



r = -0,9

Correlação negativa moderada



r = -0,5

E A COVARIÂNCIA?

Covariância

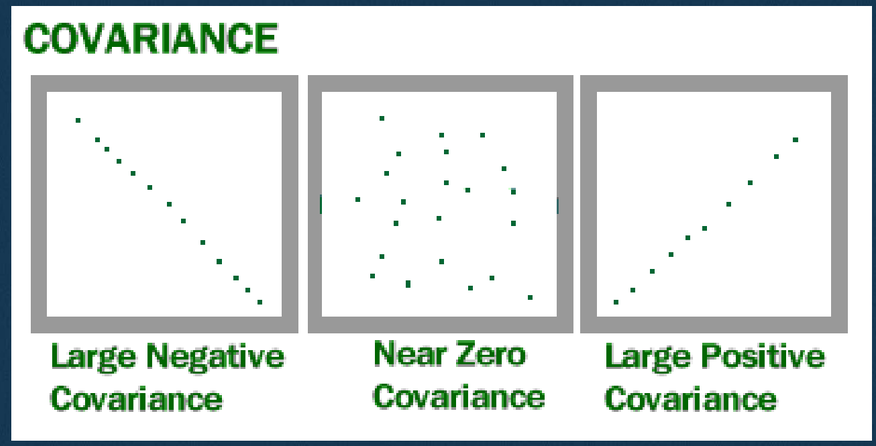
A covariância é uma medida numérica que descreve somente a direção da relação linear entre duas variáveis.

Direção:

Covariância>0: relação linear positiva

Covariância<0: relação linear negativa

Covariância=0: sem relação linear



Covariância

- O valor da covariância varia de - infinito a + infinito.
- A covariância não é muito usada, já que só nos diz a direção da associação e não o quão forte ou fraca ela é.
- A correlação é recomendada por dar as duas informações.

APLICAÇÕES

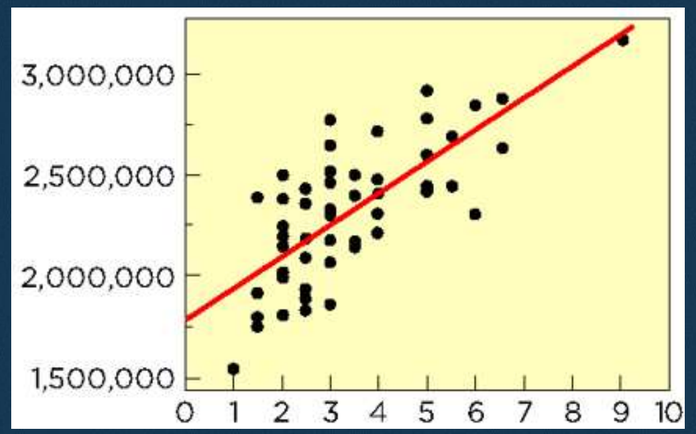
Aplicação

É melhor abrir uma nova loja mais próximo ou mais distante dos meus concorrentes?

Posso coletar dados para verificar se existe uma associação entre as vendas de uma loja e a distância do concorrente mais próximo.

Aplicação

Os seguintes dados foram coletados de 55 lojas de uma rede: o total de vendas no último ano e a distância para o concorrente mais próximo, em kilômetros.



Diversificação

Ao criar um portfólio, investidores tentam diversificar escolhendo um conjunto de investimentos cujo retornos não sejam relacionados, ou "correlacionados".

Se os investimentos não são correlacionados, e um deles está em queda, o outro pode ainda estar performando bem; isso faz com que a performance geral do portfólio aprimore.

Diversificação

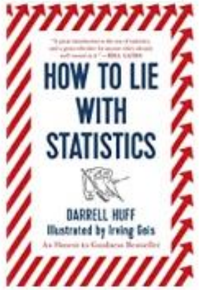
Exemplo: a correlação entre dois fundos é de 85,6%.

Análise: O coeficiente indica que existe uma relação forte entre os dois fundos: no ano que uma está em ascensão, a outra também está. Nos anos que uma está em queda, a outra também está.

Comprar os dois fundos não lhe daria uma diversificação, já que os fundos caminham em uma mesma direção.

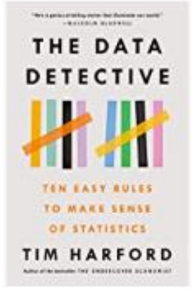
Sistema de Recomendação

Frequently bought together



+

+



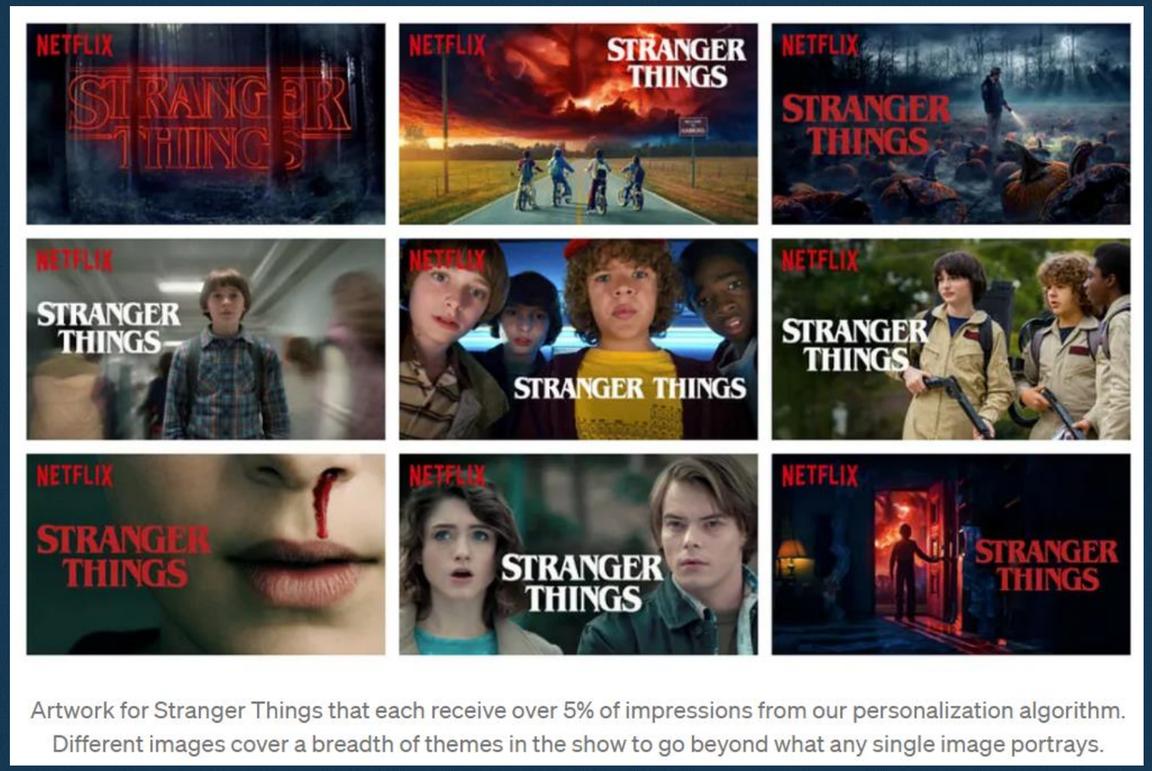
Total price: $38.58

Total Points: 54 pt

Add all three to Cart

- [x] **This item:** How to Lie with Statistics by Darrell Huff Paperback $8.89 18 pts
- [x] Naked Statistics: Stripping the Dread from the Data by Charles Wheelan Paperback $11.69
- [x] The Data Detective: Ten Easy Rules to Make Sense of Statistics by Tim Harford Paperback $18.00 36 pts

Personalização da Netflix

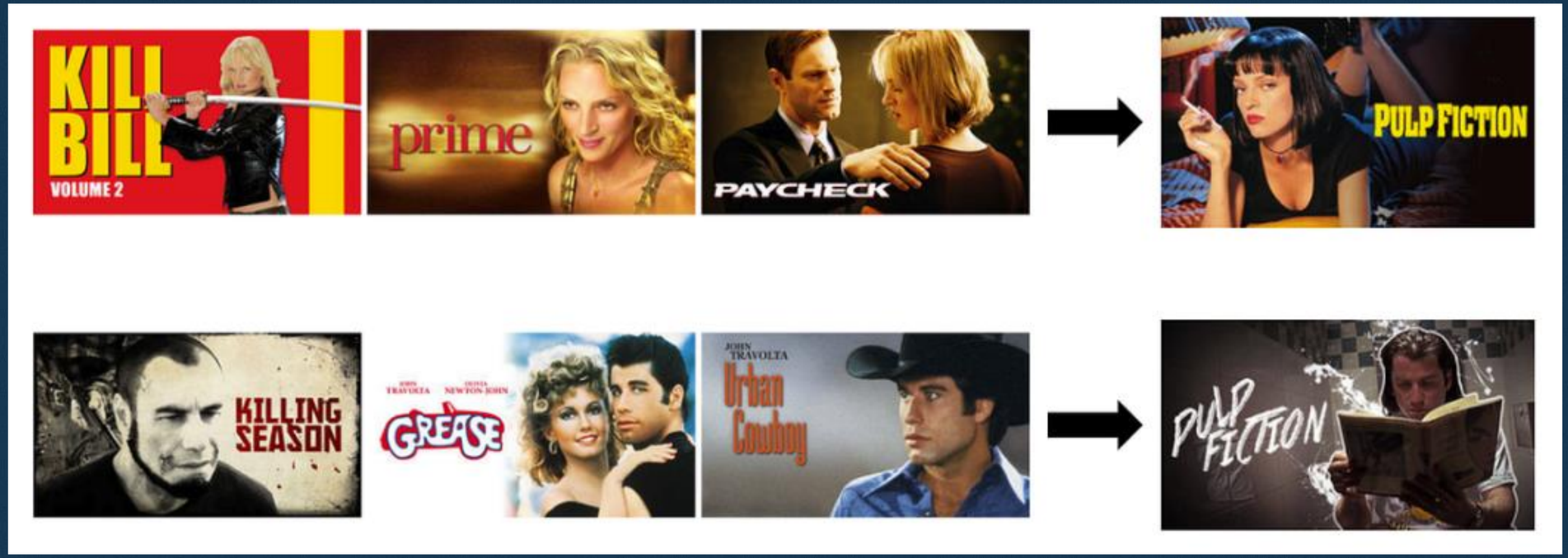


Artwork for Stranger Things that each receive over 5% of impressions from our personalization algorithm. Different images cover a breadth of themes in the show to go beyond what any single image portrays.

https://netflixtechblog.com/artwork-personalization-c589f074ad76

Personalização da Netflix

Personalização da Netflix



Caso Target

“Como a Target descobriu que uma adolescente estava grávida antes que seu pai soubesse”

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=62a1d8916668

Vendas Combinadas

Podemos entender a correlação entre vendas de produtos para fazer uma venda combinada (venda casada) [Técnica de Market Basket ou Associação].

No Walmart, encontraram que cerveja e fraldas são compradas juntas.

Interpretação: já que os pais não podem mais sair para um bar por causa do bebê, eles estavam comprando mais cerveja ao comprar fraldas.

Vendas Combinadas

Não é prático: logisticamente é impossível um mercado posicionar fraldas na área que vende álcool e também não teria sentido colocar cerveja na área de bebês.

Essa história é uma “lenda urbana”, inventada para fins ilustrativos.

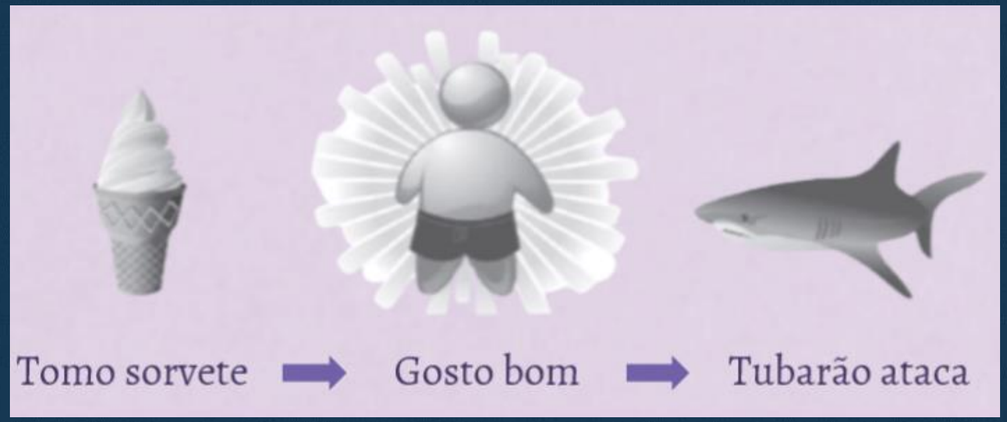
Não temos como determinar uma causalidade.

CORRELAÇÃO E CAUSALIDADE

Correlação x Causalidade

Considere esse fato: “O consumo de sorvete está relacionado à ataques de tubarão”.

A explicação causal seria que tomar sorvete nos faz ter um gosto bom para os tubarões:



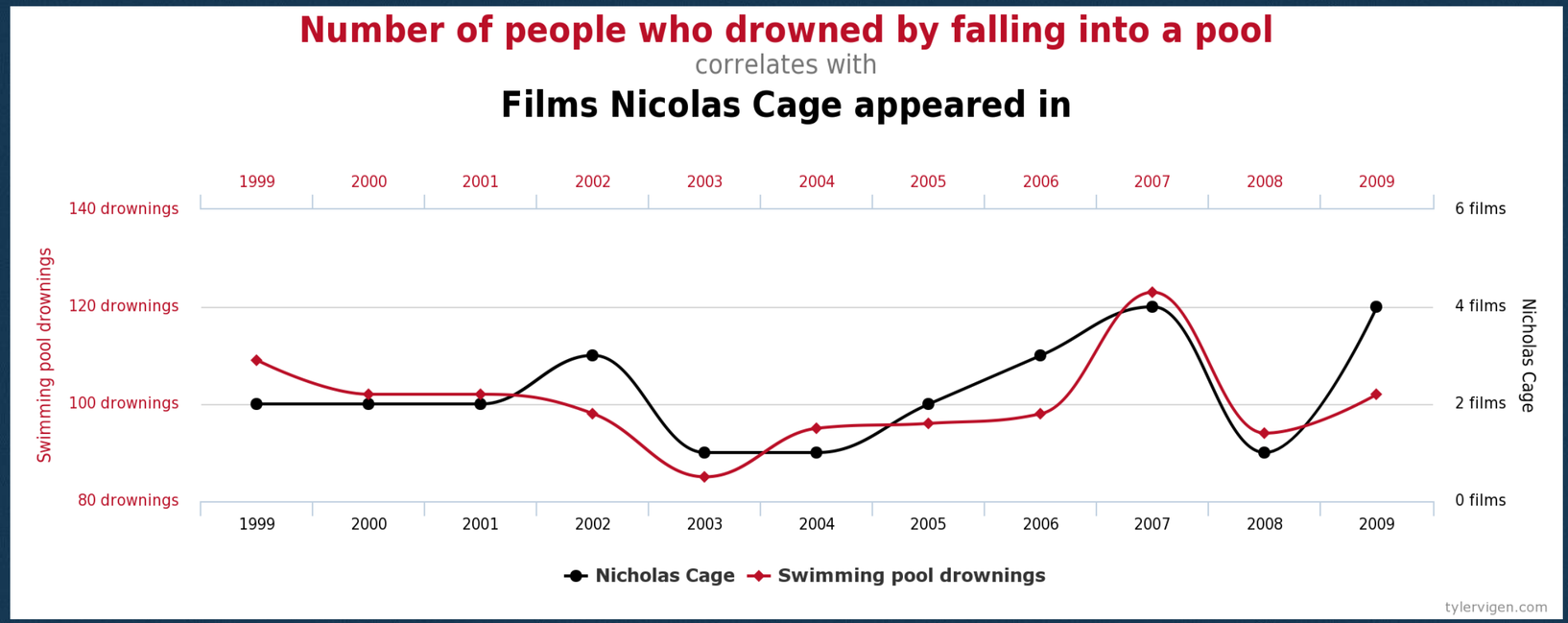
Causalidade?

Suponha que A e B tenham correlação positiva.

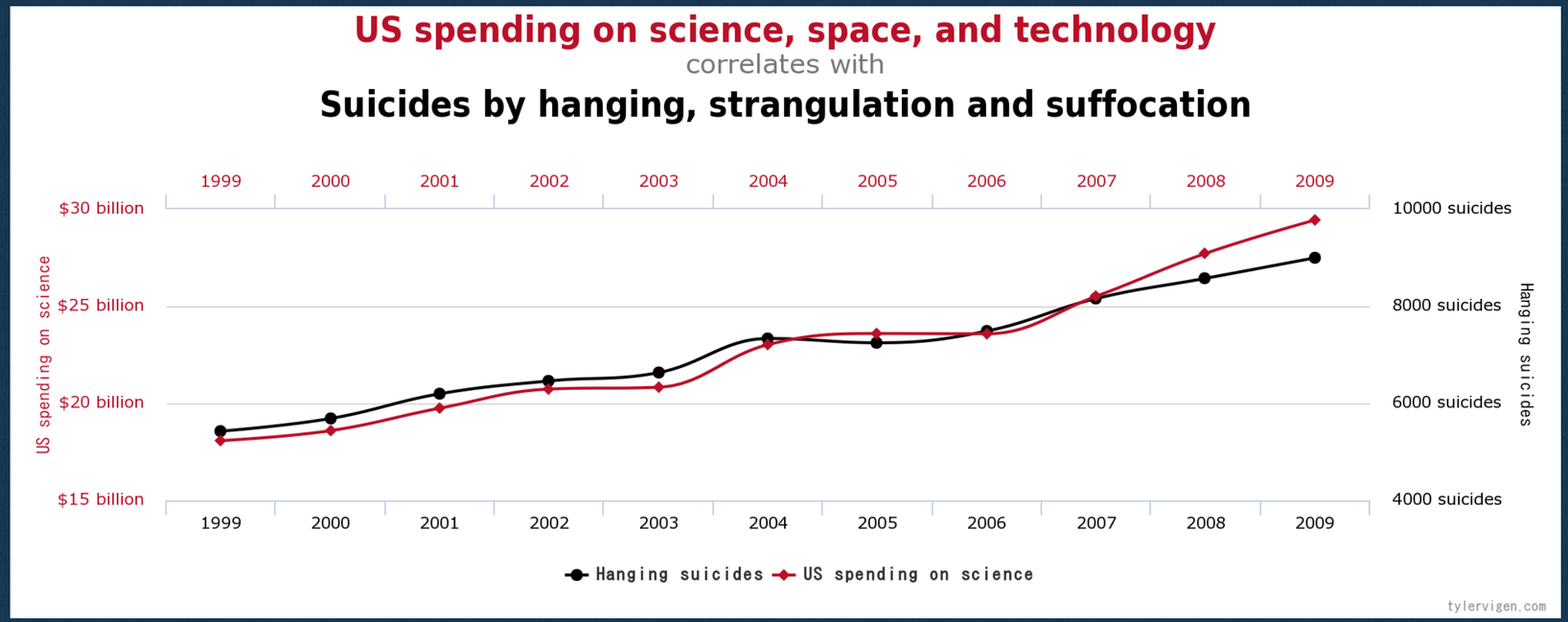
Temos 4 possibilidades:

- A causa B
- B causa A
- A e B ocorrem por causa de uma terceira variável
- A e B ocorrem por coincidência

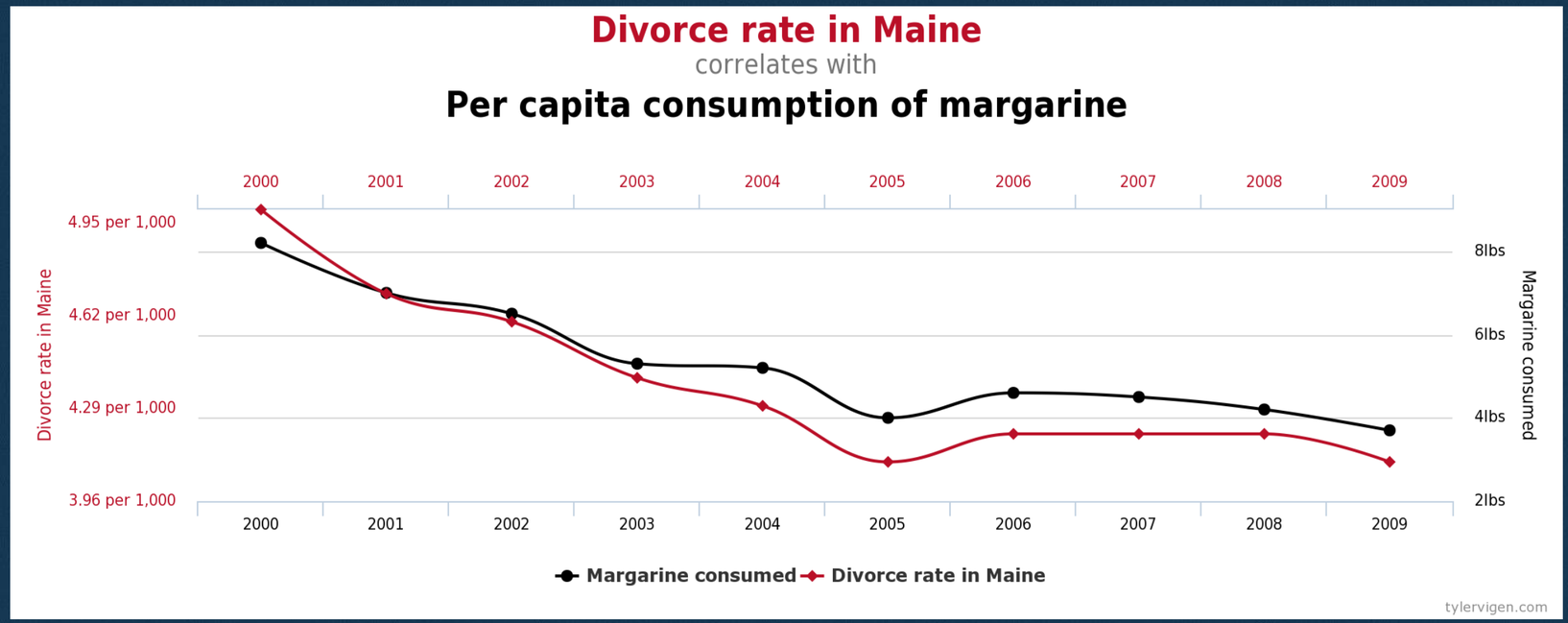
Correlações espúrias



Correlações espúrias



Correlações espúrias



Correlações espúrias

Site de correlações espúrias:

https://www.tylervigen.com/spurious-correlations

Causalidade?

Medir causalidade não é uma tarefa simples.

Alguns métodos que podem ser usados são: Experimentos Controlados Aleatorizados, Diferenças em Diferenças, Instrumental Variables (IV), Propensity Score Matching.

https://matheusfacure.github.io/python-causality-handbook/landing-page.html

https://github.com/rdemarqui/python-causality-handbook-ptbr