Анализ данных

Рита Голуб, Яндекс

ЛЭШ ILE 2022

Что изучает анализ данных?

Взаимосвязи между показателями

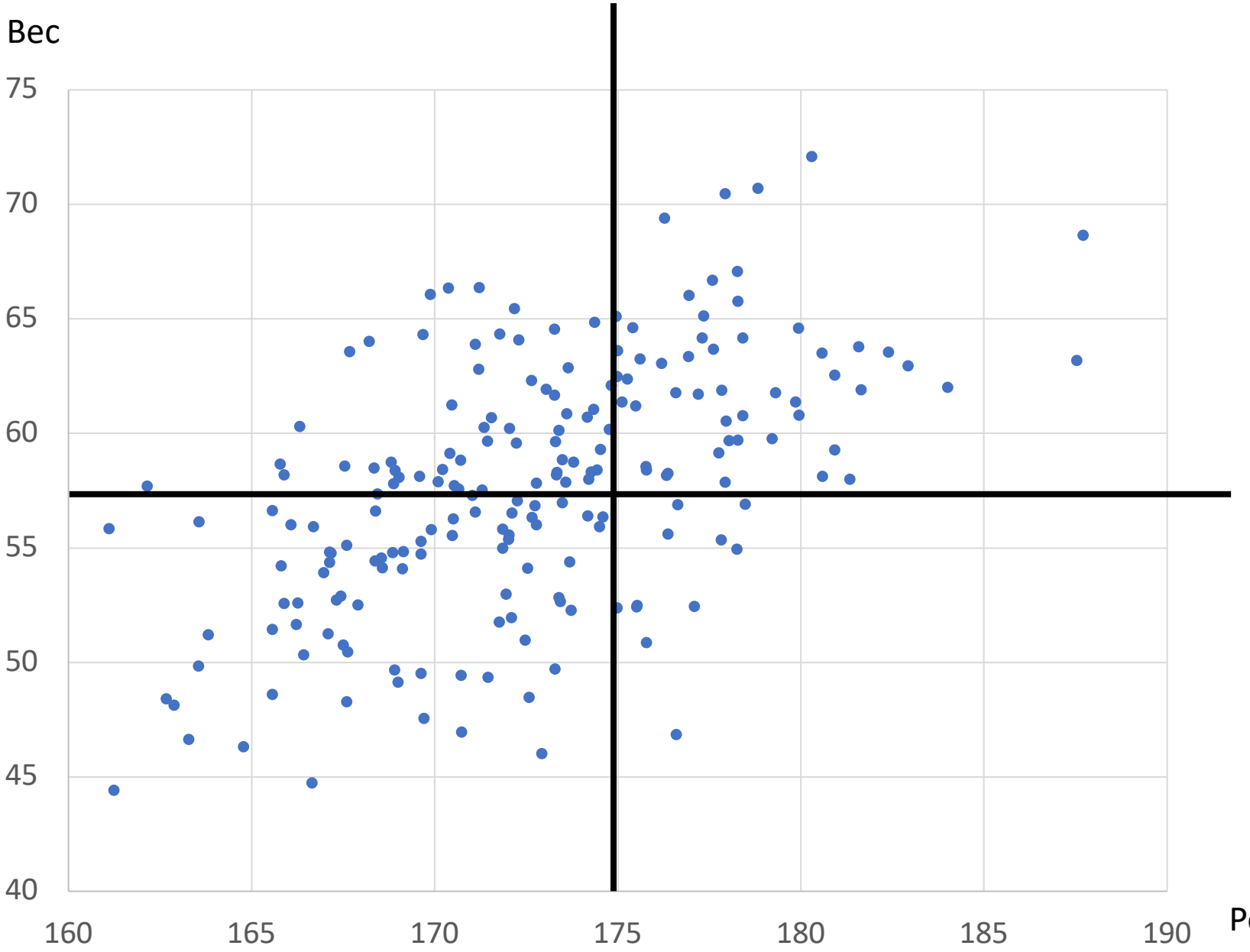
Какие взаимосвязи изучается при анализе данных?

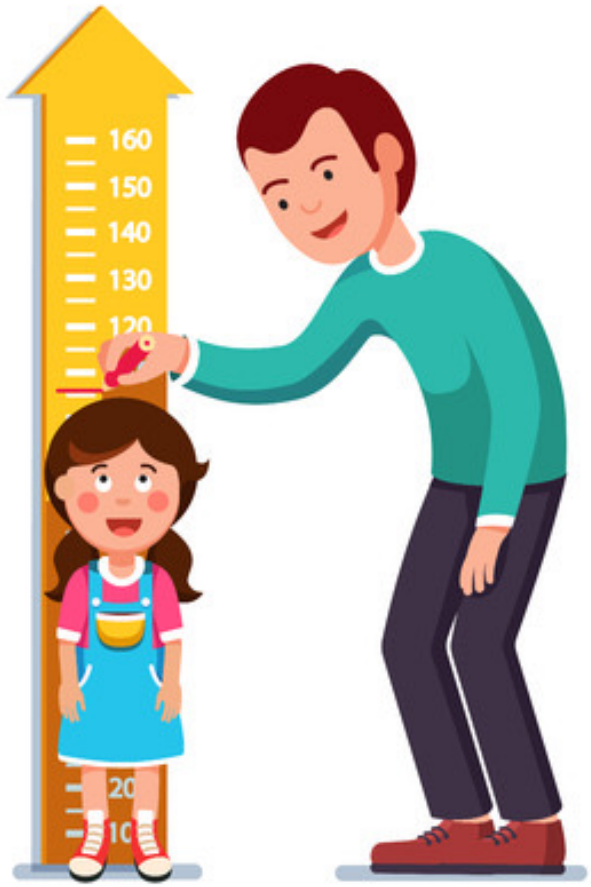
✓ Время года и температура воздуха

✓ День недели и загруженность дорог

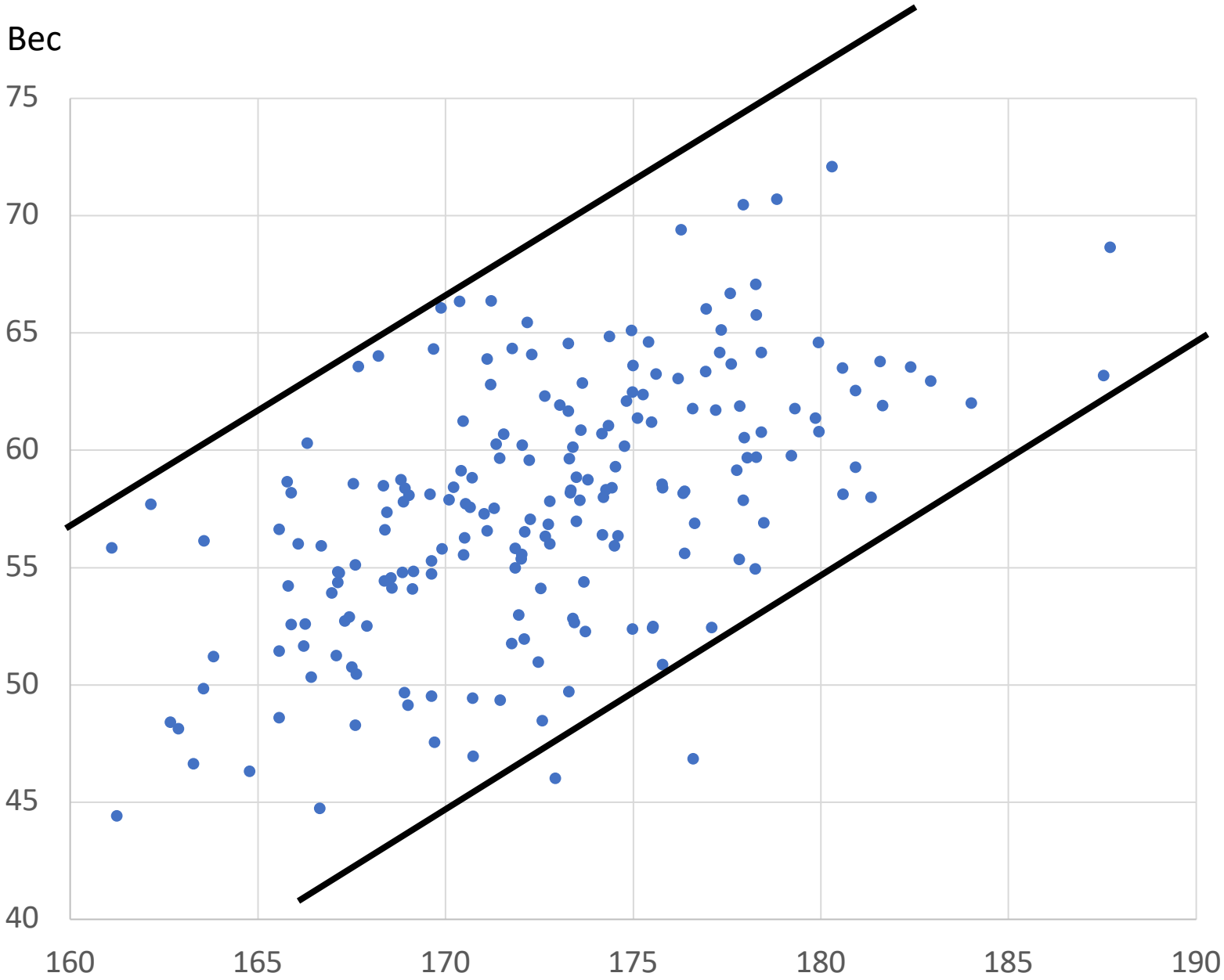
✓ в год и возраст

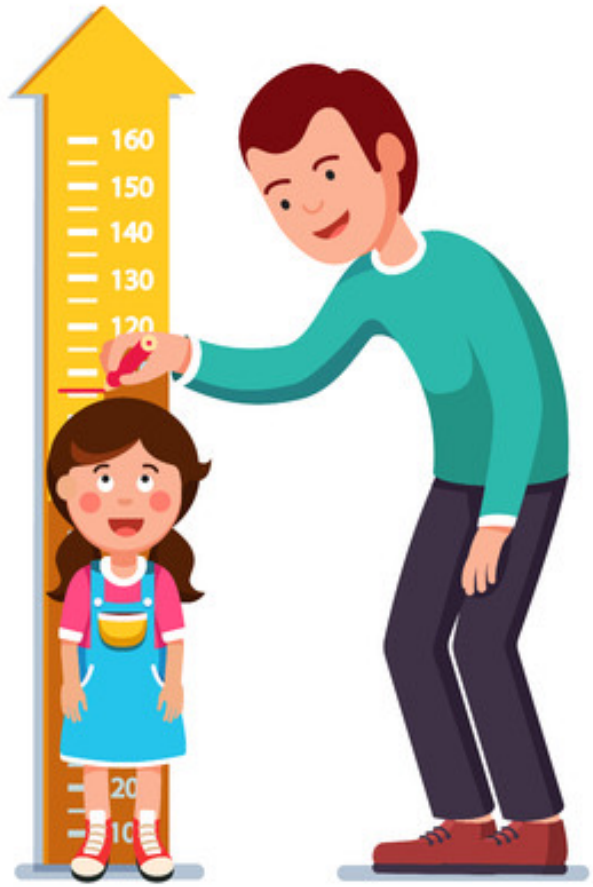
Рост-вес


Вес
75
70
65
60
55
50
45
40
160
165
170
175
180
185
190
Рост

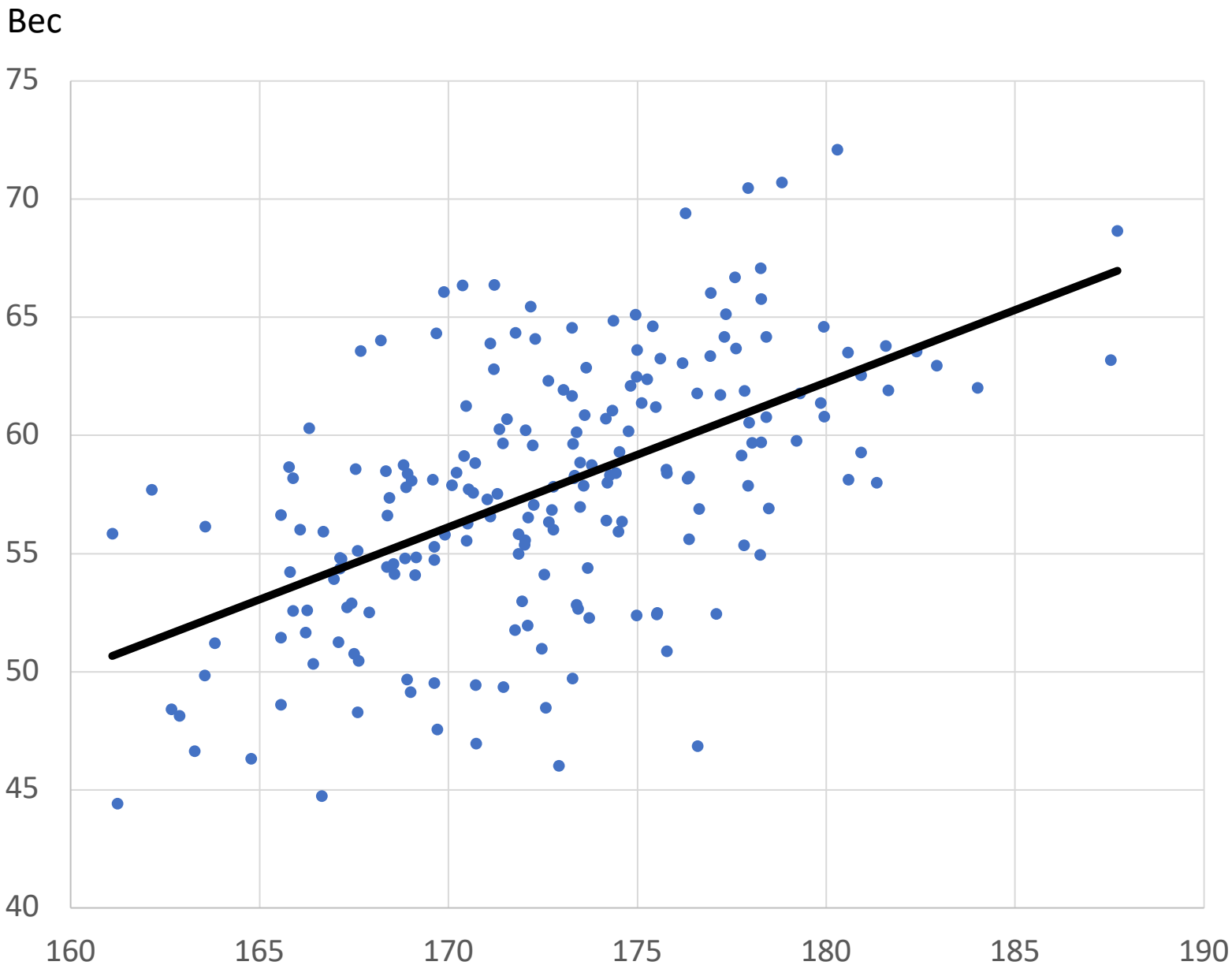

160
150
140
130
120
20
10

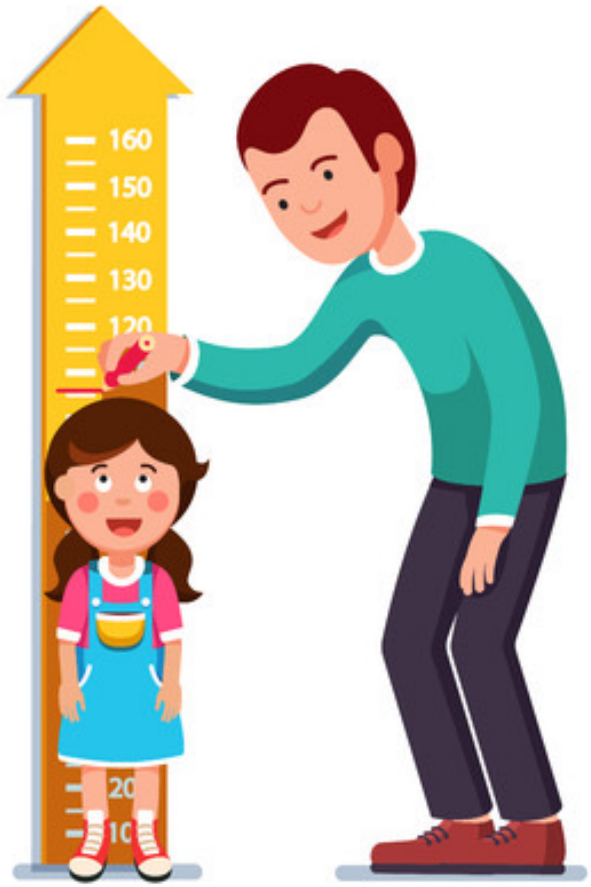
Рост-вес


Вес
75
70
65
60
55
50
45
40
160
165
170
175
180
185
190
Рост

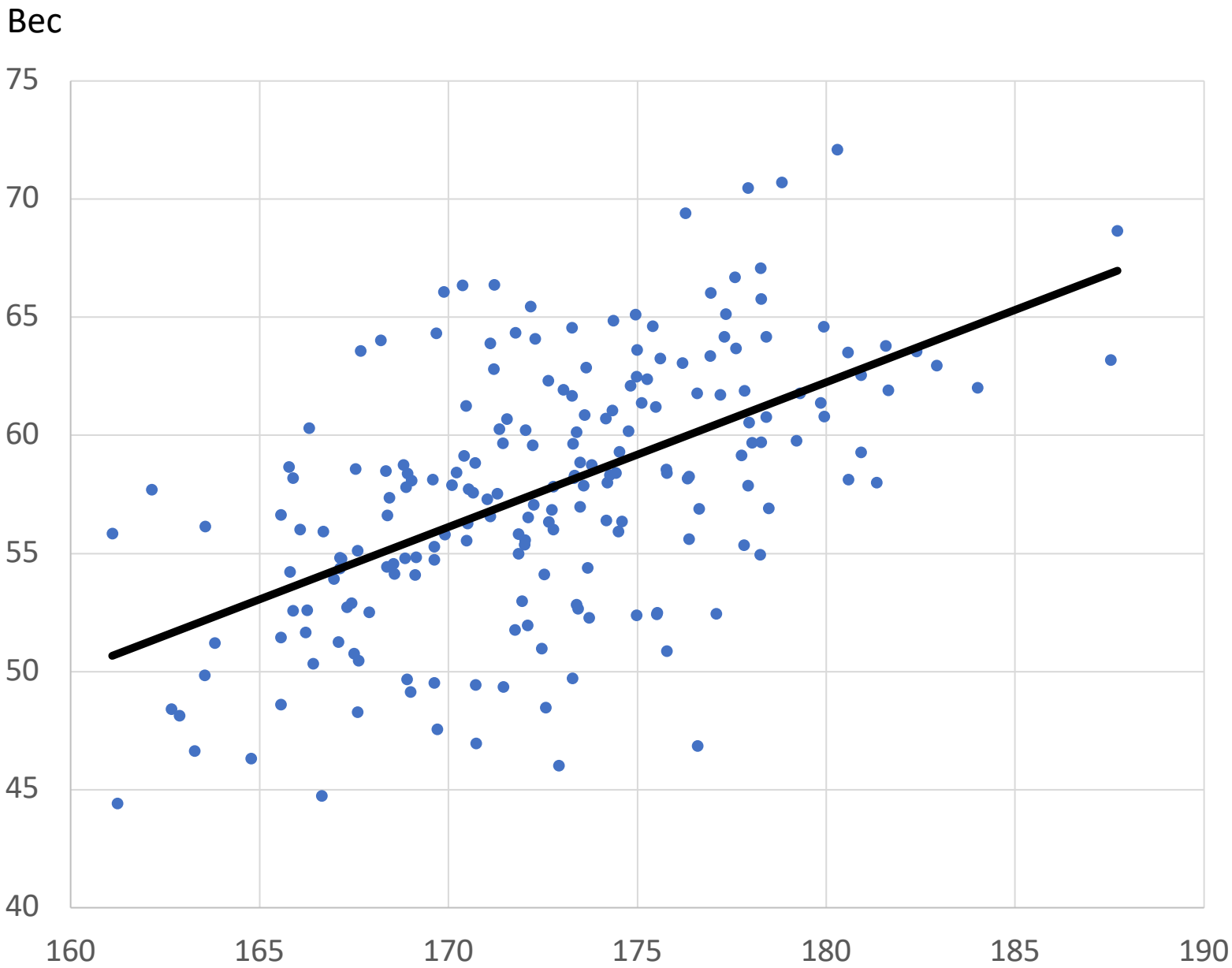

160
150
140
130
120
20
10

Рост-вес


Вес
75
70
65
60
55
50
45
40
160
165
170
175
180
185
190
Рост


160
150
140
130
120
20
10

Рост-вес



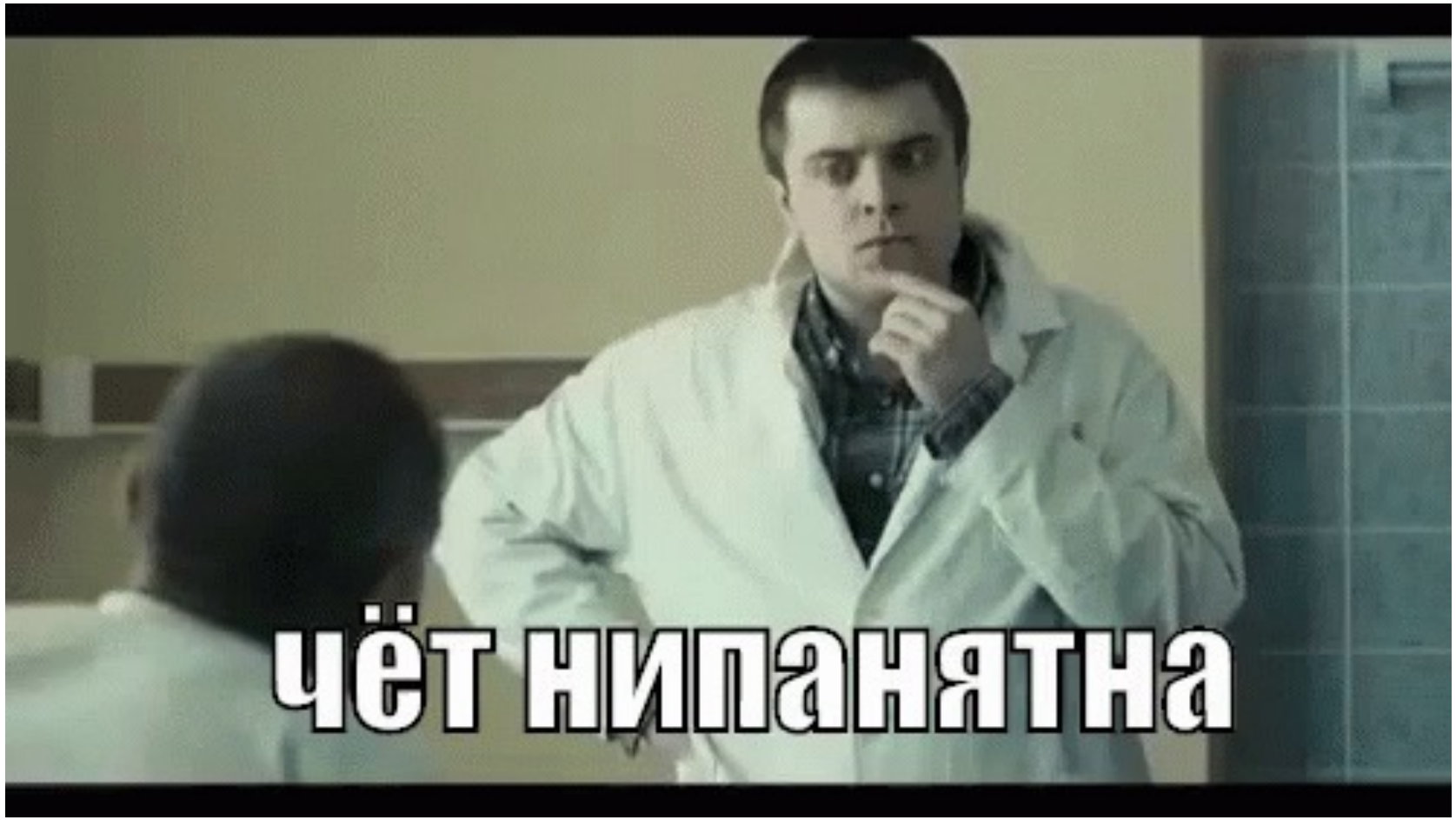
Как численно оценить линию тренда?

Коэффициент корреляции

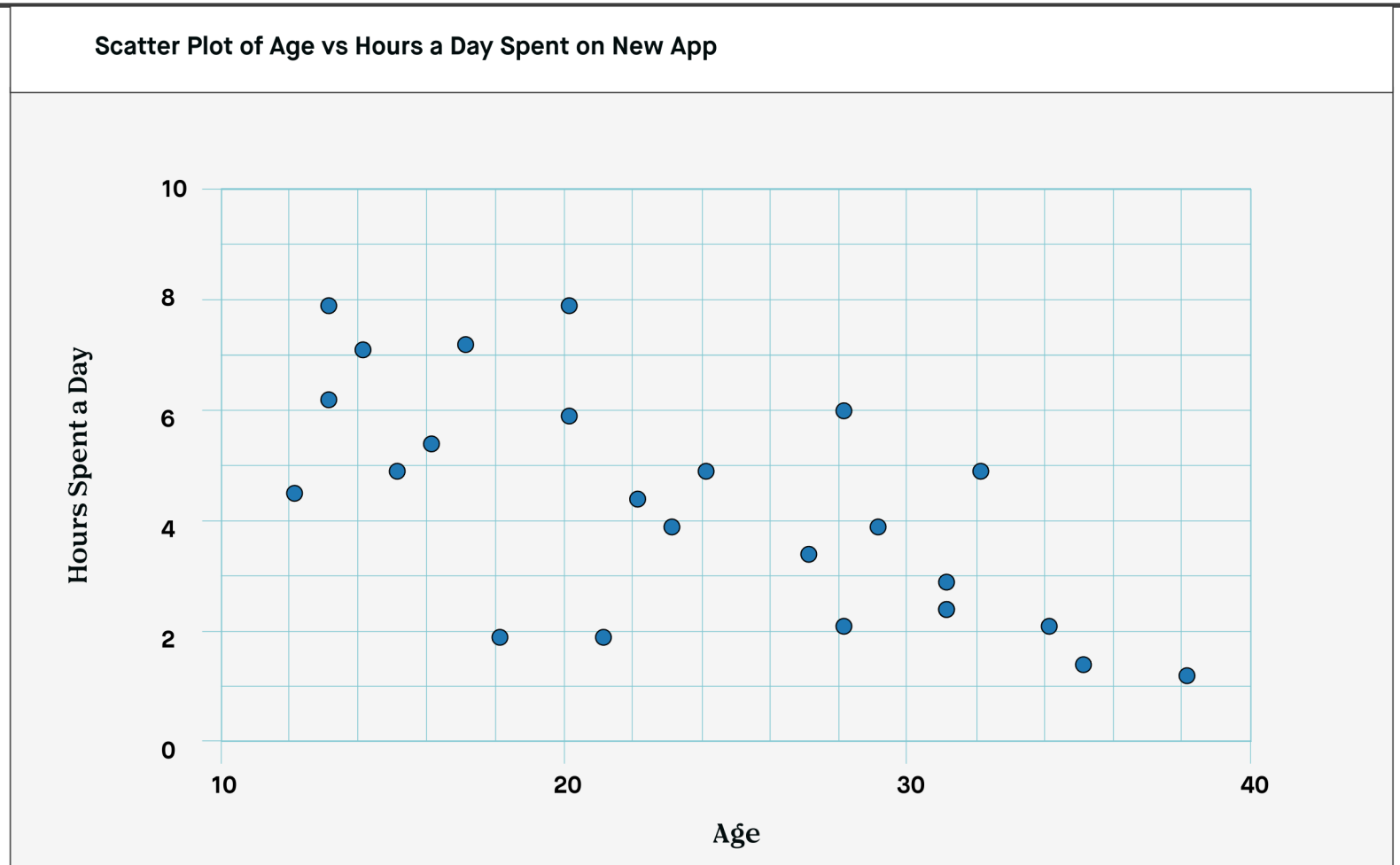
Corr = 0,56

И что делать с этим числом?
Как это интерпретировать?

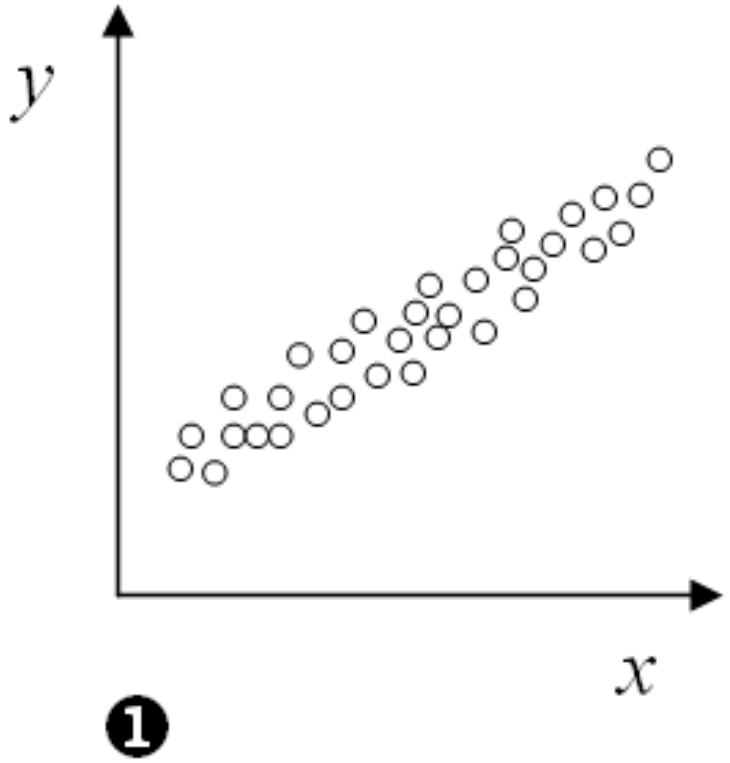
Corr = 0,56



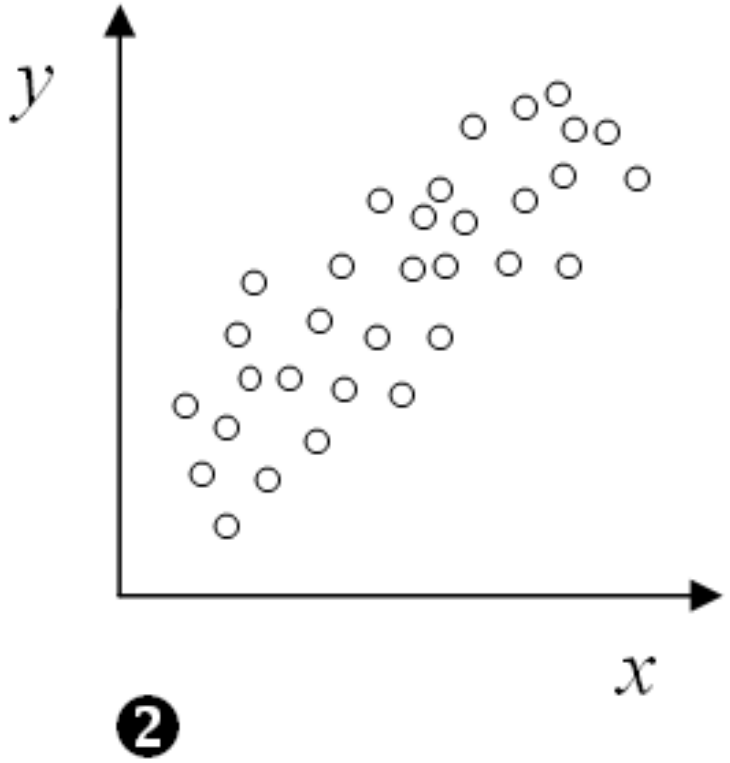
Больше примеров! Отрицательная корреляция



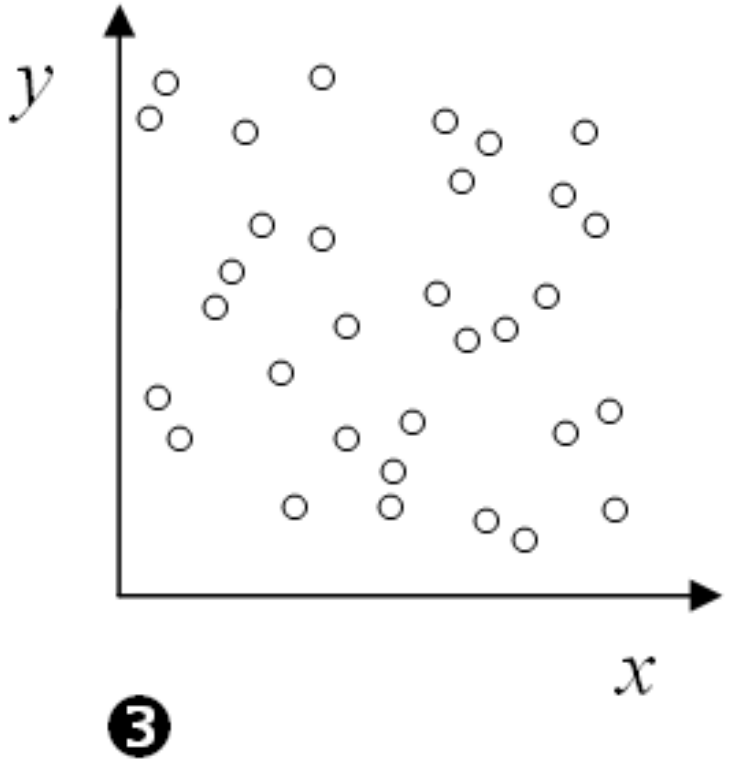
Больше примеров! Слабая корреляция



Corr ≈ 1

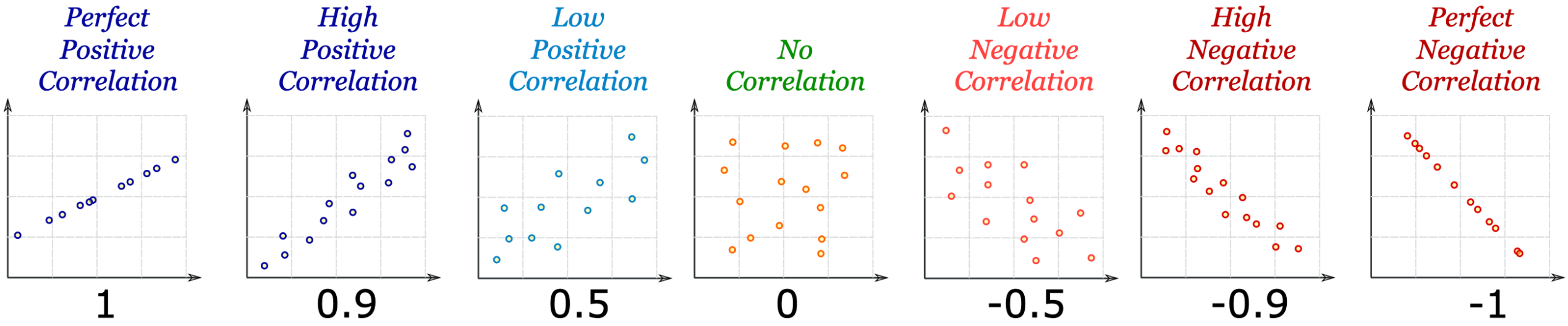


Corr ≈ 0.5



Corr ≈ 0

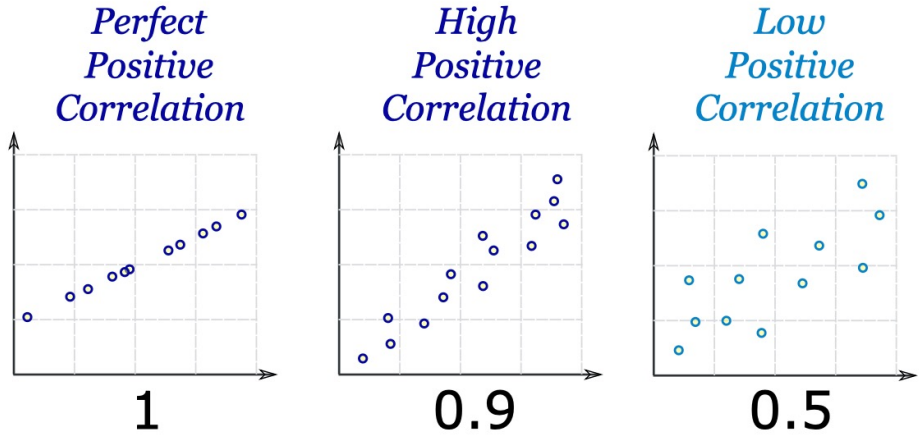
Больше примеров! Summary



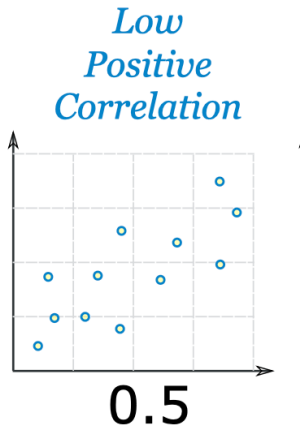
И что делать с этим числом? Как это интерпретировать?

Corr = 0,56

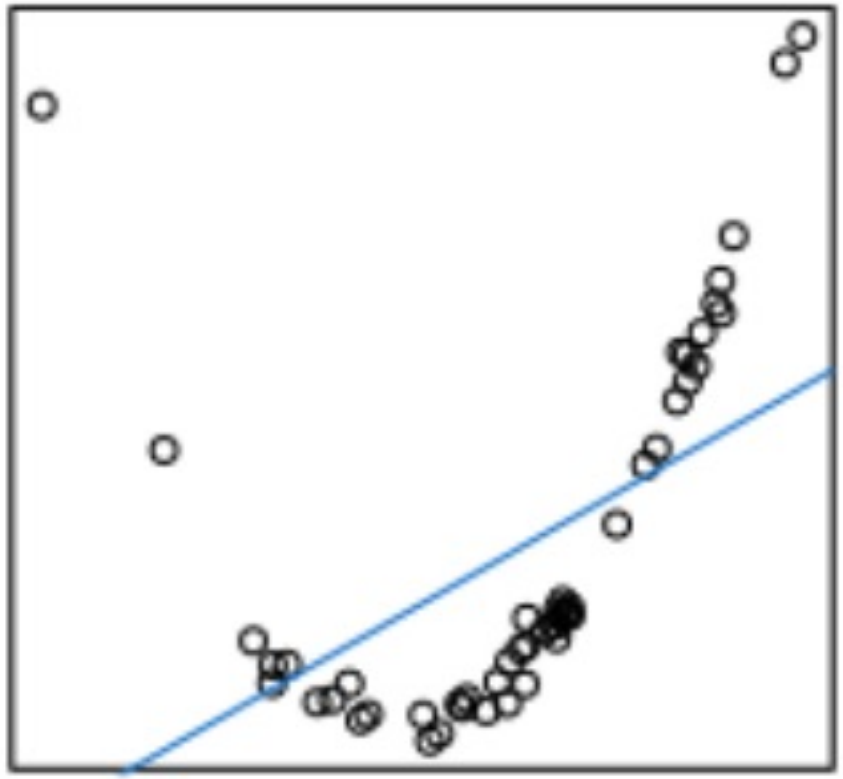
- Corr = 0,56 > 0 ⇒ положительная взаимосвязь

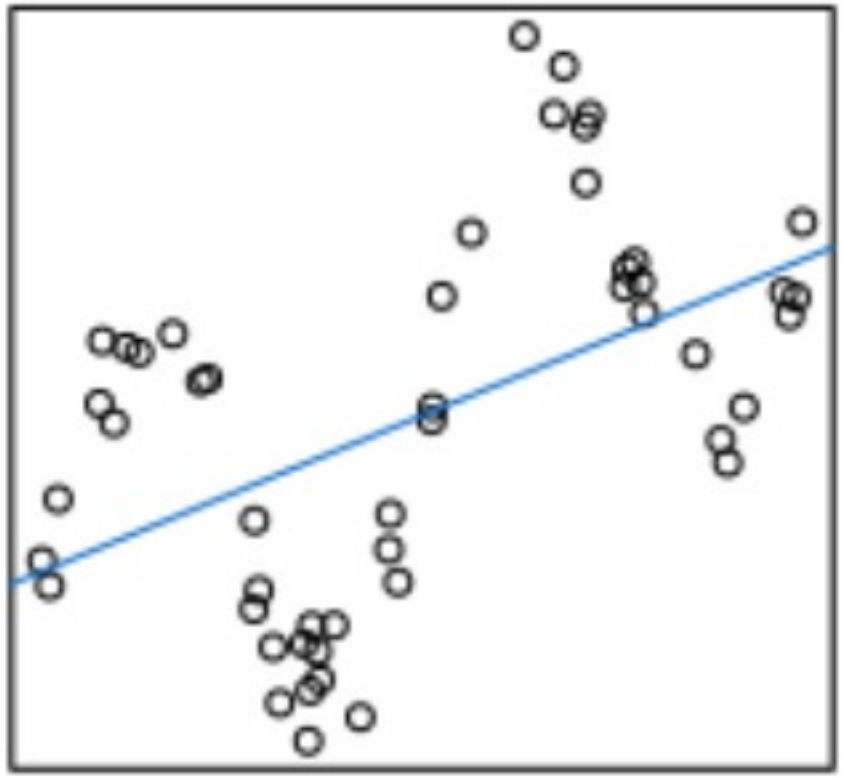


- Corr = 0,56 < 1 ⇒ вес зависит не только от роста

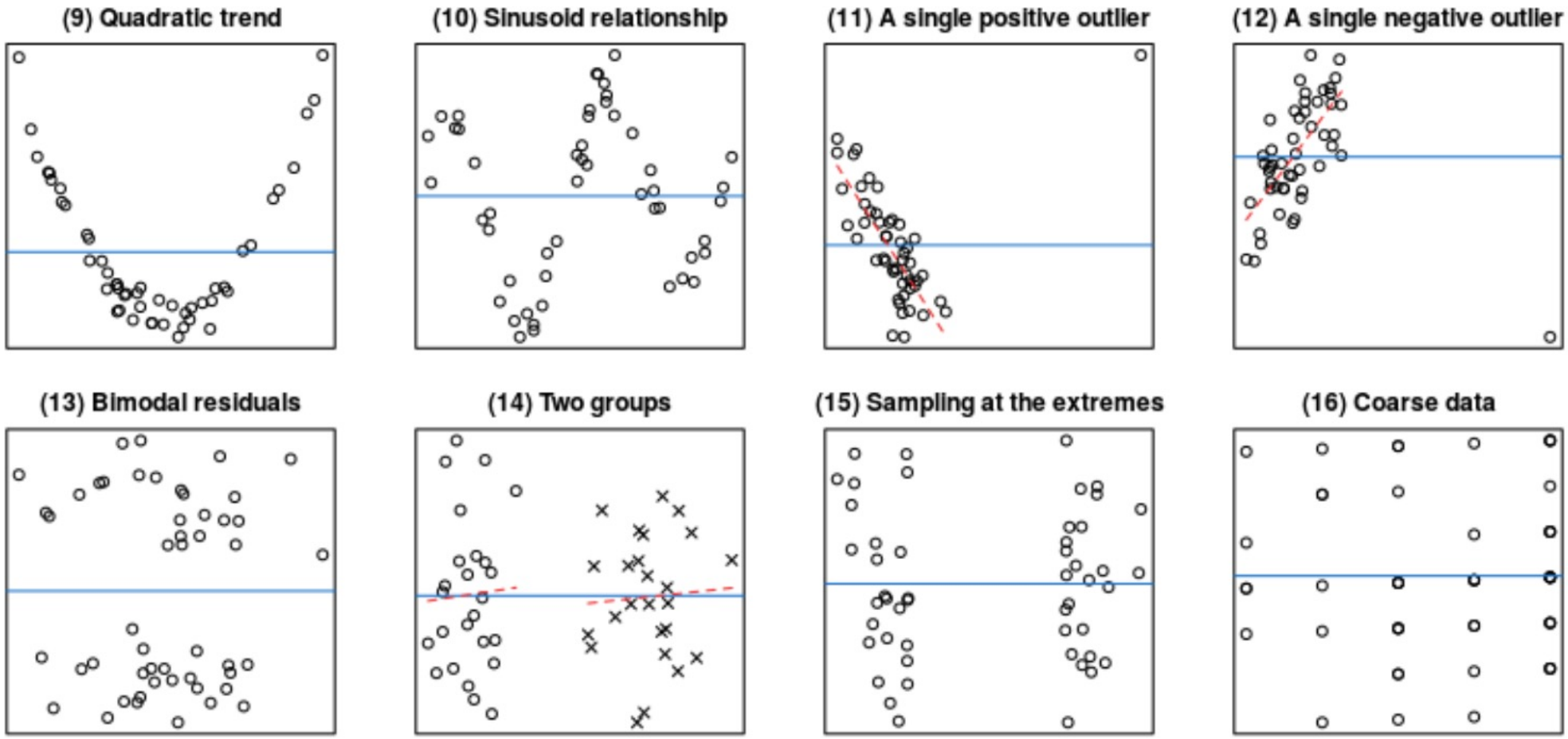


А какая корреляция здесь?

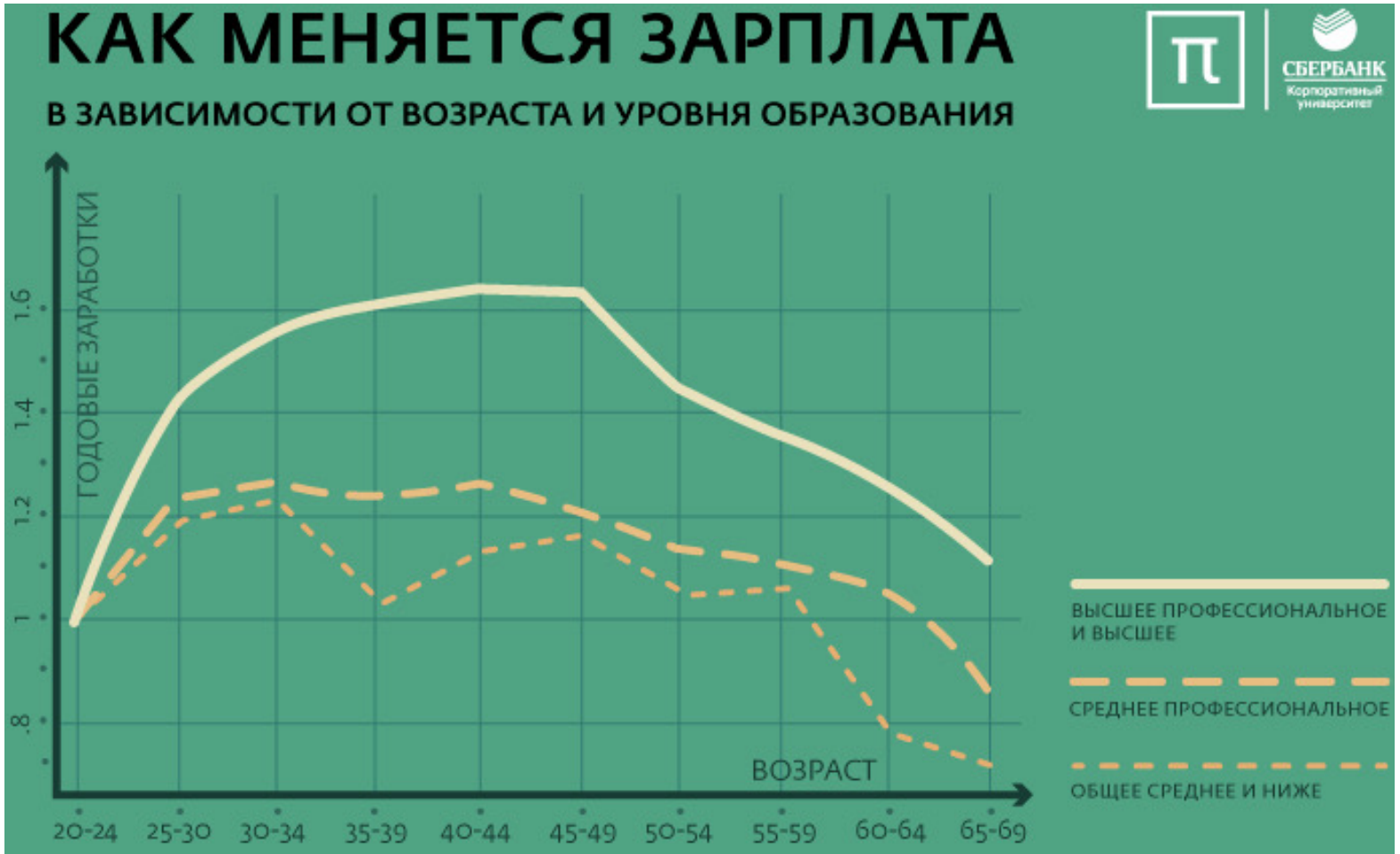




Corr = 0. Неужели нет взаимосвязи?

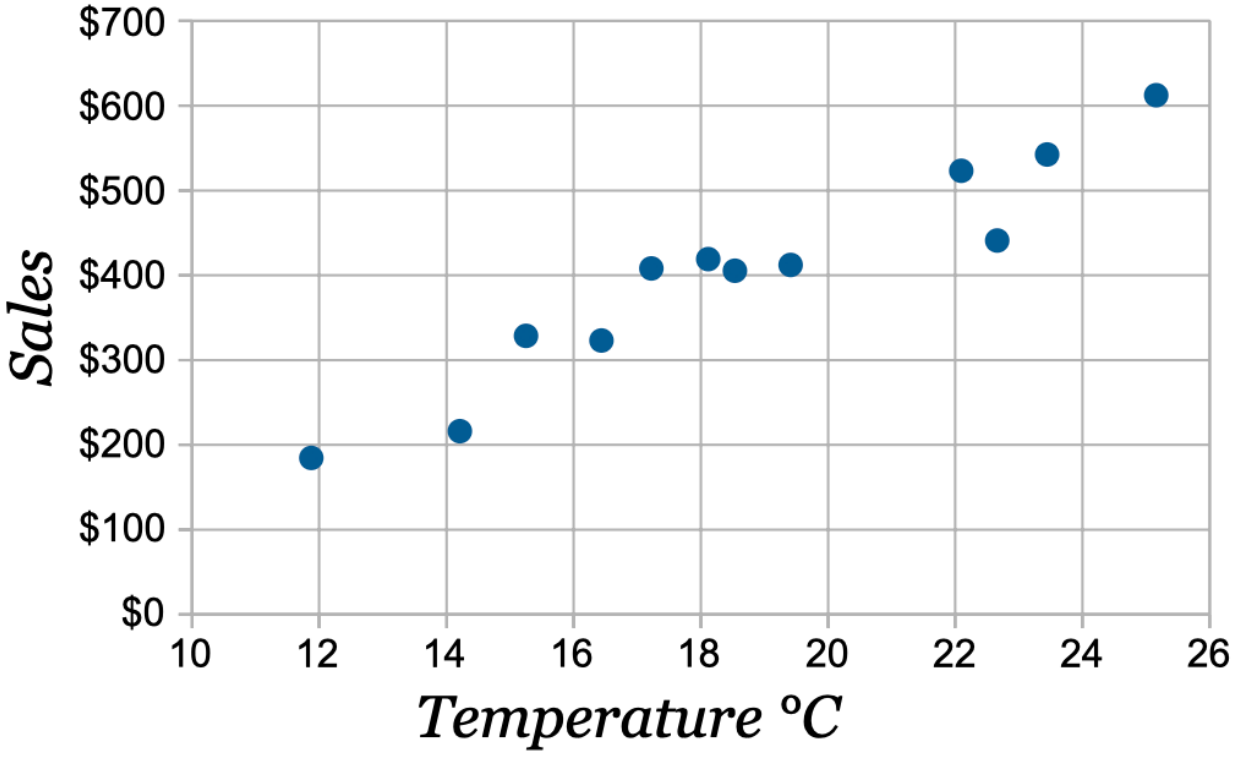


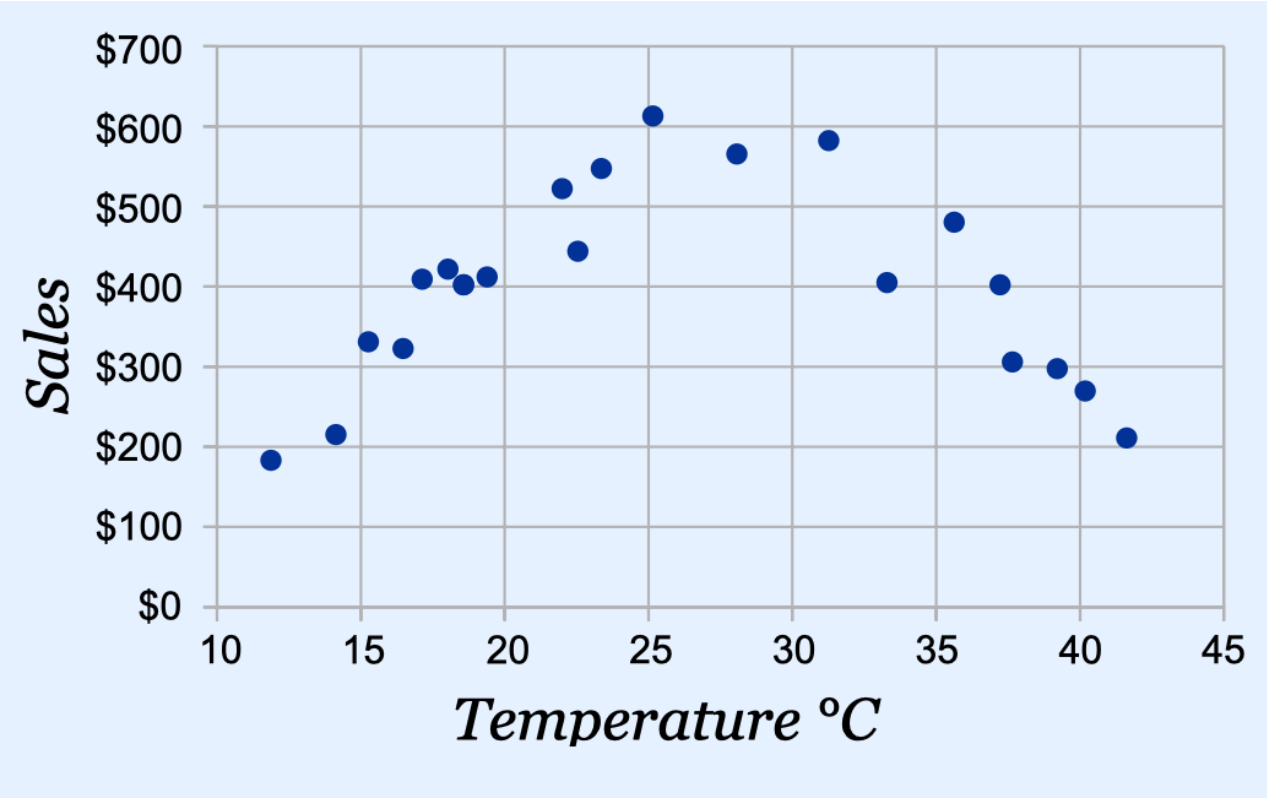
Больше примеров! Нелинейная взаимосвязь



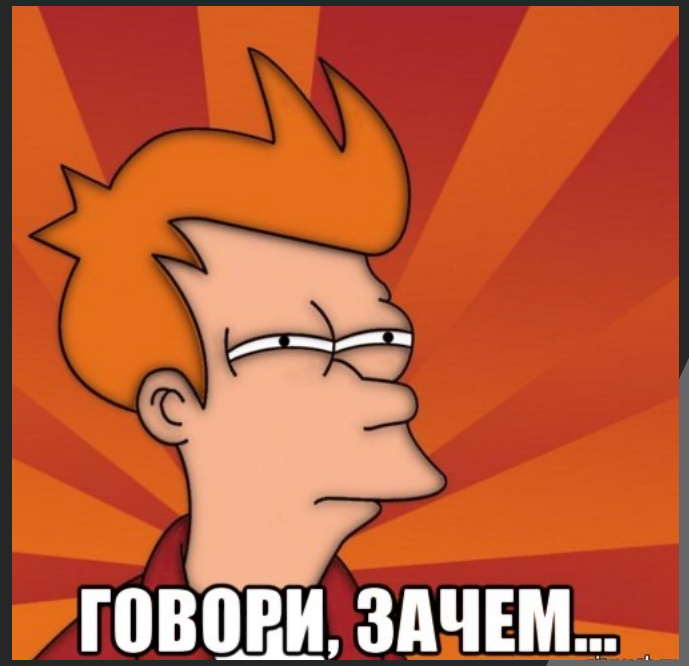
Больше примеров! Нелинейная взаимосвязь

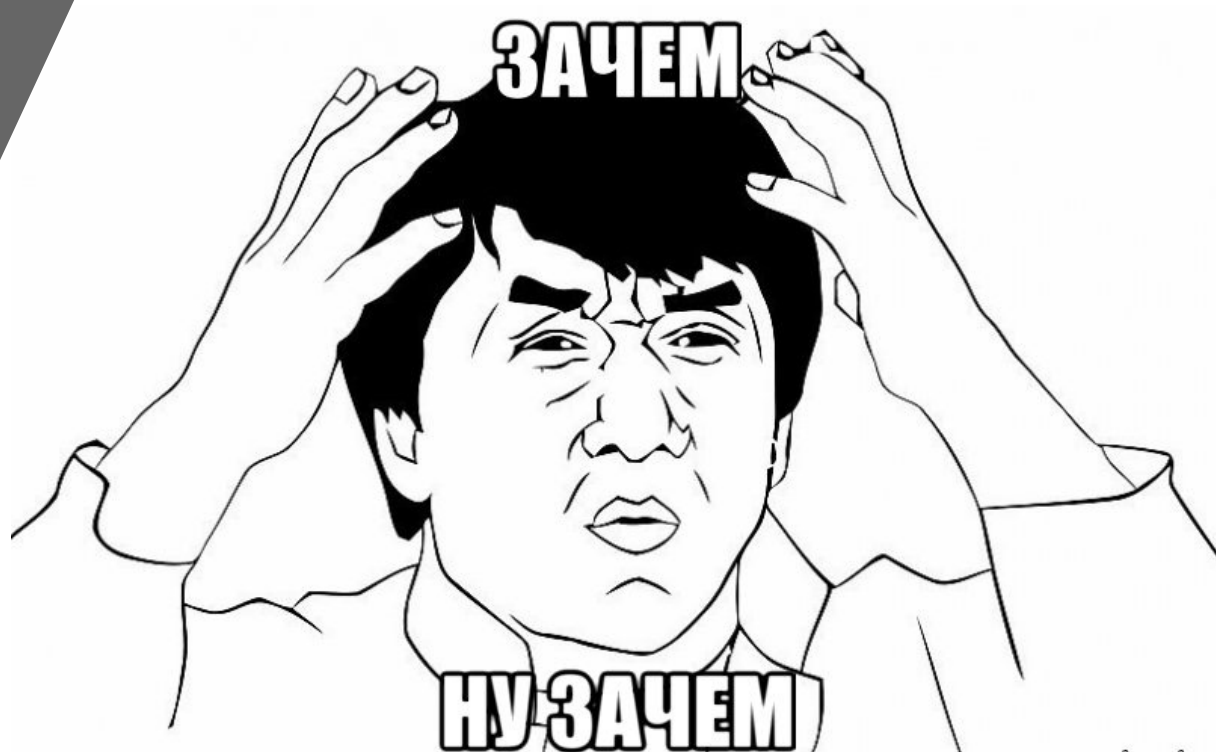
Продажи мороженого

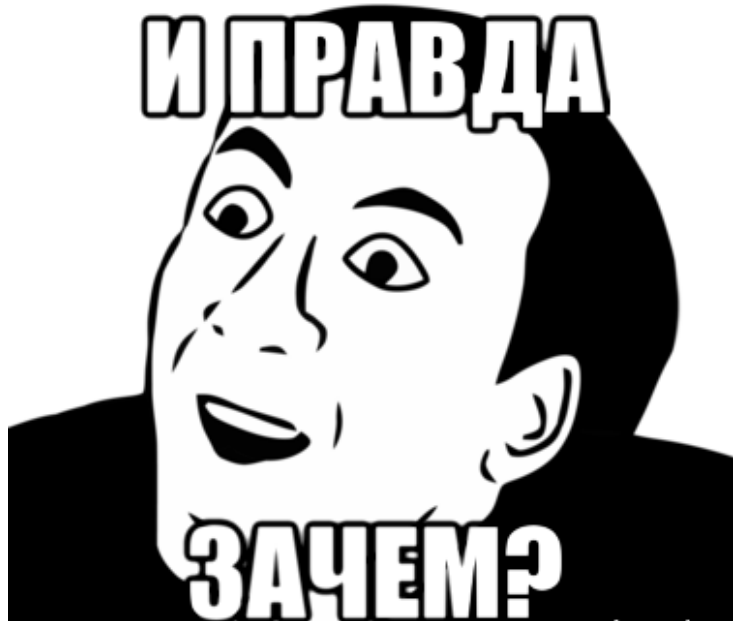




Зачем нужен анализ данных?

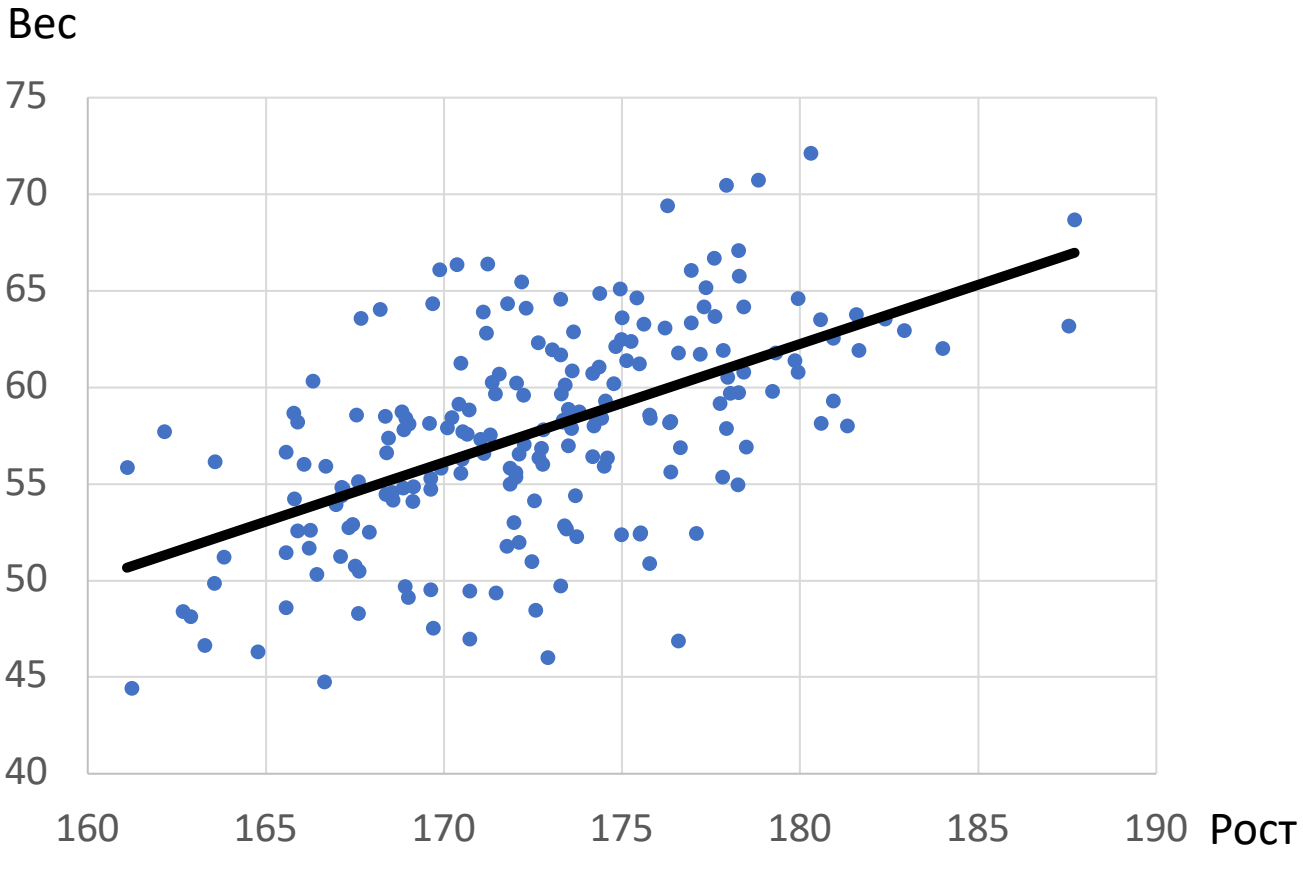


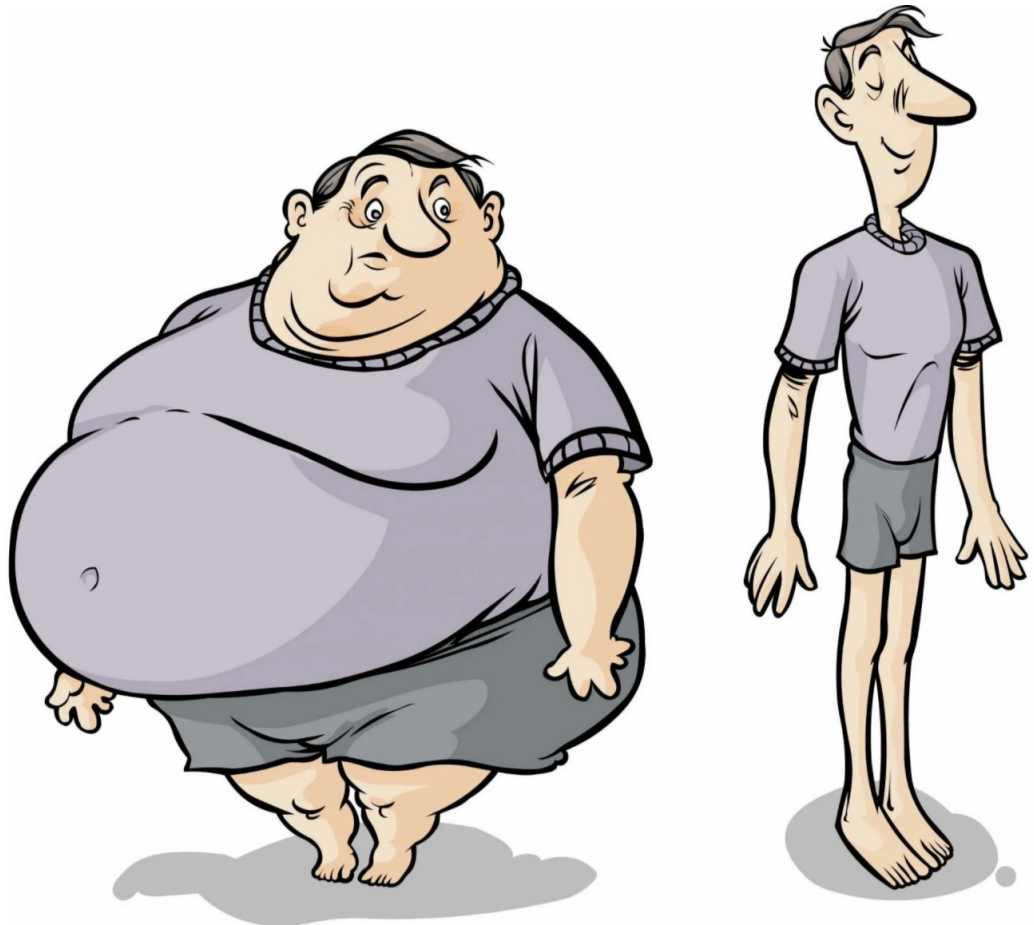




Рост-вес

- Увеличение веса ⇒ увеличение роста?
- Увеличение роста ⇒ увеличение веса?





Полиция в США

Количество полицейских в городе и количество преступлений скоррелировано

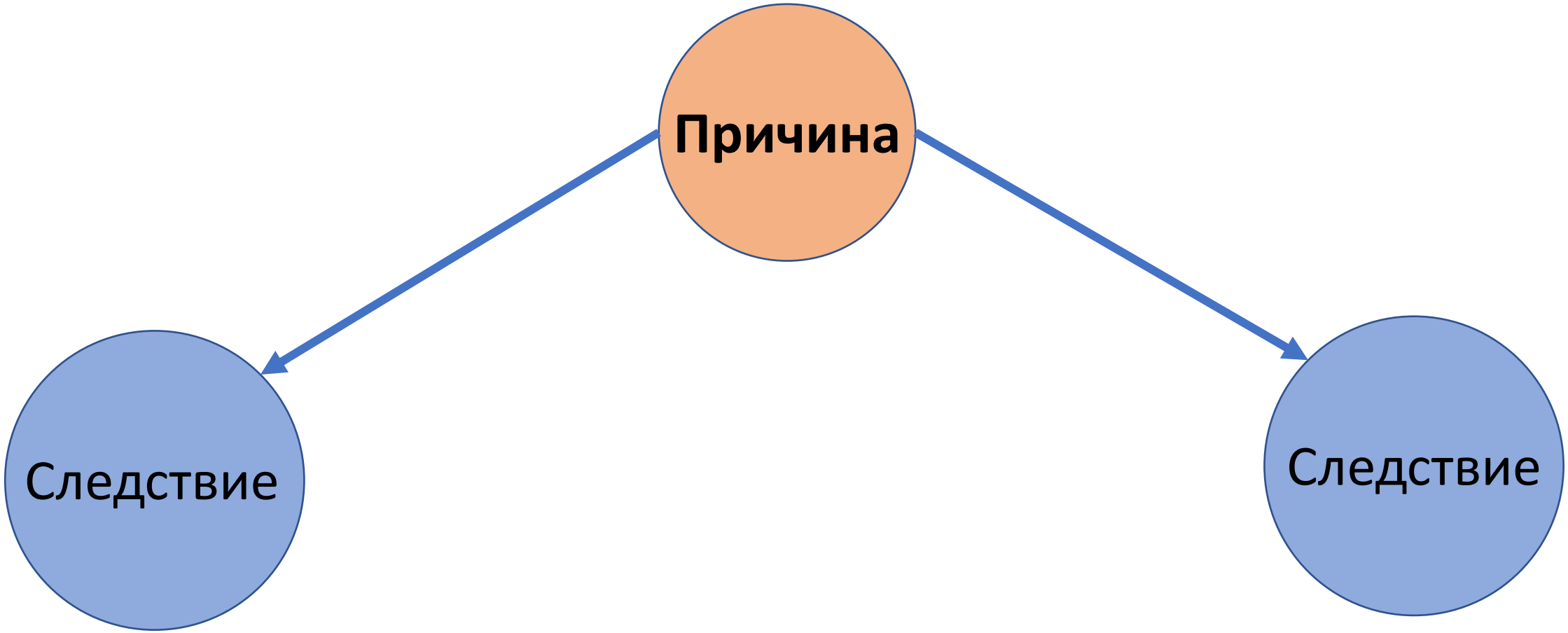
Избегание врачей в средневековье

После появления врача в поселении многие люди там умирали

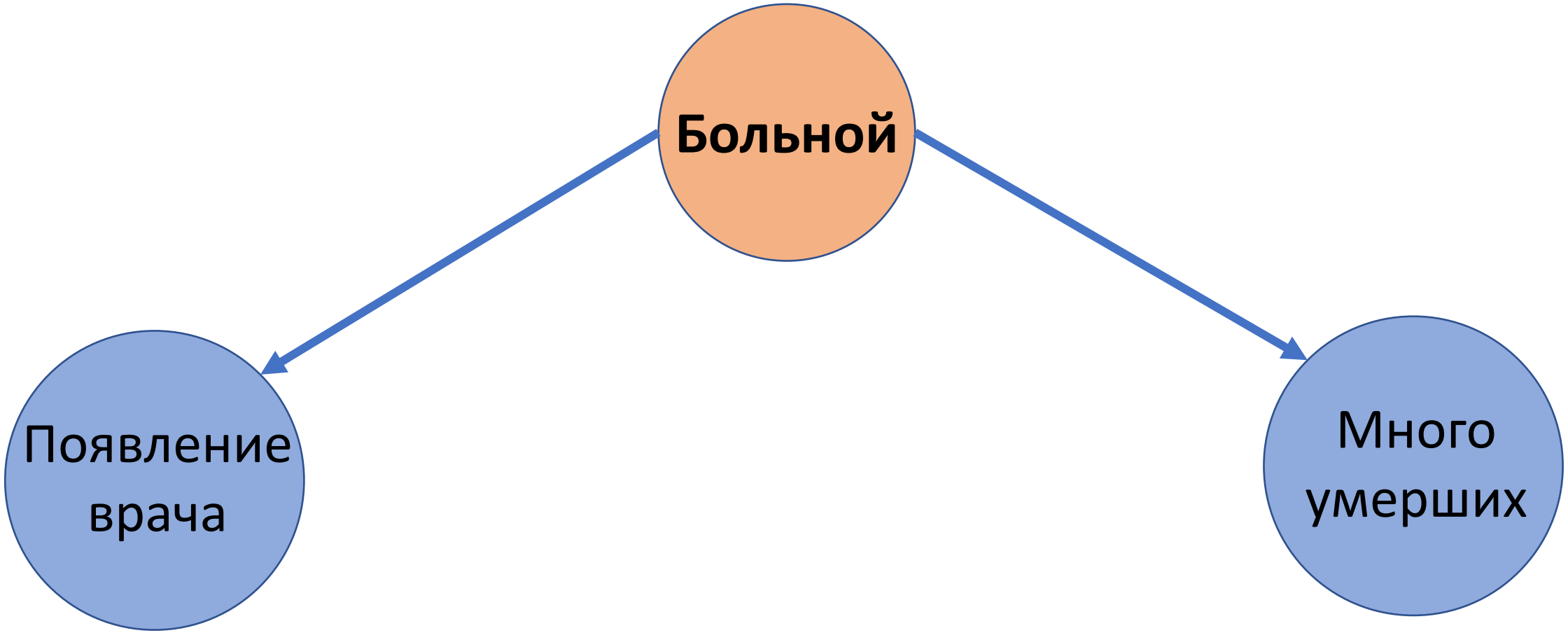
Примеры прямой взаимосвязи

Примеры прямой взаимосвязи

Одна причина и несколько следствий

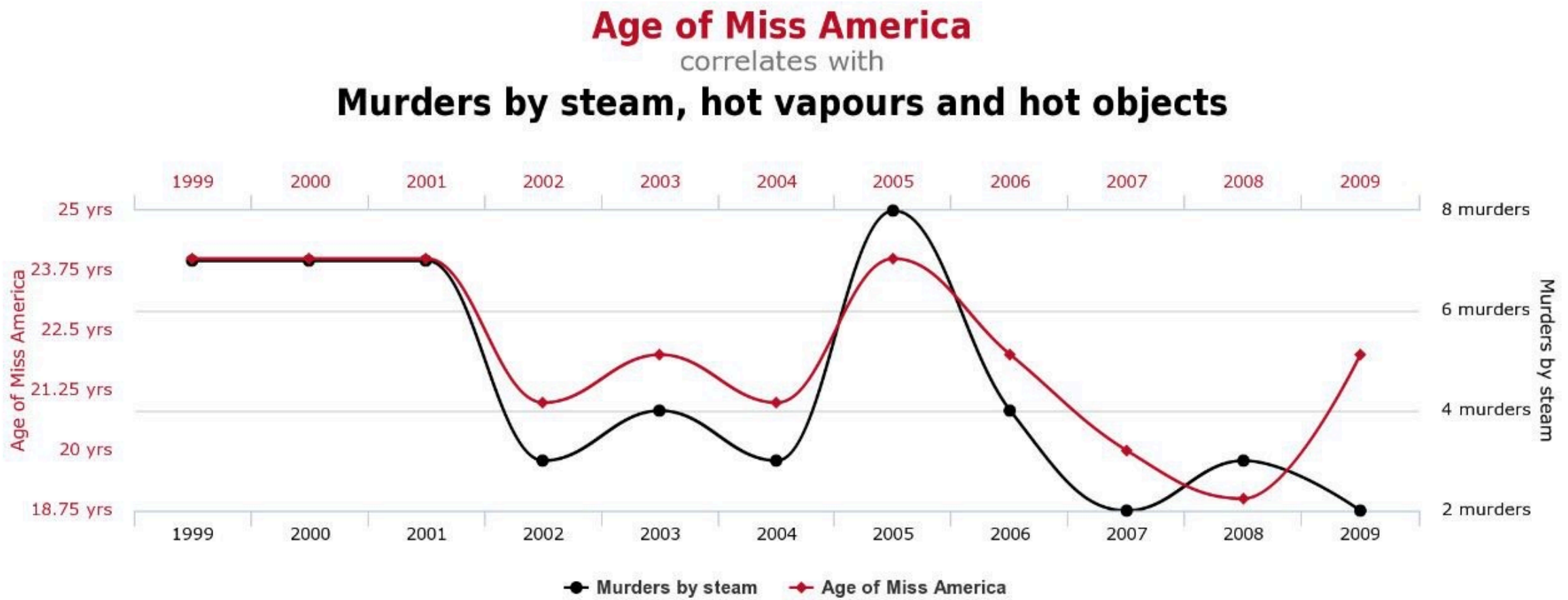


Одна причина и несколько следствий



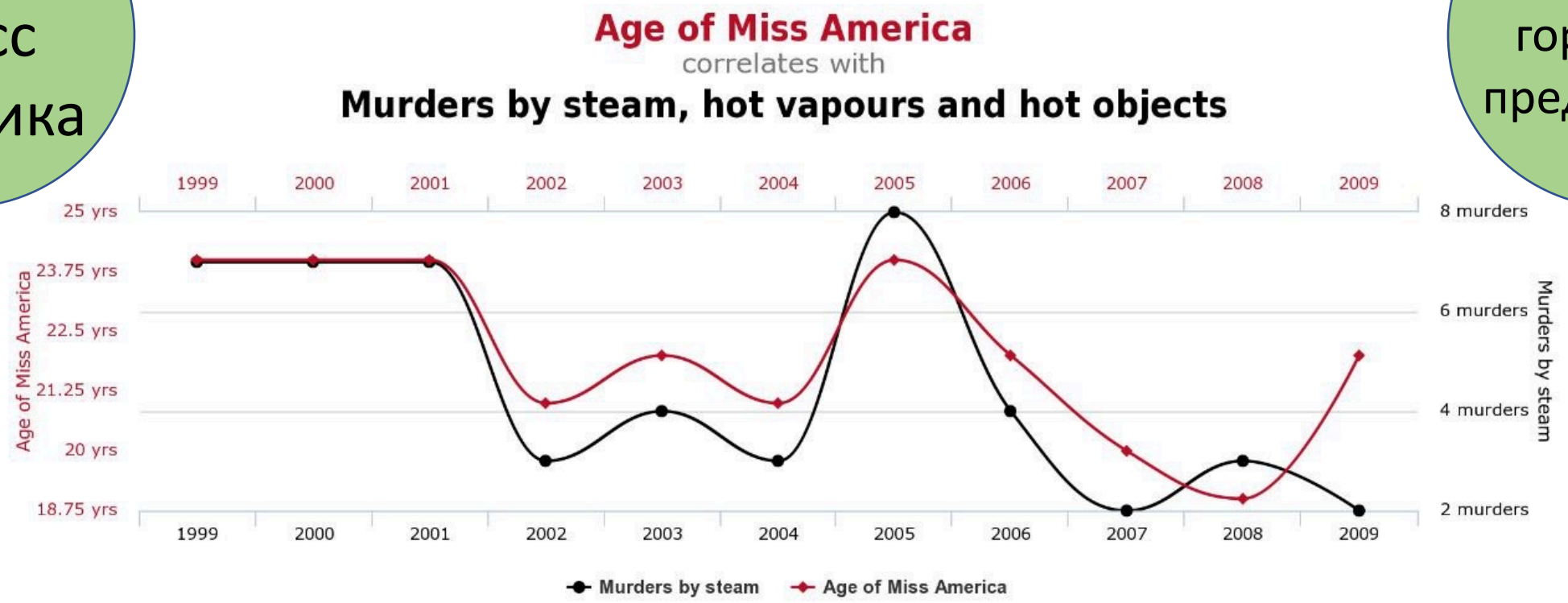
Примеры ложной взаимосвязи

Примеры ложной взаимосвязи



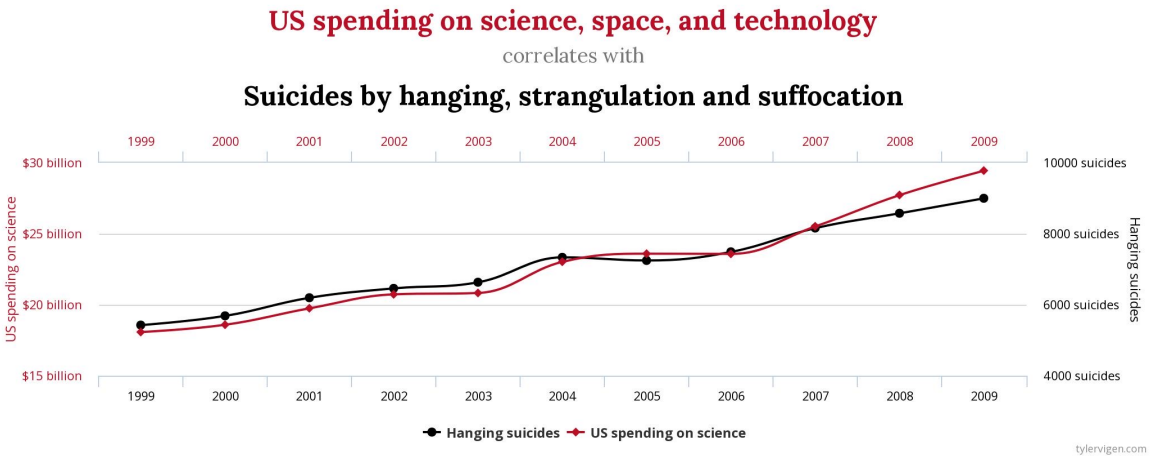
Примеры ложной взаимосвязи

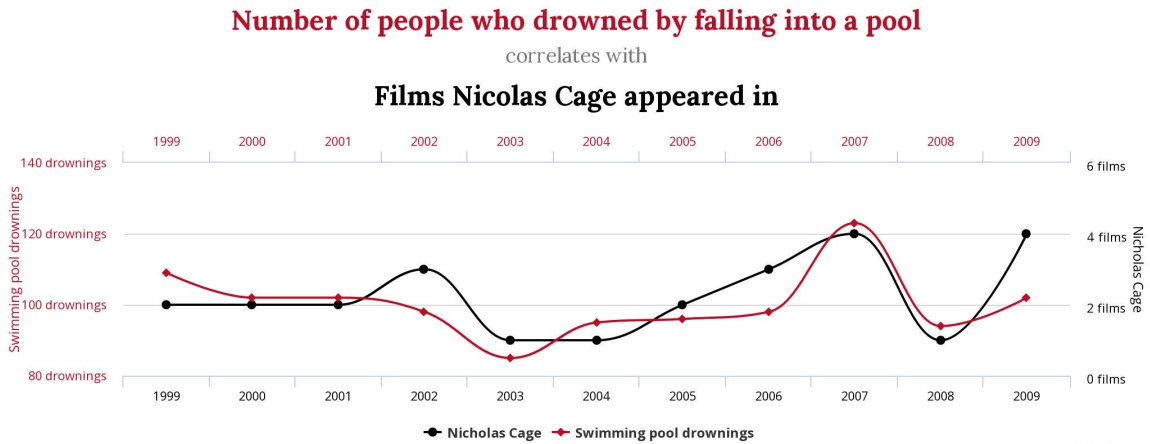
Возраст Мисс Америка

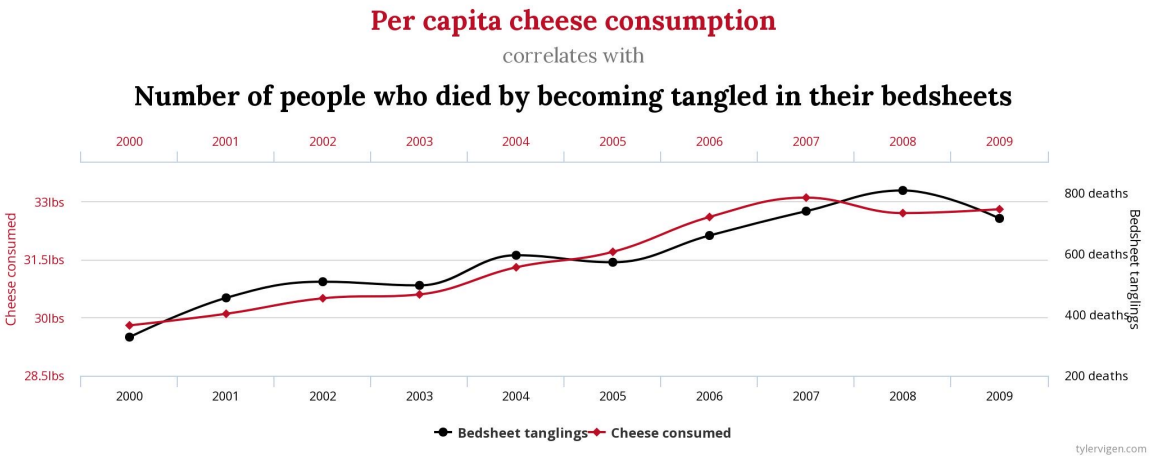


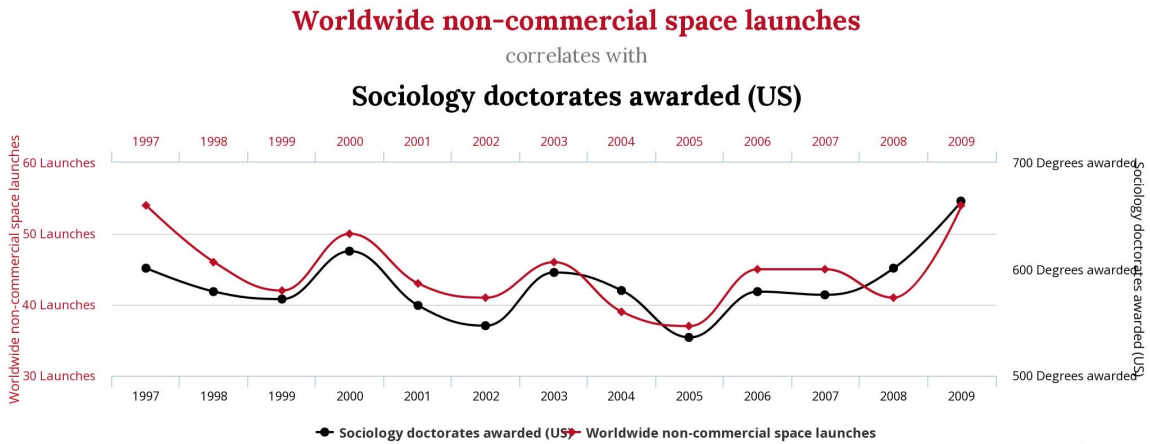
Убийства горячими предметами

Примеры ложной взаимосвязи









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

Зацикленная взаимосвязь

Зацикленная взаимосвязь

Как понять, какой именно вид взаимосвязи

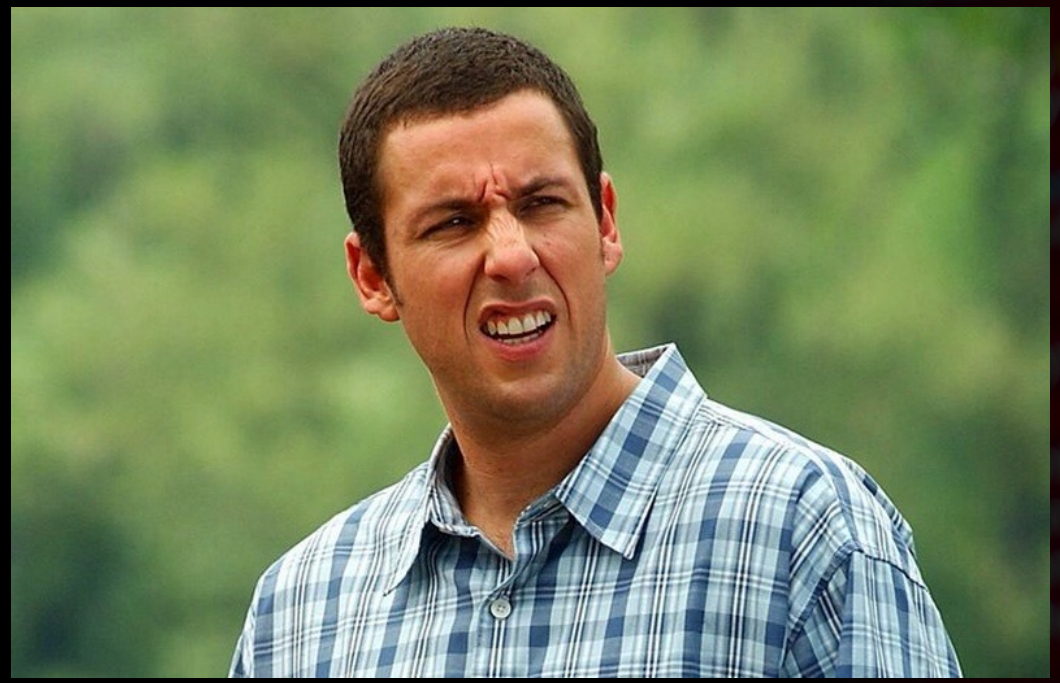
Как понять, какой именно вид взаимосвязи

Работники на складе

- Сборка заказов на складе состоит из нескольких этапов (отбор, сортировка, **упаковка**, доставка)
- Улучшаем процесс на этапе **упаковки**
- Количество **упакованных** коробок за единицу времени увеличилось на 10%

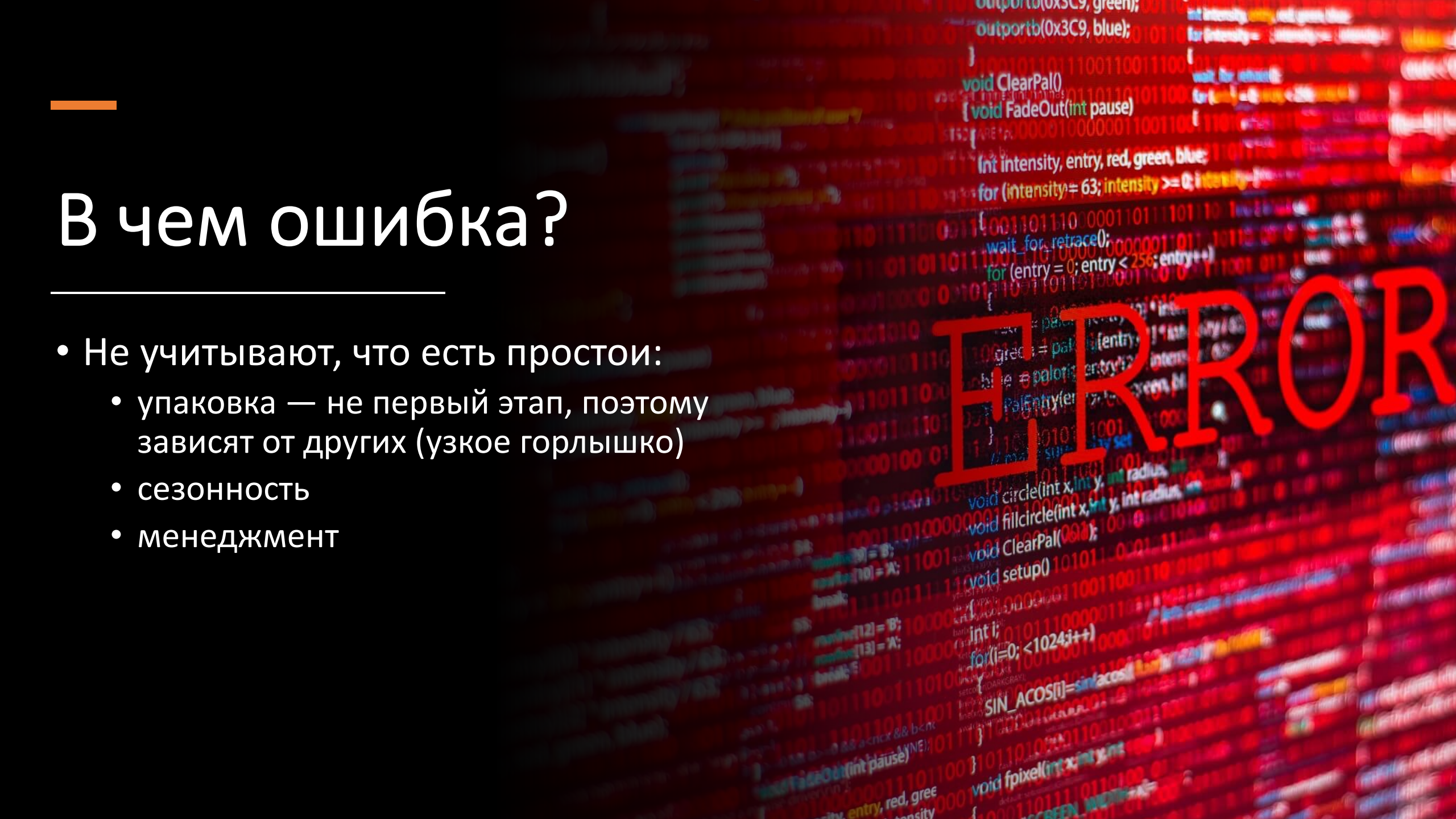
Авторитетный аналитик ЕК сказал, что это цифра завышена

В чем ошибка?



В чем ошибка?

- Не учитывают, что есть простои:
  - упаковка — не первый этап, поэтому зависят от других (узкое горлышко)
  - сезонность
  - менеджмент



А как можно было?

Смотреть на увеличение скорости при максимальной производительности

И где это применять?

- В жизни!
- На олимпиадах!
  - DANO
  - Яндекс.Учебник

Яндекс Учебник

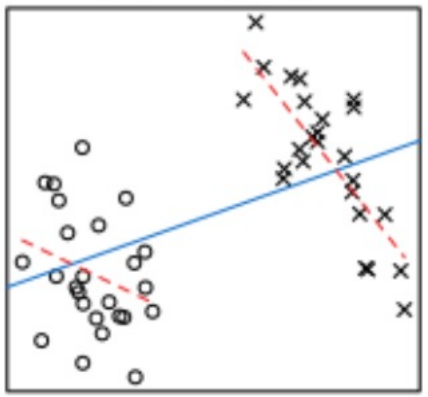
DANO

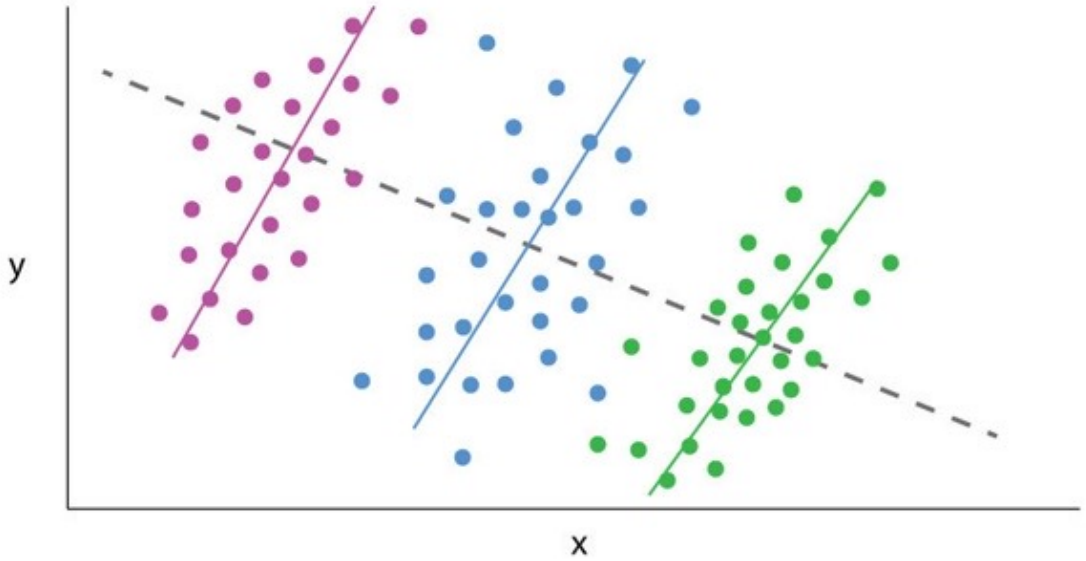
Анализ данных

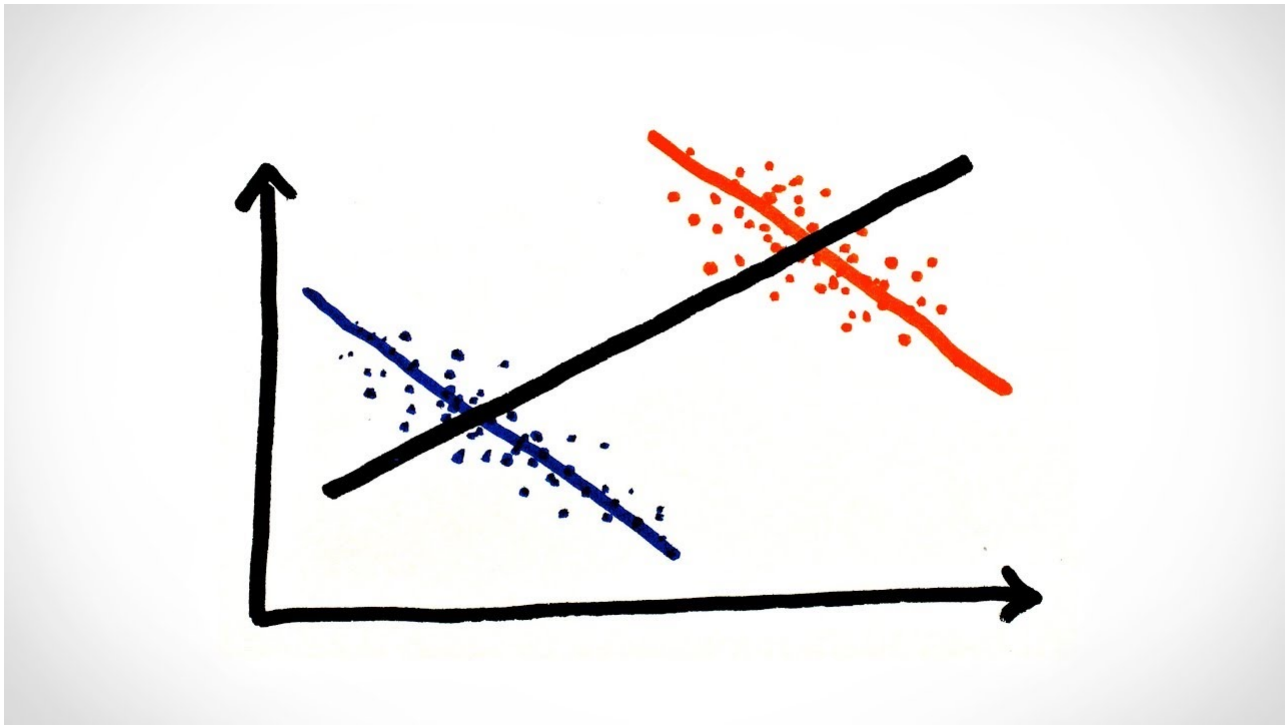
Рита Голуб, Яндекс

ЛЭШ ILE 2022

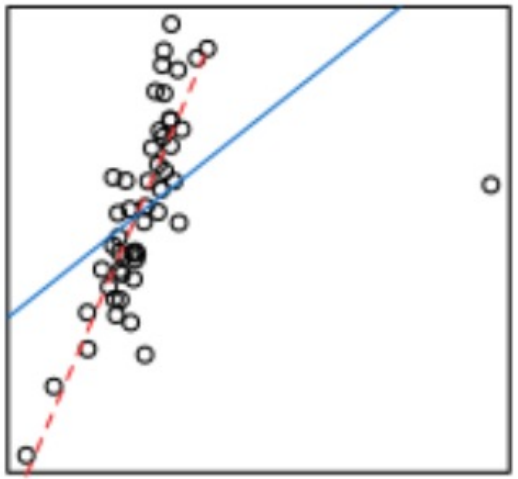
Парадокс Симпсона

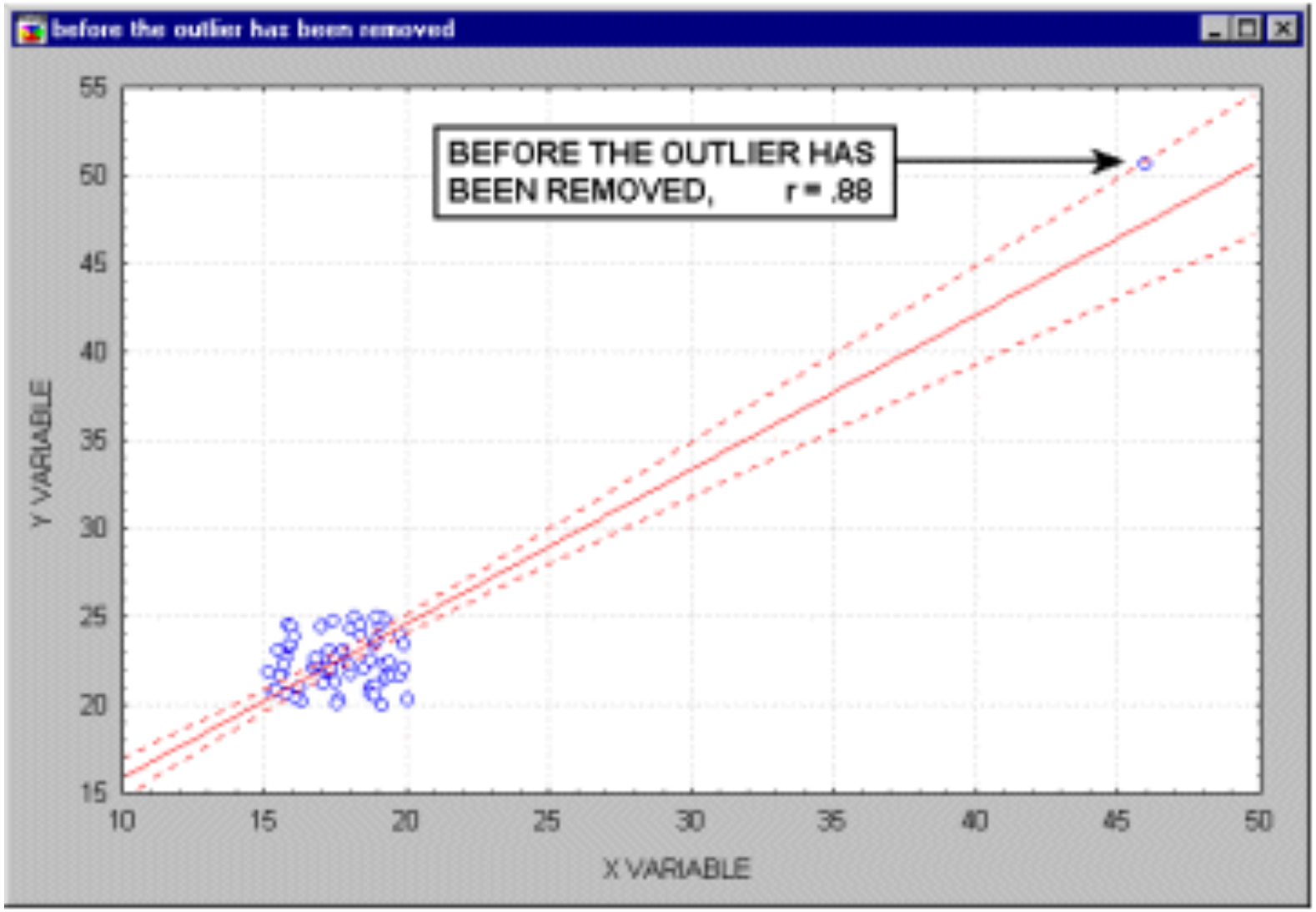






Выбросы




before the outlier has been removed
BEFORE THE OUTLIER HAS
BEEN REMOVED, r = .88
Y VARIABLE
X VARIABLE
55
50
45
40
35
30
25
20
15
10
15
20
25
30
35
40
45
50