

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

✓ ~~Поча недели и загруженность дорог~~
Irkutsk Russia Average Monthly Temperatures

✓ Загруженность дорог Москвы в течение недели

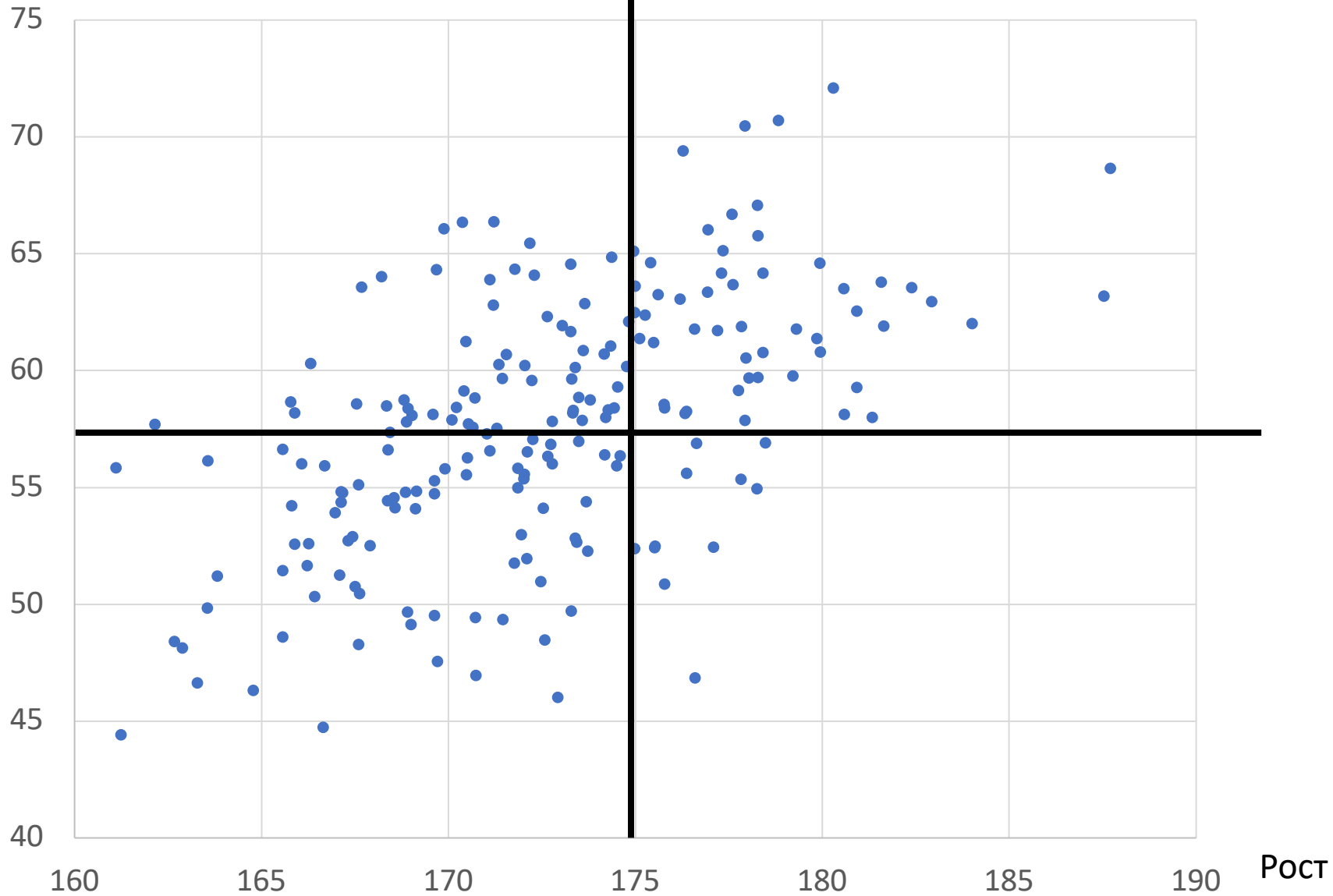


Воскресенье

в год и возраст

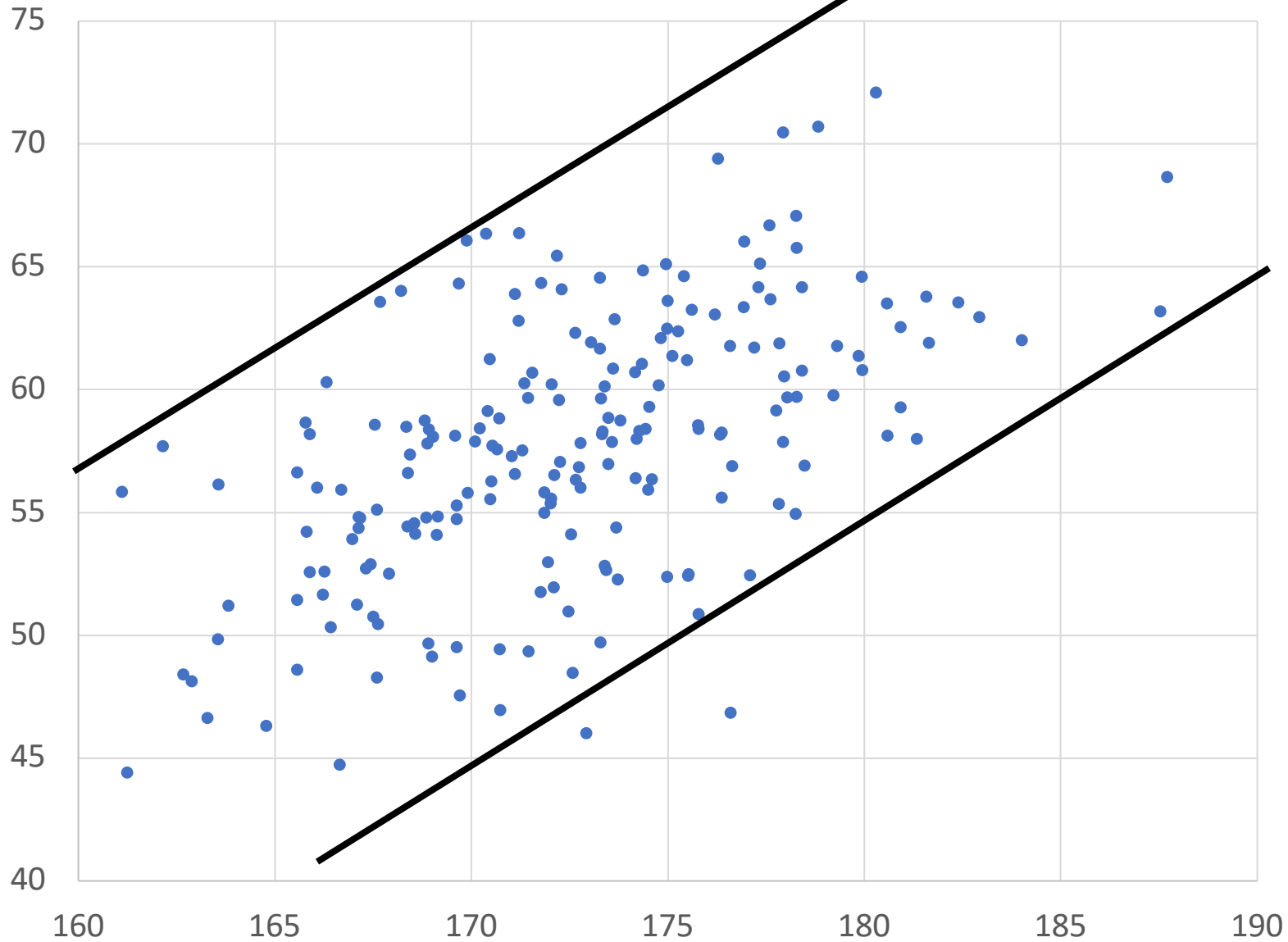
Рост-вес

Вес



Рост-вес

Вес

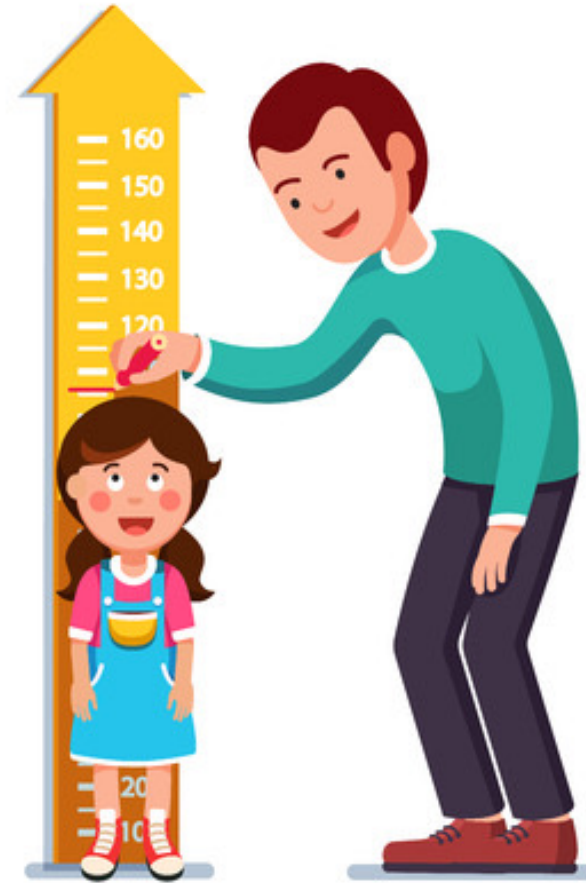
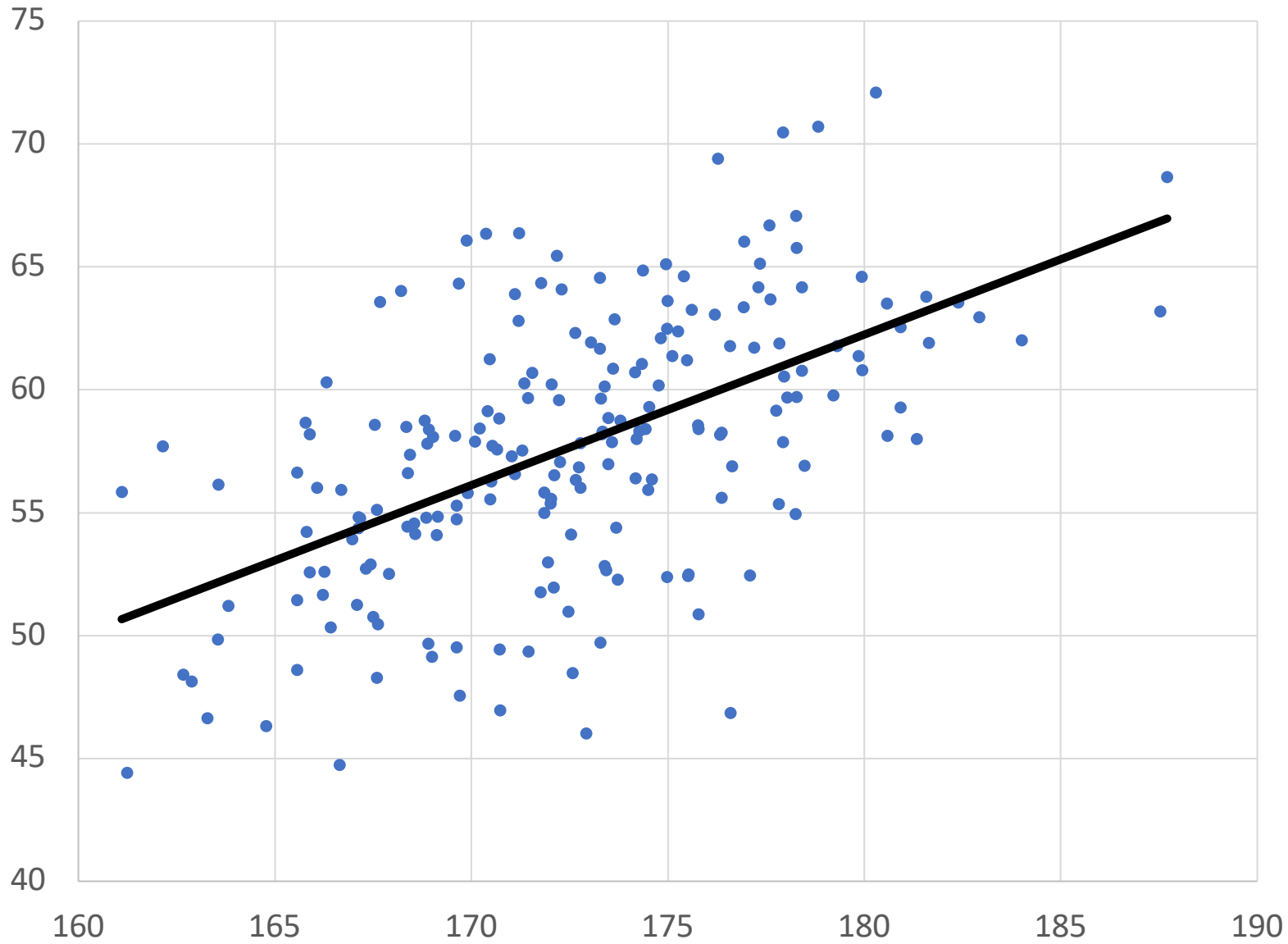


Рост



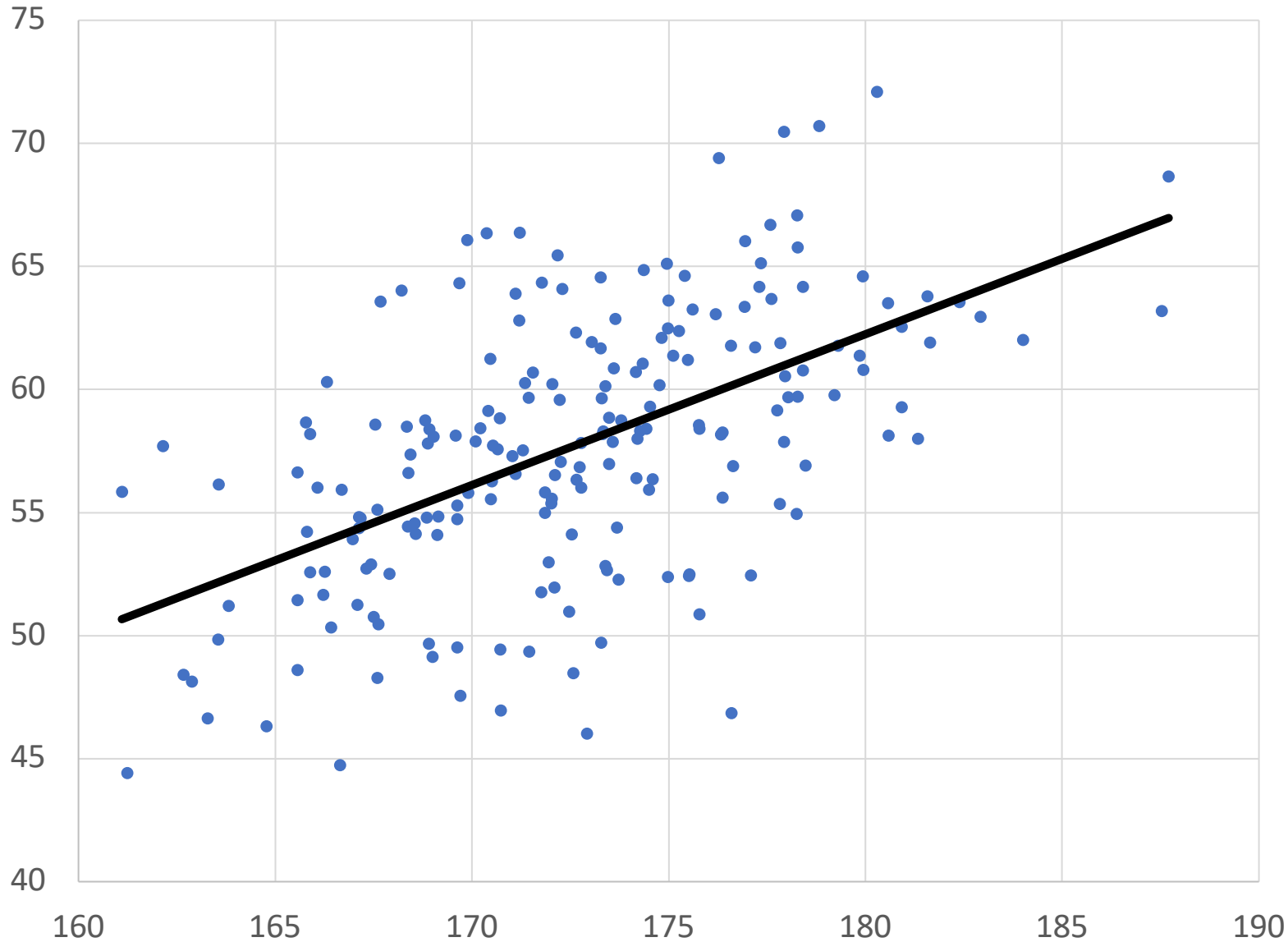
Рост-вес

Вес



Рост-вес

Вес



Как численно
оценить линию
тренда?

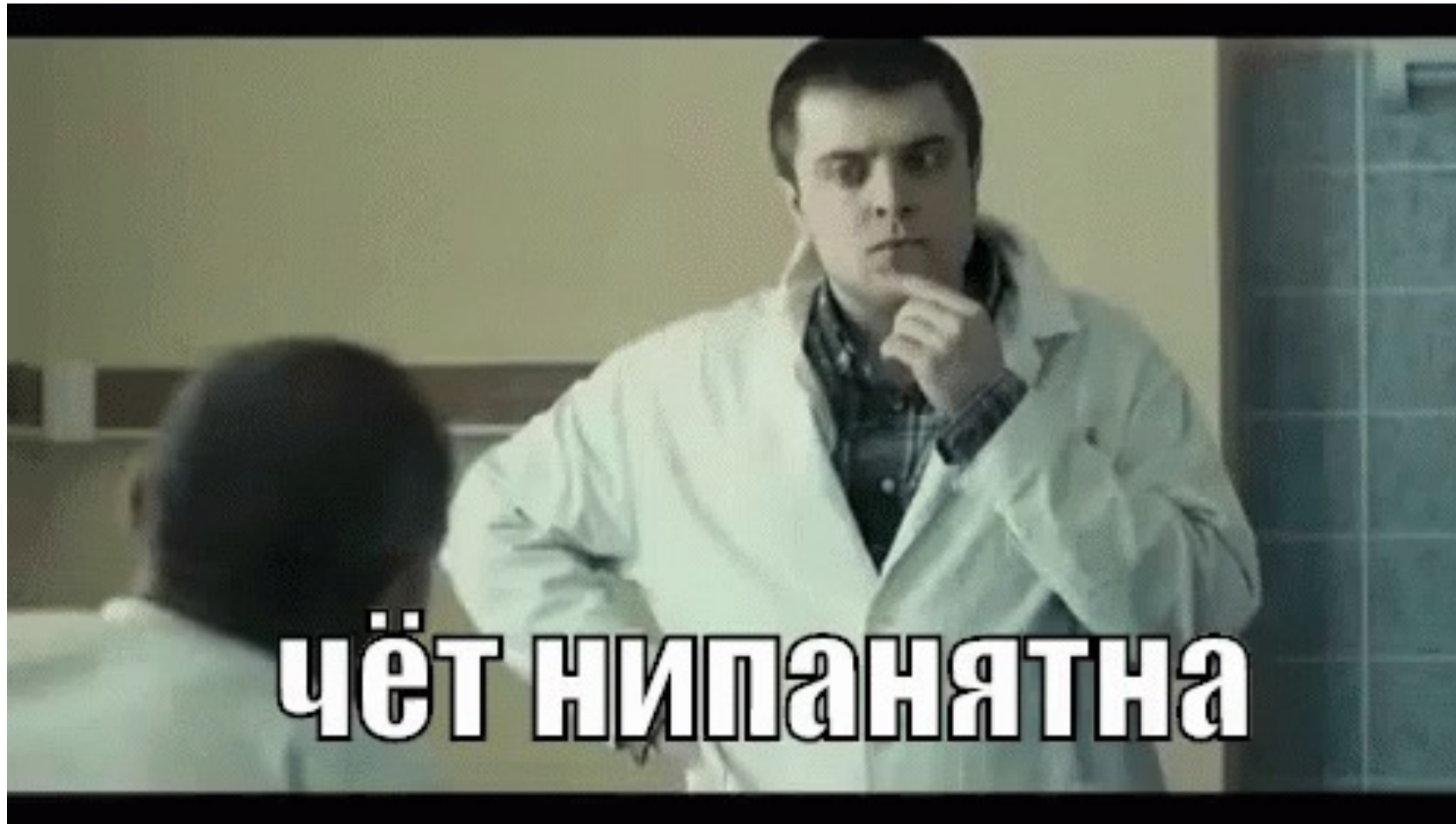
Коэффициент
корреляции

Corr = 0,56

Рост

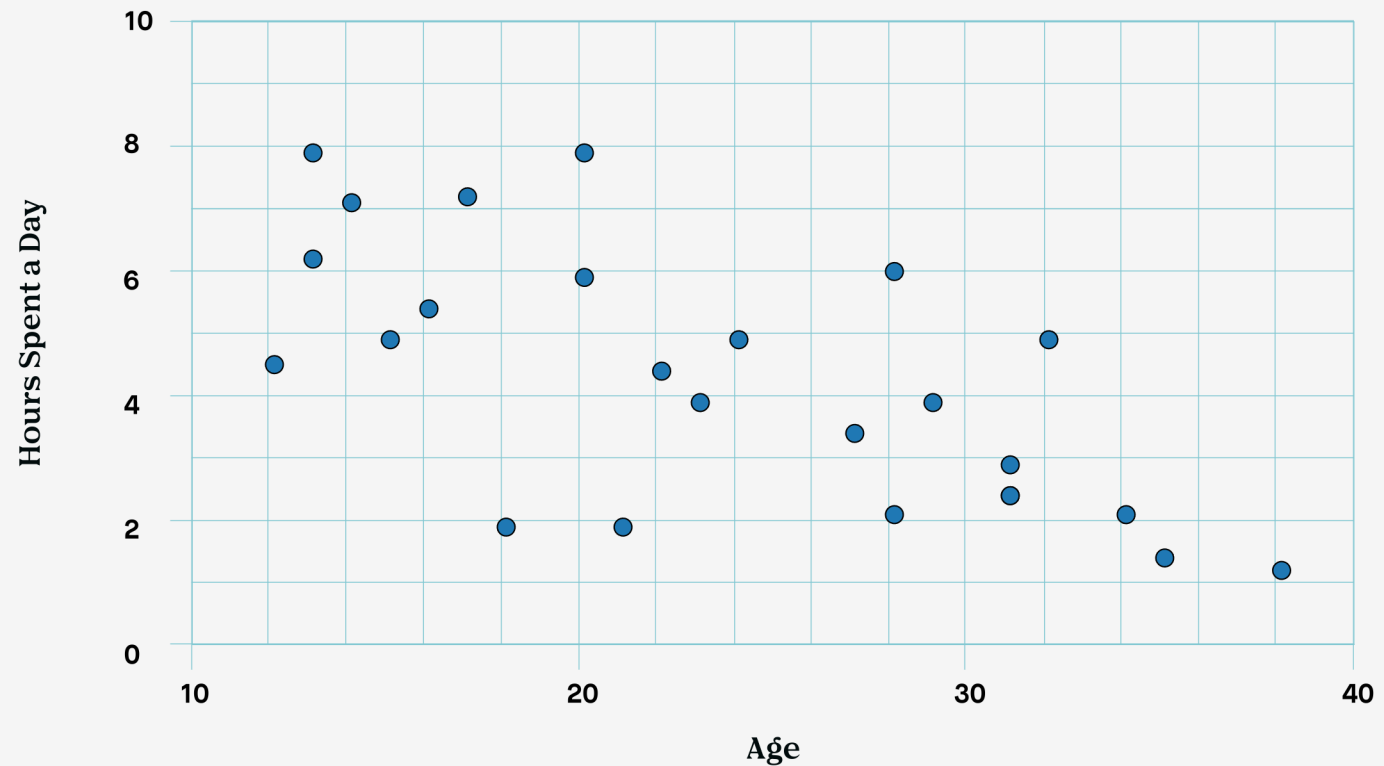
И что делать с этим числом?
Как это интерпретировать?

Corr = 0,56

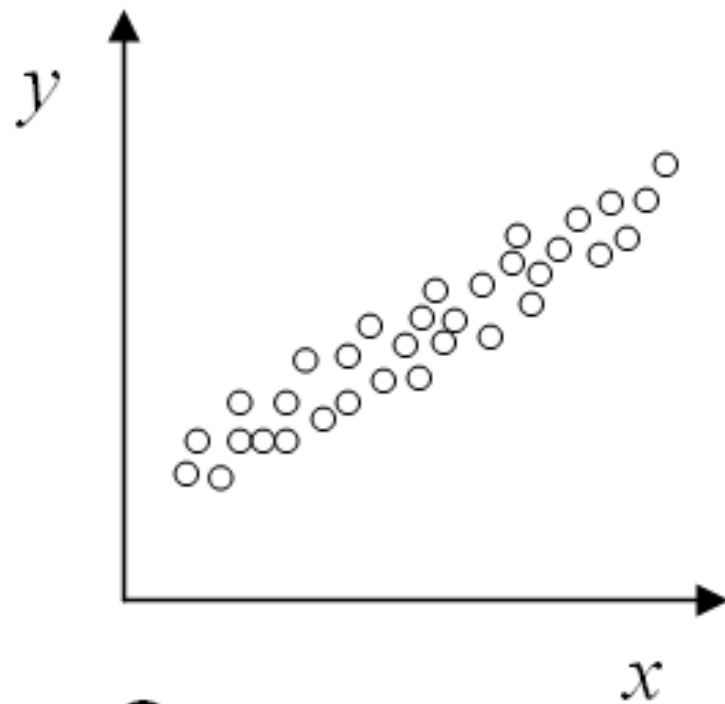


Больше примеров! Отрицательная корреляция

Scatter Plot of Age vs Hours a Day Spent on New App

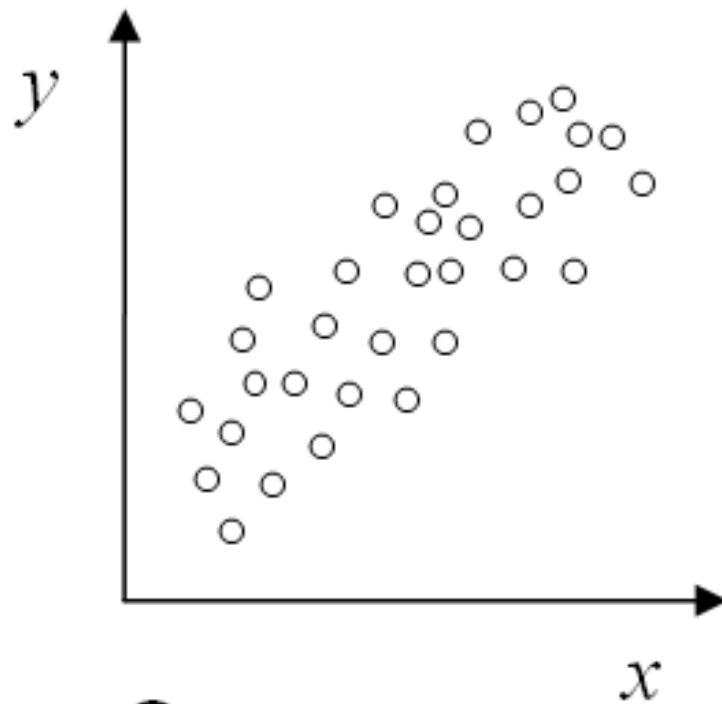


Больше примеров! Слабая корреляция



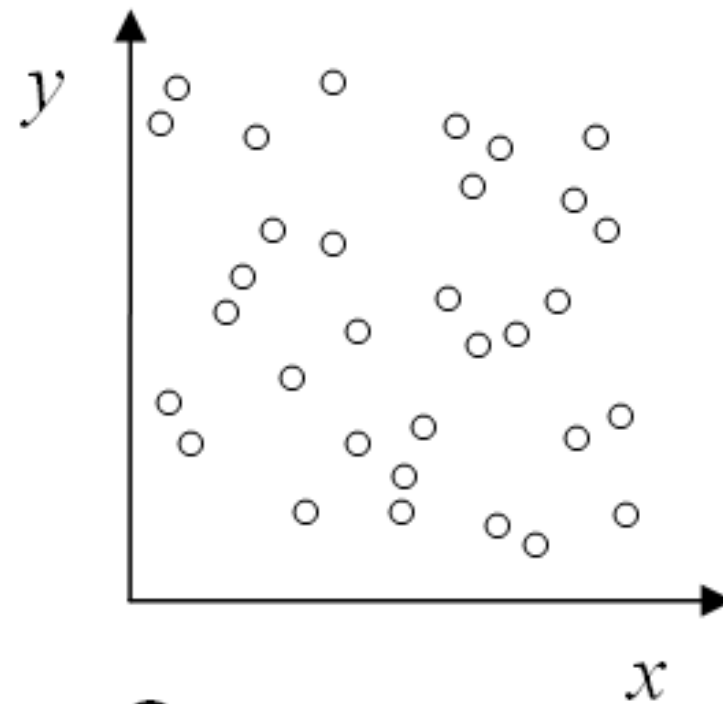
1

Corr \approx 1



2

Corr \approx 0.5

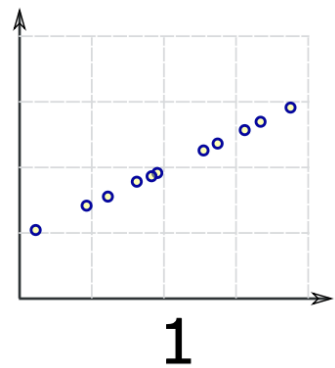


3

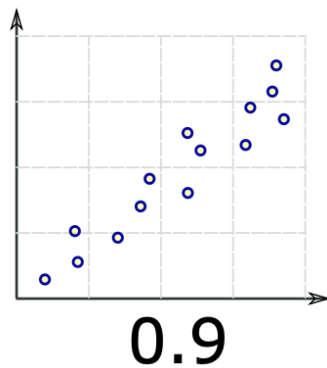
Corr \approx 0

Больше примеров! Summary

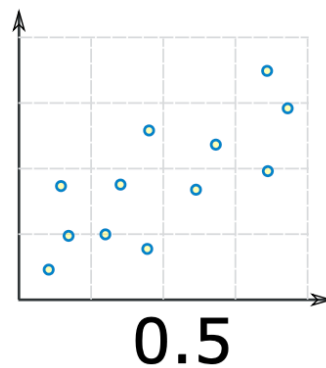
*Perfect
Positive
Correlation*



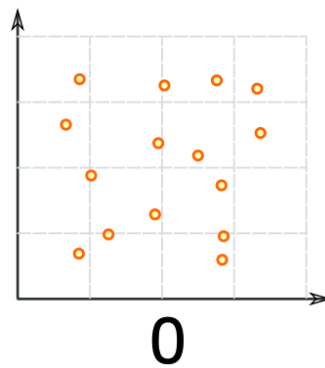
*High
Positive
Correlation*



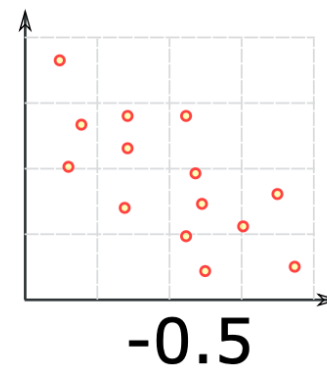
*Low
Positive
Correlation*



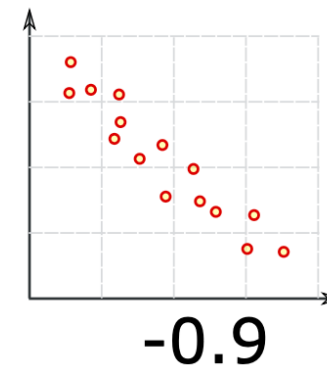
*No
Correlation*



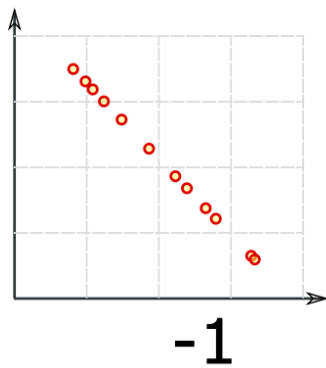
*Low
Negative
Correlation*



*High
Negative
Correlation*



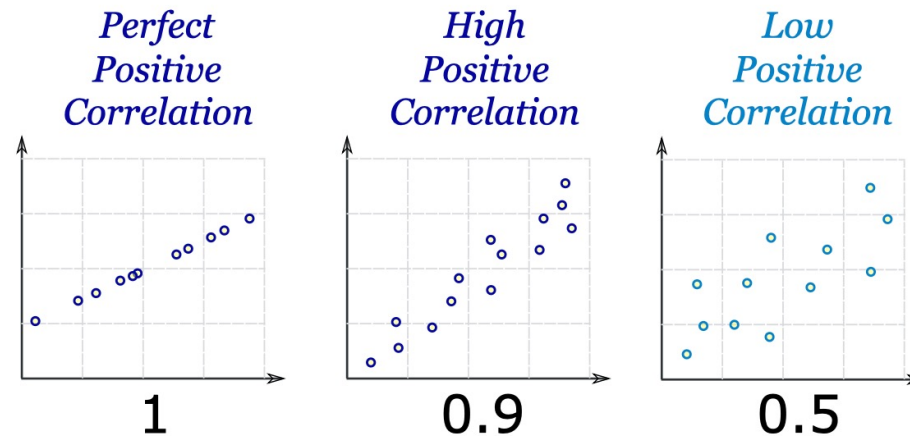
*Perfect
Negative
Correlation*



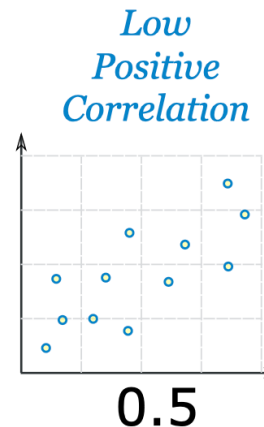
И что делать с этим числом?
Как это интерпретировать?

Corr = 0,56

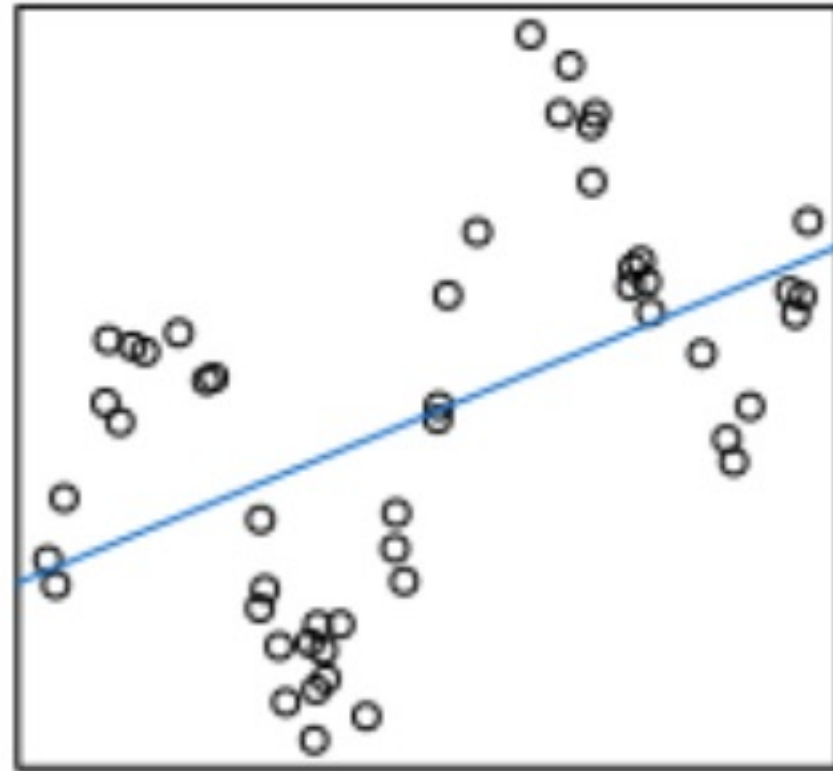
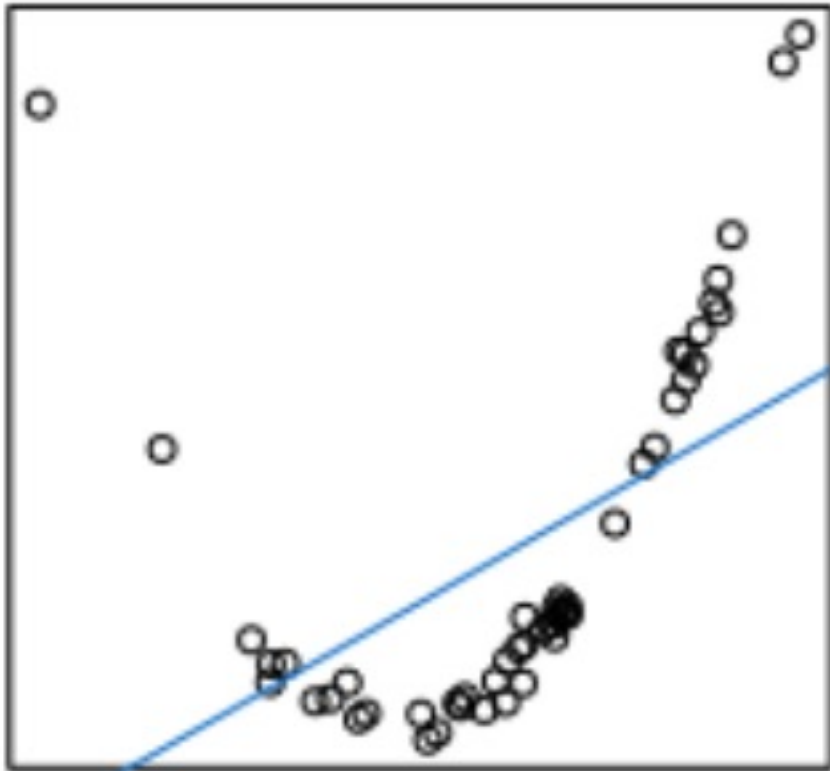
- $\text{Corr} = 0,56 > 0 \Rightarrow$ положительная взаимосвязь



- $\text{Corr} = 0,56 < 1 \Rightarrow$ вес зависит не только от роста

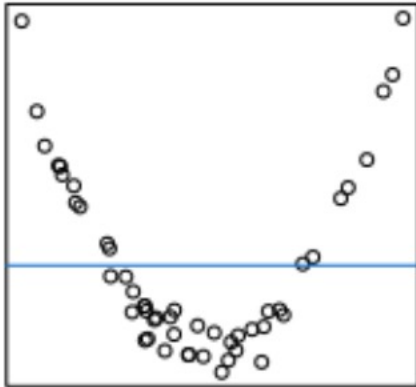


А какая корреляция здесь?

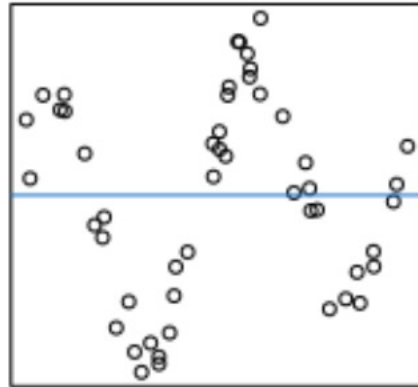


Corr = 0. Неужели нет взаимосвязи?

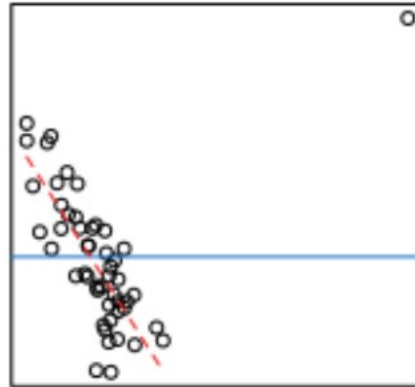
(9) Quadratic trend



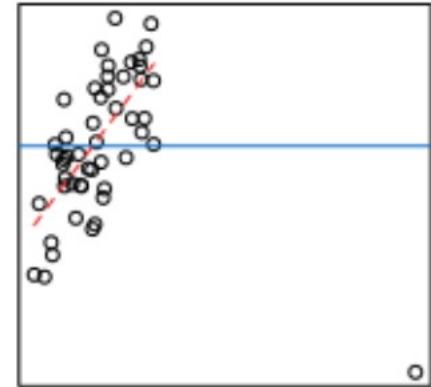
(10) Sinusoid relationship



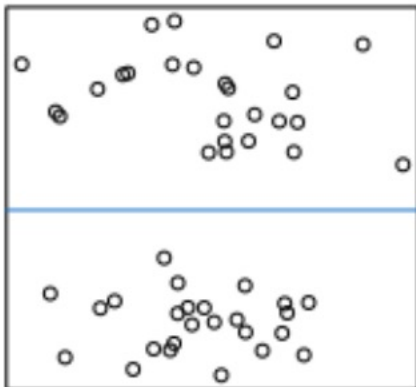
(11) A single positive outlier



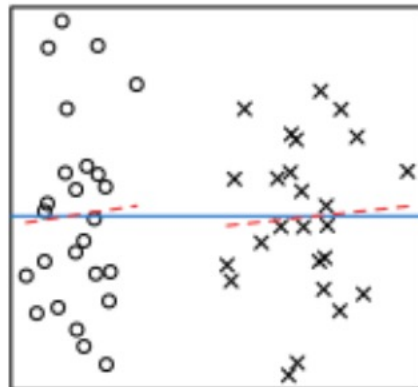
(12) A single negative outlier



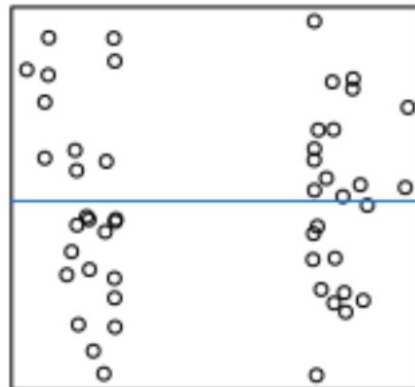
(13) Bimodal residuals



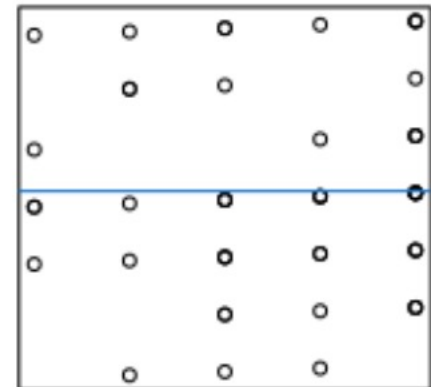
(14) Two groups



(15) Sampling at the extremes



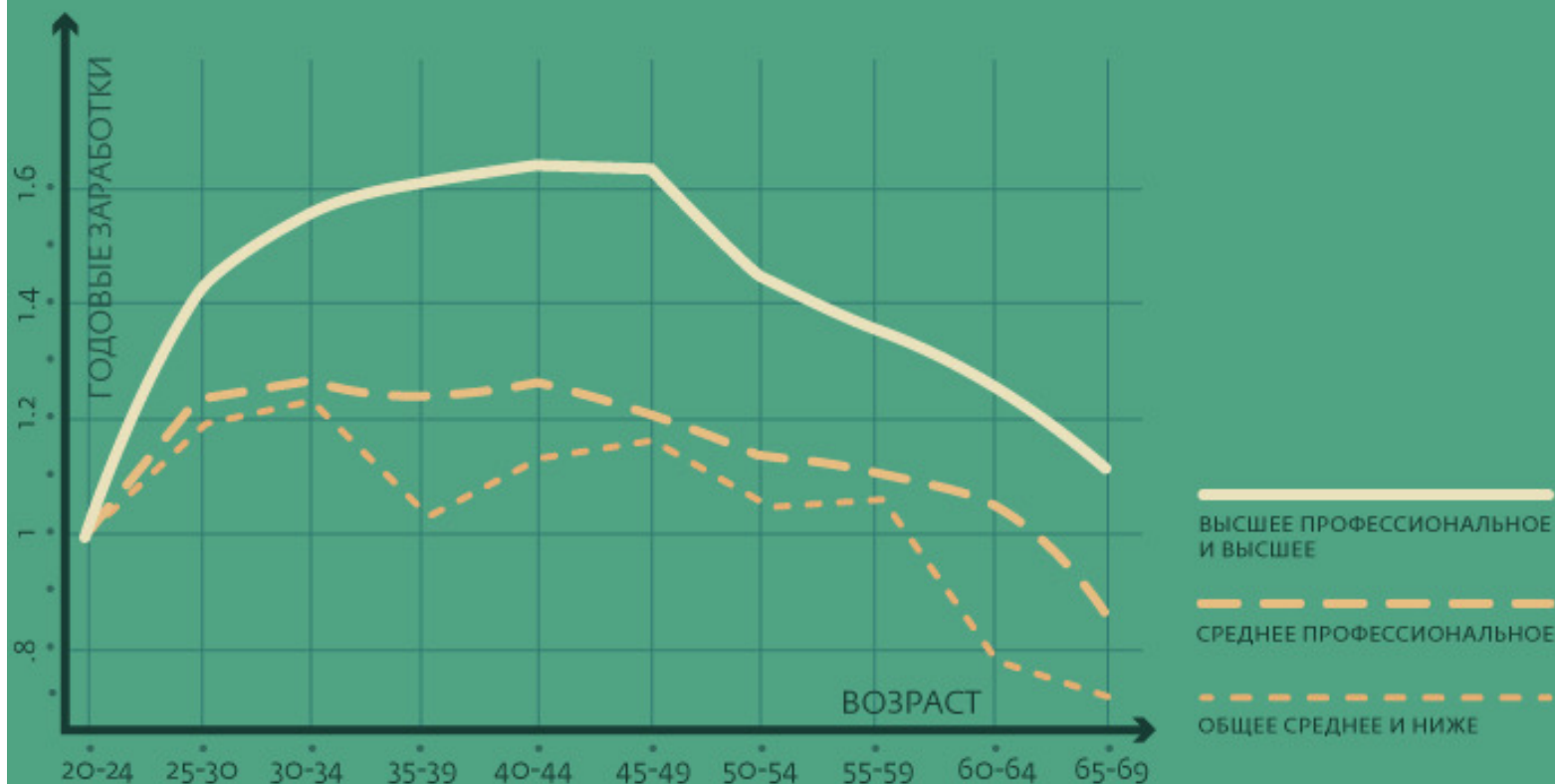
(16) Coarse data



Больше примеров! Нелинейная взаимосвязь

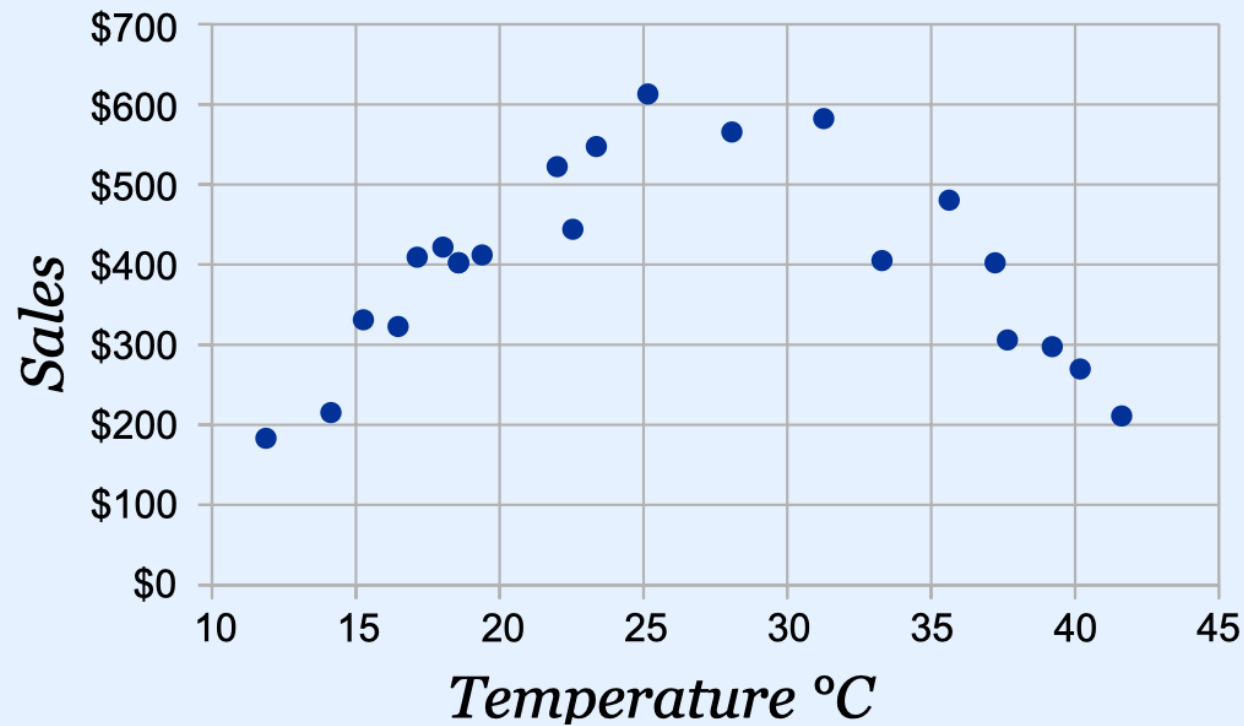
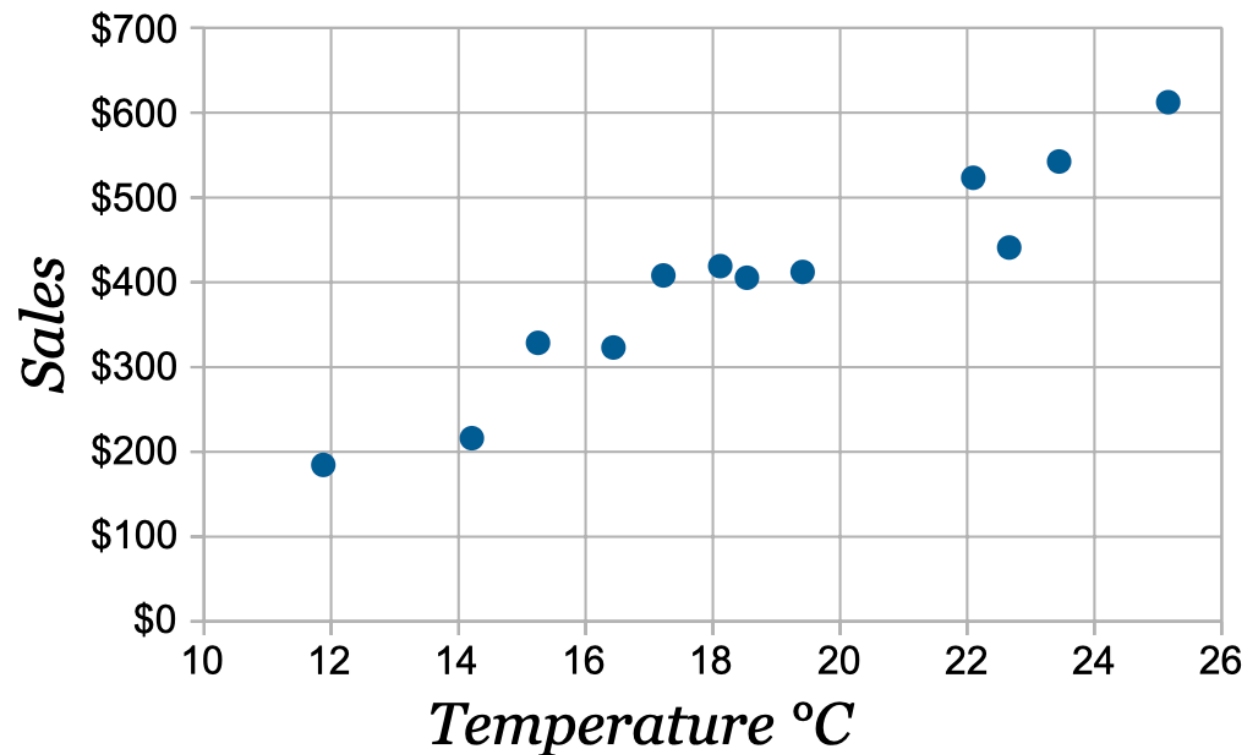
КАК МЕНЯЕТСЯ ЗАРПЛАТА

В ЗАВИСИМОСТИ ОТ ВОЗРАСТА И УРОВНЯ ОБРАЗОВАНИЯ

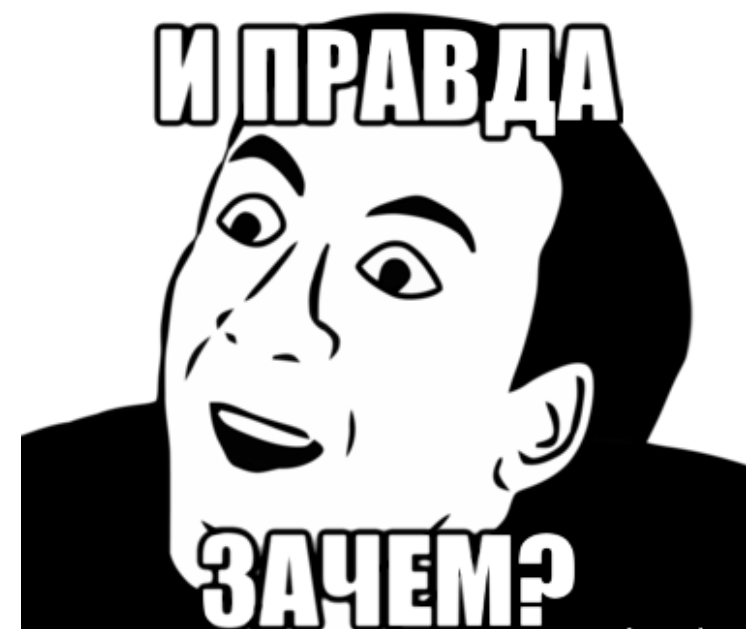
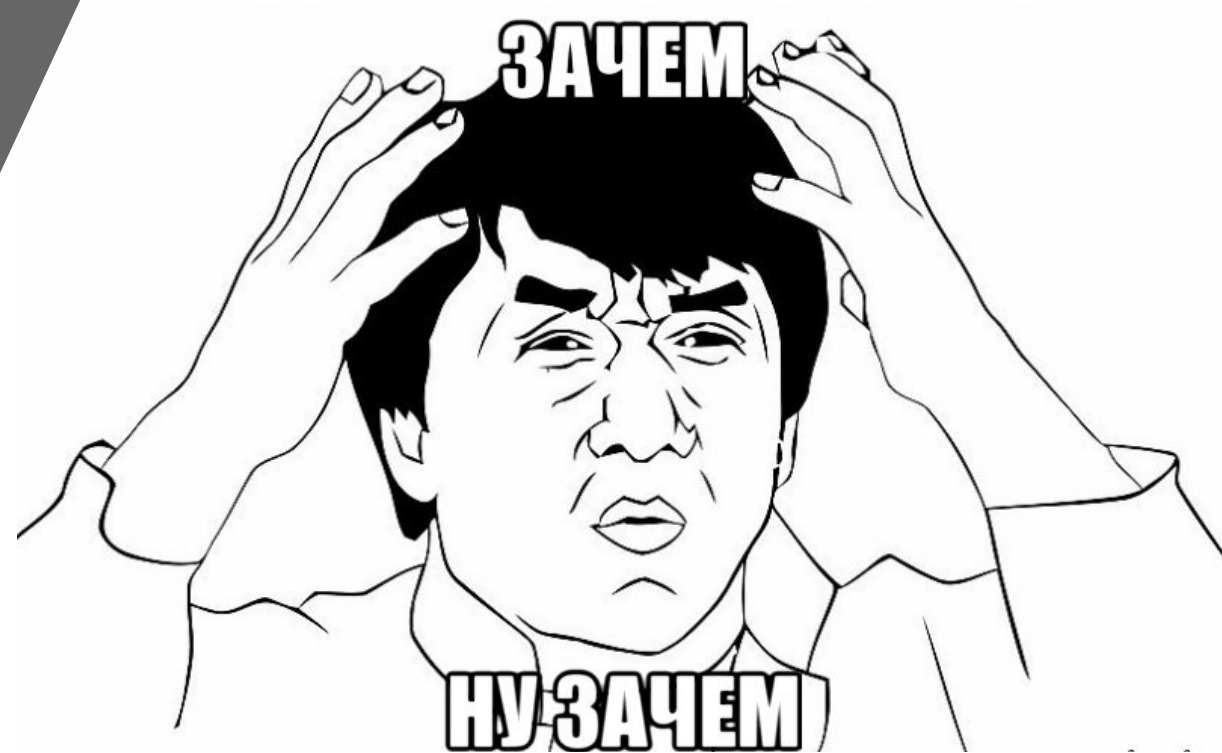
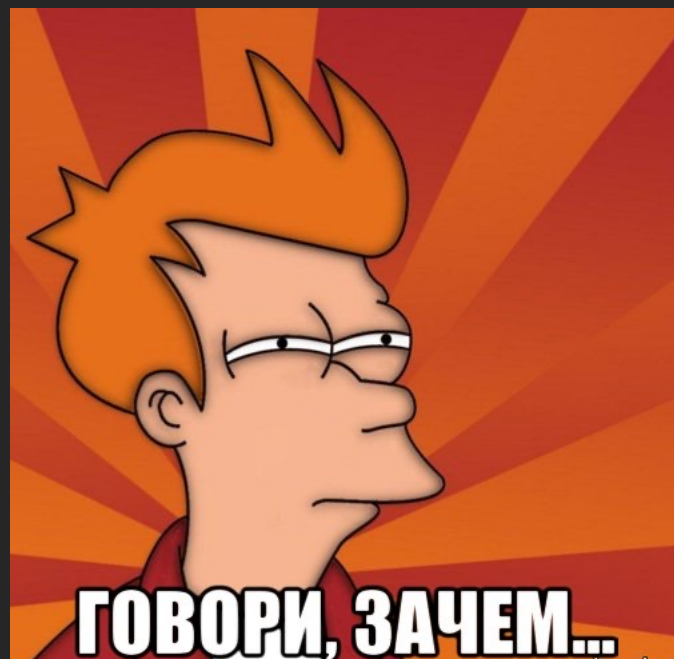


Больше примеров! Нелинейная взаимосвязь

Продажи мороженого

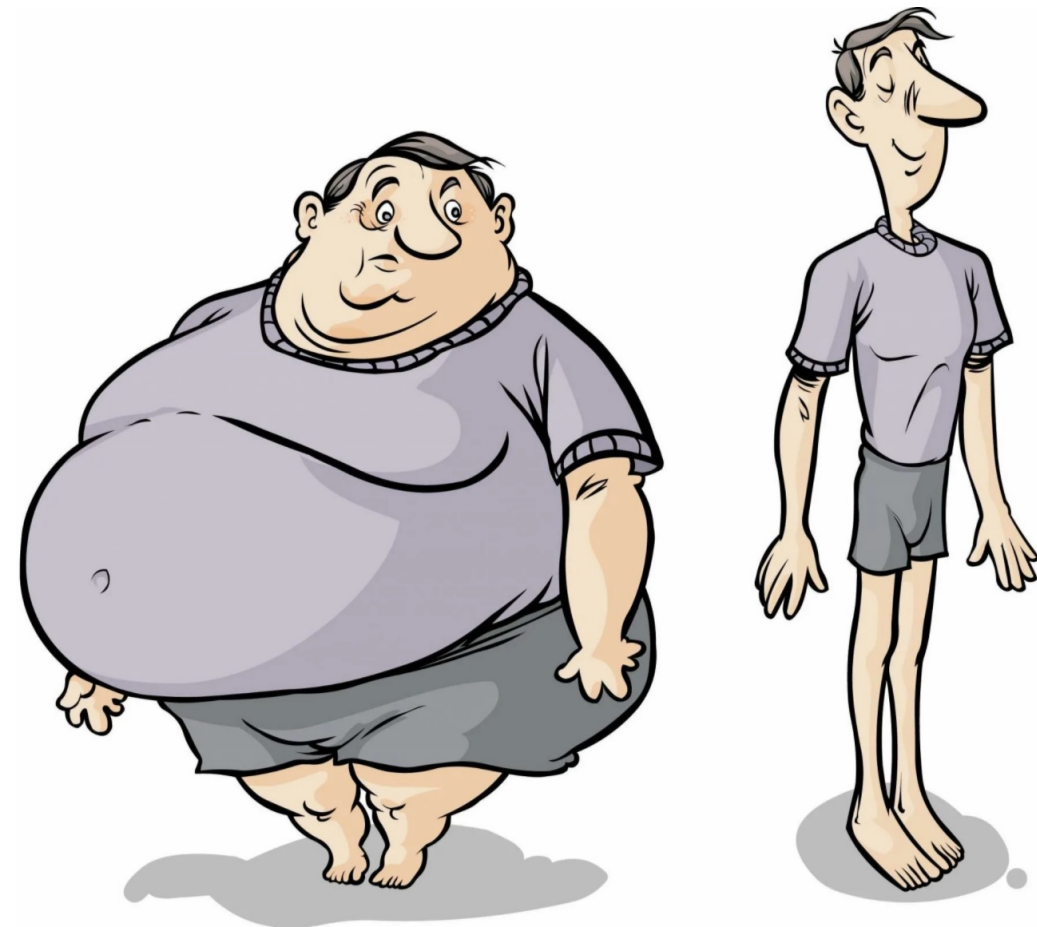
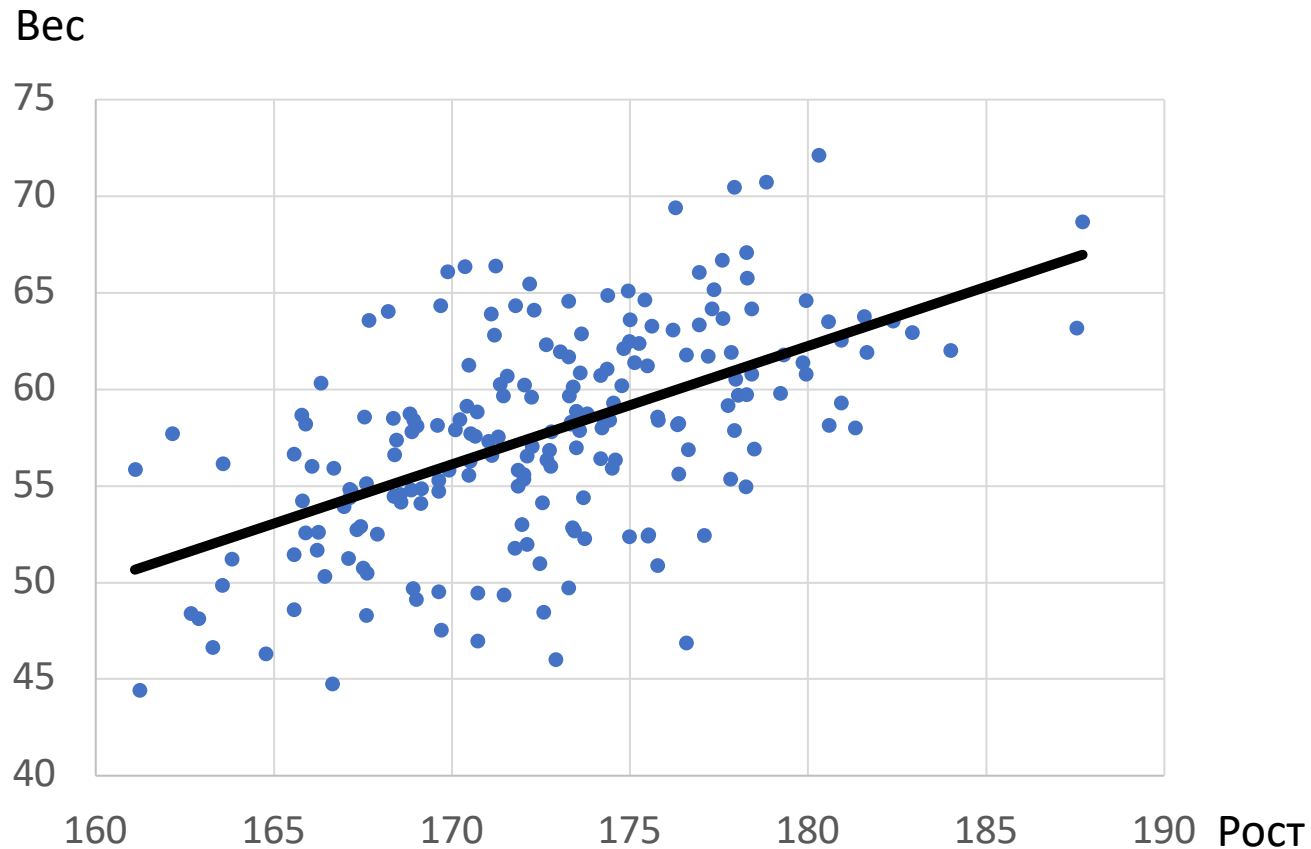


Зачем нужен
анализ
данных?



Рост-вес

- Увеличение веса \Rightarrow увеличение роста?
- Увеличение роста \Rightarrow увеличение веса?



Two police officers in dark blue uniforms and sunglasses stand in a nightclub. The background is filled with blurred lights and people, creating a bokeh effect. The lighting is predominantly blue and purple.

Полиция в США

Количество полицейских в городе и количество преступлений скоррелировано

Избегание врачей в средневековье

После появления врача в поселении многие люди там умирали



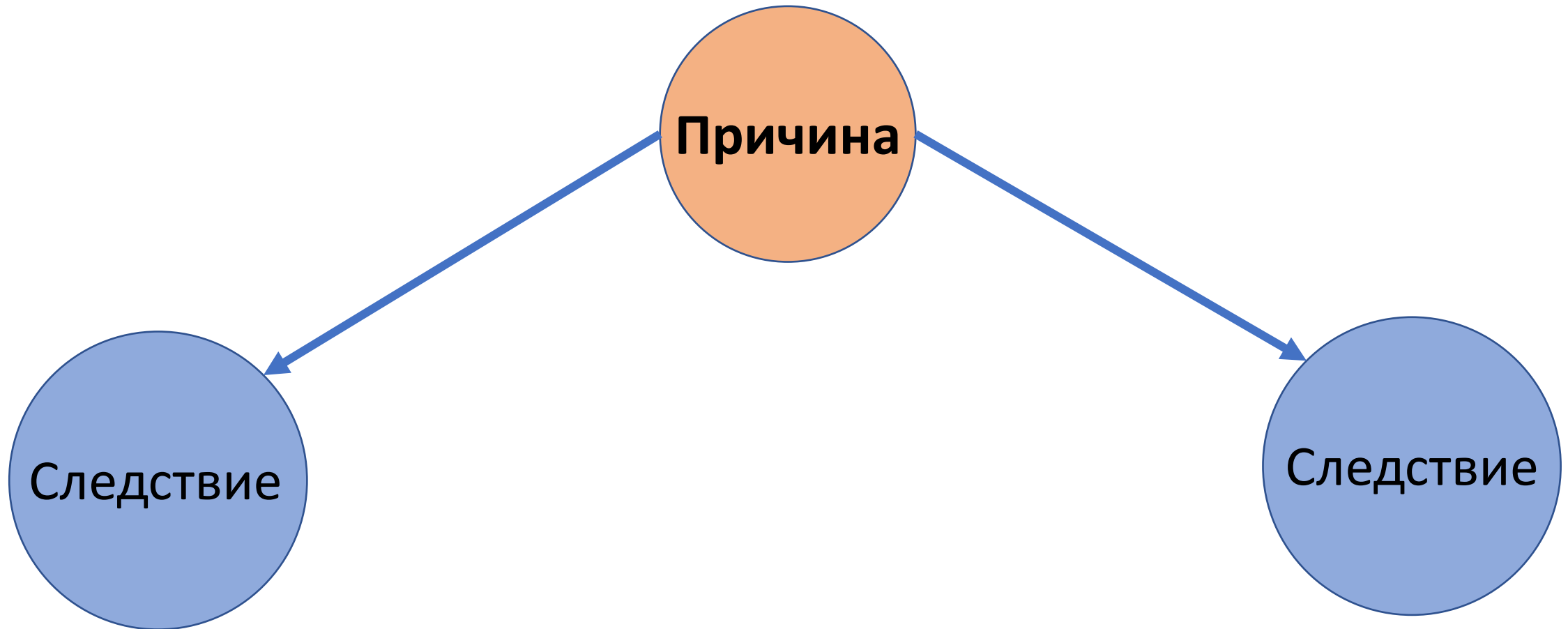
Примеры прямой взаимосвязи



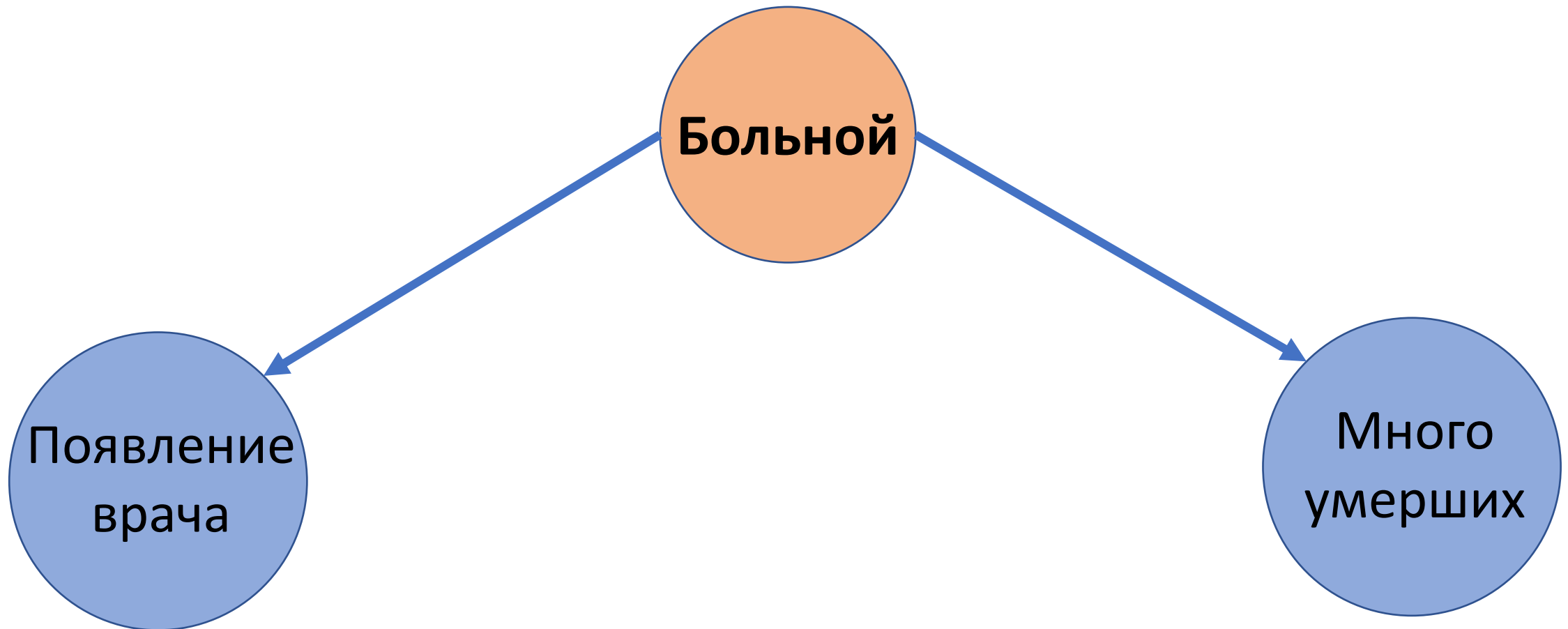
Примеры прямой взаимосвязи



Одна причина и несколько следствий



Одна причина и несколько следствий



Примеры ложной взаимосвязи



Просто
факт



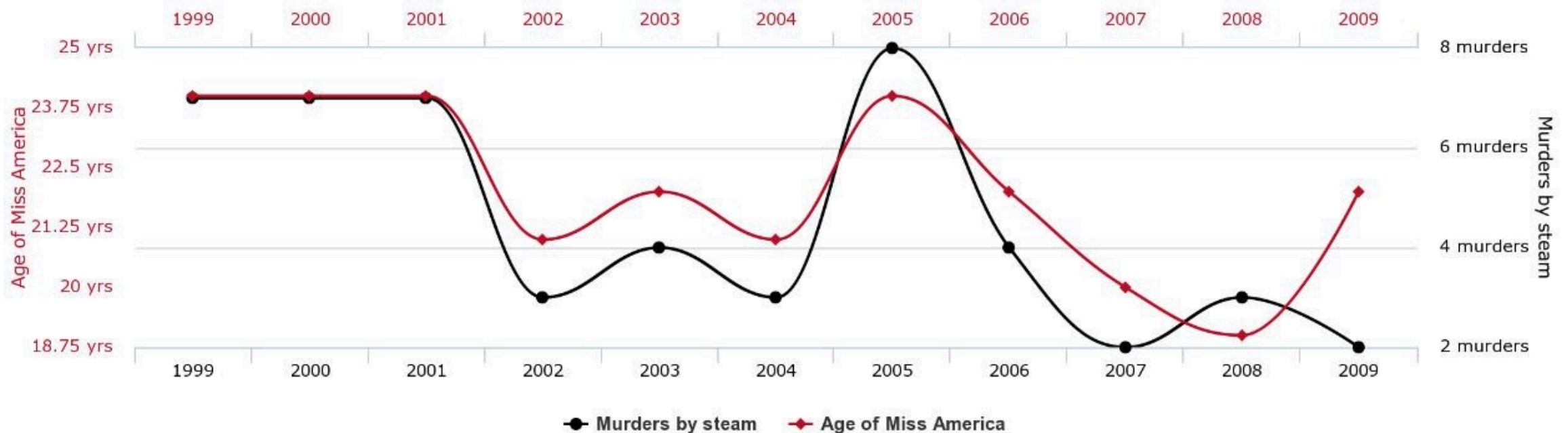
Просто
факт

Примеры ложной взаимосвязи

Age of Miss America

correlates with

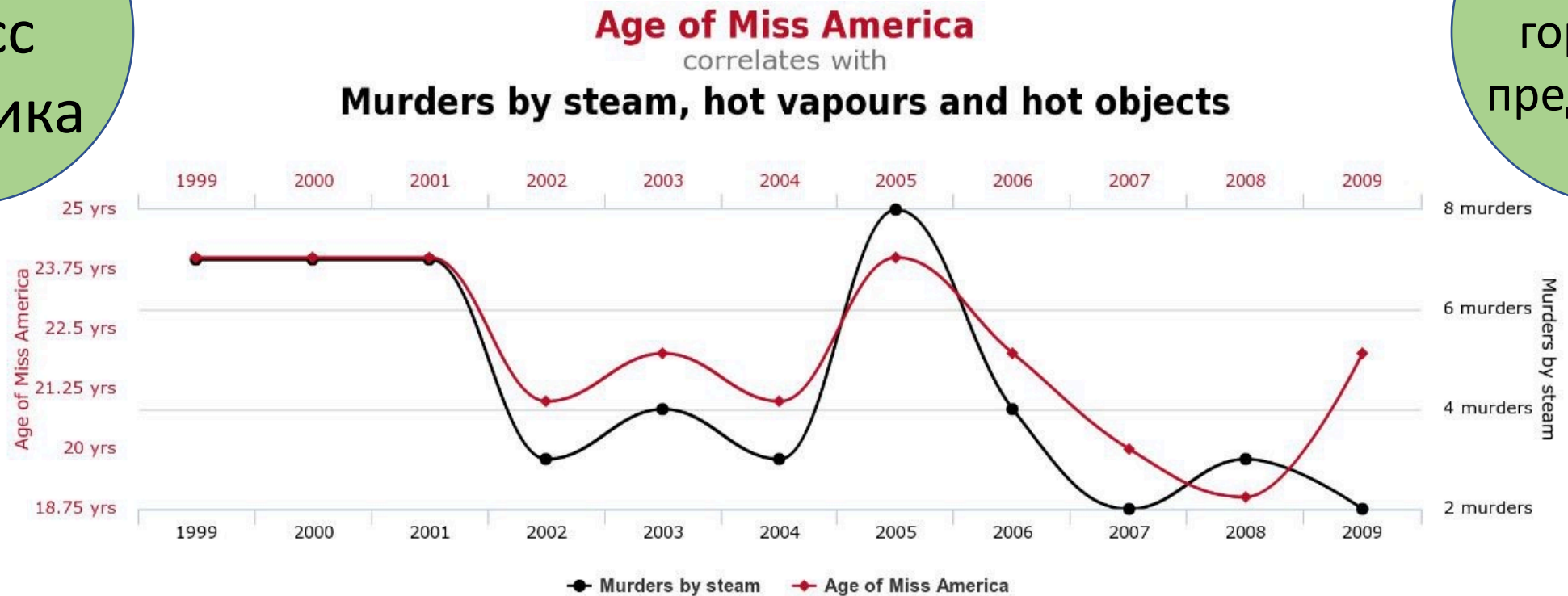
Murders by steam, hot vapours and hot objects



Примеры ложной взаимосвязи

Возраст
Мисс
Америка

Убийства
горячими
предметами

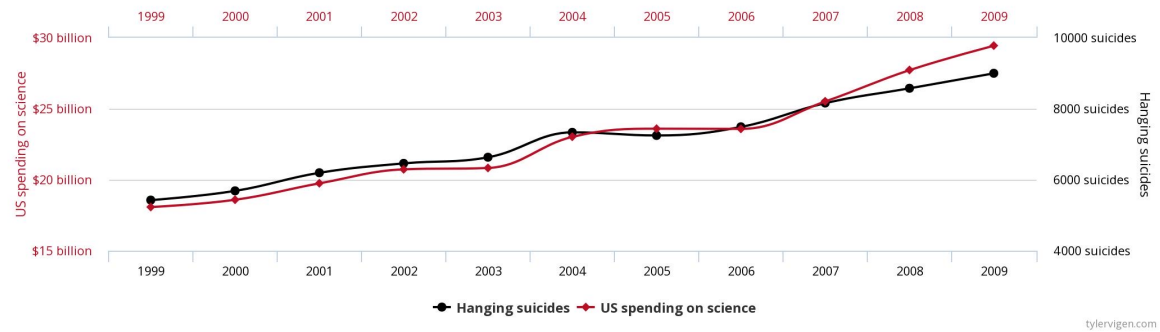


Примеры ложной взаимосвязи

US spending on science, space, and technology

correlates with

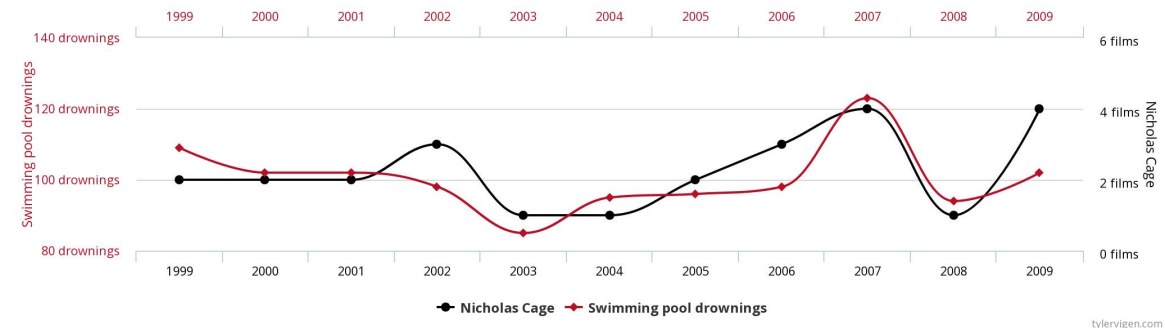
Suicides by hanging, strangulation and suffocation



Number of people who drowned by falling into a pool

correlates with

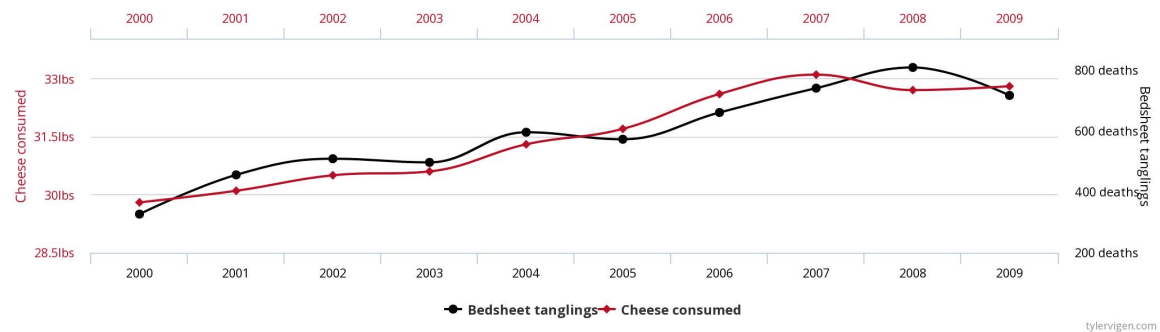
Films Nicolas Cage appeared in



Per capita cheese consumption

correlates with

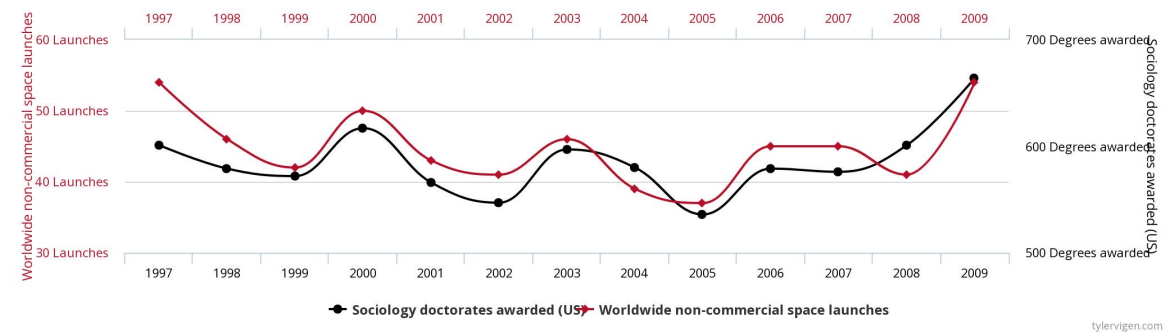
Number of people who died by becoming tangled in their bedsheets



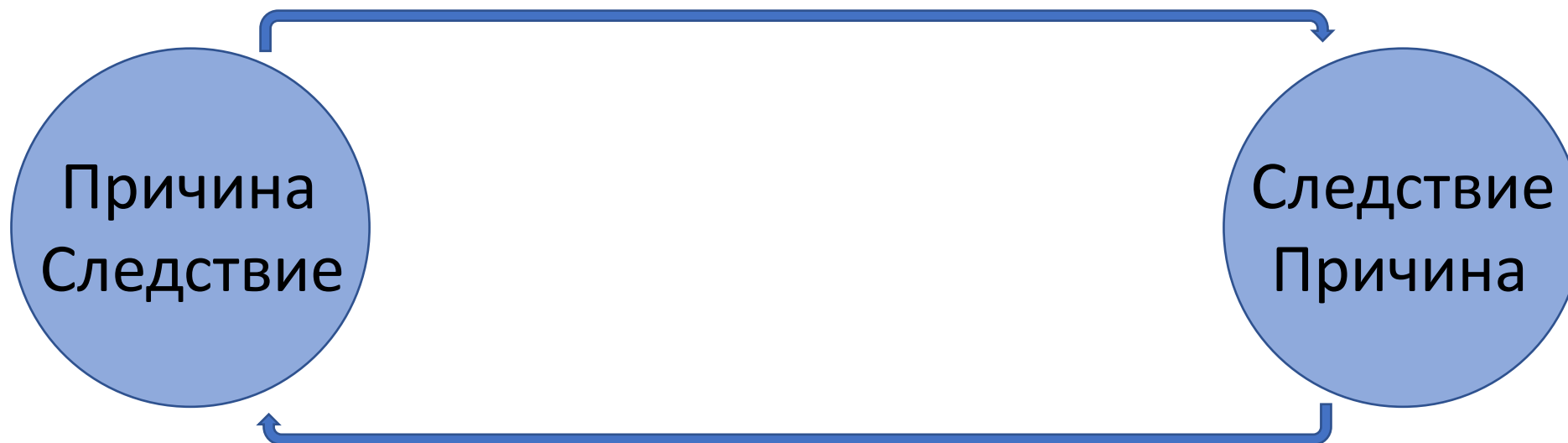
Worldwide non-commercial space launches

correlates with

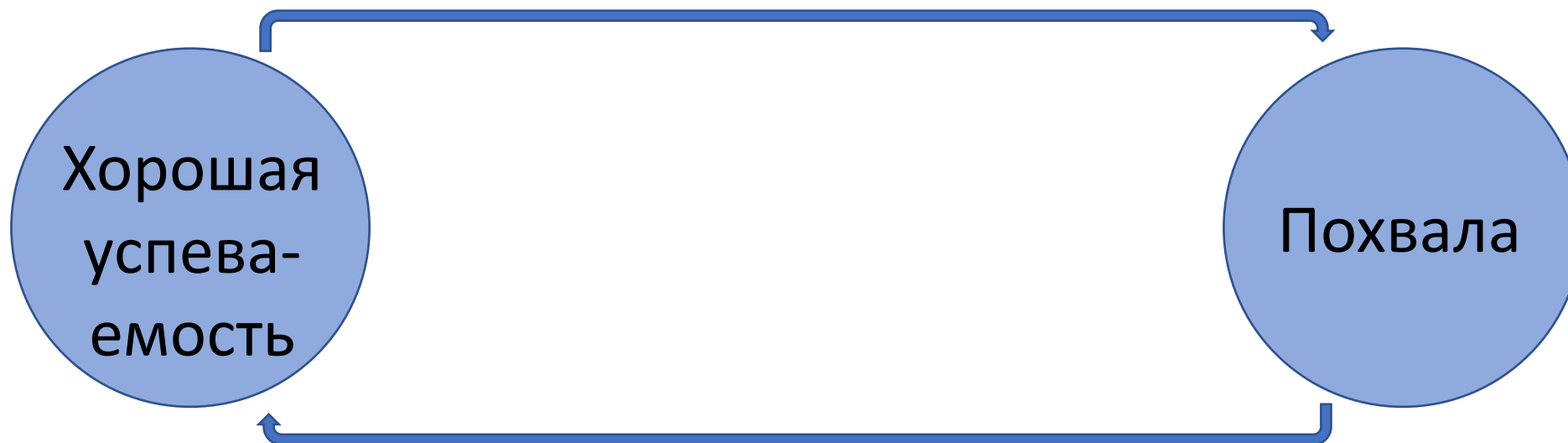
Sociology doctorates awarded (US)



Зацикленная взаимосвязь



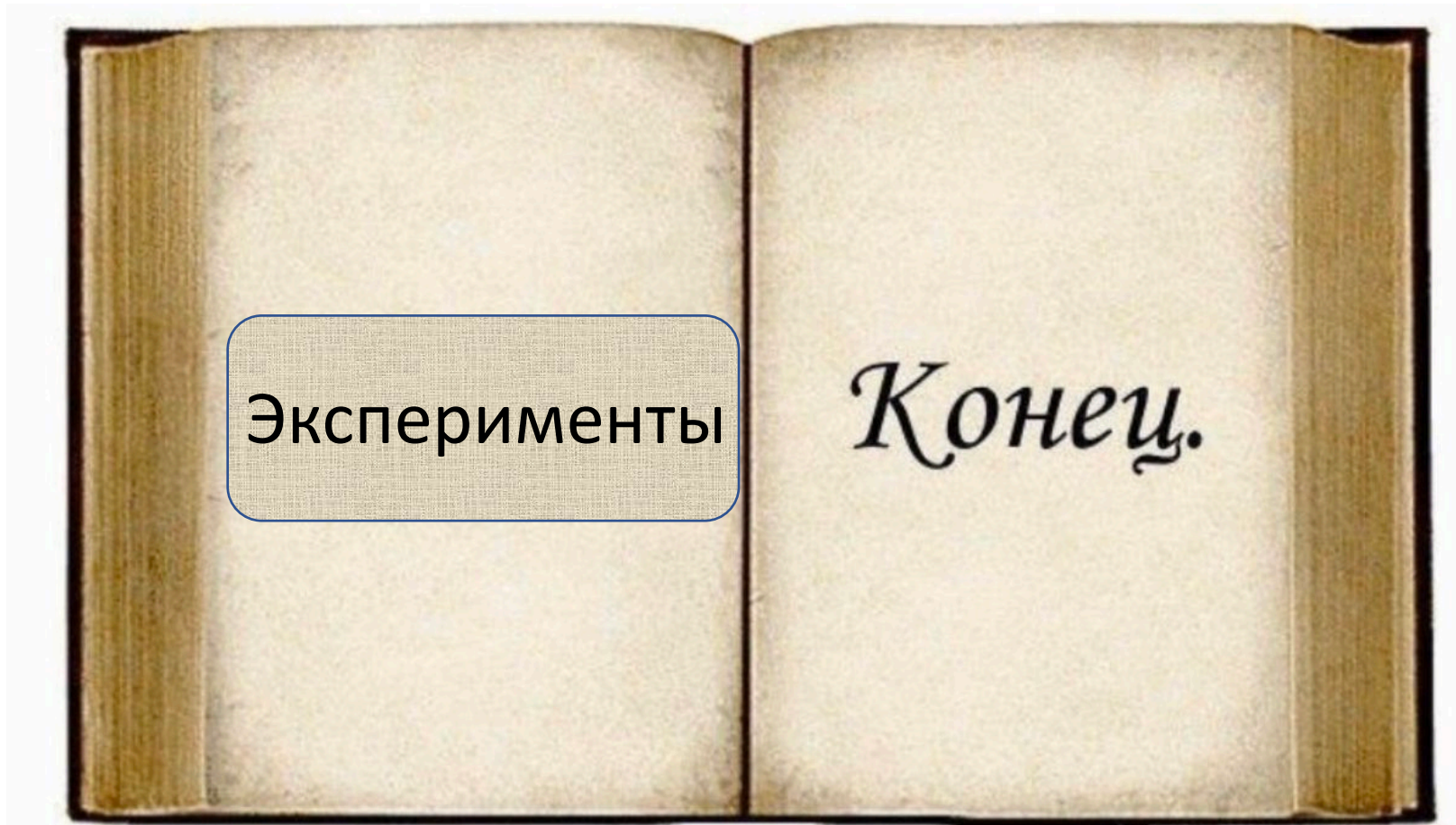
Зацикленная взаимосвязь



Как понять, какой именно вид взаимосвязи



Как понять, какой именно вид взаимосвязи



Эксперименты

Конец.

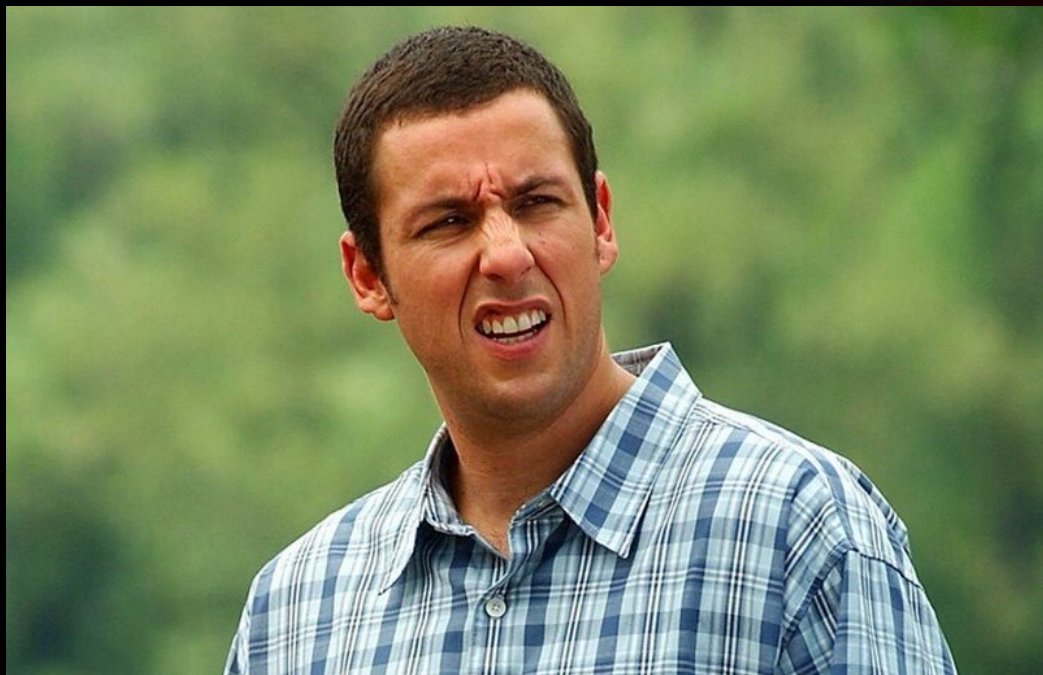
Работники на складе

- 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?



ERROR



В чем ошибка?

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

ERROR



А как можно
было?

Смотреть на увеличение
скорости при максимальной
производительности

И где это применять?

- В жизни!
- На олимпиадах!
 - [DANO](#)
 - [Яндекс.Учебник](#)



Яндекс § Учебник



Анализ данных

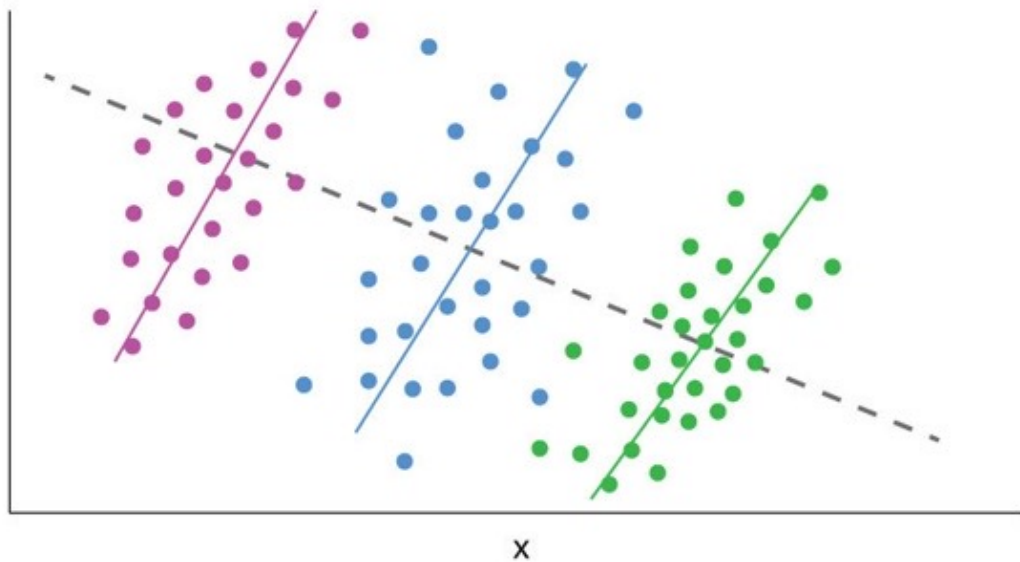
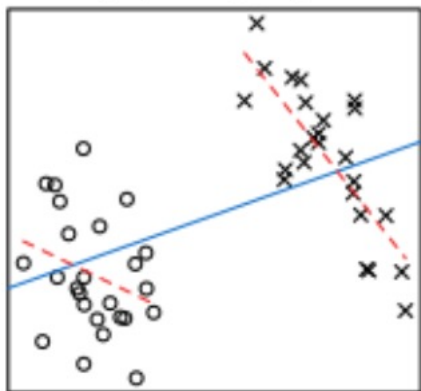
Рита Голуб, Яндекс



ЛЭШ ILE 2022



Парадокс Симпсона



Выбросы

