

Основы анализа и визуализации данных для медиа

Лекция 7
23 сентября 2019

Рогович Татьяна
rogovich@gmail.com

Корреляция и тренды.

Структура лекции

- Так как мы теперь знаем (я надеюсь), что такие меры центральной тенденции и разброса, то сегодня мы поговорим о таких штуках, как корреляция и регрессионная линия, и как их интерпретировать и использовать в своих визуализациях.

Контрольная работа

2 октября
10.30

Покровский б–
р, д.11 R305

Как пишем?

Письменная контрольная длительностью одну пару.

Чем можно пользоваться?

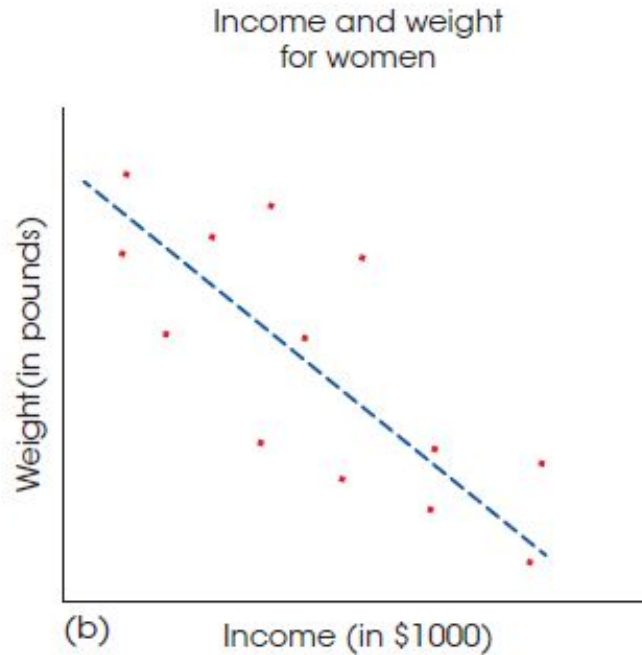
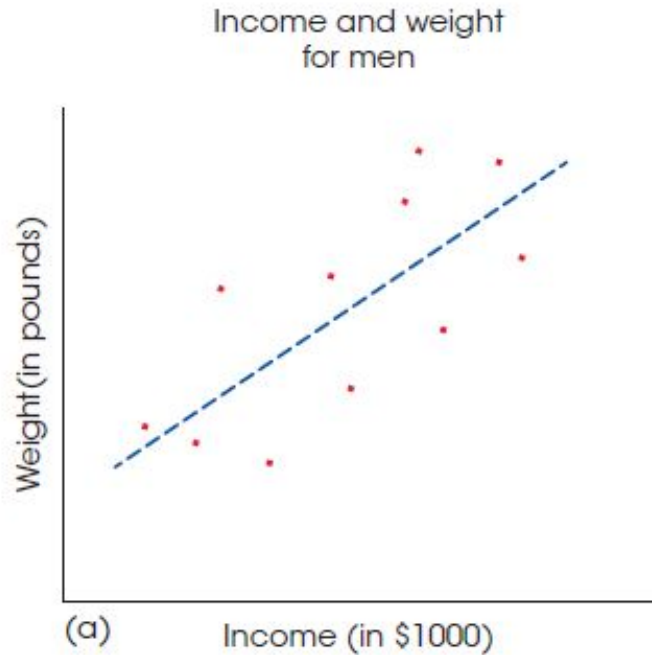
- Калькулятором на телефоне (но лучше принести обычный калькулятор, если есть возможность)
- Шпаргалкой. А5 с одной стороны, написано вами лично (распечатки и ксерокопии не допускаются).
- Захватите пару цветных ручек или ручку и карандаш (чтобы набрасывать графики).

Как готовиться?

- Материалы лекций 1-7, семинары 1-3.
- Прорешайте [пробный вариант](#) (и попробуйте уложиться в указанное время)
- Возьмите у Миши Филатова консультацию

А теперь про
корреляцию :)

Анализируем тренд



Корреляция

Статистическая взаимосвязь между двумя случайными величинами.

Многие переменные в реальном мире сильно коррелируют друг с другом, но эти связи могут объясняться случаем, влиянием других переменных или другими неизвестными причинами.

Направление корреляции

В положительной корреляции, обе переменные изменяются в одном и том же направлении: с увеличением значения переменной X , переменная Y также увеличивается; когда переменная X уменьшается, то и переменная Y тоже уменьшается.

В отрицательной корреляции, обе переменные изменяются в разных направлениях. С увеличением переменной X , переменная Y уменьшается. Это обратная взаимосвязь.

Сила корреляции

Идеальная корреляция всегда определяется 1.00 и указывает на идеально согласованную связь.

Для корреляции 1.00 (или -1.00), каждое изменение в X , сопровождается идеально предсказуемым изменением в Y . С другой стороны, корреляция 0 указывает на полное отсутствие связи. В корреляции 0, точки скапливаются случайно без какой-либо тенденции. Промежуточные значения между 0 и 1 указывают на некоторую степень связности.

Корреляцию меньше 0.5. считают слабой, от 0.5 до 0.7 - средней.

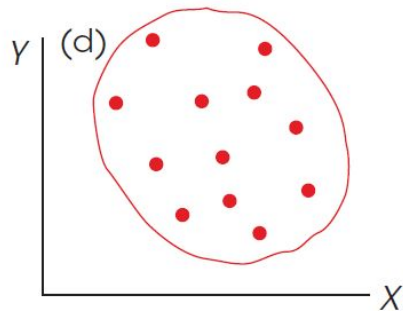
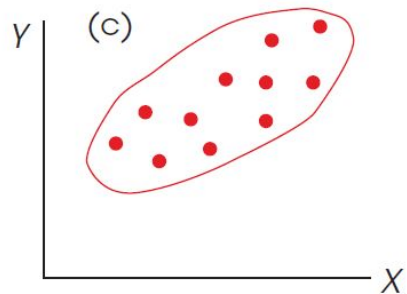
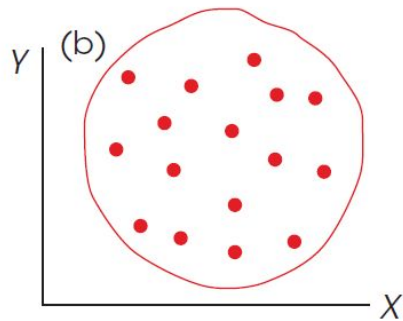
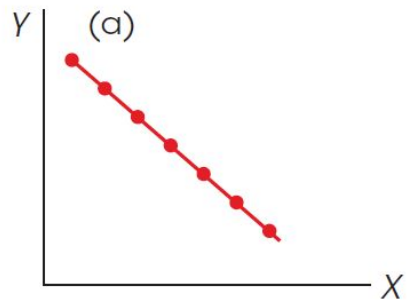
<http://guessthecorrelation.com/>

Форма корреляции

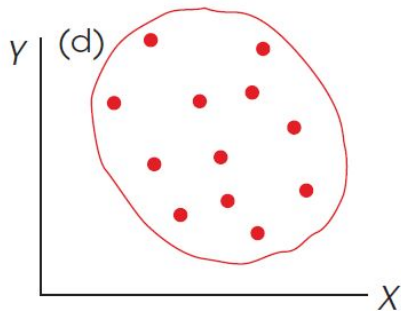
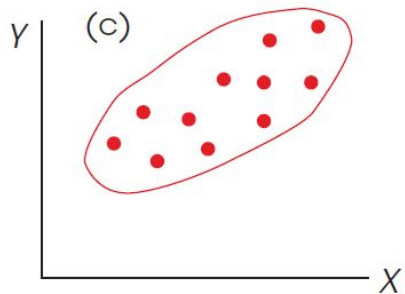
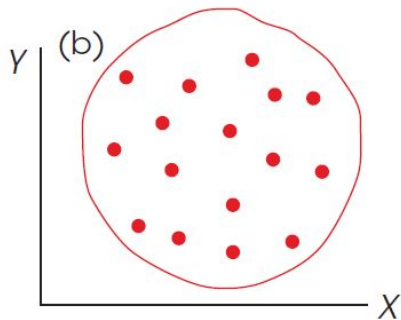
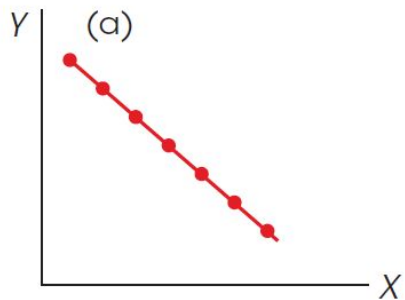
В вышеприведенном примере с зарплатой и весом, связь имела линейный вид, т.е. точки на диаграмме рассеивания, как правило, скапливались вдоль прямой линии. Мы нарисовали линию через центр скопления точек, для того чтобы показать их связь. Корреляцию в основном используют для измерения линейной связи. Однако, существуют и другие виды, для измерения которых используются специальные корреляции.

Но чаще всего говорят именно о линейной корреляции или корреляции Пирсона.

Угадай корреляцию



Угадай корреляцию



а) негативная идеальная корреляция
с коэффициентом - 1.0

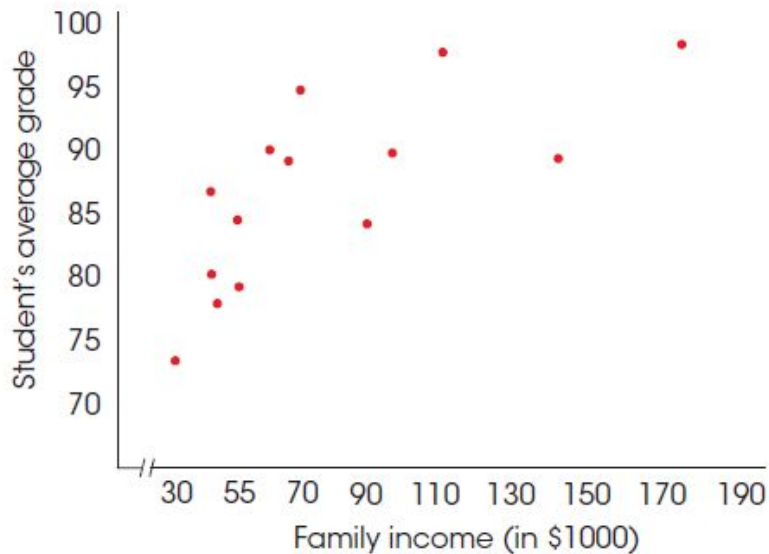
б) отсутствие корреляции,
коэффициент близок к нулю

с) позитивная корреляция,
примерно 0.7 (уже ближе к
сильной!)

д) очень слабая негативная
корреляция,
примерно -0.2

Есть ли взаимосвязь?

Person	Family Income (in \$1000)	Student's Average Grade
A	31	72
B	38	86
C	42	81
D	44	78
E	49	85
F	56	80
G	58	91
H	65	89
I	70	94
J	90	83
K	92	90
L	106	97
M	135	89
N	174	95



Когда и почему используем корреляцию?

Предсказания и исследование данных

Когда можно использовать одну переменную для предсказания другой переменной (например, результаты ЕГЭ -> средний балл в вузе)

Валидность

Например, чтобы оценить новую шкалу психологического теста - смотрим корреляцию ее результатов с результатами похожих тестов

Надежность

Чтобы понять, что шкала не подвержена ошибке - измеряем одного и того же испытуемого дважды, смотрим корреляцию результатов.

Как посчитать корреляцию?

Вы на самом деле уже умеете это делать!

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Коэффициент корреляции Пирсона. SS - sum of squares (сумма квадратов), ее формулу можно найти в семинаре 3!

$$SS_x = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

Аналогично
и для y

$$SS_{xy} = \sum_{i=1}^n (x_i y_i) - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

Пример

Таблица 7.1. Баллы за разделы Академического оценочного теста, направленные на проверку вербальных и математических умений

Ученик	Речь	Математика
1	490	560
2	500	500
3	530	510
4	550	600
5	580	600
6	590	620
7	600	550
8	600	630
9	650	650
10	700	750

$$SS_x = 3\,390\,500 - \frac{5\,790^2}{10} = 38\,090$$

$$SS_y = 3\,612\,500 - \frac{5\,970^2}{10} = 48\,410$$

$$SS_{xy} = 3\,494\,000 - \frac{(5\,790)(5\,970)}{10} = 37\,370$$

$$r = \frac{37\,370}{\sqrt{(38\,090)(48\,410)}} = 0.87$$

И коэффициент корреляции нужно возвести в квадрат (получим число в процентах - насколько процентов наши переменные изменяются вместе) и проверить на статистическую значимость.

$$R^2 = 0.7569$$

<https://poznayka.org/s71645t1.html>

<https://www.labirint.ru/books/450022/>

Линейная регрессия

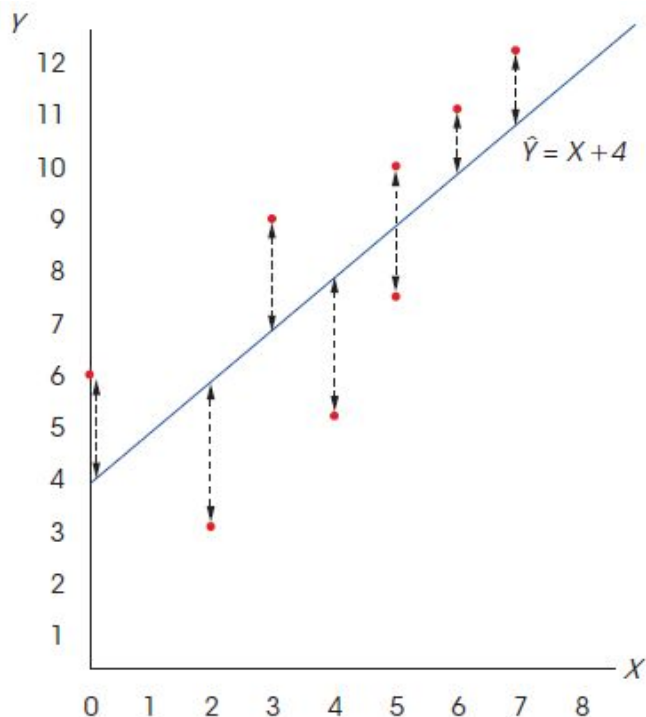
Одно из применений корреляции - предсказание.
Если между переменными существуют видимо
устойчивая связь,
то мы можем формализовать ее и предсказать
значение
переменной y от переменной x .

В целом, линейные отношения между двумя
переменными описываются
через уравнение

$$Y = bX + a,$$

где b - коэффициент, насколько изменяется Y
на изменение X на одну единицу, а a - константа..

Линейная регрессия

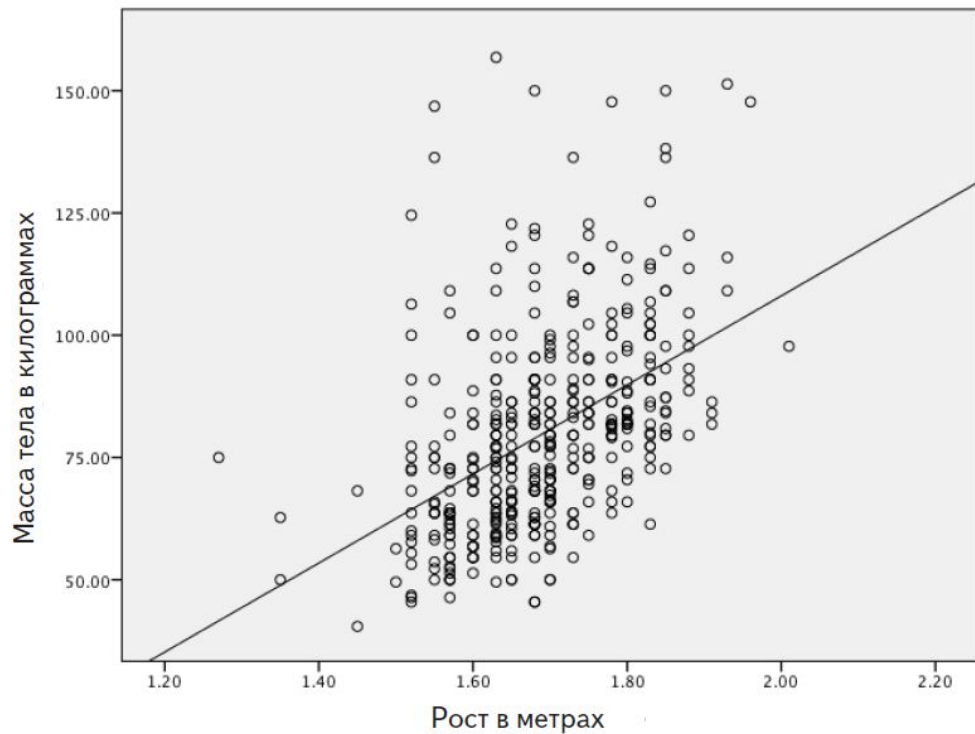


Линия регрессии предсказывает отношение между y и x .

Но она их усредняет (см. график), поэтому говорят об ошибке предсказания - насколько предсказанные значения отличаются от значений в данных.

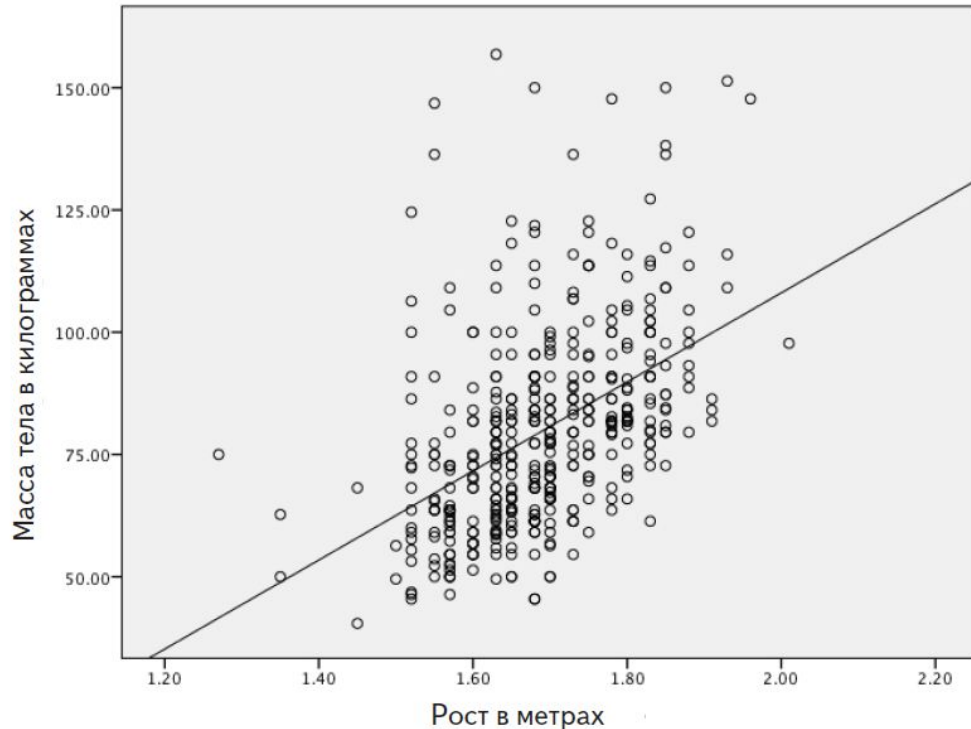
В примере слева коэффициент $X = 1$, а значит y изменяется на одну единицу на одну единицу изменения в X .

Пример



$$y = 91x - 74$$

Пример



$$y = 91x - 74$$

Коэффициент наклона равен 91, константа – 74. Для нахождения предсказываемого значения массы тела человека подставьте вместо x его рост в метрах, и останется только произвести вычисления. Это уравнение предсказывает, что человек с ростом 1,8 метра будет иметь массу тела 89,8 кг, потому что:

$$y = 91(1,8) - 74 = 89,8.$$

Перед тем, как строить уравнение регрессии для данных, нужно ознакомиться с допущениями и историей про статистическую значимость.

(на самом деле это не страшно)

<https://habr.com/ru/post/350668/>

Пример

https://nplus1.ru/material/2019/09/19/toloka?utm_source=Telegram&utm_medium=Social

Мезонин

Математика

20:38 | 19 Сен. 2019

Сложность | 0

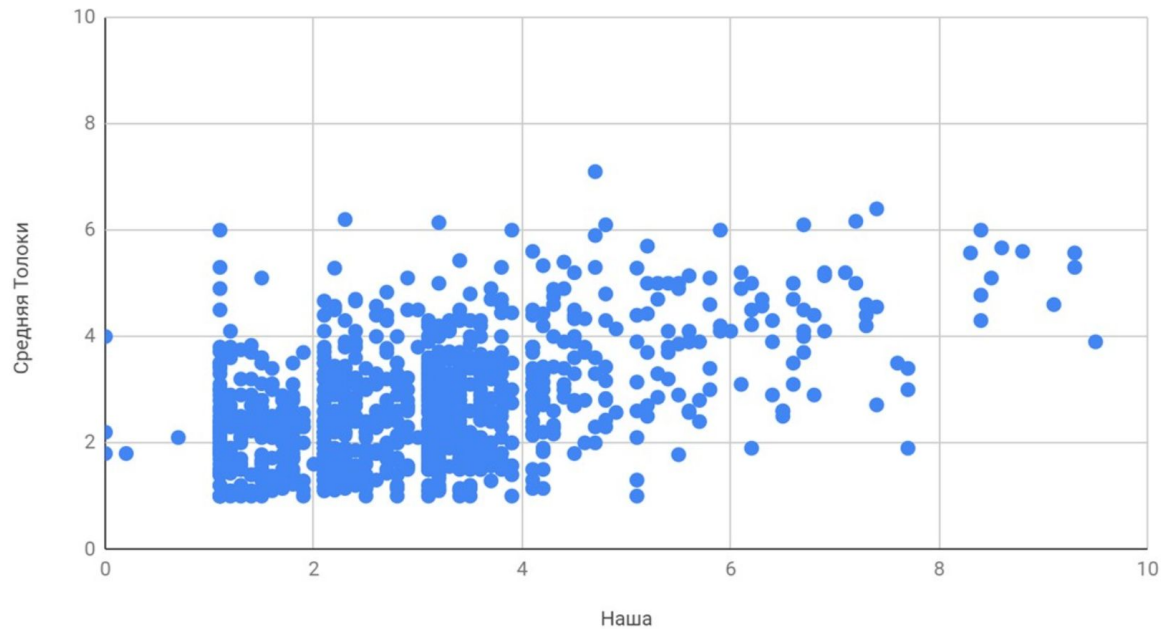


Деконструкция сложности

Яндекс.Толока помогла понять, легко ли читать новости на $N + 1$

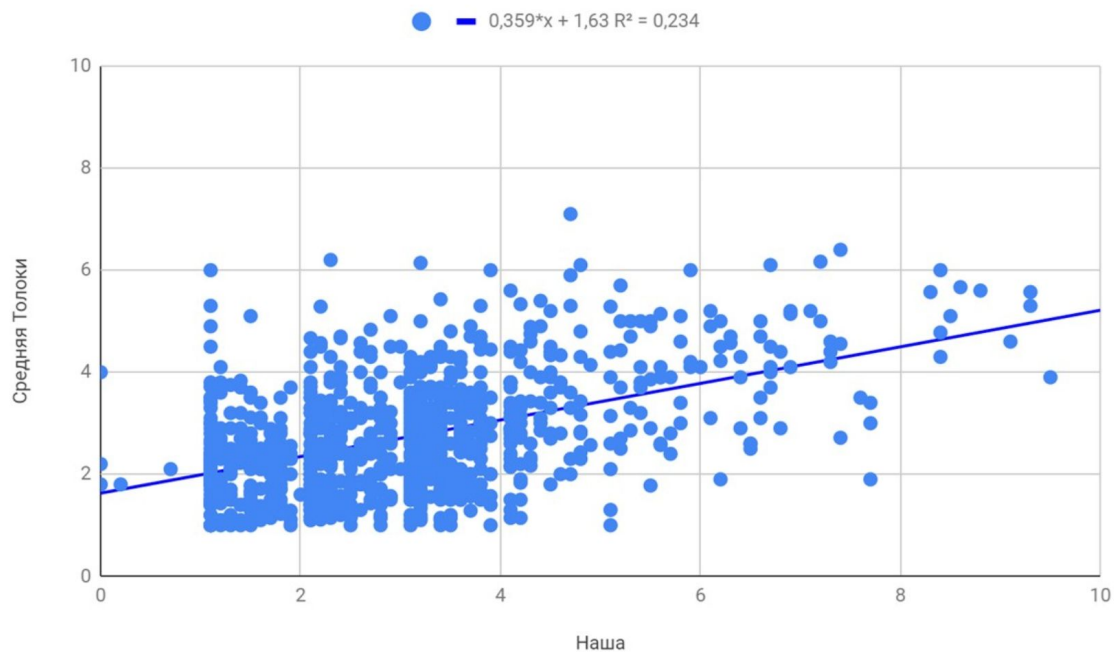
Пример

Средняя оценка Толоки vs. Оценка редакторов N + 1

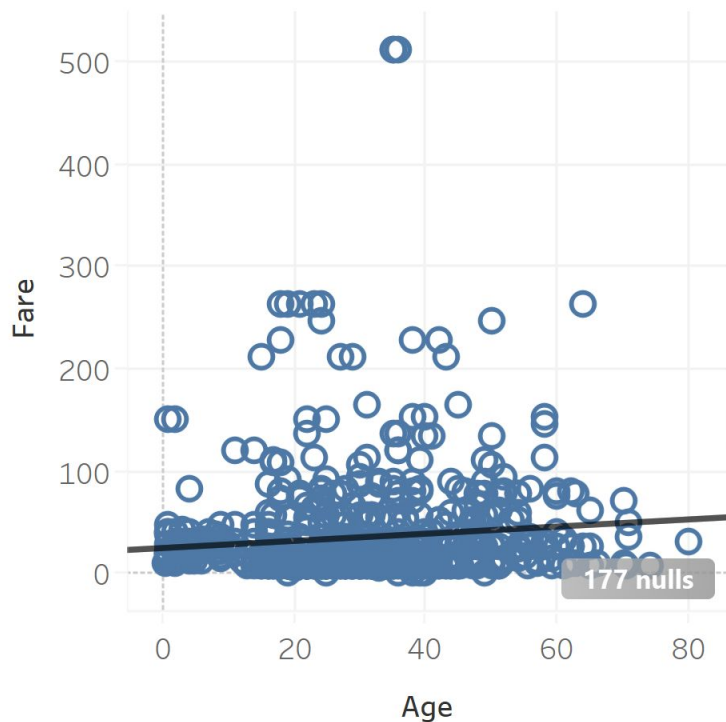


Пример

Средняя оценка Толоки vs. Оценка редакторов N + 1



Корреляция в Tableau



Fare = 0.349964*Age + 24.3009
R-Squared: 0.0092288
P-value: 0.0102163

Data

Analytics

Summarize

- Constant Line
- Average Line
- Median with Quartiles
- Box Plot
- Totals

Model

- Average with 95% CI
- Median with 95% CI
- Trend Line**
- Forecast
- Cluster

Интерпретируем корреляцию

Интерпретируем R^2

R коэффициент возведенный в квадрат - количество вариативности в двух переменных, которые объясняет корреляция. Можно переводить в проценты. Например, в предыдущем примере, можно сказать, что оценка по речевому экзамену на 75 % объясняет изменения в оценке по математике.

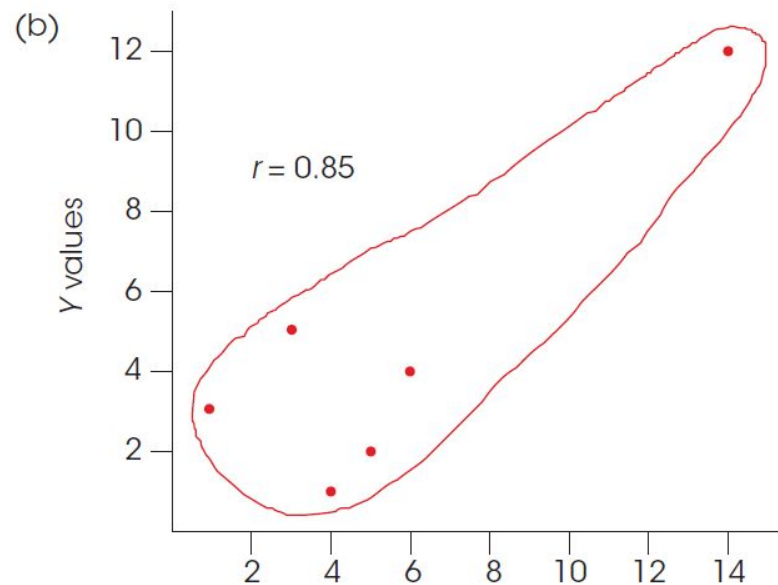
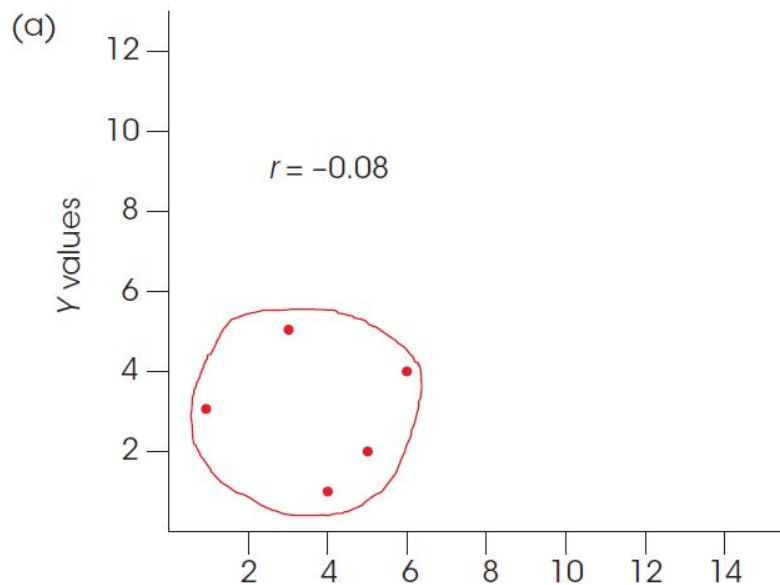
Выбросы

Один или два сильных выброса могут сильно повлиять на значение корреляции.

Отношение между переменными

Корреляция просто показывает нам отношение между переменными!
Не объясняет его! Не доказывает причинно-следственную связь.

Влияние выбросов



Correlation does not mean causation!



Альтернативные причины?

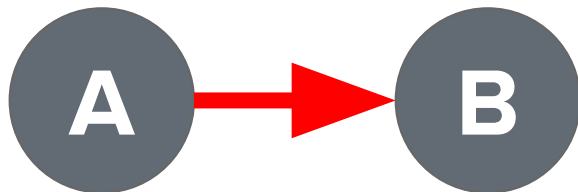
- Есть сильная связь между продажами мороженого и числом утонувших, так что причина этого в том, что люди идут купаться слишком рано после того, как съели мороженое, у них сводит мышцы, и они тонут.
- У школьников есть сильная связь между результатом теста на словарный запас и размером обуви, что можно объяснить тем, что у высоких людей мозг больше, и поэтому они могут запомнить больше слов.
- Число аистов в регионе сильно связано с уровнем рождаемости, так что, очевидно, аисты и правда приносят детей.
- Мэр города заметил сильную корреляцию между победами местной спортивной команды в соревнованиях и парадами и решил проводить больше парадов, чтобы улучшить результаты местных команд.

Альтернативные причины

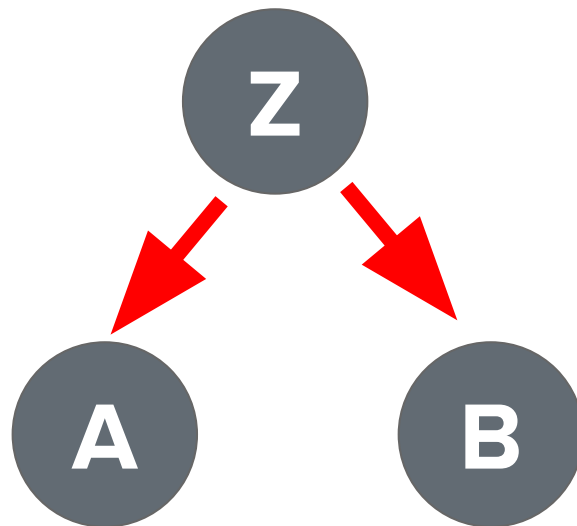
- И потребление мороженого, и плавание более обычны в теплое время года, так что очевидная связь объясняется влиянием третьего фактора, температуры (или времени года).
- Исследование проводили на школьниках, а их возраст не учитывали. Вероятно, старшие дети окажутся выше (с большим размером обуви) и будут иметь более обширный словарный запас, чем младшие дети; таким образом, наблюдаемая связь обусловлена третьей переменной, возрастом.
- Аисты чаще встречаются в сельской местности, а рождаемость также обычно выше вне городов, так что связь объясняется влиянием другого фактора, типа местности.
- Это обращенная причинно-следственная связь – парады проводят после побед в чемпионатах, так что успешный сезон для команд – это причина парадов, а не проведение парадов улучшает их результаты.

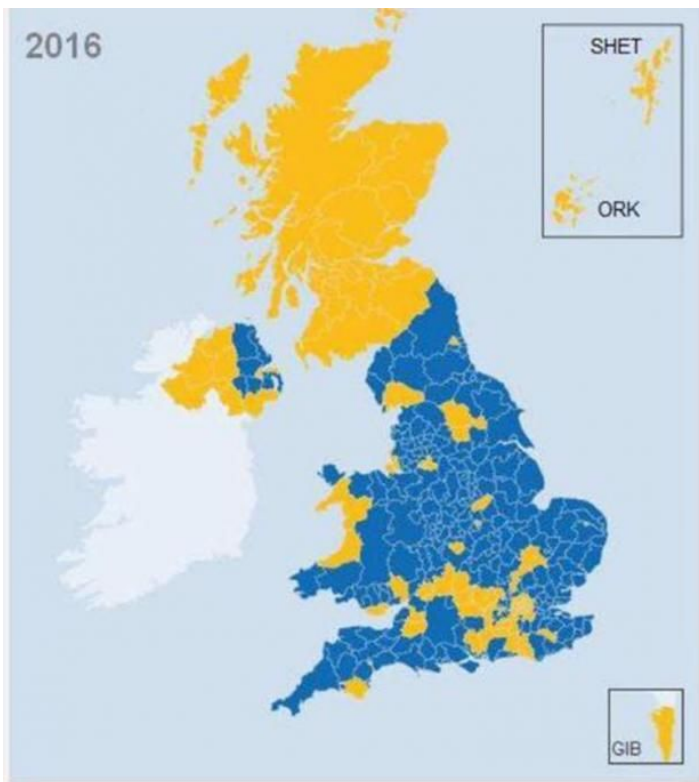
Искажающий фактор

Фактор Z, который одновременно
влияет на переменные A и B.

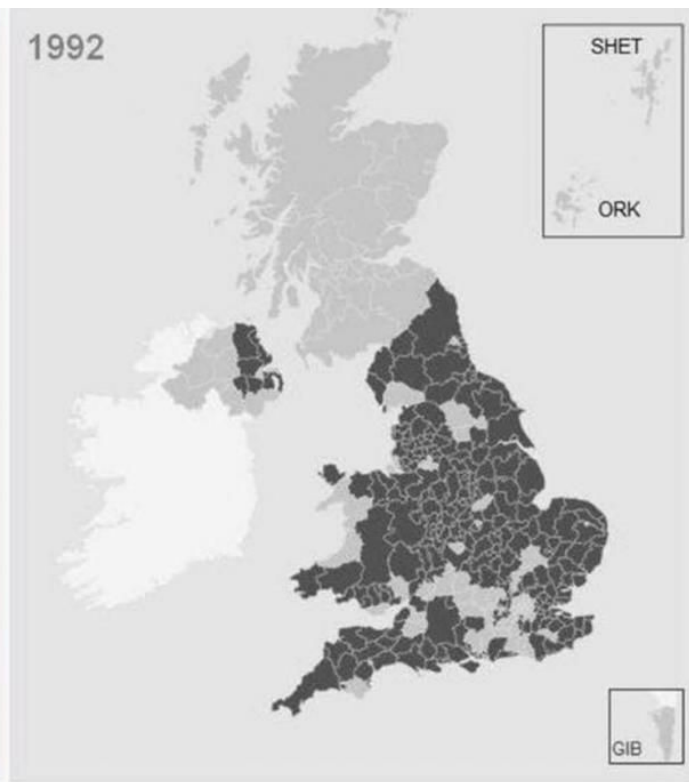


или





Key:
■ Majority leave ■ Majority remain



Key:
■ BSE-Areas ■ BSE-Free-Areas

Это графики
голосования за
брекзит в 2016
году и график
распространения
коровьего
бешенства в 1992
году.

Какая может быть
третья переменная?

Ложная корреляция (Spurious correlation)

Но иногда корреляция действительно может быть случайной и не иметь даже общей переменной. Нужно избегать соблазна ее использовать.

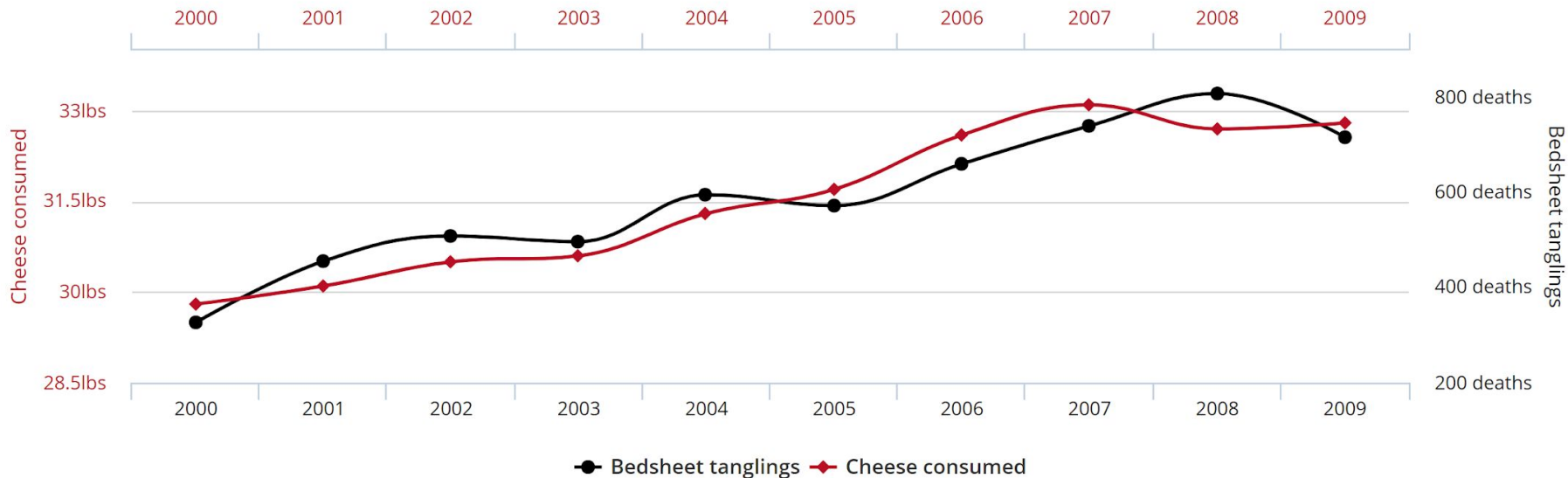


Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% ($r=0.947091$)

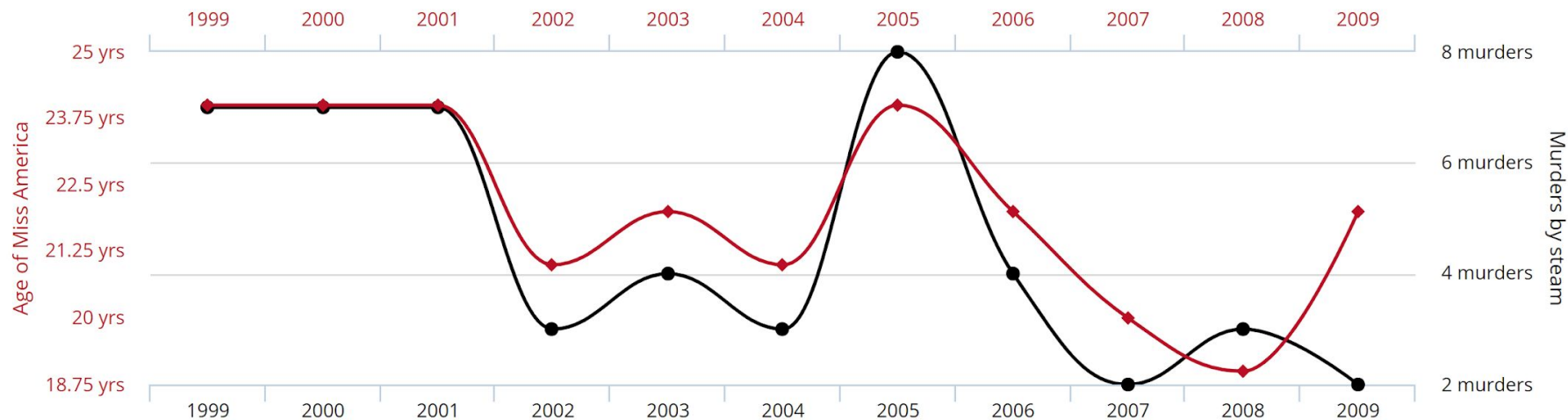


tylervigen.com

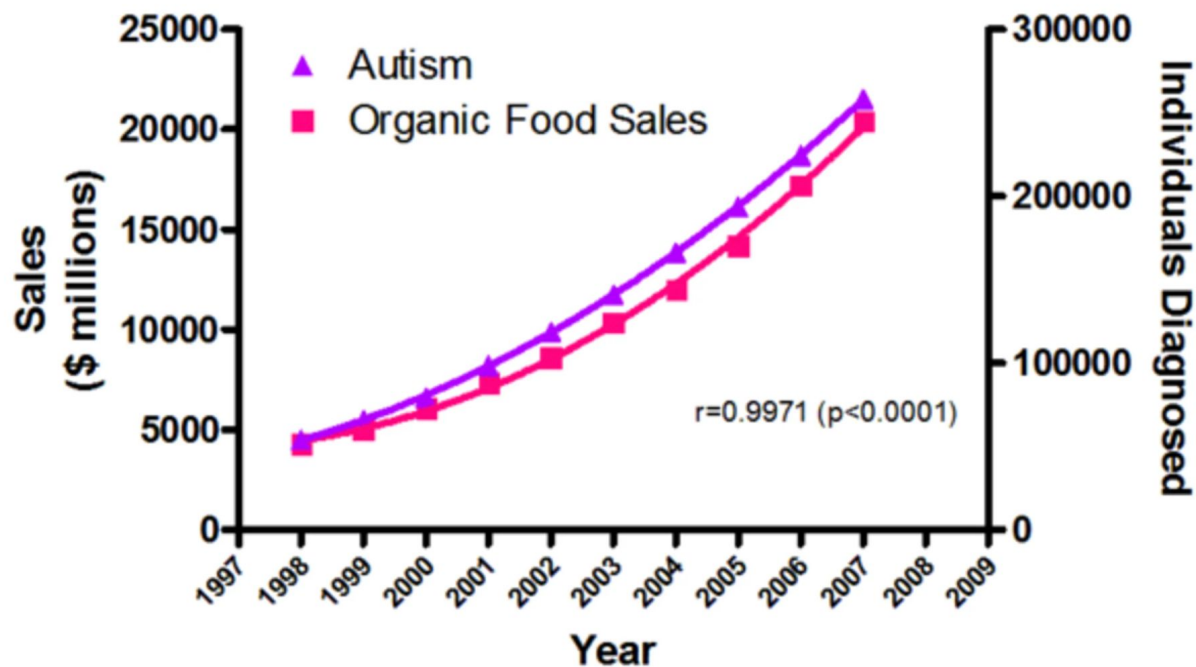


Age of Miss America correlates with Murders by steam, hot vapours and hot objects

Correlation: 87.01% ($r=0.870127$)



The real cause of increasing autism prevalence?

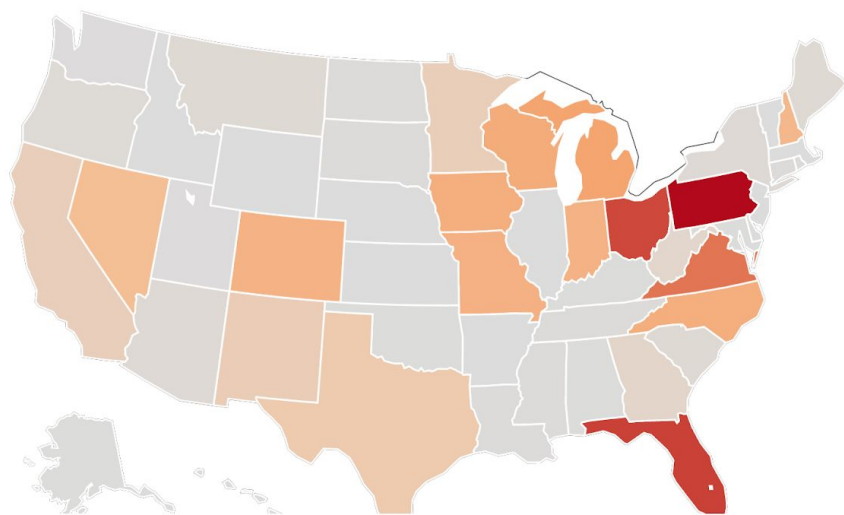


Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act"

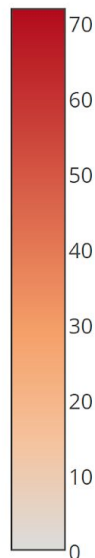
Но и у такой корреляции, например, нет научного объяснения.

И для разнообразия настоящая корреляция

2008 U.S. Presidential Campaign Ad Spending by State
(Hover for details)

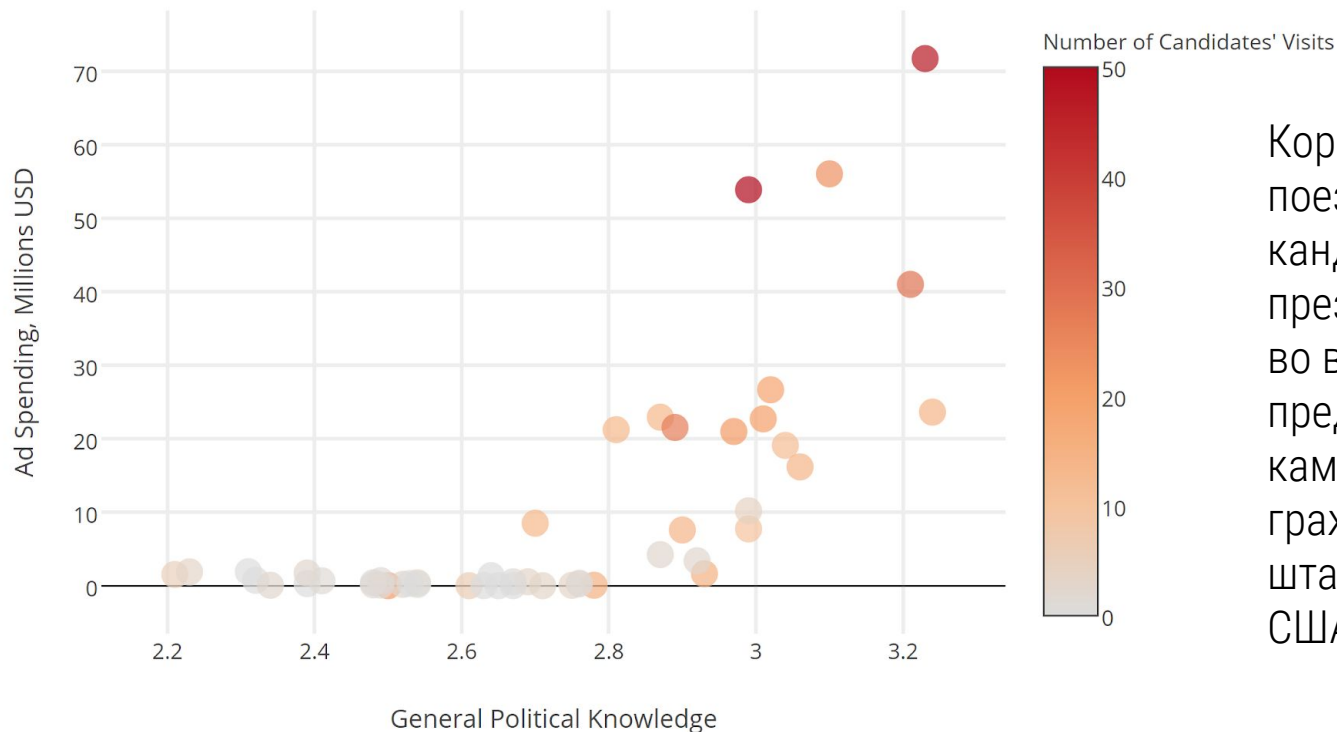


Millions USD



Предвыборные президентские кампании в США проходят на самом деле не по всей стране, а в нескольких штатах, в которых кандидаты имеют шанс повлиять на исход голосования.

И настоящие корреляции



Корреляция между поездками в штаты кандидатов в президенты во время предвыборной кампании и знаниях граждан этого штатов о политике США.