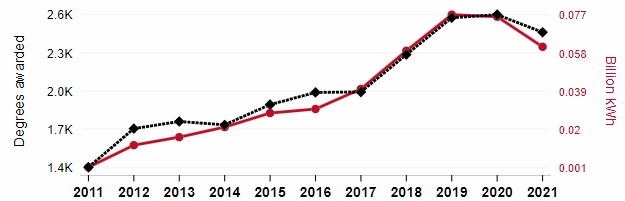
***Степени младшего специалиста, присуждаемые в области музыки и танцев и Солнечная энергия, вырабатываемая на Коста-Рика***



Корреляция r = 0,9876857 ( коэффициент корреляции Пирсона ).

Корреляция — это мера того, насколько переменные движутся вместе. Если оно равно 0,99, то когда одно растет, другое растет. Если оно равно 0,02, связь очень слабая или отсутствует. Если оно равно -0,99, то когда один растет, другой падает. Если оно равно 1,00, вы, вероятно, испортили свою корреляционную функцию.

r 2 = 0,9755230 ( Коэффициент детерминации )

. Это означает, что 97,6% изменения одной переменной (т. е. солнечной энергии, вырабатываемой в Коста-Рике) предсказуемо на основе изменения другой (т. е. степени младшего специалиста, присуждаемой в области музыки и танцев). за 11 лет с 2011 по 2021 год.

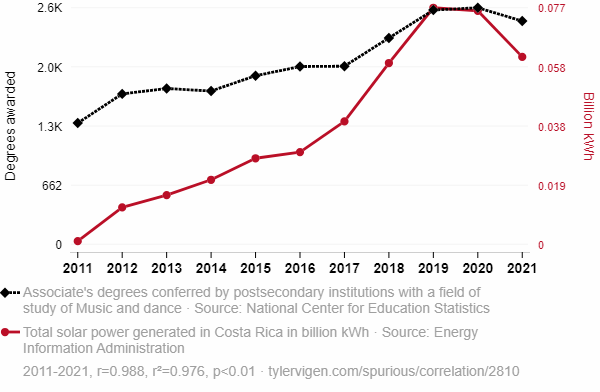
p <0,01, что является статистически значимым ( тест значимости нулевой гипотезы ).

Значение p составляет 1,5E-8. Показать Значение p является мерой того, насколько вероятно, что мы случайно найдем такой экстремальный результат. Примечание . В среднем вы обнаружите сильную корреляцию 0,99 в 1,5E-6% случайных случаев. Другими словами, если вы коррелируете 68 142 327 случайных величин с теми же 10 степенями свободы, обратите внимание , вы случайным образом ожидаете найти такую ​​сильную корреляцию, как эта. [ 0,95, 1 ] 95% доверительный интервал корреляции (с использованием z-преобразования Фишера ) Подробнее о доверительном интервале Все значения для лет, включенных выше:

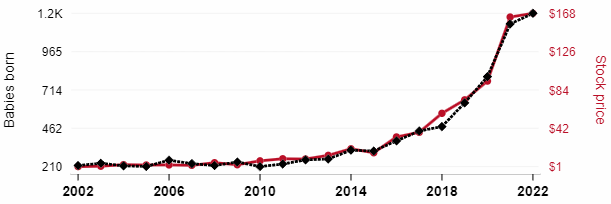
|  | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Присуждение степеней младшего специалиста в области музыки и танцев (присуждение степеней) | 1356 | 1683 | 1743 | 1715 | 1886 | 1989 | 1993 | 2309 | 2621 | 2647 | 2498 |
| Солнечная энергия, вырабатываемая в Коста-Рике (млрд кВтч) | 0,001 | 0,012 | 0,016 | 0,021 | 0,028 | 0,03 | 0,04 | 0,059 | 0,077 | 0,076 | 0,061 |

Почему это работает

1. Сбор данных: в моей базе данных 25 153 переменных. Я сравниваю все эти переменные друг с другом, чтобы найти те, которые случайно совпадают. Это 632 673 409 корреляционных вычислений! Это называется «извлечением данных». Вместо того, чтобы начать с гипотезы и проверить ее, я злоупотребил данными, чтобы увидеть, какие корреляции выявляются. Это опасный способ анализа, потому что любой достаточно большой набор данных совершенно случайным образом приведет к сильным корреляциям.
2. Отсутствие причинно-следственной связи: вероятно, нет прямой связи между этими переменными, несмотря на то, что говорит ИИ выше. Ситуация усугубляется тем, что я использовал «Годы» в качестве базовой переменной. За год происходит много событий, не связанных друг с другом! В большинстве исследований вместо «одного года» в качестве изучаемого объекта используется что-то вроде «один человек».
3. Наблюдения не являются независимыми. Для многих переменных последовательные годы не являются независимыми друг от друга. Если группа людей постоянно что-то делает каждый день, нет никаких оснований думать, что они внезапно изменят то, как они это делают 1 января. Простое вычисление значения p -примечания не учитывает это, поэтому математически это выглядит так: менее вероятно, чем это есть на самом деле.
4. Ось Y не начинается с нуля: я обрезал оси Y на графике выше. Я также использовал линейный график, который делает визуальную связь более заметной, чем она того заслуживает. Примечание. Математически то, что я показал, верно, но намеренно вводит в заблуждение. Ниже представлена ​​та же диаграмма, но обе оси Y начинаются с нуля.



***Популярность имени Стив и Цена акций Amazon.com (AMZN)***



Корреляция r = 0,9958805 ( [коэффициент корреляции Пирсона](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient) ).

Корреляция — это мера того, насколько переменные движутся вместе. Если оно равно 0,99, то когда одно растет, другое растет. Если оно равно 0,02, связь очень слабая или отсутствует. Если оно равно -0,99, то когда один растет, другой падает. Если оно равно 1,00, вы, вероятно, испортили свою корреляционную функцию.

r *2* = 0,9917779 ( [Коэффициент детерминации](https://en.wikipedia.org/wiki/Coefficient_of_determination) )

. Это означает, что 99,2% изменения одной переменной *(т. е. цены акций Amazon.com (AMZN))* можно предсказать на основе изменения другой *(т. е. популярности имени Стиви). )* в течение 21 года с 2002 по 2022 год.

p <0,01, что является статистически значимым ( [тест значимости нулевой гипотезы](https://en.wikipedia.org/wiki/P-value) )

. Значение *p* составляет 2,8E-21. *Показать* Значение *p* является мерой того, насколько вероятно, что мы случайно найдем такой экстремальный результат. *Примечание* . В среднем вы обнаружите сильную корреляцию, равную 1, в 2,8–19% случайных случаев. Другими словами, если вы сопоставили 354 152 883 479 132 766 208 случайных величин с теми же 20 степенями свободы, *обратите внимание,* вы случайным образом ожидали бы найти такую ​​сильную корреляцию, как эта. [ 0,99, 1 ] 95% [доверительный интервал](https://en.wikipedia.org/wiki/Confidence_interval) корреляции (с использованием [z-преобразования Фишера](https://en.wikipedia.org/wiki/Fisher_transformation) ) *Подробнее о доверительном интервале* Все значения для лет, включенных выше:

All values for the years included above: *Note*

|  | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Popularity of the first name Stevie (Babies born) | 210 | 227 | 254 | 260 | 318 | 312 | 379 | 444 | 473 | 629 | 801 | 1147 | 1217 |
| Amazon.com's stock price (AMZN) (Stock price) | 6.81 | 9.07 | 8.79 | 12.8 | 19.94 | 15.63 | 32.81 | 37.9 | 58.6 | 73.26 | 93.75 | 163.5 | 167.55 |

Почему это работает

1. Сбор данных: в моей базе данных 25 153 переменных. Я сравниваю все эти переменные друг с другом, чтобы найти те, которые случайно совпадают. Это 632 673 409 корреляционных вычислений! Это называется «извлечением данных». Вместо того, чтобы начать с гипотезы и проверить ее, я злоупотребил данными, чтобы увидеть, какие корреляции выявляются. Это опасный способ анализа, потому что любой достаточно большой набор данных совершенно случайным образом приведет к сильным корреляциям.
2. Отсутствие причинно-следственной связи: вероятно, *нет* прямой связи между этими переменными, несмотря на то, что говорит ИИ выше. Ситуация усугубляется тем, что я использовал «Годы» в качестве базовой переменной. За год происходит много событий, не связанных друг с другом! В большинстве исследований вместо «одного года» в качестве изучаемого объекта используется что-то вроде «один человек».
3. Наблюдения не являются независимыми. Для многих переменных последовательные годы не являются независимыми друг от друга. Если группа людей постоянно что-то делает каждый день, нет никаких оснований думать, что они внезапно *изменят* то, как они это делают 1 января. Простое вычисление *значения p* -примечания не учитывает это, поэтому математически это выглядит так: менее вероятно, чем это есть на самом деле.
4. Диковинные выбросы: в этих данных есть «выбросы». *Примечание.* Они выделяются на диаграмме рассеяния выше: обратите внимание на точки, которые находятся далеко от других точек. Я намеренно неправильно обработал выбросы, из-за чего корреляция выглядит очень сильной.