Статистический анализ данных

Занятие 6

КАРЛ ПИРСОН

АНГЛИЙСКИЙ МАТЕМАТИК, СТАТИСТИК, БИОЛОГ И ФИЛОСОФ, ОСНОВАТЕЛЬ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Карл Пирсон опубликовал основополагающие труды по математической статистике (более 400 работ по этой теме). Разработал теорию корреляции, критерии согласия, алгоритмы принятия решений и оценки параметров.

Случайные величины

– это значения некоторых свойств объектов, которые мы исследуем.

Корреляционный анализ

– статистический метод изучения взаимосвязи между двумя и более случайными величинами.

Суть корреляционного анализа

ВЫЯВЛЕНИЕ И ОЦЕНКА СИЛЫ СВЯЗИ МЕЖДУ СЛУЧАЙНЫМИ ВЕЛИЧИНАМИ (ПРИЗНАКАМИ), КОТОРЫЕ ХАРАКТЕРИЗУЮТ НЕКОТОРЫЙ РЕАЛЬНЫЙ ПРОЦЕСС.

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

ПРИМЕНЯЕТСЯ ТОЛЬКО ДЛЯ АНАЛИЗА СВЯЗИ КОЛИЧЕСТВЕННЫХ И/ИЛИ КАЧЕСТВЕННЫХ

ПОРЯДКОВЫХ ПРИЗНАКОВ

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

(СТЕПЕНЬ ВЗАИМОСВЯЗИ), МОЖЕТ ПРИНИМАТЬ ЗНАЧЕНИЯ ОТ -1 ДО +1:

Если коэффициент отрицательный – зависимость обратная, т.е. увеличение одной величины приводит к уменьшению второй и наоборот.

Если коэффициент положительный – зависимость прямая, т.е. увеличение одного показателя приводит к увеличению второго и наоборот

Прямая связь

Представьте, что вы аналитик в автобусном парке. Водитель в конце смены сдает вам маршрутный лист, где есть данные о времени движения по маршруту и пройденное расстояние. Все данные внесены в таблицу и проведен анализ. Посмотрите на результат.

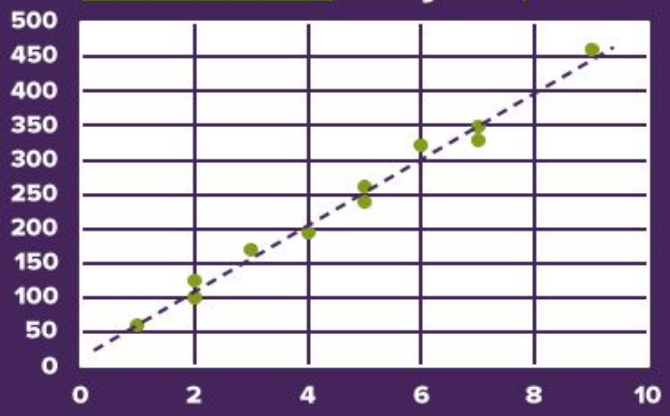
время	путь
2	100
3	170
1	60
2	101
4	195
5	239
9	461
7	349
2	125
6	322
7	329
5	261

КОРРЕЛЯЦИЯ

0,99553

сильная прямая

$y=50{,}591*x$



ВЫЯВЛЕНА СВЯЗЬ ПУТИ ОТ ВРЕМЕНИ

корреляция более **0,75**

УРАВНЕНИЕ ЗАВИСИМОСТИ

показывает нам среднюю скорость около **51 км/ч**

Обратная связь

Вклад в будущее

АКАДЕМИЯ искусственного интеллекта для школьников

Представьте, что вы собираетесь в поход и надо рассчитать продуктовый запас. Для этого взяли данные о других походах, где отражается время в пути и запас энергетических сил. Все данные внесены в таблицу и проведен анализ. Посмотрите на результат.

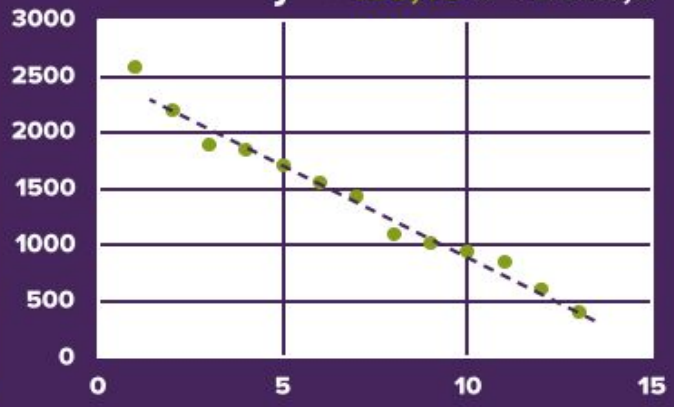
время	путь
1	2583
2	2196
3	1891
4	1850
5	1713
6	1560
7	1435
8	1101
9	1015
10	943
11	851
12	611
13	402

КОРРЕЛЯЦИЯ

- 0,9893

сильная, обратная

$y = -163{,}45x + 2540{,}3$



ВЫЯВЛЕНА СВЯЗЬ ЖИЗНЕННОЙ ЭНЕРГИИ ОТ ВРЕМЕНИ В ПУТИ

корреляция более **0,75**

ПОДУМАЙТЕ, ЧТО НАМ ПОКАЗЫВАЮТ ПОЛУЧЕННЫЕ ДАННЫЕ:

- 163,45 и 2540,3

Отсутствие связи

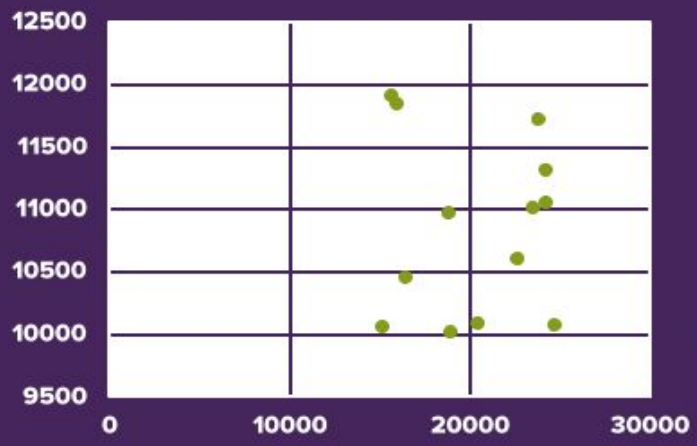
Представьте, что у папы и мамы разное образование, квалификация и прочее. У вас есть информация об их заработной плате за несколько месяцев. Все данные внесены в таблицу и проведен анализ. Посмотрите на результат.

ЗП папы	ЗП мамы
23465	11016
20432	10090
15635	11913
24190	11323
18905	10016
24209	11057
23756	11722
15930	11853
18786	10972
16408	10461
24654	10075
15130	10061
22652	10607

КОРРЕЛЯЦИЯ

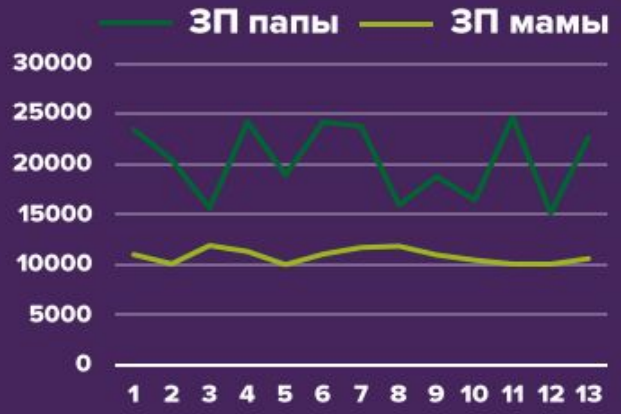
- 0,02026

отсутствует



Вклад в будущее

АКАДЕМИЯ



СВЯЗЬ НЕ ВЫЯВЛЕНА МЕЖДУ ЗП МАМЫ И ЗП ПАПЫ

корреляция менее 0,5

ПОДУМАЙТЕ, ПОЧЕМУ МЕЖДУ ЭТИМИ ДАННЫМИ НЕТ СВЯЗИ

Сила зависимости

PERFECT	+1	-1
STRONG	+0,9 +0,8 +0,7	-0,9 -0,8 -0,7
MODERATE	+0,6 +0,5 +0,4	-0,6 -0,5 -0,4
WEAK	+0,3 +0,2 +0,1	-0,3 -0,2 -0,1
ZERO	0	

Сила зависимости определяется по модулю коэффициента корреляции. Чем больше значение, тем сильнее изменение одной величины влияет на другую. Исходя из этого, при нулевом коэффициенте можно утверждать, что взаимосвязь отсутствует.

Практические задания