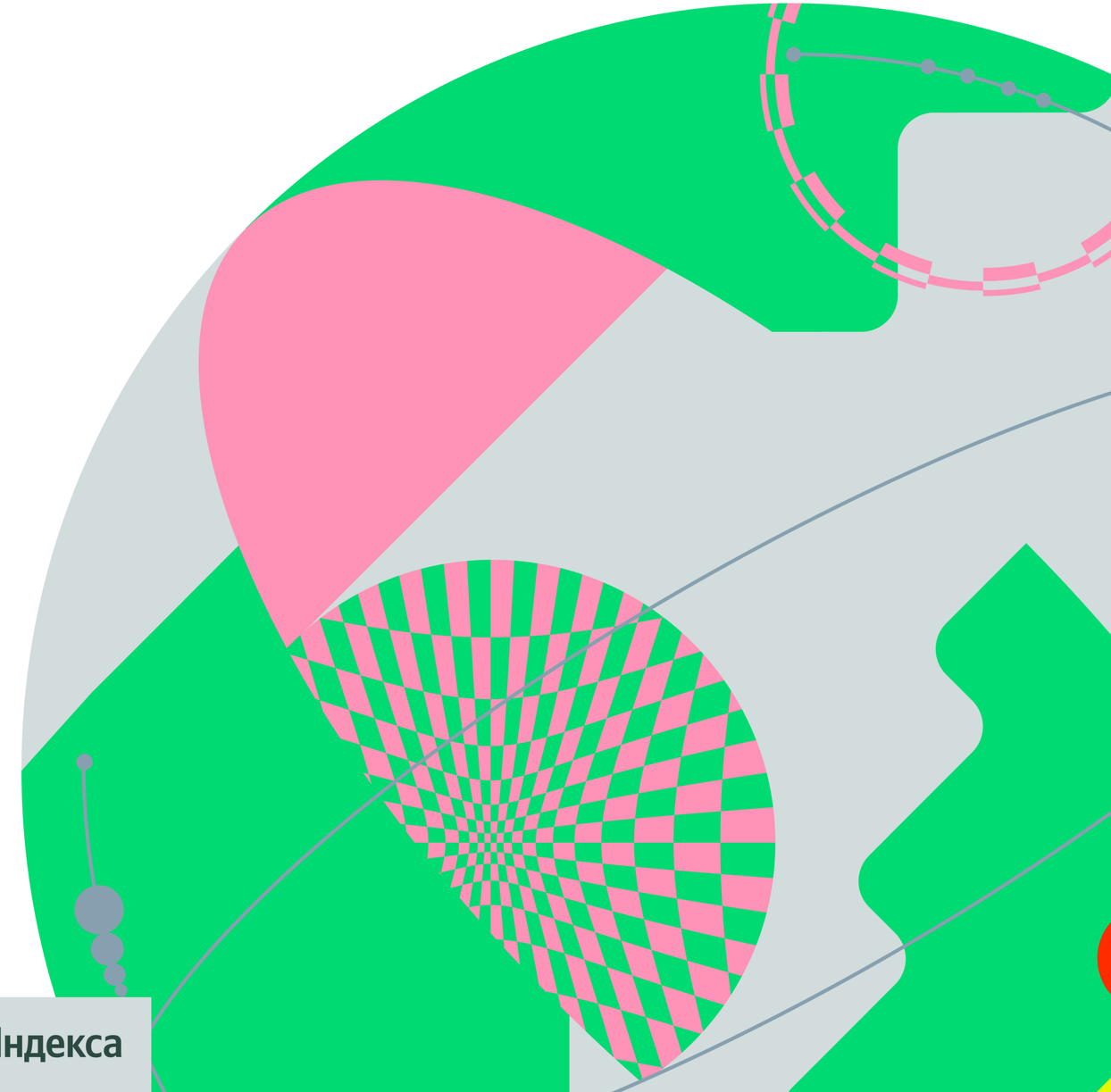
Взаимосвязь факторов друг с другом



Академия Яндекса

- Аналитическое мышление
- Технический навык
- Математика и алгоритмы
- Визаулизация, умение презентовать результат

Вместо предисловия

ума умным
наберёшься
сам поведёшься
С

Абсолютно не намекаю на нашу встречу))

Вместо предисловия

ума умным
наберёшься
сам поведёшься
С

С умным поведёшься
сам
ума наберёшься

Абсолютно не намекаю на нашу встречу))
Просто хочу показать какая интересная аналитическая история скрыта в русских пословицах. Видимо, аналитик – древнейшая профессия☺

ЧТО ПОСЕЕШЬ, ТО ПОЖНЕШЬ.

ЕСЛИ НЕТ ХОРОШИХ МЫСЛЕЙ, ТО И НЕТ ХОРОШИХ РЕЗУЛЬТАТОВ.

Что было раньше – курица или яйцо?

Как вы думаете,
что объединяет все приведённые пословицы?

Что было раньше – курица или яйцо?

Одинаковая аналитическая конструкция: *если-то* или (что для нас аналитиков должно быть привычнее) *причина-следствие*.

ПРИЧИНА → **СЛЕДСТВИЕ**

Любая аналитическая задача для любой аналитической специальности как раз ведь и вскрывает эту связь между причиной и следствием.

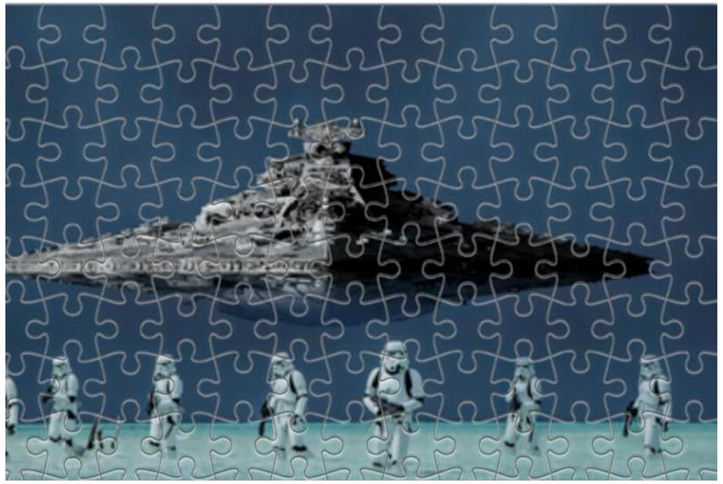
Академия Яндекса

А пазлы то причём тут)

В одной книге мне встретился такой интересный пример: (далее абзац плагиата)

Представьте, что вам предложили тысячу элементов для составления пазла, но на коробке при этом нет изображения того, что должно в итоге получиться. По мере сортировки элементов вы выделили группу элементов голубого цвета. Вероятно, это небо. Группа элементов зеленого цвета может изображать траву. Вот вы нашли глаз. Но чей — животного или человека? У вас появляется смутное представление о картинке в целом, но не хватает деталей.

Детали возникают, когда вы начинаете соединять смежные элементы, например, элементы с изображением глаза и элементы с изображением уха



Что дальше

Давайте рассмотрим эту ситуацию с точки зрения практической аналитики.
(Прошлые наши встречи были посвящены продуктовой и маркетинговой аналитики. Сегодня посвятим себя промышленной)

Предположим, вы работаете на ремонтной базе Kelvas. Там есть разные цеха, разные конвейерные линии, рядом - верфи по производству космических судов. Конечно же всё автоматизировано, работают роботы.

Но цикл постоянного улучшения качества Деминга был распространён, как религия. Поэтому данные о процессах ремонта и производства собирались бесперебойно.

было замечено:

время ремонта кораблей постоянно увеличивается

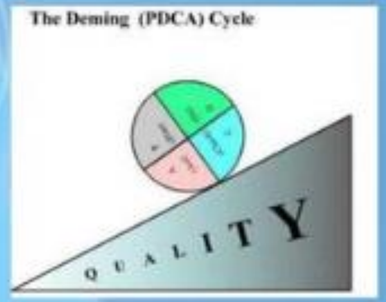
Постоянное улучшение качества

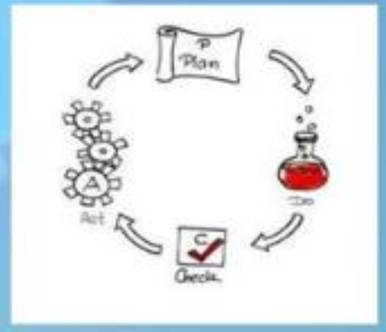
КРУГ КАЧЕСТВА

Цикл Деминга включает:

- ***Plan*** – планирование
- ***Do*** – выполнение
- ***Check*** – проверка
- ***Act*** – воздействие (корректирующее действие)

❖ **PDCA цикл (Plan-Do-Check-Act):** планирование – осуществление – проверка – претворение в жизнь) является широко распространенным методом непрерывного улучшения качества. При помощи постоянных проверок до, во время и после процесса производства, воспитания ответственности за качество и, прежде всего, при помощи постоянного аудита процесса производства могут быть обнаружены слабые места в разных процессах на предприятии. PDCA служит именно для обнаружения причин брака и поддержки всего процесса вплоть до устранения дефектов.

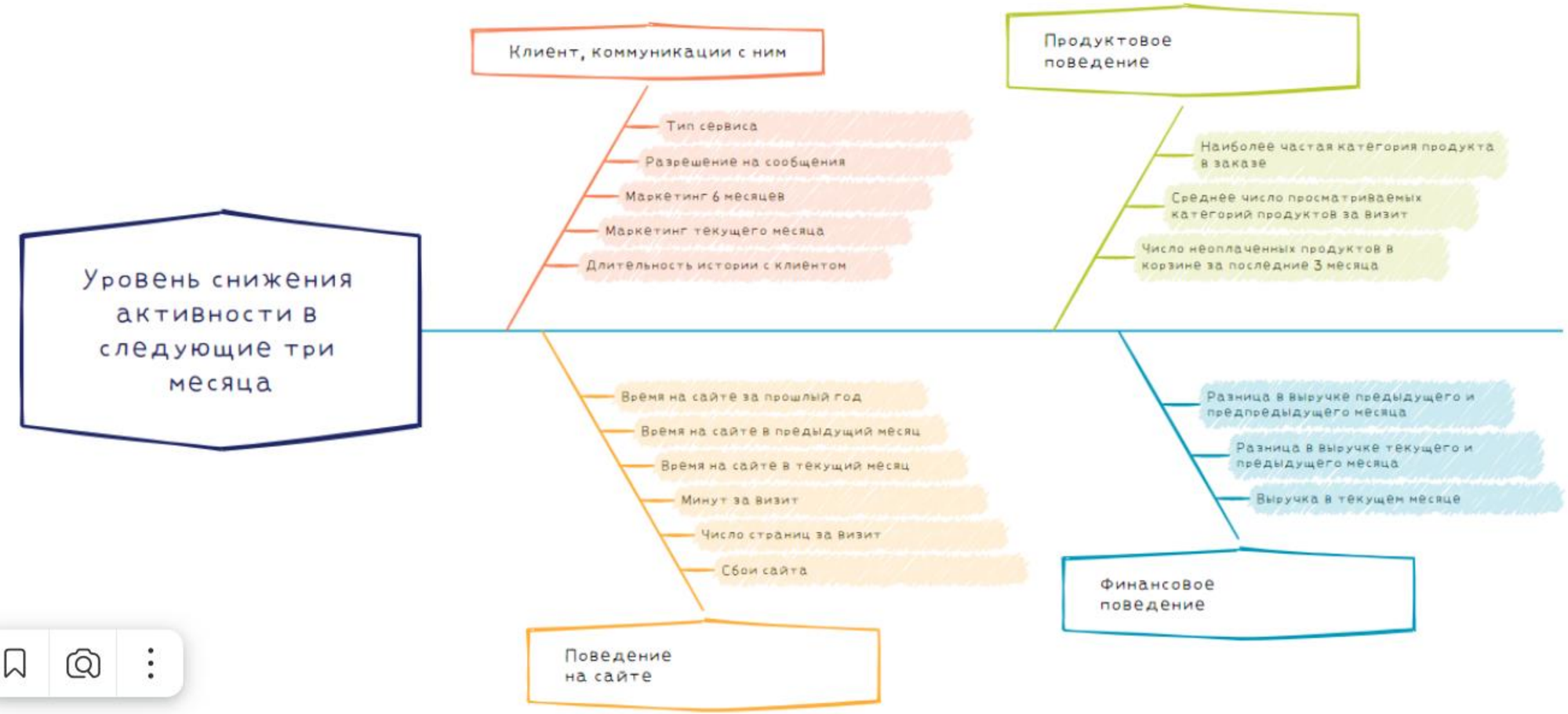




Что с базой?

И так, было замечено, что время ремонта кораблей постоянно увеличивается. Поставили задачу – найти причину и устранить её.

С чего начнём?



Визуальный метод анализа причинно-следственных связей, популярный в управлении бизнесом. Диаграмма названа в честь профессора Токийского университета Каору Исикава. Из-за внешнего вида диаграмму иногда называют «рыбий скелет» или «рыбья кость» (англ. fishbone diagram).

Итоги

правило убедительности обозначенных причин
ДОВЕРИЕ и АРГУМЕНТАЦИЯ.

Иными словами, если доказательства причин влияния убедительны, то переходим к этапу **принятия решения**

А что значит «убедительная связь»

В мысли выше, есть интересный для нас момент – «…убедительная связь….» А что это значит? Как минимум, математически доказать наличие взаимосвязи. Часто это и будет верной аргументацией к выводам.

Т.е. нужно найти математическую взаимосвязь между причиной и следствием, между фактором и целевой.

А как мы говорили на первой (или второй встрече) – разные типы шкал по-разному «работают» с математикой.

По этой причине, когда встаёт задача оценить причинно-следственную связь (или по-другому – оценить связь между факторами), то, исходя из наличия двух больших групп типов факторов (количественная и качественная(категориальная/номинальная)), принято рассматривать три варианта её решения.

- когда причина и следствие в количественной шкале,
- когда причина и следствие – в номинальной шкале,
- когда причина и следствие – в разных шкалах.

Далее, последовательно рассмотрим каждый тип взаимосвязи.

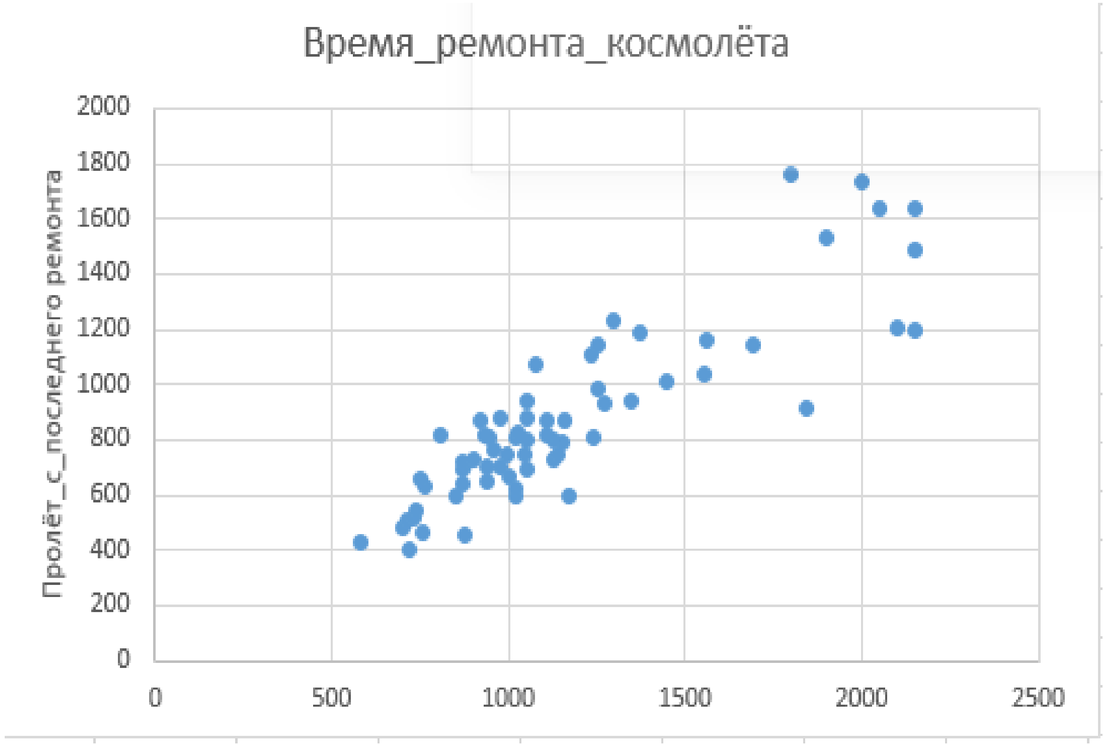
Данные с ремонтной базы

Пролёт_последнего_ремонта	Время_ремонта_космолёта	Наличие_пробоин	Тип_космолёта	Номер_бригады
720	398	нет	Венатор	бригада_2
580	426	нет	Венатор	бригада_2
875	456	нет	Венатор	бригада_3
759	461	нет	Венатор	бригада_3
699	481	нет	Венатор	бригада_2
710	504	нет	Корвет_C90	бригада_2
729	513	да	Корвет_C90	бригада_2
739	541	нет	Венатор	бригада_3
1170	600	да	Корвет_C90	бригада_3
1020	600	нет	Венатор	бригада_3
850	600	нет	Венатор	бригада_1
1020	626	нет	Венатор	бригада_2
766	634	нет	Венатор	бригада_3
870	638	нет	Венатор	бригада_3
940	647	нет	Венатор	бригада_3
749	656	нет	Венатор	бригада_2
1000	668	нет	Венатор	бригада_3
1049	690	нет	Венатор	бригада_3
869	694	нет	Корвет_C90	бригада_2

Количественная-количественная

Исследуем связь между «временем_ремонта» и «налёта_от последнего_ремонта».

Как и в прошлый раз, когда говорили о нормальности распределения, начнём с графической интерпретации. Для этого используется график «Диаграмма рассеяния». Обычно он носит ещё название скатерплот (именно так называются и функции в некоторых библиотеках).



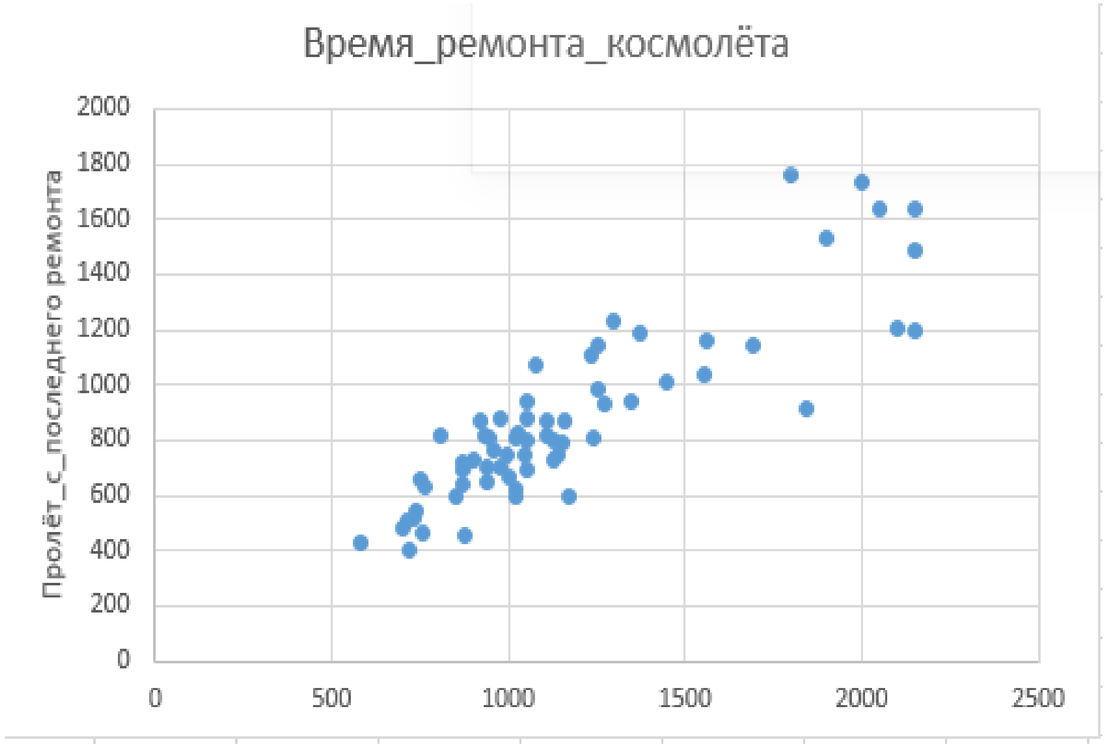
Видим, что чем больше один фактор, тем больше другой. Можем предположить, что чем больше «пролёт_с_последенего ремонта», тем больше «Время_проведения_ремонта». Это логично.

Количественная-количественная

Графически это увидели.

Есть ли метод, описывающий связь факторов в виде числа. Т.е. есть ли аналитический метод.&*

Есть. Он называется «коэффициент линейной корреляции».
А раз коэффициент, то значит есть и математическая формула:



$$\frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{Sxy}{\sqrt{Sxx\ Syy}}$$

Давайте разберём на не большом примере этот расчёт – он не сложный. А потом уже сутейно вникнем в закулисье цифр и чисел.

Количественная-количественная

$$\frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{Sxy}{\sqrt{Sxx\ Syy}}$$

	Пролёт_последнего_ремонта	Время_ремонта_космолёта					
	x	y	x - среднее	y - среднее	(x - среднее)^2	(y - среднее)^2	(x - среднее)* (y - среднее)
	580	426	-250	-119	62385	14070	29626
	875	456	45	-89	2046	7853	-4008
	759	461	-71	-84	5008	6992	5917
	699	481	-131	-64	17101	4047	8319
	710	504	-120	-41	14345	1650	4864
	729	513	-101	-32	10154	1000	3186
	739	541	-91	-4	8239	13	328
	1170	600	340	55	115757	3067	18844
	1020	600	190	55	36188	3067	10536
	850	600	20	55	409	3067	1120
	1020	626	190	81	36188	6623	15482
	766	634	-64	89	4067	7990	-5700
	870	638	40	93	1619	8721	3757
Сумма	10787	7080	0	0	313504	68159	92272
Среднее	830	545			S_{xx}	S_{yy}	S_{xy}

$$\frac{S_{xy}}{\sqrt{(S_{xx} * S_{yy})}} = 0.63$$

В общем случае формула записывается чуть иначе. Там будут ещё иные термины (ковариация, например). Но сейчас обойдёмся без неё. Что тут видно? В числителе – сумма произведений разностей факторов. В знаменателе – произведение стандартных отклонений каждого фактора.

А если проще, то коэффициент корреляции определяется в статистике как измерение силы связи между двумя переменными и их ассоциации друг с другом.

А если ещё проще, коэффициент корреляции рассчитывает эффект изменения одной переменной при изменении другой переменной.

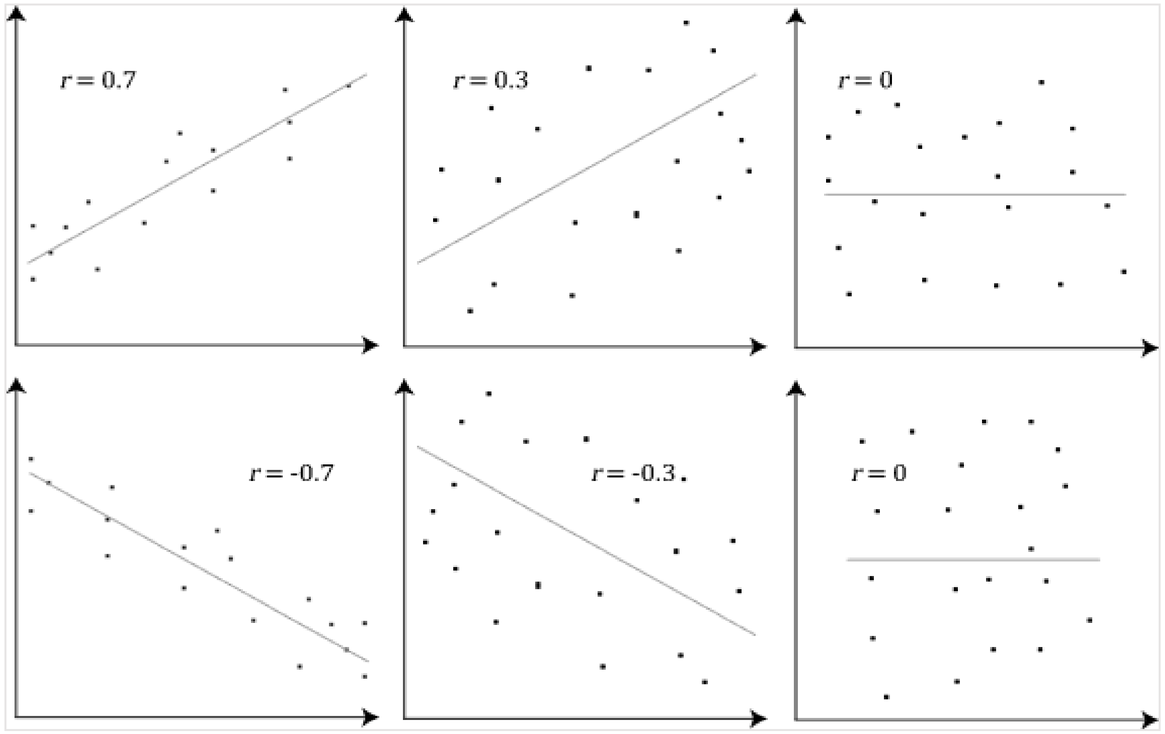
Что мы и проговорили выше, глядя на график.

Количественная-количественная

Предельные значения этого коэффициента (по модулю) такие:

Шкала Чеддока для классификации силы связи

Величина коэффициента множественной корреляции R	Оценка силы связи
0,1—0,3	Слабая
0,3—0,5	Умеренная
0,5—0,7	Заметная
0,7—0,9	Высокая
0,9—0,99	Весьма высокая



Исходя из новых знаний, что можем сказать о связи между «временем_ремонта» и «налёта_от последнего_ремонта»?

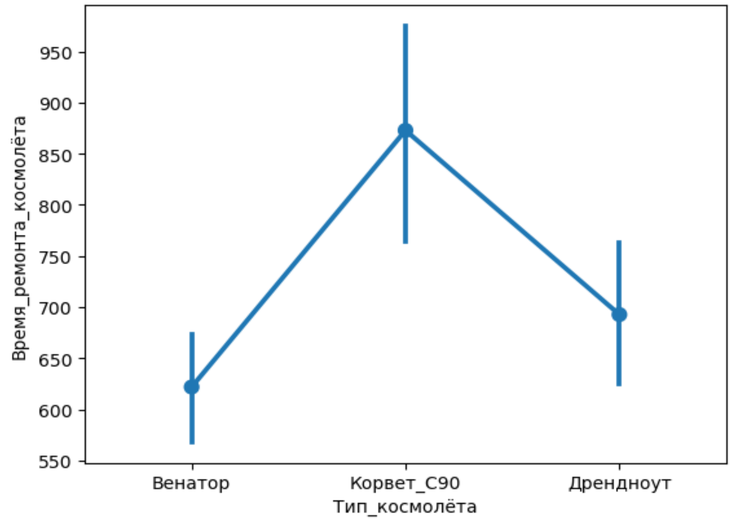
Инструмент очень удобный. Но, как у любого инструмента, у него есть ограничения по использованию. Обязательно их рассмотрим, но позже. Когда изучим остальные типы взаимосвязи между факторами.

Количественная – номинальная

А теперь давай исследовать два фактора разных типов шкал. Посмотрим, как связаны «время_ремонта» и «тип_космолёта». Данные по этим двум факторам выделим в отдельную таблицу:

Время_ремонта_космолёта	Тип_космолёта
398	Венатор
426	Венатор
456	Дрендноут
461	Венатор
481	Венатор
504	Корвет_С90
513	Корвет_С90
541	Венатор
600	Корвет_С90
600	Венатор
600	Дрендноут
626	Венатор
634	Венатор
638	Дрендноут
647	Венатор
656	Дрендноут
668	Венатор
690	Венатор
694	Корвет_С90
700	Венатор
701	Дрендноут
707	Венатор

	Венатор	Корвет_С90	Дрендноут	
	398	504	456	
	426	513	600	
	461	600	638	
	481	694	656	
	541	810	701	
	600	816	721	
	626	820	725	
	634	866	750	
	647	867	800	
	668	930	880	
	690	1035		
	700	1076		
	707	1161		
	731	1193		
	743	1209		
	750			
	768			
Сумма	10571	13094	6927	30592
Средне	622	873	693	729



Количественная - номинальная

	Венатор	Корвет_С90	Дрендноут			Венатор	Корвет_С90	Дрендноут	
	398	504	456			50097	136112	56027	
	426	513	600			38347	129552	8593	
	461	600	638			25864	74493	2992	
	481	694	656			19831	32017	1347	
	541	810	701			6532	3961	69	
	600	816	721			476	3241	801	
	626	820	725			17	2802	1043	
	634	866	750			148	48	3283	
	647	867	800			634	35	11513	
	668	930	880			2132	3257	35081	
	690	1035				4648	26266		
	700	1076				6112	41236		
	707	1161				7255	82982		
	731	1193				11920	102443		
	743	1209				14684	112941		
	750					16429			
	768					21368			
Сумма	10571	13094	6927	30592	Сумма дисперсий	226494	751385	120750	1098630
Среднее	622	873	693	729					
Всего ремонтов	17	15	10						

Шаг 1:

Расчёт разброса данных внутри групп. Примерно подобное мы делали, когда говорили о корреляции. Помните

(x - среднее)^2	(y - среднее)^2

И величина 226494 + 751385 + 120750 = 1098630 - называется внутригрупповая дисперсия. Т.е. мы получили сумму квадратов отклонений каждого наблюдения от среднего в своей группе. Важная величина по сути своей в нашей жизни. Ещё поговорим об этом на практике.

Шаг 2:

Рассчитаем межгрупповую дисперсию – меру разброса между группами космолётов (между категориями). Для этого надо найти стандартные отклонения средних значений от общего среднего значения, умножить их на количество данных в соответствующей ….. и вычислить сумму полученных произведений. Т.е.

Кол-во случаев ремонта «Венанта» * (Ср.знач. «Венанта» - общ.ср.знач)^2 +
+ Кол-во случаев ремонта «Коврет_С90» * (Ср.знач. «Коврет_С90» - общ.ср.знач)^2 + Кол-во случаев ремонта «Дрендноут» * (Ср.знач. «Дрендноут» - общ.ср.знач)^2

17*(622 - 729)^2 + 15*(873 - 729)^2 + 10*(693 - 729)^2 = 519204
Межгрупповая дисперсия = 519204

Количественная – номинальная

Шаг 3:
Корреляционное значение Eta между факторами:

$$\frac{\text{Межгрупповая дисперсия}}{\text{Внутригрупповая дисперсия} + \text{Межггрупповая дисперсия}} = \frac{1098630}{1098630 + 519204} = 0{,}32$$

Eta = корень(0,32) = 0,56

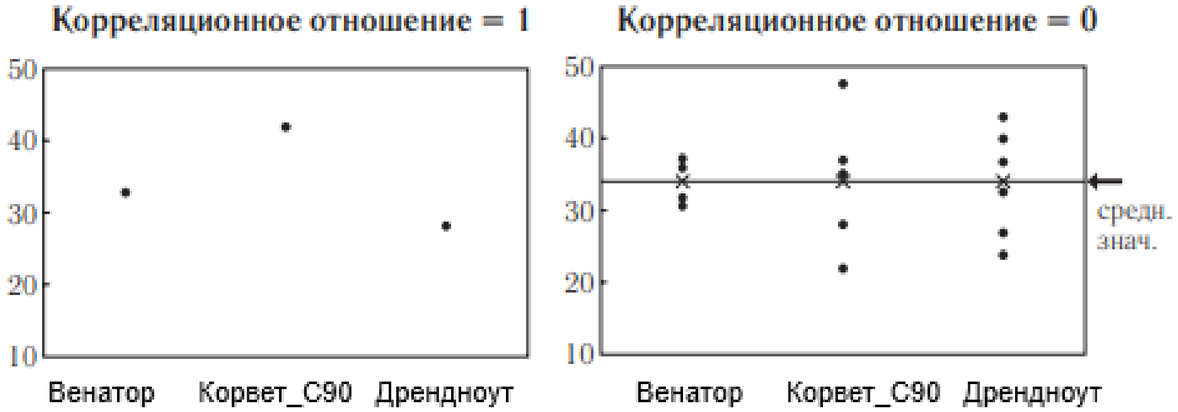
Пределы изменения данной корреляции от 0 до 1. А о силе связи принято рассуждать примерно в такой терминологии.

Корреляционное отношение		Вывод о степени взаимосвязи	Вывод о наличии взаимосвязи
1,0—0,8	⇒	Очень тесная	Есть
0,8—0,5	⇒	Достаточно тесная	
0,5—0,25	⇒	Слабая	
< 0,25	⇒	Очень слабая	Нет

Получается, что в нашем случае связь между «тип_космолёта» и «временем_ремонта» .....

Количественная – номинальная

Давайте посмотрим крайние (0 и 1) значения корреляций:



Корреляционное отношение = 1

⇕

Данные внутри каждой группы одинаковы

⇕

Внутригрупповая дисперсия = 0

Корреляционное отношение = 0

⇕

Средние значения всех групп одинаковы

⇕

Межгрупповая дисперсия = 0

$$\frac{\text{Межгрупповая дисперсия}}{\text{Внутригрупповая дисперсия} + \text{Межггрупповая дисперсия}} =$$

В разобранном подходе мы изучили метод корреляции - сравнения дисперсий.

Надо признать, что с практической стороны, этот метод корреляции не часто используется. Можно сказать совсем мало используется. Интереснее ответить на другой вопрос: два космолёта имеет разное или одинаковое время ремонта. Если разное – почему? Этот подход мы разберём в следующих встречах.

А пока, рассмотрим последний вид взаимосвязи между факторами – когда тип шкалы - номинальный (качественный) и у одного и у другого фактора

Номинальная - номинальная

Проверим, есть ли связь между «наличием_пробоин» и «типом_космолёта». Данные выглядят так:

Наличие_пробоин	Тип_космолёта
нет	Венатор
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Корвет_С90
да	Корвет_С90
нет	Венатор
да	Корвет_С90
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Корвет_С90
нет	Венатор
да	Дрендноут
нет	Венатор

А значение меры корреляции между факторами носит название – коэффициент корреляции Крамера. (V-крамера, к-т независимости). Рассчитаем его. Для этого, пройдём последовательно несколько шагов.

Шаг 1:

Составим табличку частот.
(В литературе также можно встретить название – таблица сопряжённости). Частоты – это просто подсчёт обозначенных случаев. Например, космолёт «Венатор» ремонтировался 5 раз.

		Венатор	Корвет_С90	Дрендноут	ИТОГО
	да	5	4	2	11
	нет	33	14	8	55
ИТОГО		38	18	10	66

Шаг 2:

Рассчитаем теоретические частоты. В формуле приведён расчёт для «Венатор» с пробоинами «да»

Теоретические частоты – это значения в таблице в случае, когда связи между рассматриваемыми факторами не наблюдается.

	Венатор	Корвет_С90	Дрендноут	ИТОГО
да	6	3	2	11
нет	32	15	8	55
ИТОГО	38	18	10	66

$$\frac{\text{Число пробоин(да) * кол-во космолётов в группе}}{\text{Общее число отремантироанных космолётов}}$$

$$\frac{11*38}{66} = 6$$

Номинальная – номинальная

Проверим, есть ли связь между «наличием_пробоин» и «типом_космолёта». Данные выглядят так:

Наличие_пробоин	Тип_космолёта
нет	Венатор
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Корвет_С90
да	Корвет_С90
нет	Венатор
да	Корвет_С90
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Корвет_С90
нет	Венатор
да	Дрендноут
нет	Венатор

Шаг 3:

Найдём разницу между теоретическими и эмпирическими частотами.
На картинке дана общая формула и расчёт для ячейки «Венатор» с пробоинами «да».

	Венатор	Корвет_С90	Дрендноут	ИТОГО
да	0,28	0,33	0,07	11
нет	0,06	0,07	0,01	55
ИТОГО	38	18	10	66

(эмпирическая частота - теоретическая частота)^2
Теоретическая частота

$$\frac{(5-6)^2}{6} = 0{,}28$$

Шаг 4:

Посчитать сумму всех разниц отношения частот. Данная сумма носит название «критерий согласия Пирсона или Хи-квадрат. (что это и зачем – на следующих встречах обсудим).

А пока, посчитаем.
Хи-квадрат = 0,28 + 0,33 + 0,07 + 0,06 + 0,07 + 0,01 = 0,82

Чем больше разница между факторами, тем больше значение Хи-квадрат.

Академия Яндекса

Номинальная – номинальная

Проверим, есть ли связь между «наличием_пробоин» и «типом_космолёта». Данные выглядят так:

Наличие_пробоин	Тип_космолёта
нет	Венатор
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Корвет_С90
да	Корвет_С90
нет	Венатор
да	Корвет_С90
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Дрендноут
нет	Венатор
нет	Венатор
нет	Корвет_С90
нет	Венатор
да	Дрендноут
нет	Венатор

Шаг 5:

Расчёт коэффициента корреляции Крамера

$$\sqrt{\frac{\chi_0^2}{n \times (\min\{\text{кол-во строк в таблице}; \text{кол-во столбцов в таблице}\} - 1)}},$$

, где

n – общее число строк в таблице (число ремонтов в нашем примере)

min (кол-во строк в таблице; кол-во столбцов в таблице) – надо взять меньше из двух чисел. В нашем примере – 2 (количество строк))

$$\sqrt{\frac{0{,}82}{66*(2-1)}} = 0{,}012$$

Коэффициент корреляции Крамера равен 0,012.

Много это или мало? Это очень мало. Граничные значения имеют значения 0 и 1.

Итого, вывод такой: «Тип_космолёта» и «кол-во_пробоин» не связаны между собой.

Академия Яндекса

Начало аналитической работы в питоне

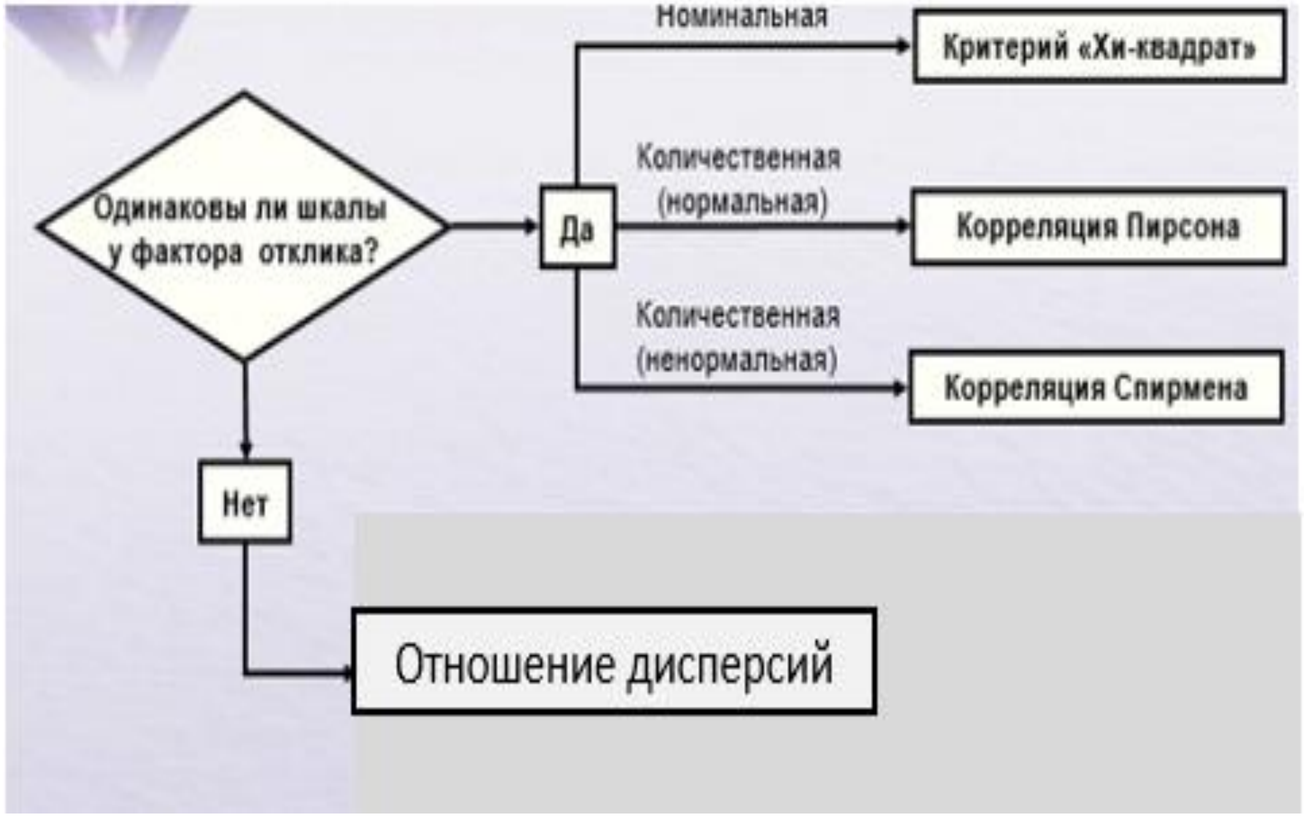
Вид функции распределения вероятностей	Значение коэффициента корреляции		Отсутствие какой-либо связи между двумя переменными величинами	Переменные величины максимально тесно связаны
	max	min		
Коэффициент линейной корреляции	−1	1	0	−1 или 1
Корреляционное отношение	0	1	0	1
Коэффициент корреляции Крамера	0	1	0	1

К-т линейной корреляции – для поиска взаимосвязи между количественными данными.

Корреляционное отношение дисперсий Eta - для поиска взаимосвязи между данными в разных шкалах.

К-т корреляции Крамера - для поиска взаимосвязи между данными в качественными (номинальными) данными.

Взаимосвязь факторов

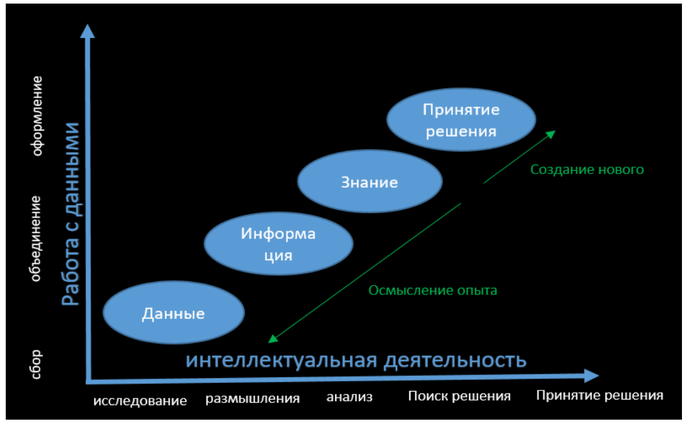


Взаимосвязь факторов

ВАЖНО:

Мы начали встречу с обсуждения причины и следствия. Нам важно понимать, что корреляция НЕ даёт понимание что на что влияет – т.е. что есть причина, а что следствие.

Мы видим только связь в цифре и в графике.
А вот как разбираться с причинами и следствием – это сам аналитик решает из контекста своей области деятельности.



Что понадобится для выполнения задач

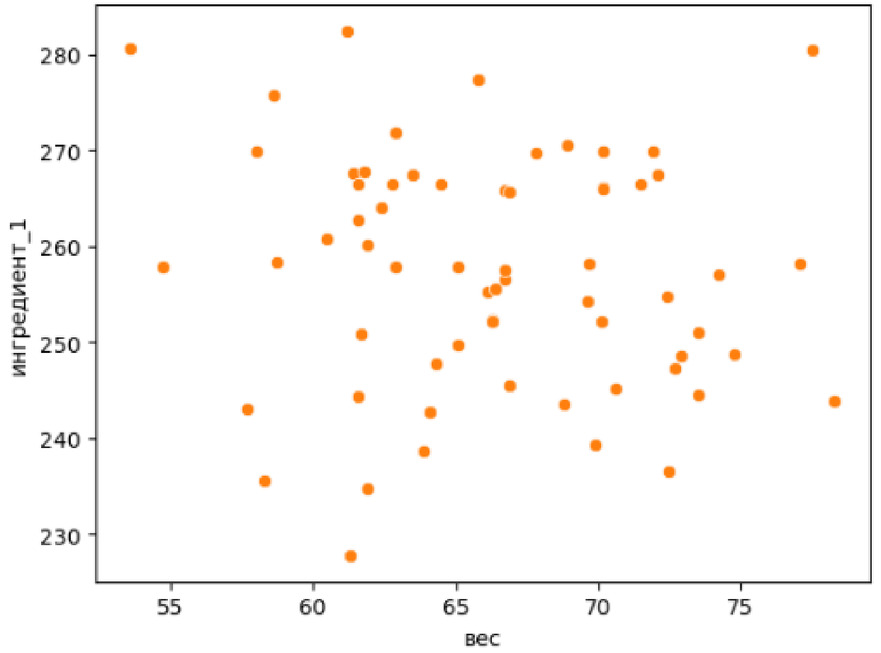
1. Корреляция количественная-количественная corr(*метод корреляции Спирмена или Пирсона*)
2. Метод сортировки sort()

3. Корреляция количественная-номинальная: запрограммировать формулу «отношение дисперсий»

4. Корреляция номинальная-номинальная: запрограммировать формулу «к-т Крамера»
   - Создать таблицу сопряжённости: pd.crosstab(x, y)
   - Расчёт Хи-квадрат: scipy.stats.chi2_contingency()

5. График скатерплот.
   - Использовать библиотеку сиборн.
   - «\n» перенос текстовой строки в заголовке

```
1 sns.scatterplot(data = df, x = 'вес',  y = 'ингредиент_1')
2 plt.show()
```

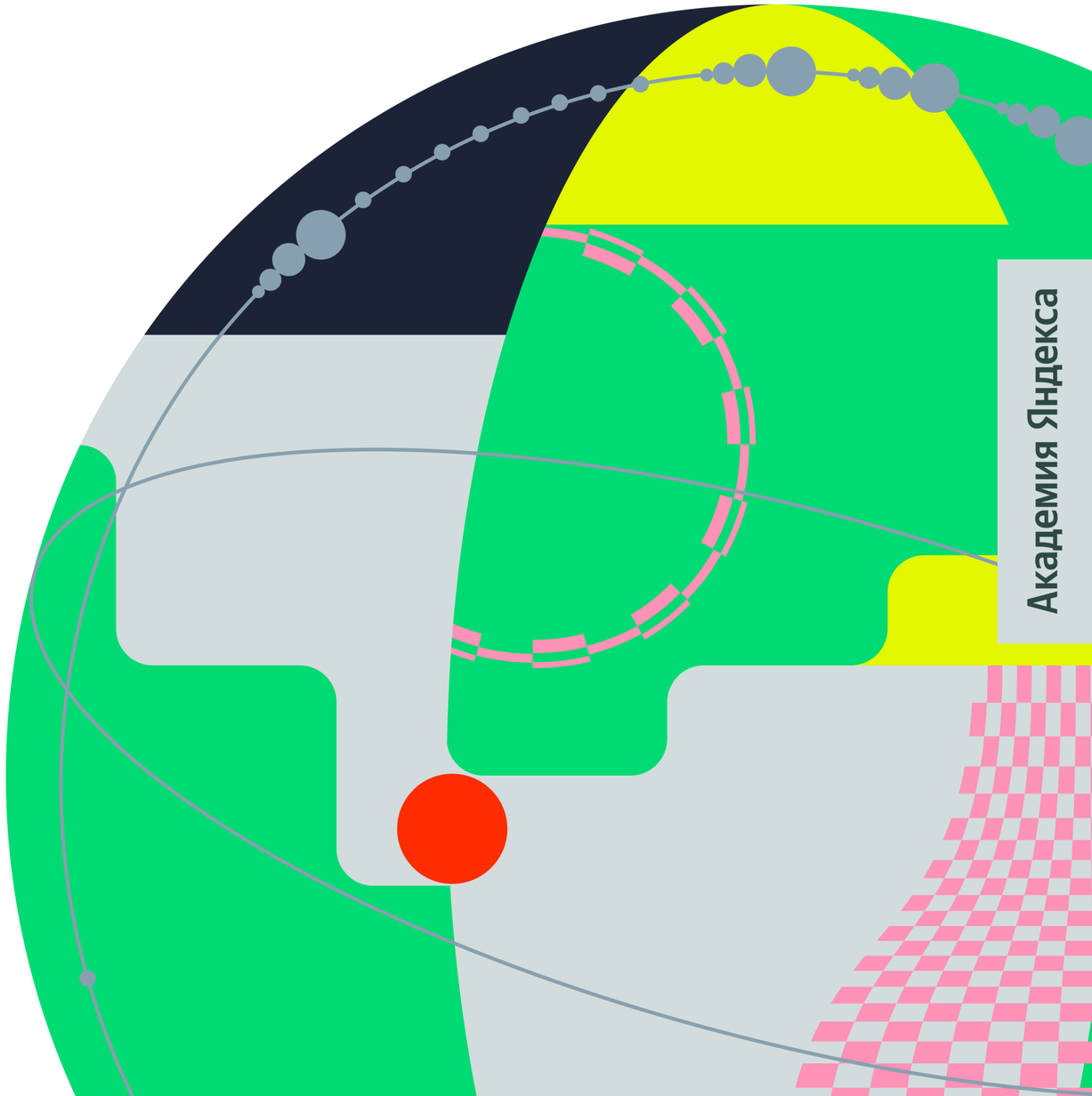


Дополнительный материал

МАТЕМАТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ЭКСПЕРТНЫХ ОЦЕНОК

Анализ экспертных оценок: https://medstatistic.ru/articles/Chegodaev_ekspertnye_ocenki.pdf

Спасибо за внимание



Академия Яндекса позволяет школьникам и студентам освоить востребованные IT-профессии по программам, разработанным экспертами компании