

Проведение статистического исследования в Excel

Александра Калинина
PhD Management
CMO at IPification



Проверка связи





Если у вас нет звука:

- убедитесь, что на вашем устройстве и на колонках включён звук
- обновите страницу вебинара (или закройте страницу и заново присоединитесь к вебинару)
- откройте вебинар в другом браузере
- перезагрузите компьютер (ноутбук) и заново попытайтесь зайти



Поставьте в чат:

-  если меня видно и слышно
-  если нет

Правила участия

- 1 Приготовьте блокнот и ручку, чтобы записывать важные мысли и идеи
- 2 Продолжительность вебинара – 2 часа
- 3 Вы можете писать свои вопросы в чате
- 4 После каждого этапа работы мы уделим время ответам на вопросы
- 5 Финальный датасет со всеми вычислениями есть в личном кабинете



Александра Калинина

О спикере:

- Директор по маркетингу IPification, Hong Kong
- PhD Management



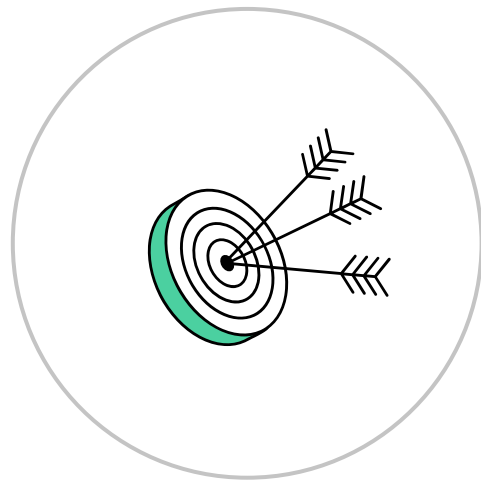
Вспоминаем прошлые занятия

- Что такое статистика и для чего она нужна?
- Основные статистические показатели.
Виды распределений
- Регрессионный анализ и меры связи.
Исследование данных
- Работа со статистическими гипотезами и
основы АВ-тестирования



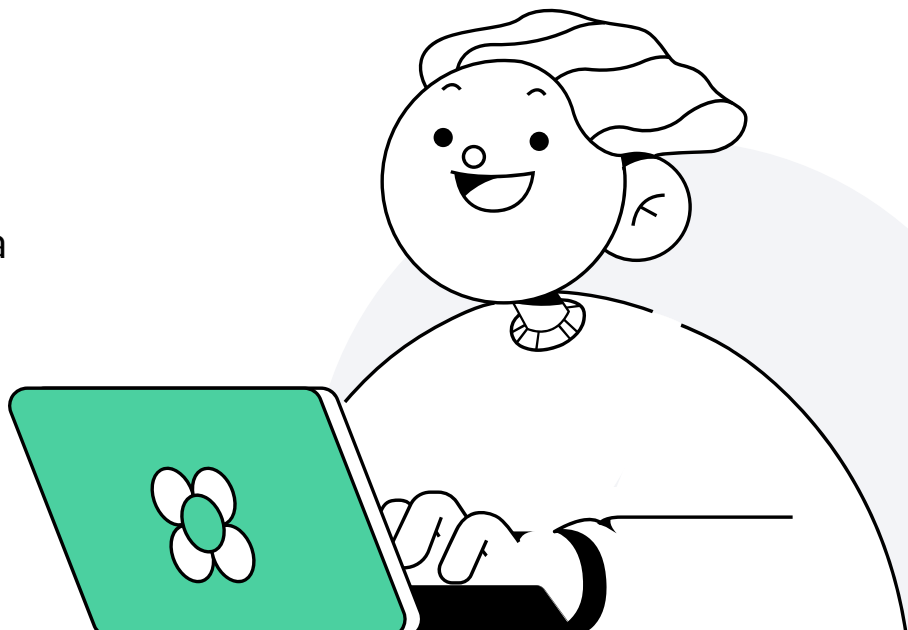
Цели занятия

- Вычислить описательную статистику, используя инструменты Excel
- Применить инструменты Excel для поиска и обработки выбросов в данных
- Определить характер взаимосвязи между переменными по визуализации
- Рассчитать коэффициент корреляции, используя инструменты Excel
- Провести регрессионный анализ в Excel и исследовать характеристики точности модели
- Построить точечные и интервальные прогнозы результирующей переменной



План занятия

- 1 Вычисление описательной статистики
- 2 Поиск выбросов в данных
- 3 Формулирование гипотез
- 4 Расчёт корреляции
- 5 Проведение регрессионного анализа



Вычисление описательной статистики



1

Данные

Наша задача: предсказать цену апартаментов в США по доходу, размеру квартиры и году постройки. Первый лист датасета — данные, с которыми мы работаем

X_1 X_2 X_3 X_4 X_5 Y

	A	B	C	D	E	F	G
1	Средний доход в районе	Средний возраст зданий в районе	Среднее количество комнат в районе	Среднее количество спален в районе	Количество жителей района	Цена	Адрес
2	\$ 79 545,5	5,7	7,0	4,1	23086,8	\$ 1 059 033,6	208 Michael Ferry Apt, 674Laurabury, NE 37010-5101
3	\$ 79 248,6	6,0	6,7	3,1	40173,1	\$ 1 505 890,9	188 Johnson Views Suite 079Lake Kathleen, CA 48958
4	\$ 61 287,1	5,9	8,5	5,1	36882,2	\$ 1 058 988,0	9127 Elizabeth StravenueDanielstown, WI 06482-3489
5	\$ 63 345,2	7,2	5,6	3,3	34310,2	\$ 1 260 616,8	USS BarnettFPO AP 44820
6	\$ 59 982,2	5,0	7,8	4,2	26354,1	\$ 630 943,5	USNS RaymondFPO AE 09386
7	\$ 80 175,8	5,0	6,1	4,0	26748,4	\$ 1 068 138,1	06039 Jennifer Islands Apt, 443Tracyport, KS 16077
8	\$ 64 698,5	6,0	8,1	3,4	60828,2	\$ 1 502 055,8	4759 Daniel Shoals Suite 442Nguyenburgh, CO 20247
9	\$ 78 394,3	7,0	6,6	2,4	36516,4	\$ 1 573 936,6	972 Joyce ViaductLake William, TN 17778-6483
10	\$ 59 927,7	5,4	6,4	2,3	29387,4	\$ 798 869,5	USS GilbertFPO AA 20957
11	\$ 81 885,9	4,4	8,2	6,1	40150,0	\$ 1 545 154,8	Unit 9446 Box 0958DPO AE 97025
12	\$ 80 527,5	8,1	5,0	4,1	47224,4	\$ 1 707 045,7	6368 John Motorway Suite 700Janetbury, NM 26854
13	\$ 50 593,7	4,5	7,5	4,5	34344,0	\$ 663 732,4	911 Castillo Park Apt, 717Davisborough, PW 78603
14	\$ 39 033,8	7,7	7,3	3,1	39220,4	\$ 1 042 814,1	209 Natasha Stream Suite 961Huffmanland, NE 52457
15	\$ 73 163,7	6,9	6,0	2,3	32326,1	\$ 1 291 331,5	829 Welch Track Apt, 992North John, AR 26532-5136
16	\$ 69 391,4	5,3	8,4	4,4	35521,3	\$ 1 402 818,2	PSC 5330, Box 4420APO AP 08302
17	\$ 73 091,9	5,4	8,5	4,0	23929,5	\$ 1 306 674,7	2278 Shannon ViewNorth Carriemouth, NM 84617
18	\$ 79 707,0	5,1	8,2	3,1	39717,8	\$ 1 556 786,6	064 Hayley UnionsNicholsborough, HI 44161-1887
19	\$ 61 929,1	4,8	5,1	4,3	24595,9	\$ 528 485,2	5498 Rachel LocksNew Gregoryshire, PW 54755
20	\$ 63 508,2	5,9	7,2	5,1	35719,7	\$ 1 019 425,9	Unit 7424 Box 2786DPO AE 71255

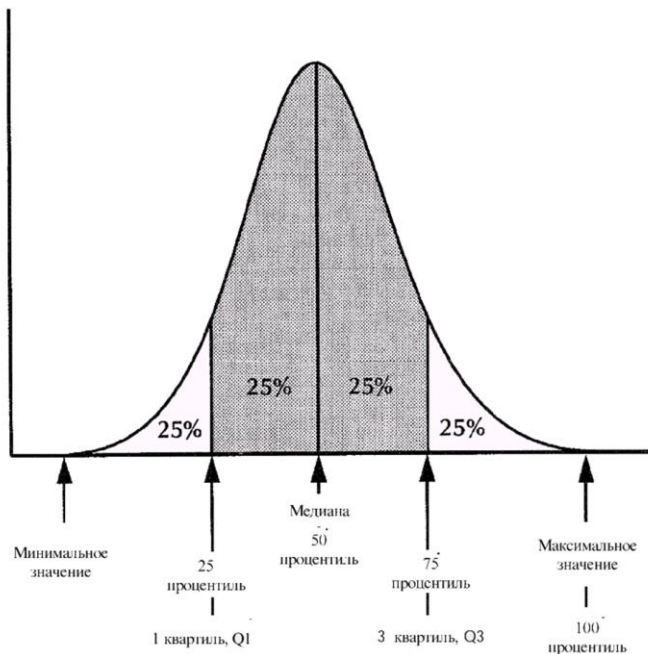


**Какие параметры
описательной статистики
будем вычислять?**

Описательная статистика: что вычисляем?

Показатели положения, которые описывают положение данных (или середины совокупности) на числовой оси:

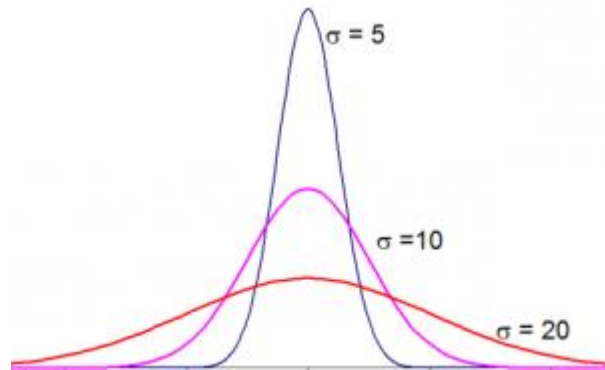
- Минимальный и максимальный элементы выборки
- Выборочные верхний и нижний квартили
- Выборочное среднее
- Выборочная медиана
- Выборочная мода



Описательная статистика: что вычисляем?

Показатели разброса, которые описывают степень разброса данных относительно своего центра (насколько кучно основная масса данных группируется около середины совокупности):

- Дисперсия выборки
- Выборочное среднее квадратическое отклонение (СКО, стандартное отклонение)
- Размах
- Минимум
- Максимум
- Коэффициент эксцесса



Описательная статистика: что вычисляем?

Показатели асимметрии, которые описывают симметричность распределения данных около своего центра:

- Коэффициент асимметрии
- Положение выборочной медианы относительно выборочного среднего и относительно выборочных квартилей
- Гистограмма

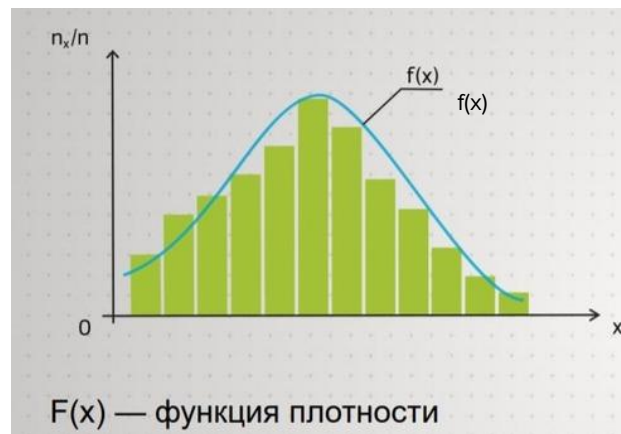


Описательная статистика: что вычисляем?

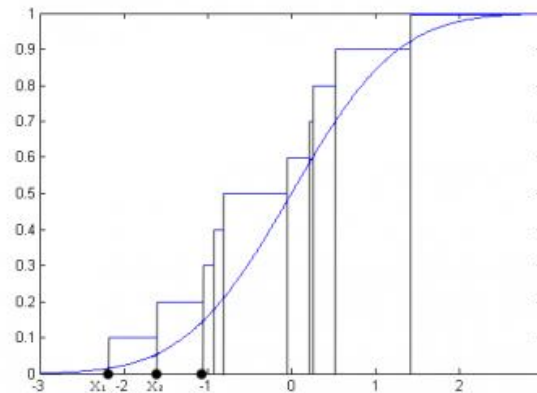
Показатели, описывающие закон распределения. Они дают представление о законе распределения данных:

- Гистограмма
- Выборочная функция распределения
- Таблица частот

Гистограмма



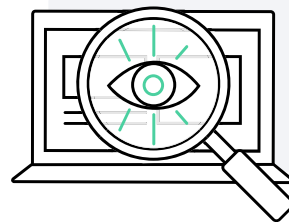
Выборочная функция распределения



Практика

Рассчитаем описательную статистику в Excel

Построим гистограмму, кумуляту и ящик с усами в Excel



Описательная статистика: как вычисляем?

Во вкладке «Данные» выбираем опцию «Анализ данных». Во всплывающем окне будет доступен выбор инструментов анализа

The screenshot shows the Microsoft Excel interface with the 'Данные' (Data) tab selected in the ribbon. The 'Анализ данных' (Data Analysis) button is highlighted with a red rectangle. Below the ribbon, the 'Анализ данных' (Data Analysis) task pane is open, displaying a list of analysis tools. The 'Описательная статистика' (Descriptive Statistics) option is selected and highlighted in blue. The background shows a spreadsheet with data for average income and age in a district.

	A	B
1	Средний доход в районе	Средний возраст зданий в районе
2	\$ 17 796,6	4,9
3	\$ 35 454.7	6.9

Анализ данных

Инструменты анализа

- Двухфакторный дисперсионный анализ без повторений
- Корреляция
- Ковариация
- Описательная статистика**
- Экспоненциальное сглаживание
- Двухвыборочный F-тест для дисперсии
- Анализ Фурье
- Гистограмма
- Скользящее среднее
- Генерация случайных чисел

ОК Отмена Справка

Описательная статистика: как вычисляем?

Выбираем инструмент «Описательная статистика» и настраиваем входной интервал

The screenshot shows the Microsoft Excel interface with the 'Данные' (Data) ribbon selected. The 'Анализ данных' (Data Analysis) group contains the 'Описательная статистика' (Descriptive Statistics) tool. The dialog box for 'Описательная статистика' is open, showing the following settings:

- Входные данные** (Input data):
 - Входной интервал:
 - Группирование: ☒ по столбцам, ☐ по строкам
 - ☐ Метки в первой строке
- Параметры вывода** (Output options):
 - ☐ Выходной интервал:
 - ☒ Новый рабочий лист:
 - ☐ Новая рабочая книга
 - ☒ Итоговая статистика
 - ☐ Уровень надежности: %
 - ☒ К-ый наименьший:
 - ☒ К-ый наибольший:

The background shows a spreadsheet with columns labeled 'Средний доход' (Average income) and 'Средний возраст' (Average age). The data rows show values for these metrics across 13 rows.

Описательная статистика: как вычисляем?

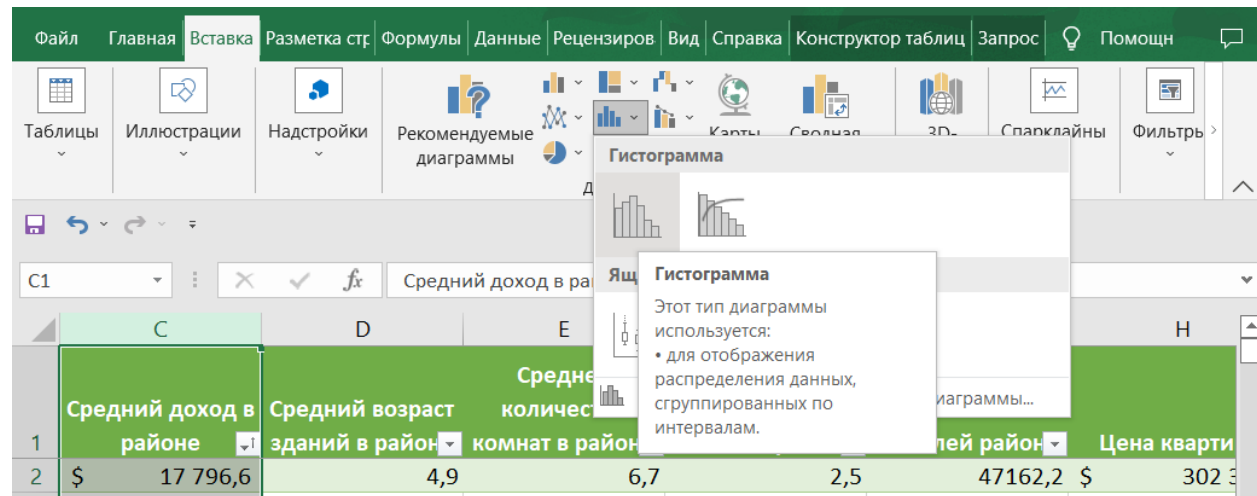
Получаем результат на новом листе:

		X_1	X_2	X_3	X_4	X_5	Y
	A	B	C	D	E	F	G
1		Средний доход в районе	Средний возраст зданий в районе	Среднее количество комнат в районе	Среднее количество спален в районе	Количество жителей района	Цена квартиры
2							
3	Среднее	\$ 68 583,1	6,0	7,0	4,0	36 163,5	\$ 1 232 072,7
4	Стандартная ошибка s/\sqrt{n}	\$ 150,7	0,0	0,0	0,0	140,4	\$ 4 993,8
5	Медиана	\$ 68 804,3	6,0	7,0	4,1	36 199,4	\$ 1 232 669,4
6	Мода	#Н/Д	#Н/Д	#Н/Д	4,4	#Н/Д	#Н/Д
7	Стандартное отклонение выборки s	\$ 10 658,0	1,0	1,0	1,2	9 925,7	\$ 353 117,6
8	Дисперсия выборки s^2	113 592 776,7	1,0	1,0	1,5	98 518 530,2	124 692 058 202,2
9	Экцесс	0,0	-0,1	-0,1	-0,7	-0,0	-0,1
10	Асимметричность	-0,0	-0,0	-0,0	0,4	0,1	-0,0
11	Интервал (простой размах)	\$ 89 905,1	6,9	7,5	4,5	69 449,1	\$ 2 453 126,9
12	Минимум	\$ 17 796,6	2,6	3,2	2,0	172,6	\$ 15 938,7
13	Максимум	\$ 107 701,7	9,5	10,8	6,5	69 621,7	\$ 2 469 065,6
14	Сумма	\$ 342 915 544,9	29 886,1	34 939,0	19 906,7	180 817 580,2	\$ 6 160 363 270,7
15	Счет n	5 000	5 000	5 000	5 000	5 000	5 000
16	Наибольший(1250) 75 персентиль Q_3	\$ 75 786,3	6,7	7,7	4,5	42 873,1	\$ 1 471 746,6
17	Наименьший(1250) 25 персентиль Q_1	\$ 61 478,6	5,3	6,3	3,1	29 403,5	\$ 997 452,5

Описательная статистика: гистограмма

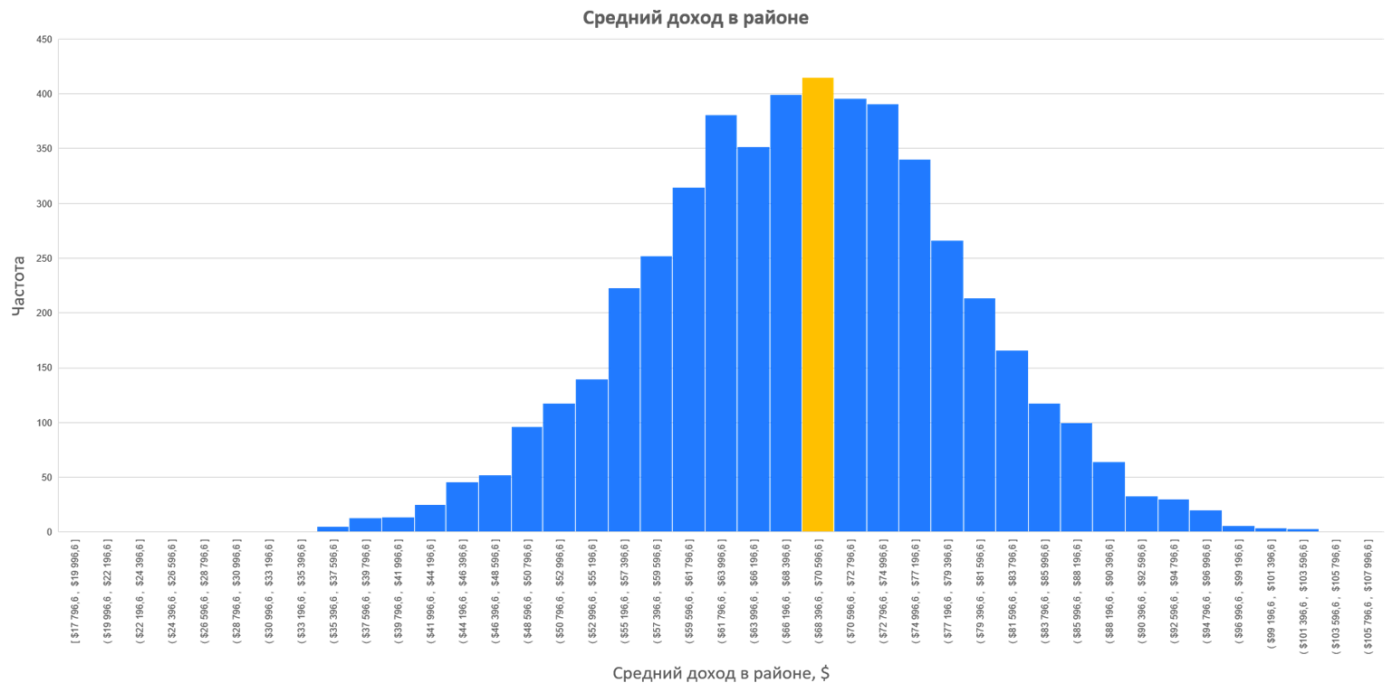
Строим столбчатую диаграмму, которая показывает частоту повторяемости значений. Гистограмма позволяет:

- наглядно представить тенденции изменения измеряемых параметров и зрительно оценить закон их распределения
- быстро определить центр, разброс и форму распределения



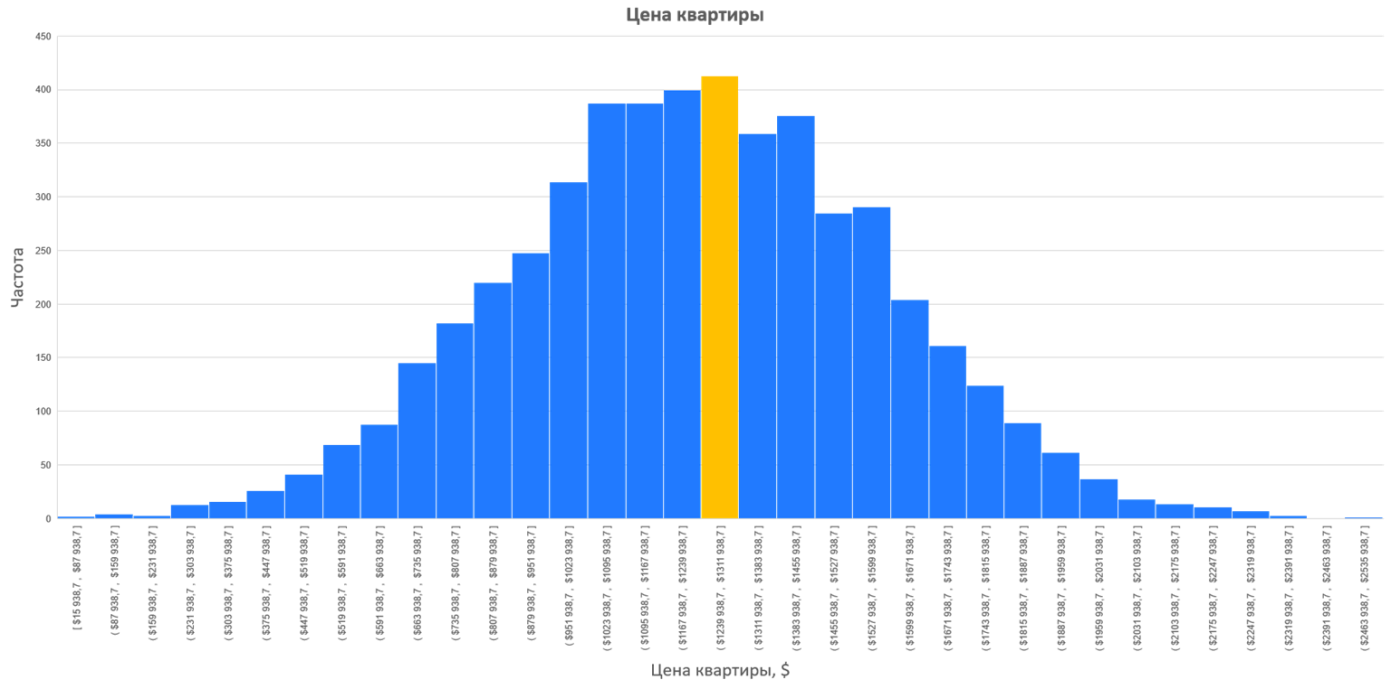
Описательная статистика: гистограмма

На гистограмме жёлтым цветом обозначен модальный интервал — интервал, частота которого максимальна относительно других интервалов



Описательная статистика: гистограмма

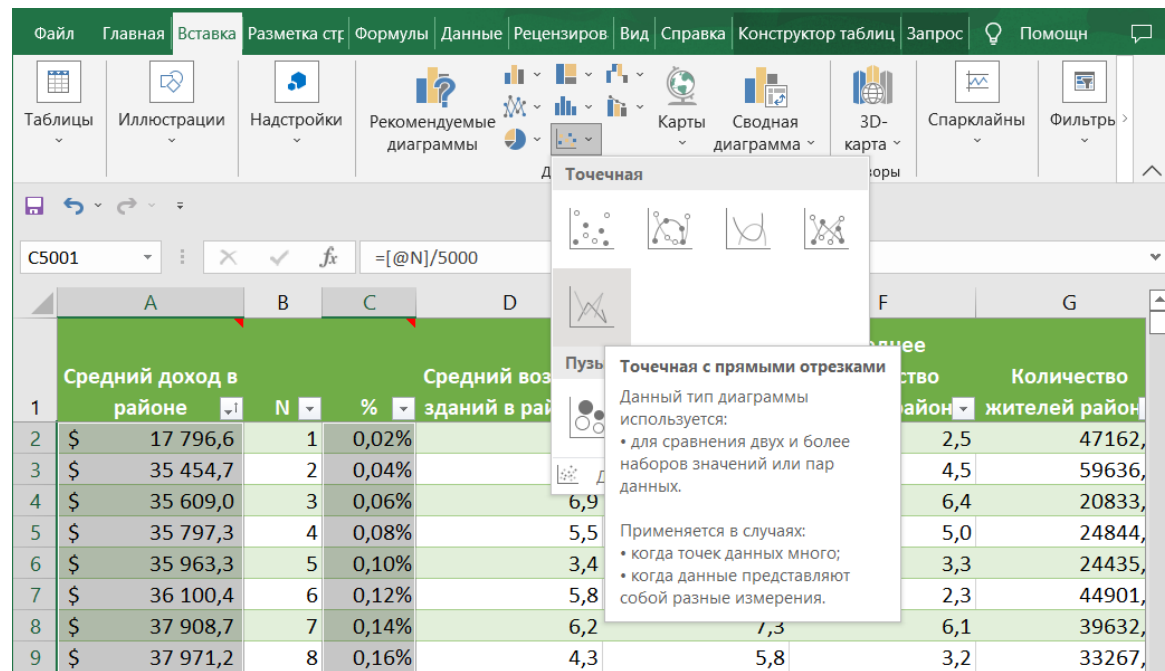
На гистограмме жёлтым цветом обозначен модальный интервал — интервал, частота которого максимальна относительно других интервалов



Описательная статистика: кумулята

Строим кумуляту, график накопленных относительных частот.

Кумулята — это экспериментальная оценка формы графика функции распределения



Описательная статистика: кумулята

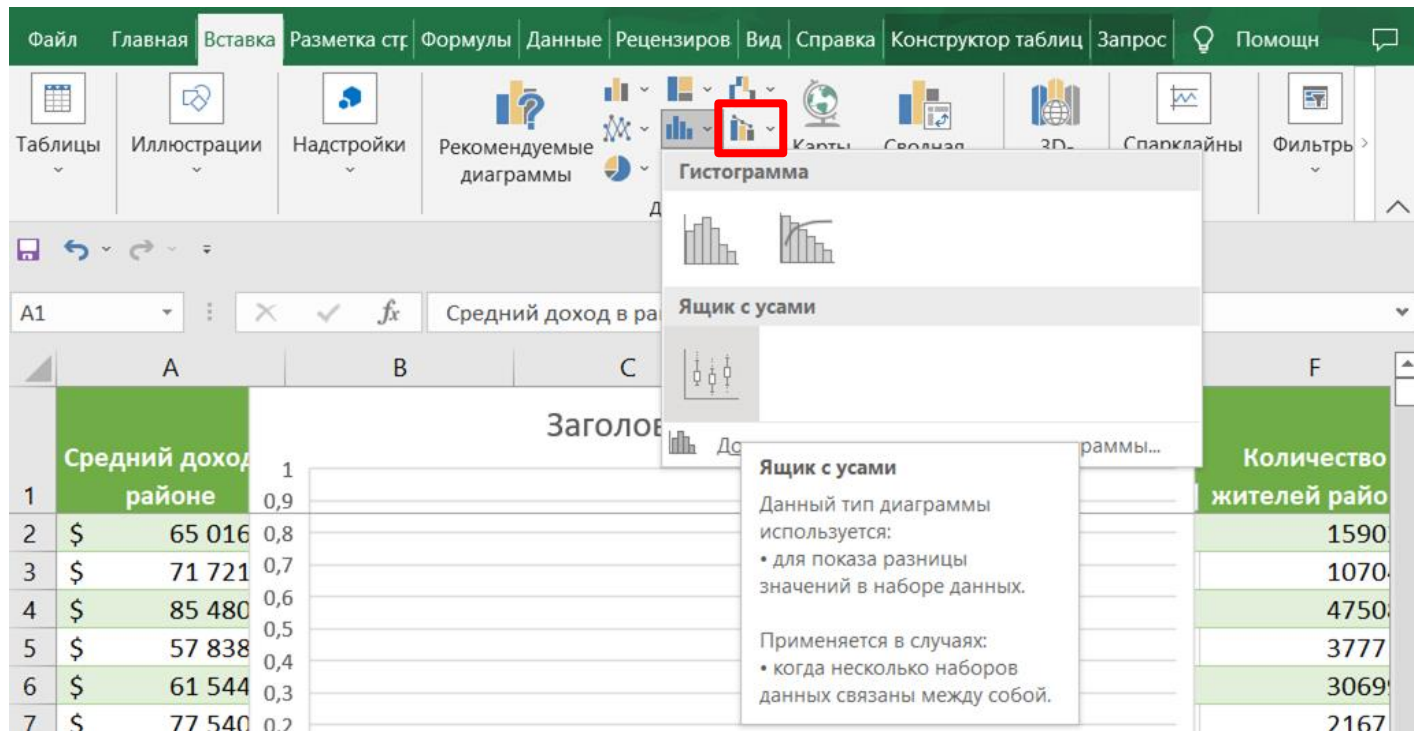


Описательная статистика: кумулята



Ящик с усами

Найдём инструмент «Ящик с усами» во вкладке «Вставка»



Гистограмма

Ящик с усами

Ящик с усами

Данный тип диаграммы используется:

- для показа разницы значений в наборе данных.

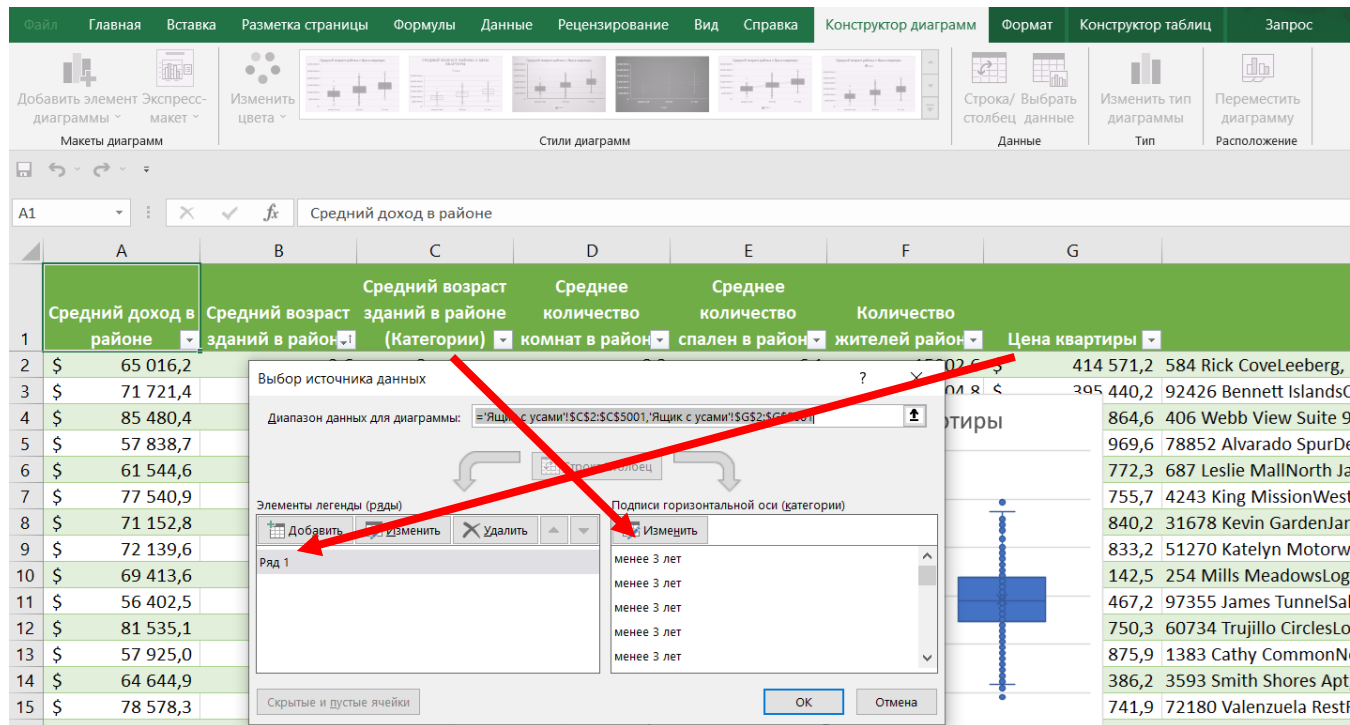
Применяется в случаях:

- когда несколько наборов данных связаны между собой.

	Средний доход в районе	Количество жителей района
1	0,9	1590
2	0,8	1070
3	0,7	4750
4	0,6	3777
5	0,5	3069
6	0,4	2167
7	0,3	
8	0,2	

Ящик с усами

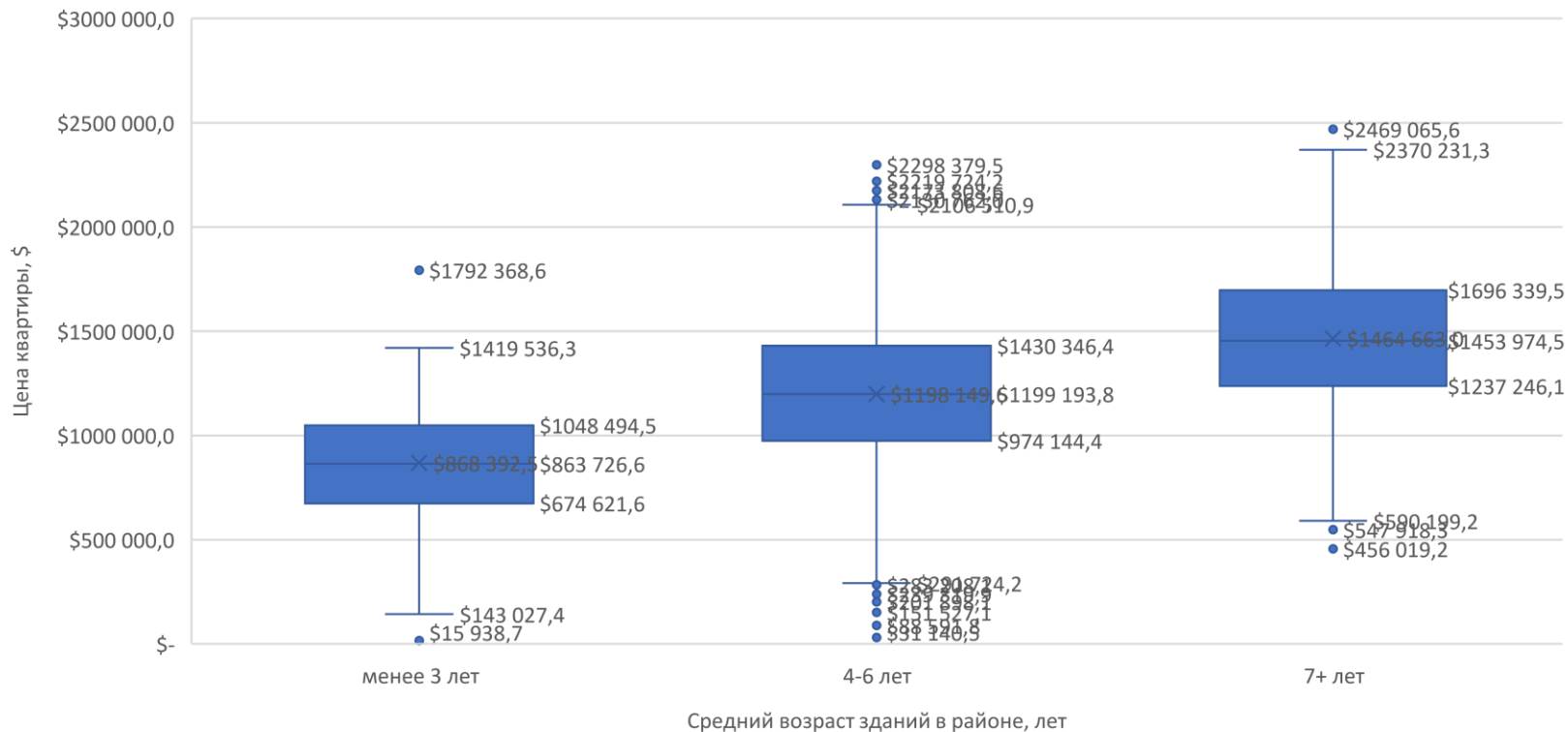
Настроим элементы легенды и подписи к горизонтальной оси



Ящик с усами

Получаем результат:

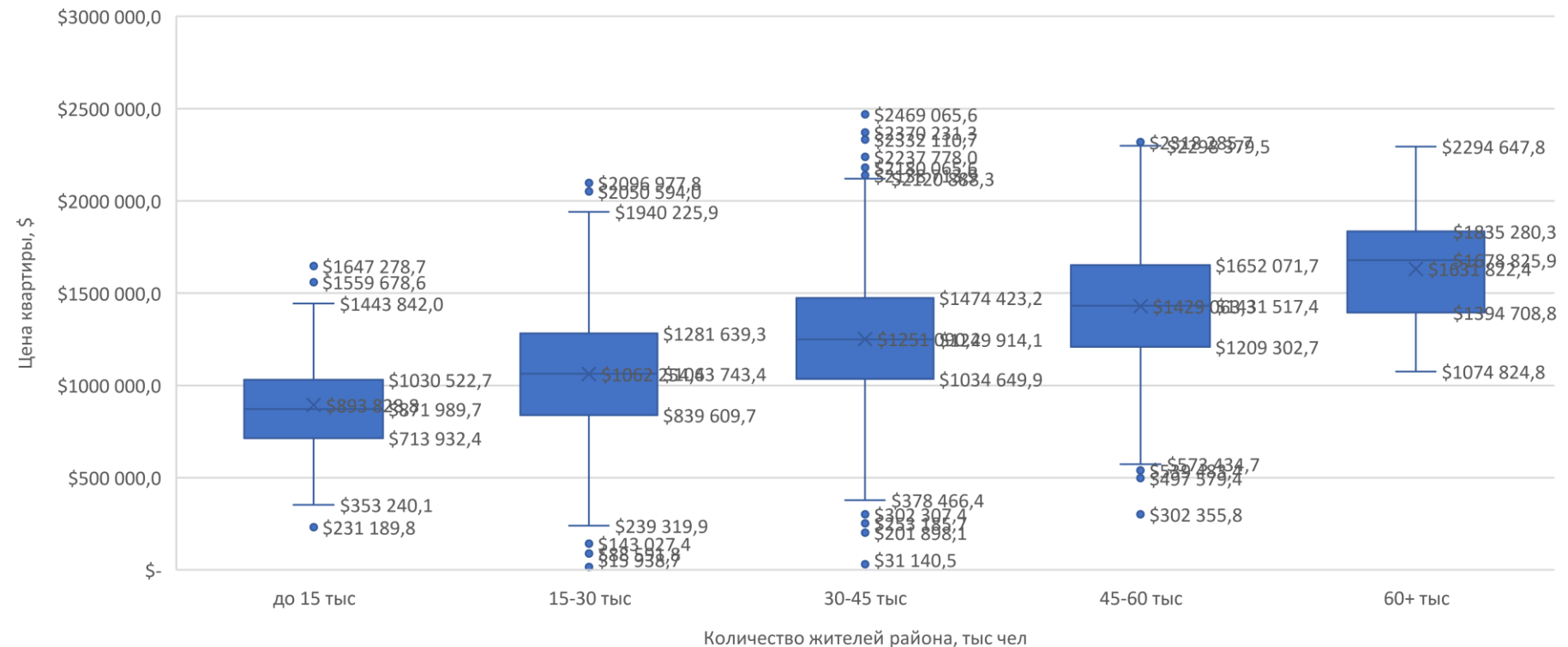
Средний возраст района x Цена квартиры



Ящик с усами

Получаем результат:

Количество жителей района x Цена квартиры





Ваши вопросы?

Поиск выбросов в данных



2

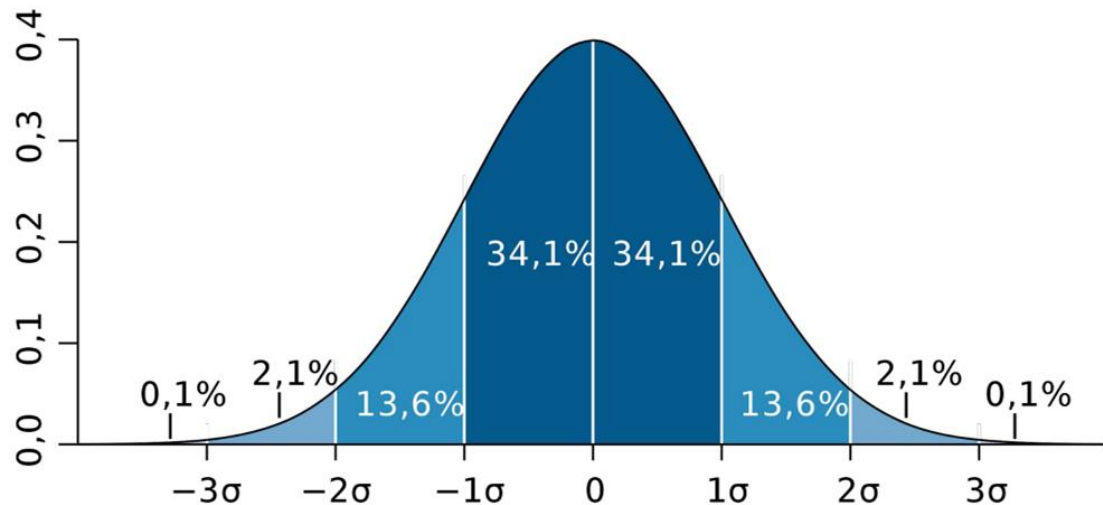


**Какими двумя способами
можно найти выбросы
в данных?**

Выбросы: правило трёх сигм

Правило трёх сигм:

Если данные распределены нормально, то всё, что расположено за пределами расстояния 3 сигм (стандартного отклонения) можно считать выбросами



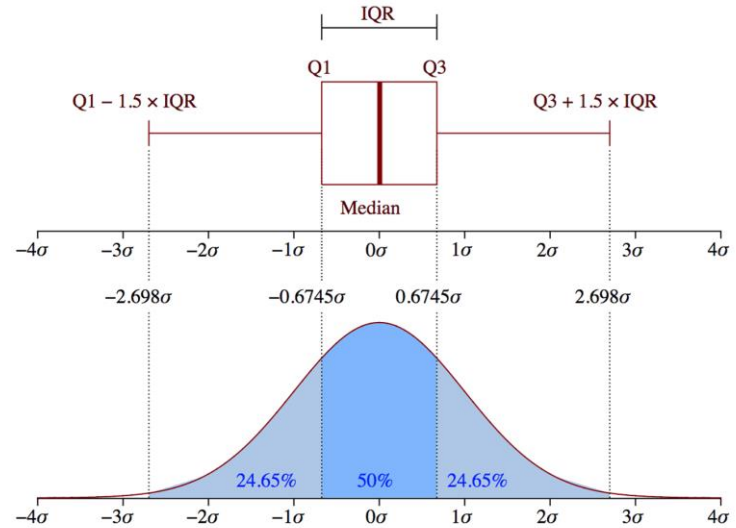
Выбросы: «заборы Тьюки»

Этот метод обнаружения выбросов основан на межквартильном размахе.

Всё, что не попадает в заданные диапазоны, считаем выбросом:

Lower = $Q1 - 1.5 \text{ IQR}$

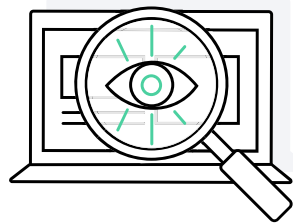
Higher = $Q3 + 1.5 \text{ IQR}$



Практика

Применим правило трёх сигм в Excel

Определим выбросы методом «заборы Тьюки» в Excel



Выбросы: три сигмы vs «заборы Тьюки»

Полученные разными способами результаты будут немного отличаться:

<i>Выбросы. Правило 3х сигм</i>	Средний доход в районе	Цена квартиры
Среднее	\$ 68,583.1	\$ 1,232,072.7
Стандартное отклонение выборки s	\$ 10,658.0	\$ 353,117.6
-3 сигмы	\$ 36,609.1	\$ 172,719.8
+3 сигмы	\$ 100,557.1	\$ 2,291,425.5
<i>Выбросы. Метод «заборы Тьюки»</i>	Средний доход в районе	Цена квартиры
Q1	\$ 61,478.6	\$ 997,452.5
Q3	\$ 75,786.3	\$ 1,471,746.6
Межквартильный размах	\$ 14,307.7	\$ 474,294.1
lower = Q1 - 1.5 IQR	\$ 40,017.1	\$ 286,011.4
higher = Q3 + 1.5 IQR	\$ 97,247.9	\$ 2,183,187.6

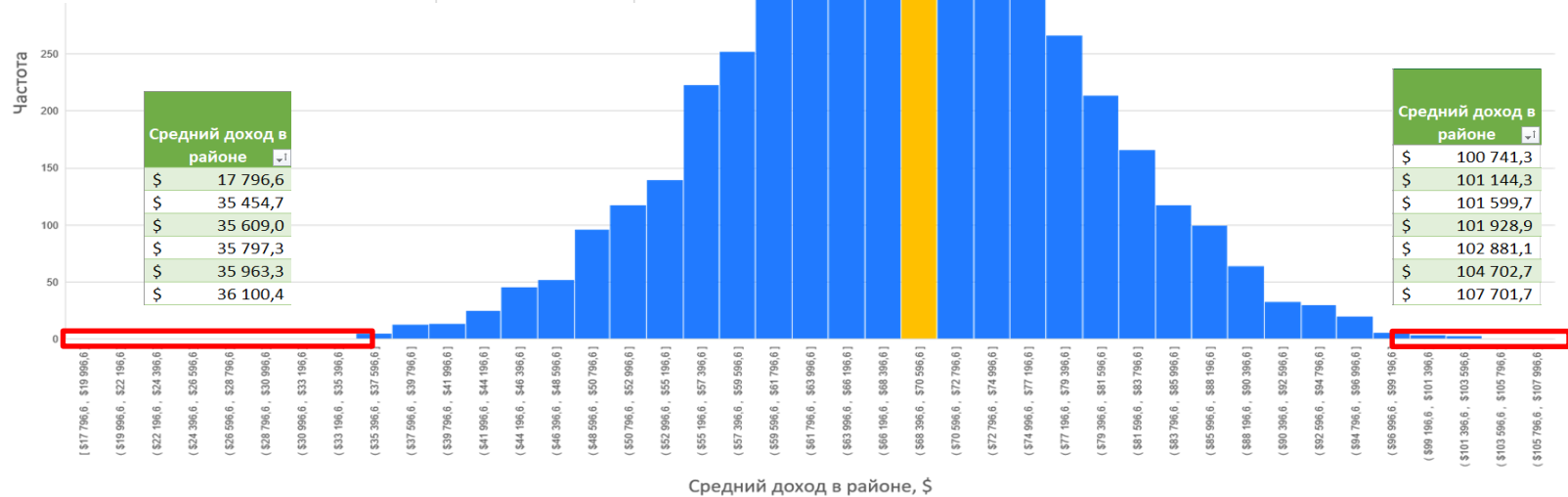
Выбросы: гистограмма

Всё, что расположено за пределами расстояния 3 сигм можно считать выбросами

Выбросы. Правило 3х сигм	Средний доход в районе
Среднее	\$ 68 583,1
Стандартное отклонение выборки s	\$ 10 658,0
-3 сигмы	\$ 36 609,1
+3 сигмы	\$ 100 557,1



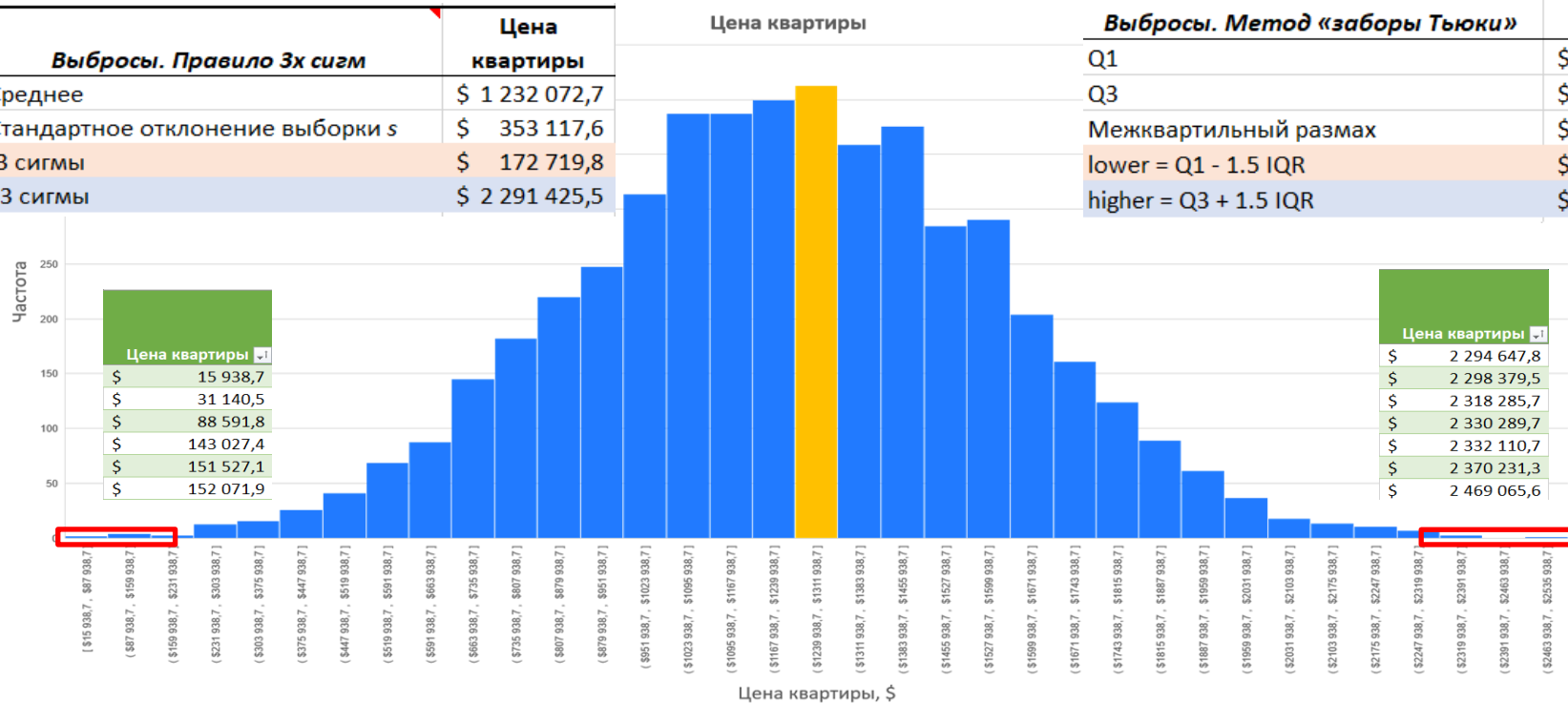
Выбросы. Метод «заборы Тьюки»	Средний доход в районе
Q1	\$ 61 478,6
Q3	\$ 75 786,3
Межквартильный размах	\$ 14 307,7
lower = Q1 - 1.5 IQR	\$ 40 017,1
higher = Q3 + 1.5 IQR	\$ 97 247,9



Выбросы: гистограмма

Всё, что расположено за пределами расстояния 3 сигм можно считать выбросами _____

Выбросы. Правило 3х сигм	Цена квартиры
Среднее	\$ 1 232 072,7
Стандартное отклонение выборки s	\$ 353 117,6
-3 сигмы	\$ 172 719,8
+3 сигмы	\$ 2 291 425,5



Выбросы. Метод «заборы Тьюки»	Цена квартиры
Q1	\$ 997 452,5
Q3	\$ 1 471 746,6
Межквартильный размах	\$ 474 294,1
lower = Q1 - 1.5 IQR	\$ 286 011,4
higher = Q3 + 1.5 IQR	\$ 2 183 187,6



Ваши вопросы?

Перерыв



Формулирование гипотез



3

Что хотим выяснить?

- Какой фактор (какие факторы) влияет (влияют) на цену квартиры?

	A	B	C	D	E	F
1	Средний доход в районе	Средний возраст зданий в районе	Среднее количество комнат в районе	Среднее количество спален в районе	Количество жителей района	Цена
	X_1	X_2	X_3	X_4	X_5	Y

Гипотезы

Гипотеза N°1:

H_1 : Чем выше средний доход в районе (X_1), тем выше цена квартиры (Y)

Это наша тестируемая гипотеза. На языке статистики она называется альтернативной гипотезой H_1

Нулевая гипотеза H_0 предполагает обратное: либо чем выше доход, тем ниже цена квартиры, либо вообще нет какого-либо значимого влияния среднего дохода на цену квартиры

Гипотезы

Гипотеза N°2:

H_2 : Чем больше комнат (X_3), тем выше цена квартиры (Y)



Гипотеза N°3, N°4...
Ваши варианты?



Ваши вопросы?

Расчёт корреляции



4

Корреляция Пирсона

Это параметрический показатель линейной корреляции

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

где \bar{X} и \bar{Y} — это средние по выборкам

Коэффициент корреляции Пирсона помогает нормировать значение ковариации, поделив её на произведение среднеквадратических отклонений случайных величин



**Что показывает
коэффициент
корреляции?**

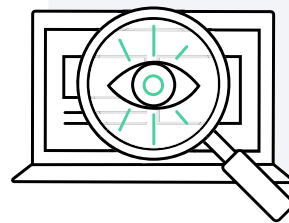
Как измерить корреляцию

- Коэффициент корреляции показывает величину взаимосвязи двух переменных между собой
- Варьируется от -1 (отрицательная корреляция) до +1 (положительная корреляция)

Величина коэффициента корреляции	0,1–0,3	0,3–0,5	0,5–0,7	0,7–0,9	0,9–1.0
Характеристика силы связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

Практика

Рассчитаем корреляцию в Excel



Корреляционная таблица

Выбираем инструмент «Корреляция» и настраиваем входной интервал

The screenshot shows the Microsoft Excel interface with the 'Данные' (Data) ribbon selected. The 'Анализ данных' (Data Analysis) group contains the 'Корреляция' (Correlation) tool. The 'Корреляция' dialog box is open, showing the following settings:

- Входные данные** (Input data):
 - Входной интервал:
 - Группирование: ☒ по столбцам, ☐ по строкам
 - ☐ Метки в первой строке
- Параметры вывода** (Output options):
 - ☐ Выходной интервал:
 - ☒ Новый рабочий лист:
 - ☐ Новая рабочая книга

The background spreadsheet shows a table with the following data:

	Средний доход в районе	Средний возраст зданий в районе
1		
2	\$ 47 320,7	3,6
3	\$ 37 971,2	4,3
4	\$ 60 167,7	4,6

Корреляционная таблица

Получаем результат на новом листе:

	A	B	C	D	E	F	G
1		<i>Средний доход в районе</i>	<i>Средний возраст зданий в районе</i>	<i>Среднее количество комнат в районе</i>	<i>Среднее количество спален в районе</i>	<i>Количество жителей района</i>	<i>Цена квартиры</i>
2	Средний доход в районе	1,00					
3	Средний возраст зданий в районе	0,00	1,00				
4	Среднее количество комнат в районе	-0,01	-0,01	1,00			
5	Среднее количество спален в районе	0,02	0,01	0,46	1,00		
6	Количество жителей района	-0,02	-0,02	0,00	-0,02	1,00	
7	Цена квартиры	0,64	0,45	0,34	0,17	0,41	1,00



Ваши вопросы?

Проведение регрессионного анализа



5



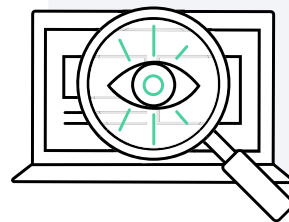
**Зачем нужен
регрессионный анализ?**

Основные задачи регрессионного анализа

- Определение вида и формы зависимости
- Оценка параметров уравнения регрессии
- Проверка значимости уравнения регрессии
- Проверка значимости отдельных коэффициентов уравнения
- Построение интервальных оценок коэффициентов
- Исследование характеристик точности модели
- Построение точечных и интервальных прогнозов результирующей переменной

Практика

Проведём регрессионный анализ инструментами Excel



Определение вида и формы зависимости

Найдём инструмент «Точечная диаграмма» во вкладке «Вставка»

The screenshot shows the Microsoft Excel interface with the 'Вставка' (Insert) ribbon active. The 'Точечная' (Scatter) chart type is selected from the 'Рекомендуемые диаграммы' (Recommended Charts) group. A tooltip is displayed over the 'Точечная' icon, providing details about its usage.

Точечная

Данный тип диаграммы используется:

- для сравнения двух и более наборов значений или пар значений;
- для отображения взаимосвязи между рядами значений.

Применяется в случаях:

- когда данные представляют собой разные измерения.

The background spreadsheet contains the following data:

	A	B	C	E	F
	Средний доход в районе	Средний возраст зданий в районе	Среднее количество комнат в районе	Среднее количество квартир в районе	Цена квартиры
1					
2	\$ 17 796,6	4,9		47162,2	\$ 302 3
3	\$ 35 454,7	6,9		59636,4	\$ 1 077 8
4	\$ 35 609,0	6,9	7,8	20833,0	\$ 449 3
5	\$ 35 797,3	5,5	7,8	24844,2	\$ 299 8
6	\$ 35 963,3	3,4	8,3	24435,8	\$ 143 0

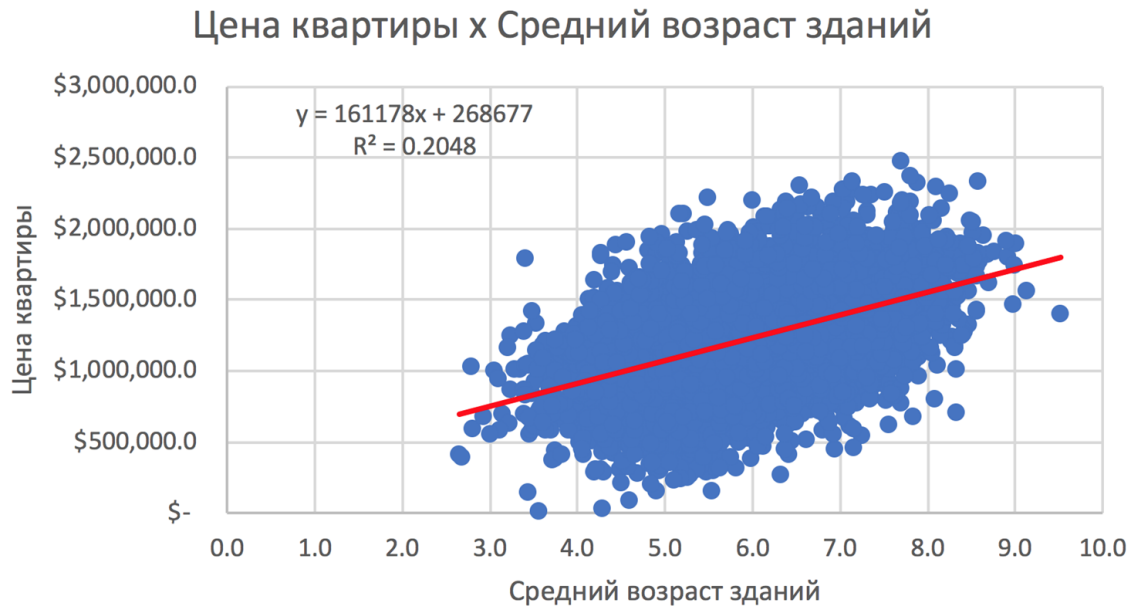
Определение вида и формы зависимости

На графике видно, что между признаками x (средний доход) и y (цена квартиры) действительно наблюдается линейная связь



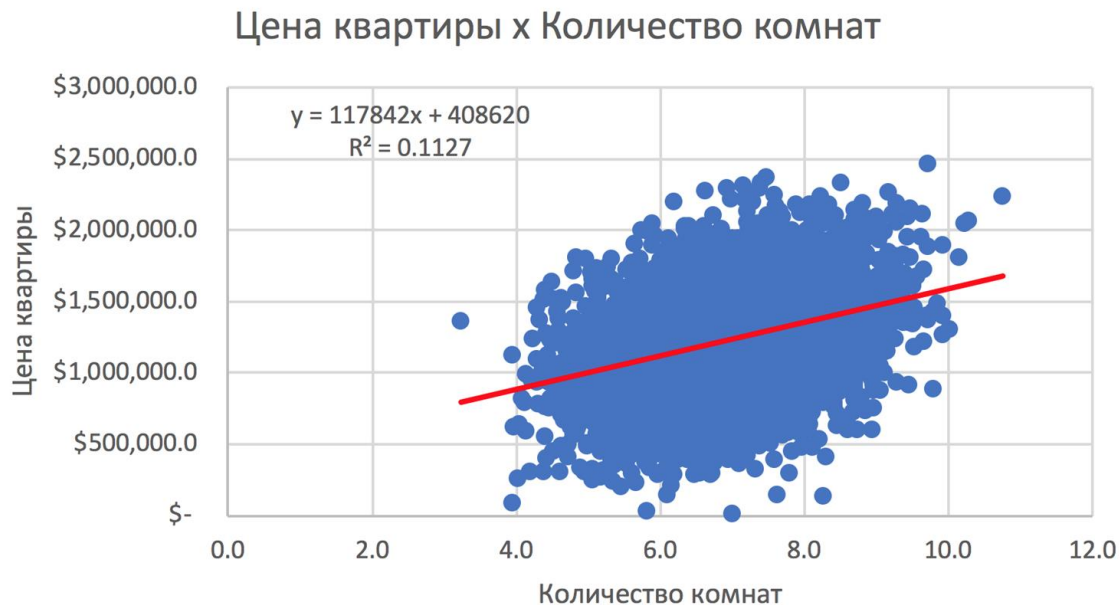
Определение вида и формы зависимости

На графике видно, что между признаками x (средний возраст зданий) и y (цена квартиры) действительно наблюдается линейная связь



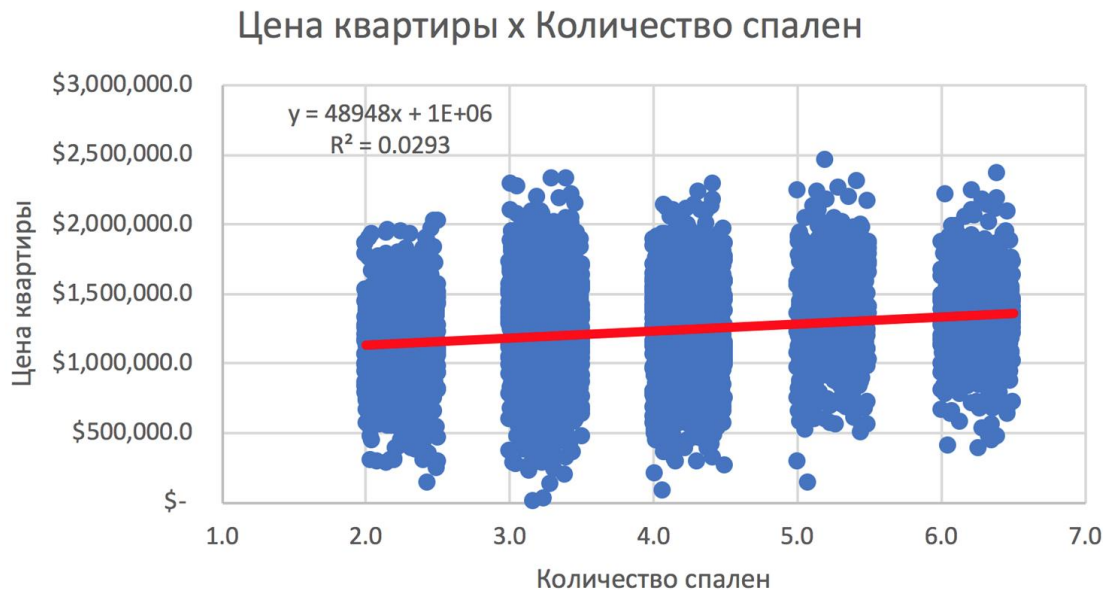
Определение вида и формы зависимости

На графике видно, что между признаками x (количество комнат) и y (цена квартиры) действительно наблюдается линейная связь



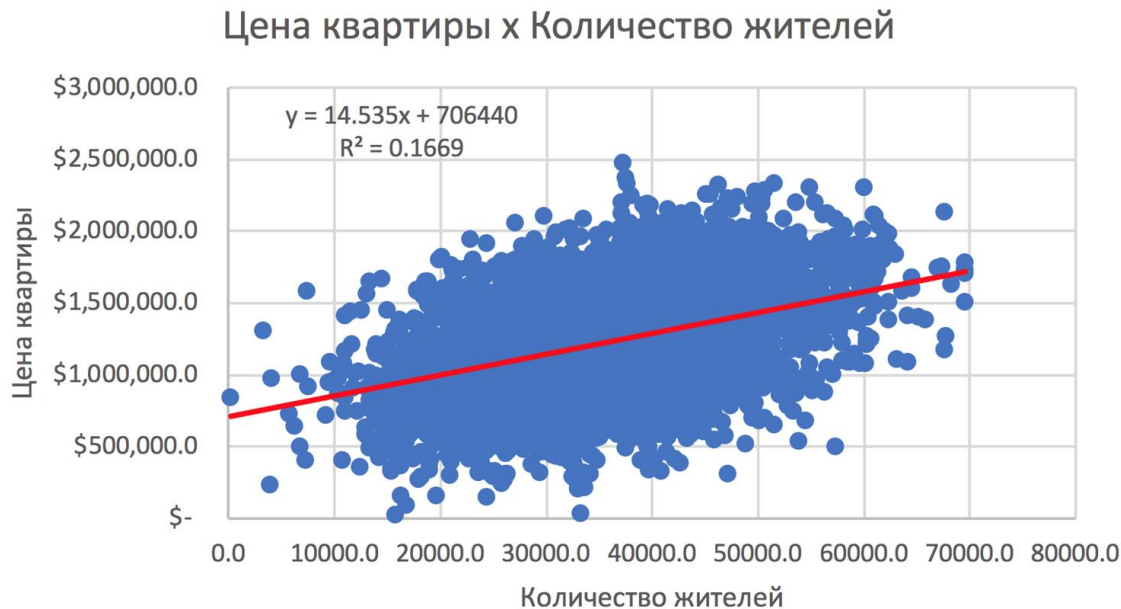
Определение вида и формы зависимости

На графике видно, что между признаками x (количество спален) и y (цена квартиры) наблюдается некоторая линейная связь, но из дальнейшего регрессионного анализа мы увидим, что она статистически незначима



Определение вида и формы зависимости

На графике видно, что между признаками x (количество жителей) и y (цена квартиры) действительно наблюдается линейная связь



Прогнозирование

- Модель

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon_i$$

α — константа

β — параметры модели

x — факторы модели

ε — случайная ошибка модели

Прогнозирование

- Модель

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon_i$$

- Точечный прогноз

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

$\hat{\alpha}$ — константа

$\hat{\beta}$ — параметры модели

x — факторы модели

Интервальное прогнозирование

- Точность прогноза определяется шириной доверительного интервала

$$\hat{\alpha} - t_{\text{кр}} \hat{S}_{\alpha} \leq \alpha \leq \hat{\alpha} + t_{\text{кр}} \hat{S}_{\alpha}$$

$$\hat{\beta} - t_{\text{кр}} \hat{S}_{\beta} \leq \beta \leq \hat{\beta} + t_{\text{кр}} \hat{S}_{\beta}$$

$\hat{\alpha}$ — константа

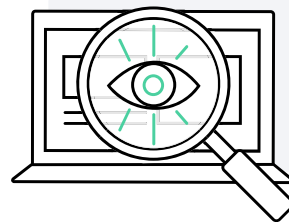
$\hat{\beta}$ — параметры модели

$t_{\text{кр}} \hat{S}_{\alpha}$ — табличное значение t-критерия \times стандартная ошибка $\left(\frac{\sigma}{\sqrt{n}}\right)$ константы

$t_{\text{кр}} \hat{S}_{\beta}$ — табличное значение t-критерия \times стандартная ошибка $\left(\frac{\sigma}{\sqrt{n}}\right)$ параметра

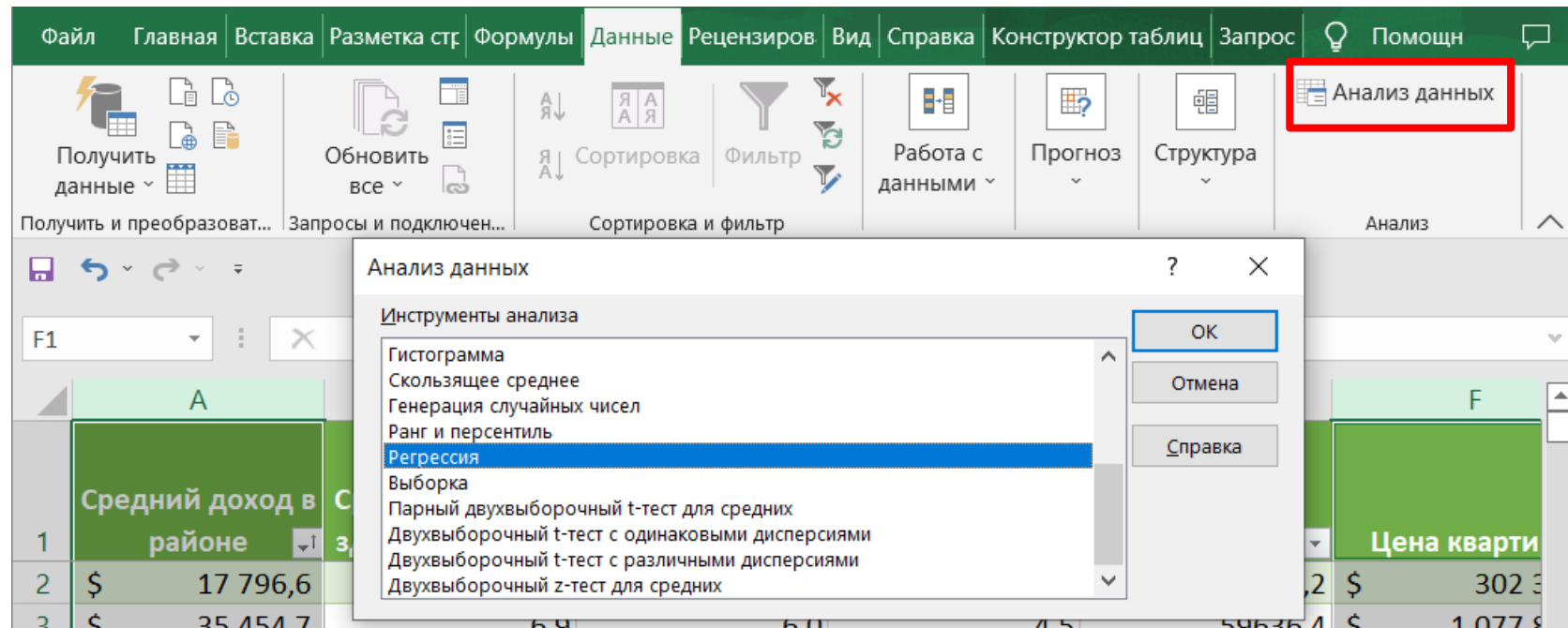
Практика

Вычислим простую линейную регрессию и множественную регрессию в Excel



Простая линейная регрессия

Во вкладке «Данные» выбираем опцию «Анализ данных». Во всплывающем окне будет доступен выбор инструментов анализа



Простая линейная регрессия

Выбираем инструмент «Регрессия», настраиваем входные интервалы и уровень надёжности

The screenshot displays the Microsoft Excel interface with the 'Data' ribbon active. The 'Data Analysis' task pane is open, and the 'Regression' tool is selected. The 'Regression' dialog box is shown with the following settings:

- Входные данные (Input data):**
 - Входной интервал Y:
 - Входной интервал X:
 - ☐ Метки
 - ☒ Уровень надежности: %
 - ☐ Константа - ноль
- Параметры вывода (Output options):**
 - ☐ Выходной интервал:
 - ☒ Новый рабочий лист:
 - ☐ Новая рабочая книга
- Остатки (Residuals):**
 - ☒ Остатки
 - ☒ Стандартизованные остатки
 - ☒ График остатков
 - ☒ График подбора
- Нормальная вероятность (Normal probability):**
 - ☒ График нормальной вероятности

The background shows a table with two columns: 'Средний доход в районе' (Average income in the district) and 'Цена квартиры' (Apartment price). The data is as follows:

Средний доход в районе	Цена квартиры
17 796,6	302 3
35 454,7	1 077 8
35 609,0	449 3
35 797,3	299 8
35 963,3	143 0
36 100,4	599 5
37 908,7	880 4
37 971,2	31 1
38 122,5	899 6
38 139,9	723 7
38 530,1	1 267 9

Простая линейная регрессия

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	Регрессионная статистика						
4	Множественный R	0,6397					
5	R-квадрат	0,4093					
6	Нормированный R-квадрат	0,4091					
7	Стандартная ошибка	271432,1479					
8	Наблюдения	5000					
9							
10	Дисперсионный анализ						
11		df	SS	MS	F	Значимость F	
12	Регрессия	1	255105895277055,00	255105895277055,00	3462,56	0,00	
13	Остаток	4998	368229703675944,00	73675410899,55			
14	Итого	4999	623335598952999,00				
15							
16		Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
17	Y-пересечение	-221579,48	25000,13	-8,86	0,00	-270590,70	-172568,25
18	X1 - Средний доход в районе	21,20	0,36	58,84	0,00	20,49	21,90



Оценка качества прогнозной модели с помощью R-квадрат

$$Y = -221\,579 + 21,20 \times X_1$$

Интервальные оценки коэффициентов

Множественная линейная регрессия

Во входном интервале X должны быть значения всех наших переменных от X_1 до X_5

Регрессия

Входные данные

Входной интервал Y:

Входной интервал X:

☐ Метки

☐ Константа - ноль

☒ Уровень надежности: %

Параметры вывода

☐ Выходной интервал:

☒ Новый рабочий дист.:

☐ Новая рабочая книга

Остатки

☒ Остатки ☒ График остатков

☒ Стандартизованные остатки ☒ График подбора

Нормальная вероятность

☒ График нормальной вероятности

OK Отмена Справка

X_1, X_2, X_3, X_4, X_5

	Средний доход в районе	Средний доход в районе
1		
2	\$ 17 796,6	
3	\$ 35 454,7	
4	\$ 35 609,0	
5	\$ 35 797,3	
6	\$ 35 963,3	
7	\$ 36 100,4	
8	\$ 37 908,7	
9	\$ 37 971,2	
10	\$ 38 122,5	
11	\$ 38 139,9	

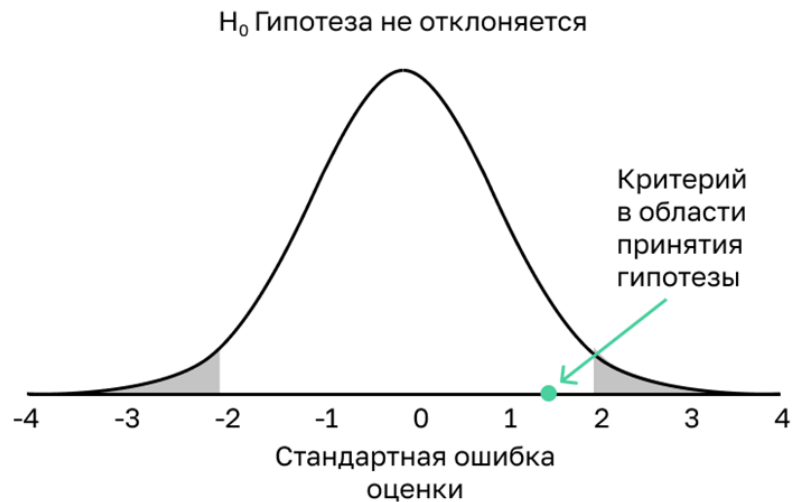
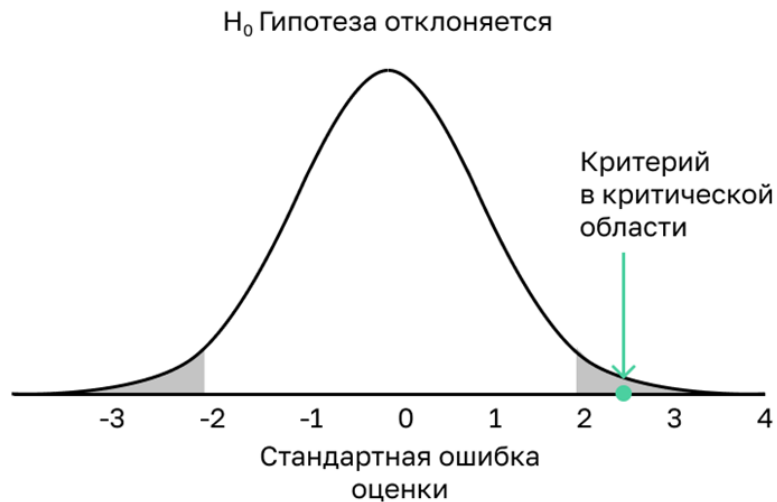
Цена квартиры
302 3
1 077 8
449 3
299 8
143 0
599 5
880 4
31 1
899 6
723 7

Множественная линейная регрессия

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	Регрессионная статистика						
4	Множественный R	0,9581					
5	R-квадрат	0,9180					
6	Нормированный R-квадрат	0,9179					
7	Стандартная ошибка	101153,4118					
8	Наблюдения	5000					
9							
10	Дисперсионный анализ						
11		df	SS	MS	F	Значимость F	
12	Регрессия	5	572236927386735,00	114447385477347,00	11185,23	0,00	
13	Остаток	4994	51098671566264,50	10232012728,53			
14	Итого	4999	623335598952999,00				
15							
16		Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
17	Y-пересечение	-2637299,03	17157,81	-153,71	0,00	-2670935,87	-2603662,19
18	X1 - Средний доход в районе	21,58	0,13	160,66	0,00	21,31	21,84
19	X2 - Средний возраст зданий в районе	165637,03	1443,41	114,75	0,00	162807,30	168466,75
20	X3 - Среднее количество комнат в районе	120659,95	1605,16	75,17	0,00	117513,13	123806,77
21	X4 - Среднее количество спален в районе	1651,14	1308,67	1,26	0,21	-914,43	4216,71
22	X5 - Количество жителей района	15,20	0,14	105,39	0,00	14,92	15,48

$$Y = -2\,637\,299 + 21,58 \times X_1 + 165\,637 \times X_2 + 120\,660 \times X_3 + 1\,651 \times X_4 + 15 \times X_5$$

Проверка значимости уравнения регрессии



Проверка значимости уравнения регрессии

H_0 : уравнение незначимо, $b_i = 0$

H_1 : уравнение значимо, $b_i \neq 0$

$$F_{\text{набл}} = \frac{R^2}{1 - R^2} \times \frac{n - m - 1}{m}$$

m – число параметров при переменных x

n – число наблюдений

$$F_{\text{набл}} = 11\,185,23$$

Проверка значимости уравнения регрессии

H_0 : уравнение незначимо, $b_i = 0$

H_1 : уравнение значимо, $b_i \neq 0$

$$F_{\text{набл}} = \frac{R^2}{1 - R^2} \times \frac{n - m - 1}{m}$$

m – число параметров при переменных x

n – число наблюдений

$$F_{\text{набл}} = 11\,185,23$$

$$F_{\text{табл}} = 2,21$$

$F_{\text{набл}} > F_{\text{табл}}$, значит, H_0 отклоняется

ν_2	α	ν_1 (число степеней свободы)									
		1	2	3	4	5	6	7	8	9	10
40	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
80	.10	2.77	2.37	2.16	2.02	1.93	1.85	1.80	1.75	1.72	1.69
	.05	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95
	.01	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55
100	.10	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.70	1.67
	.05	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92
	.01	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51
120	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
200	.10	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63
	.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
	.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41
∞	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Проверка значимости отдельных коэффициентов уравнения

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	Регрессионная статистика						
4	Множественный R	0,9581					
5	R-квадрат	0,9180					
6	Нормированный R-квадрат	0,9179					
7	Стандартная ошибка	101153,4118					
8	Наблюдения	5000					
9							
10	Дисперсионный анализ						
11		df	SS	MS	F	Значимость F	
12	Регрессия	5	572236927386735,00	114447385477347,00	11185,23	0,00	
13	Остаток	4994	51098671566264,50	10232012728,53			
14	Итого	4999	623335598952999,00				
15							
16		Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
17	Y-пересечение	-2637299,03	17157,81	-153,71	0,00	-2670935,87	-2603662,19
18	X1 - Средний доход в районе	21,58	0,13	160,66	0,00	21,31	21,84
19	X2 - Средний возраст зданий в районе	165637,03	1443,41	114,75	0,00	162807,30	168466,75
20	X3 - Среднее количество комнат в районе	120659,95	1605,16	75,17	0,00	117513,13	123806,77
21	X4 - Среднее количество спален в районе	1651,14	1308,67	1,26	0,21	-914,43	4216,71
22	X5 - Количество жителей района	15,20	0,14	105,39	0,00	14,92	15,48

p-value

Связь p-value и уровней значимости и доверия

Уровень значимости α — это вероятность отвергнуть нулевую гипотезу при условии, что она верна

Связь p-value и уровней значимости и доверия

Уровень значимости α — это вероятность отвергнуть нулевую гипотезу при условии, что она верна

p-value — вероятность получить наблюдаемое или ещё большее отклонение оценки от гипотезы, если она (гипотеза) верна

Связь p-value и уровней значимости и доверия

Уровень значимости α — это вероятность отвергнуть нулевую гипотезу при условии, что она верна

p-value — вероятность получить наблюдаемое или ещё большее отклонение оценки от гипотезы, если она (гипотеза) верна

- Если **p-value < уровня значимости α** , нулевую гипотезу можно отвергнуть и принять альтернативную гипотезу

Связь p-value и уровней значимости и доверия

Уровень значимости α — это вероятность отвергнуть нулевую гипотезу при условии, что она верна

p-value — вероятность получить наблюдаемое или ещё большее отклонение оценки от гипотезы, если она (гипотеза) верна

- Если $p\text{-value} < \text{уровня значимости } \alpha$, нулевую гипотезу можно отвергнуть и принять альтернативную гипотезу
- Если **p-value > уровня значимости α** , оснований отвергать нулевую гипотезу нет

Построение точечных и интервальных прогнозов результирующей переменной

Задание

Спрогнозировать Цену квартиры Y при:

- Средний доход в районе $X_1 = \$ 100\,000$
- Средний возраст зданий в районе $X_2 = 3,5$
- Среднее количество комнат в районе $X_3 = 7$
- Среднее количество спален в районе $X_4 = 4$
- Количество жителей района $X_5 = 25\,000$

Построение точечного прогноза результатирующей переменной

Формула точечного прогноза:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

$\hat{\alpha}$ — константа

$\hat{\beta}$ — параметры модели

x — факторы модели

Построение точечного прогноза результатирующей переменной

Формула точечного прогноза:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

Подставляем значения коэффициентов:

$$Y = -2\,637\,299 + 21,58 \times X_1 + 165\,637 \times X_2 + 120\,660 \times X_3 + 1\,651 \times X_4 + 15 \times X_5$$

Построение точечного прогноза результатирующей переменной

Формула точечного прогноза:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

Подставляем значения коэффициентов:

$$Y = -2\,637\,299 + 21,58 \times X_1 + 165\,637 \times X_2 + 120\,660 \times X_3 + 1\,651 \times X_4 + 15 \times X_5$$

Подставляем значения X (из условия):

$$Y = -2\,637\,299 + 21,58 \times 100\,000 + 165\,637 \times 3,5 + 120\,660 \times 7 + 1\,651 \times 4 + 15 \times 25\,000 = \mathbf{\$1\,331\,478,3}$$

Построение интервального прогноза результатирующей переменной

Формула доверительного интервала:

$$\hat{\alpha} + \hat{\beta}_i x_i - t_{\text{кр}} \hat{S}_{y(x)} \leq \alpha + \beta_i x_i \leq \hat{\alpha} + \hat{\beta}_i x_i + t_{\text{кр}} \hat{S}_{y(x)}$$

$\hat{\alpha}$ — константа

$\hat{\beta}$ — параметры модели

x — факторы модели

$t_{\text{кр}} \hat{S}_{y(x)}$ — табличное значение t-критерия × стандартная ошибка $\left(\frac{\sigma}{\sqrt{n}}\right)$ регрессии

Построение интервального прогноза результатирующей переменной

Формула доверительного интервала:

$$\hat{\alpha} + \hat{\beta}_i x_i - t_{\text{кр}} \hat{S}_{y(x)} \leq \alpha + \beta_i x_i \leq \hat{\alpha} + \hat{\beta}_i x_i + t_{\text{кр}} \hat{S}_{y(x)}$$

Подставляем полученное значение и вычисляем (95%-ый доверительный интервал):

$$1\,331\,478,3 - 1,96 \times 101153,4 \leq \alpha + \beta_i x_i \leq 1\,331\,478,3 + 1,96 \times 101153,4$$

$$1\,133\,217,6 \leq \alpha + \beta_i x_i \leq 1\,529\,739,0$$

Построение интервального прогноза результатирующей переменной

Формула доверительного интервала:

$$\hat{\alpha} + \hat{\beta}_i x_i - t_{кр} \hat{S}_{y(x)} \leq \alpha + \beta_i x_i \leq \hat{\alpha} + \hat{\beta}_i x_i + t_{кр} \hat{S}_{y(x)}$$

Подставляем полученное значение и вычисляем:

$$1\,331\,478,3 - 1,96 \times 101\,153,4 \leq \alpha + \beta_i x_i \leq 1\,331\,478,3 + 1,96 \times 101\,153,4$$

$$1\,133\,217,6 \leq \alpha + \beta_i x_i \leq 1\,529\,739,0$$

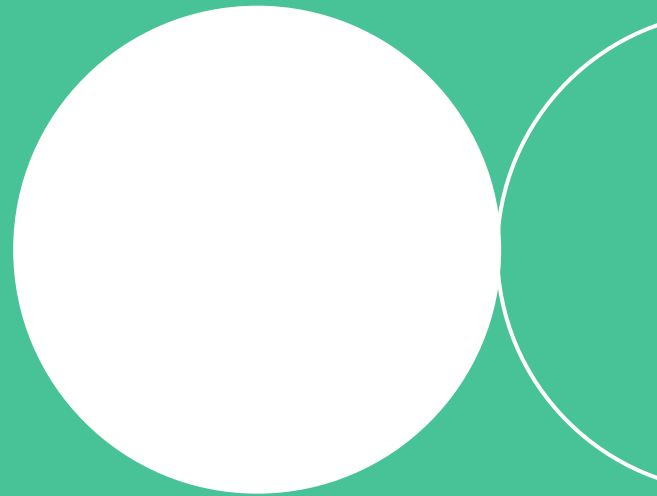
Интервальный прогноз (95%-ый доверительный интервал):

\$1 133 217,6 – \$1 529 739,0

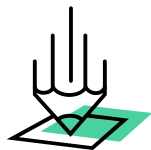


Ваши вопросы?

Итоги



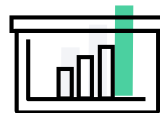
Основные тезисы



Сделали описательную статистику и проанализировали данные



Нашли выбросы в данных



Провели корреляционный анализ



Провели регрессионный анализ, оценили качество прогнозной модели



Сделали прогноз на основе модели



Протестировали гипотезы и визуализировали данные

Проведение статистического исследования в Excel

Александра Калинина
PhD Management
CMO at IPification

