

ОБРАБОТКА И МОДЕЛИРОВАНИЕ ДАННЫХ В MS EXCEL



Екатерина
Золотарева

ТЕМА 2. ПРЕДОБРАБОТКА ДАННЫХ

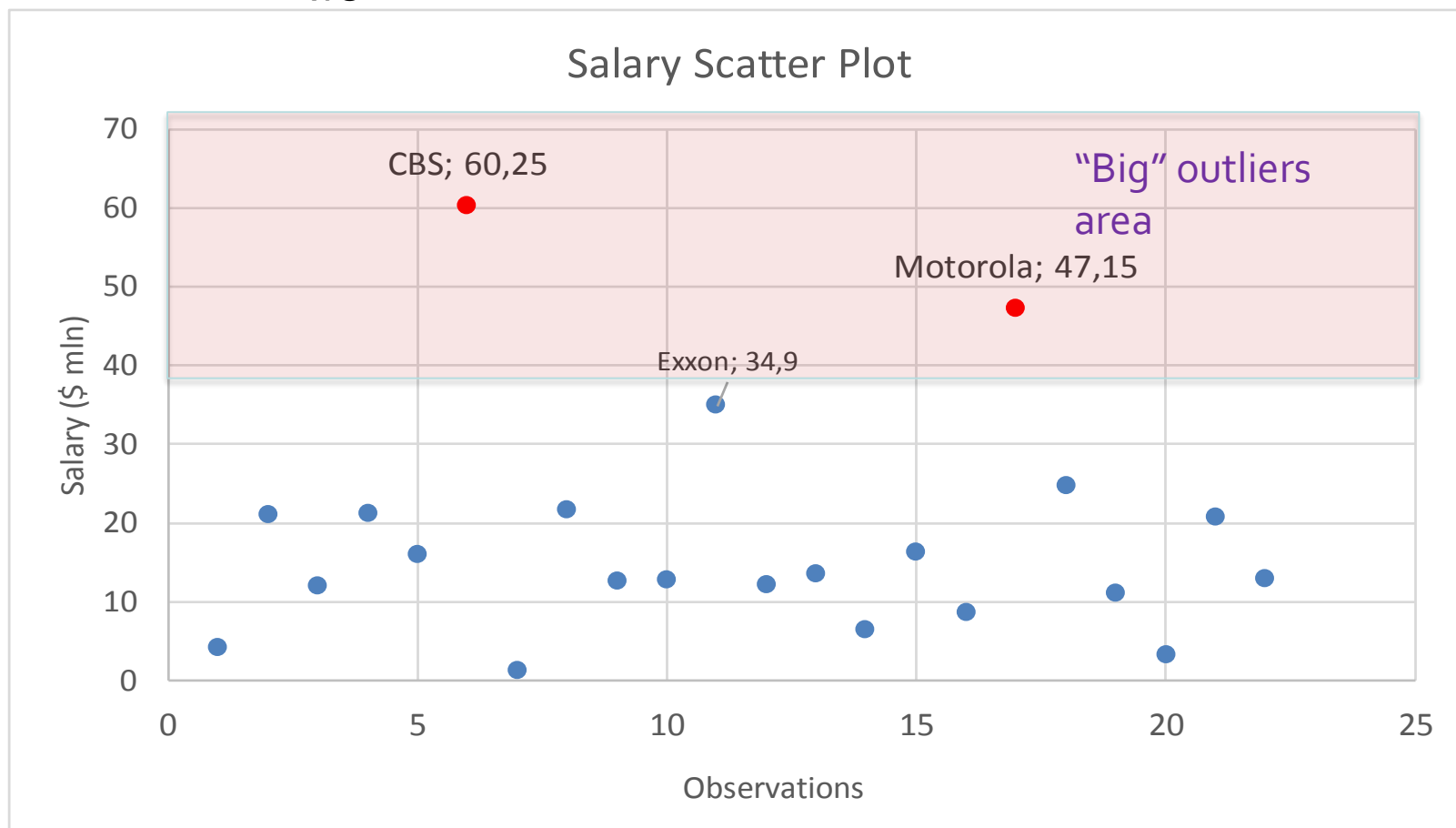
Предобработка данных

- Выбросы
- Пропуски
- Дубликаты
- Синтетические признаки

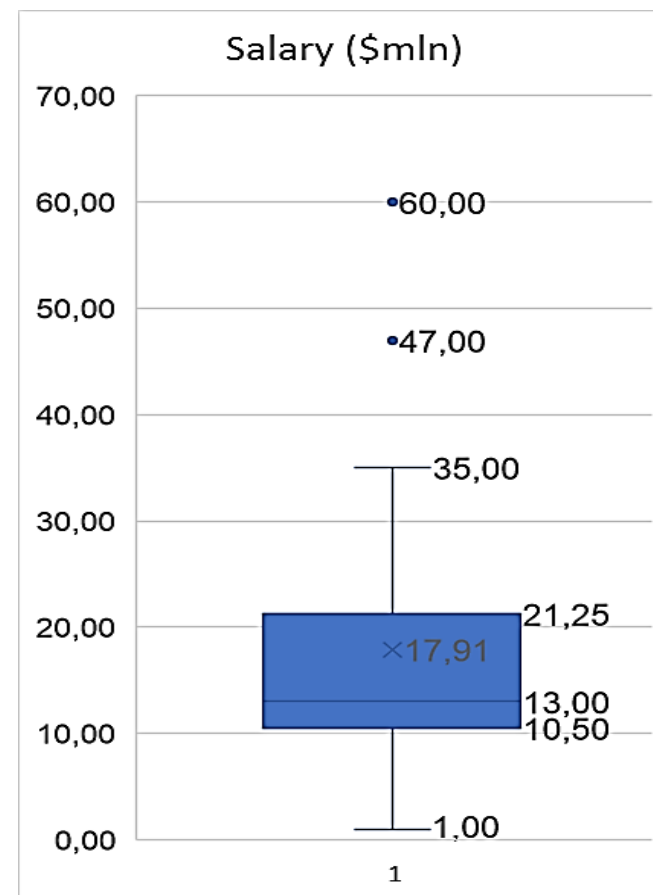
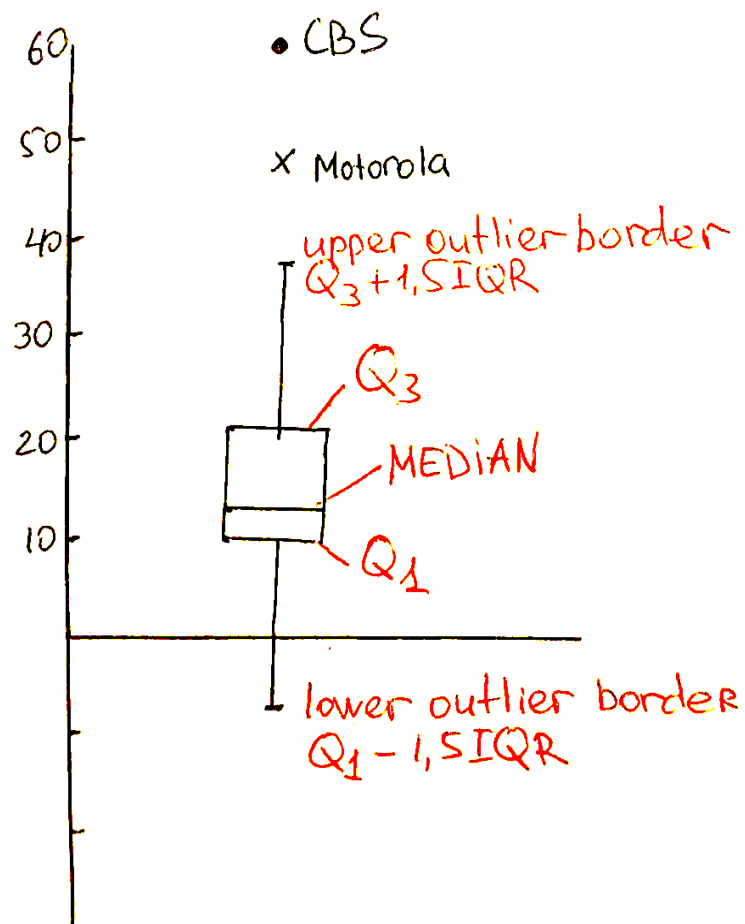
Данные могут быть
неточными,
неполными,
противоречивыми,
разнородными,
косвенными....
И иметь гигантские
объемы

Выбросы

Выбросы – значения признака, не попавшие в отрезок $[x_{0,25} - 1,5 \text{ IQR}; x_{0,75} + 1,5 \text{ IQR}]$



Ящик с усами (box-whiskers plot)



Выбросы сверху – положительная асимметрия



Причины выбросов и работа с ними

- Аномальные явления
- Ошибки ввода
- Преднамеренное искажение

Чем плохи выбросы?



Заменить на границы
 $[x_{0,25} - 1,5 \text{ IQR}; x_{0,75} + 1,5 \text{ IQR}]$



Работать как с пропуском

Пример: Кофеварки

Дата	Платформа	Серебристые кофеварки	Красные кофеварки	Курс USD/RUR	Расходы на рекламу	Цена	Выброс - серебр.	Пропущ - серебр.	Пропущ - реклама
01.07.2017	VK	97	67	59,3862	9 000	2500	0	0	0
02.07.2017	VK	98	67	59,3862	9 000	2500	0	0	0
03.07.2017	VK	110	77	59,3862	10 400	2500	0	0	0
04.07.2017	Facebook	134	99	58,9695	9 800	2500	0	0	0
05.07.2017	Facebook	159	118	59,2295	13 500	2500	0	0	0
06.07.2017	Facebook	103	69	59,5787	9 000	2500	0	0	0
06.07.2017	Facebook	103	69	59,5787	9 000	2500	0	0	0
07.07.2017	Facebook	143	101	60,2426	13 500	2500	0	0	0
08.07.2017	Facebook	123	86	60,3792	11 300	2500	0	0	0
09.07.2017	Facebook	111	95	60,3792	12 600	2500	0	1	0
10.07.2017	Facebook	140	98	60,3792	13 100	2500	0	0	0
11.07.2017	Facebook	162	120	60,3014	13 500	2500	0	0	0
12.07.2017	Facebook	111	95	60,7397	9 900	2500	0	1	0
13.07.2017	Facebook	109	75	60,6227	9 900	2500	0	0	0
14.07.2017	Facebook	122	85	60,1836	11 300	2500	0	0	0
15.07.2017	Facebook	98	62	59,8806	10 800	5000	1	0	0
16.07.2017	Facebook	81	50	59,8806	9 000	5000	1	0	0
17.07.2017	Facebook	115	76	59,8806	12 600	5000	0	0	0
18.07.2017	VK	131	92	59,0657	12 200	5000	0	0	0
19.07.2017	VK	122	85	59,3705	11 300	5000	0	0	0
20.07.2017	VK	71	42	59,2418	10 800	5000	0	0	1
21.07.2017	VK	83	50	59,0823	9 000	5000	0	0	0
22.07.2017	VK	112	75	58,9325	10 800	5000	0	0	0
23.07.2017	VK	120	82	58,9325	11 700	5000	0	0	0
24.07.2017	VK	121	82	58,9325	11 700	5000	0	0	0
25.07.2017	VK	156	113	59,6572	13 500	5000	0	0	0
26.07.2017	VK	176	129	59,8185	15 800	3500	0	0	0
27.07.2017	VK	104	68	59,9102	9 900	3500	0	0	0
28.07.2017	VK	96	63	59,4102	9 000	3500	0	0	0
29.07.2017	VK	100	66	59,5436	9 500	3500	0	0	0
30.07.2017	Facebook	88	57	59,5436	8 100	3500	0	0	0
31.07.2017	Facebook	76	47	59,5436	6 800	3500	0	0	0
Медианы		111			10 800				

Пропуски

- Пустое значение
- #ДЕЛ/0, #Н/Д, #ИМЯ?, #ПУСТО!, #ЧИСЛО!, #ССЛЫКА!, #ЗНАЧ!
- NA, NaN
- 9999,99999999

Причины пропусков и работа с ними

- Неизвестны в принципе
- Утрачены
- Появились при удалении выброса

Чем плохи пропуски?

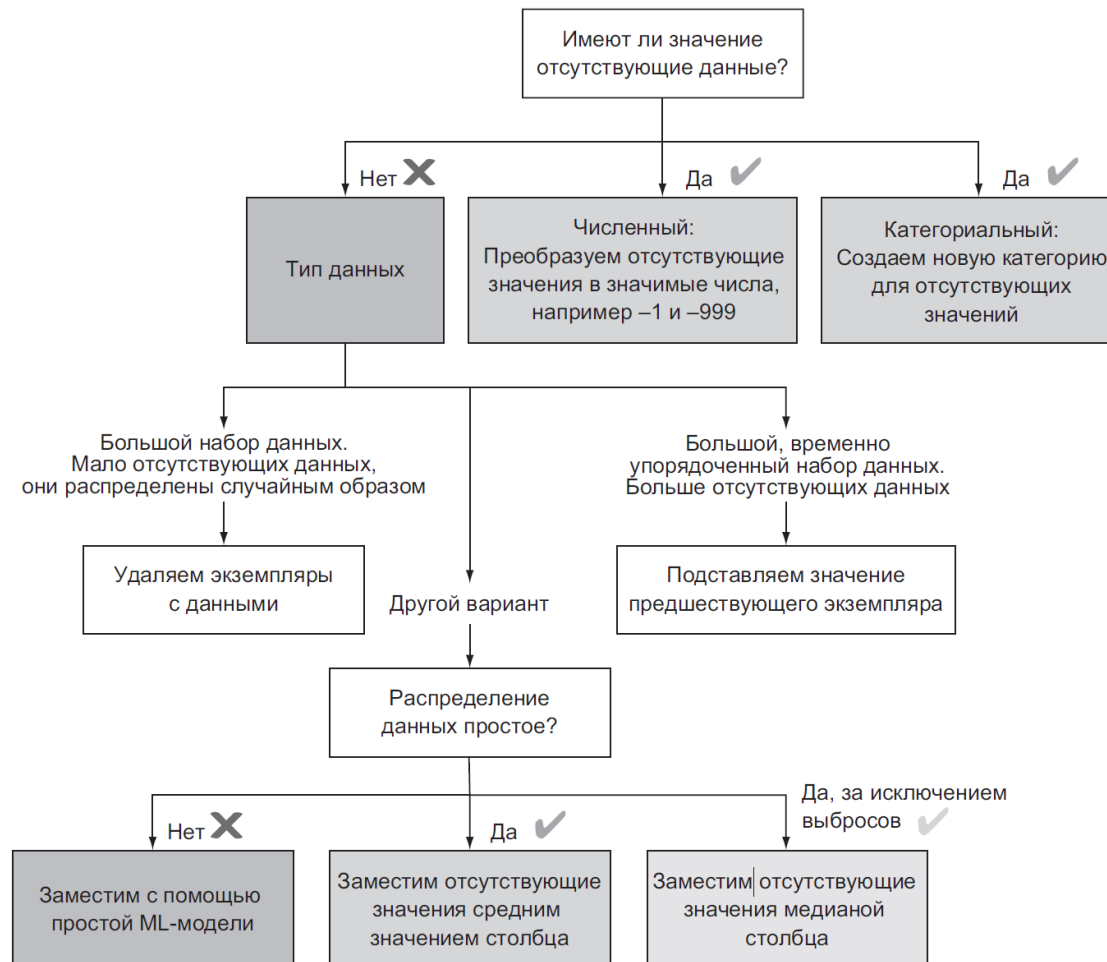


Интерполяция (каким-то способом)



Удаление всей строки

Работа с пропусками



Источник:

Бринк Х., Ричардс Дж., Феверолф М.
Машинное обучение. — СПб.: Питер, 2017.

Илл. 2.9. Полная диаграмма решений для обработки отсутствующих значений при подготовке данных к ML-моделированию

Пример: Кофеварки

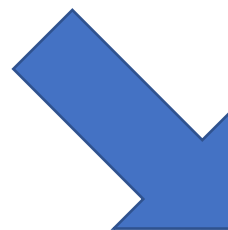
Дата	Платформа	Серебристые кофеварки	Красные кофеварки	Курс USD/RUR	Расходы на рекламу	Цена	Выброс - серебр.	Пропущ - серебр.	Пропущ - реклама
01.07.2017	VK	97	67	59,3862	9 000	2500	0	0	0
02.07.2017	VK	98	67	59,3862	9 000	2500	0	0	0
03.07.2017	VK	110	77	59,3862	10 400	2500	0	0	0
04.07.2017	Facebook	134	99	58,9695	9 800	2500	0	0	0
05.07.2017	Facebook	159	118	59,2295	13 500	2500	0	0	0
06.07.2017	Facebook	103	69	59,5787	9 000	2500	0	0	0
06.07.2017	Facebook	103	69	59,5787	9 000	2500	0	0	0
07.07.2017	Facebook	143	101	60,2426	13 500	2500	0	0	0
08.07.2017	Facebook	123	86	60,3792	11 300	2500	0	0	0
09.07.2017	Facebook	111	95	60,3792	12 600	2500	0	1	0
10.07.2017	Facebook	140	98	60,3792	13 100	2500	0	0	0
11.07.2017	Facebook	162	120	60,3014	13 500	2500	0	0	0
12.07.2017	Facebook	111	95	60,7397	9 900	2500	0	1	0
13.07.2017	Facebook	109	75	60,6227	9 900	2500	0	0	0
14.07.2017	Facebook	122	85	60,1836	11 300	2500	0	0	0
15.07.2017	Facebook	98	62	59,8806	10 800	5000	1	0	0
16.07.2017	Facebook	81	50	59,8806	9 000	5000	1	0	0
17.07.2017	Facebook	115	76	59,8806	12 600	5000	0	0	0
18.07.2017	VK	131	92	59,0657	12 200	5000	0	0	0
19.07.2017	VK	122	85	59,3705	11 300	5000	0	0	0
20.07.2017	VK	71	42	59,2418	10 800	5000	0	0	1
21.07.2017	VK	83	50	59,0823	9 000	5000	0	0	0
22.07.2017	VK	112	75	58,9325	10 800	5000	0	0	0
23.07.2017	VK	120	82	58,9325	11 700	5000	0	0	0
24.07.2017	VK	121	82	58,9325	11 700	5000	0	0	0
25.07.2017	VK	156	113	59,6572	13 500	5000	0	0	0
26.07.2017	VK	176	129	59,8185	15 800	3500	0	0	0
27.07.2017	VK	104	68	59,9102	9 900	3500	0	0	0
28.07.2017	VK	96	63	59,4102	9 000	3500	0	0	0
29.07.2017	VK	100	66	59,5436	9 500	3500	0	0	0
30.07.2017	Facebook	88	57	59,5436	8 100	3500	0	0	0
31.07.2017	Facebook	76	47	59,5436	6 800	3500	0	0	0
Медианы		111			10 800				

Дубликаты (и противоречия)

Повторяющиеся значения в отдельных столбцах
или в строке целиком



Противоречит логике
(например,
повторяющийся ключ)



Не противоречит

Чем плохи дубликаты?

Пример: Кредиты и депозиты

<i>Дата</i>	<i>Число заключенных кредитных договоров</i>	<i>Число заключенных депозитных договоров</i>
01.08.2017	126	75
02.08.2017	123	0
02.08.2017	123	80
	136	84
04.08.2017		75
05.08.2017	128	73
05.08.2017	128	73

Пример: Акции

	Date	Open	High	Low	Close	Volume	IDselect	Type	Username	File
26615	2007-05-16	15.49710	15.545700	14.908600	15.33430	1.317850e+08	4.0	Trend	DVM	AAPL UW Equity_mapped (4).csv
464991	2007-05-16	15.49710	15.545700	14.908600	15.33430	1.317850e+08	1.0	Trend	NZ	AAPL UW Equity_mapped (5).csv
485835	2007-05-16	15.50427	15.547127	14.774271	15.33427	2.819257e+08	NaN	NaN	VB	AAPL_daily_10Y_mapped.csv
931436	2007-05-16	15.49710	15.545700	14.908600	15.33430	1.317850e+08	10.0	Trend	Bark	AAPL UW Equity_mapped (6).csv
1388054	2007-05-16	15.49710	15.545700	14.908600	15.33430	1.317850e+08	8.0	Trend	Bragin	AAPL UW Equity_mapped (3).csv
1586002	2007-05-16	15.49710	15.545700	14.908600	15.33430	1.317850e+08	NaN	NaN	SV	AAPL UW Equity_mapped.csv

Синтетические признаки

Синтетические признаки – признаки, полученные путем преобразования исходных признаков, например:

- цена → доходность
- расходы, доходы → прибыль
- таймстамп → время суток, день недели и пр.

По смыслу

Синтетические признаки

Синтетические признаки – признаки, полученные путем преобразования исходных признаков, например:

- z-преобразование $\frac{X - \bar{x}}{s_X}$
- Min-max преобразование $\frac{X - x_{min}}{x_{max} - x_{min}}$
- Логарифмирование, возведение в квадрат, корень

Стандартизация

Пример: Кофеварки

Дата	Платформа	Серебристые кофеварки	Красные кофеварки	Курс USD/RUR	Расходы на рекламу	Цена	Выброс - серебр.	Пропущ - серебр.	Пропущ - реклама
01.07.2017	VK	97	67	59,3862	9 000	2500	0	0	0
02.07.2017	VK	98	67	59,3862	9 000	2500	0	0	0
03.07.2017	VK	110	77	59,3862	10 400	2500	0	0	0
04.07.2017	Facebook	134	99	58,9695	9 800	2500	0	0	0
05.07.2017	Facebook	159	118	59,2295	13 500	2500	0	0	0
06.07.2017	Facebook	103	69	59,5787	9 000	2500	0	0	0
06.07.2017	Facebook	103	69	59,5787	9 000	2500	0	0	0
07.07.2017	Facebook	143	101	60,2426	13 500	2500	0	0	0
08.07.2017	Facebook	123	86	60,3792	11 300	2500	0	0	0
09.07.2017	Facebook	111	95	60,3792	12 600	2500	0	1	0
10.07.2017	Facebook	140	98	60,3792	13 100	2500	0	0	0
11.07.2017	Facebook	162	120	60,3014	13 500	2500	0	0	0
12.07.2017	Facebook	111	95	60,7397	9 900	2500	0	1	0
13.07.2017	Facebook	109	75	60,6227	9 900	2500	0	0	0
14.07.2017	Facebook	122	85	60,1836	11 300	2500	0	0	0
15.07.2017	Facebook	98	62	59,8806	10 800	5000	1	0	0
16.07.2017	Facebook	81	50	59,8806	9 000	5000	1	0	0
17.07.2017	Facebook	115	76	59,8806	12 600	5000	0	0	0
18.07.2017	VK	131	92	59,0657	12 200	5000	0	0	0
19.07.2017	VK	122	85	59,3705	11 300	5000	0	0	0
20.07.2017	VK	71	42	59,2418	10 800	5000	0	0	1
21.07.2017	VK	83	50	59,0823	9 000	5000	0	0	0
22.07.2017	VK	112	75	58,9325	10 800	5000	0	0	0
23.07.2017	VK	120	82	58,9325	11 700	5000	0	0	0
24.07.2017	VK	121	82	58,9325	11 700	5000	0	0	0
25.07.2017	VK	156	113	59,6572	13 500	5000	0	0	0
26.07.2017	VK	176	129	59,8185	15 800	3500	0	0	0
27.07.2017	VK	104	68	59,9102	9 900	3500	0	0	0
28.07.2017	VK	96	63	59,4102	9 000	3500	0	0	0
29.07.2017	VK	100	66	59,5436	9 500	3500	0	0	0
30.07.2017	Facebook	88	57	59,5436	8 100	3500	0	0	0
31.07.2017	Facebook	76	47	59,5436	6 800	3500	0	0	0
Медианы		111			10 800				

Предобработка данных с помощью текстовых функций

Product Data

54482100AFES | CONTROLLER SERVER 1TB H | 304.00
54482100JCP9 | DESKTOP UNIT | 225.00
54482700BAAS | DESKTOP WINDOWS 8.1 SERVER | 2302.00
54482600BAAS | DESKTOP WINDOWS 8.1 WKST | 355.00
54482100BAAS | DESKTOP WINDOWS 10 | 182.00
54482200BAAS | DESKTOP WINDOWS DESKTOP OS | 255.00
54482500BAAS | DESKTOP WINDOWS OS | 354.00
54483000BAAS | MINITOWER NO OS | 1840.00
54483000KEBB | MINI TOWER | 2550.00

	A	B
1		
2		Product Data
3		54482100AFES CONTROLLER SERVER 1TB H 304.00
4		54482100JCP9 DESKTOP UNIT 225.00
5		54482700BAAS DESKTOP WINDOWS 8.1 SERVER 2302.00
6		54482600BAAS DESKTOP WINDOWS 8.1 WKST 355.00
7		54482100BAAS DESKTOP WINDOWS 10 182.00
8		54482200BAAS DESKTOP WINDOWS DESKTOP OS 255.00
9		54482500BAAS DESKTOP WINDOWS OS 354.00
10		54483000BAAS MINITOWER NO OS 1840.00
11		54483000KEBB MINI TOWER 2550.00

Предобработка данных с помощью текстовых функций

Функция	Описание
ПЕЧСИМВ	Удаляет из текста все непечатаемые символы.
СЦЕП	Объединяет текст из нескольких диапазонов или строк, но не добавляет разделитель или аргументы IgnoreEmpty.
СЦЕПИТЬ	Объединяет несколько текстовых элементов в один.
СОВПАД	Проверяет идентичность двух текстовых значений.
НАЙТИ, НАЙТИ	Ищет вхождения одного текстового значения в другом (с учетом регистра).
ФИКСИРОВАННЫЙ	Форматирует число и преобразует его в текст с заданным числом десятичных знаков.
ЛЕВСИМВ	Возвращают крайние слева знаки текстового значения.
ДЛСТР	Возвращают количество знаков в текстовой строке.
ПСТР	Возвращают заданное число знаков из строки текста, начиная с указанной позиции.
ЧЗНАЧ	Преобразует текст в число независимо от языкового стандарта.
ЗАМЕНИТЬ	Заменяют знаки в тексте.
ПРАВСИМВ	Возвращают крайние справа знаки текстовой строки.
ПОИСК	Ищут вхождения одного текстового значения в другом (без учета регистра).
ПОДСТАВИТЬ	Заменяет в текстовой строке старый текст новым.
ТЕКСТ	Форматирует число и преобразует его в текст.
ОБЪЕДИНИТЬ	Объединяет текст из нескольких диапазонов или строк, вставляя между текстовыми значениями указанный разделитель. Если в качестве разделителя используется пустая текстовая строка, функция эффективно объединит диапазоны.
СЖПРОБЕЛЫ	Удаляет из текста пробелы.
ЗНАЧЕН	Преобразует текстовый аргумент в число.

Пример: ПЕЧСИМВ и СЖПРОБЕЛЫ

A	B	C
1		
2	Product Data	Data Cleaning
3	54482100AFES CONTROLLER SERVER 1TB H 304.00	=TRIM(CLEAN(B3))
4	54482100JCP9 DESKTOP UNIT 225.00	=TRIM(CLEAN(B4))
5	54482700BAAS DESKTOP WINDOWS 8.1 SERVER 2302.00	=TRIM(CLEAN(B5))
6	54482600BAAS DESKTOP WINDOWS 8.1 WKST 355.00	=TRIM(CLEAN(B6))
7	54482100BAAS DESKTOP WINDOWS 10 182.00	=TRIM(CLEAN(B7))
8	54482200BAAS DESKTOP WINDOWS DESKTOP OS 255.00	=TF
9	54482500BAAS DESKTOP WINDOWS OS 354.00	=TF
10	54483000BAAS MINITOWER NO OS 1840.00	=TF
11	54483000KEBB MINI TOWER 2550.00	=TF

Raw Data		Nonprintable Characters and Excess Spaces removed	
A	B		C
1			
2	Product Data		Data Cleaning
3	54482100AFES CONTROLLER SERVER 1TB H 304.00	54482100AFES CONTROLLER SERVER 1TB H 304.00	
4	54482100JCP9 DESKTOP UNIT 225.00	54482100JCP9 DESKTOP UNIT 225.00	
5	54482700BAAS DESKTOP WINDOWS 8.1 SERVER 2302.00	54482700BAAS DESKTOP WINDOWS 8.1 SERVER 2302.00	
6	54482600BAAS DESKTOP WINDOWS 8.1 WKST 355.00	54482600BAAS DESKTOP WINDOWS 8.1 WKST 355.00	
7	54482100BAAS DESKTOP WINDOWS 10 182.00	54482100BAAS DESKTOP WINDOWS 10 182.00	
8	54482200BAAS DESKTOP WINDOWS DESKTOP OS 255.00	54482200BAAS DESKTOP WINDOWS DESKTOP OS 255.00	
9	54482500BAAS DESKTOP WINDOWS OS 354.00	54482500BAAS DESKTOP WINDOWS OS 354.00	
10	54483000BAAS MINITOWER NO OS 1840.00	54483000BAAS MINITOWER NO OS 1840.00	
11	54483000KEBB MINI TOWER 2550.00	54483000KEBB MINI TOWER 2550.00	

Предобработка данных с помощью логических и информационных функций

Логические

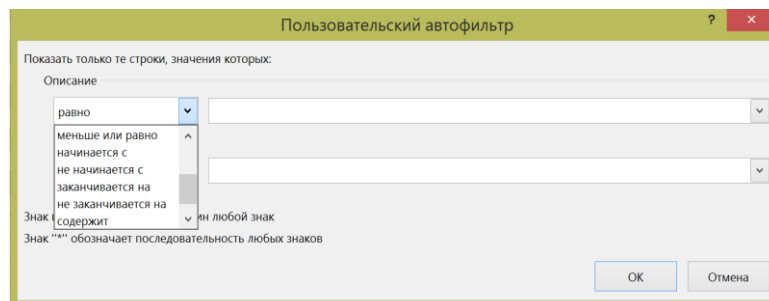
- ЕСЛИ
- И, ИЛИ, ИСКИЛИ
- ИСТИНА, ЛОЖЬ
- НЕ
- ЕСЛИОШИБКА, ЕСНД

Информационные

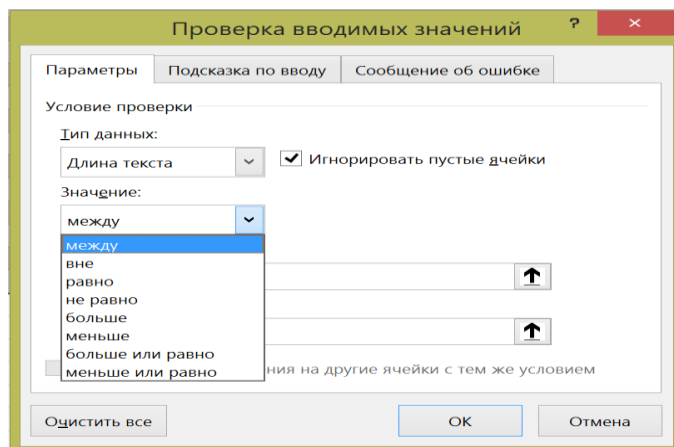
- ЕПУСТО, ЕОШ, ЕОШИБКА, ЕНД
- ЯЧЕЙКА, ТИП, ТИП.ОШИБКИ
- ЕТЕКСТ, ЕЧИСЛО, ЕЛОГИЧ

Пример: фильтр, форматирование и проверка данных

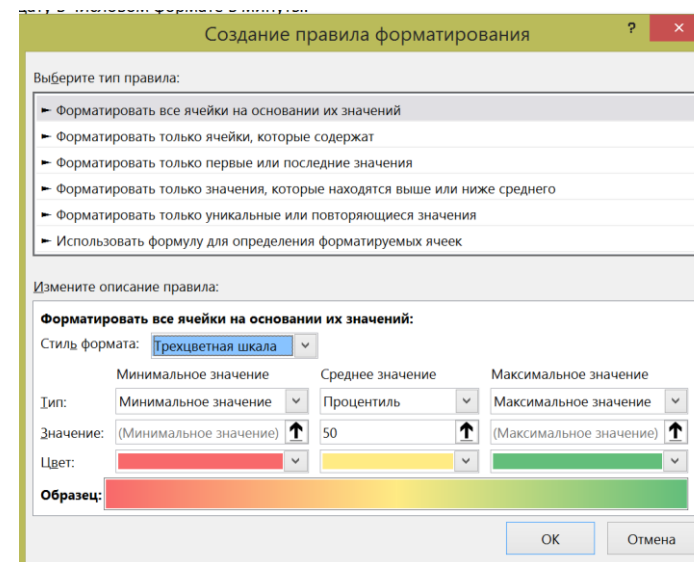
Фильтрация



Проверка данных



Условное форматирование



Предобработка данных с помощью функций времени

Функция	Описание
ДАТА	Возвращает заданную дату в числовом формате.
РАЗНДАТ	Вычисляет количество дней, месяцев или лет между двумя датами. Эта функция полезна в формулах для расчета возраста.
ДЕНЬ	Преобразует дату в числовом формате в день месяца.
ДНИ	Возвращает количество дней между двумя датами.
ЧАС	Преобразует дату в числовом формате в часы.
МИНУТЫ	Преобразует дату в числовом формате в минуты.
МЕСЯЦ	Преобразует дату в числовом формате в месяцы.
ЧИСТРАБДНИ	Возвращает количество полных рабочих дней между двумя датами.
ТДАТА	Возвращает текущую дату и время в числовом формате.
СЕКУНДЫ	Преобразует дату в числовом формате в секунды.
ВРЕМЯ	Возвращает заданное время в числовом формате.
ВРЕМЗНАЧ	Преобразует время из текстового формата в числовой.
СЕГОДНЯ	Возвращает текущую дату в числовом формате.
ДЕНЬНЕД	Преобразует дату в числовом формате в день недели.
РАБДЕНЬ	Возвращает дату в числовом формате, отстоящую вперед или назад на заданное количество рабочих дней.
ГОД	Преобразует дату в числовом формате в год.

Пример: синтетический признак дня недели

Данные		
14.02.2008		
Формула	Описание (результат)	Результат
=ДЕНЬНЕД(A2)	День недели с числом от 1 (воскресенье) до 7 (суббота) (5)	5
=ДЕНЬНЕД(A2; 2)	День недели с числом от 1 (понедельник) до 7 (воскресенье) (4)	4
=ДЕНЬНЕД(A2; 3)	День недели с числом от 0 (понедельник) до 6 (воскресенье) (3)	3

Предобработка данных с помощью Power Query

Файл Главная страница Преобразование Добавление столбца Просмотр

Группировать по Использовать первую строку в качестве заголовков Таблица

Транспонировать
Обратить строки
Считать строки

Тип данных: Текст
Определить тип данных
Переименовать

Любой столбец

Объединить столбцы
Извлечь
Выполнить анализ

Столбец "Текст"

Статистика
Стандартный
Научный

Столбец "Количество"

Тригонометрические
Округление
Информация

Столбец "Дата и время"

Структурированный столбец

Запросы [2]
Table 0
Table 4

АВС	Функция	АВС	Описание
1	ДАТА		Возвращает заданную дату в числовом формате.
2	РАЗДАТ		Вычисляет количество дней, месяцев или лет между двумя датами...
3	ДАТАЗНАЧ		Преобразует дату из текстового формата в числовой.
4	ДЕНЬ		Преобразует дату в числовом формате в день месяца.
5	ДНИ		Возвращает количество дней между двумя датами.
6	ДНЕЙ360		Вычисляет количество дней между двумя датами на основе 360-д...
7	ДАТАМЕС		Возвращает дату в числовом формате, отстоящую на заданное чис...
8	КОНМЕСЯЦА		Возвращает дату в числовом формате для последнего дня месяца,...
9	ЧАС		Преобразует дату в числовом формате в часы.
10	НОМНЕДЕЛИ.ISO		Возвращает номер недели по ISO для заданной даты.
11	МИНУТЫ		Преобразует дату в числовом формате в минуты.
12	МЕСЯЦ		Преобразует дату в числовом формате в месяцы.
13	ЧИСТРАБДНИ		Возвращает количество полных рабочих дней между двумя датами.
14	ЧИСТРАБДНИ.МЕЖД		Возвращает количество полных рабочих дней в интервале между ...
15	ТДАТА		Возвращает текущую дату и время в числовом формате.
16	СЕКУНДЫ		Преобразует дату в числовом формате в секунды.
17	ВРЕМЯ		Возвращает заданное время в числовом формате.
18	ВРЕМЗНАЧ		Преобразует время из текстового формата в числовой.
19	СЕГОДНЯ		Возвращает текущую дату в числовом формате.
20	ДЕНЬНЕД		Преобразует дату в числовом формате в день недели.
21	НОМНЕДЕЛИ		Преобразует дату в числовом формате в число, которое указывает...
22	РАБДЕНЬ		Возвращает дату в числовом формате, отстоящую вперед или наза...
23	РАБДЕНЬ.МЕЖД		Возвращает числовое значение даты, предшествующей заданном...
24	ГОД		Преобразует дату в числовом формате в год.
25	ДОЛЯГОДА		Возвращает долю года, которую составляет количество дней межд...

Параметры запроса

СВОЙСТВА

Имя
Table 4

Все свойства

ПРИМЕНЕННЫЕ ШАГИ

Источник

Навигация

Измененный тип