



Качество данных

Визуализатор позволяет производить комплексную оценку качества данных для каждого поля.

Интерфейс

Операции

-  **Сводка** — происходит выбор полей для анализа данных, отображаются общие показатели качества данных и обнаруженные проблемы в полях данных.
-  **Дискретные** — обзор качества данных для дискретных полей.
-  **Непрерывные** — обзор качества данных для непрерывных полей.
-  **Показатели** — выбор и настройка показателей:
 -  Выбор показателей — выбираются показатели, которые требуются пользователю для оценки качества данных.
 -  Настройки показателей — выбор метода идентификации, настройка параметров метода идентификации, настройка максимального количества пропусков.
-  **Абсолютные или относительные значения** — переключает режим отображения количественных значений с абсолютного на процентный.
-  **Экспорт** — позволяет экспортировать файл с отчетом. Экспортируется только текущая таблица (Сводка, Дискретные, Непрерывные). В данном пункте меню будет отображаться название того формата, который выбран для экспорта:
 - Экспорт в XLS(html);
 - Экспорт в HTML.
-  **Детализация** — показывает входной набор данных, отфильтрованный по заданному показателю для выбранного поля. Например, при выборе ячейки *Пропуски* будут показаны все записи входного набора, которые в выбранном поле имеют пропуски. Детализация доступна только для таблиц *Дискретные* и *Непрерывные*.
-  **Гистограмма** — отображает панель с гистограммой по выбранному полю. Гистограмма доступна только для таблиц *Дискретные* и *Непрерывные*.

Характеристики набора данных

- **Метод определения нетипичных значений** — показывает метод идентификации. По умолчанию устанавливается метод *Стандартное отклонение*
- **Столбцов** — общее количество столбцов в исследуемом наборе.
- **Строк** — общее количество записей в исследуемом наборе.

- **Заполненных полей** — количество полей, которые не содержат пропущенные значения, в процентном соотношении.
- **Полных записей** — количество записей, которые не содержат пропущенные значения, в процентном соотношении.
- **Пригодных столбцов** — количество пригодных столбцов. Столбец считается непригодным, если:
 - процент *Пропусков* выше показателя *Максимальный процент пропусков* (данный показатель задаётся в *Настройках показателей*);
 - *Индекс качества* у дискретного поля равен 0 (поле содержит константу).
- **Индекс RPV** — отношение количества записей к количеству столбцов.

Настройка таблиц

При открытии визуализатора *Качество данных* отображается таблица *Сводка* со списком полей. С помощью флагов можно выбирать поля, которые понадобятся пользователю для обработки. Таблицы *Дискретные* и *Непрерывные* в этот момент не доступны.

Сводка Дискретные Непрерывные				
№	☐	Тип	Метка	Вид
0	☒	ab	Код магазина	
1	☐	ab	Название магазина	
2	☒	31	Дата (месяц)	
3	☐	31	Дата (день)	
4	☒	ab	Тип дня недели	
5	☐	12	Количество покуп...	
6	☒	9.0	Выручка магазина	
7	☒	9.0	Средний чек	

Рисунок 1. Выбор полей для обработки

Доступные показатели:

Показатель	Вид данных	Значение
Индекс качества		Отражает разнообразие значений в поле
<u>Гистограмма</u>		Предпросмотр гистограммы

Показатель	Вид данных	Значение
Экстремальные		Количество записей, в которых значение является экстремальным
Пустые		Количество записей с пустыми значениями
Пробелы в конце		Количество записей со значениями, которые оканчиваются пробельным символом (показатель не включает в себя, записи, состоящие только из пробельных символов)
Диапазон значений		Диапазон значений в формате: <i>первое ... последнее значение</i> в отсортированном списке значений
Нулевые		Количество записей, в которых значения равны нулю
<u>Диаграмма размаха</u>		Считает показатели: <i>Наименьший выброс, Нижний квартиль, Медиана, Верхний квартиль, Наибольший выброс, Наблюдаемый максимум</i>
Минимум		Минимальное значение
Среднее		Среднее значение
<u>Мода</u>		Наиболее часто встречающееся значение ¹
Пропуски		Количество записей, в которых отсутствует значение для данного поля
Значения		Отображает диаграмму, в которой показано разнообразие значений по строкам
Выбросы		Количество записей, в которых значение является выбросом
Пробельные		Количество записей со значениями, состоящими только из пробельных символов
Длины строк		Диапазон длин строк в формате: <i>минимальная длина - максимальная длина</i>
Отрицательные		Количество записей с отрицательными значениями
Бесконечности		Количество записей, в которых значения равны бесконечности
Уникальные		Количество уникальных значений

Показатель	Вид данных	Значение
Максимум		Максимальное значение
<u>Медиана</u>	 	Медиана для данного поля
Монотонность	 	Показывает однообразие значений поля (<i>Убывающая, Возрастающая, Не монотонная</i>)

¹. Если в поле значений моды несколько, то показывается первое из этих значений, и отображается  иконка мультимодальности.↵

Настройки показателей

В *Настройках показателей* выбирается Метод идентификации и его параметры.

Показатели

Выбор показателей

Настройки показателей

Выбросы и экстремальные значения

Метод идентификации:

☒ Стандартное отклонение
☐ Интерквартильная ширина

Стандартное отклонение

Выбросы:

3

Экстремальные значения:

5

Интерквартильная ширина

Выбросы:

1,5

Экстремальные значения:

3

Максимальный процент пропусков:

50

Применить

Отменить

Рисунок 2. Настройки показателей

После выбора полей и показателей для анализа качества данных нужно активировать кнопку  *Рассчитать статистики*.

Теперь становятся доступными таблицы **Дискретные** и **Непрерывные**.

При нажатии на кнопку  можно сортировать поля по выбранному показателю, а также убрать/добавить столбцы с показателями.

Экстремальные	▼	Выбросы	Отрицательные	Нулевые	Максим...
0.20%		Сортировать по возрастанию		0.00%	125
0.42%		Сортировать по убыванию		0.00%	592376,1
0.51%		Сбросить сортировку		0.00%	29037,6
		 Столбцы >		☒ Пропуски ☒ Гистограмма ☒ Экстремальные ☒ Выбросы ☒ Отрицательные ☒ Нулевые ☐ Бесконечности ☐ Диаграмма размаха ☐ Минимум ☒ Максимум ☒ Среднее ☒ Медиана ☒ Мода ☒ Монотонность	

Рисунок 3. Сортировка и удаление/добавление столбцов

При выборе ячейки и нажатии кнопки  Детализация внизу открывается окно со всеми записями из этой ячейки.

#	12 CODE	ab OUTLOOK	9.0 TEMP	9.0 HUMIDI...	0/1 WIN...	ab CLASS	12 CLASS_NUM
1	1	sunny	75	70	Истина	Play	1
2	2	sunny	80	90	Истина	Don't Play	0
3	3	sunny	85	85	Ложь	Don't Play	0
4	4	sunny	72	95	Ложь	Don't Play	0
5	5	sunny	69	70	Ложь	Play	1
6	6	overcast	72	90	Истина	Play	1
7	7	overcast	83	78	Ложь	Play	1
14	8	overcast	64	65	Истина	Play	1

Рисунок 4. Детализация ячейки

При выборе ячейки и нажатии на кнопку  Гистограмма в правой части интерфейса открывается окно с более детальной гистограммой. Для дискретных полей можно выбирать значения для отображения:

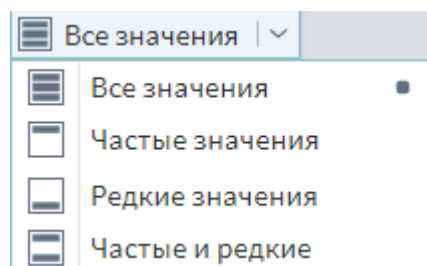


Рисунок 5. Возможные значения

Проблемы

В колонке *Проблемы* отображается общее значение проблемных строк по данному полю.

Виды проблем

- **Пропуски** — количество записей, содержащих пропущенные значения.
- **Пустые** — количество записей, содержащих пустые значения.
- **Нули** — количество записей, содержащих нулевые значения.
- **Пробелы** — количество записей, содержащих только пробельные символы (пробел, неразрывный пробел, табуляция и т.д.)
- **Бесконечности** — количество записей, содержащих значение *Бесконечность*.
- **Константа** — указывает, что поле содержит только одно уникальное значение, при этом в поле могут быть пропуски.
- **Уникальное** — указывает, что каждое значение в этом поле является уникальным.
- **Ошибка конвертации** — появляется, если поле содержит значения, которые не могут быть приведены к типу данных поля.
- **Не вычислено** — отмечает поля, которые не содержат рассчитанных показателей.
- **Нет данных** — проблема проявляется при анализе пустого набора данных.
- **Экстремальные** — количество записей, в которых значение является экстремальным.
- **Выбросы** — количество записей, в которых значение является выбросом.
- **Отрицательные** — количество записей с отрицательными значениями.
- **Пробелы в конце** — количество записей со значениями, которые оканчиваются пробельным символом (показатель не включает в себя, записи, состоящие только из пробельных символов).

Рядом со значением *Вида проблем* указывается процентное соотношение проблемных строк к общему количеству строк, а в скобках количество строк с такой проблемой.

Проблемы	Виды проблем
100,00%	Пустые - 100,00% (149) Константа
100,00%	Не вычислено Ошибка конвертации
68,46%	Пропуски - 44,97% (67) Выбросы - 1,34% (2) Нули - 22,15% (33)
63,76%	Пропуски - 61,74% (92) Пробелы - 2,01% (3)
63,09%	Пробелы - 51,01% (76) Пробелы в конце - 12,08% (18)
22,15%	Пропуски - 14,09% (21) Нули - 8,05% (12)
19,46%	Пустые - 9,40% (14) Пробелы - 10,07% (15)
10,74%	Экстремальные - 5,37% (8) Бесконечности - 5,37% (8)
6,71%	Экстремальные - 3,36% (5) Пробелы - 3,36% (5)
1,34%	Экстремальные - 0,67% (1) Выбросы - 0,67% (1)
0,00%	Уникальное
0,00%	Константа

Рисунок 6. Виды проблем