

Заполнение пропусков

Компонент предназначен для автоматического заполнения пропущенных значений в наборах данных. Для каждого столбца исходного набора данных пользователь может выбрать наиболее подходящий метод заполнения пропусков. Пропусками считаются Null-значения.

Примечание: узел не обрабатывает поля с переменным типом данных (см. [типы данных](#)).

Порты

Вход

-  Входной источник данных (таблица данных).

Выход

-  Выходной набор данных (таблица данных).

Мастер настройки

Настройка соответствия между столбцами проводится в интерфейсах [таблица](#) или [связи](#).

Настройка заполнения пропусков

- **Исходные данные упорядочены** — установку данного флага следует производить в том случае, когда известно, что данные являются упорядоченными. Например, временной или иной ряд, значения которого упорядочены по возрастанию или убыванию (например, по дате или времени). Для упорядоченных и неупорядоченных данных могут применяться различные методы заполнения пропусков.
- **Допустимый процент пропусков** — задает максимальный процент пропусков в поле, при котором узел еще будет их заполнять. Если фактическая доля пропусков в поле больше указанного значения, узел пропускает обработку этого поля и оставляет все пропущенные значения как есть. Например, если значение настройки **Допустимый процент пропусков**

указано 55%, узел будет заполнять пропуски в тех полях, где их 55% или менее. Если в каком-то поле пропущено 70% значений, система оставит его без изменений.

Крайние значения:

- **0%** — пропуски будут заполняться только если пропущенных значений в указанном столбце строго 0%. Поскольку заполнять там нечего, фактически массового заполнения не произойдет.
- **100%** — пропуски будут заполняться во всех полях, где есть хотя бы одно значение. Если поле полностью пустое (100% пропусков), оно не будет заполнено, так как невозможно вычислить значение для замены.

Примечание: процент пропусков округляется до целых чисел. Например, если в поле из 120 строк пропущено 1 значение (0,83%), алгоритм округлит это до 0%, но заполнит пропуск, даже если параметр установлен в 0%.

- **Random seed** — начальное число (целое, положительное), которое используется для инициализации генератора псевдослучайных чисел. Последовательность чисел генератора полностью определяется начальным числом. Если генератор повторно инициализируется с тем же начальным числом, он выдаст ту же последовательность чисел.

Параметр влияет на порядок замены случайными значениями. Можно повторить результат выполнения узла, если подать те же данные и выставить тот же random seed.

Для параметра доступны следующие команды:

- Всегда случайно — начальное число всегда будет случайным.
- Генерировать — сгенерируется новое начальное число.
- Копировать — в буфер обмена будет скопировано указанное значение.
- **Область настройки методов обработки пропусков** — содержит список полей, доступных для обработки, с указанием вида данных. Для каждого поля можно установить флаг, задающий необходимость обработки, и затем выбрать метод заполнения пропусков.

Доступны следующие методы обработки:

- **Заменять предыдущим значением** — выявленные пропуски заменяются предыдущим известным (непустым) значением из того же столбца. Если в начале набора данных встречаются пропуски, они остаются незаполненными до первого непустого значения.
- **Заменять средним** — выявленные пропуски заменяются средним значением столбца.
- **Заменять медианой** — выявленные пропуски заменяются медианой, вычисленной по столбцу.
- **Заменять наиболее вероятным** — выявленные пропуски заменяются наиболее вероятным значением по столбцу, замена производится на среднее значение из наиболее вероятного интервала. Число интервалов варьируется в зависимости от объема выборки — чем она больше, тем больше интервалов. Метод обработки вычисляется по-разному в зависимости от вида данных:
 - Для дискретных данных пропуски заполняются наиболее часто встречающимся значением. Если таких значений несколько, выбирается последнее среди значений с одинаковой максимальной частотой.
 - Для непрерывных данных пропуски заполняются средним значением из наиболее частого интервала. Если таких интервалов несколько, выбирается первый интервал среди

интервалов с одинаковой максимальной частотой.

- **Заменять на 0** — выявленные пропуски заменяются значением 0.
- **Заменять случайными значениями** — выявленные пропуски заменяются случайными значениями, которые формируются из диапазона от наименьшего до наибольшего значения столбца.
- **Линейная интерполяция** — выявленные пропуски заменяются на вычисленные промежуточные значения линейной функции, полученной по известным значениям (условно проведя прямую между ними).
- **Кубическая интерполяция** — выявленные пропуски заменяются на промежуточные значения кубического сплайна (сплайны третьей степени с непрерывной первой производной), полученного по известным значениям.
- **Сплайн-интерполяция** — выявленные пропуски заменяются на промежуточные значения сплайн-функции, полученной по известным значениям.
- **Оставить без изменения** — выявленные пропуски заполняться не будут.
- **Удалять записи** — строки с выявленными пропусками исключаются из выходного набора данных.
- **Заменять на "Не задано"** — выявленные пропуски заменяются значением "Не задано" или на указанное пользователем значение, которое можно задать, нажав на кнопку

Для каждого поля спектр доступных методов определяется тремя характеристиками данных одновременно (см. [данные](#)):

- Упорядоченностью
- Типом
- Видом

Таблица применимости по этим характеристикам:

Метод	Неупорядоченный набор		Упорядоченный набор	
	 Дискретны й	 Непрерывны й	 Дискретны й	 Непрерывны й
Заменять предыдущим значением	0/1 31 90 12 ab	0/1 31 90 12 ab	0/1 31 90 12 ab	0/1 31 90 12 ab
Заменять средним	31	31 90 12	31	31 90 12
Заменять медианой	31 90 12	31 90 12	31 90 12	31 90 12
Заменять наиболее вероятным	0/1 31 90 12 ab	31 90 12	0/1 31 90 12 ab	31 90 12
Заменять на 0	90 12	90 12	90 12	90 12
Заменять случайными значениями	0/1 31 90 12 ab	31 90 12	0/1 31 90 12 ab	31 90 12
Линейная интерполяция			31 90 12	31 90 12

Кубическая интерполяция			$\overline{31} \ 9.0 \ 12$	$\overline{31} \ 9.0 \ 12$
Слайн-интерполяция			$\overline{31} \ 9.0 \ 12$	$\overline{31} \ 9.0 \ 12$
Оставить без изменения	$0/1 \ \overline{31} \ 9.0 \ 12$ ab	$\overline{31} \ 9.0 \ 12$		
Удалять записи	$0/1 \ \overline{31} \ 9.0 \ 12$ ab	$\overline{31} \ 9.0 \ 12$		
Заменять значением «Не задано»	ab		ab	

Смотри также:

- [Редактирование выбросов](#)