



Чем вызваны ошибки в данных

Никто не вносит ошибки в набор данных умышленно. Пропуски, дубликаты и неверные значения всплывают не просто так, а по каким-то причинам — их надо устанавливать и разбираться.

Ошибки делятся на две большие группы: те, что возникли по технологическим причинам, и вызванные человеческим фактором.

Технологические ошибки возникают при копировании, считывании, смене формата. Сбой может произойти даже во время записи данных.

Например, изучая *try-except*, вы столкнулись с технологическими ошибками:

```
position = [ ['2019-05-01', '- 6'], ['2019-05-02', '+5'], ['2019-05-03', '5'], ['2019-05-04', '4'], ['2019-05-05', '5'], ['2019-05-06', '5'], ['2019-05-07', '4'], ['2019-05-08', 'Error 5'], ['2019-05-09', '3'], ['2019-05-10', '3'], ]
```

Python ▾

Люди ошибаются и делают это постоянно. Дайте человеку возможность ввести данные, и он обязательно даст промашку: опечатается в тексте, укажет другой порядок величин, перепутает числа или категориальные значения. Эти ошибки вызваны **человеческим фактором**.

Не укажете, как именно заполнять поля в форме регистрации, получите ошибки: пустое значение или буквы вместо возраста, а телефон — вместо email.

Увидели пропуск — подумайте, нормально ли это. Сколько вообще пропусков может быть в этом столбце? К примеру, в списке с электронными адресами пользователей, согласных на рассылку, будет много пропусков. Далеко не все предоставляют email.

Если обнаружили проблему, надо точно понимать, кому именно о ней сообщать.

Чтобы ошибки не повторялись, утвердите регламент выгрузки данных с коллегами. Все следуют договорённостям, а проблемы всё же возникают? Значит, причина не в «кривой» выгрузке. Скорее всего, вы столкнулись с технологической ошибкой.