Конспект

Ручной поиск дубликатов

Часто при анализе данных возникают дубликаты. Если их не найти, то анализ данных может привести к некорректным результатам.

Дубликаты можно искать двумя способами.

Способ 1. Ранее вы уже знакомились с методом `duplicated()`. В сочетании с методом `sum()` он возвращает количество дубликатов. Если выполнить метод `duplicated()` без суммирования, на экране будут отображены все строки. Там, где есть дубликаты, будет значение `True`, где дубликата нет — `False`.

Способ 2. Вызвать метод `value_counts()`, возвращающий уникальные значения с их частотой. Его применяют к объекту `Series`. Результат работы метода — список пар «значение-частота», отсортированный по убыванию. Интересующие вас дубликаты будут в начале списка.

Ручной поиск дубликатов с учётом регистра

Дубликаты в строковых данных требуют особого внимания, поскольку регистр имеет значение: заглавная `'A'` и строчная `'a'` с точки зрения Python — разные символы, но имеют одинаковое значение — буква А.

Чтобы учесть такие дубликаты, все символы в строке приводят к нижнему регистру вызовом метода `lower()`.

В pandas символы приводят к нижнему регистру методом с похожим синтаксисом: `str.lower()`.