# 2. GIGO

Когда вы приступаете к работе над аналитической задачей, должны быть готовы не только вы (чтобы придумать, как найти зависимости или красиво визуализировать данные), но и сами данные.

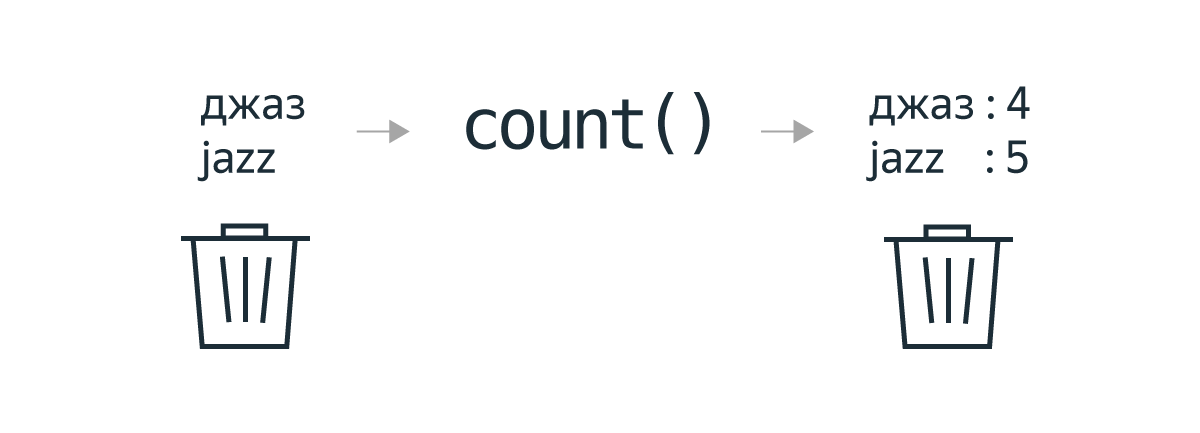
Процесс подготовки данных для дальнейшего анализа называется **предобработка**. Заключается она в поиске проблем, которые могут быть в данных, и в решении этих проблем.

В информатике работает принцип GIGO (от англ. garbage in — garbage out, буквально «мусор на входе — мусор на выходе»). Это значит, что при ошибках во входных данных даже правильный алгоритм работы выдаёт неверные результаты.

Посмотрите на этот срез данных для нашей задачи:



Что сразу же бросается в глаза? В столбце genre для одного жанра есть два названия: джаз и jazz. Если принять всё как есть, подсчёт прослушанных джазовых композиций даст два ответа — для жанра джаз и для жанра jazz, которые представляют одну и ту же сущность. Это помешает сделать правильные выводы об интересе пользователей к джазу, и может повлечь неверные решения.



Нам предстоит очистить данные от мусора, чтобы эта аббревиатура расшифровывалась более приятно: *good in — good out*, т.е. «хорошо на входе — хорошо на выходе».

Подвох может крыться не только в содержании столбца, но даже в его названии. Вы уже умеете получать данные из столбцов таблицы, чтобы, к примеру, посчитать количество уникальных пользователей. Попробуем получить доступ к столбцу «user\_id», как учили:

print(df['user\_id'])

-----------------------------------------------------

KeyError Traceback (most recent call last)

/usr/local/lib/python3.6/dist-packages/pandas/core/indexes/base.py in get\_loc(self, key, method, tolerance)

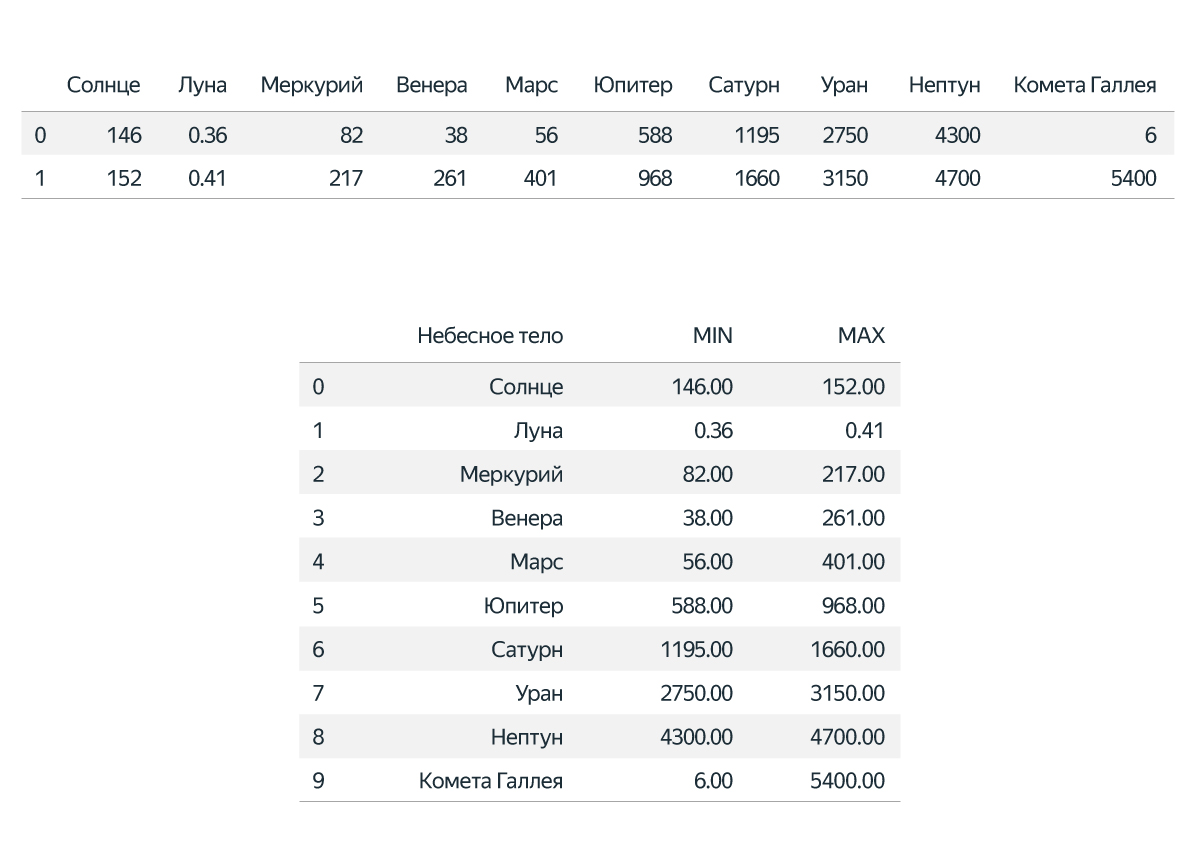
На первый взгляд кажется, что мы ввели имя столбца правильно, но при тщательном изучении можно заметить пробелы в начале. Обычно их порождает человеческий фактор — при именовании нажимают клавишу пробела или копируют лишние символы из текста технического задания. Ещё вариант: при выгрузке данных из какой-нибудь информационной системы заголовки сформировались вот так плохо. Каждое название бывает несчастливо по-своему — в одном столбце всё нормально, в другом 5 пробелов в начале и 1 в конце, в третьем 3 пробела в начале.

Трудно определить количество пробелов на глаз, чтобы задать точное правило их ликвидации, но ещё более сумасбродное решение — всякий раз копировать значимые символы названия столбца вручную. Это серьёзная проблема; избавляться от лишних пробелов учатся в первую очередь.

Вообще сложности с данными бывают двух видов:

* данные содержат мусор;
* данные корректны, но представлены в такой форме, что малопригодны для анализа.

Например, таблица с данными о расстоянии от Земли до небесных тел (в миллионах километров) может быть представлена в двух вариантах:



Как вы думаете, какой вариант лучше подходит для анализа?