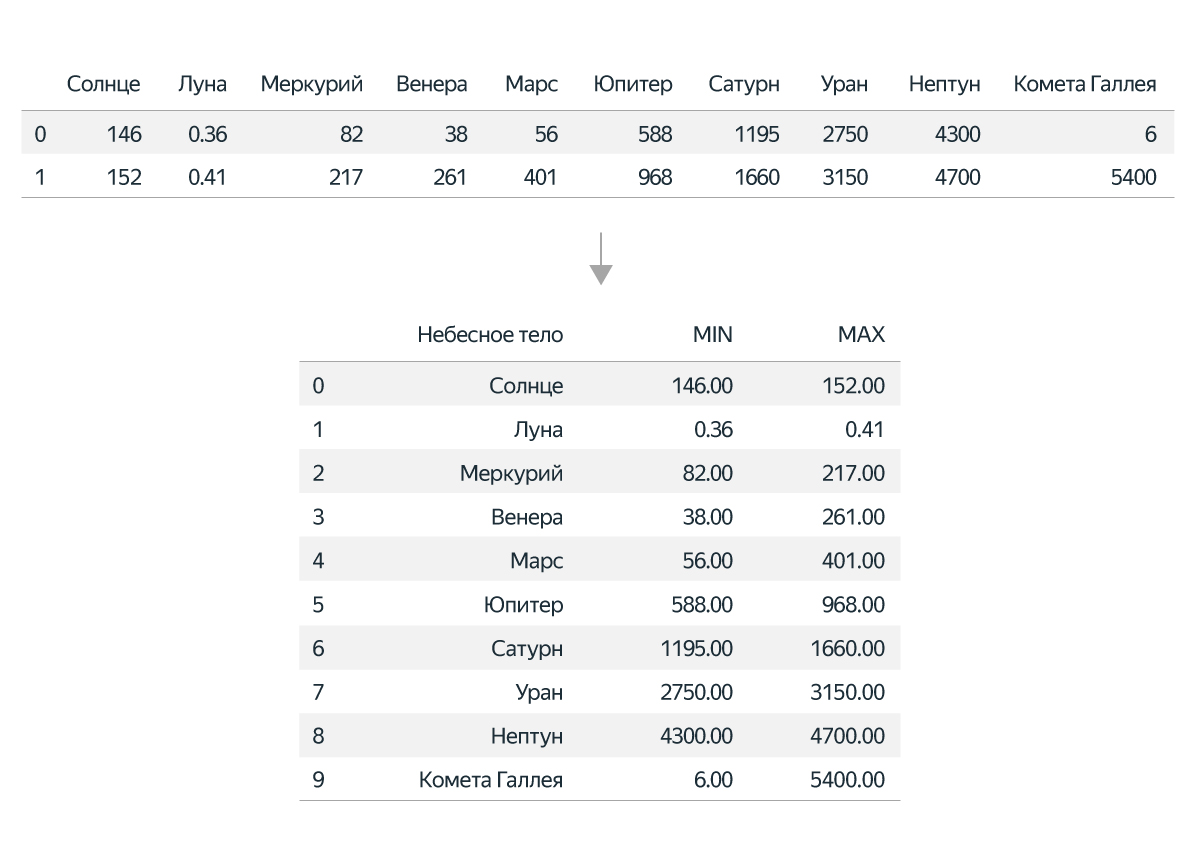
## 3. Охота на мусор



В верхней таблице не разобрать, какая строка за что отвечает, тогда как в нижней всё более или менее ясно. Лучше, если в каждом столбце хранятся значения одной переменной, а каждая строка содержит одно наблюдение, к которому привязаны значения разных переменных.

Строки нижней таблицы можно представить как наблюдения за одним небесным телом: в переменных сохраним его название, минимальное и максимальное расстояния до Земли.

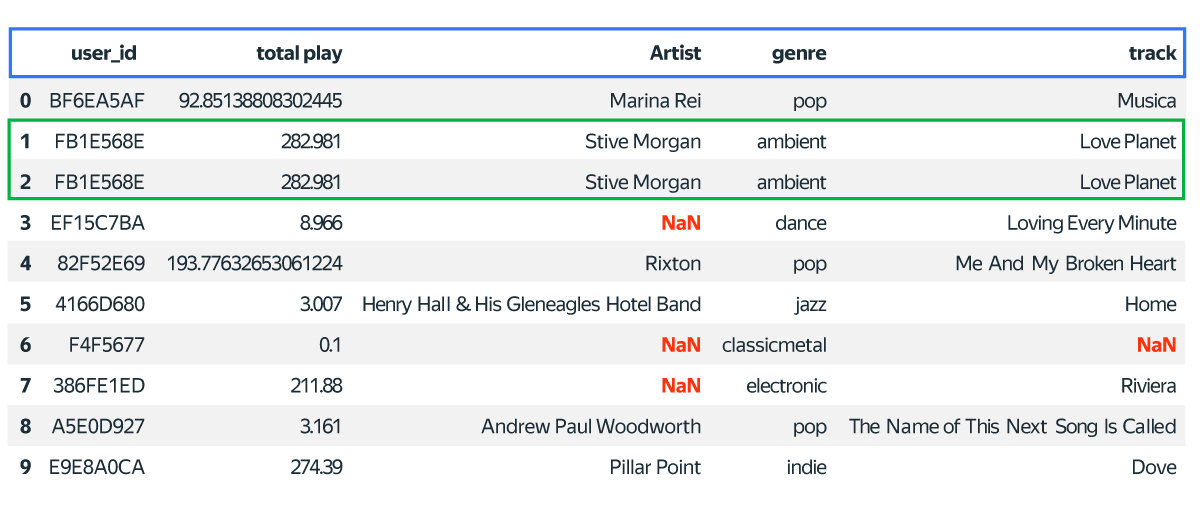
Верхняя таблица не совсем безнадёжна — такой вариант может быть представлен в качестве визуального отчета. Международные организации часто публикуют данные своих ежегодников в подобной форме.

С точки зрения организации данных менеджер Яндекс Музыки дал нам хорошую таблицу, но мусор в ней всё-таки есть. Понадобится выполнить ряд операций.

Посмотрите на первые 10 строк. Какие необычные явления вы можете заметить?



Проблемы с названиями столбцов уже известны. Теперь обратите внимание на повторяющиеся значения в столбце Artist: какие-то NaN, и это точно не музыкальная группа. А ещё нужно разобраться со строками 1 и 2. Там полный повтор.



Имейте в виду, что аналитик не только «чистит» данные, но и сообщает команде о проблемах, чтобы их причины были установлены. Каждый сорт мусора появляется в таблице не просто так.

В этой теме мы раскроем базовые механизмы борьбы с часто встречаемыми проблемами. Из них самые очевидные, но в то же время самые опасные:

• некорректное именование столбцов  
• дублирование значений  
• отсутствующие значения (NaN)

Механизмы борьбы с ними — основа, с которой вы можете начать своё развитие как специалист по предобработке данных.

Для начала вызовем метод info(), чтобы просмотреть сводку по всему набору данных.