## 2. С чего начинается исследование

Идею объединения сервисов Музыка и Радио тестировали на небольшой группе пользователей. Результаты сведены в csv-файл, который вам предстоит изучить. Итог анализа таких данных — это **метрики**: величины, значения которых отражают пользовательские впечатления. Одна из важнейших — happiness (англ. happiness, «счастье»). Здесь это среднее время, которое пользователь слушает музыку в течение выбранного периода времени (в нашей задаче — за сутки). Чем дольше пользователь слушает музыку, тем он довольнее. Ваша задача: найти значение happiness и посмотреть, как оно поменялось с прошлого эксперимента.

В анализе данных важно наглядное представление результатов, чтобы их мог оценить заказчик. Это бизнес, где на кону громадный трафик, серверные мощности, личное время многотысячной аудитории. И одновременно это исследование. Как в настоящей экспериментальной науке, ответ непредсказуем. Его точность зависит от умелого владения статистическими методами и от качества исходных данных. Перед тем, как начинать считать, проверьте, грамотно ли ваши данные подготовлены.

**Ознакомление с данными: в предыдущих сериях.**

Перед тем, как браться за статистику, нужно:

**1. Прочесть исходный файл и превратить его в структуру данных**

**2. Посмотреть на данные**

**3. Оценить качество предподготовки**

**1. Прочесть исходный файл и превратить его в структуру данных**

К заданию прилагается файл в формате *csv*, где все значения разделены запятыми. Это наши исходные данные. Чтобы применить к ним все возможности языка Python и библиотеки Pandas, надо импортировать эту библиотеку и сохранить её в переменной. По сокращённому названию панельных данных (**p**anel **d**ata), с которых начиналась Pandas, эту переменную принято называть *pd*:

import pandas as pd

Для чтения csv-файла в библиотеке Pandas есть готовая функция — метод read\_csv(). Как и все методы, он вызывается записью через точку после имени своего объекта. В скобках указывается аргумент (параметр) метода. У read\_csv() это имя файла с данными. Прочтение превращает файл в структуру данных DataFrame. Имя переменной, в которой эта структура данных сохраняется, чаще всего df либо отражает тематику данных:

df = pd.read\_csv('music\_log.csv')

**2. Посмотреть на данные**

Вывести на экран таблицу и оценить данные:

print(df)

Как правило, таблица очень велика. Практичнее запросить определённое количество первых строк, методом head():

print(df.head(15)) *# выведет первые 15 строк таблицы*

**3. Оценить качество предподготовки**

Нужно убедиться в том, что данные прошли предподготовку. По крайней мере, не должно быть пропусков и повторов. Пропущенные и неопределённые значения выявляет метод isna(), а суммарное количество таких значений — метод sum(). Обратите внимание: мы записали вызов обоих методов в одну строку, разделив их точкой. Python сначала вызовет метод isna(), а затем результаты его работы передаст методу sum().

print(df.isna().sum())

Повторяющиеся строки — дубликаты — выявляются методом duplicated() и подсчитываются тем же sum():

print(df.duplicated().sum())