# Лабораторная работа № 2

## **Тема**: Предобработка данных.

Для выполнения данной лабораторной работы можно использовать любой найденный датасет с пропусками, а если пропусков нет, то необходимо их сгенерировать. Также в датасете помимо числовых данных должны присутствовать текстовые, для дальнейшей замены их на числа. Если ваш датасет из первой лабораторной соответствует этим критериям, то можете использовать его.

Вот пара ссылок на датасеты с пропусками:

https://www.kaggle.com/datasets/ander289386/cars-germany
https://www.kaggle.com/datasets/arnabchaki/fitness-trackers-products-ecommerce

Пример предобработки данных можно посмотреть в файле *Titanic_missing_data.html* в папке *Examples*.

**Задание**:

1. Выявите пропуски данных несколькими способами (визуальный, расчетный…)

*При удалении (замене) пропусков необходимо рассуждать: можно ли удалить данный параметр и чем целесообразно заменять пропуски данных в конкретных параметрах, руководствуясь описанием параметров датасета и предметной областью.*

2. Исключите строки и столбцы с наибольшим количеством пропусков.
3. Произведите замену оставшихся пропусков на логически обоснованные значения.
4. Постройте гистограмму распределения исходного датасета до и после обработки пропусков. Сделайте выводы как обработка данных повлияла на их распределение.
5. Проверьте датасет на наличие выбросов, удалите найденные аномальные записи.
6. Приведите все параметры к числовому виду (кодирование текстовых данных).
7. Сохраните обработанный датасет.

**Вопросы**:

1. Перечислите различные способы обнаружения пропущенных данных.
2. Как можно определить тип данных каждого признака?
3. Приведите пример категориальных данных.
4. Какими способами можно закодировать категориальные данные?
5. Как работает One-Hot Encoding?
6. Какие еще встречаются ошибки данных помимо пропусков и выбросов?