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*.

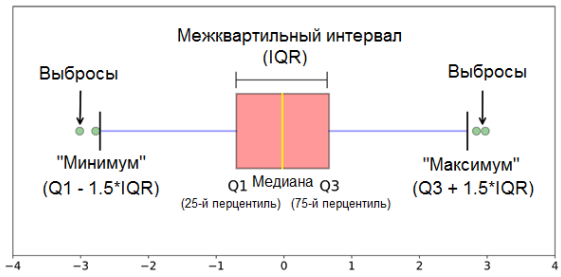
Задание:

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. Проверьте датасет на наличие выбросов, удалите найденные аномальные записи.

Выбросы можно обнаружить разными способами, на рисунке приведен самый наглядный из них (box-plot) →



Предобработка данных проводится для дальнейшей работы с этими данными в библиотеке машинного обучения SciKit Learn. Методы этой библиотеки не работают с текстовыми данными, поэтому текстовые данные необходимо проанализировать и:

- *если они являются неважными, то их удаляют;*
- *если они важны, то их приводят к числовому виду* ↓

5. Приведите категориальные параметры к числовому виду (это могут быть названия стран, дни недели, марки машин, другие признаки, записанные текстом).
   Методы LabelEncoder и OneHotEncoder из библиотеки машинного обучение SciKit Learn. Можно почитать пример здесь https://habr.com/ru/articles/456294/
6. Проведите нормализацию данных.
7. Сохраните обработанный датасет.

В итоге должен получиться набор данных, содержащий только числовые нормализованные данные.

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. Как работает нормализация данных?
7. Какие еще встречаются ошибки данных помимо пропусков и выбросов?