Для выполнения проекта вам необходимо создать конвейер очистки данных (data cleaning), используя как Python, так и R. Конвейр должен работать на любой таблице с количественными и качественными переменными.

Основное задание (20 баллов) состоит в проведении следующих операций:

- *Удаление дубликатов.* Дубликаты являются одной из самых распространенных проблем с качеством данных. Они могут привести к искажению результатов анализа, поэтому важно удалить их из набора данных. Необходимо удалить дублирующие друг друга записи.
- *Обработка пропущенных значений.* Пропущенные значения могут возникнуть по многим причинам, например, из-за ошибок ввода данных, потери данных или ошибок в программном обеспечении. Для корректного анализа данных необходимо определить и обработать пропущенные значения. Более конкретно, если количество пропущенных значений превышает 25%, то соответствующий столчбец данных требуется удалить; иначе для количетственных данных пропущенные значения нужно заменить на среднее арифметическое, а для качественных - на моду.
- *Обработка выбросов.* Выбросы - это значения, которые сильно отличаются от остальных значений в наборе данных. Они могут возникать из-за ошибок измерения или ввода данных, их необходимо обнаруживать и обрабатывать, чтобы они не искажали результаты анализа. Для избавления от выбросов используйте следующий метод - считайте выбросом всё, что лежит дальше одного межквартильного расстояния (IQR) от медианы. Каждый выброс необходимо заменить на минимальное/максимальное значение не являющееся выбросом.
- *Сохранение* обработанных данных в указанной пользователем локации.

Методы требуется протестировать (с демонстрацией) на самостоятельно найденных данных.

В качестве источников можно использовать, например:

https://www.kaggle.com/

https://archive.ics.uci.edu/ml/index.php

Дополнительное задание #1 (10 баллов)

Оформите ваш обработчик в виде функции, которому передаётся путь до таблицы и добавьте для него настройки:

- Позвольте изменять после какого порога мы выбрасываем столбец с большим количеством пропущенных значений (по-умолчанию 25%).
- Позвольте настраивать на каком расстоянии от медианы данные начинают считаться выбросами.
- Придумайте и добавьте ещё один способ избавления от выбросов.

Дополнительное задание #2 (10 баллов)

Финальным шагом является создание визуализаций сравнения исходных и очищенных данных. Вы можете использовать любые библиотеки для создания

визуализаций. В качестве примеров приведём Matplotlib для Python и ggplot2 для R.

- Конкретный тип графиков остаётся на ваше усмотрение.
- Один график для одного столбца данных.
- Графики должны демонстрировать то, как повлияла на значения столбцов очистка данных.

В качестве простого примера, на одном графике можно изобразить две гистограммы - до и после очисти данных.