**Задание 1. Исследование набора данных**

В ходе выполнения данного задания выбирается набор данных, который будет использоваться в дальнейшем при исследовании алгоритмов кластеризации и классификации.

Любая задача анализа данных предполагает предобработку и статистический анализ данных, это позволяет в дальнейшем выбрать подходящую модель и алгоритм анализа данных.

Задание состоит из последовательного выполнения следующих подзадач:

1. Создать Jupyter Notebook, переименовать его «Lab 1, № Группы, ФИО»

1. Выбор исследуемого датасета.

Возможные датасеты хорошо представлены тут:

1. <https://github.com/awesomedata/awesome-public-datasets>
2. https://www.kaggle.com/datasets

Рекомендация к выбору датасета: предпочтительнее числовые данные, с разметкой.

2. Для каждого датасета представить краткое его описание в вашем Jupyter Notebook:

* 1. предметная область, источник данных, характер данных (реальные или имитационные)
  2. какие атрибуты представлены в датасете, их тип (числовой, строковый (категории)), что они обозначают
  3. есть ли описание задачи анализа, если есть - представить

3. Для каждого атрибута нужно определить:

1. среднее значение, ско
2. построить гистограмму распределения значений, определить есть ли выбросы
3. есть ли пропущенные значение, сколько
4. предложить вариант обработки пропущенных значений

4. Определить корреляцию между параметрами

1. какие атрибуты высококоррелированы, определить характер корреляции
2. какие атрибуты не имеют корреляцию
3. постройте графики рассеивания (предпочтительнее матрицу графиков рассеивания)
4. проанализируйте полученные результаты.

Отчет должен включать описания выполнения каждой подзадачи.

**Дополнительная литература**

1. Поиск выбросов https://medium.com/swlh/identify-outliers-with-pandas-statsmodels-and-seaborn-2766103bf67c
2. Анализ и визуализация https://www.analyticsvidhya.com/blog/2021/02/an-intuitive-guide-to-visualization-in-python/
3. Хорошее описание корреляции https://stepik.org/lesson/8086/step/1?unit=1365