# Проект 2. Разведывательный анализ данных

## Цель проекта

Выявить параметры будущей модели для предсказания результатов сдачи гос. экзамена по математике в зависимости от условий жизни школьников.

## Краткое описание.

Проект производит разведывательный анализ данных датасета, в котором собраны данные об условиях жизни учащихся и результатах сдачи гос. экзамена по математике.

При анализе данных в исходном датасете находилось 395 строк и 30 столбцов.

Из 30 столбцов в датасете, 17 это строковые данные и 13 столбцов с числовыми данными

В 27 столбцов есть отсутствующие данные (NaN). Столбцы с полностью заполненными данными: ['school', 'sex', 'age'].

Отдельно проводим анализ числовых данных и строковых.

В результате анализа числовых данных были выявлены выбросы в столбцах age, fedu, famrel, absences. Данные с выбросами были исключены из датасета. Также из датасета были исключены данные с неопределенными значениями (NaN) в результатах сдачи гос. экзамена по математике, т.к. данные в ключевом параметре должны трактоваться однозначно.

Некорректно заполненных числовых данных, таких как пустые или нечисловые значения, не выявлено.

В результате корреляционного анализ количественных переменных, из датасета был исключен только один параметр studytime\_granular, т.к. он полностью коррелирует с обратным знаком с другим параметром studytime.

В результате анализа столбцов c номинативными данными, некорректно заполненных данных, таких как пустые или не входящие в список допустимых значений для конкретных столбцов, не выявлено.

Анализ графиков зависимости результата гос. экзамена по математике от номинативных параметров, показал, что все номинативные параметры в большей или меньшей степени влияют на этот результат. Согласно теста Стьюдента, статистически значимые различия распределения результата гос. экзамена по математике найдены для следующих столбцов:

['sex', 'address', 'mjob', 'paid', 'higher', 'romantic']

## Итог проекта.

Итоговый датасет для дальнейшего построения модели включает следующие данные:

Data columns (total 18 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 age 384 non-null int64

1 medu 381 non-null float64

2 fedu 360 non-null float64

3 traveltime 356 non-null float64

4 studytime 377 non-null float64

5 failures 362 non-null float64

6 famrel 357 non-null float64

7 freetime 375 non-null float64

8 goout 377 non-null float64

9 health 369 non-null float64

10 absences 373 non-null float64

11 score 384 non-null float64

12 sex 384 non-null object

13 address 369 non-null object

14 mjob 365 non-null object

15 paid 345 non-null object

16 higher 364 non-null object

17 romantic 353 non-null object

## Улучшение проекта при его развитии.

В процессе анализа приходится выполнять повторяющиеся операции над множеством столбцов, и анализировать часто однотипные данные и графики. Предполагаю, что для этого возможно разработать и использовать специальные функции.