**Основные цели и задачи проекта.**

Отследить влияние условий жизни учащихся в возрасте от 15 до 22 лет на их успеваемость по математике

**Краткая информация о данных.**

1 school — аббревиатура школы, в которой учится ученик

2 sex — пол ученика

3 age — возраст ученика

4 address — тип адреса ученика

5 famsize — размер семьи

6 Pstatus — статус совместного жилья родителей

7 Medu — образование матери

8 Fedu — образование отца

9 Mjob — работа матери

10 Fjob — работа отца

11 reason — причина выбора школы

12 guardian — опекун

13 traveltime — время в пути до школы

14 studytime — время на учёбу помимо школы в неделю

15 failures — количество внеучебных неудач

16 schoolsup — дополнительная образовательная поддержка

17 famsup — семейная образовательная поддержка

18 paid — дополнительные платные занятия по математике

19 activities — дополнительные внеучебные занятия

20 nursery — посещал детский сад

21 higher — хочет получить высшее образование

22 internet — наличие интернета дома

23 romantic — в романтических отношениях

24 famrel — семейные отношения

25 freetime — свободное время после школы

26 goout — проведение времени с друзьями

27 health — текущее состояние здоровья

28 absences — количество пропущенных занятий

29 score — баллы по госэкзамену по математике

**Этапы работы над проектом.**

Первичная обработка данных.

Просмотр распределения признаков для числовых переменных.

Устранение выбросов.

Оценка количества уникальных значений для номинативных переменных.

Преобразование данные

Корреляционный анализ количественных переменных

Отбор некоррелирующих переменных.

Анализ номинативных переменных

Устранение малозначимых номинативных

Формулировка выводов

**Ответы на вопросы саморефлексии.**

Проект выполнен исходя из имеющихся знание и навыков, а хотелось бы попробовать большее и видеть более адекватные результаты.

Хотелось бы лучше знать и понимать методы нормализации и методы замены пропусков, условия их использования.

**Выводы проекта.**

В процессе работы с сетом, дубликатов вне найдено.

При поверхностной проверке содержимого в столбцах fedu и famrel были выявлены некорректные данные. Ошибка исправлена путём замены на моду, так как показатели являются категориальными.

Также на этом же этапе обнаружена колонка studytime granular. Она оставлена для дальнейшего анализа с целью выявления на результирующую графу.

Поиск пропусков не выявил необходимости удаления каких-либо параметров, так как критического содержания пропусков (более 30%) не найдено.

В предобработке в целевом показателе score найдены строки с пропусками, поэтому они удалены.

Показатели, где пропусков чуть более 10% оставлены для оценки дальнейшего влияния на результат.

Пропуски данных в категориальных параметрах заменены на моду, а в числовых – на медиану.

Просмотр гистограмм численных параметров выявил, что лишь некоторые показатели имеют тенденцию к нормальному распределению, такие как freetime, goout и score. Большинство параметров (age, traveltime, studytime, failures, absences) распределены с правосторонней ассиметрией, а показатели medu, studytime granular, famrel - с левосторонней. При этом параметры score и absences имеет явные аномалии.

Анализ выбросов показал, что чистых числовых данных не так много, это – medu, goout, health и score. Количество выбросов по показателям famrel и failures дают право включать их в модель на основании возможности моделирования их выбросов.

Просмотр боксплотов показал, что многие признаки нуждаются в объединении категории (age, medu, fedu, traveltime, freetime, absences), после которого они могут адекватно работать в модели. Показатели failures и famrel могут быть значимы, так имеют разную величину «усов» у категорий. Показатели studytime, studytime granular, goout, наоборот, не эффективны для модели, так как их категории имеют тенденцию к идентичности.

Корреляционный анализ показал, что studytime и studytime granular идентичны, что делает невозможным включения их обоих в модель. Также medu и fedu имеют высокий уровень корреляции (59%), который может возрасти при очистке данных и может поставить вопрос об удалении одного из них и в силу идентичности.

На целевой показатель наиболее всего влияют medu, fedu, age и failures. При этом на успеваемость ребёнка положительно влияет высокий уровень образования родителей, меньшее количество пропусков и небольшой возраст самого ребёнка.

Корреляционный анализ показывает, что большинство данных слабо связаны между собой и нуждаются в обработке. При этом, на корреляционную матрицу не повлияли ни замена пропусков на нулевые значения, ни различные варианты очистки выбросов (при удалении строк не более 10%), коэффициенты выросли не более 2%.

Просмотр диаграмм категориальных нечисловых параметров показал, что визуально выборка делиться полярно наиболее всего по таким признаком, как school, address, pstatus, fjob, schoolsup, nursery, higher, internet.

Просмотр боксплотов категориальных нечисловых параметров позволяет выбрать показатели, которые могут быть значимы для модели. Это – mjob, fjob, schoolsup, higher. Однако формализованный подход (тест Стьюдента) выделил значимыми address, mjob, higher, romantic.

Представленный вариант анализа данных носит промежуточный характер, который объясняется отсутствием необходимых знаний и навыков. На данном этапе, агрегируя все выводы, в модель следует включить score (как целевой показатель), address, mjob, higher, romantic (по результатам реализации критерия Стьюдента), age, failures, medu (по результатам корреляционного анализа), famrel (как показатель с наибольшим количеством выбросов, которые возможно моделировать), traveltime, freetime, absences (как показатели, которые могут быть значимы после объединения категорий данных).