**Цель данной работы** - отследить влияние условий жизни учащихся в возрасте от 15 до 22 лет на их успеваемость по математике, чтобы на ранней стадии выявлять студентов, находящихся в группе риска.

**Задача** – провести EDA и определить, какие условия жизни учащихся в возрасте от 15 до 22 лет влияют на успеваемость по математики.

Анализ выполнен в файле: 'EDA unit 2.ipynb’

**Ход работы:**

**Загружаем и просматриваем данные:**

Датафрейм с загруженными данными - stud

Мы имеем выборку из 395 учащихся. Полные данные содержатся в колонках: school, sex и age.

В остальных колонках имеются пропуски до 11%.

После беглого просмотра дорабатываем столбцы:

fedu – значение 40.0 (1 раз) - похоже на ошибку 4.0 (так как диапазон значений 0-4)

исправить на 4.0

famrel – значение -1.0 (1 раз) – похоже на ошибку 1.0 (так как диапазон значений 1-5)

исправить на 1.0

absences – имеется два нереальных значения 212 и 385. Напрашивается предположение – выбросы. Это числовой столбец. При расчете среднего они будут искажать результат. Так как их всего 2 – удаляем.

**Подготовка данных к анализу:**

Рассмотрим столбец score (оценки):

Удалим строки с NaN в столбце с оценками 'score', так как они наши ориентиры (по ним будем выявлять влияние различных условий) . Их всего 6 шт из 393 (1,5%)

Построим гистограмму распределения.

Оценки в пределах значений, указанных в сопровождении к данным. Выбросов нет.

* 37 строк имеет значение 0 - по каким-то причинам они не сдавали или не сдали госэкзамен (их можно не рассматривать для анализа влияния различных условий).
* большинство получило 50-55 баллов (оценки кратны 5).

Поделим оценки на группы 0 - (0), 1 - (20-30), 2 - (35 - 45), 3 - (50-60), 4 - (65-75), 5 - (80-100) и добавим столбец.

Сформируем отдельный датафрейм из учеников, у которых оценка выше 0 - score\_group\_1\_5. В него входят группы оценок от 1 до 5.

Данные типа float64 не будем преобразовывать в int64. Так легче ориентироваться в сводных таблицах. Они не путаются с группами оценок.

**Работаем с пропусками NaN**

Анализируем по score\_group\_1\_5, а правим одинаково score\_group\_1\_5 и stud (так как мы не знаем на какую оценку они бы написали и не можем определить повлиял ли это фактор на них или нет. Ориентир наш на тех, у кого есть баллы выше 0.

Анализ будем производить по количеству уникальных значений в столбце (сколько раз это значение встречается (в абсолютных величинах и в процентах). А так же по сводной таблице, где эти значения распределены по группам оценок.

Если в колонке имелись пропуски, то после каждого заполнения будем повторно формировать сводную таблицу, чтоб убедиться, что соотношение показателей не уехало.

Получилось заполнить пропуски во всех столбцах, кроме fedu и paid. В них просто так не получится заполнить без искажения результата.

### Проводим анализ: как разные условия оказывают влияние на успеваемость.

### Посмотрим коэффициент корреляции и сводные таблицы по числовым столбцам.

Есть ли между ними взаимосвязь и какая. В основном они у нас ранговые (от меньшего к большему, от худшего к лучшему), кроме пропусков занятий (фактическое количество пропусков).

Нас в основном интересует связь всех столбцов со столбцом score.

Видим слабую взаимосвязь между всеми показателями со столбцом score.

Наиболее выражена отрицательная корреляция между результатами экзаменов и количеством внеучебных неудач (score и failures -0,29). И между пропусками занятий и результатами экзамена(absences и score -0.21). Получается чем больше ребенок терпит неудач и чем больше пропусков, тем хуже в итоге оценка по госэкзамену.

Отрицательная корреляция по колонке goout (-0.18). Чем больше времени ребенок проводит с друзьями, тем хуже результат экзаена.

Так же слабая положительная корреляция со score по показателям medu (0,18) и fedu (0.14) - что в принципе логично. Уровень образования родителей может оказывать влияние на успеваемость по предмету.

age (-0.13) Возраст , отрицательная корреляция. Чем меньше возраст, тем лучше результат.

studytime (0.12) Время на учебу помимо школы.

studytime, granular (-0.12) неизвестно, что за показатель.

Отметим эти показатели для дальнейшей проверки. И более детально рассмотрим в сводных таблицах.

age - С 18 лет уже хуже имеют результат по экзамену. В основном хорошие и лучшие баллы (группа 4 и 5) имеют 16-летние.

Как и показала корреляция - этот показатель влияет на результат

medu - Образование матери оказывает влияние на результат. В группах с высокими оценками у мам высшее образование встречается чаще.

fedu - По образованию отца картина более сглажена. Там не выявлено такой сильной разницы. Хотя тоже есть. Чем лучше образование родителей, тем чаще дети получают более высокие баллы.

traveltime - В большинстве случаев школа находится близко от дома (время пути менее 15 минут). Какой либо закономерности не выявлено.

studytime - В большенстве своем ученики тратят в среднем 2-5 часов в неделю (%2.0) на учебу помимо школы. Но у тех, чей результат по экзамену выше тратят времени больше (если рассмотреть столбец %3.0 и %4.0 в процентном соотношении по группам 23.1% -> 0 и 12.8% -> 4,3% соответсвенно). Влияние на результат есть.

failures - Количество внеучебных неудач оказывает влияние на результат. Это видели раньше при расчете корреляции - получили отрицательную. Дети, отличники 97,4% не сталкивались с неудачами , и только 2,6% детей имели одну. Тогда как дети с плохими результатами чаще сталкивались с внеучебными неудачами.

studytime, granular - У большинства значение -6.0. Но зависимость имеется. По значению -12.0 (12.8(5) -> 4.3(1)), -9.0 (23.1(5) -> 15.2(1)) и противоположная зависимость -6.0 (41(5) -> 65.2(1)), -3.0(23.1(5) -> 30.4(1))

famrel - Семейные отношения у большинства детей хорошие. У 35,9% отличников по экзамену отношения в кругу семьи отличные, тогда как у тех, кто имеет результат плохой всего 21,7% детей имеют отличные взаимоотношения. Средние результаты распределились между ними. Корреляция низкая. Посмотрим на тесте.

freetime - По свободному времени после школы нет какой-то четкой взаимосвязи. 39,7% детей имеет среднее количество этого времени. Совсем мало свободного времени имеют 4,6% детей и совсем много 9,7%.

goout - Проведение времени с друзьями тоже нет четкой зависимости. Но дети с отличным результатом его проводят меньше, чем с плохим результатом. Но корреляция показала результат -0.18. Если проверку тестом пройдет, то будем учитывать.

health - В районе 39% ребят имеют очень хорошее здоровье. Плохое и не очень чаще встречается у 4 и 5 групп, реже 1, а в 2 и 3 всего 6-10% имеют плохое и не очень здоровье. Как таковая зависимость не улавливается.

absences - Чем выше результат, тем в среднем ребята меньше пропускали занятия. Это подтверждает и отрицательный коэффициент корреляции.

### Пробежимся по номинативным переменным с помощью графика боксплот и сводных таблиц.

school - Школу GP выбирают больше, чем MS (88.6 против 11,4). Хотя в группах, прослеживается закономерность - группы с более высокими баллами гр.4 и гр.5 на 5-7% чаще выбирают школу GP. Но мы не можем судить по выбору школы. Так как процент детей, учащийся в школах MS мал.

sex - Внутри одного пола распределения одинаковое (коробочки одного размера). Но мальчики лучше усваивают математику. И в сводной таблице прослеживается линейная связь, что мальчики имеют более высокие баллы чаще, чем девочки, особенно в 5 группе (отличники). А очень плохо сдают в основном девочки (1 группа). Этот показатель влияет на результат экзамена.

address - 79% учащихся живет в городе. Но если сравнивать по группам, то группы 4 и 5 имеет более высокий процент с признаком город. И чем ниже группа (хуже результат), тем процент проживающих в городе ниже, а за городом соответственно выше. Стоит отметить этот показатель.

famsize - Семьи, где больше трех человек встечаются чаще, чем 3 и менее человека. Распределены в основном одинакого, за исключением 5 группы. Здесь больший процент, по сравнению с 1-4 граппами, маленьких семей. Влияние на результат он не оказывает.

pstatus - В основном 90% имеют полные семьи. Но в группах 1 и 5 этот процент немного ниже. А в группах 2 и 4 немного выше. У учащихся, у которых родители живут раздельнно результаты по 25% квартилю выше и нижний ус на 10 балов выше, чем если родители живут вместе. Медиана у них одинаковая. Никакой четкой закономерности я здесь не вижу. Наерное не стоит учитывать этот фактор.

mjob - График отличаются сильно по всем показателям: медиана, коробочка, усы. Это подсказывает, что данный показатель оказывает влияние на результат.

fjob - Графики так же, как и у работы матери, отличаются сильно по всем показателям: медиана, коробочка, усы. Отцы в своем большенстве работают на других работах, не выделенныз в выборке и меньше сидят дома по сравнению с мамами. Особенности: среди отличников папы учителя втречаются в 3-4 раза чаще, чем среди остальных групп и . Наверное папы являются преподавателями технических специальностей. Предположим, что данный показатель так же оказывает влияние на результат.

reason - Выбор школы исходя из репутации по медиане имеет более высокий показатель, но и размах у нее в обе стороны одинаковый, что по 25 и 75 квартилю, что и по усам. По сводной таблице видим выбор школы чаще всего основывается на образовательной программе. Каких то конкретных закономерностей, влияющих на результат, не выявлено.

guardian - Опекунами детей в основном является мать (72%). Отец намного реже (22%). Но есть закономерность. При более высоких баллах опекунство отцом встречается чаще, чем при низких баллах. У матерей наоборот, при более низникх баллах встречается чаще (78%), чем при более высоких(69), а в средних (2-4 группы) распределено равномерно. У учеников (с опекуном мать) разброс результатов больше всего. Тут неоднозначно. Посмотрим как этот показатель пройдет тест.

schoolsup - Дополнительной образовательной поддержкой чаще пользуются ученики, у которых результат экзамена ниже. У кого результат выше среднего обычно не нуждаются в ней. Те, кто находится в группе 5 (отличники) ей не пользовались.Прослеживается линейная связь от гр.5 к гр.1 (0 -> 4.4% -> 16% -> 21.2% -> 34.8%) Существенный показатель.

famsup - Семейной образовательной поддержкой пользовались все ученики примерно одинакого ~ 66%. Только группа 1 на 13% чаще обраращалась к родительской поддержки. По этому показателю невозможно судить о результатах экзамена.

paid - Дополнительные платные занятия чаще брали дети средних результатов - группы 2, 3 и 4. Особенно хорошисты - 4. Неоднозначно он влияет на резльтат.

activities - Дополнительными внучебными занятиями занималась примерно половина детей в равной степени во всех группах, кроме 1. У них на 19% меньше посещает доп занятия - всего 34.8%. У посещающий дополнительные внеучебные занятия оценки ниже 50 получают реже. Не влияет на результат.

nursery - 80.6% детей посещали детский сад. Меньше всего в 3 группе. Распределение в форме параболы. У тех, кто пощещал детский сад вариативнойсть результата экзамена выше. Медина на одном уровне с теми, кто не посещал. Но, как мне кажеться влияние тут незначительное.

higher - У тех, кто желает получить высшее образования в основной массе оценки выше. Но раздрос по усам гораздо больше. У тех, кто не планирует получать высшее образование результат находится на среднем уровне 40-60 баллов в основном, и реже 35 и 65. Из сводной таблице видим, что получить высшее образование желают 96,3%. Со 2 по 4 группу есть желающие получить средне-профессиональное образование по линейной зависимости 1.1(4) -> 4.6(3) -> 9.1(2). А в 5 и 1 группах все желают получить высшее образование. Так что он не может по нему оценивать влияние.

internet - Наличие интернета дает более широкий размах как и в плюс 75 квартиль выше и ус верхний, так и в некоторой степени в минус - ус нижний до самого низа (интернет явно не для учебы используют) Интернет есть у 86,3% детей. Зависимость по параболе. Больше процент детей имеют интернет в 1 и 5 группе и меньше всего процент детей в 3 группе. Разница небольшая. Не будем рассматривать этот фактор как влияющий на результат.

romantic - В романических отношениях медиана на одном уровне у кого есть, и у кого нет. Но по разбосу нижнего и верхнего квартиля получается, что у кого нет романтическиих отношений результат лучше. По сводной таблице видим, что романтических отношения больший процент детей во 2 группе (ниже среднего результат), а меньший в 1 (где худший результат). Неоднозначный показатель.

**Теперь проверим, есть ли статистическая разница в распределении оценок по всем признакам с помощью теста Стьюдента**

Нулевая гипотеза: распределения оценок учащихся по экзамену по различным параметрам неразличимы.

Исходя из проведенного анализа, мое предположение, что влияние на результат экзамена по математике оказывают следующие условия:

1. 'sex',
2. 'age'
3. 'address'
4. 'medu'
5. 'fedu',
6. 'mjob'
7. 'fjob'
8. 'guardian' - под ?
9. 'studytime',
10. 'failures'
11. 'schoolsup'
12. 'studytime, granular'
13. 'famrel' - под ?
14. 'goout'
15. 'absences'

**Результат проверки:**

Найдены статистически значимые различия для колонки sex

Найдены статистически значимые различия для колонки age

Найдены статистически значимые различия для колонки address

Найдены статистически значимые различия для колонки medu

Найдены статистически значимые различия для колонки fedu

Найдены статистически значимые различия для колонки mjob

Найдены статистически значимые различия для колонки studytime

Найдены статистически значимые различия для колонки failures

Найдены статистически значимые различия для колонки schoolsup

Найдены статистически значимые различия для колонки studytime, granular

Найдены статистически значимые различия для колонки goout

Не прошли проверку колонки:

1. 'fjob'
2. 'guardian'
3. 'famrel'
4. 'absences'

Если в 'guardian' и 'famrel' я сомневалась и тест их отбросил. То работа отца и пропуски я думаю надо будет включить в отборочную таблицу. Но при этом в дальнейшем использовании учесть, что тест не нашел в них значимых различий.

По итогу были созданы два датафрейма на основе stud и score\_group\_1\_5 с колонками, которые влияют на результат.

stud\_for\_model - все учащиеся

score\_1\_5\_for\_model - с учащимися у которых оценки выше 0

**Вывод:**

В предоставленных данных достаточно мало пустых значений не более 11,4 % по столбцу.

Выбросы найдены только в одном столбце absences (количество пропусков) – 2 шт, что позволяет сделать вывод о том, что данные достаточно чистые.

Самые важные параметры, которые предлагается использовать в дальнейшем для прогнозирования результатов по госэкзамену по математике и иных исследований в этом направлении: 'sex', 'age', 'address', 'medu', 'fedu', 'mjob', 'studytime', 'failures', 'schoolsup', 'studytime, granular', 'goout'.

А так же: 'absences' и 'fjob', но учитывать, что они не прошли тест-проверку.