Задачей данного проекта было выявить ключевые факторы, которые оказывают заметное влияние на оценки по математике, особенно негативное влияние.

Набор данных представляет собой выборку по 390+ ученикам, в которых представлены их оценки по математике и еще 28 сопутствующих характеристик условий жизни и учебы учеников.

Основными этапами работы над проектом можно разбить на такие этапы:

1. Изучение описательной части к набору данных
2. Ознакомление с данными: их набор, заполнение, разброс, пр.
3. Изучение корреляции показателей между собой и выявление характеристик с наибольшим показателем корреляции
4. Дальнейшее углубленное изучение ограниченного набора характеристик с наибольшей корреляцией к оценкам.
5. Формирование выводов.
6. Какова была ваша роль в команде?

На данный момент не могу сказать, что я работаю в полноценной команде. Скорее меня привлекают на точечные технические задачи по анализу данных.

1. Какой частью своей работы вы остались особенно довольны?

Матрица корреляции всего набора данных. Их оценки подтвердились дальнейшим изучением характеристик.

3. Что не получилось сделать так, как хотелось? Над чем ещё стоит поработать?

В рамках модуля надо более детально разобраться с тестированием гипотез.

4. Что интересного и полезного вы узнали в этом модуле?

Применение статистических методов.

5. Что является вашим главным результатом при прохождении этого проекта?

В целом то, что удалось пройти все задания по модулю.

6. Какие навыки вы уже можете применить в текущей деятельности?

Пока применяю общие навыки Python.

1. Планируете ли вы дополнительно изучать материалы по теме проекта?

Хочу, посмотрим, что получиться.

Можно сделать следующие выводы из анализа данных по учащимся:

1. Признаки, которые оказывают влияние на оценку по математике (score):
   1. Failures – демонстрирует наибольшую корреляцию , а значит и влияние на score. Из анализа нет возможности точно определить природу такой связи. Можно предположить на основании косвенных признаков: failures так же демонстрирует корреляцию с показателем age – чем старше ученик, тем больше активность во внеучебных мероприятиях и значит количественно больше неудач.
   2. Age – не совсем прямолинейная зависимость. Коэффициент корреляции показывает обратную зависимость – то есть с увеличением возраста ухудшается оценка. График boxplot так же это подтверждает но только для возраста 19 лет. Группа в 20 лет показывает значительно выше среднюю оценку. Хотя таких студентов только 3. Предположение: возрастная тенденция – тинейджеров тянет к социализации и как следствие меньше времени на обучение и формируются другие приоритеты.
   3. Goout – из пункта b ожидаемо получаем тенденцию того, что при увеличении проведения времени с друзьями ухудшается оценка. Считаю, ту прямая связь – уменьшается время на учебу. Это подтверждает следующий пункт
   4. Studytime – с увеличением времени на обучение с 1-2,5 часов до 5-10 часов средняя оценка возрастает на 6-7 балов.
   5. Medu – можно назвать неожиданно значимым фактором на улучшение средней оценки. Тут прослеживается четкая тенденция – с повышением образования у матери повышается средний бал. При этом образование отца не демонстрирует такого сильного влияния.
   6. Higher – следующий по значимости фактор влияния на оценку. Желание получить высшее образование стимулирует на лучшую учебу.
   7. Romantic – оказывает отрицательное влияние на среднюю оценку, коэффициент корреляции обратный. При этом это не связано с тем, что такие ученики проводят больше времени с друзьями (соответствующий коэффициент корреляции стремятся к 0). Можем предположить, что это связано с психологическим состоянием и не возможностью сконцентрироваться.
   8. Address – ученики из города показывают оценки лучше чем их сверстники за городом. При этом прослеживается связь с наличием интернета только у городских учеников. На загородных сверстников отсутствие интернета не оказывает значимого влияния на среднюю оценку.
2. Остальные факторы можно убрать из анализа как не оказывающие значимого влияния на результат.