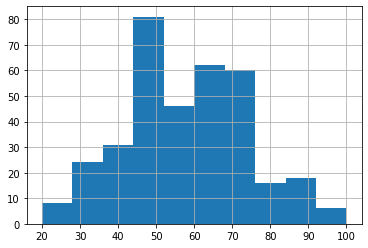
Проверил на корректность возраст. Все в пределах 15-22 года.

Удалил всех учеников, у которых балл = 0. Считаю, что такие ученики либо не явились на экзамены, либо по тем или иным причинам их не писали. Информацию о них сохранил в фрейме data\_no\_score, чтобы в дальнейшем проанализировать отдельно и, возможно, выявить зависимость с тем, почему ученики не сдают экзамены. После удаления «нулей» минимальный балл ученика составил 20.

Распределение оценок учеников:



В дальнейшем предполагаю попробовать проанализировать все возможные параметры на предмет корреляции с оценкой за экзамен, а также разбить всех учеников на группы согласно полученному баллу (см картинку выше) и проанализировать эти группы в отдельности в том числе на предмет корреляции всех параметров внутри них.

Ввиду большого количества пропусков в столбцах address и traveltime (тип адреса городской/деревенский, время пути до школы) было решено их заполнить следующим образом. Было найдено среднее время пути для деревенского и городского жителя (среднее в смысле представленных данных, то есть – от 1 до 4). Все пустые адреса проверялись по соответствию имеющегося времени пути среднему в соответствующей группе. Все пропуски адресов для конкретных учеников были заполнены в зависимости от того, к какой группе (городской тип проживания или за городом) ближе время пути до школы соответствующего ученика. Аналогично были заполнены пропуски во времени пути. В зависимости от адреса проживания ставилось среднее значение времени для соответствующей группы людей (что привело к появлению дробных значений в столбце с временем пути. Но на методы вычисления это обстоятельство не влияет). Оставшихся три ученика, у которых не были заполнены ни время пути до школы, ни тип адреса проживания было решено удалить. Тем более, что в задании в качестве цели указано желание выявлять проблемных учеников на ранних стадиях, а все три означенных ученика имеют средние (55 - 75) баллы за экзамен.

Удалил (всего несколько учеников) тех, у кого не был заполнен балл за экзамен, потому что анализировать их не имеет смысла.

Сделал гистограммы всех численных переменных. У некоторых даже визуально видны выбросы.

На выбросы анализирую следующим образом. Во-первых – математически сообразно формуле выбросов, во-вторых – с точки зрения поставленной задачи. Если с точки зрения задания данные имеют место быть, то выбросами их считать не буду даже если они таковыми являются математически. В частности, сразу же можно опустить показатель возраста. С точки зрения математики там есть выброс, но все значения соответствуют интервалу возрастов, рассматриваемом в задании. Так что никаких учеников из имеющихся данных убирать не буду.

В частности, многие показатели (к примеру, образование родителей) разбиты по группам дискретно. Поэтому выбросы могут исключать в том числе и существенные группы данных.

**!!!!! Примечание !!!!!** Обнаружил, что после изменений параметров (удаления некоторых данных, в основном – выбросов) данные, выдаваемые программой меняются. То есть – считаются новые выбросы, отличные от предыдущих и прочее. Это нормально, но из-за того, что я обновлял ячейки ноутбука после удаления лишних данных неочевидно, какие выбросы и неточности были в данных изначально. Эта информация отражена мной в комментариях к соответствующих ячейках.

С выбросами в численных данных разобрался. Некоторые данные удалил.

Столбец **studytime, granular идентичен studytime** в том смысле, что не несёт в себе новой информации (корреляция 100%). Кроме того, нет его описания в задании. Удалил.

Чтобы приступить к анализу корреляций осталось заполнить или удалить пропуски в данных. В численных данных буду заполнять средним/медианой/модой. Точнее укажу в комментариях непосредственно в коде ноутбука. В номинативных переменных стану ставить скорее всего самые часто встречающиеся. Точнее так же буду писать в коде.

Примечание. В идеале считаю, что нужно проанализировать все данные и заполнить не просто средними или самыми часто встречающимися, но наиболее характерные согласно другим имеющимся данным с наибольшей корреляцией. То есть, если переменная, чьего значения не достаёт, сильно коррелирует с другой переменной, то нужно заполнять пропуск в соответствие с другими учениками, имеющими аналогичные параметры с высокой корреляцией. Для численных переменных коэффициенты корреляции могли бы выступить в качестве весовых коэффициентов. Но в имеющихся данных незначительное количество пропусков, поэтому в таких сложностях нет необходимости.

Решил, что если в столбце большое количество пропусков (примерно больше 5-6), то буду заполнять их медианными значениями, а если небольшое – то модой.

Решил удалить ученика №236, так как у него не хватает целых трёх численных данных.

**Пропуски заполнил**

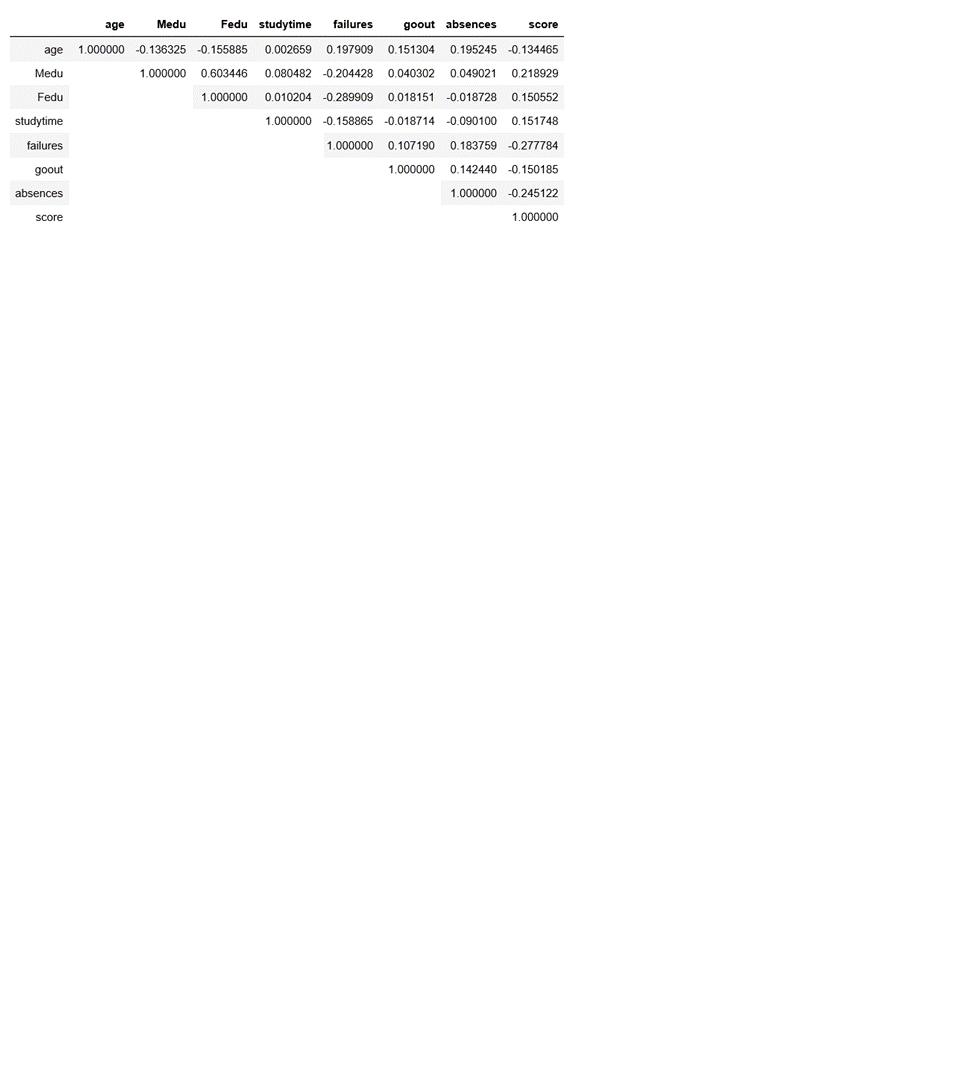
После заполнения пропусков в данных, осталось проанализировать корреляции. Кроме того предполагаю после анализа выделить учеников с низким баллом (меньше 30, к примеру) и отдельно проанализировать корреляции с ними, так как цель – выявить закономерности, приводящие учеников к плохому результату (это вместо разбиения на малые группы, предложенного мной в начале работы). И обстоятельства, не влияющие на учеников с хорошим, могут тем не менее иметь большую значимость на меньшей выборке. Кроме того нужно также посмотреть учеников, которые не явились на экзамен и были выделены в отдельный датафрейм в начале работы.

Получившиеся корреляции численных значений:

|  | **age** | **Medu** | **Fedu** | **traveltime** | **studytime** | **failures** | **famrel** |  | **freetime** | **goout** | **health** | **absences** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 1.000000 | -0.136325 | -0.155885 | 0.110284 | 0.002659 | 0.197909 | 0.084343 |  | 0.023685 | 0.151304 | -0.055550 | 0.195245 | -0.134465 |
| Medu | -0.136325 | 1.000000 | 0.603446 | -0.171306 | 0.080482 | -0.204428 | -0.021531 |  | 0.034662 | 0.040302 | -0.027976 | 0.049021 | 0.218929 |
| Fedu | -0.155885 | 0.603446 | 1.000000 | -0.155176 | 0.010204 | -0.289909 | -0.033327 |  | 0.009396 | 0.018151 | 0.000942 | -0.018728 | 0.150552 |
| traveltime | 0.110284 | -0.171306 | -0.155176 | 1.000000 | -0.109505 | 0.117293 | -0.022116 |  | 0.010088 | 0.038463 | 0.015498 | -0.074865 | -0.076893 |
| studytime | 0.002659 | 0.080482 | 0.010204 | -0.109505 | 1.000000 | -0.158865 | 0.065248 |  | -0.143910 | -0.018714 | -0.065974 | -0.090100 | 0.151748 |
| failures | 0.197909 | -0.204428 | -0.289909 | 0.117293 | -0.158865 | 1.000000 | 0.006340 |  | 0.116182 | 0.107190 | 0.060334 | 0.183759 | -0.277784 |
| famrel | 0.084343 | -0.021531 | -0.033327 | -0.022116 | 0.065248 | 0.006340 | 1.000000 |  | 0.165712 | 0.032474 | 0.089942 | -0.106403 | 0.055443 |
| freetime | 0.023685 | 0.034662 | 0.009396 | 0.010088 | -0.143910 | 0.116182 | 0.165712 |  | 1.000000 | 0.294914 | 0.064900 | 0.080196 | 0.006913 |
| goout | 0.151304 | 0.040302 | 0.018151 | 0.038463 | -0.018714 | 0.107190 | 0.032474 |  | 0.294914 | 1.000000 | 0.004581 | 0.142440 | -0.150185 |
| health | -0.055550 | -0.027976 | 0.000942 | 0.015498 | -0.065974 | 0.060334 | 0.089942 |  | 0.064900 | 0.004581 | 1.000000 | -0.005698 | -0.084214 |
| absences | 0.195245 | 0.049021 | -0.018728 | -0.074865 | -0.090100 | 0.183759 | -0.106403 |  | 0.080196 | 0.142440 | -0.005698 | 1.000000 | -0.245122 |
| ***score*** | -0.134465 | 0.218929 | 0.150552 | -0.076893 | 0.151748 | -0.277784 | 0.055443 |  | 0.006913 | -0.150185 | -0.084214 | -0.245122 | 1.000000 |

Как видно – есть совсем слабо влияющие на итоговый балл параметры. В частности, буду считать такими те, у которых по модулю коэффициент корреляции меньше 10%. То есть: **traveltime, famrel, freetime, health.** В этой связи считаю допустимым удалить данные столбцы из-за низкой ценности этих данных.

После удаления лишних данных зависимости следующие:



Дальше необходимо посмотреть учеников по нечисловым признакам. Решил в первую очередь разделить их по полу и посмотреть влияние разных факторов на результаты экзаменов.

На основании полученных графиков можно сделать следующие предварительные выводы:

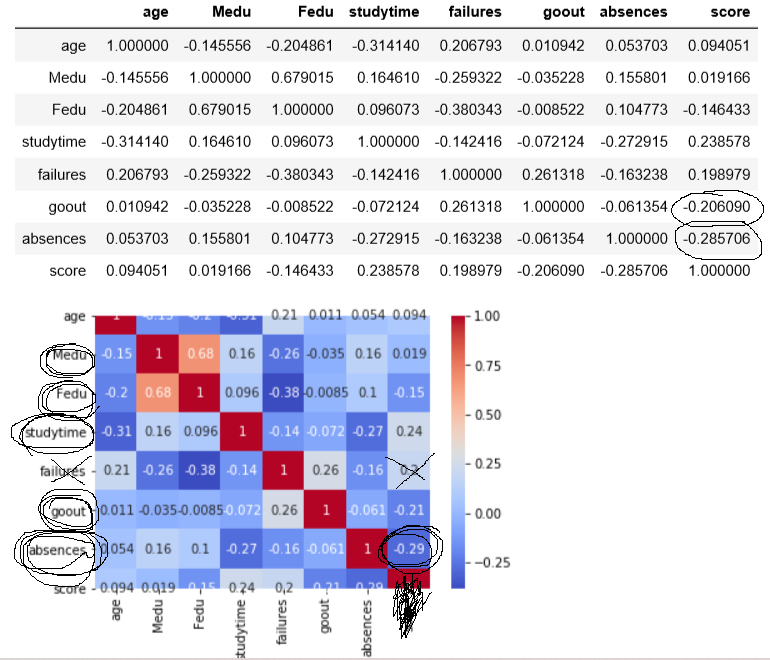
1. Ученики, живущие в черте города, имеют лучшие результаты. Особенно это касается учеников мужского пола.
2. Размер семьи на экзамен влияет незначительно.
3. Дети разведённых родителей учатся лучше, если они мужского пола, и хуже, если женского.
4. Работа родителей влияет на результаты учеников, но не оказывает кардинального влияния. То есть не «загоняет» никого в зону риска. Видно небольшое исключение: мать-домохозяйка у учеников мужского пола может оказать существенно положительное влияние на успеваемость.
5. Причина выбора школы почти не оказывает влияния на экзамен. Небольшое положительное влияние есть у тех, кто выбирал школу по её репутации.
6. Пол опекуна влияния не оказывает, но для девочек предпочтительно, чтобы их опекал кто-то из родителей, а не другой человек. У всех, кого опекают не родители (девочки) очень средние результаты, но в основной массе ниже, чем у тех, кого опекают родители.
7. Дополнительная образовательная поддержка (schoolsup) отрицательно влияет на результаты экзаменов. Возможно – результат ошибочный (потому что с точки зрения здравого смысла довольно странный. Другая мысль – дополнительную поддержку оказывают детям с проблемами во время обучения, отчего они в целом слабее других уже только из-за факта того, что их именно по этому критерию и отбирали)
8. Семейная образовательная поддержка не сильно влияет на результаты экзаменов.
9. Дополнительные платные занятия не сильно влияют на результаты экзаменов.
10. Дополнительные внеучебные занятия положительное влияние оказывают на учеников мужского пола.
11. На девочек посещение детского сада не сильно влияет, но имеется некое отрицательное воздействие, у мальчиков – наоборот – имеется положительное влияние (но тоже не сильно значительное)
12. Желание получить высшее образование оказывает однозначное положительное влияние на учеников мужского пола. Для женского недостаточно данных для оценки (нет тех, кто не хочет)
13. Наличие доступа к интернету оказывает положительное влияние на результаты экзаменов для обоих полов.

Итого, если искать группу риска (речь пока что идёт только о качественных параметрах), то следует присмотреться:

Женский пол: раздельно живущие родители; опекун – не кто-то из родителей; получает дополнительную образовательную поддержку; посещала детский сад.

Мужской пол: живёт вне городской черты; родители живут вместе; мать не является домохозяйкой; нет дополнительных внеучебных занятий; не посещал детский сад.

Теперь хочу проанализировать группу учеников, получивших меньше сорока баллов. Считаю это значение достаточно низким.



Есть некоторые существенные различия. Образование матери не влияет на группу с низким итоговым баллом. В отличие от общей статистики образование отца влияет отрицательно. Большее положительное влияние ученого времени помимо школы. В отличие от общей статистики количество внеучебных неудач имеет положительное влияние. Большее отрицательное влияние проводимого с друзьями времени и пропущенных занятий.

Дополнительно по качественным признакам можно сказать, анализируя группу с плохими результатами, что на девочек негативно влияет отец, занимающийся домашним хозяйством. В остальном ничего добавить по качественным признакам нельзя.

Посмотрел учеников, не явившихся на экзамен. Общего у них почти ничего нет. Можно выделить то, что большинство из многодетных (GT3) семей, всем от 15 до 19 лет (при этом в основном от 15 до 18, 19 – довольно редкая цифра), у большинства родители живут вместе. Почти у всех школа GP, большинство таких учеников городские жители, у большинства опекуном является мать, у них мало внеучебных неудач (0-1 подавляющее большинство), почти у всех отсутствует дополнительная образовательная поддержка, почти все посещали детский сад, и почти все хотят получить высшее образование, у них практически нет пропущенных занятий.

***Выводы.***

*Группы риска.*

Женский пол: раздельно живущие родители; опекун – не кто-то из родителей; получает дополнительную образовательную поддержку; посещала детский сад; отец занимается домашним хозяйством.

Мужской пол: живёт вне городской черты; родители живут вместе; мать не является домохозяйкой; нет дополнительных внеучебных занятий; не посещал детский сад.

+ Для обоих полов: нет доступа к интернету, мало учебного времени помимо школы, много времени проводит с друзьями, много пропущенных занятий. Образованный отец (Только для группы риска! В общем случае оказывает положительное, а не отрицательное влияние).

Кроме того – для учеников, не явившихся на экзамен.

Можно выделить то, что большинство из многодетных (GT3) семей, всем от 15 до 19 лет (при этом в основном от 15 до 18, 19 – довольно редкая цифра), у большинства родители живут вместе. Почти у всех школа GP, большинство таких учеников городские жители, у большинства опекуном является мать, у них мало внеучебных неудач (0-1 подавляющее большинство), почти у всех отсутствует дополнительная образовательная поддержка, почти все посещали детский сад, и почти все хотят получить высшее образование, у них практически нет пропущенных занятий.