Анализ успеваемости студентов

Кокорев Артём БПИ235

Поиск данных

Поиск данных

https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

Student Performance Factors

Что внутри?

6607 строк и 20 значений в каждой

Hours_Studied - Количество часов, потраченных на обучение (int)
Attendance - Процент посещенных занятий (int)
Parental_Involvement - Уровень участия родителей в образовании студента (Low, Medium, High)
Access_to_Resources - Доступность образовательных ресурсов (Low, Medium, High)
Extracurricular_Activities - Участие в дополнительных занятиях (Yes, No)
Sleep_Hours - Среднее количество часов сна за ночь (int)
Previous_Scores - Средние оценки предыдущих экзаменов (int)
Motivation_Level - Уровень мотивации студента (Low, Medium, High)
Internet_Access - Доступ в интернет (Yes, No)
Tutoring_Sessions - Количество занятий с репетитором в месяц (int) метрическая
Family_Income - Уровень дохода семьи (Low, Medium, High)
Teacher_Quality - Качество преподавания учителями (Low, Medium, High), 78 пропусков
School_Type - Тип школы (Public, Private)
Peer_Influence - Влияние сверстников на академическую жизнь (Positive, Neutral, Negative)
Physical_Activity - Среднее количество часов физической активности в неделю (int)
Learning_Disabilities - Наличие проблем с обучаемостью (Yes, No)
Parental_Education_Level - Образование родителей (High School, College, Postgraduate), 90 пропусков
Distance_from_Home - Расстояние от дома до школы (Near, Moderate, Far), 67 пропусков
Gender - пол студента (Male, Female).
Exam_Score - Оценка за экзамен (int)

Цель

Выявить ключевые факторы, влияющие на академическую успеваемость студентов, и определить их относительную важность.

Задачи:

1. Определить влияние временных затрат на обучение на итоговые результаты

2. Выявить роль социально-экономических факторов в успеваемости

3. Оценить влияние качества преподавания и доступности ресурсов

4. Исследовать взаимосвязь между физической активностью и академическими результатами

5. Определить влияние родительского участия на успеваемость

Гипотезы

H1: Существует положительная корреляция между количеством часов обучения и итоговой оценкой

H2: Существует положительная корреляция между уровнем родительского участия и итоговой оценкой

H3: Существует положительная корреляция между доступностью образовательных ресурсов и итоговой оценкой

H4: Существует разница в успеваемости между государственными и частными школами

H5: Существует положительная корреляция между количеством часов сна и итоговой оценкой

H6: Существует положительная корреляция между количеством занятий с репетитором и итоговой оценкой

H7: Студенты с хорошими оценками в прошлом экзамене показывают лучшие результаты

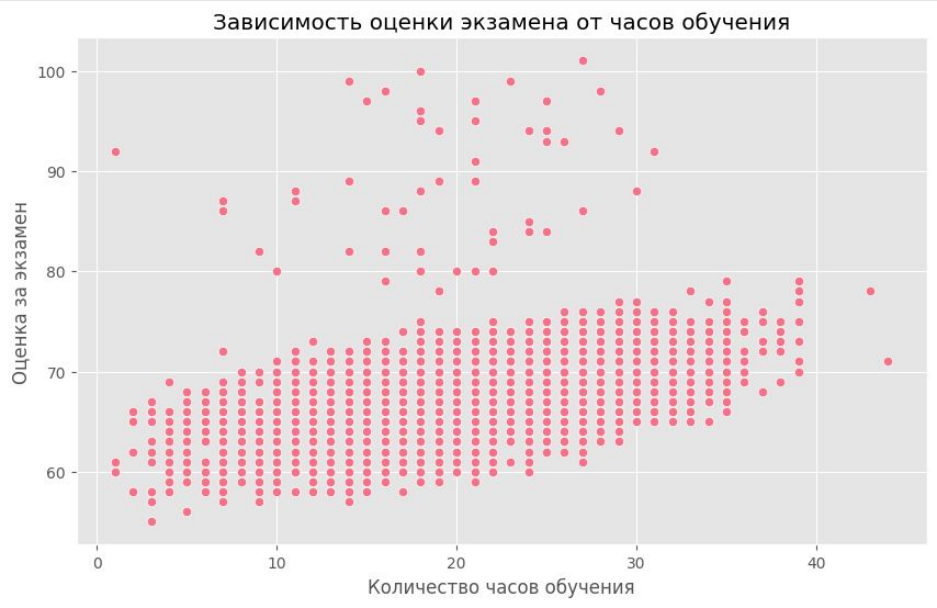
H8: Существует корреляция между влиянием сверстников и итоговой оценкой

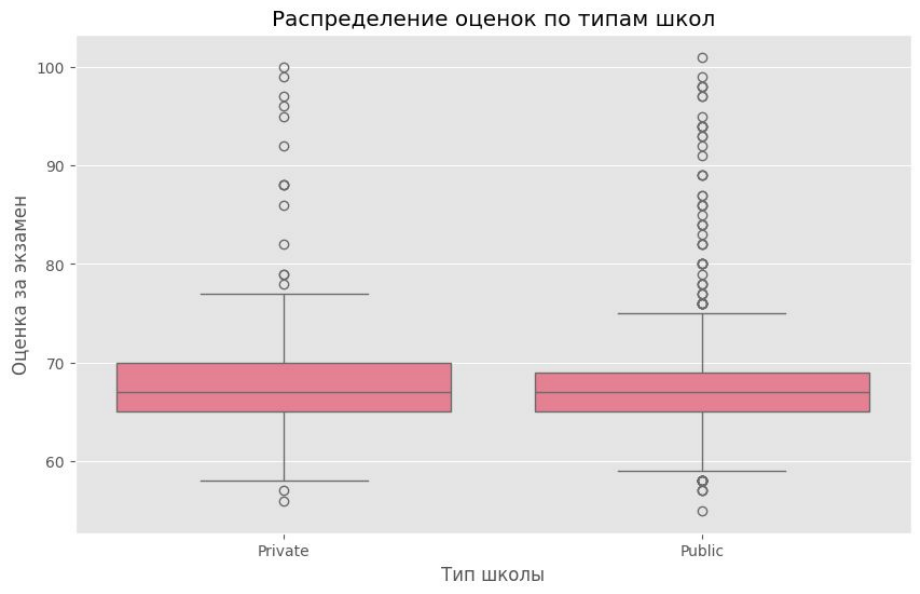
H9: Сущетсвует корреляция между родительским образованием и итоговой оценкой

H10: Существует корреляция между уровнем мотивации и итоговой оценкой

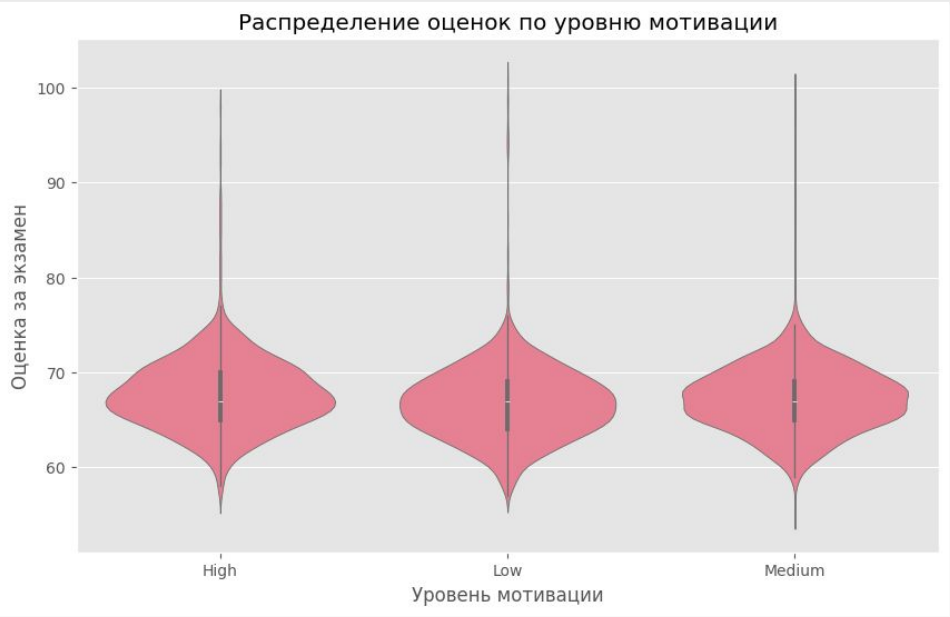
Ps Везде я буду рассматривать влияние на оценку за экзамен, поэтому зависимая переменная будет метрической. Если вторая переменная номинальная, то я буду использовать однофакторный дисперсионный анализ (ANOVA), если бинарная, то t-тест для независимых выборок. Если метрическая, то корреляционный анализ Пирсона.

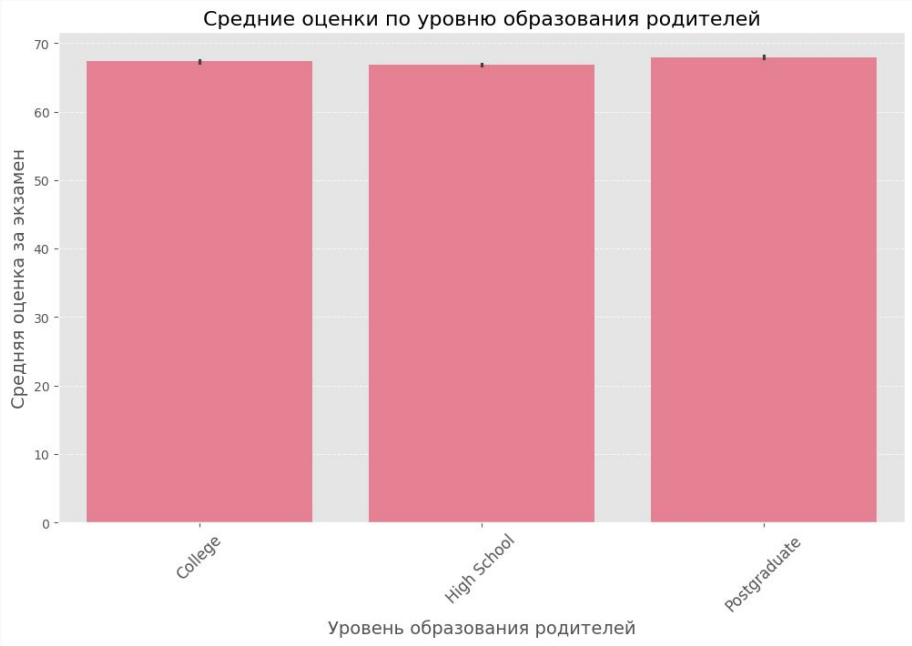
Интересные графики



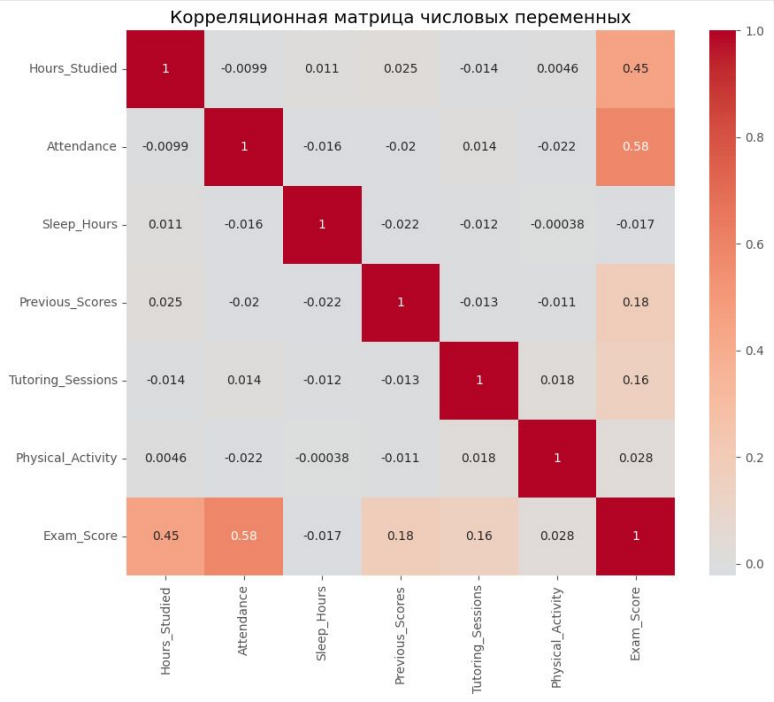


Интересные графики

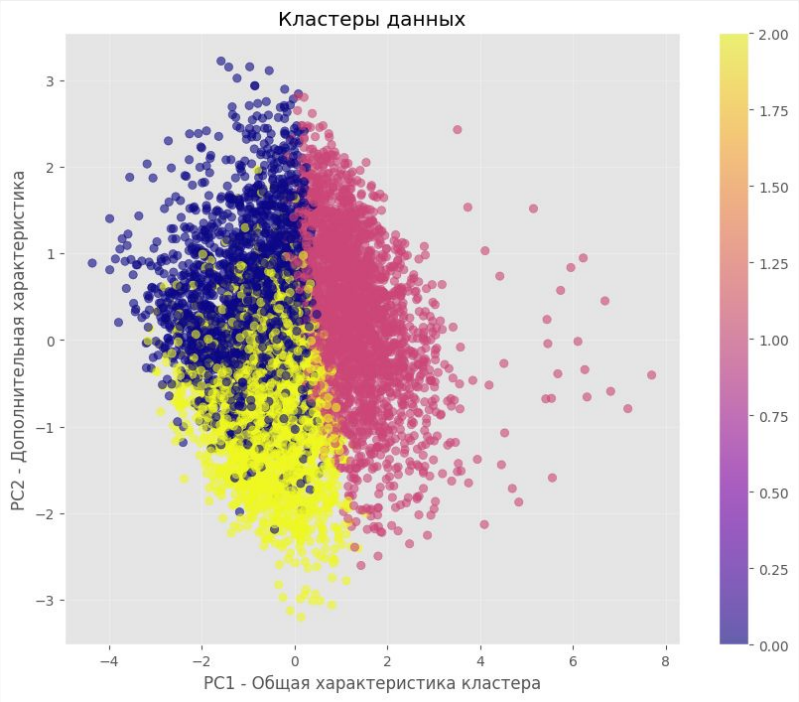


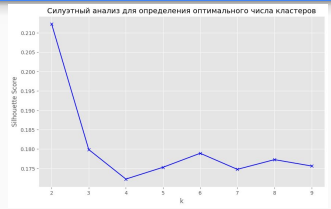


Корреляция


Корреляционная матрица числовых переменных
Hours_Studied
Attendance
Sleep_Hours
Previous_Scores
Tutoring_Sessions
Physical_Activity
Exam_Score
1 -0.0099 0.011 0.025 -0.014 0.0046 0.45
-0.0099 1 -0.016 -0.02 0.014 -0.022 0.58
0.011 -0.016 1 -0.022 -0.012 -0.00038 -0.017
0.025 -0.02 -0.022 1 -0.013 -0.011 0.18
-0.014 0.014 -0.012 -0.013 1 0.018 0.16
0.0046 -0.022 -0.00038 -0.011 0.018 1 0.028
0.45 0.58 -0.017 0.18 0.16 0.028 1
1.0
0.8
0.6
0.4
0.2
0.0

Кластеры





У меня вышло три кластера:

0 кластер - бездельники с низким посещением, временем на учёбу и плохими старыми оценками, но высокой физической активностью.
1 кластер - отличники, посещающие занятия, изучающие материал и имеющие хорошие оценки.
2 кластер - скатывающиеся вниз студенты, с низким посещением, временем на учёбу, но высокими старыми оценками.

Вернемся к гипотезам

H1: количество часов обучения влияет на оценку за экзамен как главный предикат в обеих регрессиях.

H2: родительское участие в образовании практически не влияет на оценку за экзамен, это было доказано в графиках и сводных таблицах.

H3: влияния доступности образовательных ресурсов не было выявлено.

H4: разницы между государственными и частными школами практически нет(десятые доли), это было доказано в графиках и сводных таблицах.

H5: влияния количества сна не было выявлено.

H6: количество занятий с репетитором очень сильно влияет на оценку за экзамен, это был главный предикат в регрессиях.

H7: влияние прошлых оценок не было выявлено.

H8: влияние сверстников очень слабо влияет на оценку за экзамен, сводные таблицы дали разницу чуть больше 1 процента.

H9: родительское образование не влияет на оценку за экзамен, это было доказано в графиках.

H10: мотивация практически не влияет на оценку за экзамен, это было доказано в графиках.

Выводы

Цель исследования явно достигнута, регрессионные модели показывают хорошие результаты и подсвечивают главные предикаты.

Мне показалось неожиданным, что практически ничего кроме посещения и услуг репетиторства глобально не влияет на оценку за экзамен.

Все социально экономические факторы вместе влияют меньше, чем репетиторство.

Отсутствие влияния родительского участия и образования меня очень удивило, но я думаю эти факторы сопряжены с использованием услуг репетиторства.

Анализ успеваемости студентов

Кокорев Артём БПИ235