МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра Информационных систем

ОТЧЁТ О ПРОВЕДЕНИИ ИССЛЕДОВАНИЯ НА ТЕМУ: «ЗАВИСИМОСТЬ МЕЖДУ РЕЗУЛЬТАТАМИ ОБУЧЕНИЯ СТУДЕНТОВ И ИХ ОКРУЖАЮЩЕЙ СРЕДОЙ»

Выполнил студент
гр. 0373
Гейченко Евгений

Преподаватель _____ Татчина Я.А.

Санкт-Петербург
2025

Тема: исследование зависимости между итоговыми оценками студентов и их повседневным стилем жизни

Цель: получение практических навыков применения библиотеки pandas, анализа данных с использованием ЯП Python 3.11.11

Введение

Специалисты компании Amazon, одного из пионеров использования DS для оптимизации бизнеса, выделяют четыре вида анализа данных, который реализуется в рамках DS:

1. *Описательный анализ* (Descriptive analysis) изучает данные, чтобы получить представление о том, что произошло или происходит в среде данных. Он широко использует визуализацию данных, такую как круговые диаграммы, столбчатые диаграммы, линейные графики, таблицы или сгенерированные повествования. Например, служба бронирования авиабилетов может регистрировать такие данные, как количество билетов, забронированных каждый день. Описательный анализ выявит всплески бронирования, спады бронирования и высокоэффективные месяцы для этой службы.

2. *Диагностический анализ* (Diagnostic analysis) предполагает глубокое погружение или подробное изучение данных для понимания того, почему что-то произошло. Он характеризуется такими методами, как детализация, обнаружение данных, интеллектуальный анализ данных и корреляции. При этом над заданным набором данных могут быть выполнены множественные преобразования, чтобы обнаружить уникальные закономерности в каждом из этих методов. Например, служба бронирования авиабилетов может выделить месяц, в котором произошел всплеск бронирования. Это может привести к выявлению закономерности: многие клиенты приезжают в определенный город, чтобы посетить ежемесячное спортивное мероприятие.

3. *Предиктивный анализ* (Predictive analysis) использует исторические данные для составления точных прогнозов относительно закономерностей данных, которые могут возникнуть в будущем. Для этого применяются такие методы, как машинное обучение, прогнозирование, сопоставление с образцом

и предиктивное моделирование. В каждом из этих методов компьютеры обучаются обратному проектированию причинно-следственных связей в данных. Например, служба полетов может использовать DS для прогнозирования закономерностей бронирования рейсов на следующий год в начале каждого года. Компьютерная программа или алгоритм могут просматривать прошлые данные и прогнозировать всплески бронирования для определенных направлений в мае. Предвидя будущие потребности своих клиентов в поездках, компания может начать целевую рекламу для этих городов с февраля.

4. *Предписывающая аналитика* (Prescriptive analysis) выводит предиктивные данные на новый уровень. Она не только предсказывает, что, скорее всего, произойдет, но и предлагает оптимальный ответ на это. Она может анализировать потенциальные последствия различных вариантов и рекомендовать наилучший курс действий. Он использует анализ графов, моделирование, сложную обработку событий, нейронные сети и рекомендательные механизмы из машинного обучения. В примере с бронированием авиабилетов предписывающий анализ может рассмотреть предыдущие маркетинговые кампании, чтобы максимально использовать преимущества предстоящего всплеска бронирований. Специалист по данным может прогнозировать результаты бронирования для разных уровней маркетинговых расходов на разных маркетинговых каналах. Эти прогнозы данных дадут компании по бронированию авиабилетов большую уверенность в своих маркетинговых решениях.

Ход работы

В процессе исследования влияния различных факторов окружающей среды на итоговые результаты студентов будут использованы 2 набора данных, заимствованных с интернет-ресурса https://www.kaggle.com/, доступных по ссылкам:

1) https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style

2) https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset

Первый набор данных является искусственно сгенерированным, содержит 10 000 записей и имеет структуру, представленную в таблице 1. Второй набор данных был получен от реального образовательного учреждения, содержит 5 000 записей и имеет структуру, представленную в таблице 2.

Таблица 1 – Структура набора данных на 10 000 записей

Название, метка(column)	Значение	Тип
Student_ID	Уникальный идентификатор для каждого студента	Строка
Age	Возраст студента (18-30 лет)	Число
Gender	Пол: Мужской, Женский или Другой	Строка
Study_Hours_per_Week	Часы, затраченные на учебу в неделю (5-50 часов)	Число
Preferred_Learning_Style	Предпочитаемый стиль обучения: Визуальный, Аудиальный, Чтение/Письмо, Кинестетический	Строка
Online_Courses_Completed	Количество завершенных онлайн-курсов (0-20)	Число

Название, метка(column)	Значение	Тип
Participation_in_Discussions	Активно ли студент участвует в обсуждениях (Да/Нет)	Булева
Assignment_Completion_Rate (%)	Процент выполненных заданий (50%-100%)	Число
Exam_Score (%)	Оценка студента за финальный экзамен (40%-100%)	Число
Attendance_Rate (%)	Процент посещаемости занятий (50%-100%)	Число
Use_of_Educational_Tech	Использует ли студент образовательные технологии (Да/Нет)	Булева
Self_Reported_Stress_Level	Уровень стресса студента (Низкий, Средний, Высокий)	Строка
Time_Spent_on_Social_Media (hours/week)	Часы, проведенные в социальных сетях за неделю (0-30 часов)	Число
Sleep_Hours_per_Night	Средняя продолжительность сна (4-10 часов)	Число
Final_Grade	Оценка, присвоенная на основе результата экзамена (A, B, C, D, F)	Символ

Таблица 2 - Структура набора данных на 5 000 записей

Название, метка(column)	Значение	Тип
Student_ID	Уникальный идентификатор для каждого студента	Строка
First_Name	Имя студента.	Строка
Last_Name	Фамилия студента	

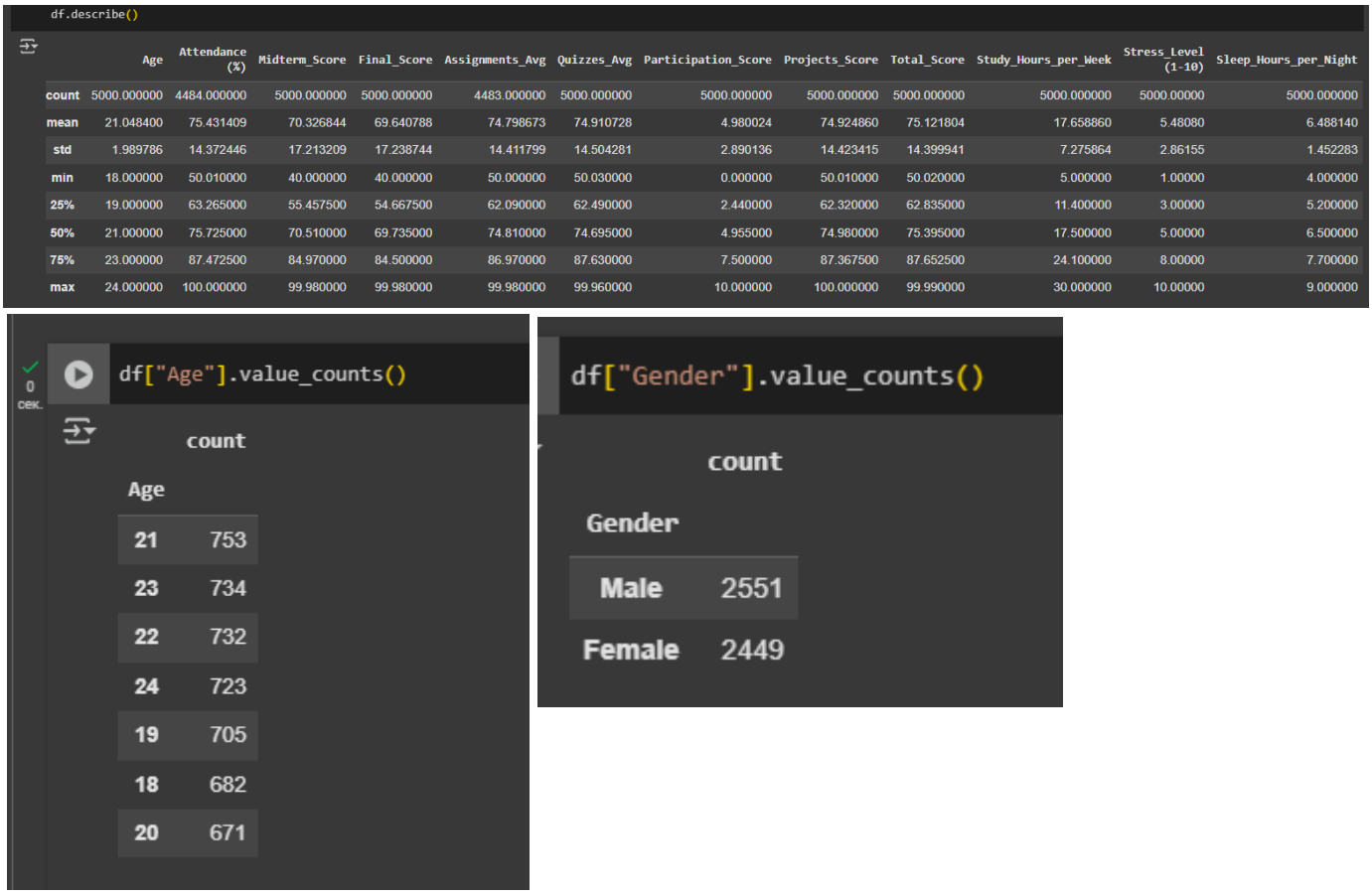
Название, метка(column)	Значение	Тип
Email	Контактный адрес электронной почты	Строка
Gender	Пол: Мужской, Женский, Другой	Строка
Age	Возраст студента	Число
Department	Факультет студента	Число
Attendance (%)	Процент посещаемости (0-100%)	Число
Midterm_Score	Оценка за промежуточный экзамен (из 100).	Число
Final_Score	Оценка за финальный экзамен (из 100).	Число
Assignments_Avg	Средняя оценка всех заданий (из 100).	Число
Quizzes_Avg	Средняя оценка за тесты (из 100).	Число
Participation_Score	Оценка на основе участия в классе (0-10).	Число
Projects_Score	Оценка за проекты (из 100).	Число
Total_Score	Взвешенная сумма всех оценок.	Число
Grade	Буквенная оценка (A, B, C, D, F).	Символ
Study_Hours_per_Week	Среднее количество часов учебы в неделю.	Число
Extracurricular_Activities	Участвует ли студент во внеучебных мероприятиях (Да/Нет).	Булева
Internet_Access_at_Home	Есть ли у студента доступ в интернет дома? (Да/Нет).	Булева
Parent_Education_Level	Высший уровень образования родителей (Нет, Средняя школа, Бакалавр, Магистр, Доктор философии).	Строка
Family_Income_Level	Низкий, Средний, Высокий.	Строка

Название, метка(column)	Значение	Тип
Stress_Level (1-10)	Уровень стресса по самооценке (1: Низкий, 10: Высокий).	Число
Sleep_Hours_per_Night	Среднее количество часов сна за ночь.	Число

Целью исследования является выявление взаимосвязей и зависимостей, следовательно будет применён описательный анализ данных.

Рассмотрим набор данных с количеством записей в 5000:

В наборе данных представлены записи о результатах обучения студентов в возрасте от 18 до 24 женского и мужского пола (рис.1).



df.describe()

	Age	Attendance (%)	Midterm_Score	Final_Score	Assignments_Avg	Quizzes_Avg	Participation_Score	Projects_Score	Total_Score	Study_Hours_per_Week	Stress_Level (1-10)	Sleep_Hours_per_Night
count	5000.000000	4484.000000	5000.000000	5000.000000	4483.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.00000	5000.000000
mean	21.048400	75.431409	70.326844	69.640788	74.798673	74.910728	4.980024	74.924860	75.121804	17.658860	5.48080	6.488140
std	1.989786	14.372446	17.213209	17.238744	14.411799	14.504281	2.890136	14.423415	14.399941	7.275864	2.86155	1.452283
min	18.000000	50.010000	40.000000	40.000000	50.000000	50.030000	0.000000	50.010000	50.020000	5.000000	1.00000	4.000000
25%	19.000000	63.265000	55.457500	54.667500	62.090000	62.490000	2.440000	62.320000	62.835000	11.400000	3.00000	5.200000
50%	21.000000	75.725000	70.510000	69.735000	74.810000	74.695000	4.955000	74.980000	75.395000	17.500000	5.00000	6.500000
75%	23.000000	87.472500	84.970000	84.500000	86.970000	87.630000	7.500000	87.367500	87.652500	24.100000	8.00000	7.700000
max	24.000000	100.000000	99.980000	99.980000	99.980000	99.960000	10.000000	100.000000	99.990000	30.000000	10.00000	9.000000

Рисунок 1 – Описание набора данных

С целью определения корреляций между различными столбцами построим тепловые карты, используя функции corr() и phik_matrix() (рис.2, рис. 3)

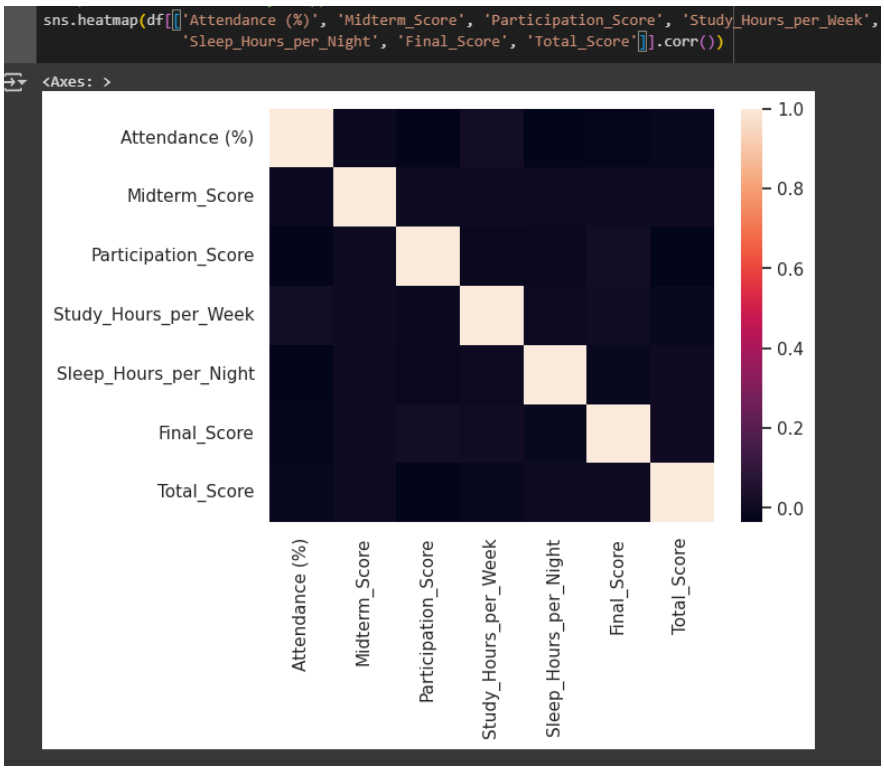


Рисунок 2 – Корреляция между столбцами набора данных с использованием функции corr()

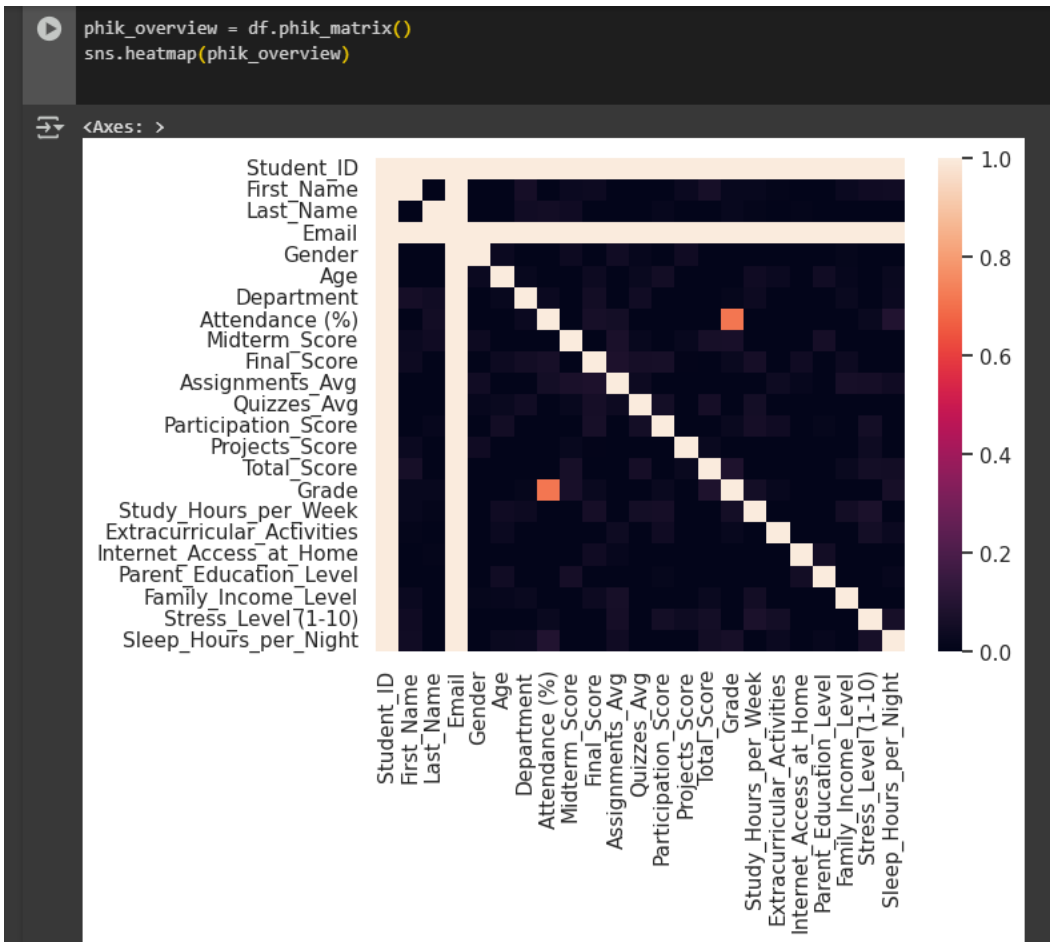


Рисунок 3 – Корреляция между столбцами набора данных с использованием функции phik_matrix()

По полученным тепловым картам видно, что все зависимости выражены слабо за исключением зависимости оценки от посещаемости. Рассмотрим данную зависимость более подробно, построив гистограмму с разбиением по полу (рис. 4). По данной гистограмме видно, что студенты с высокой посещаемостью в среднем получают более высокие баллы, чем студенты с более низкой посещаемостью.

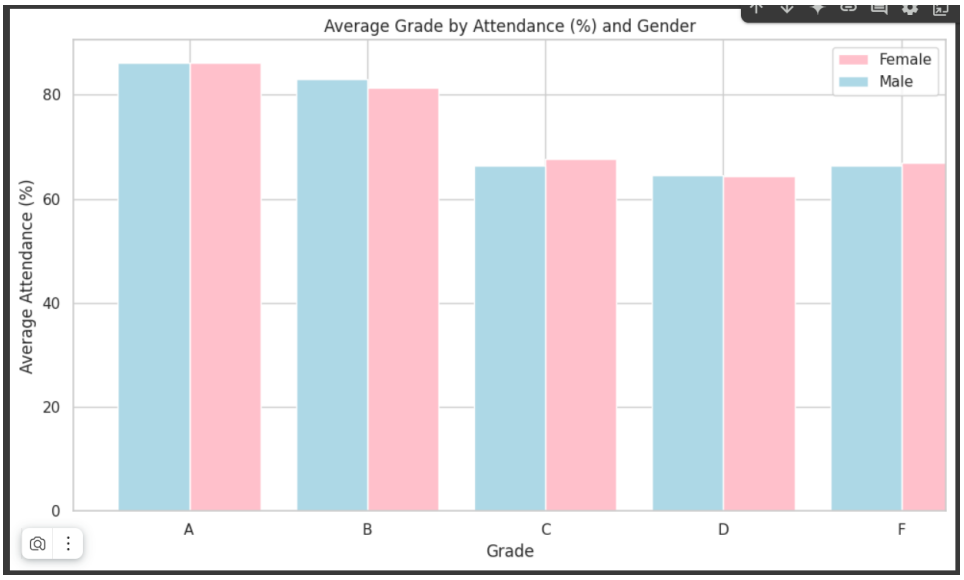


Рисунок 4 – Посещаемость студентов с различными оценками

Рассмотрим зависимость оценок от времени, затрачиваемого на обучение (рис. 5)

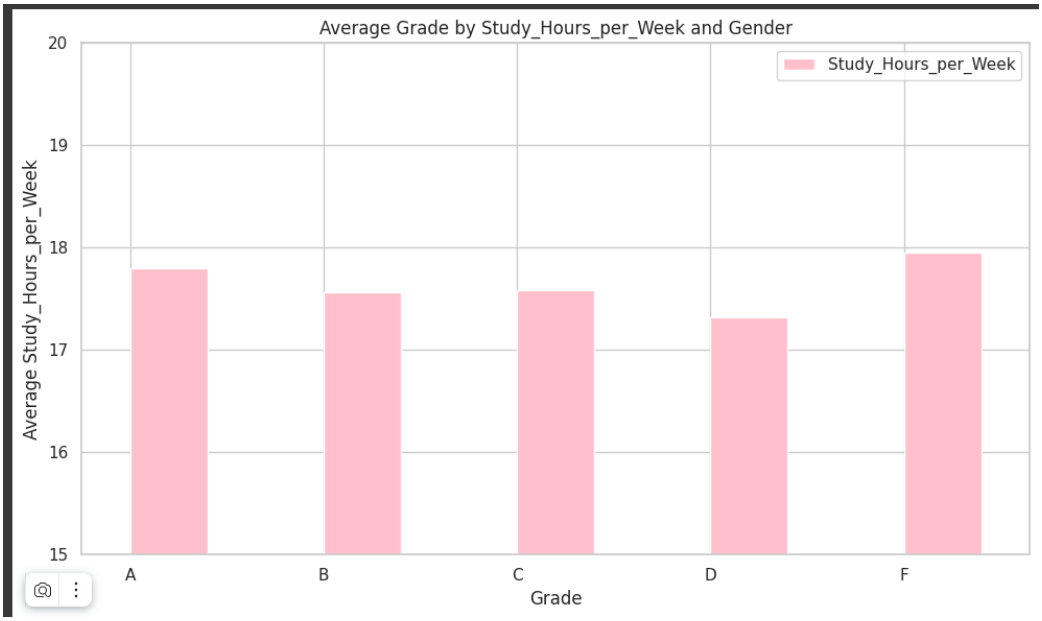


Рисунок 5 – Зависимость оценок от среднего времени, затрачиваемого на обучение

Из рисунка 5 следует, что все группы студентов тратили примерно одинаковое количество времени на обучение. Более того, студенты с плохими оценками тратили наибольшее количество времени, что может говорить о неэффективном использовании времени, либо о незаинтересованности студентов в материале.

Далее рассмотрим влияние наличия дома у студентов доступа к интернету (рис. 6). Из гистограммы видно, что средняя итоговая оценка у студентов без доступа к интернету дома оказалась несколько выше, чем у студентов с доступом к интернету.

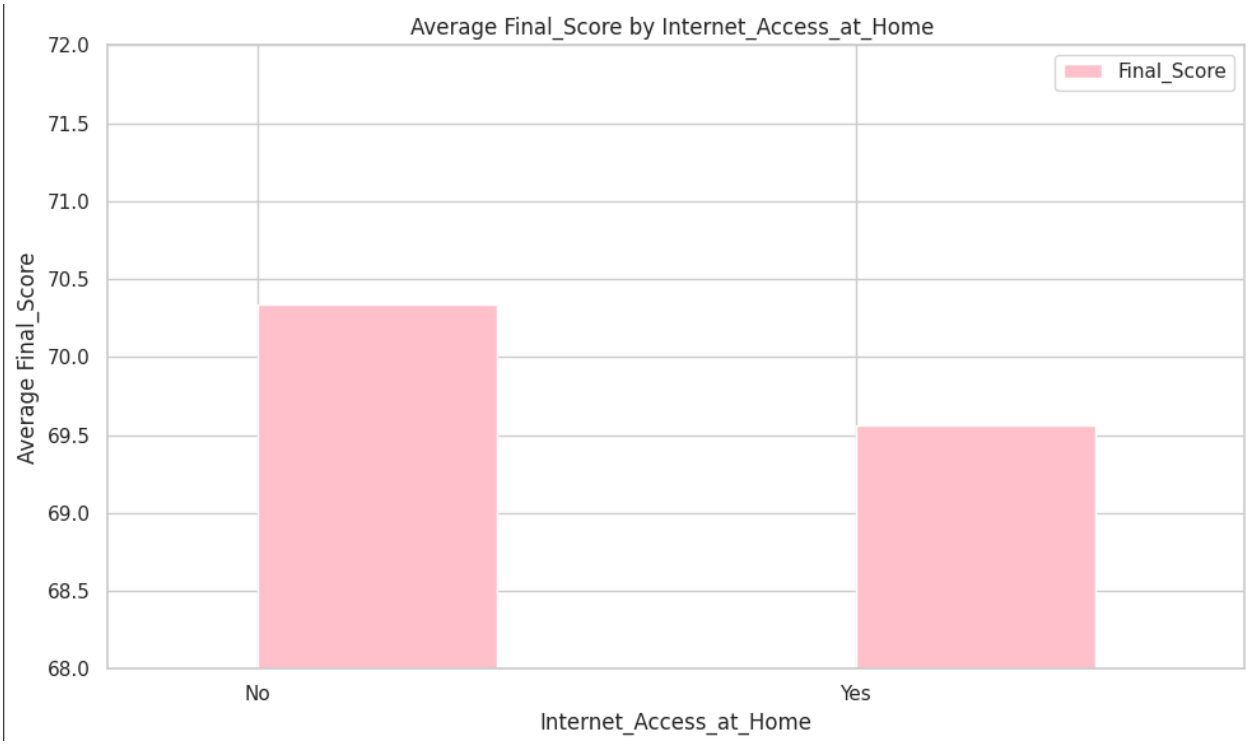


Рисунок 6 – Зависимость между итоговой оценкой и доступом к интернету дома

Проанализируем эту же зависимость, но с разбиением по полу (рис. 7): студенты женского пола в среднем показали более высокий результат по сравнению со студентами мужского пола при отсутствии интернета, тогда как для группы с доступом к интернету ситуация оказалась прямо противоположной. Следует учесть, что распределение между анализируемыми группами является достаточно неравномерным (4500/500 – рис. 8), что говорит о низкой показательности данных.

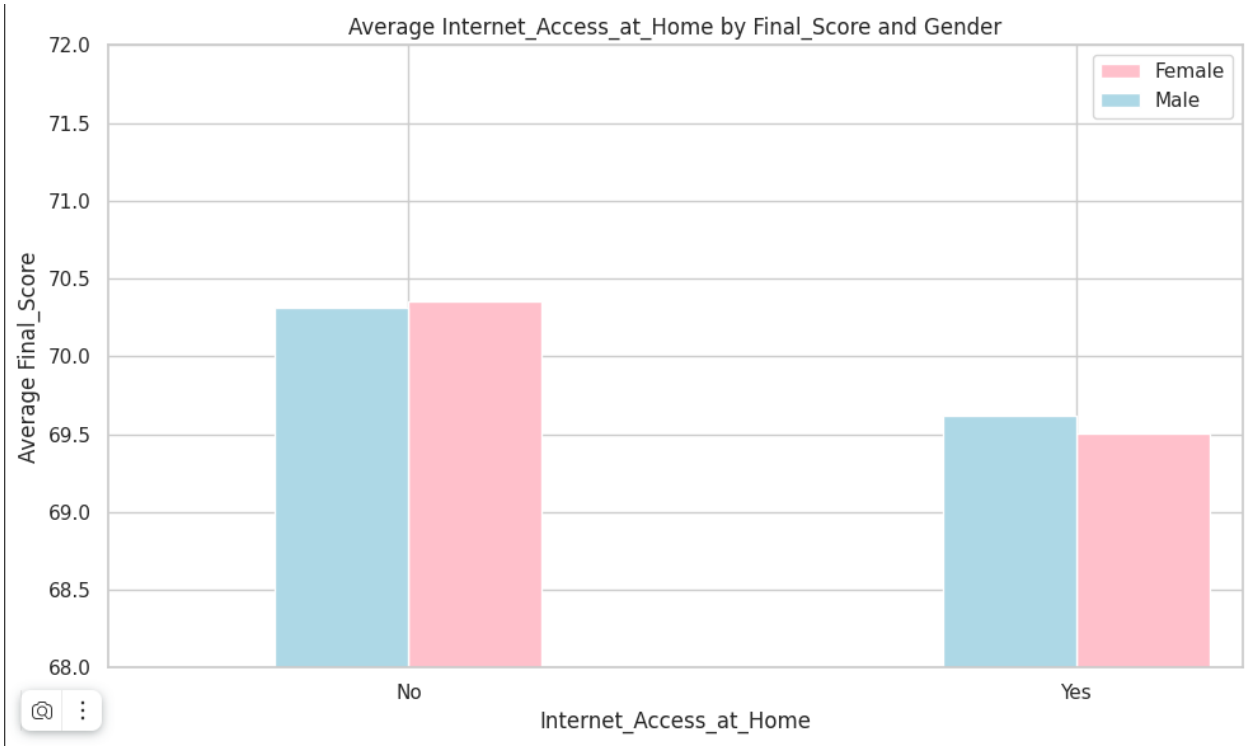


Рисунок 7 - Зависимость между итоговой оценкой и доступом к интернету дома с разбиением по полу

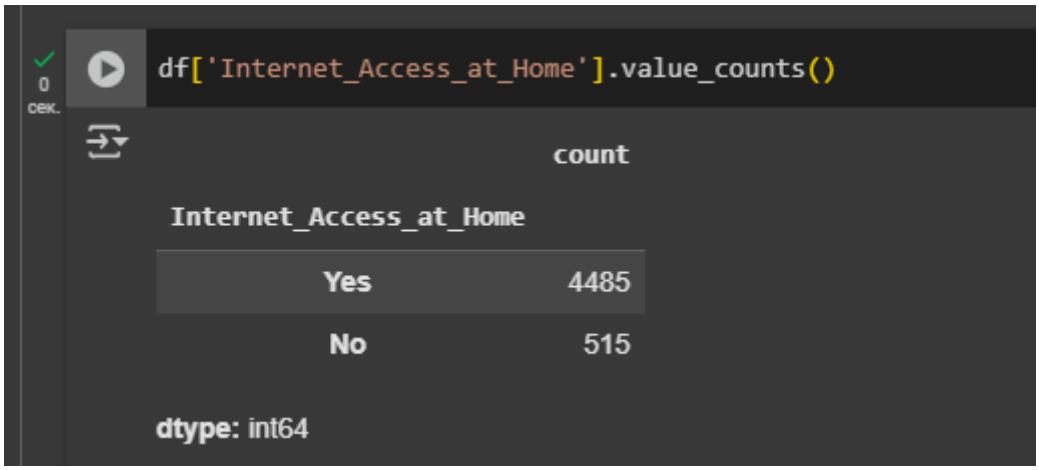


Рисунок 8 – Распределение по доступу к интернету дома

Далее рассмотрим зависимость оценок студентов от высшей ступени образования их родителей (рис. 9). Из гистограмм видно, что наивысшие оценки получили студенты, родители которых не имели высшего образования. Это может говорить о низком влиянии образования родителей на способности студентов. Более того, данная взаимосвязь может объясняться желанием студентов получить высшее образование по причине неудовлетворенности уровнем жизни семьи, где родители не имеют высшего образования.

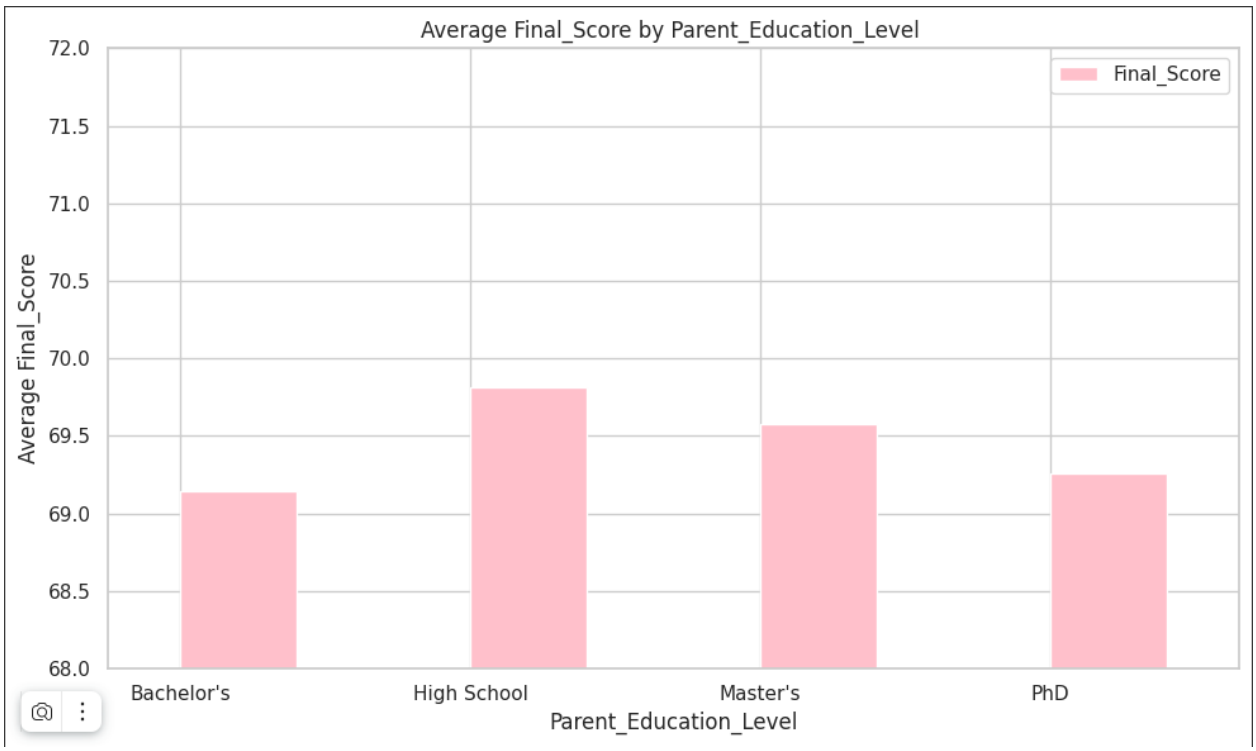


Рисунок 9 – Зависимость оценок от уровня образования родителей

Далее рассмотрим зависимость между итоговой оценкой и уровнем дохода семьи (рис. 10). По полученным гистограммам видно, что студенты из семей со средним уровнем достатка в среднем справлялись с заданиями лучше, тогда как студенты из семей с высоким и низким уровнями достатка справлялись в одинаковой степени хуже.

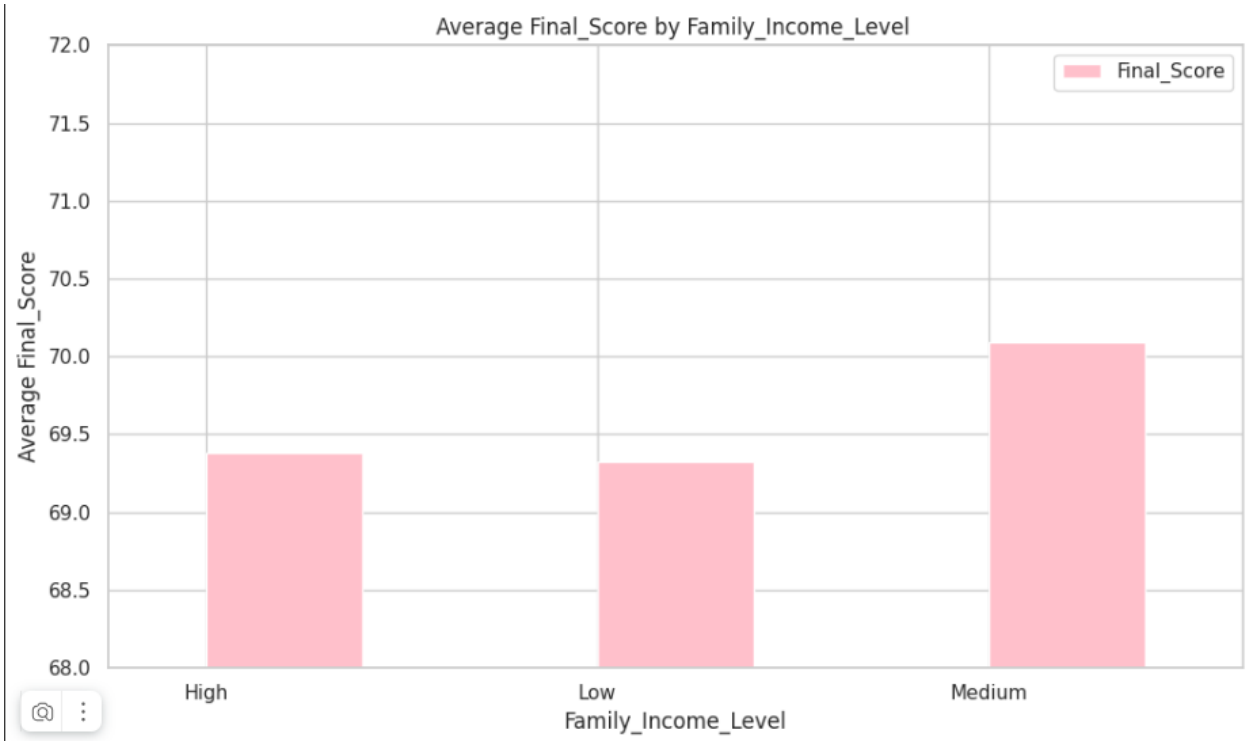


Рисунок 10 – Зависимость оценок от уровня дохода семьи

Далее рассмотрим зависимость итоговых оценок от уровня стресса студентов. Полученная гистограмма говорит о том, что уровень стресса слабо влияет на оценки студентов, при этом студенты со средним уровнем стресса (4-6) показали наивысшие результаты в итоговом тесте.

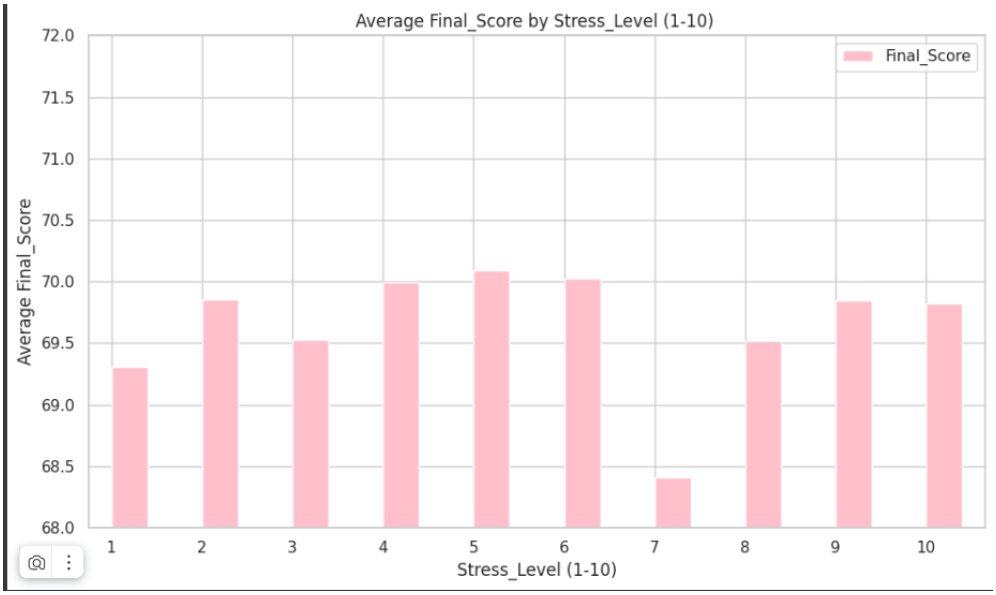


Рисунок 11 – Зависимость оценок от среднего уровня стресса студента

Далее на рисунке 12 представлена гистограмма, отражающая зависимость между оценками студентов и средней длительностью сна. Из неё видно, что студенты с наивысшими оценками (A, B) в среднем спят меньше остальных.

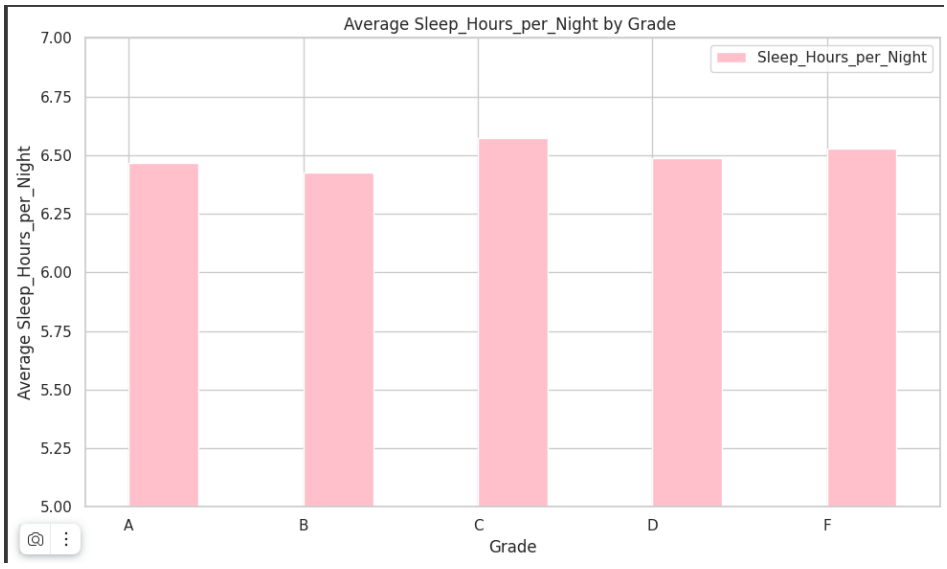
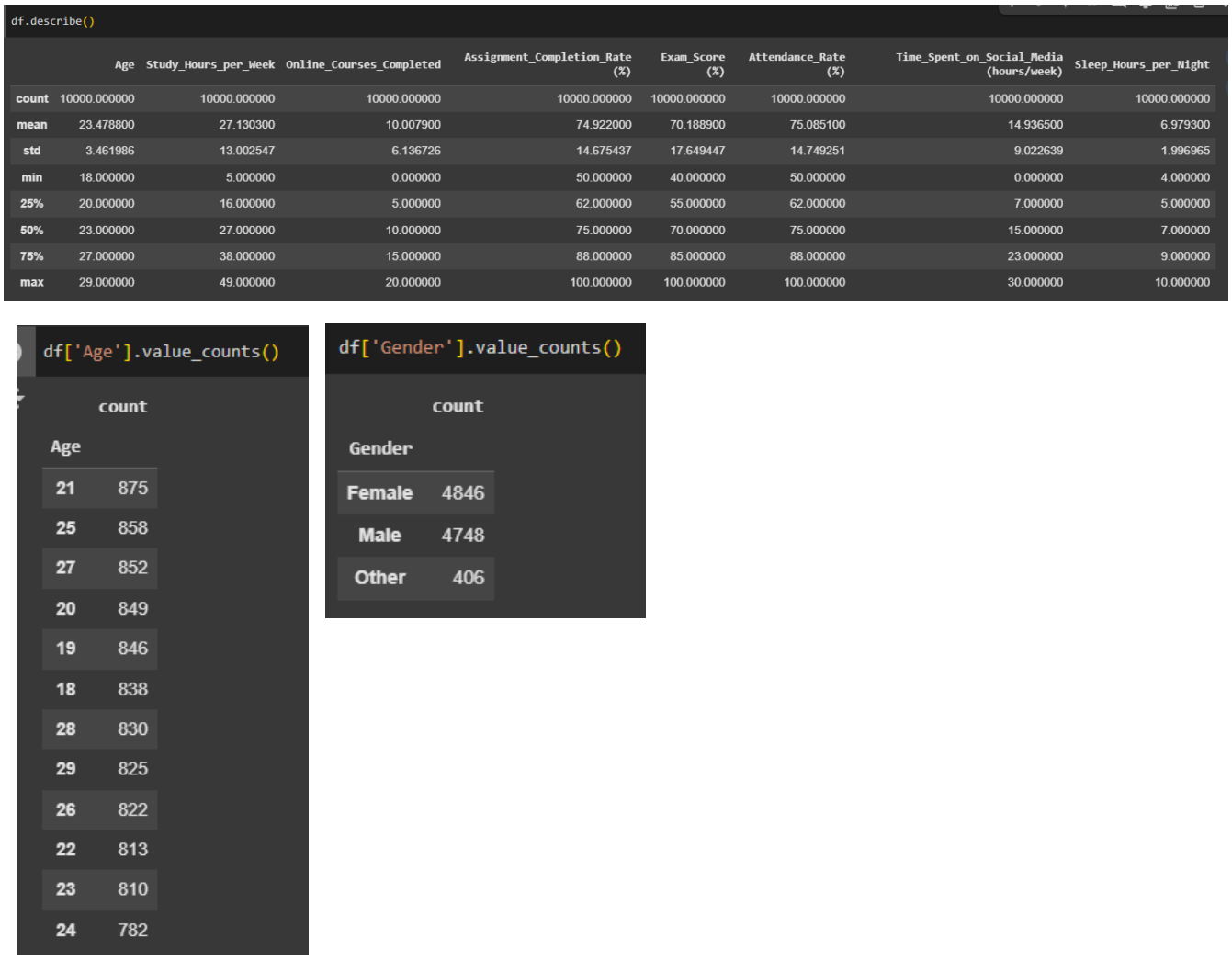


Рисунок 12 – Зависимость оценки от среднего времени сна в сутки

Рассмотрим набор данных с количеством записей в 10 000:

В наборе данных представлены записи о результатах обучения студентов в возрасте от 18 до 29 (рис.13).



```
df.describe()
```

	Age	Study_Hours_per_Week	Online_Courses_Completed	Assignment_Completion_Rate (%)	Exam_Score (%)	Attendance_Rate (%)	Time_Spent_on_Social_Media (hours/week)	Sleep_Hours_per_Night
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	23.478800	27.130300	10.007900	74.922000	70.188900	75.085100	14.936500	6.979300
std	3.461986	13.002547	6.136726	14.675437	17.649447	14.749251	9.022639	1.996965
min	18.000000	5.000000	0.000000	50.000000	40.000000	50.000000	0.000000	4.000000
25%	20.000000	16.000000	5.000000	62.000000	55.000000	62.000000	7.000000	5.000000
50%	23.000000	27.000000	10.000000	75.000000	70.000000	75.000000	15.000000	7.000000
75%	27.000000	38.000000	15.000000	88.000000	85.000000	88.000000	23.000000	9.000000
max	29.000000	49.000000	20.000000	100.000000	100.000000	100.000000	30.000000	10.000000

```
df['Age'].value_counts()
```

Age	count
21	875
25	858
27	852
20	849
19	846
18	838
28	830
29	825
26	822
22	813
23	810
24	782

```
df['Gender'].value_counts()
```

Gender	count
Female	4846
Male	4748
Other	406

Рисунок 13 – Структура набора данных

С целью определения корреляций между различными столбцами построим тепловые карты, используя функции corr() и phik_matrix() (рис.14, рис. 15)

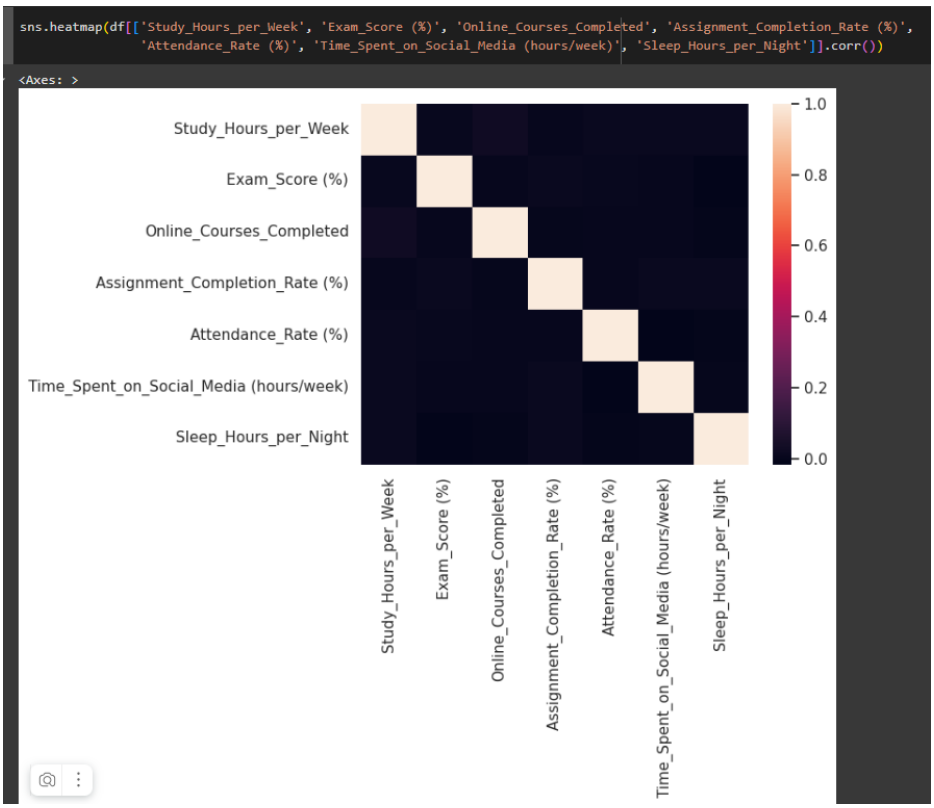


Рисунок 14 - Корреляция между столбцами набора данных с использованием функции corr()

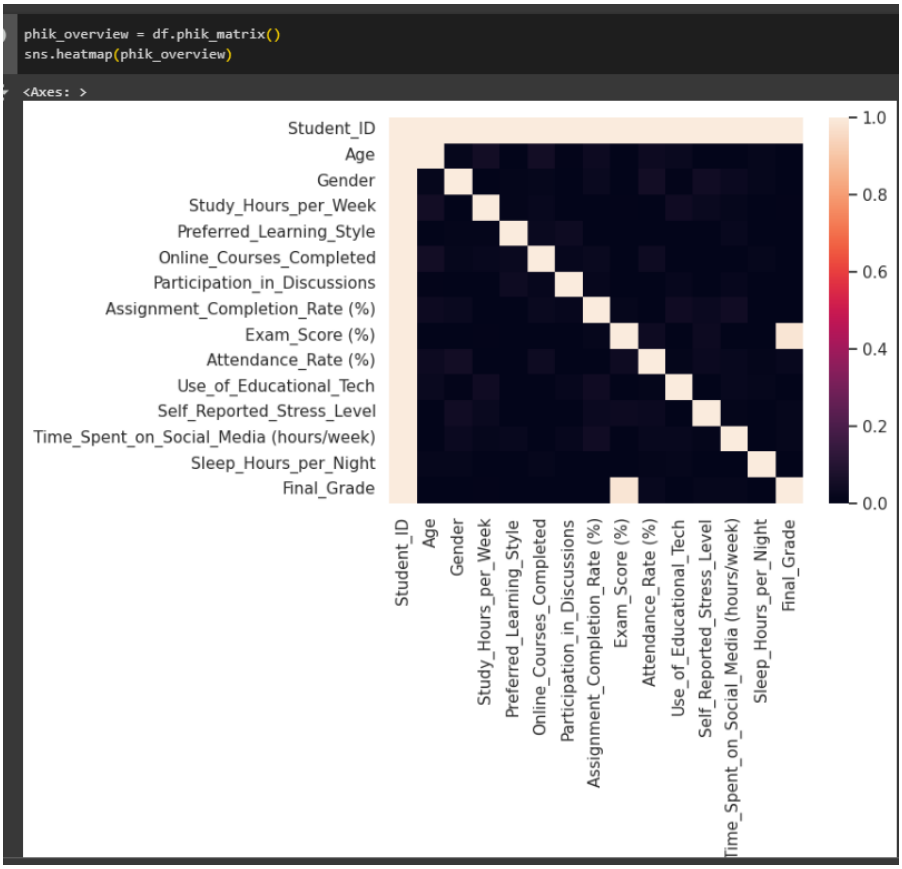


Рисунок 15 - Корреляция между столбцами набора данных с использованием функции phik_matrix()

Рассмотрим зависимость итоговых оценок от посещаемости (рис. 16, рис. 17). По гистограммам видно, что в среднем посещаемость студентов с оценкой A выше, чем в других группах, что соответствует результатам анализа первого набора данных и подтверждает связь между этими показателями.

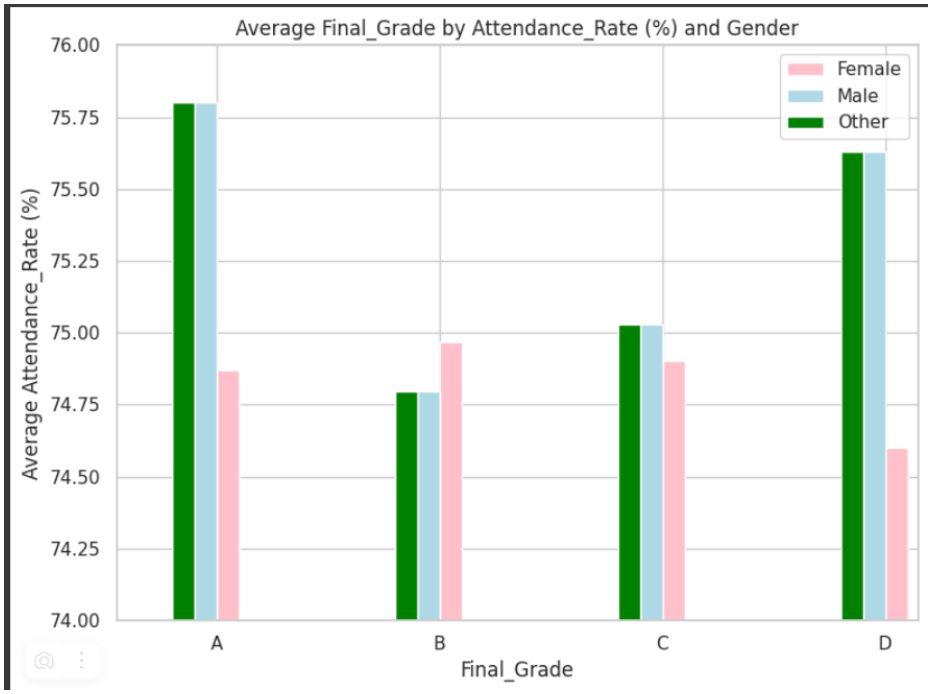


Рисунок 16 - Взаимосвязь между итоговой оценкой и посещаемостью с разбиением по полу

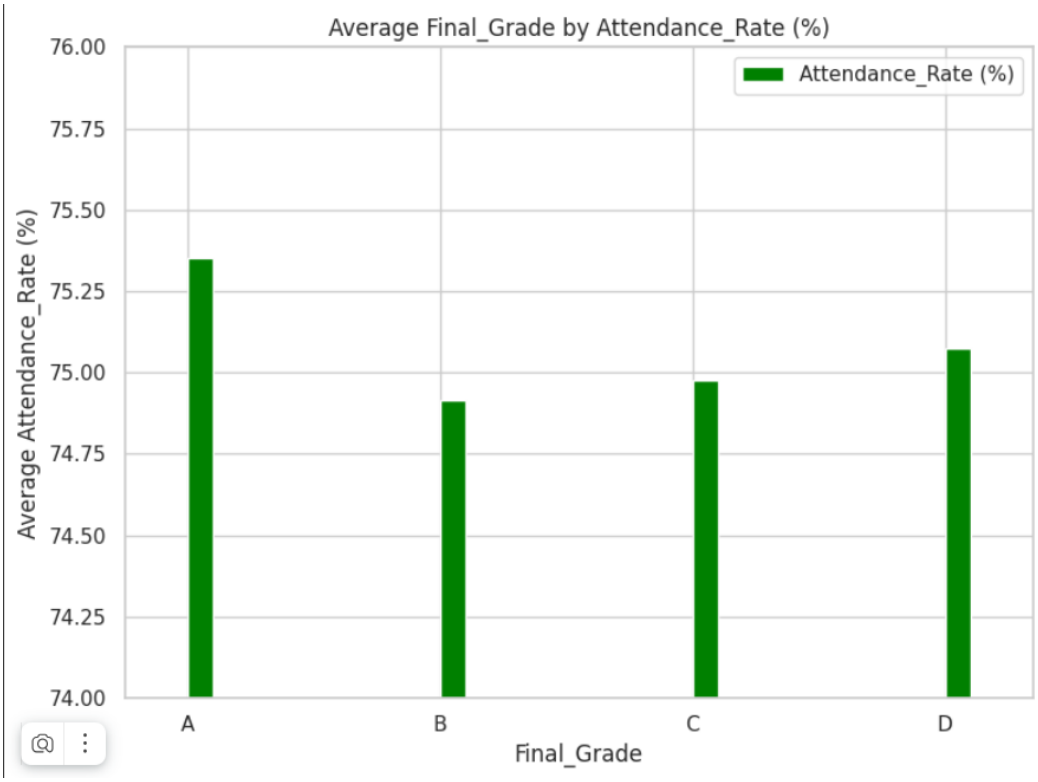


Рисунок 17 – Взаимосвязь между итоговой оценкой и посещаемостью

Далее рассмотрим взаимосвязь между итоговыми оценками и средним количеством времени, проводимом в социальных сетях (рис. 18, рис. 19). Полученные гистограммы говорят о слабой зависимости итоговых оценок от времени, затрачиваемого студентами на социальные сети.

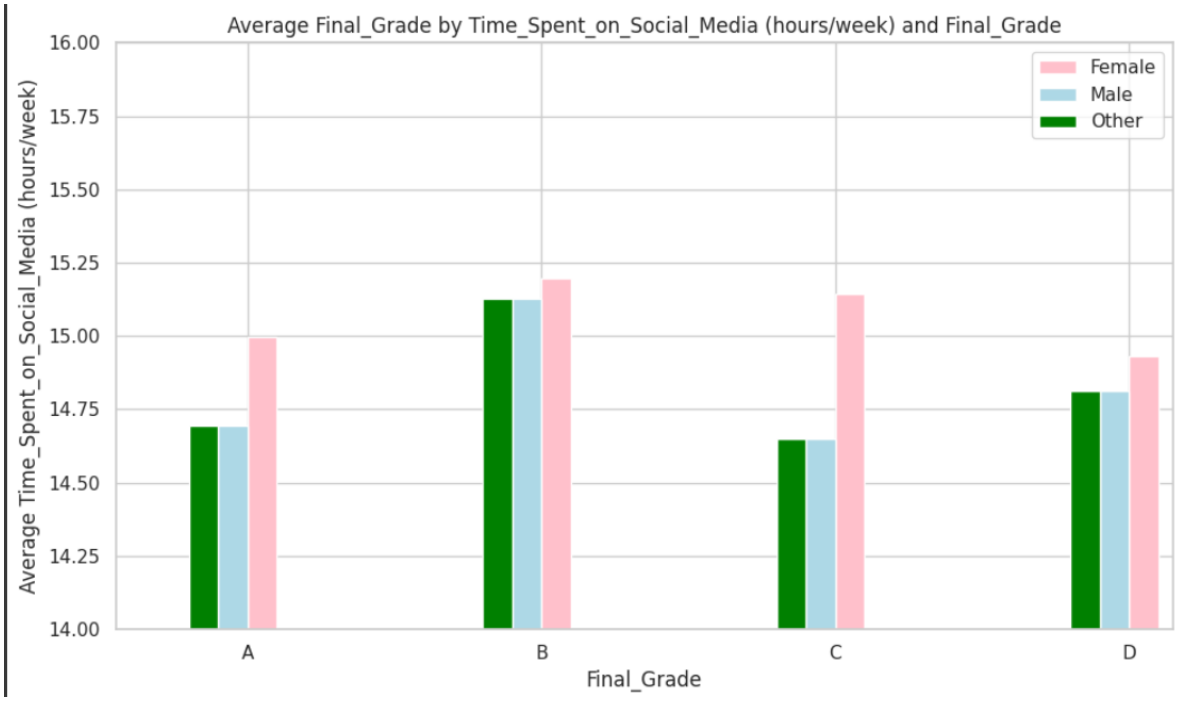


Рисунок 18 - Взаимосвязь итоговых оценок и количества потраченного времени на социальные сети с разбиением по полу

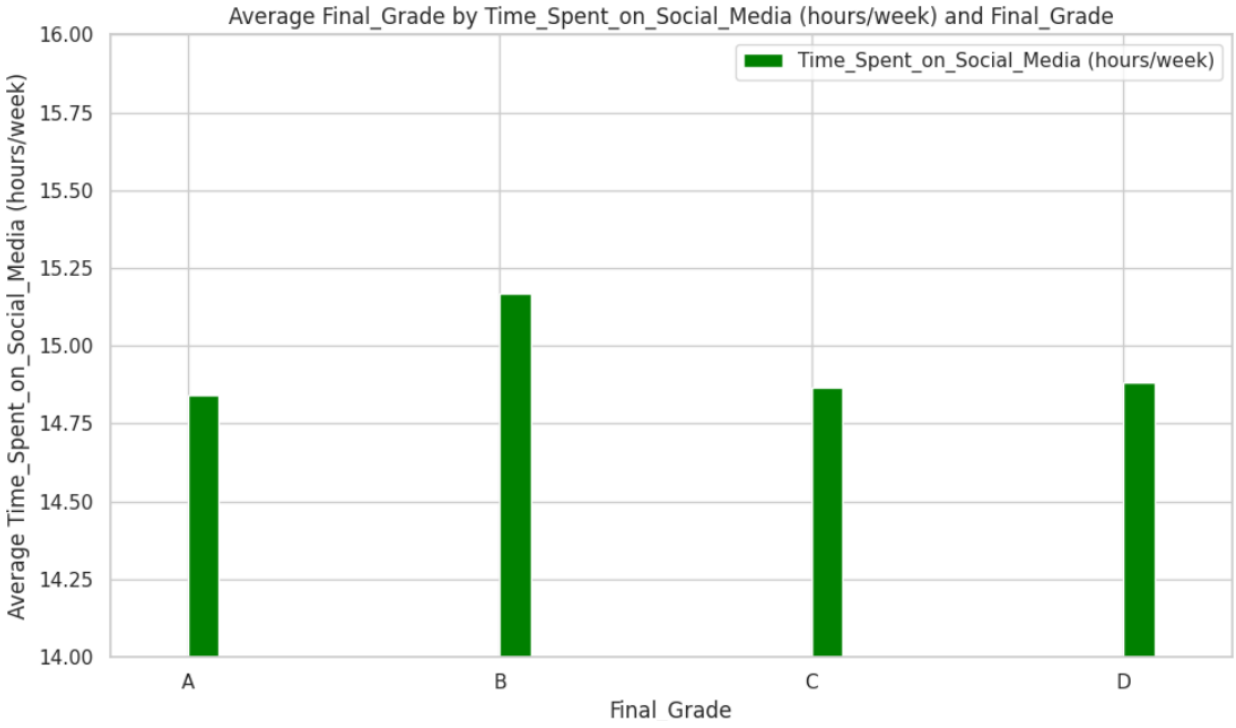


Рисунок 19 – Взаимосвязь итоговых оценок и количества потраченного времени на социальные сети

Рассмотрим взаимосвязь итоговых оценок и предпочтительного стиля обучения студентов (рис. 20, рис.21). Полученные гистограммы говорят о том, что в среднем наилучшим способом обучения является восприятие информации на слух, в то время как по отдельности студенты мужского пола показали наилучшие результаты, воспринимая информацию визуально, а студенты женского пола показали наивысшие результаты, воспринимая информацию на слух.

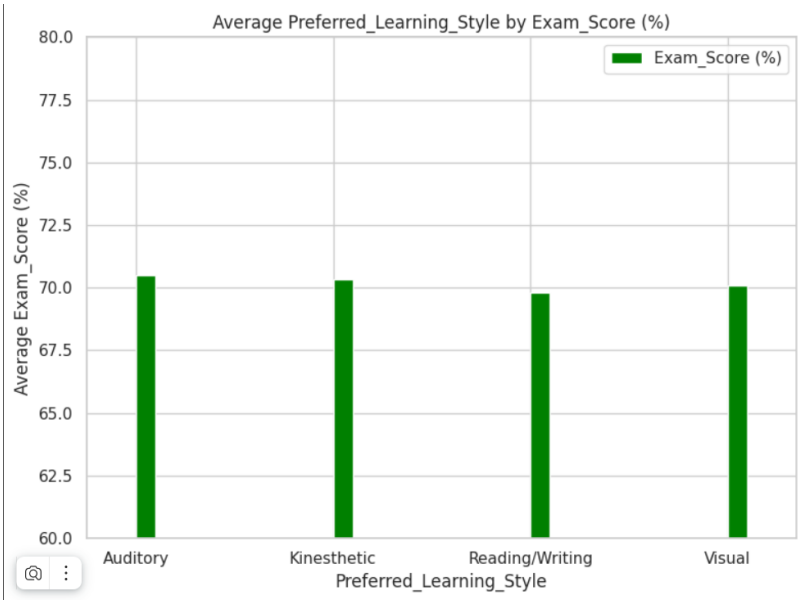


Рисунок 20 – Взаимосвязь между оценкой за экзамен и предпочтительным стилем обучения

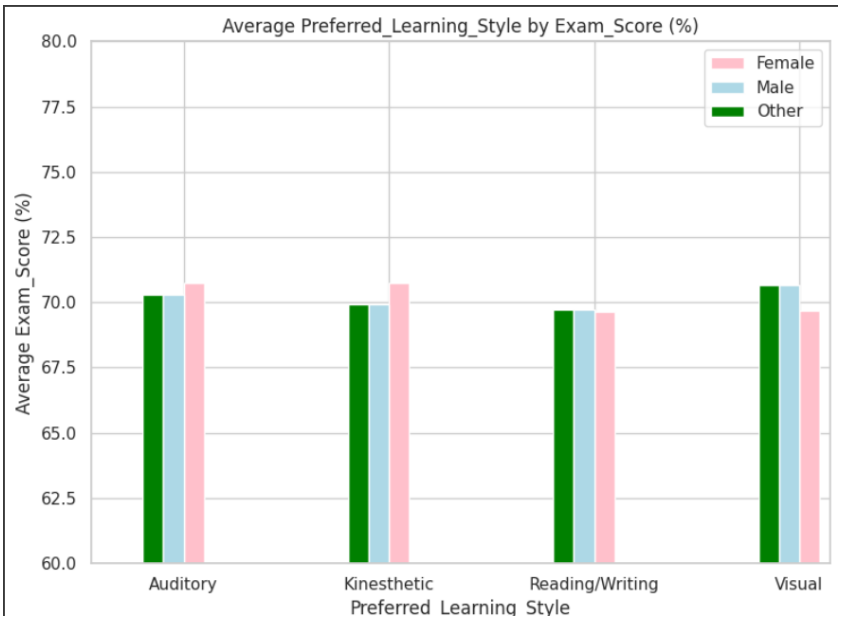


Рисунок 21 – Взаимосвязь между оценкой за экзамен и предпочтительным стилем обучения с разбиением по полу

Выводы

В результате проведенного исследования были рассмотрены взаимосвязи между итоговыми оценками студентов и различными аспектами их жизни и окружающей среды. Высокая зависимость прослеживается между итоговыми оценками и посещаемостью студентов, что говорит о важности коммуникации студентов с наставниками для эффективного усвоения материала. Наиболее эффективным методом представления новой информации для студентов мужского пола является визуализация, в то время как студенты женского пола лучше усваивают информацию на слух.

Парадоксально, но наличие высшего образования различных ступеней у родителей студентов не коррелирует с успеваемостью студентов, напротив – студенты, родители которых не имеют высшего образования показывают более высокие результаты в обучении. Высокий доход в семье так же не гарантирует высоких оценок при обучении, при этом наилучшие результаты показывают студенты из семей со средним уровнем дохода.

Было определено, что большинство взаимосвязей выражены слабо. Это может быть объяснено двумя причинами. Первая – сложность сбора данных. Поскольку информация была собрана путём анкетирования студентов, некоторые записи могут искаженно отражать реальную картину в результате преднамеренного предоставления студентами ложных данных.

Ссылка на код, использованный в ходе исследования:
https://github.com/Loomman1/StudentLifeResearch.git