Московский государственный технический университет им. Н.Э. Баумана

Факультет «Головной учебно-исследовательский и методический центр

профессиональной реабилитации лиц с ограниченными возможностями здоровья (инвалидов)»

Кафедра «Системы обработки информации и управления»



**Рубежный контроль 2**

**по дисциплине «Методы машинного обучения в АСОИУ»**

**"** **Методы обработки текстов "**

СТУДЕНТ:

студент группы ИУ5Ц-21М

Москалик А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

Москва, 2024

# Задание:

# Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета (кроме примера, который рассматривался в лекции). Классификация может быть бинарной или многоклассовой. Целевой признак из выбранного Вами датасета может иметь любой физический смысл, примером является задача анализа тональности текста.

# Необходимо сформировать два варианта векторизации признаков - на основе CountVectorizer и на основе TfidfVectorizer.

# В качестве классификаторов необходимо использовать два классификатора по варианту для Вашей группы:

# Группа - ИУ5-21М, ИУ5И-21М, ИУ5Ц-21М

# Классификатор №1 – KneighborsClassifier

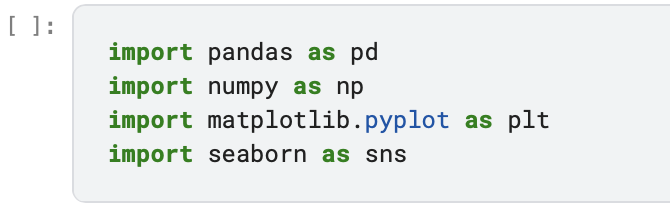
# Классификатор №2 – LogisticRegression

# 

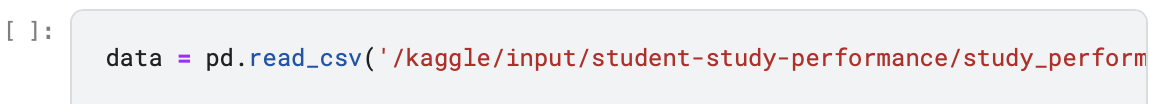
# Ход работы

# Подготовка

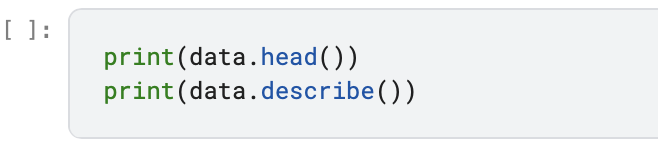
# Подключение библиотек



Загрузка датасета



1. **Предварительный анализ данных**



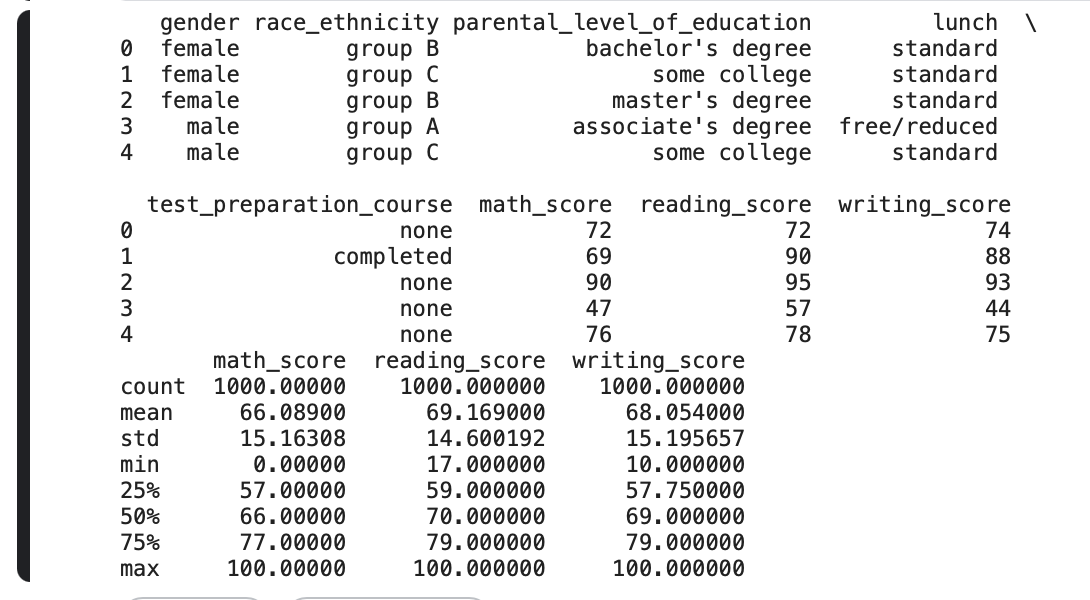


Таблица отображает два разных вида статистической информации о датасете:

*data.head()* показывает первые пять строк датасета, как

- пол (gender),

- этническая принадлежность (race/ethnicity),

- уровень образования родителей (parental level of education),

- завтрак (lunch),

- прохождение подготовительных курсов (test preparation course),

- баллы по математике (math score), чтению (reading score) и письму (writing score).

*data.describe()* предоставляет описательную статистику для всех числовых столбцов в датасете.

Это включает:

- Количество значений (count)

- Среднее значение (mean)

- Стандартное отклонение (std), показывающее разброс данных

- Минимальное значение (min) -25-й перцентиль (25%), который является первым квартилем

- Медианное значение (50%), также известное как второй квартиль

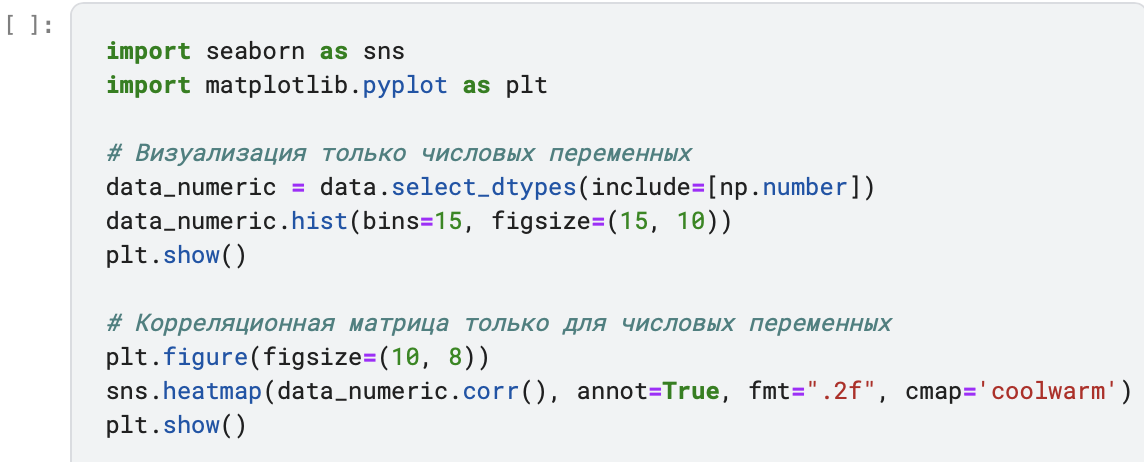
- 75-й перцентиль (75%), который является третьим квартилем

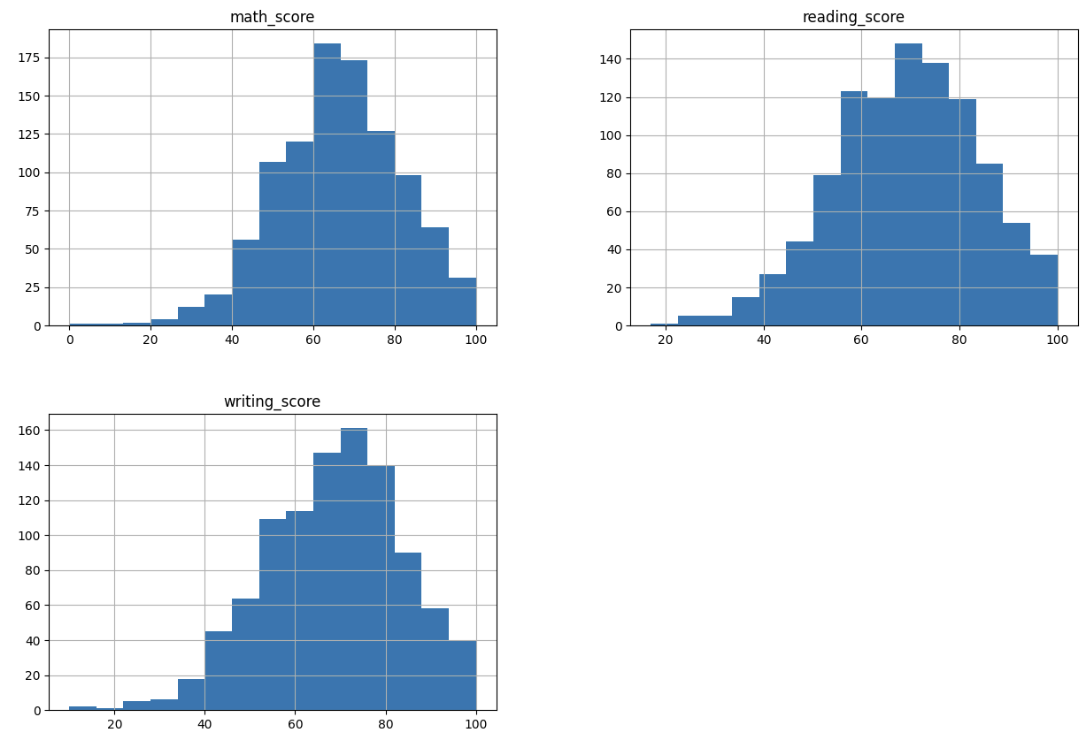
- Максимальное значение (max)

1. **Исследование данных**

Первый фрагмент кода выбирает только числовые столбцы из DataFrame *data* с помощью *select\_dtypes(include=[np.number]*). После этого для каждой числовой переменной строится гистограмма. Гистограммы показывают распределение данных по каждому числовому признаку с помощью столбиков, где высота каждого столбика соответствует количеству наблюдений в каждом интервале. Опция *bins=15* указывает, что для каждой гистограммы следует использовать 15 интервалов (столбиков), а *figsize=(15, 10)* задаёт размер всего рисунка, на котором будут расположены гистограммы.

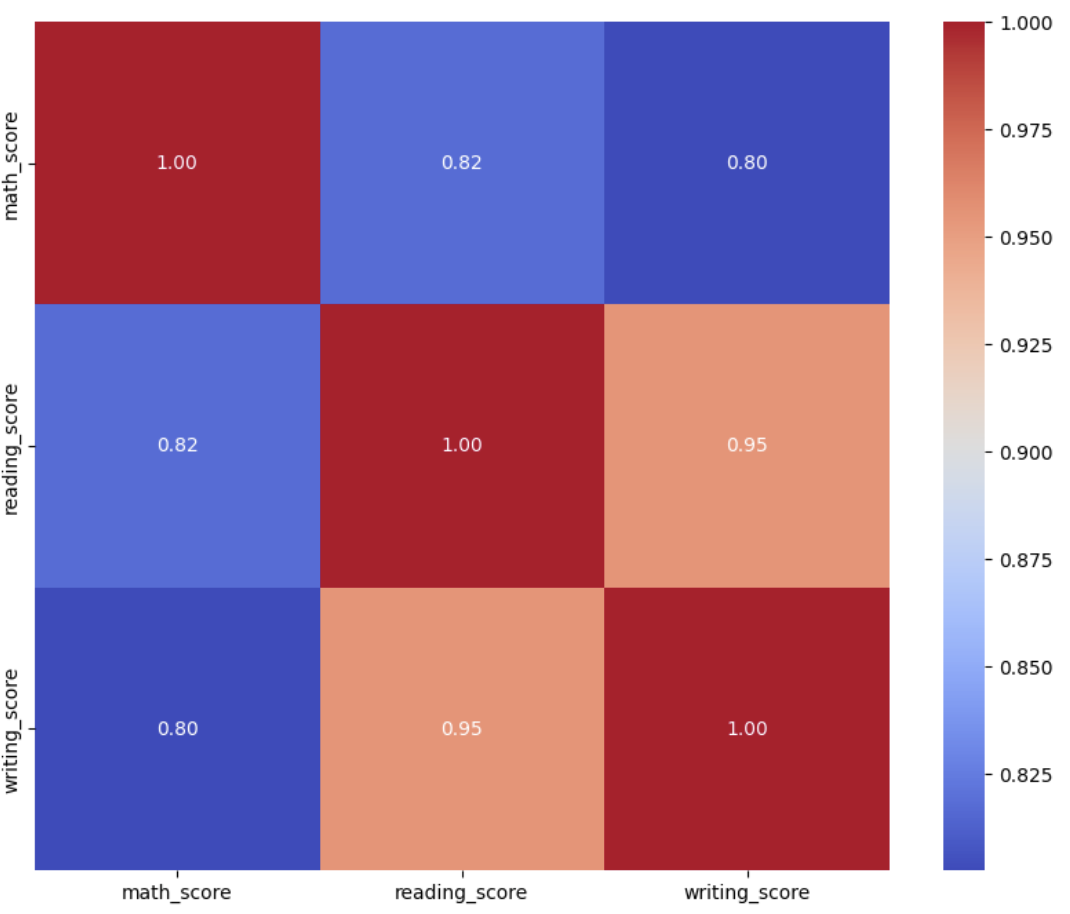
Второй фрагмент кода строит корреляционную матрицу для числовых переменных и визуализирует её с помощью тепловой карты (*heatmap*). Корреляционная матрица позволяет оценить степень линейной связи между парами переменных. Значения корреляции варьируются от -1 до 1, где 1 означает идеальную положительную корреляцию, -1 — идеальную отрицательную корреляцию, а 0 — отсутствие линейной связи. Аргументы *annot=True и fmt=".2f"* указывают на необходимость отображения числовых значений корреляций с двумя десятичными знаками на тепловой карте, а *cmap='coolwarm'* задаёт цветовую схему.



****

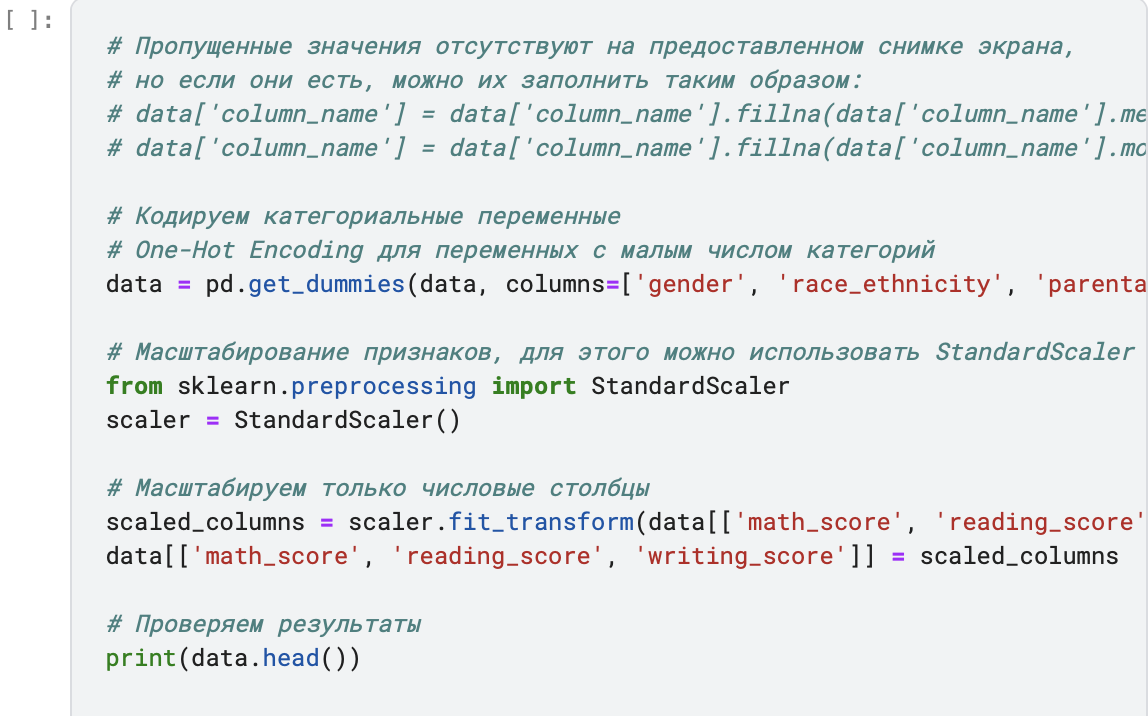
На представленных диаграммах показаны гистограммы, отображающие распределение оценок студентов по трем предметам: математике (math\_score), чтению (reading\_score) и письму (writing\_score).

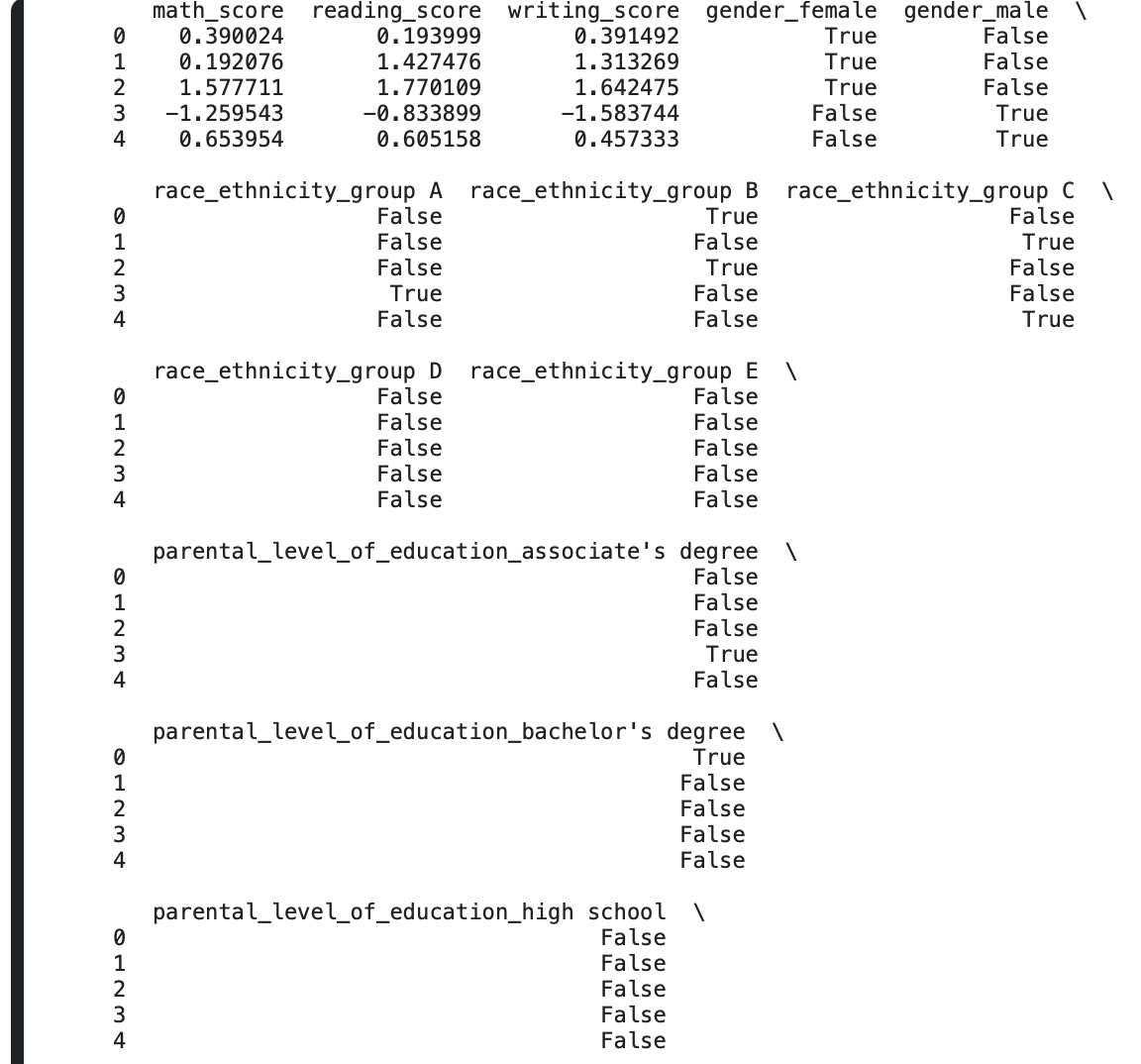
Все три гистограммы демонстрируют, что оценки сосредоточены в среднем диапазоне, с относительно меньшим количеством очень низких и очень высоких оценок. Это типично для школьных оценок, где большинство студентов проявляют средний уровень успеваемости, а оценки на краях спектра (очень высокие или очень низкие) менее обычны.

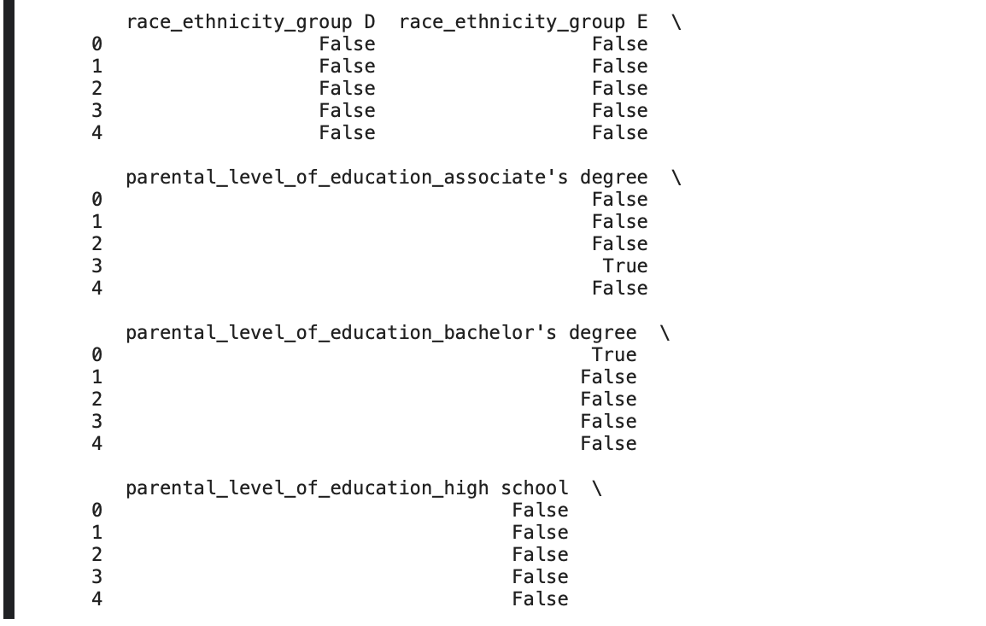
****

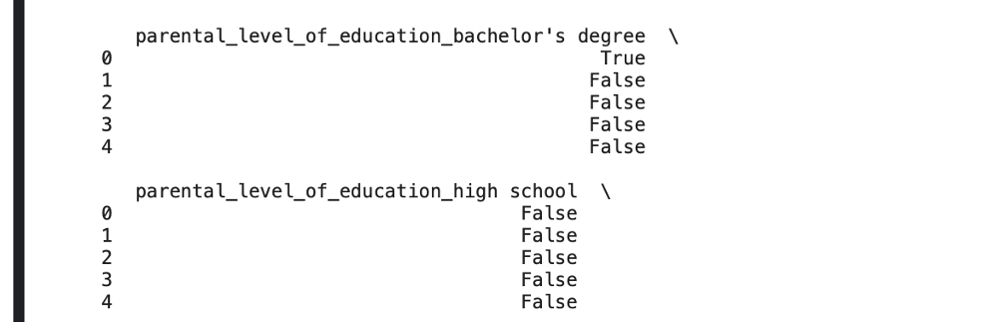
Тепловая карта показывает корреляционные связи между оценками по математике, чтению и письму. Все три переменные имеют сильную положительную корреляцию друг с другом, значения варьируются от 0.80 до 0.95. Это означает, что учащиеся, которые хорошо справляются с одним предметом, склонны показывать хорошие результаты и в других предметах. Самая сильная корреляция наблюдается между чтением и письмом (0.95), что указывает на то, что навыки в этих областях развиты схожим образом.

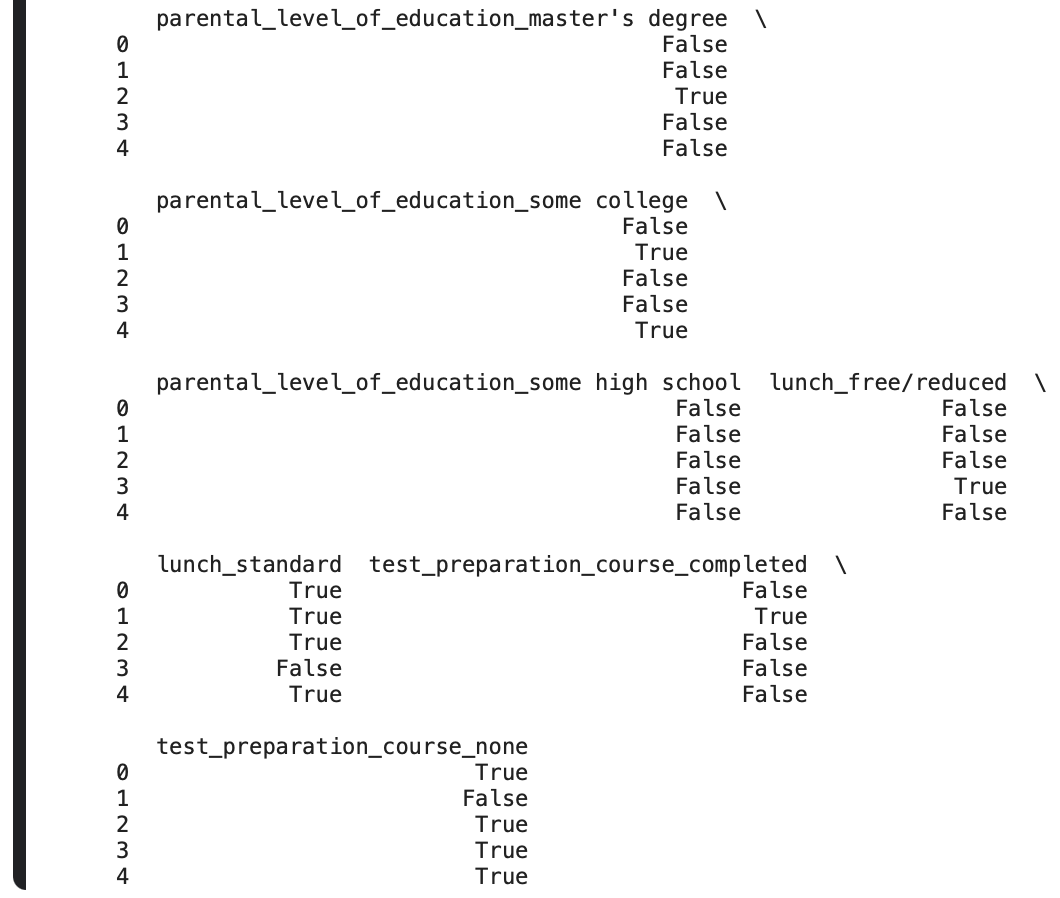
**3. Очистка данных**











Эти данные представляют обработанный датасет, где:

- math\_score, reading\_score, writing\_score - это стандартизированные баллы по математике, чтению и письму. Стандартизация преобразует распределение оценок так, чтобы среднее значение было 0, а стандартное отклонение — 1.

- gender\_female, gender\_male - это флаги, полученные после one-hot кодирования пола, где True означает, что учащийся соответствует категории, а False - не соответствует.

- столбцы (race\_ethnicity\_group A, race\_ethnicity\_group B и т.д.) также являются результатом one-hot кодирования и показывают принадлежность учащегося к определенной этнической группе.

- столбцы, связанные с parental\_level\_of\_education\_..., указывают на уровень образования родителей.

- lunch\_free/reduced, lunch\_standard - указывают на тип питания учащегося, где True означает, что данный тип питания учащегося соответствует названию столбца.

- test\_preparation\_course\_completed, test\_preparation\_course\_none - показывают, завершил ли учащийся подготовительные курсы. Значение True или False для каждой из этих dummy-переменных показывает, принадлежит ли наблюдение к соответствующей категории.

**4. Формирование гипотез и тестирование**

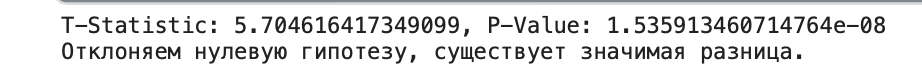
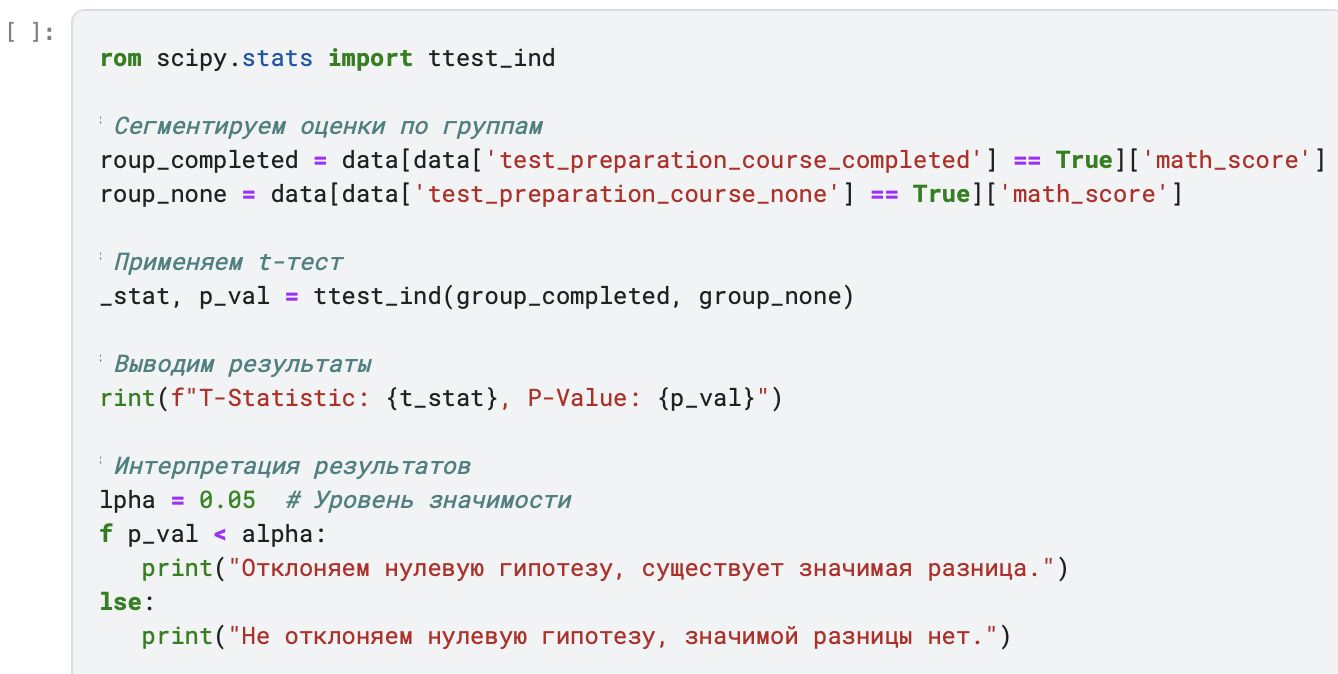
1. T-тест для сравнения математических оценок по курсу подготовки - это помогает определить, влияет ли участие в подготовительных курсах на успеваемость по математике.

2. Исследование зависимостей между категориальными и числовыми переменными - позволяет выявить, как различные категориальные признаки (например, пол или этническая принадлежность) влияют на числовые результаты (оценки).

3. Анализ распределений каждого признака в разрезе различных групп - изучает, как признаки распределяются среди разных групп, что может помочь выявить паттерны или необходимость корректировки данных.

4. Создание сводных таблиц для сравнения средних значений между различными группами - обеспечивает наглядное сравнение, которое может быть полезно для подтверждения или опровержения гипотез.

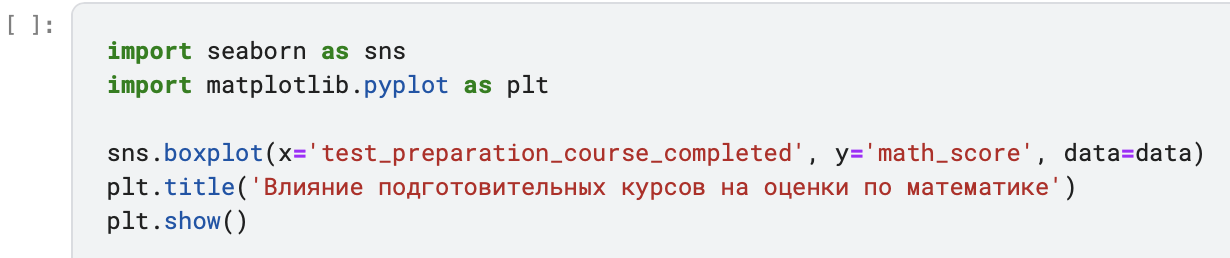
**T-тест для сравнения математических оценок по курсу подготовки**

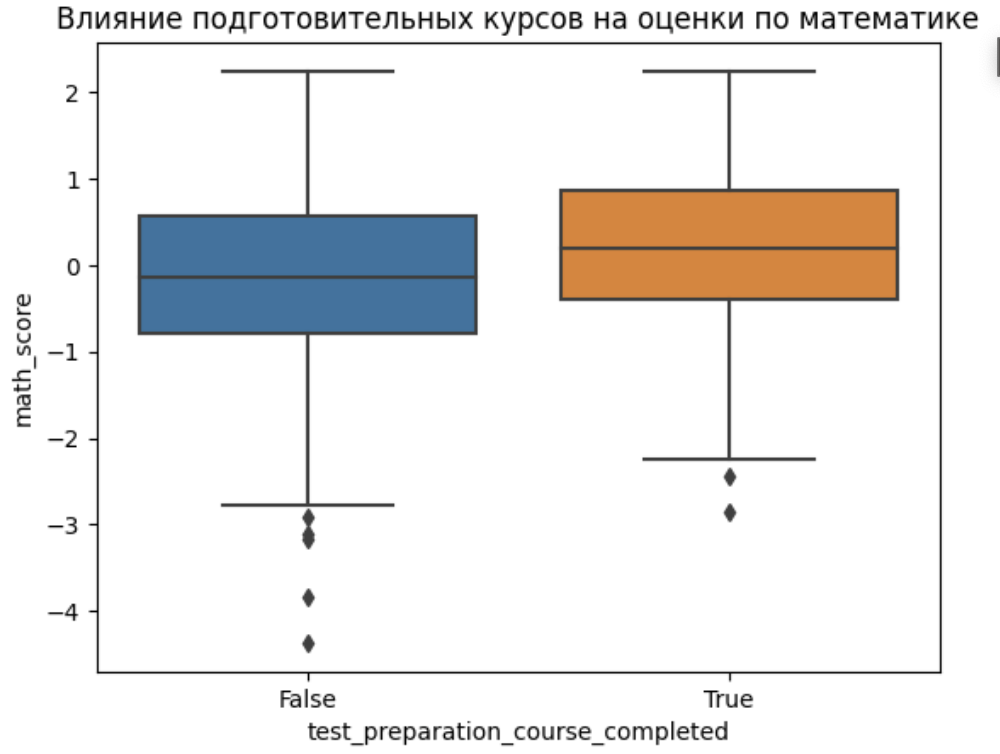


P-значение в результате говорит о вероятности получить данные, как минимум такие же экстремальные, как те, что были получены во время теста, если нулевая гипотеза верна. Очень маленькое P-значение, такое как 1.53e-08 (что эквивалентно 0.00000001539), гораздо меньше стандартного порога α = 0.05, и поэтому можно сделать вывод о том, что есть статистически значимая разница между группами.

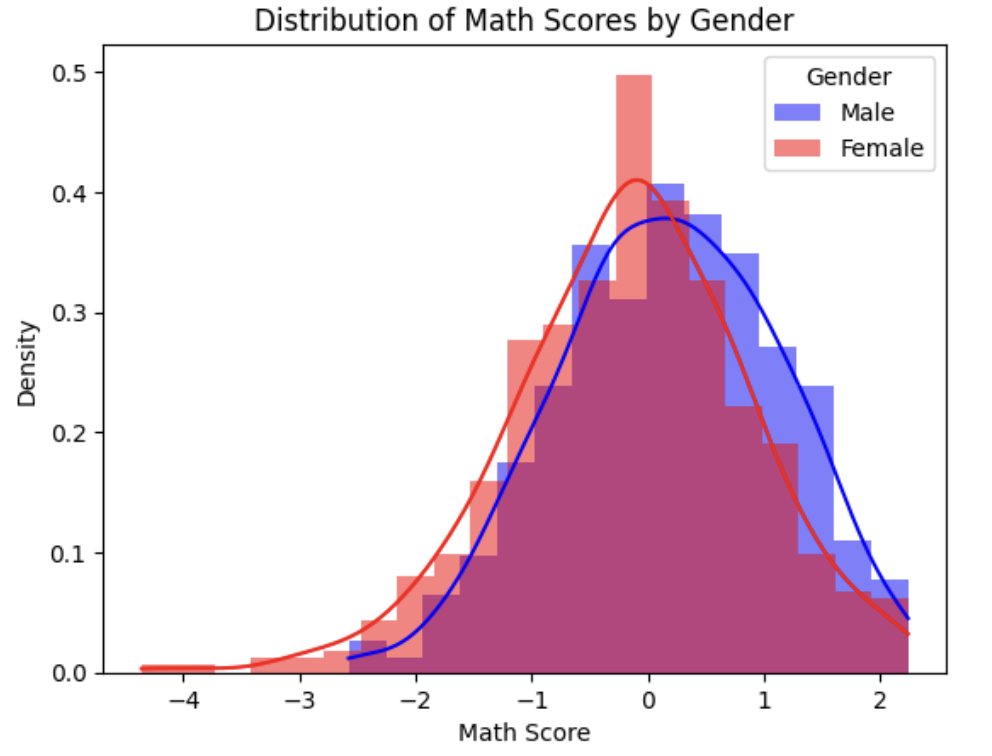
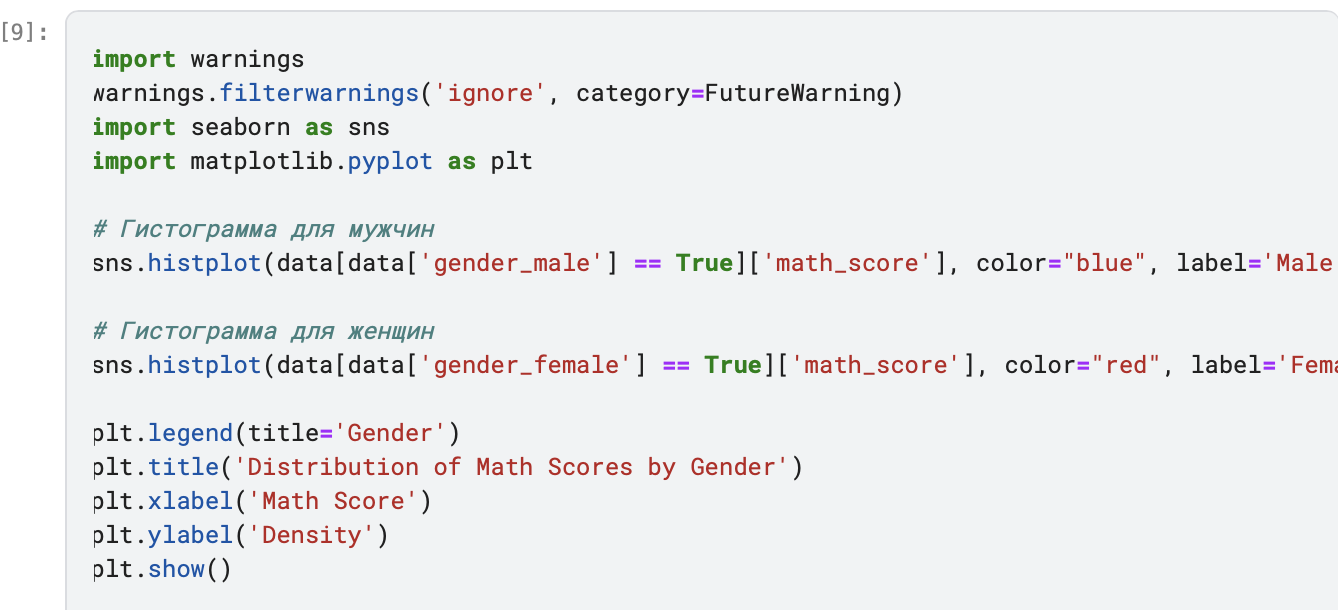
Значение T-статистики 5.704 указывает на то, насколько велико отклонение средних значений между группами в единицах стандартного отклонения. Большое значение T-статистики в сочетании с маленьким P-значением подтверждает, что различия между группами значимы с точки зрения статистики.

**Исследование зависимостей между категориальными и числовыми переменными**



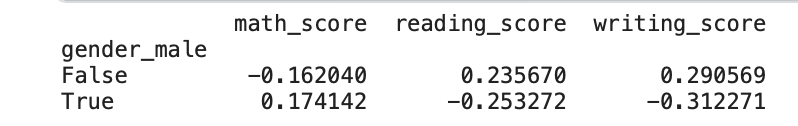
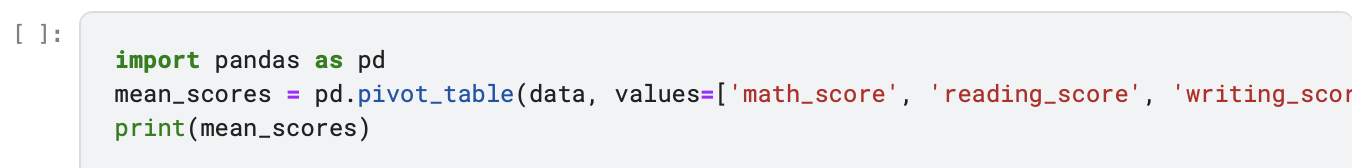


**Анализ распределений каждого признака в разрезе различных групп**



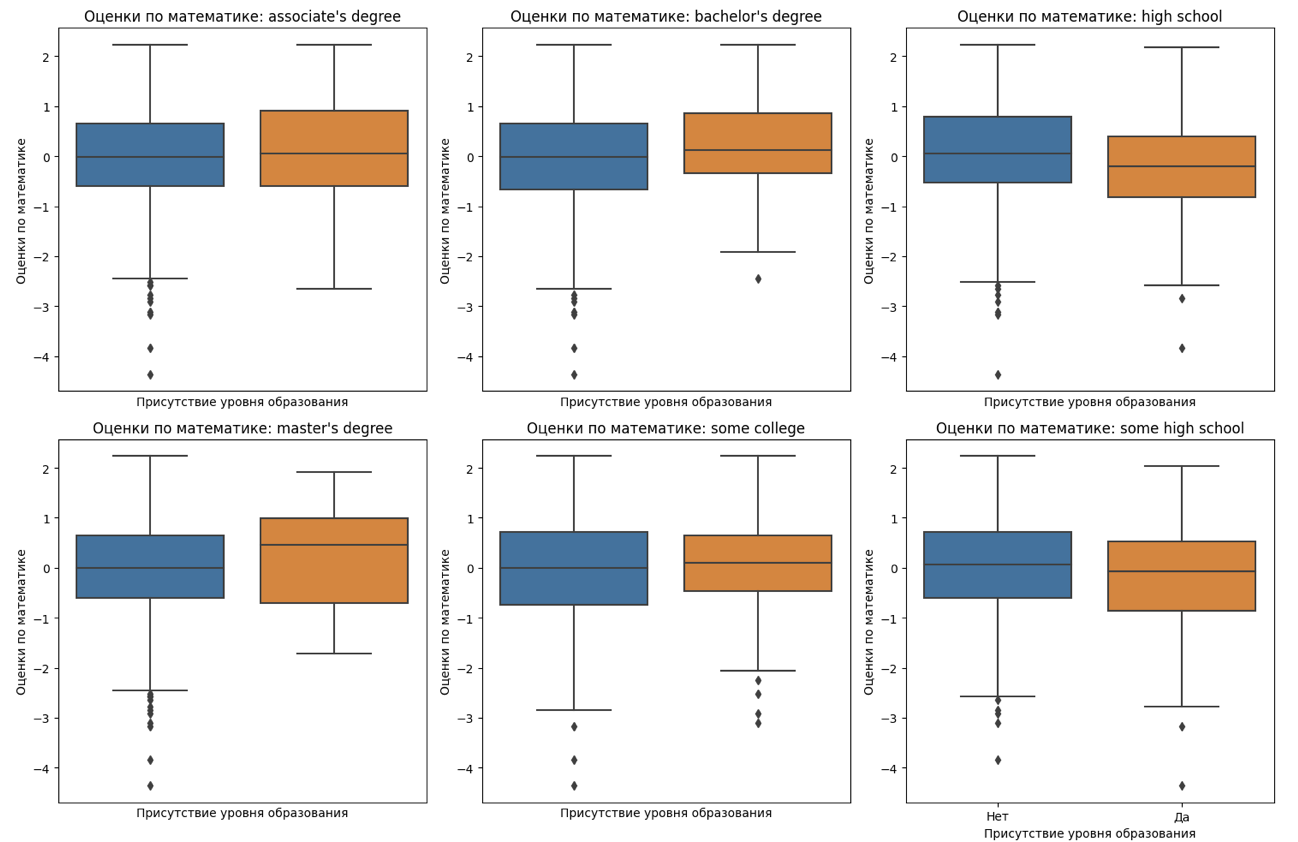
Этот код создает две наложенные гистограммы для мужчин и женщин, что позволяет визуально сравнить распределение оценок по математике. Аргумент stat="density" гарантирует, что площадь под гистограммой нормализуется, что делает сравнение более честным, особенно если размеры групп различаются.

**Создание сводных таблиц для сравнения средних значений между различными группами**



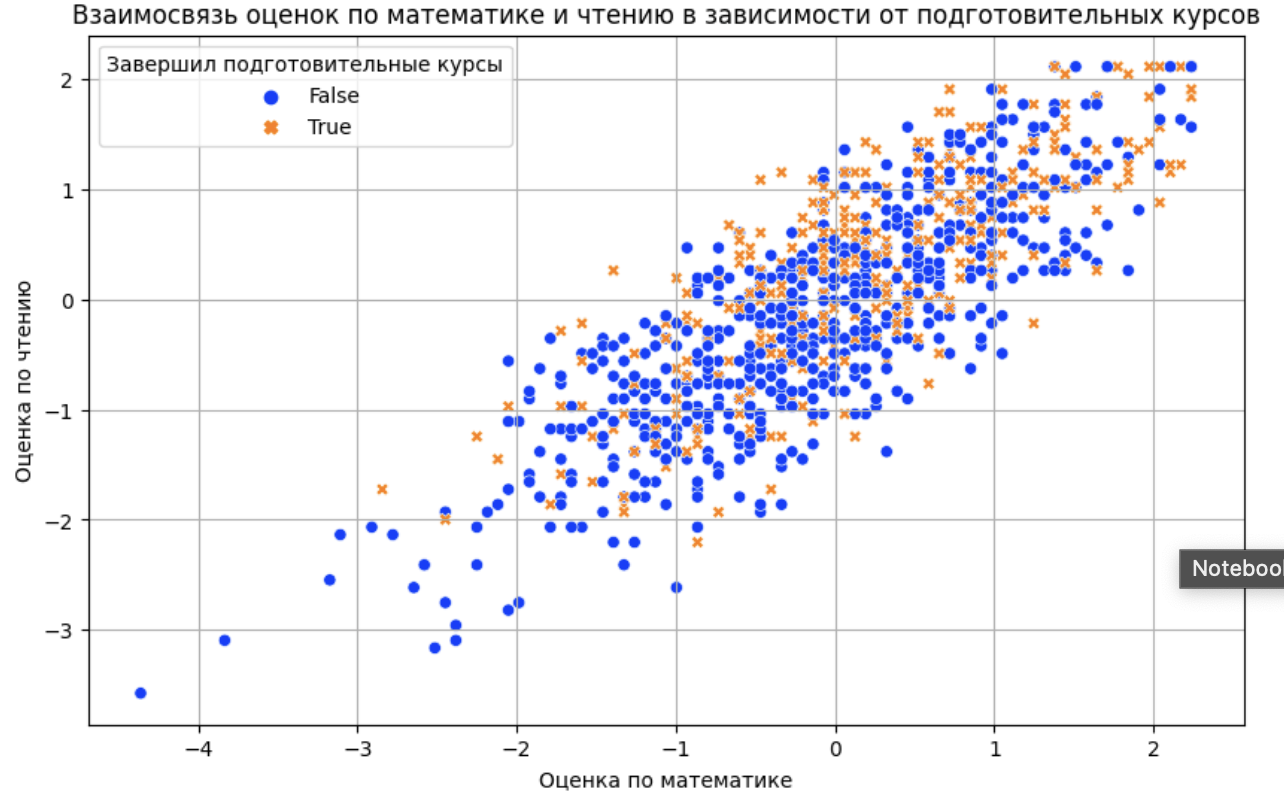
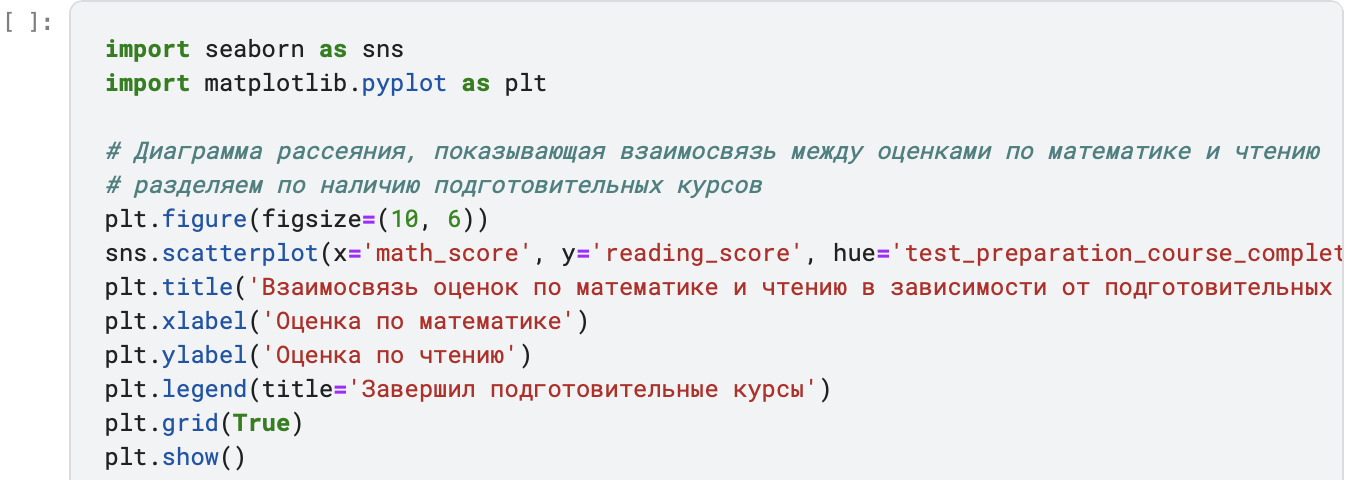
Таблица, представляет средние значения оценок по математике, чтению и письму для разных групп по полу (мужчины и женщины), где данные были стандартизированы (среднее = 0, стандартное отклонение = 1). Столбцы math\_score, reading\_score, и writing\_score показывают средние значения для каждой группы. False обозначает женщин, True — мужчин.

1. **Создание окончательной визуализации для истории**



Этот код создает отдельные диаграммы для каждой категории образования родителей и показывает распределение оценок по математике для студентов, чьи родители соответствуют этой категории образования.

Создадим диаграмму рассеяния (scatter plot), которая может помочь визуализировать взаимосвязь между различными переменными в датасете. Например, можно исследовать зависимость между оценками по математике и оценками по чтению, разделяя данные по наличию или отсутствию подготовительных курсов.



Этот график поможет визуально оценить, есть ли различия в оценках по математике и чтению между студентами, которые завершили подготовительные курсы, и теми, кто их не проходил.

**Вывод:** На основе выполненных задач, таких как анализ данных, тестирование гипотез, визуализация результатов, и сводные анализы, можно считать, что лабораторная работа по созданию "истории о данных" выполнена успешно. Был проведен комплексный анализ, используя различные методы статистической обработки и визуализации данных, что позволило выявить значимые закономерности и подтвердить сформулированные гипотезы.