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»

Лабораторная работа 1

по дисциплине «Методы машинного обучения в АСОИУ»

" Создание "истории о данных" (Data Storytelling) "

СТУДЕНТ:
студент группы ИУ5Ц-21М
Москалик А.А.

ПРЕПОДАВАТЕЛЬ:
Гапанюк Ю.Е.

Москва, 2024

Цель лабораторной работы: изучение различных методов визуализация данных и создание истории на основе данных.

Краткое описание. Построение графиков, помогающих понять структуру данных, и их интерпретация.

Задание:

Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.
Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

Ход работы

Подготовка

Подключение библиотек

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Загрузка датасета

[ ]:
```
data = pd.read_csv('/kaggle/input/student-study-performance/study_perform
```

1. Предварительный анализ данных

```
[ ]:
print(data.head())
print(data.describe())
```

```
   gender race_ethnicity parental_level_of_education         lunch  \
0  female        group B           bachelor's degree      standard
1  female        group C                some college      standard
2  female        group B            master's degree      standard
3    male        group A          associate's degree  free/reduced
4    male        group C                some college      standard

  test_preparation_course  math_score  reading_score  writing_score
0                    none          72             72             74
1               completed          69             90             88
2                    none          90             95             93
3                    none          47             57             44
4                    none          76             78             75
       math_score  reading_score  writing_score
count  1000.00000    1000.000000    1000.000000
mean     66.08900      69.169000      68.054000
std      15.16308      14.600192      15.195657
min       0.00000      17.000000      10.000000
25%      57.00000      59.000000      57.750000
50%      66.00000      70.000000      69.000000
75%      77.00000      79.000000      79.000000
max     100.00000     100.000000     100.000000
```

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)

2. 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.

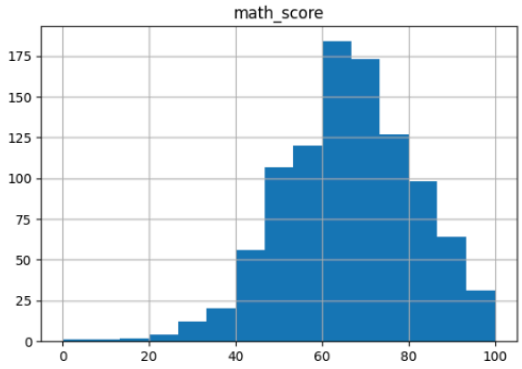
Второй фрагмент кода строит корреляционную матрицу для числовых переменных и визуализирует её с помощью тепловой карты (*heatmap*). Корреляционная матрица позволяет оценить степень линейной связи между парами переменных. Значения корреляции варьируются от -1 до 1, где 1 означает идеальную положительную 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.

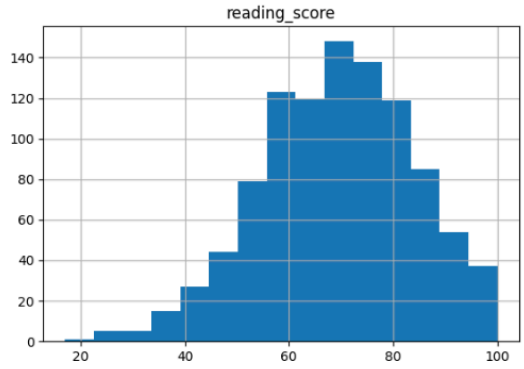
[ ]:

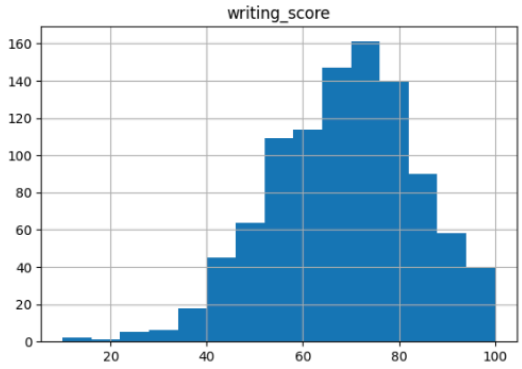
```
import seaborn as sns
import matplotlib.pyplot as plt

# Визуализация только числовых переменных
data_numeric = data.select_dtypes(include=[np.number])
data_numeric.hist(bins=15, figsize=(15, 10))
plt.show()

# Корреляционная матрица только для числовых переменных
plt.figure(figsize=(10, 8))
sns.heatmap(data_numeric.corr(), annot=True, fmt=".2f", cmap='coolwarm')
plt.show()
```

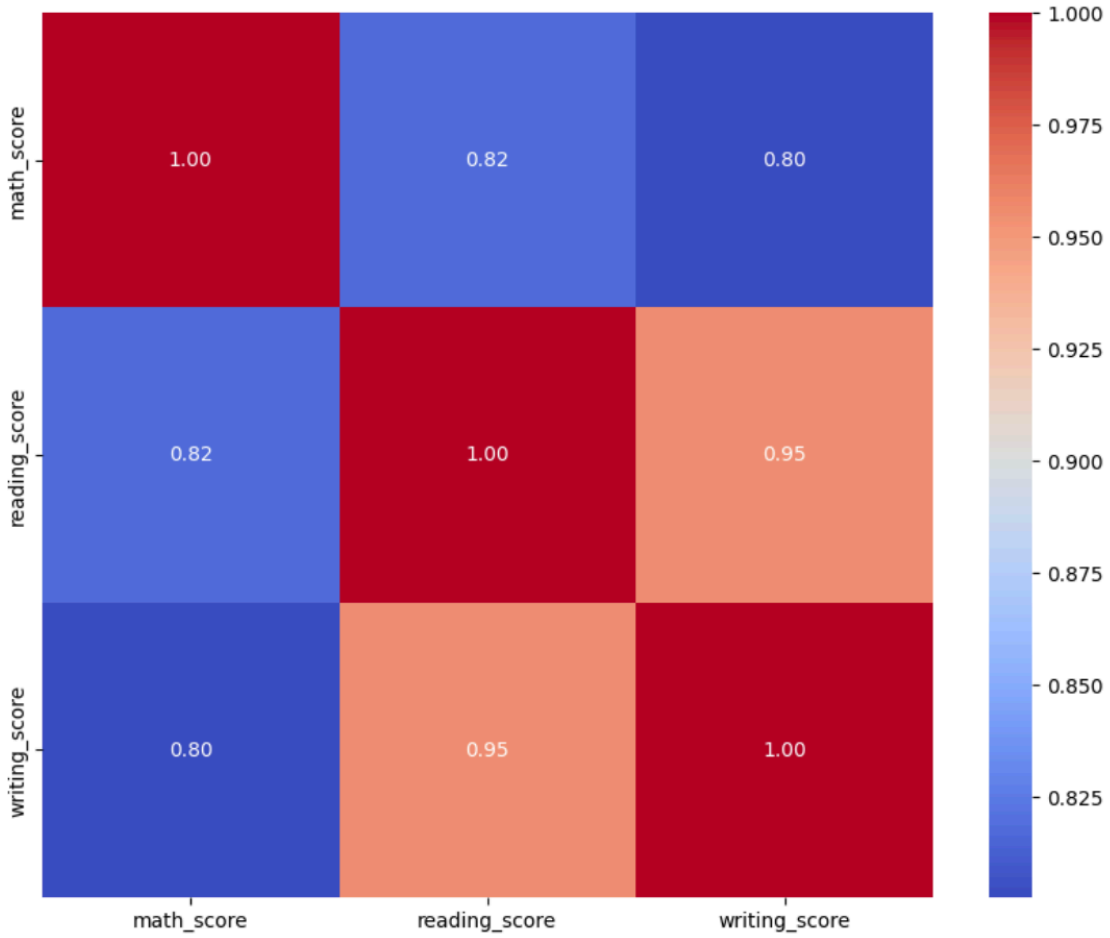






На представленных диаграммах показаны гистограммы, отображающие распределение оценок студентов по трем предметам: математике (math_score), чтению (reading_score) и письму (writing_score).

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

[ ]:

```
# Пропущенные значения отсутствуют на предоставленном снимке экрана,
# но если они есть, можно их заполнить таким образом:
# data['column_name'] = data['column_name'].fillna(data['column_name'].me
# data['column_name'] = data['column_name'].fillna(data['column_name'].mo

# Кодируем категориальные переменные
# One-Hot Encoding для переменных с малым числом категорий
data = pd.get_dummies(data, columns=['gender', 'race_ethnicity', 'parenta

# Масштабирование признаков, для этого можно использовать StandardScaler
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()

# Масштабируем только числовые столбцы
scaled_columns = scaler.fit_transform(data[['math_score', 'reading_score'
data[['math_score', 'reading_score', 'writing_score']] = scaled_columns

# Проверяем результаты
print(data.head())
```

```
   math_score  reading_score  writing_score  gender_female  gender_male  \
0    0.390024       0.193999       0.391492           True        False
1    0.192076       1.427476       1.313269           True        False
2    1.577711       1.770109       1.642475           True        False
3   -1.259543      -0.833899      -1.583744          False         True
4    0.653954       0.605158       0.457333          False         True

   race_ethnicity_group A  race_ethnicity_group B  race_ethnicity_group C  \
0                   False                    True                   False
1                   False                   False                    True
2                   False                    True                   False
3                    True                   False                   False
4                   False                   False                    True

   race_ethnicity_group D  race_ethnicity_group E  \
0                   False                   False
1                   False                   False
2                   False                   False
3                   False                   False
4                   False                   False

   parental_level_of_education_associate's degree  \
0                                           False
1                                           False
2                                           False
3                                            True
4                                           False
```

```
   parental_level_of_education_bachelor's degree  \
0                                           True
1                                          False
2                                          False
3                                          False
4                                          False

   parental_level_of_education_high school  \
0                                    False
1                                    False
2                                    False
3                                    False
4                                    False
```

```
   parental_level_of_education_master's degree  \
0                                        False
1                                        False
2                                         True
3                                        False
4                                        False

   parental_level_of_education_some college  \
0                                     False
1                                      True
2                                     False
3                                     False
4                                      True

   parental_level_of_education_some high school  lunch_free/reduced  \
0                                         False               False
1                                         False               False
2                                         False               False
3                                         False                True
4                                         False               False

   lunch_standard  test_preparation_course_completed  \
0            True                              False
1            True                               True
2            True                              False
3           False                              False
4            True                              False

   test_preparation_course_none
0                          True
1                         False
2                          True
3                          True
4                          True
```

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. Т-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

```
[ ]:
rom scipy.stats import ttest_ind

 Сегментируем оценки по группам
roup_completed = data[data['test_preparation_course_completed'] == True]['math_score']
roup_none = data[data['test_preparation_course_none'] == True]['math_score']

 Применяем t-тест
_stat, p_val = ttest_ind(group_completed, group_none)

 Выводим результаты
rint(f"T-Statistic: {t_stat}, P-Value: {p_val}")

 Интерпретация результатов
lpha = 0.05  # Уровень значимости
f p_val < alpha:
   print("Отклоняем нулевую гипотезу, существует значимая разница.")
lse:
   print("Не отклоняем нулевую гипотезу, значимой разницы нет.")
```

```
T-Statistic: 5.704616417349099, P-Value: 1.535913460714764e-08
Отклоняем нулевую гипотезу, существует значимая разница.
```

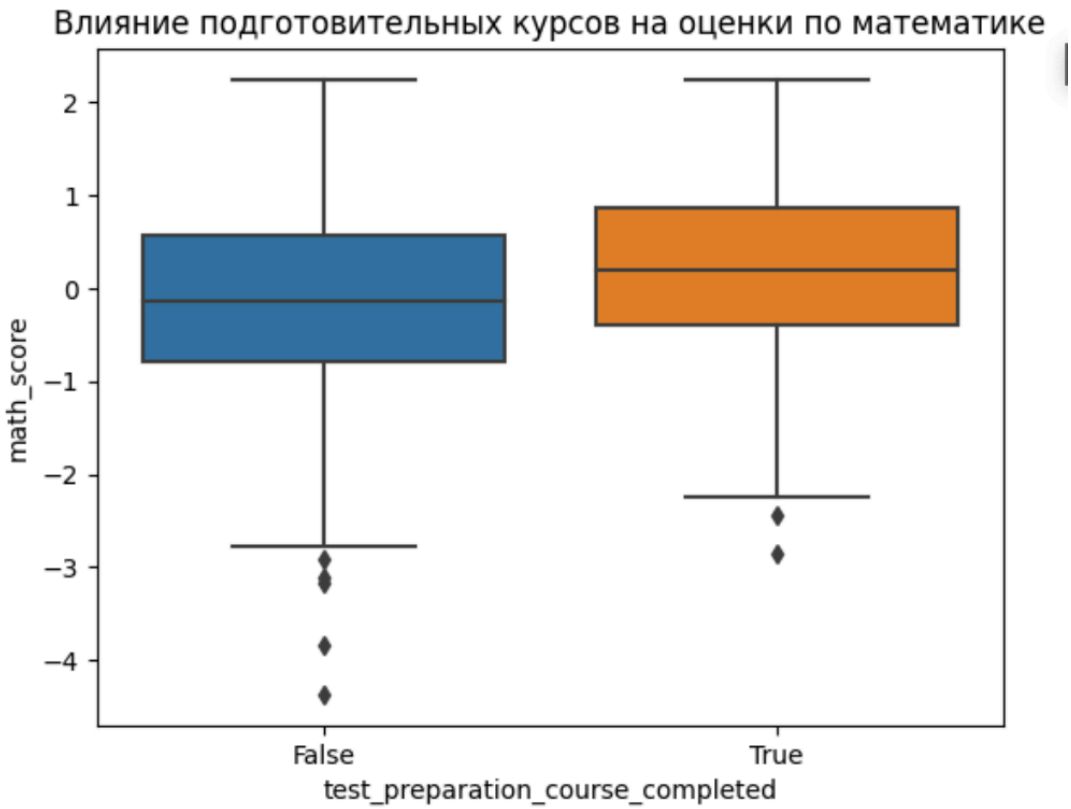
P-значение в результате говорит о вероятности получить данные, как минимум такие же экстремальные, как те, что были получены во время теста, если нулевая гипотеза верна. Очень маленькое P-значение, такое как 1.53e-08 (что эквивалентно 0.00000001539), гораздо меньше стандартного порога $\alpha = 0.05$, 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

```
[ ]:
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(x='test_preparation_course_completed', y='math_score', data=data)
plt.title('Влияние подготовительных курсов на оценки по математике')
plt.show()
```



Анализ распределений каждого признака в разрезе различных групп

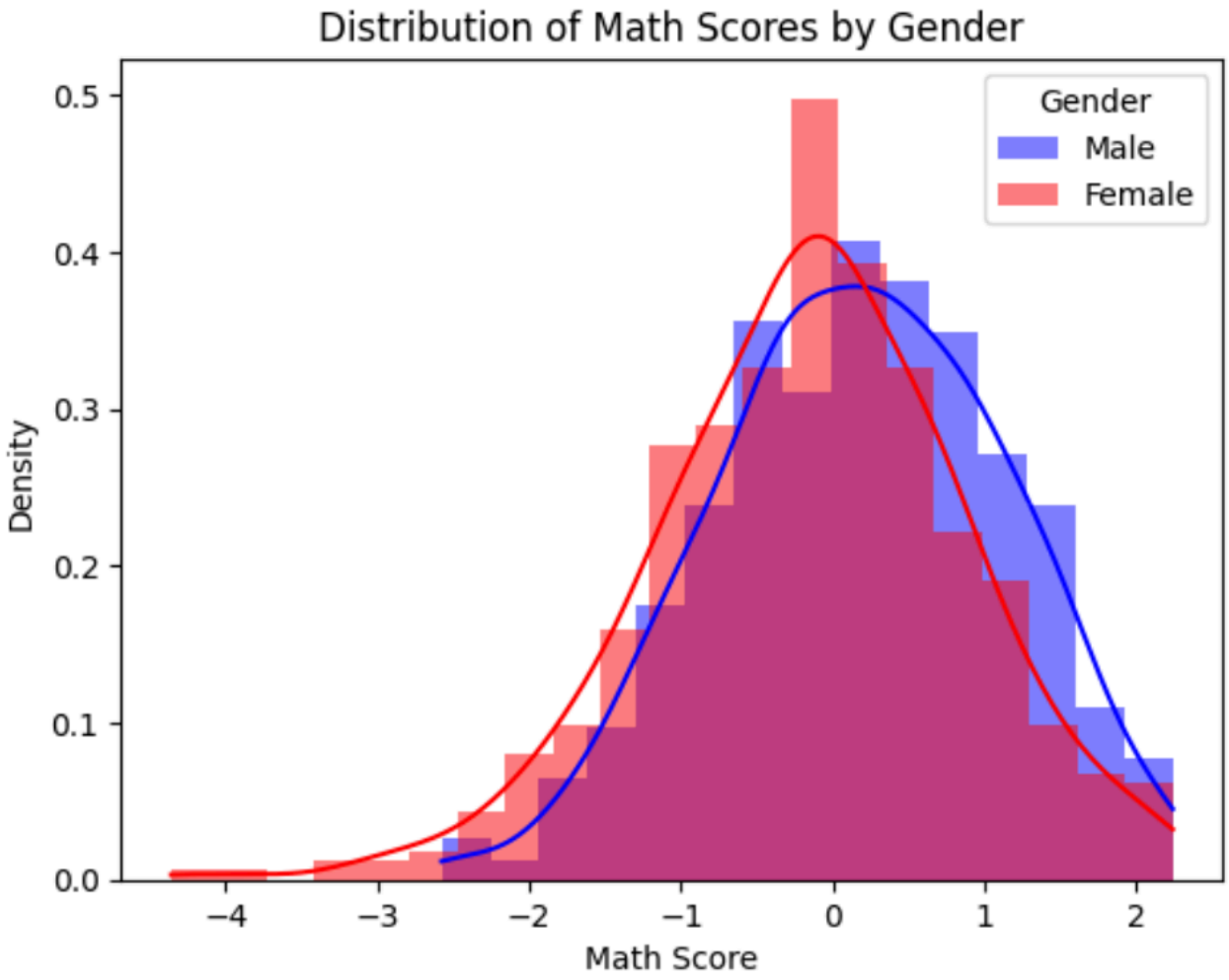
[9]:

```
import warnings
warnings.filterwarnings('ignore', category=FutureWarning)
import seaborn as sns
import matplotlib.pyplot as plt

# Гистограмма для мужчин
sns.histplot(data[data['gender_male'] == True]['math_score'], color="blue", label='Male

# Гистограмма для женщин
sns.histplot(data[data['gender_female'] == True]['math_score'], color="red", label='Fema

plt.legend(title='Gender')
plt.title('Distribution of Math Scores by Gender')
plt.xlabel('Math Score')
plt.ylabel('Density')
plt.show()
```



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

```
[ ]: import pandas as pd
     mean_scores = pd.pivot_table(data, values=['math_score', 'reading_score', 'writing_scor
     print(mean_scores)
```

```
             math_score  reading_score  writing_score
gender_male
False         -0.162040       0.235670       0.290569
True           0.174142      -0.253272      -0.312271
```

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.

5. Создание окончательной визуализации для истории

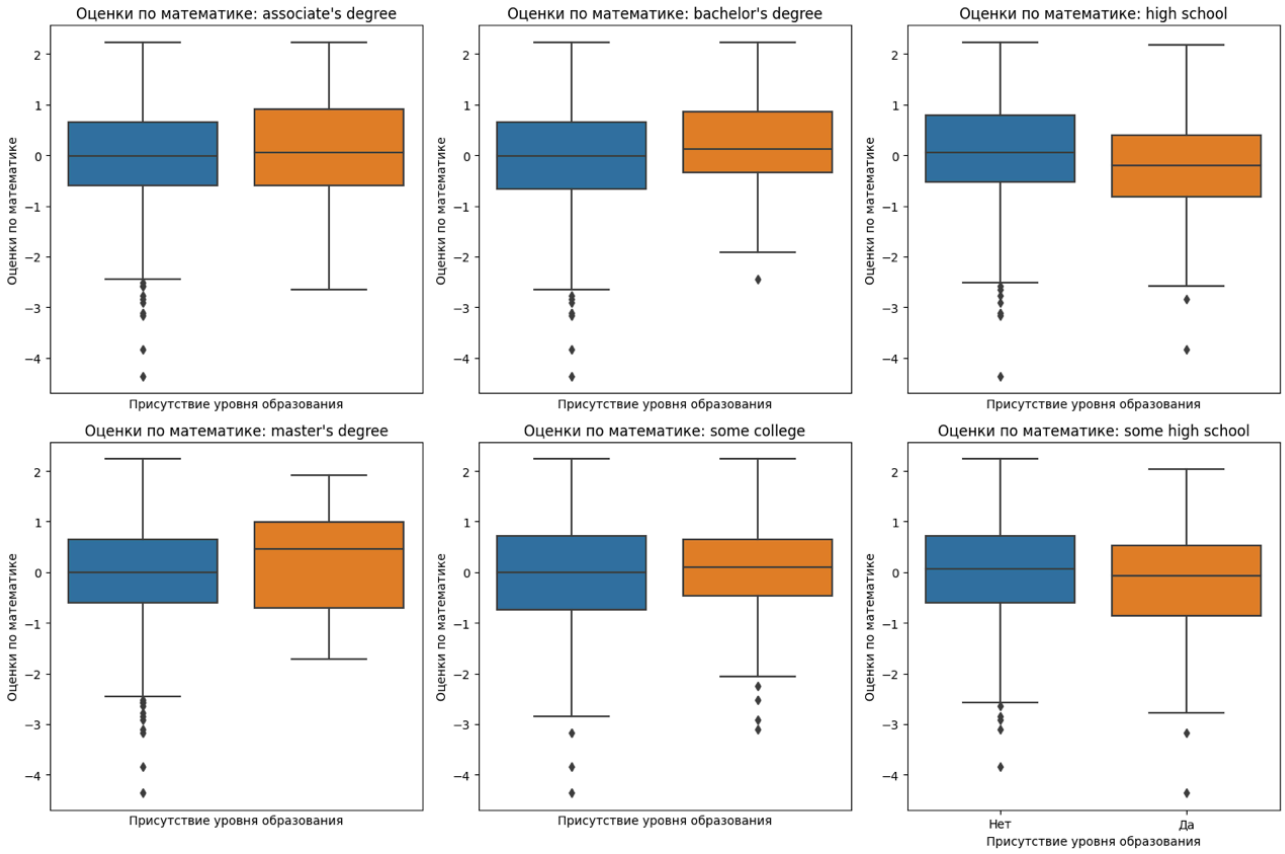
[ ]:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Предполагая, что вы преобразовали 'parental_level_of_education' в столбцы формата 'pa
education_levels = [col for col in data.columns if col.startswith('parental_level_of_ed

plt.figure(figsize=(15, 10))
for i, level in enumerate(education_levels, 1):
    plt.subplot(2, 3, i)  # Предполагая, что у нас не больше 6 категорий образования
    sns.boxplot(x=data[level], y=data['math_score'])
    plt.title(f'Оценки по математике: {level.split("_")[-1]}')
    plt.xlabel('Присутствие уровня образования')
    plt.ylabel('Оценки по математике')
    if i == len(education_levels):
        plt.xticks([0, 1], ['Нет', 'Да'])
    else:
        plt.xticks([])  # Скрываем подписи x-оси для наглядности

plt.tight_layout()
plt.show()
```



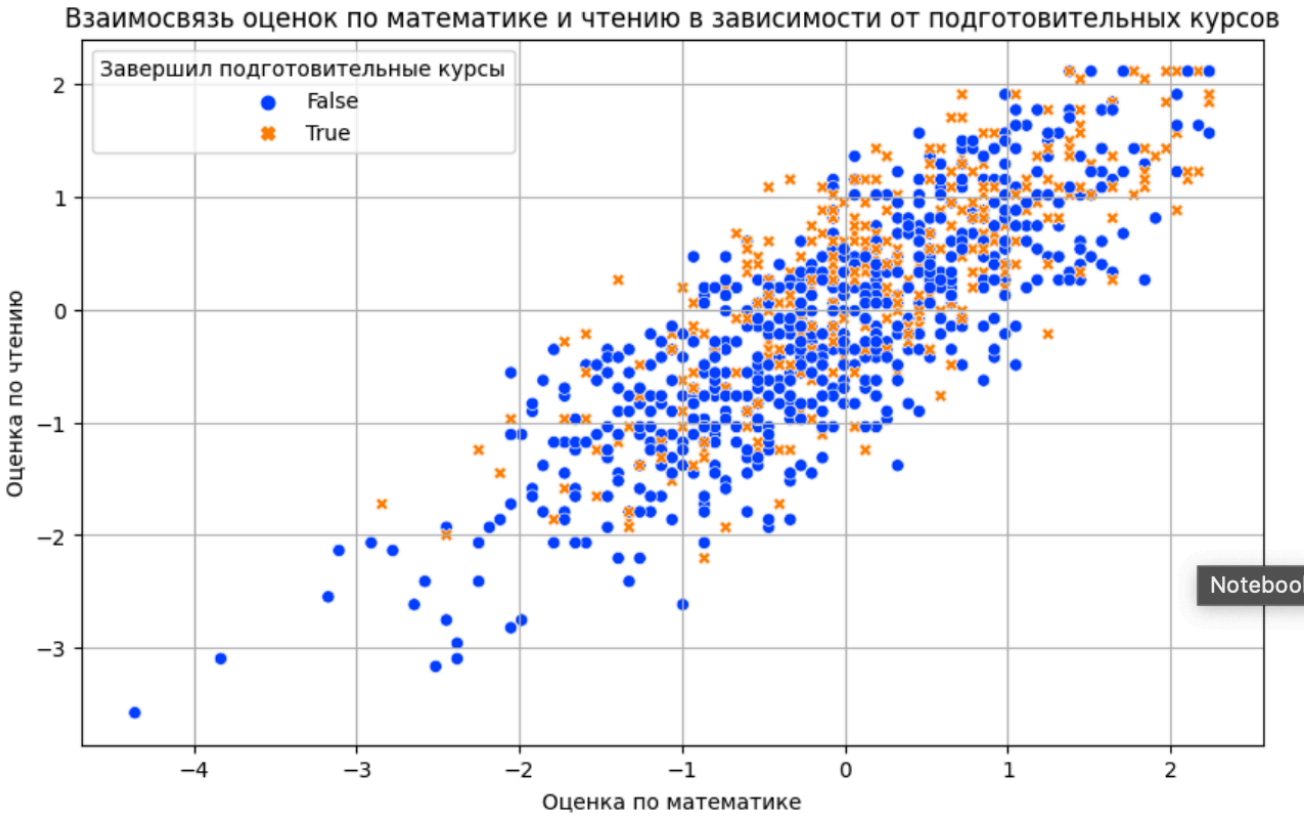
Этот код создает отдельные диаграммы для каждой категории образования родителей и показывает распределение оценок по математике для студентов, чьи родители соответствуют этой категории образования.

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.

[ ]:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Диаграмма рассеяния, показывающая взаимосвязь между оценками по математике и чтению
# разделяем по наличию подготовительных курсов
plt.figure(figsize=(10, 6))
sns.scatterplot(x='math_score', y='reading_score', hue='test_preparation_course_complet
plt.title('Взаимосвязь оценок по математике и чтению в зависимости от подготовительных
plt.xlabel('Оценка по математике')
plt.ylabel('Оценка по чтению')
plt.legend(title='Завершил подготовительные курсы')
plt.grid(True)
plt.show()
```



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.