Университет ИТМО

Практическая работа №4

по дисциплине «Визуализация и моделирование»

Автор: Костылев Иван Михайлович
Поток: 1.1
Группа: K3240
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Описание датасета

Датасет состоит из данных о студентах, их родителей и оценок, полученных ими по различным предметам.

Всего записей: 1000

Формальное описание

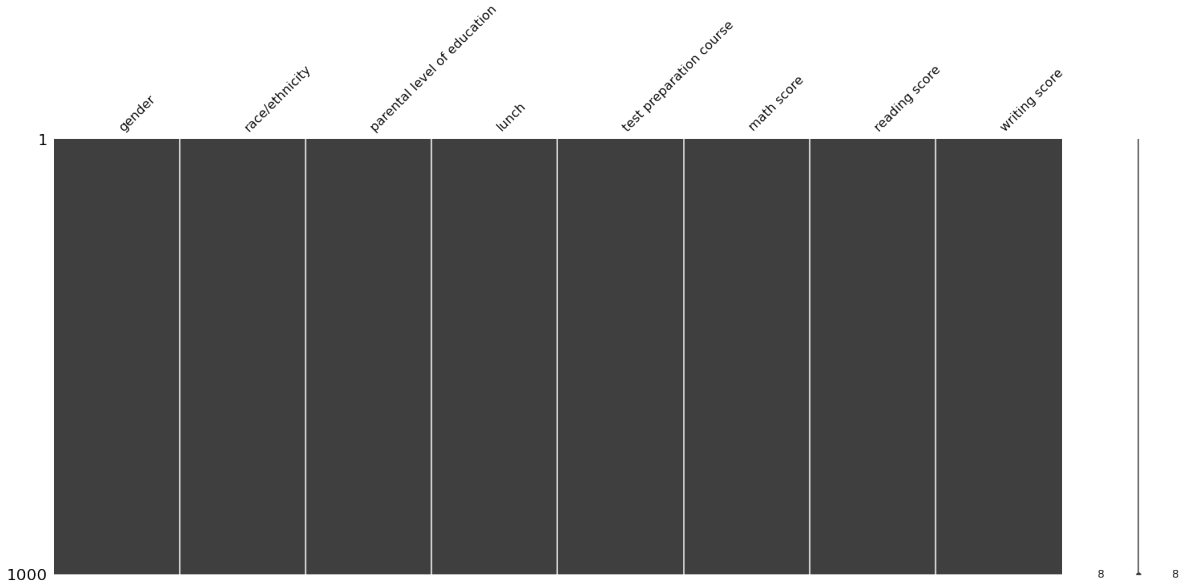
Столбец	Описание	Значения	Формат	Шкала
gender	пол студента	male / female	текст	Качеств номинальная
race/ethnicity	расовая классификация	group A / group B / group C / group D	текст	Качеств номинальная
parental level of education	уровень образования родителей	collegue / school / bachelor's degree / others	текст	Качеств номинальная
lunch	оплата обеда	standart / free/reduced	текст	Качеств номинальная
test preparation	подготовка к тесту	none / completed	текст	Качеств номинальная
math score	оценка по математике	0..100	целое число	Колич относительная
reading score	оценка по чтению	0..100	целое число	Колич относительная
writting score	оценка по письму	0..100	целое число	Колич относительная

Предобработка данных

Подобранный датасет не нуждался в предобработке на устранение пустых ячеек. Категориальные данные будем приводить из строк к числовому типу по мере необходимости.

Описательная статистика 2.0

Здесь можно лишь показать, что данные являются цельными, без пустых ячеек.

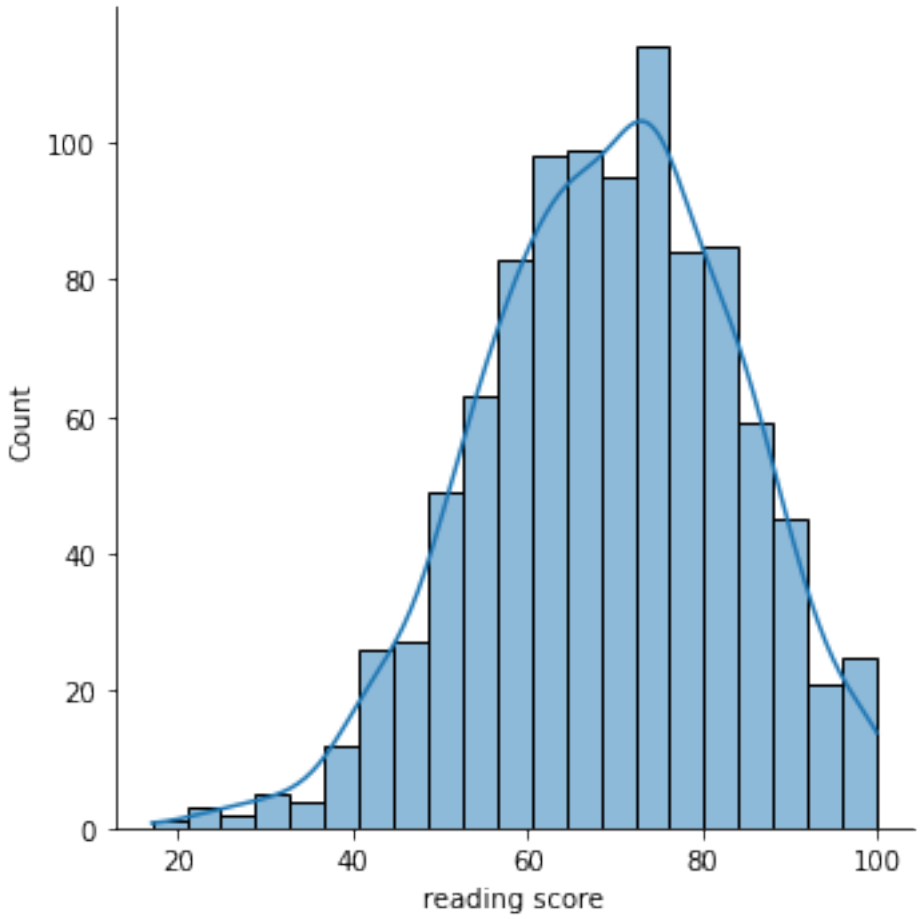


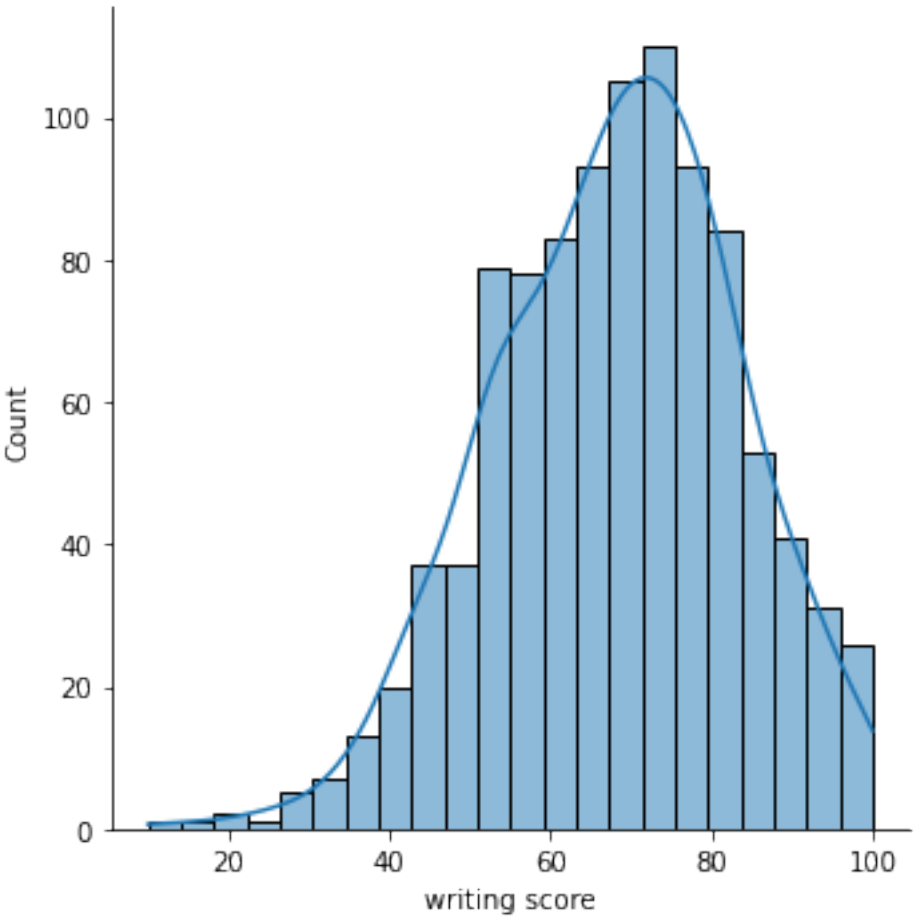
С предыдущего этапа не произошло никаких качественных изменений в данных, которые могли бы позволить получить какое-то новое понимание о данных, поэтому проводить описательную статистику снова не имеет смысла.

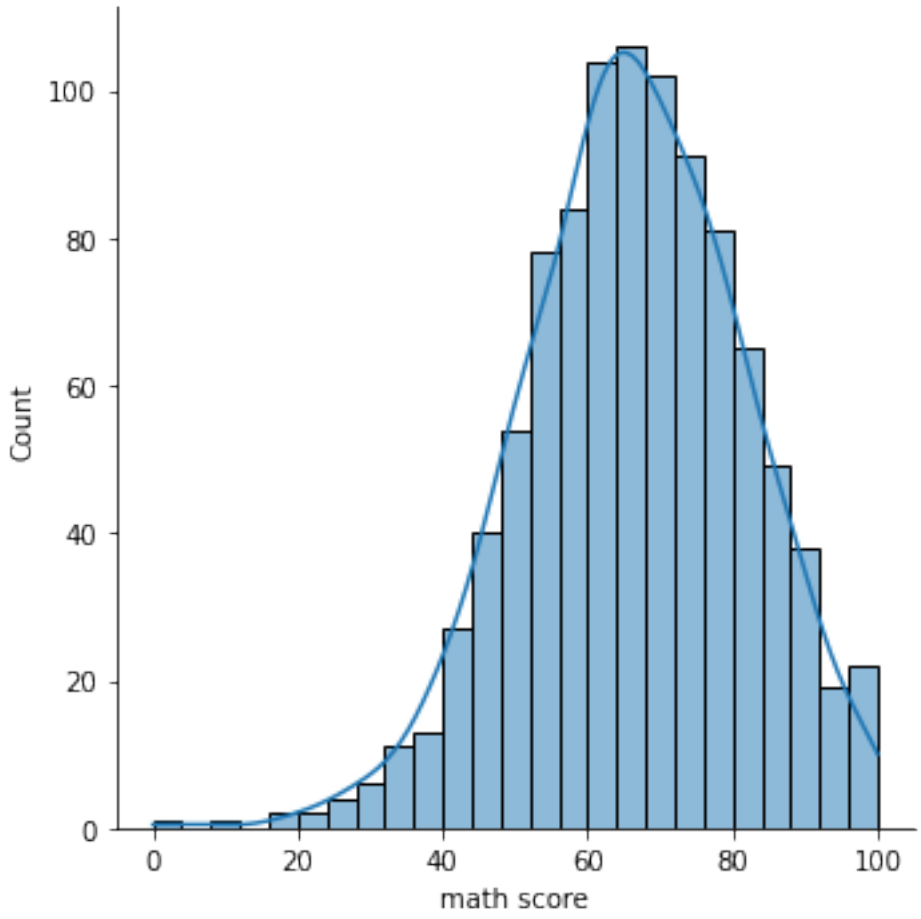
Новые гипотезы

1. Гипотеза: распределение оценок по предметам соответствует нормальному распределению

```
sns.displot(df, x=READING, kde=True)
sns.displot(df, x=WRITING, kde=True)
sns.displot(df, x=MATH, kde=True)
```


Count
0
20
40
60
80
100
20
40
60
80
100
reading score


Count
0
20
40
60
80
100
20
40
60
80
100
writing score

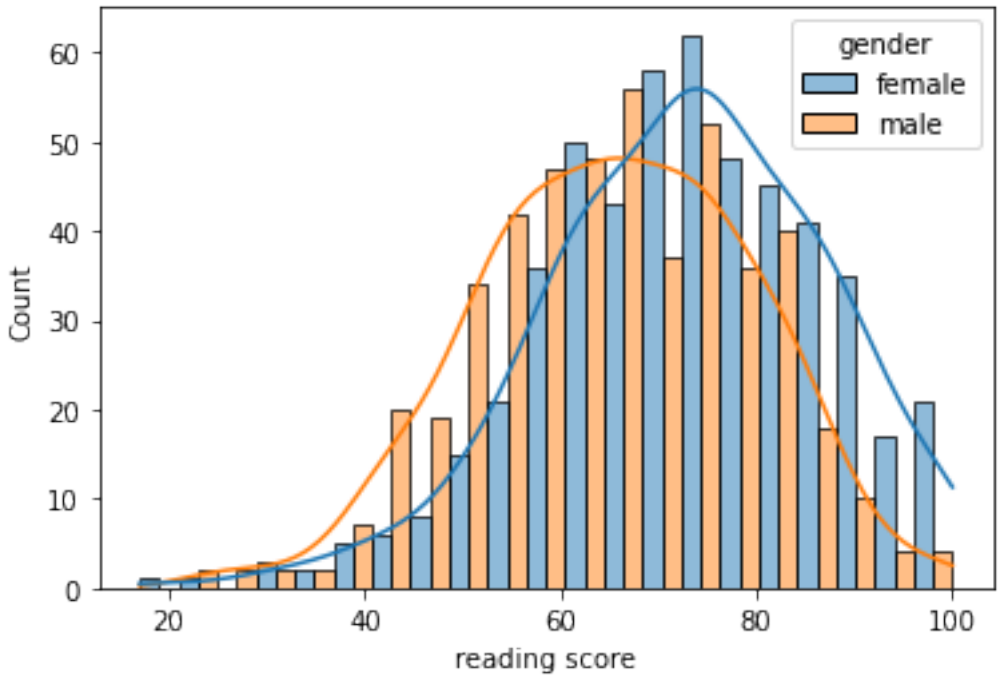


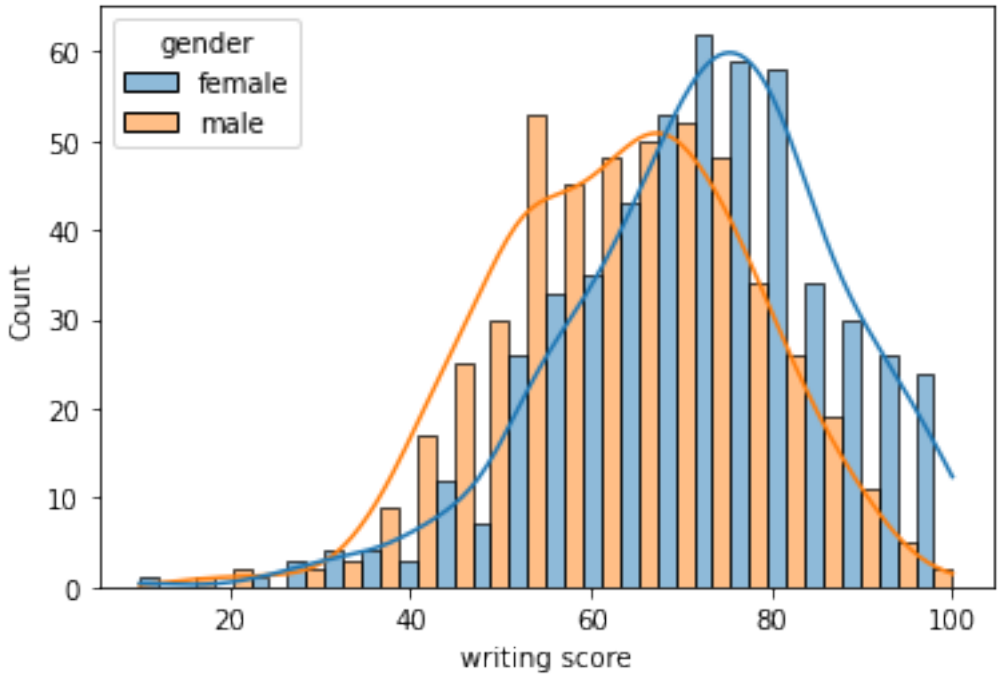
Вывод: распределение оценок действительно является похожим на нормальное. Теперь хотелось бы посмотреть более детально на данный график, как хорошо справлялись мужчины и женщины.

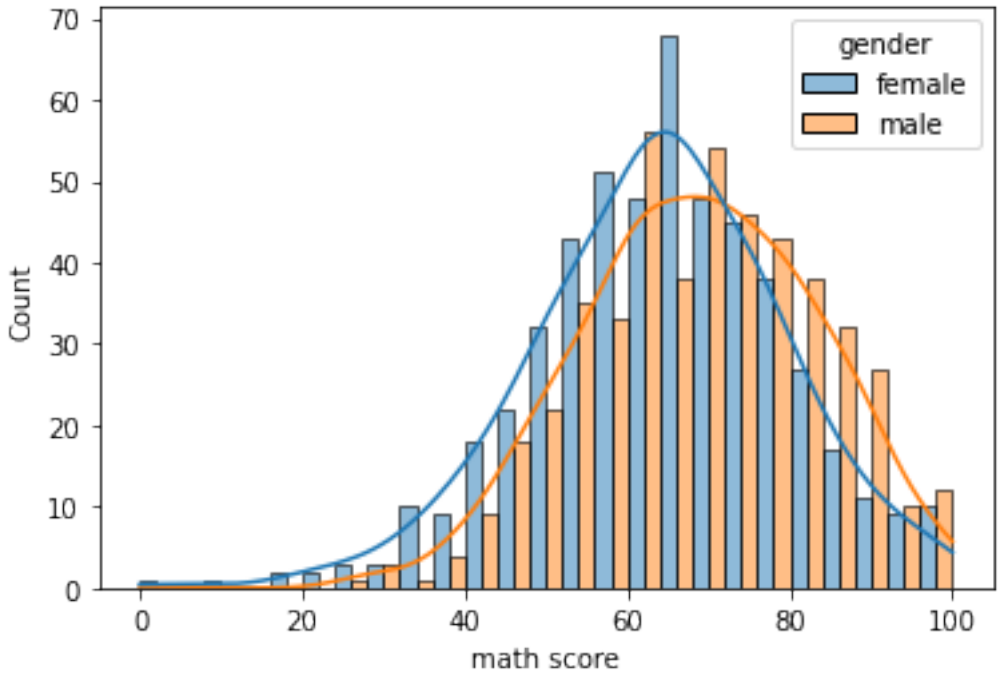
Из предыдущих исследований было видно, что женщины, в среднем, лучше справляются с работами (суммарный балл выше в среднем, чем у мужчин).

2. Вопрос: каково распределение баллов по предметам? Везде ли лидируют женщины?

Построим аналогичный график, который будет отдельно показывать количество текущего балла у мужчин и женщин.


gender
female
male
Count
reading score
0
10
20
30
40
50
60
20
40
60
80
100


gender
female
male
Count
writing score
0
10
20
30
40
50
60
20
40
60
80
100



Здесь оранжевый график показывает распределение баллов у мужчин. Видно, что максимум правее у мужчин только в математике, т.е. можно утверждать, что мужчины справляются лучше только с математикой. Причем в этом предмете отрыв не такой большой, как в письме и чтении у женщин от мужчин.

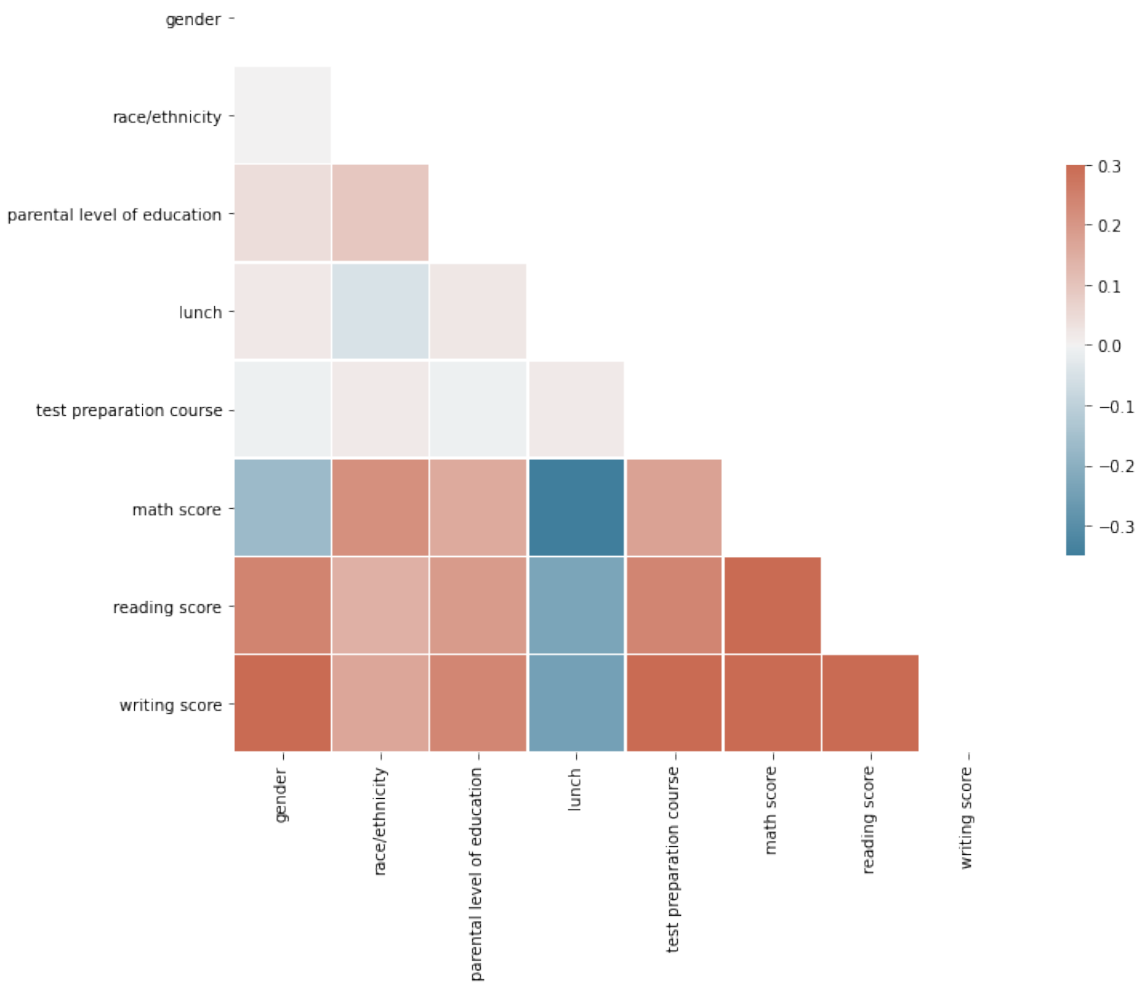
Давайте построим большую матрицу корреляции, к которой сделаем несколько гипотез и вопросов.

3. Удостоверимся, что оценки по чтению и письму коррелируют с полом студента (поскольку там явно выражено преобладание студентов-женщин по оценкам)

4. Гипотеза: Данные по уровню образования родителей коррелируются с наличием подготовки студентов к тесту. Студенты, родители которых имеют высшее образование, проходят подготовительные курсы, поскольку являются более ответственными.

5. Гипотеза: Уровень образования родителей коррелирует с

расовой принадлежностью



3. Оценки по чтению больше всего коррелируют с гендером по письму и чтению - гипотеза подтвердилась. По математике корреляция меньше, поскольку мужчины чуть лучше справлялись с предметами, тогда как по письму и чтению виделось явное преобладание баллов у женщин.

4. Удивительно, но уровень образования родителей не делает студентов "более ответственными"и не гарантирует, что они будут проходить подготовительные курсы. В то же время уровень образования родителей коррелирует с оценками по предметам достаточно хорошо.

5. На данной матрице видно, что уровень образования родителей в общем и целом каким-то образом соответствует расовой принадлежности.

Общие выводы по работе: на основе проведенного исследования мы получили новую информацию о взаимосвязи между данными, что позволит строить более конкретные гипотезы в будущем и ставить какие-либо задачи.

К вопросам и задачам на будущее я бы отнес:
1) Как именно связана расовая принадлежность и уровень образования (какая этническая группа имеет более высокий уровень образования?)

2) Как именно связаны оценки по предметам с расой / уровнем образования родителей?

3) Каким образом можно смотреть на поле lunch? Видно, что есть небольшая корреляция с подготовительным курсом и с уровнем образования родителей.