Университет ИТМО

Практическая работа №2

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

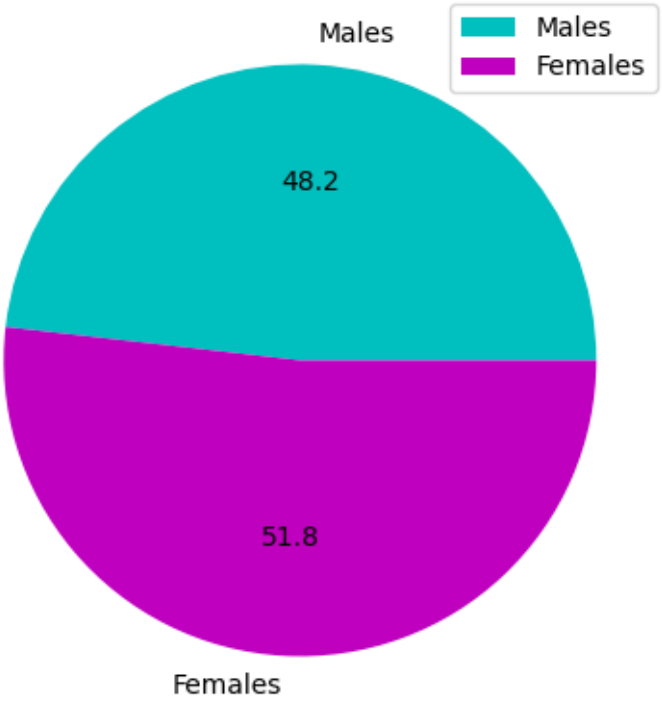
Описательная статистика

1. Соотношение студентов мужчин и женщин

Фрагмент кода, запроса:

```
male_df = df.loc[df[GENDER] == "male"]
m_num = male_df.shape[0]
female_df = df.loc[df[GENDER] == "female"]
f_num = female_df.shape[0]

gender_pie = pd.DataFrame({"": [m_num, f_num]},
                          index=["Males", "Females"])
gender_pie.plot.pie(y="",
                    colors=["c", "m"],
                    autopct="%.1f",
                    fontsize=10,
                    figsize=(5, 5))
```

График:



Вывод: можно считать равным количество студентов мужчин и женщин

Для решения следующих вопросов объявим для простоты константные переменные - названия полей

```
READING = "reading score"
WRITING = "writing score"
MATH = "math score"
GENDER = "gender"
PLE = "parental level of education"
ETGROUP = "race/ethnicity"
PREPAR = "test preparation course"
```

Следующие вопросы связаны между собой:

2. Статистика по набранным баллам

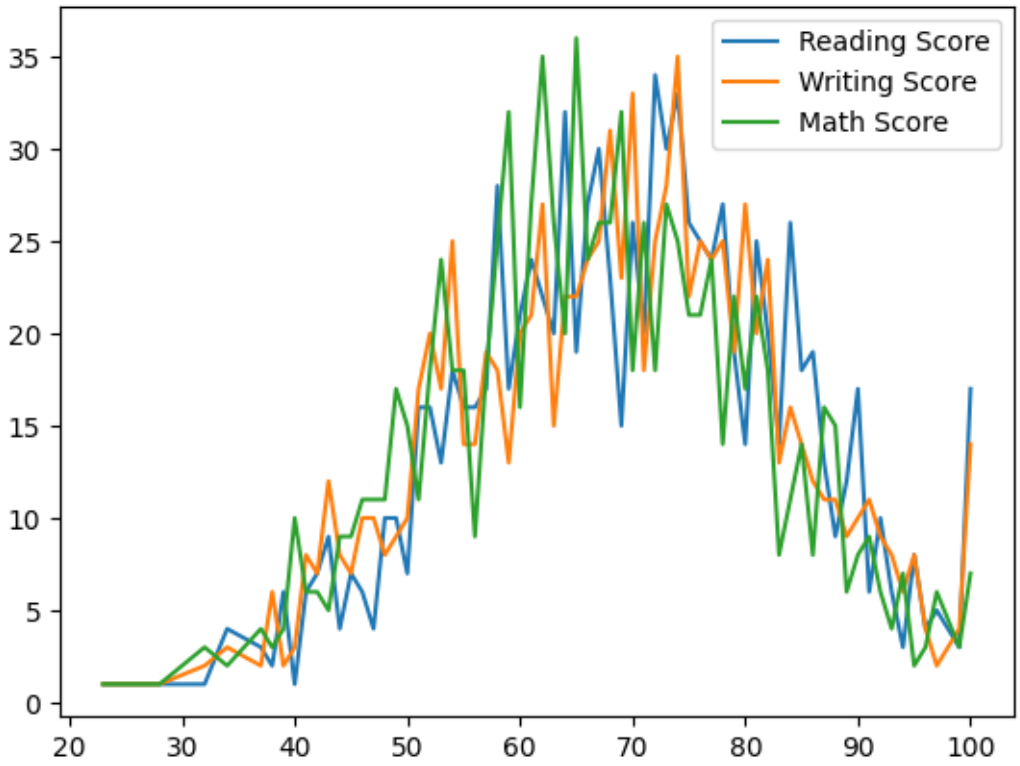
3. Соотношение средних баллов к баллам по чтению

4. Соотношение средних баллов к баллам по письму

5. Соотношение средних баллов к баллам по математике

Фрагмент кода, запроса: *Следующий код оказался слишком объемным, его можно посмотреть в исходном коде*

График (2): на рисунке по оси абсцисс - баллы от 0 до 100, а по оси ординат - количество человек, которые получили данный балл



Вывод (2): распределение по баллам примерно соответствует нормальному распределению. На данном графике не видно преобладание какого-либо предмета над другим. Для определения наиболее "успешного"(где средний балл выше) построить следующие графики:

График (3): средние оценки и оценки по чтению. На рисунке также по оси абсцисс - баллы от 0 до 100, а по оси ординат - количество человек, которые получили данный балл

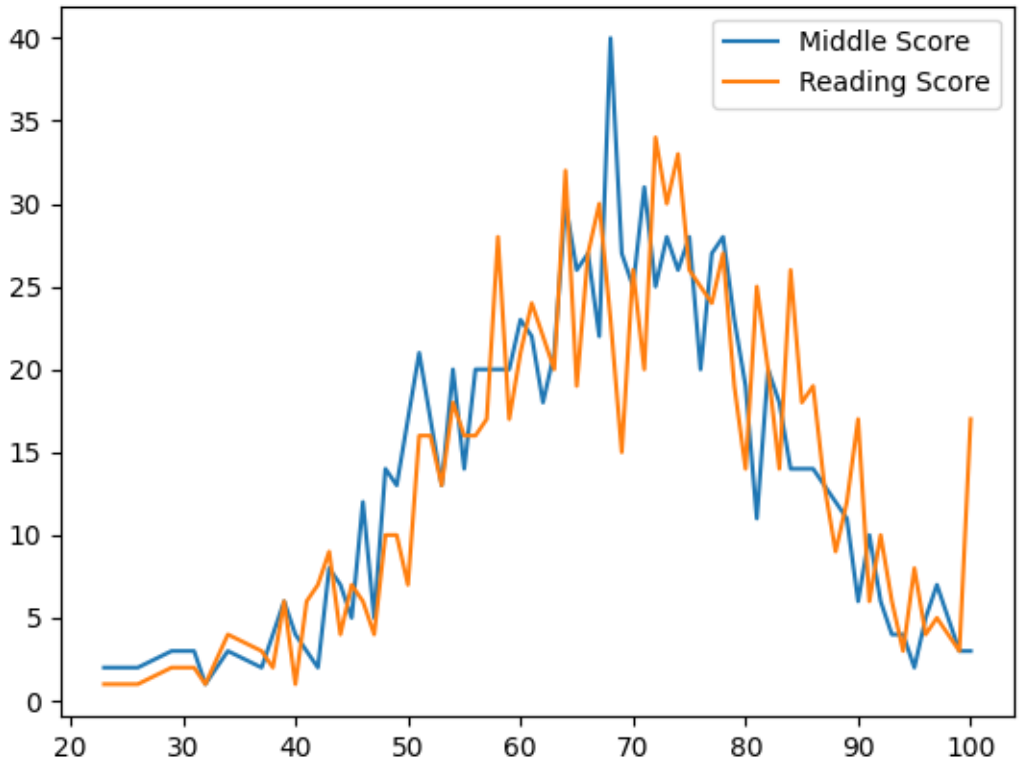


График (4): средние оценки и оценки по письму

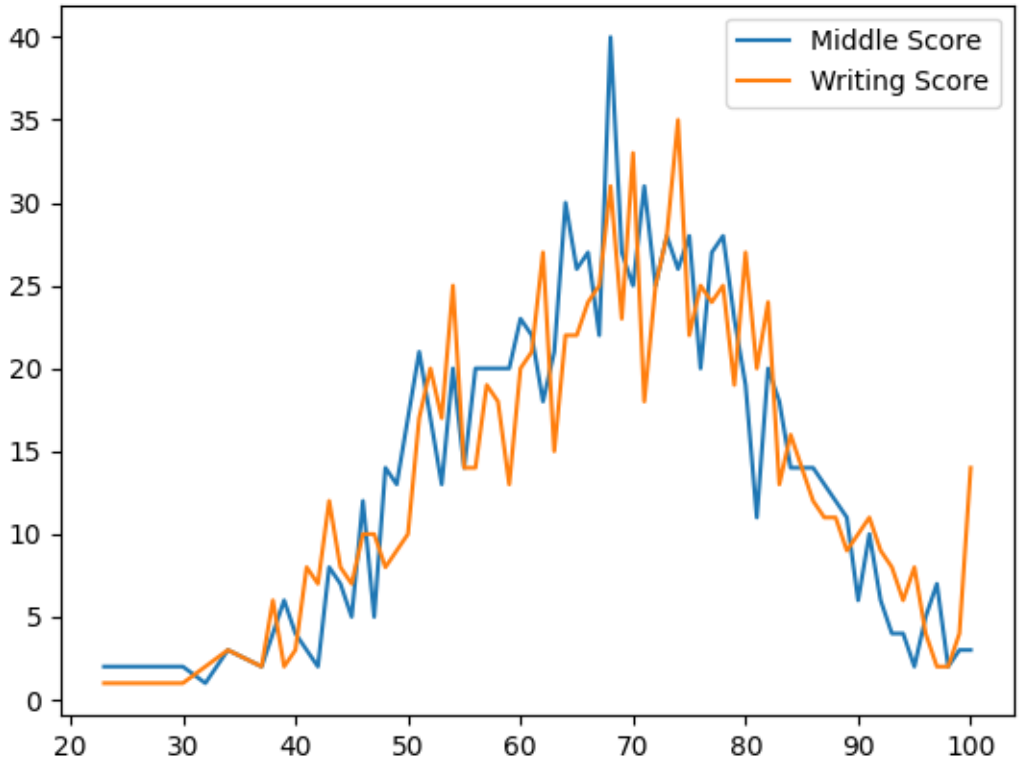
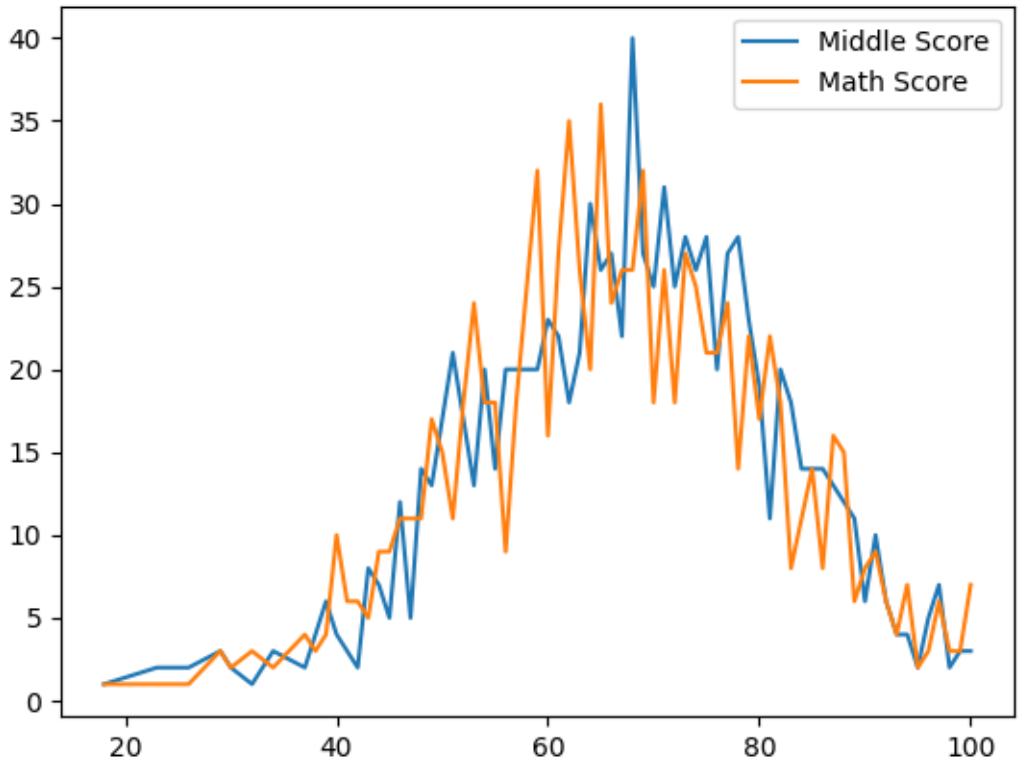


График (5): средние оценки и оценки по математике



Вывод (3-5): из графика (5) видно, что баллы по математике ниже средних по всем предметам, так как график при баллах > 70 лежит ниже графика средних баллов. *Можно построить гипотезу, что по математике средний балл ниже.* В то же самое время из графика (3) видно, что здесь максимальное количество тех, кто получил 100 баллов, а также максимальная точка смещена вправо, что позволяет *построить гипотезу, что по чтению средний балл самый высокий.*

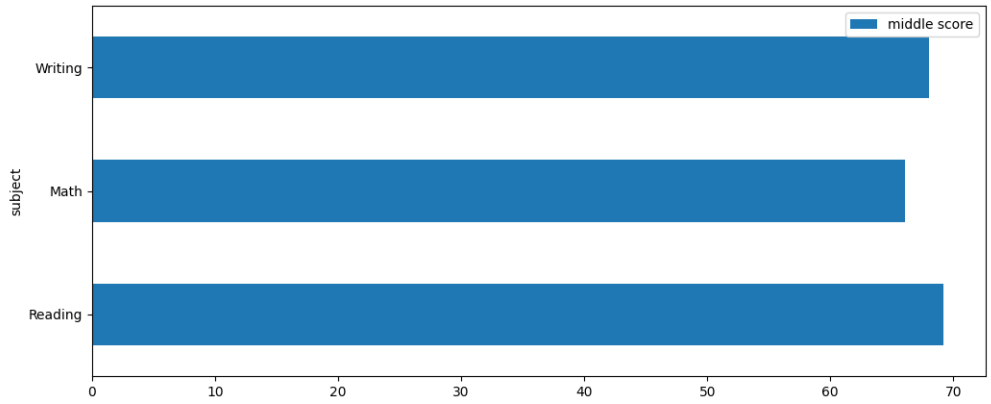
Проверим далее наши предположения

6. По какому предмету баллы, в среднем, выше?

Фрагмент кода, запроса:

```
middle_scores = [middle_reading, middle_math, middle_writing]
subjects = ["Reading", "Math", "Writing"]

middles_df = pd.DataFrame({"subject": subjects, "middle score": middle_sco
middles_df.plot.barh(x="subject", y="middle score", figsize=(12, 5))
```

График (6):



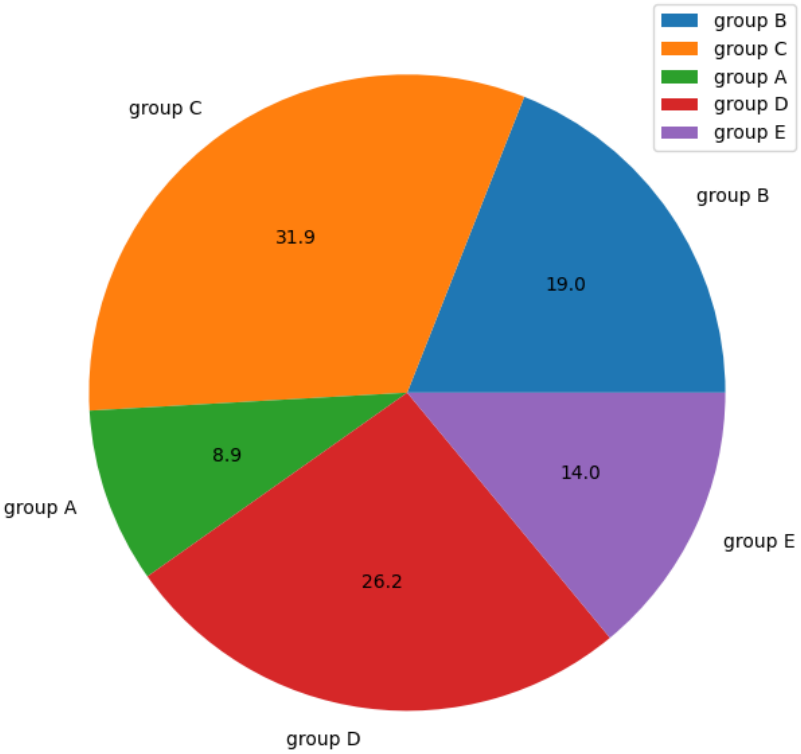
Вывод: гипотезы о том, что оценки по чтению в среднем выше, чем по остальным. Самые низкие получились по математике.

7. Каково соотношение между этническими группами студентов?

Фрагмент кода, запроса:

```
col = ETGROUP
groups_data = {name: df[col].to_list().count(name) for name in df[col].uni
groups_list = []
nums = []

for group, num in groups_data.items():
    groups_list.append(group)
    nums.append(num)

group_df = pd.DataFrame({"": nums}, index=groups_list)

group_df.plot.pie(y="",
                  autopct="%.1f",
                  fontsize=10,
                  figsize=(12, 12))
```

График (7):



Вывод: нельзя выделить какое-либо серьезное преобладание какой-либо группы (более 50 процентов). Самая распространенная группа, которая включена в статистику - группа D, самая менее распространенная - группа A.

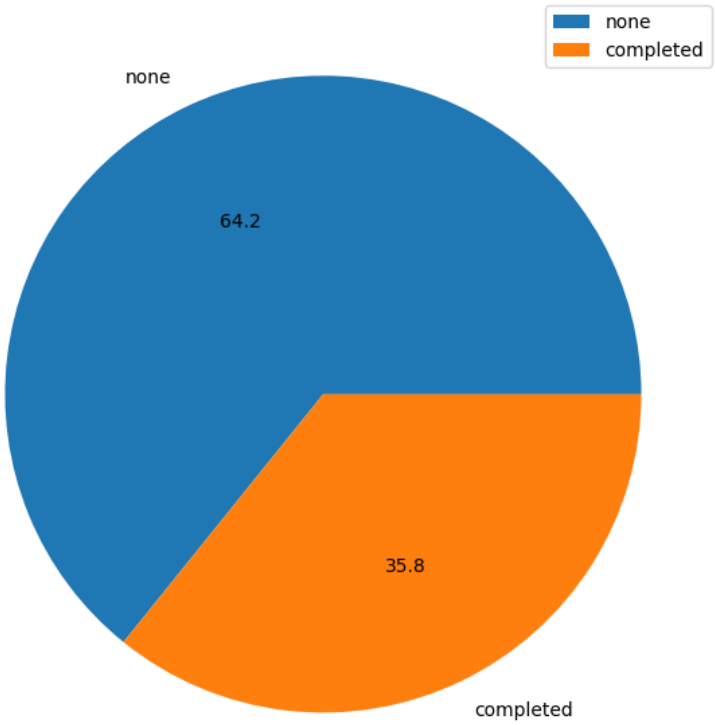
8. Как много студентов прошло подготовительный курс? 9. Существенна ли помощь подготовки?

Фрагмент кода, запроса:

```
col = PREPAR
preparation_data = {name: df[col].to_list().count(name) for name in df[col

preparations = []
nums = []

for prep, num in preparation_data.items():
    preparations.append(prep)
```

```
    nums.append(num)


prep_df = pd.DataFrame({"": nums}, index=preparations)

prep_df.plot.pie(y="",
                 autopct="%.1f",
                 fontsize=10,
                 figsize=(12, 12))
```

График (8):

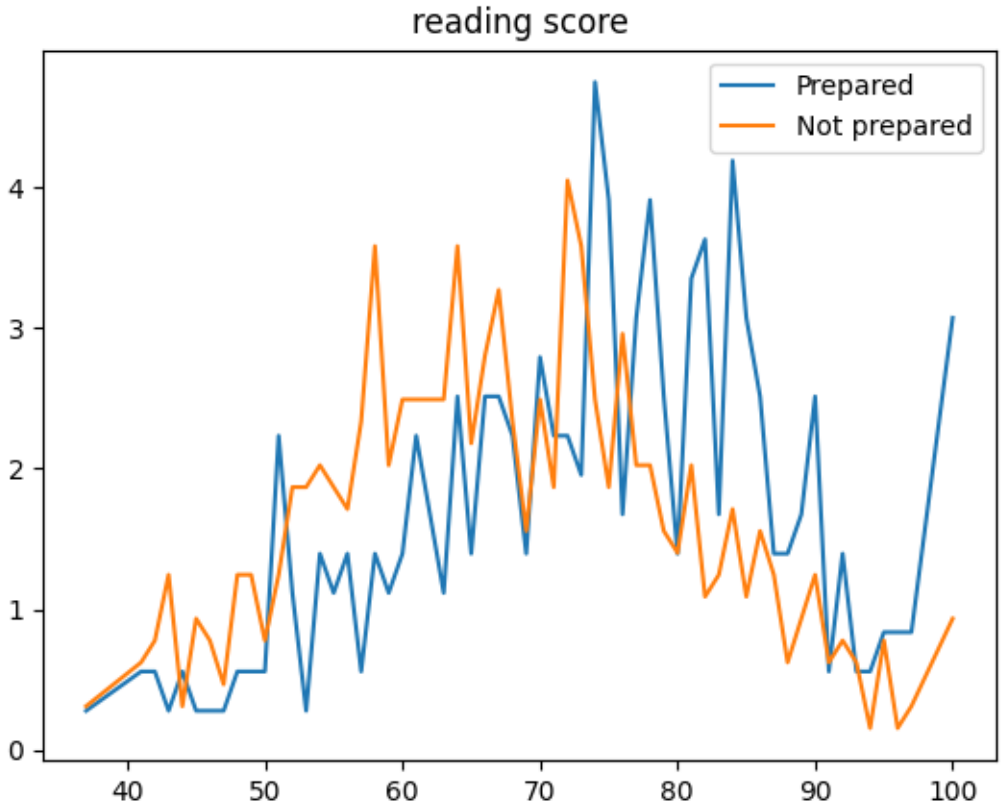


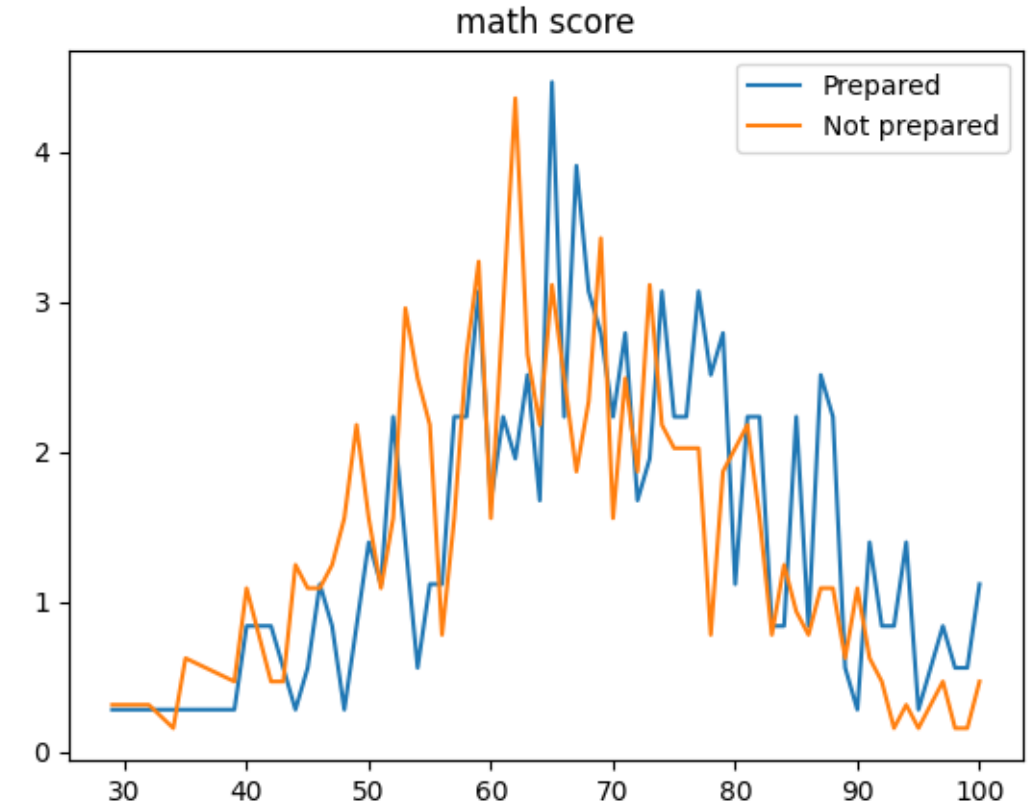
Вывод: преобладают студенты, которые не проходили подготовительный курс. Посмотрим, выше ли баллы у студентов, которые его прошли

Следующий код оказался слишком объемным, его можно посмотреть

в исходном коде

Получаем следующие графики (9): здесь значения по оси y - доля людей от общего числа подготовившихся / не подготовившихся





Вывод (9): Ярче всего видно на графике чтения, что доля студентов,

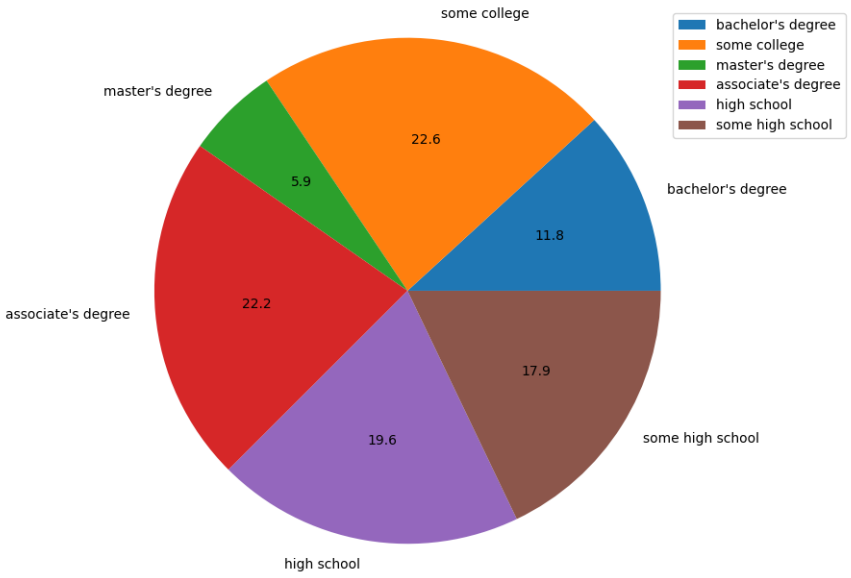
которые проходили курс, имеют более высокие баллы.

10. Соотношение студентов мужчин и женщин

Фрагмент кода, запроса:

```
or edu, num in par_edu_data.items():
        labels.append(edu)
        nums.append(num)

    par_edu_df = pd.DataFrame({"": nums}, index=labels)

    par_edu_df.plot.pie(y="",
                autopct="%.1f",
                fontsize=10,
                figsize=(12, 12))
```

График:



Вывод: можно считать равным количество всех уровней образования, кроме степени магистра (master's degree). В общем и целом идёт преобладание среднего образования (колледжей, старшая школа)