

Университет ИТМО

Практическая работа №3
по дисциплине «Визуализация и моделирование»

Автор: Костылев Иван Михайлович

Поток: 1.1

Группа: К3240

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	Качественная
race/ethnicity	расовая классификация	group A / group B / group C / group D	текст	Качественная
parental level of education	уровень образования родителей	collegue / school / bachelor's degree / others	текст	Качественная
lunch	оплата обеда	standart / free/reduced	текст	Качественная
test preparation	подготовка к тесту	none / completed	текст	Качественная
math score	оценка по математике	0..100	целое число	Количественная
reading score	оценка по чтению	0..100	целое число	Количественная
writting score	оценка по письму	0..100	целое число	Количественная

Описание проблем

Название	Описание	Формат	Шкала	Проблема	Решение
gender	пол студента	str	номинальная	текст неудобно использовать при построении модели	перевод в число
race/ethnicity	этническая принадлежность	str	номинальная	текст неудобно использовать при построении модели	перевод в число
parental level of education	уровень образования родителей	str	номинальная	текст неудобно использовать при построении модели	перевод в число
lunch	оплата обеда	str	номинальная	текст неудобно использовать при построении модели	перевод в число
test preparation course	курс по подготовке к тесту	str	номинальная	текст неудобно использовать при построении модели	перевод в число
math score	балл по математике	int	относительная	-	-
reading score	балл по чтению	int	относительная	-	-
writing score	балл по письму	int	относительная	-	-

Подготовка данных

Полный код лежит в блокноте:

<https://colab.research.google.com/drive/1V56fg3-hLgE9BnUmFQGEiFc8l21EHF0e?usp=s>

1. Обработка пустых ячеек.

В датасете отсутствуют пустые ячейки. Убеждаемся в этом:

```
df.isnull().sum()
```

Output:

```
gender                0
race/ethnicity        0
parental level of education  0
lunch                 0
test preparation course  0
math score            0
reading score         0
writing score         0
dtype: int64
```

2. Для построения моделей данных в следующих работах нам необходимо перевести текст в числовые значения.

1) Столбец 'gender' (male / female):

Категориальный признак. Нормализуем его с помощью следующей функции:

```
def norm_gender(gender: str) -> int:
    if gender == 'male':
        return 0
    else:
        return 1
```

Значению 'male' соответствует 0, 'female' - 1.

```
PREP = 'test preparation course'
df_norm = df.copy()
df_norm[PREP] = df_norm[PREP].apply(norm_preparation)
```

2) Столбец 'test preparation course' (none / completed):

Категориальный признак. Нормализуем его с помощью следующей функции:

```
def norm_preparation(is_prepared: str) -> int:
    if is_prepared == 'none':
        return 0
    else:
        return 1
```

Таким образом, значению 'none' (не проходил курс) будет соответствовать 0, 'completed' - 1.

```
PREP = 'test preparation course'
df_norm = df.copy()
df_norm[PREP] = df_norm[PREP].apply(norm_preparation)
```

Нормализация остальных данных выполнялась аналогичным способом, поскольку они также являются категориальными (кроме столбцов с оценками).

Гипотезы

Опираясь только лишь на описательную статистику из лабораторной работы №2, на данном этапе сложно построить новые гипотезы.

Хотелось бы ещё узнать немного о самом датасете, поэтому сформулируем следующие вопросы:

1. Кто лучше справлялся с задачами по предметам - мужчины или женщины?

```
width = 0.3
genders = ['male', 'female']
x = np.arange(len(genders))
scores = [sum(male_score.values())/3, sum(female_score.values())/3]
fig, ax = plt.subplots()
graph = ax.bar(x, scores, width, label='Score')
ax.set_title('Распределение оценок у мужчин и женщин')
```

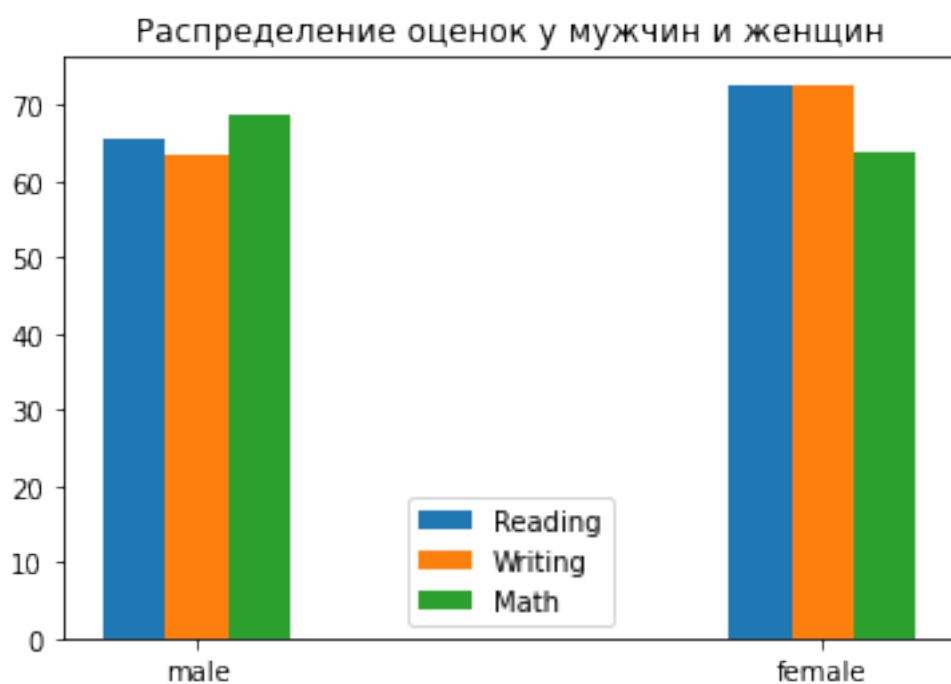
```
ax.set_xticks(x)
ax.set_xticklabels(genders)
ax.legend()
```



Если брать средние баллы, то видно, что женщины справляются с задачами лучше.

Детализируем по предметам вопрос:

2. Кто как из мужчин/женщин справлялся с задачами по каждому предмету в отдельности?



Проведем похожее исследование по поиску зависимостей между категориальными данными и числовыми

3. Гипотеза: чем выше уровень образования у родителей, тем выше средний балл студента

```
import statistics as stat

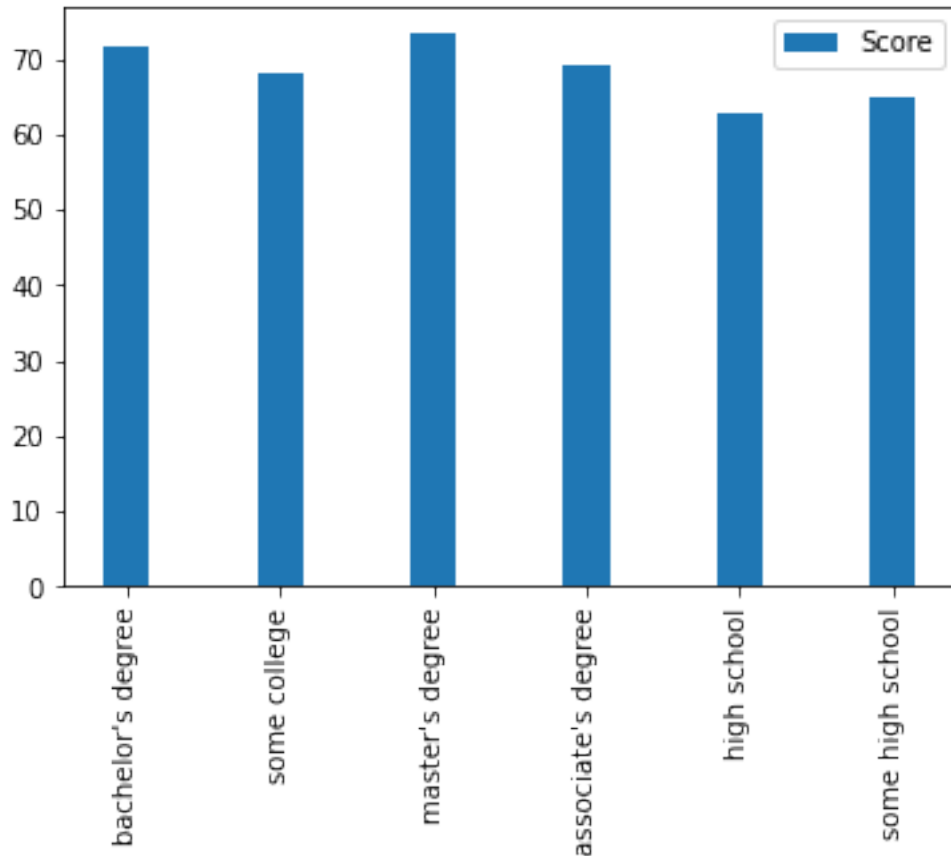
middle_score_list
df['avg_score'] = middle_score_list
parental_levels = df[PLE].unique()
level_scores = {}
scores_only = []
for level in parental_levels:
    level_scores[level] = stat.mean(list(df[df[PLE] == level]['avg_score']))
    scores_only.append(level_scores[level])

x = np.arange(len(parental_levels))
fig, ax = plt.subplots()
graph = ax.bar(x, scores_only, width, label='Score')

ax.set_title('Зависимость уровня сдачи экзамена от образования родителей')
ax.set_xticks(x)
ax.set_xticklabels(parental_levels, rotation = 'vertical')
ax.legend()
```

Из рисунка ниже видно, что самые большие баллы получили дети родителей с высшим образованием. Ранжирование уровня сдачи можно сопоставить уровню образования родителей.

Зависимость уровня сдачи экзамена от образования родителей



4. Гипотеза: баллы по математике плохо коррелируются с баллами по чтению и письму

5. Гипотеза: баллы по письму и чтению коррелируются лучше между собой и со средним значением

Построим матрицу корреляции:

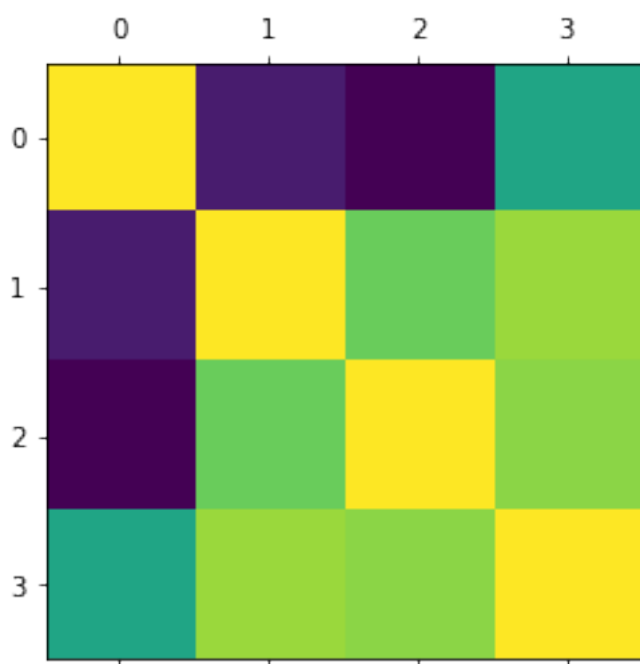
На ней

0 - математика

1 - чтение

2 - письмо

3 - среднее значение



Наши гипотезы 4 и 5 подтвердились частично:

Гипотеза (4) верна. Отсюда может следовать вывод о разделении людей на гуманитариев и технарей (т.к. видно, что оценки по гуманитарным предметам и математике коррелируются не так хорошо).

Гипотеза (5) подтвердилась частично. Письмо и чтение действительно коррелируют друг с другом, однако в сравнении со средним значением у них разброс достаточно большой.