Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика, искусственный интеллект и системы управления

КАФЕДРА Системы обработки информации и управления

Методические указания к лабораторным работам по курсу «Машинное обучение»

Лабораторная работа №1
«Создание "истории о данных" (Data Storytelling)»

Выполнил Поташников М.Д. (ИУ5-24М)

Москва, 2023 г.

ЗАДАНИЕ

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

- 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.

Сформировать отчет и разместить его в своем репозитории на github.

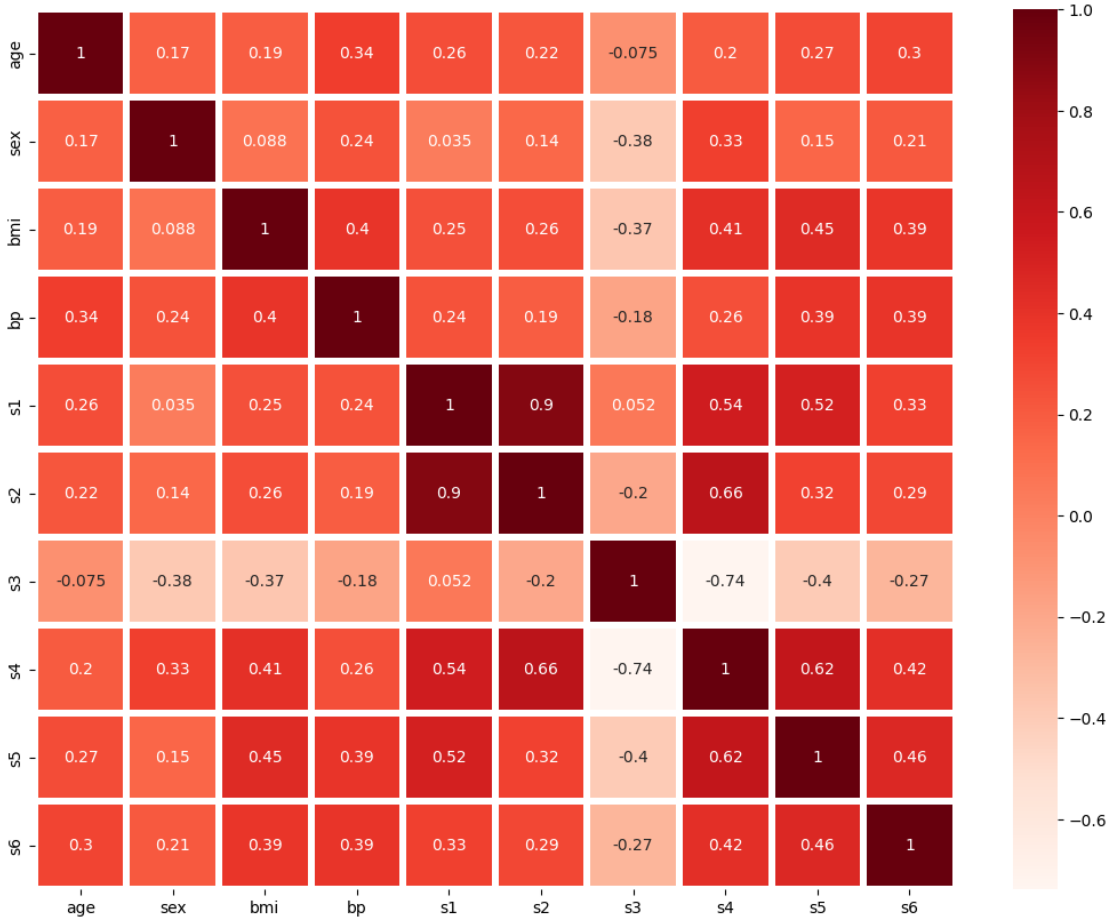
```
import sklearn
import pandas as pd
import seaborn as sns
from sklearn.datasets import load_diabetes
import matplotlib.pyplot as plt
import numpy as np

df = load_diabetes(as_frame=True, scaled=False)

data = df['data']
target = df['target']

plt.figure(figsize=(13,10))
sns.heatmap(data.corr(), cmap = "Reds", annot = True, linewidth=3)
```

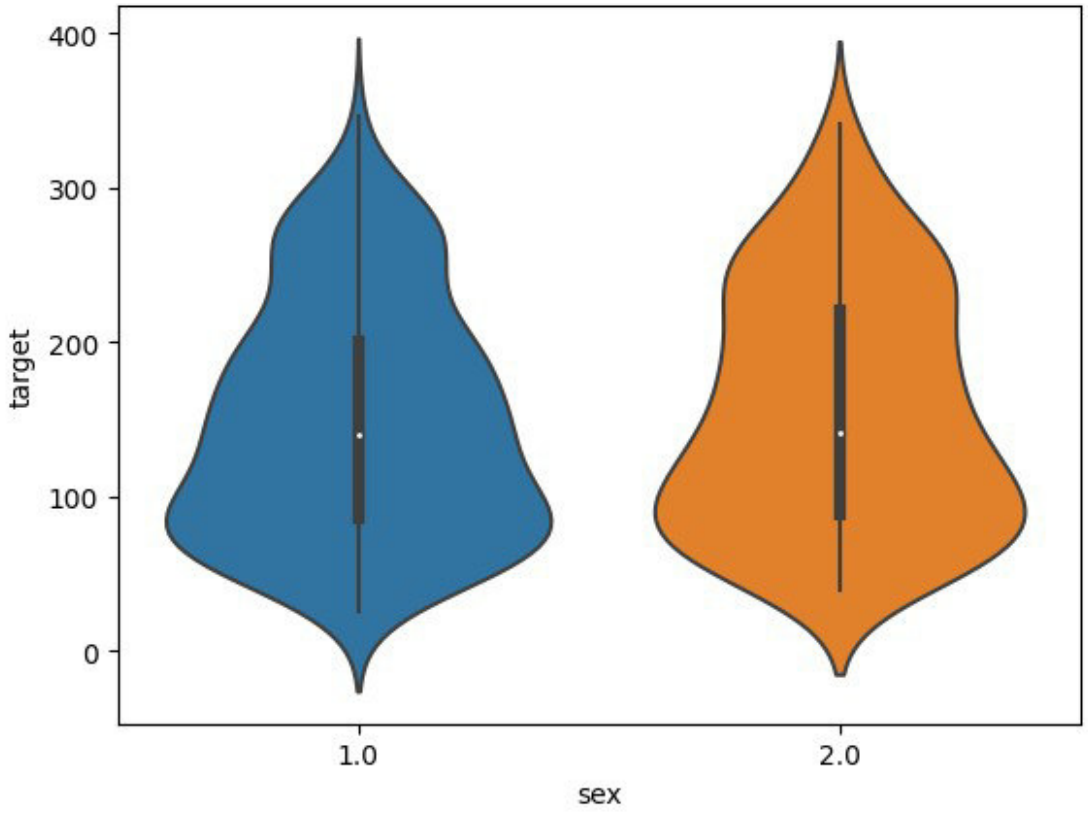
```
<Axes: >
```



По матрице корреляции видно, что данные сильно коррелируют друг с другом, особенно показатели медицинских анализов (s1-s6). Можно объяснить это тем, что у здорового человека скорее всего все анализы будут в норме, а если человек болен, то, соответственно, анализы будут отклоняться от нормы.

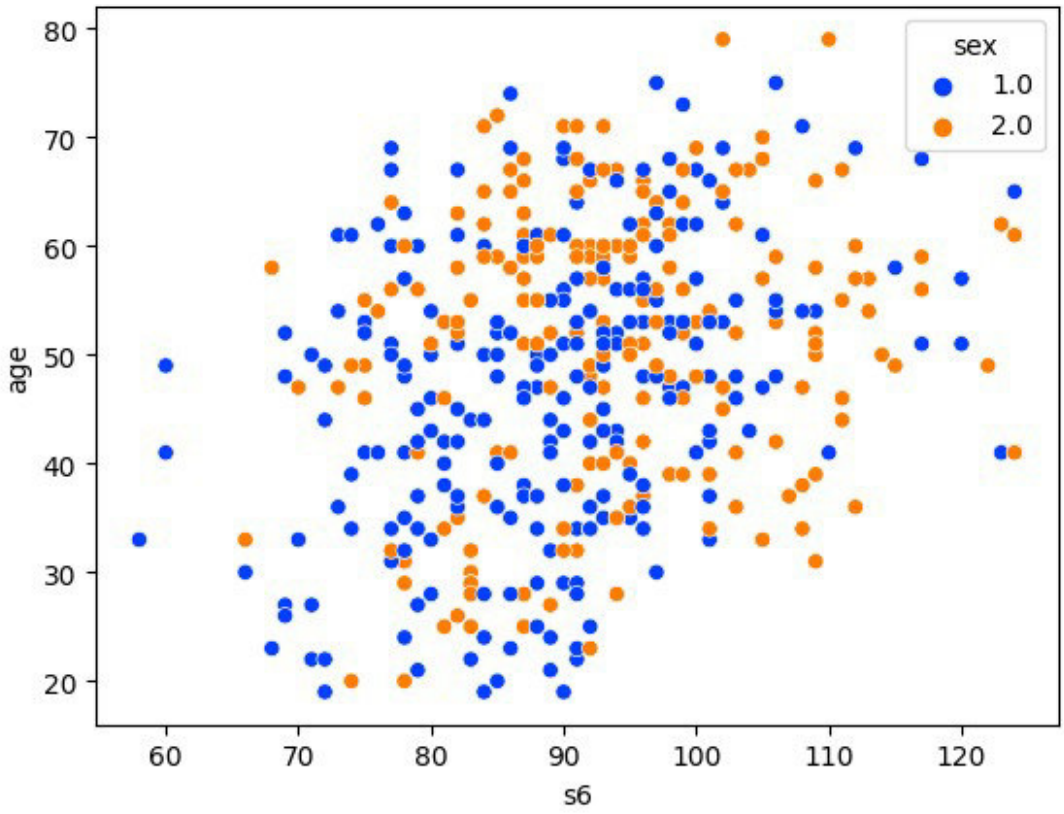
```
sns.violinplot(x=data["sex"], y=target)
```

```
<Axes: xlabel='sex', ylabel='target'>
```



В датасете параметр target представляет собой количественный показатель прогрессирования заболевания через год после исходного уровня. Как мы видим по графику, диабет прогрессирует у обоих полов примерн одинаково.

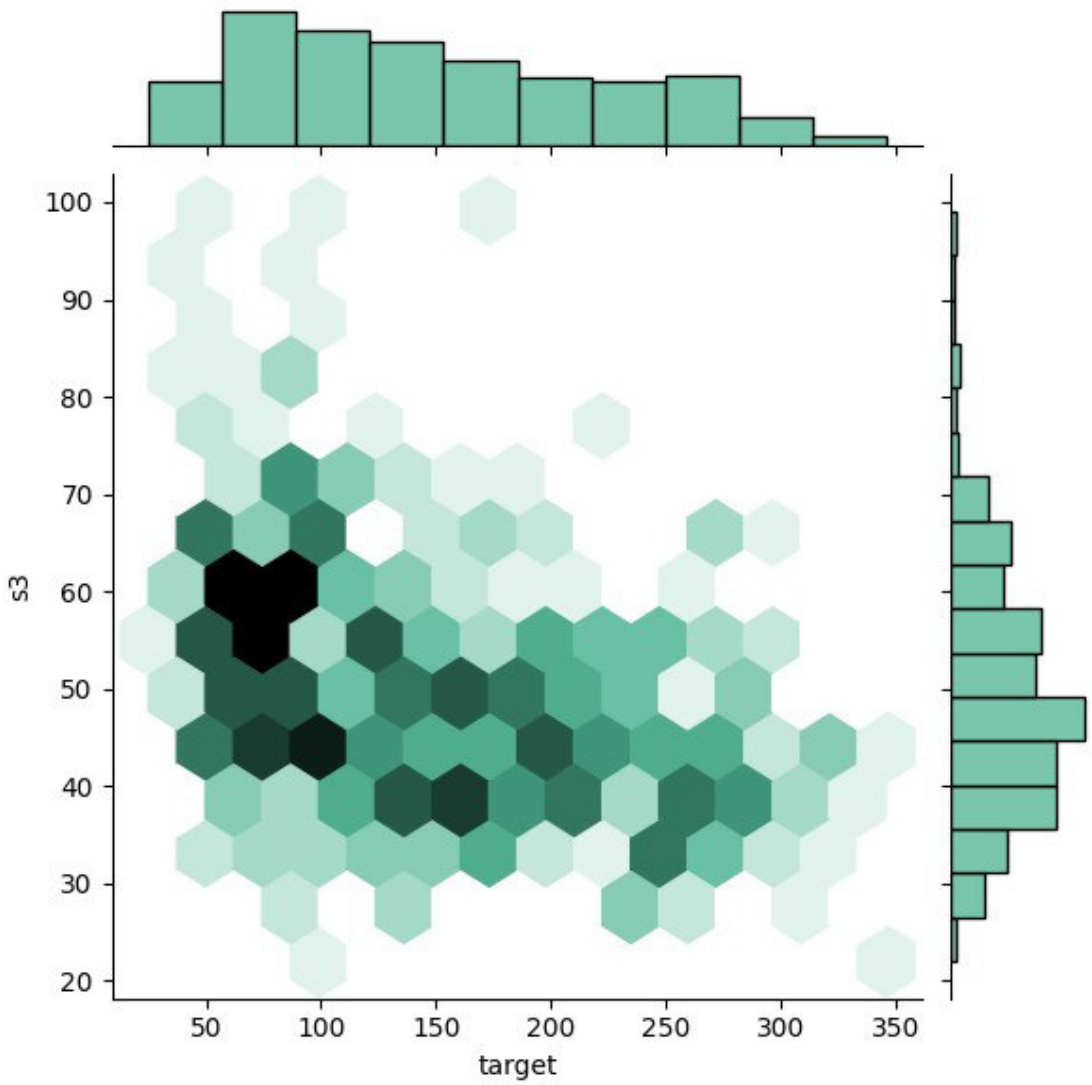
```
g = sns.scatterplot(
    x=data["s6"], y=data["age"],
    hue=data["sex"],
    palette='bright', sizes=(10, 200),
)
```



На графике представлены возраст человека и его уровень сахара в крови (по оси абсцисс). Можно увидеть, что у молодых людей высокого содержания сахара в крови не наблюдается, чего не скажешь о людях в возрасте.

```
sns.jointplot(x=target, y=data['s3'], kind="hex", color="#4CB391")
```

```
<seaborn.axisgrid.JointGrid at 0x7fd41461e9d0>
```



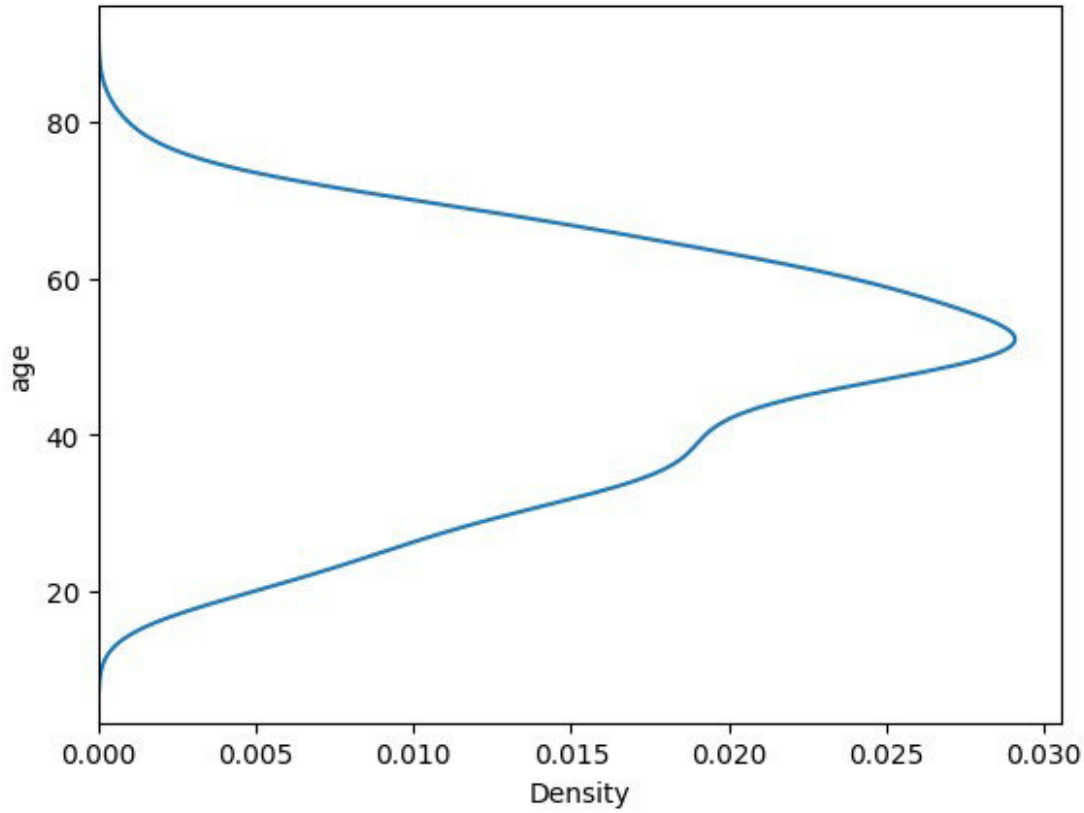
На графике по оси абсцисс отложен количественный показатель прогрессирования заболевания через год после исходного уровня, а по оси ординат - липопротеины высокой плотности (что бы это не значило). Видно, что последний параметр не сильно влияет на прогрессирования заболевания, так как он на одном уровне как у людей с сильно прогрессирующим заболеванием, так и у тех у кого оно протекает медленно.

```
sns.kdeplot(data['age'], vertical=True)
plt.show()
```

```
<ipython-input-57-a531ba994ee5>:1: UserWarning:

The `vertical` parameter is deprecated; assigning data to `y`.
This will become an error in seaborn v0.13.0; please update your code.
```

```
sns.kdeplot(data['age'], vertical=True)
```



Можно заметить по графику, что диабет беспокоит в основном людей в возрасте (около 60 лет).