МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Лабораторная работа №1
по курсу «СТРУКТУРНОЕ ПРОЕКТИРОВАНИЕ АСОИУ»

Тема: «Разведочный анализ данных. Исследование и визуализация данных»

ИСПОЛНИТЕЛЬ: Чертилин А.А.
ФИО
группа ИУ5-22М _____
подпись
"__"_____2019 г.

ПРЕПОДАВАТЕЛЬ: Гапанюк Ю.Е.
ФИО
_____
подпись
"__"_____2019 г.

Москва - 2018

Цель лабораторной работы

Цель лабораторной работы: изучение различных методов визуализация данных.

Задание

Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов на Kaggle.com. Для лабораторных работ не рекомендуется выбирать датасеты большого размера.
Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков. Сформировать отчет и разместить его в своем репозитории на github.

Описание датасета

Датасет HeHeart Disease UCI (болезни сердца)
Информация об атрибутах:
1. Возраст
2. Пол
3. Тип боли в груди (4 значения)
4. Кровяное давление в покое
5. Сыворотка холестеральная в мг / дл
6. Уровень сахара в крови натощак> 120 мг / дл
7. Результаты электрокардиографии в покое (значения 0,1,2)
8. Достигнута максимальная частота сердечных сокращений
9. Осуществление индуцированной стенокардии
10. Oldpeak = депрессия ST, вызванная физическими упражнениями относительно отдыха
11. Наклон пика упражнений сегмента ST
12. количество крупных сосудов (0-3), окрашенных по цвету

13. тал: 3 = нормально; 6 = исправленный дефект; 7 = обратимый дефект

Имена и номера социального страхования пациентов были недавно удалены из базы данных, заменены фиктивными значениями. Один файл был "обработан", тот, который содержит базу данных Кливленда. Все четыре необработанных файла также существуют в этом каталоге.

Результат выполнения

ЛР1 Чертилин Андрей

Heart Disease UCI Болезни сердца

```
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt

% matplotlib inline
sb.set(style='ticks')
```

Загрузка файла

```
data = pd.read_csv('heart.csv', sep=",")
```

Основные характеристики датасета

```
data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
0	63	1	3	145	233	1	0	150	0	2.3	0
1	37	1	2	130	250	0	1	187	0	3.5	0
2	41	0	1	130	204	0	0	172	0	1.4	2
3	56	1	1	120	236	0	1	178	0	0.8	2
4	57	0	0	120	354	0	1	163	1	0.6	2

Размер датасета, столбцы и типы

```
data.shape
```

```
(303, 14)
```

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 303
```

```
data.columns
```

```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')
```

```
data.dtypes
```

```
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
```

Пустые значения

```
for column in data.columns:
    temp_null_count = data[data[column].isnull()].shape[0]
    print('{} - {}'.format(column,temp_null_count))
```

```
age - 0
sex - 0
cp - 0
trestbps - 0
chol - 0
fbs - 0
restecg - 0
thalach - 0
exang - 0
oldpeak - 0
slope - 0
ca - 0
thal - 0
target - 0
```

Основные статистические характеристики

```
data.describe()
```

	age	sex	cp	trestbps	chol	fbs	reste
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000(

Уникальные значения для целевых признаков

std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525(

```
1 data['age'].unique()
```

```
array([63, 37, 41, 56, 57, 44, 52, 54, 48, 49, 64, 58, 50, 66, 43, 69, 59,
       42, 61, 40, 71, 51, 65, 53, 46, 45, 39, 47, 62, 34, 35, 29, 55, 60,
       67, 68, 74, 76, 70, 38, 77])
```

75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000(

```
1 data['sex'].unique()
```

```
array([1, 0])
```

Графическое исследование датасета

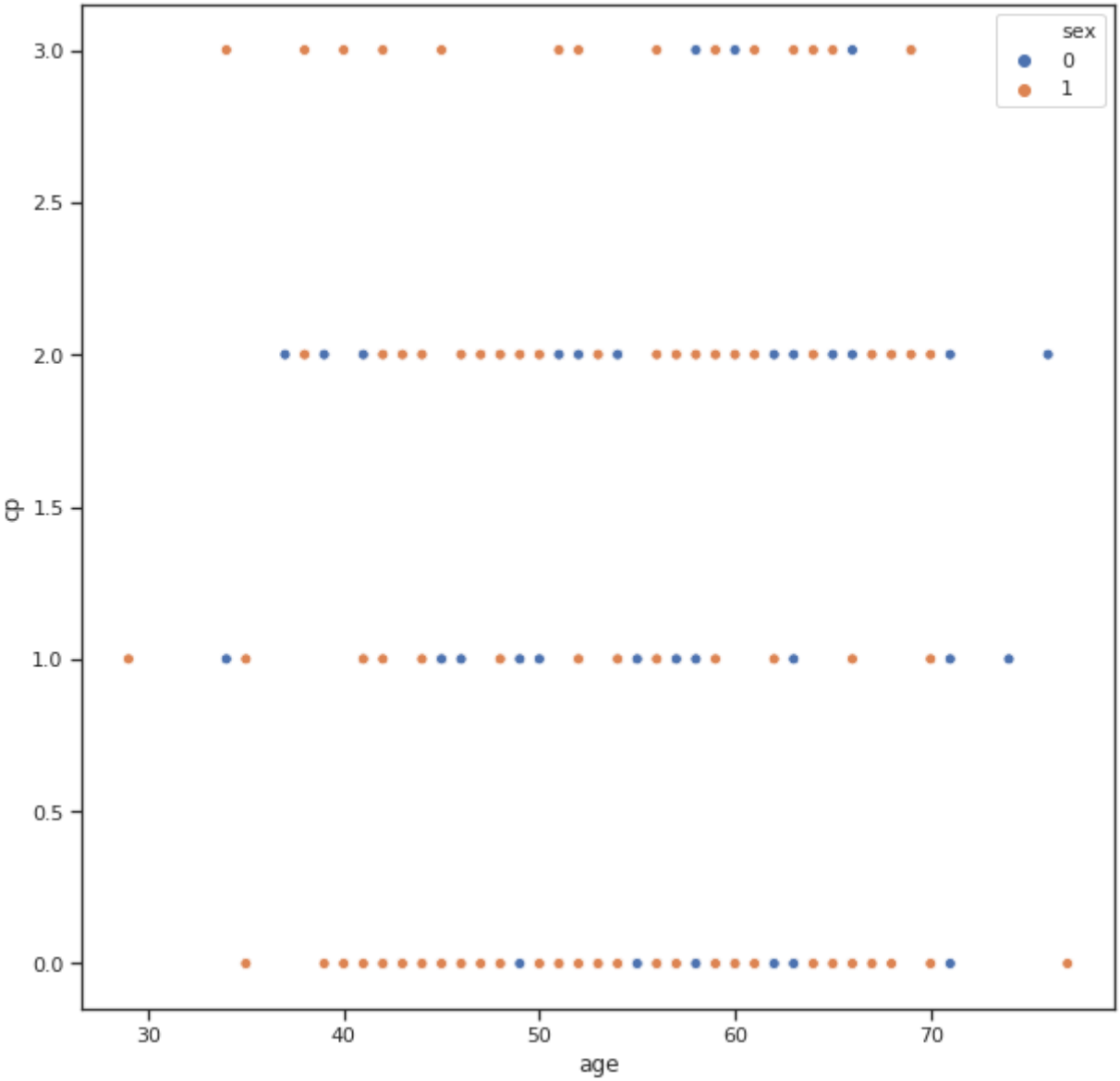
Диаграмма рассеяния

```
1 fig, ax = plt.subplots(figsize=(10,10))
2 sb.scatterplot(ax=ax, x='age', y='sex', data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f88e1eef518>
```


```
1 fig, ax = plt.subplots(figsize=(10,10))
2 sb.scatterplot(ax=ax, x='age', y='cp', data=data, hue='sex')
```

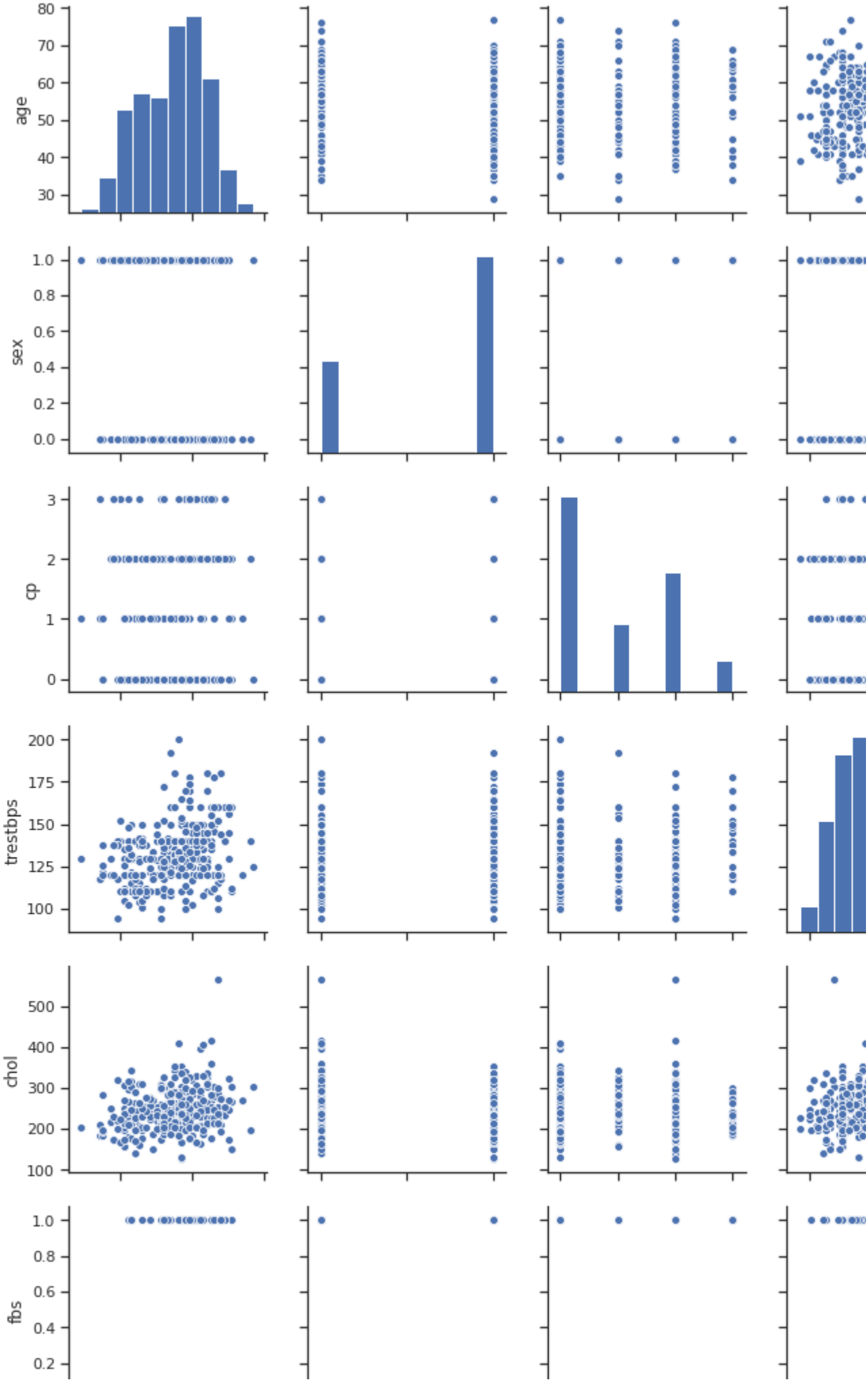
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f88e1e0dfd0>
```

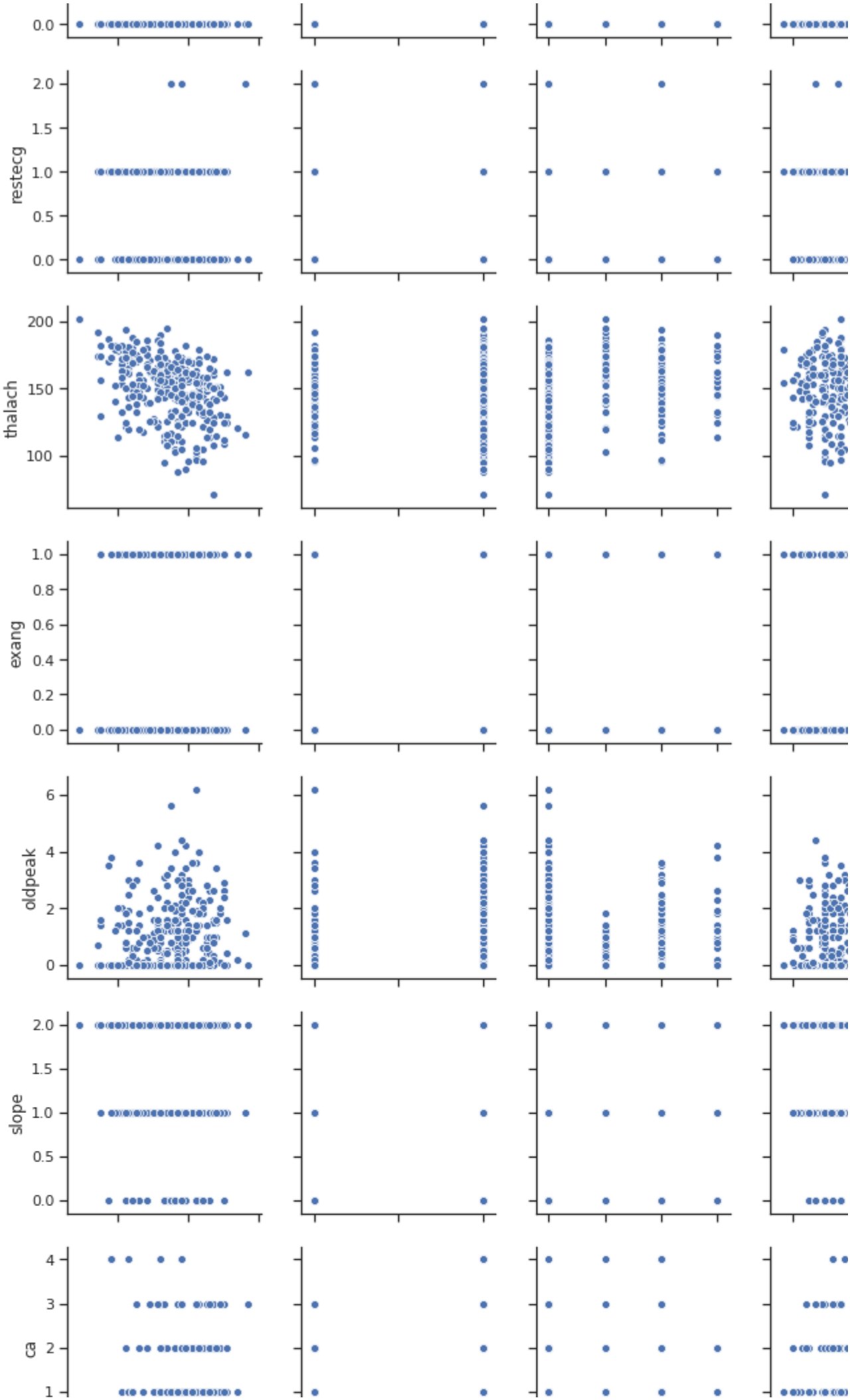


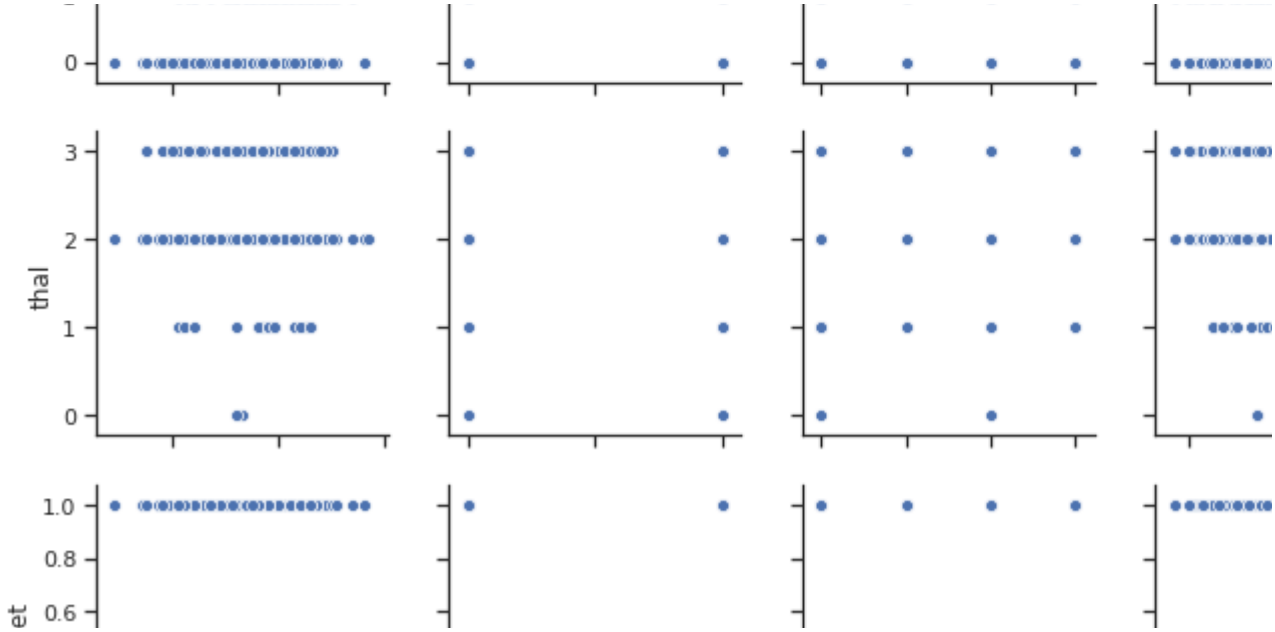
Гистограмма

```
1 sb.pairplot(data)
```

```
<seaborn.axisgrid.PairGrid at 0x7f88e1e17828>
```

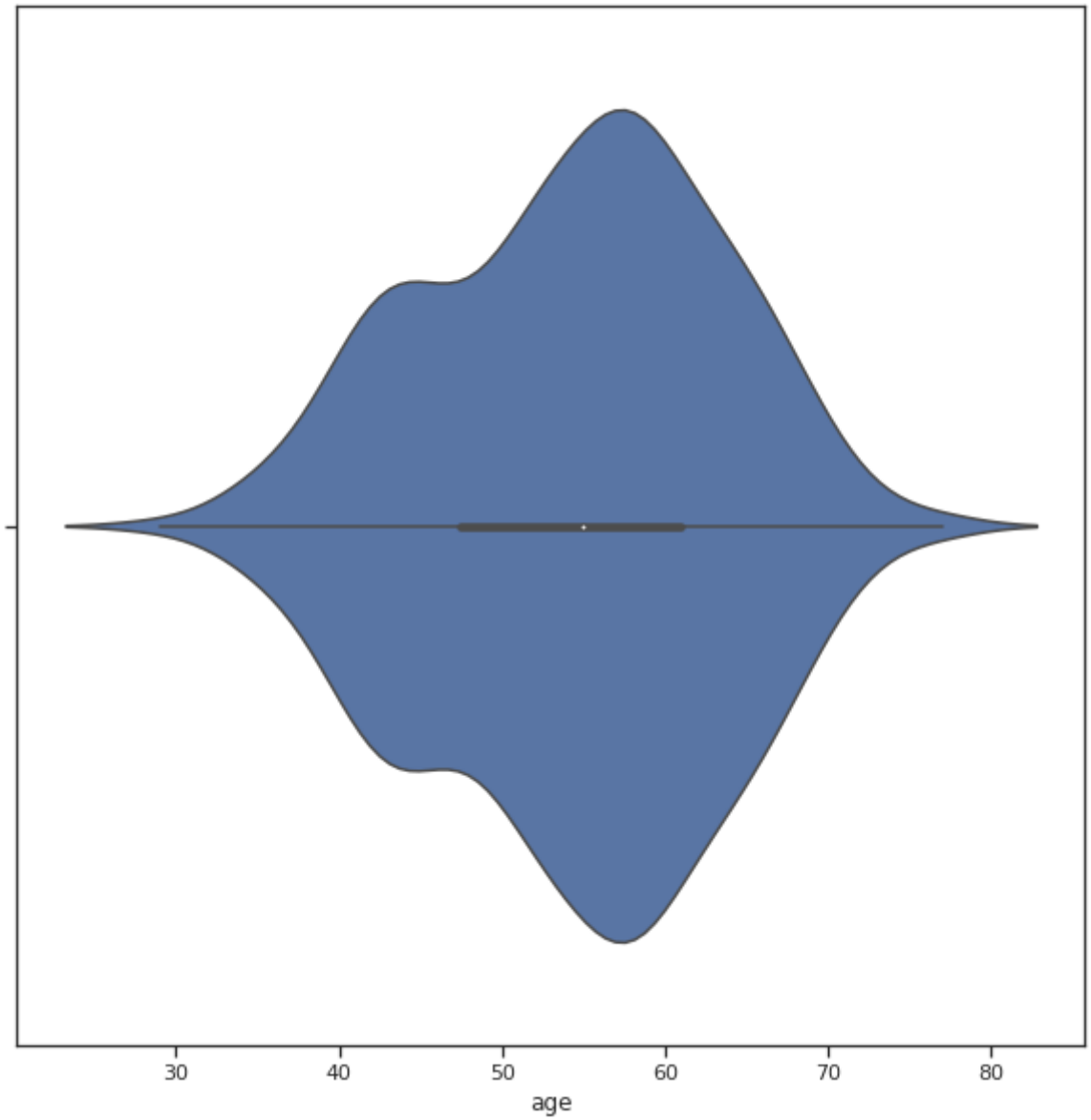



0.0
2.0
1.5
1.0
0.5
0.0
restecg
200
150
100
thalach
1.0
0.8
0.6
0.4
0.2
0.0
exang
6
4
2
0
oldpeak
2.0
1.5
1.0
0.5
0.0
slope
4
3
2
1
ca




```
fig, ax = plt.subplots(1, 1, figsize=(10,10))
sb.violinplot(x=data['age'])
```

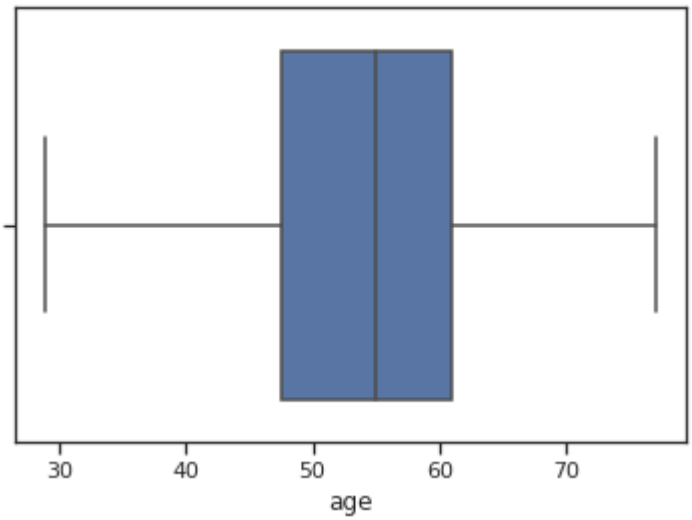
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f88dcf86278>
```



boxplot

```
1 sb.boxplot(x=data['age'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f88de635cc0>
```



Корреляция признаков

```
1 data.corr()
```

	age	sex	cp	trestbps	chol	fbs	restecg	t
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.
trestbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.
exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.
slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.
ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.
thal	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.
target	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.

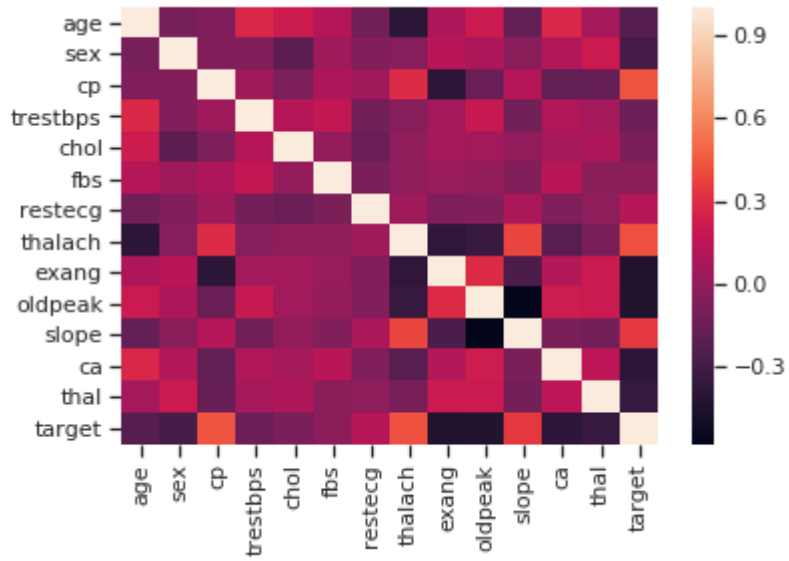
```
1 data.corr(method='pearson')
```

	age	sex	cp	trestbps	chol	fbs	restecg	th
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.
trestbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.
exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.
slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.
ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.

Корреляция с графиками

```
1 sb.heatmap(data.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f88de45a9e8>
```




```
1 fig, ax = plt.subplots(figsize=(12,12))
2 sb.heatmap(data.corr(), annot=True, fmt='.3f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f88dc91ddd8>
```