Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-21М
Маматкулов У.Б

Москва — 2021 г.

1. Цель лабораторной работы

Изучить различные методы визуализации данных [1].

2. Задание

Требуется выполнить следующие действия [1]:

• Выбрать набор данных (датасет).

• Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного 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

3. Ход выполнения работы

3.1. Текстовое описание набора данных

Сердце - это удивительный орган. Он бьется ровно, ровно, примерно от 60 до 100 раз каждую минуту. Это примерно 100 000 раз в день. Иногда твое сердце выходит из ритма. Ваш врач называет нерегулярное или неправильное сердцебиение аритмией. Аритмия (также называемая дисритмией) может вызывать неравномерное сердцебиение или сердцебиение, которое либо слишком медленное, либо слишком быстрое.

3.2. Основные характеристики набора данных

Подключим все необходимые библиотеки [1]

```
[ ]  from datetime import datetime
     import pandas as pd
     import seaborn as sns
```

```
[ ]  # Enable inline plots
     %matplotlib inline
     # Set plot style
     sns.set(style="ticks")
     # Set plots formats to save high resolution PNG
     from IPython.display import set_matplotlib_formats
     set_matplotlib_formats("retina")
```

```
[30] pd.set_option("display.width", 70)
```

```
     data = pd.read_csv("./heart.csv")
```

Настроим отображение графиков [3,4]:

Загрузим непосредственно данные[5]

```
[ ]  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
```

Посмотрим на данные в данном наборе данных:

```
[ ] data.head()
```

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

Проверим размер набора данных:

```
[ ] df = data.copy()
    df.shape
```

```
(303, 14)
```

Проверим основные статистические характеристики набора данных:

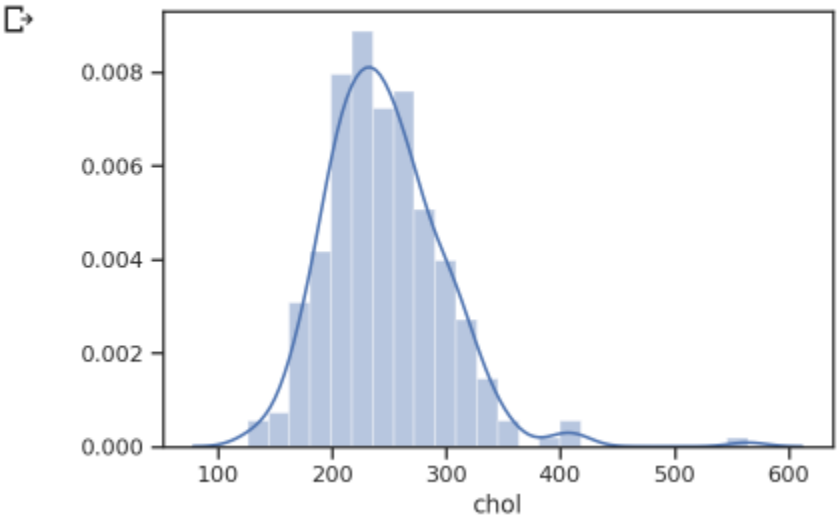
```
[ ] df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000

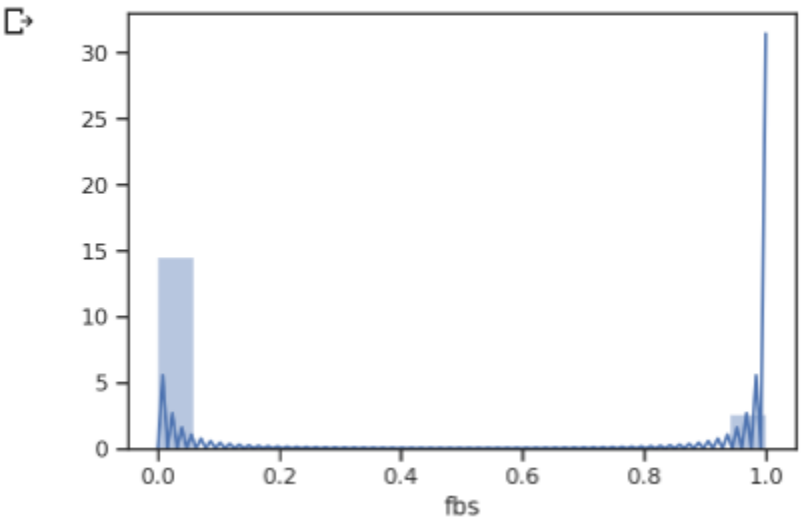
3.3. Визуальное исследование датасета

Давайте оценим распределение целевого атрибута - Рейтинг:

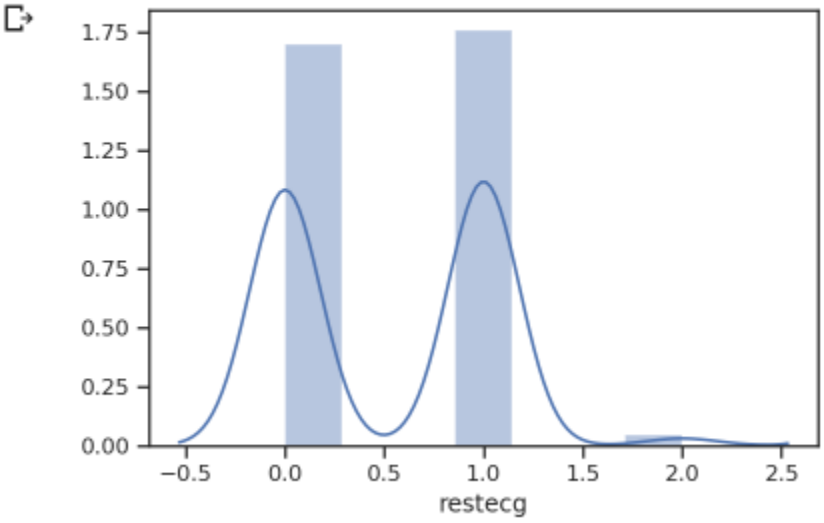
```
[ ] sns.distplot(df["chol"]);
```



```
[ ]  sns.distplot(df["fbs"]);
```

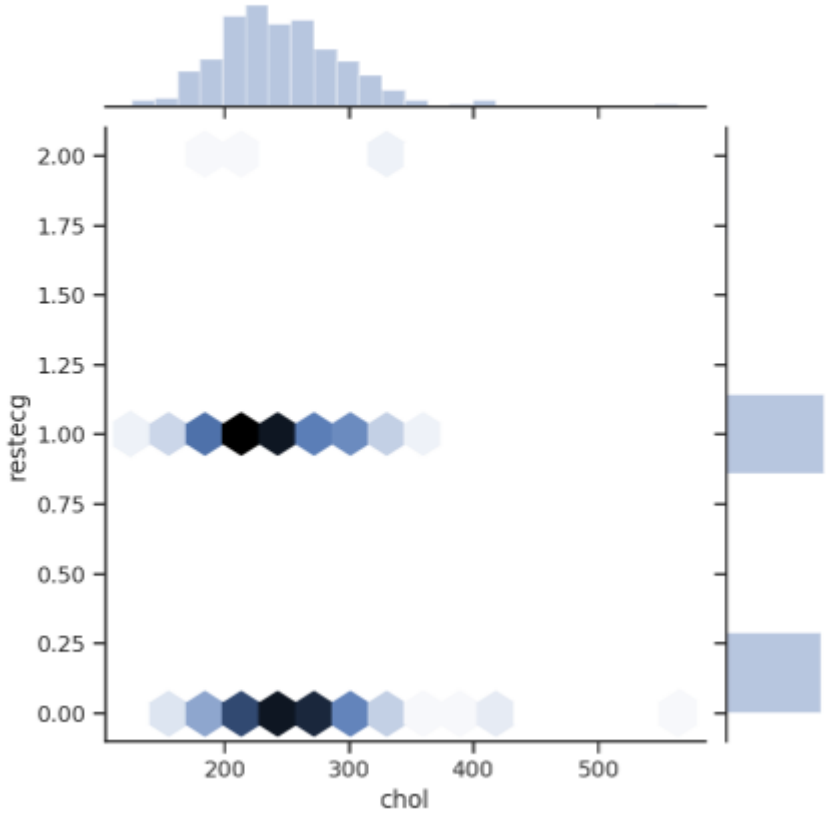



```
[ ]  sns.distplot(df["restecg"]);
```



```
[ ]  sns.jointplot(x="chol",y="restecg",data=df,kind="hex")
```

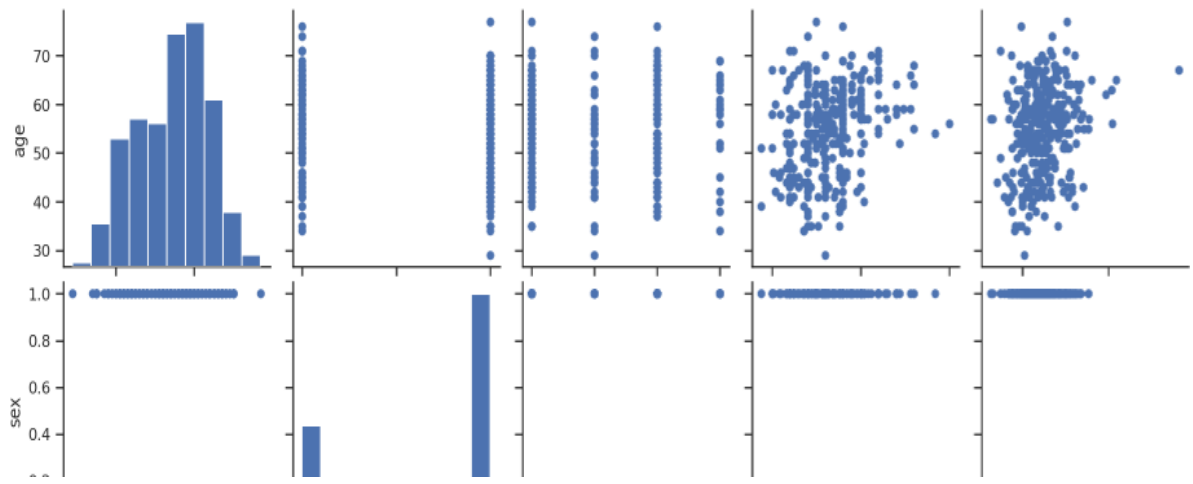
```
<seaborn.axisgrid.JointGrid at 0x7f7821abb438>
```



Построим парные диаграммы по всем показателям по исходному набору данных:

```
sns.pairplot(df, plot_kws=dict(linewidth=0));
```

```
[ ]  sns.pairplot(df,plot_kws=dict(linewidth=0));
```



3.4. Информация о корреляции признаков

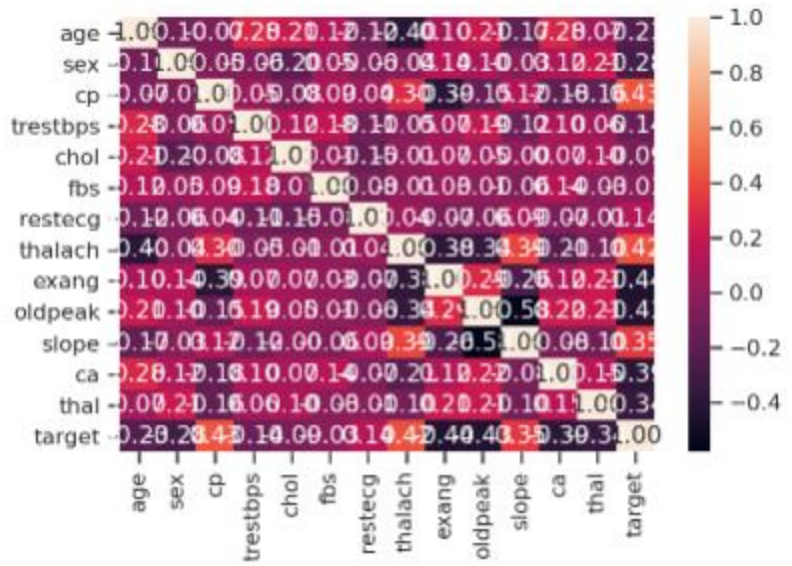
```
df.corr()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326	0.
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261	0.
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053	-0.
trestbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389	0.
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511	0.
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979	-0.
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042	-0.
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177	-0.
exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739	0.
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682	0.
slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155	-0.
ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000	0.
thal	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832	1.
target	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724	-0.

Построим корреляционную матрицу по всему набору данных:

Визуализируем корреляционную матрицу с помощью тепловой карты:

```
[35]  sns.heatmap(df.corr(),annot=True,fmt=".2f");
```



Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets