Лабораторная работа №0

Задача: научиться определять пациента с сердечными заболеваниями по его показателям

Текущий шаг: провести EDA, преобразовать признаки для будущего обучения, визуализировать данные, и на основе анализа сделать выводы ¶

In [1]:

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

In [2]:

```
heart_info = pd.read_csv("heart.csv")
heart_info.head(20)
```

Out[2]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	outp
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	
10	54	1	0	140	239	0	1	160	0	1.2	2	0	2	
11	48	0	2	130	275	0	1	139	0	0.2	2	0	2	
12	49	1	1	130	266	0	1	171	0	0.6	2	0	2	
13	64	1	3	110	211	0	0	144	1	1.8	1	0	2	
14	58	0	3	150	283	1	0	162	0	1.0	2	0	2	
15	50	0	2	120	219	0	1	158	0	1.6	1	0	2	
16	58	0	2	120	340	0	1	172	0	0.0	2	0	2	
17	66	0	3	150	226	0	1	114	0	2.6	0	0	2	
18	43	1	0	150	247	0	1	171	0	1.5	2	0	2	
19	69	0	3	140	239	0	1	151	0	1.8	2	2	2	

In [3]:

```
heart_info.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trtbps    303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalachh  303 non-null    int64  
 8   exng      303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64  
 11  caa       303 non-null    int64  
 12  thall     303 non-null    int64  
 13  output    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

In [4]:

```
heart_info.describe()
```

Out[4]:

	age	sex	cp	trtbps	chol	fbs	restecg
count	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
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000

In [5]:

```
heart_info.value_counts()
```

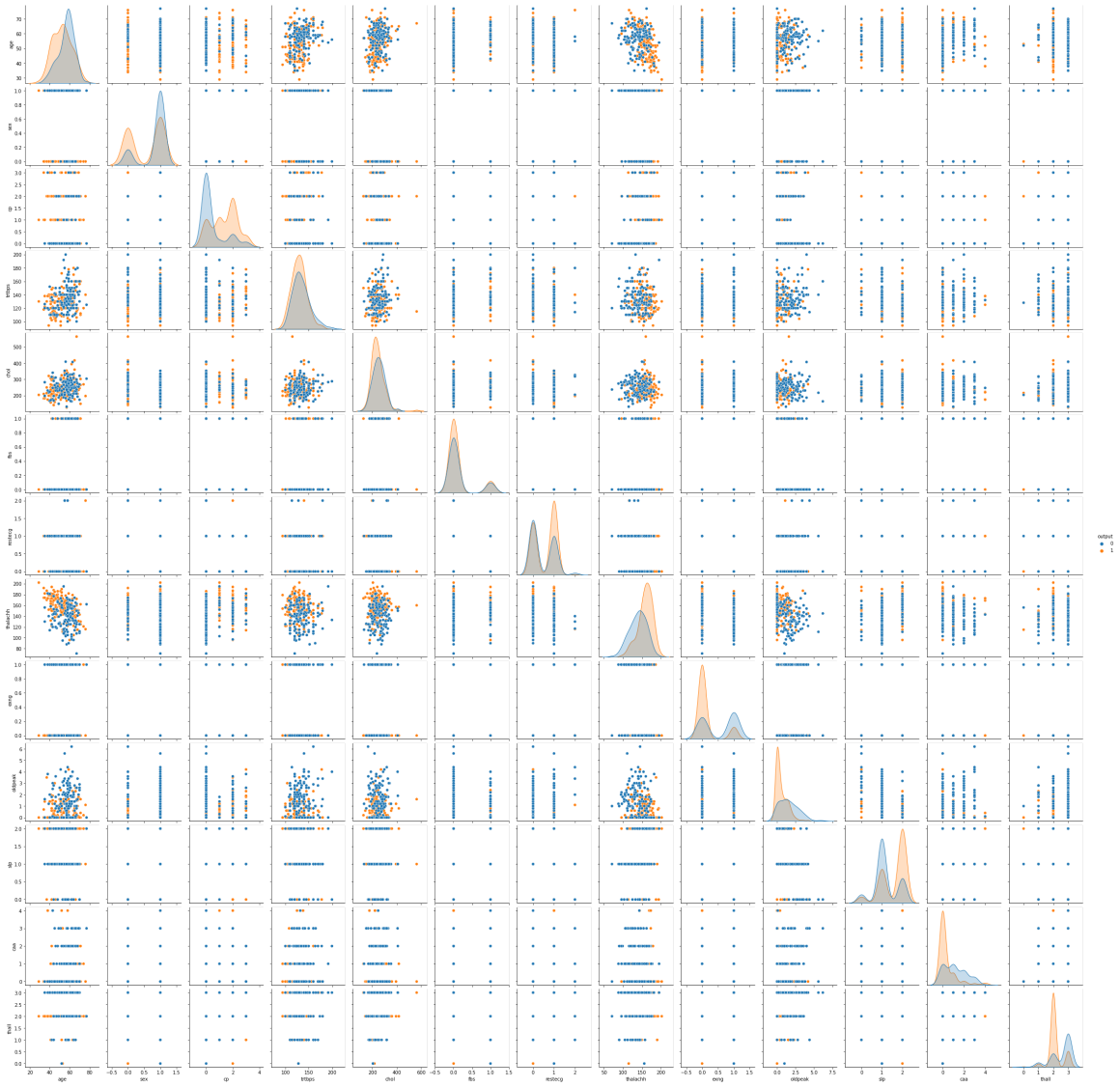
Out[5]:

```
age  sex  cp  trtbps  chol  fbs  restecg  thalachh  exng  oldpeak  slp  ca
a  thall  output
38   1    2   138     175   0    1        173       0     0.0      2    4
2      1         2
59   1    0   110     239   0    0        142       1     1.2      1    1
3      0         1
          2   126     218   1    1        134       0     2.2      1    1
1      0         1
          1   140     221   0    1        164       1     0.0      2    0
2      1         1
          0   170     326   0    0        140       1     3.4      0    0
3      0         1

..
51   1    2   94      227   0    1        154       1     0.0      2    1
3      1         1
          0   140     299   0    1        173       1     1.6      2    0
3      0         1
                    298   0    1        122       1     4.2      1    3
3      0         1
                    261   0    0        186       1     0.0      2    0
2      1         1
77   1    0   125     304   0    0        162       1     0.0      2    3
2      0         1
Length: 302, dtype: int64
```

In [6]:

```
sns.pairplot(heart_info, hue = "output")
plt.show()
```


age
sex
cp
trtbps
chol
fbs
restecg
thalachh
exng
oldpeak
slp
caa
thall
output
0
1

In [9]:

```
heart_info.corr()
```

Out[9]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalach
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.3985
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.0440
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.2957
trtbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.0466
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.0099
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.0085
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.0441
thalachh	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.0000
exng	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.3788
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.3441
slp	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.3867
caa	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.2131
thall	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.0964
output	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.4217

Вывод: **найденный мною датасет обладает дастаточным количеством данных (данные имеют категариальные признаки, одако уже преобразованны к числовому виду), матрица кореляций показывает что данные не имеют ярко-выраженной зависимости лишь по одному признаку, то же самое подтверждает график pairplot, где мы видим облака точек давольно смешанными и на первый взгляд трудно разделяемыми.**