РК ИУ5-63Б Гусева Валерия Вариант № 5

Условие задачи:

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

In [2]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [3]:
```
data = pd.read_csv('heart.csv', sep=',')
```

In [5]:
```
data.head()
```

Out[5]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	ta
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	

In [6]:
```
data.dtypes
```

Out[6]:
```
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
```

In [7]:
```
data.isnull().sum()
```

Out[7]:
```
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
dtype: int64
```

In [8]:
```
data.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   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Вывод:

Пропусков в данных не обнаруженно

Корреляционный анализ

In [13]:
```
data.corr()
```

Out[13]:

	age	sex	cp	trestbps	chol	fbs	restecg	
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	-
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	

In [19]:
```
sns.heatmap(data.corr())
```

Out[19]: <AxesSubplot:>

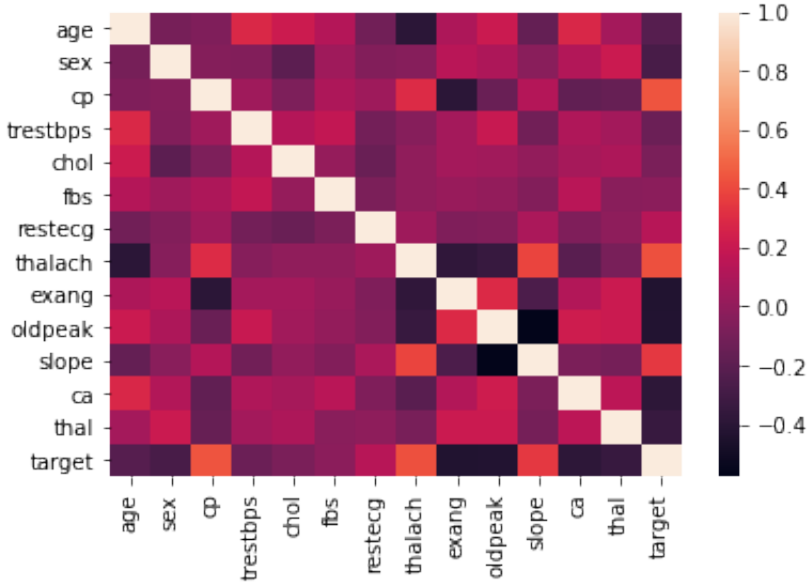
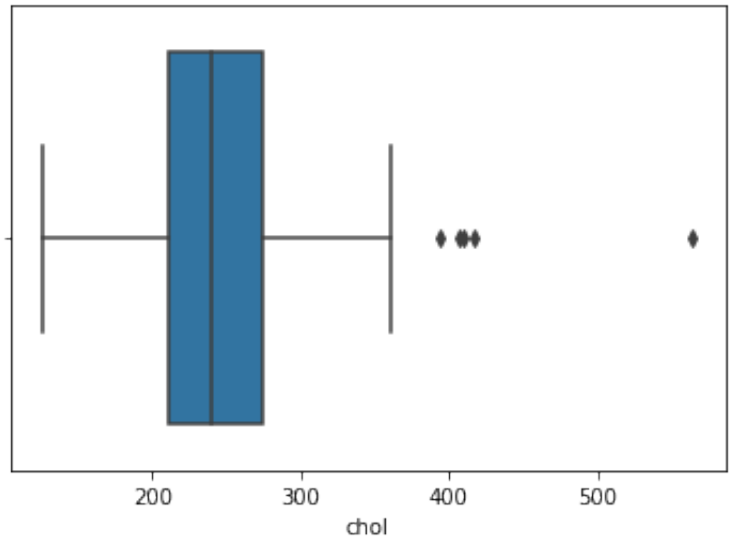


График ящик с усиками

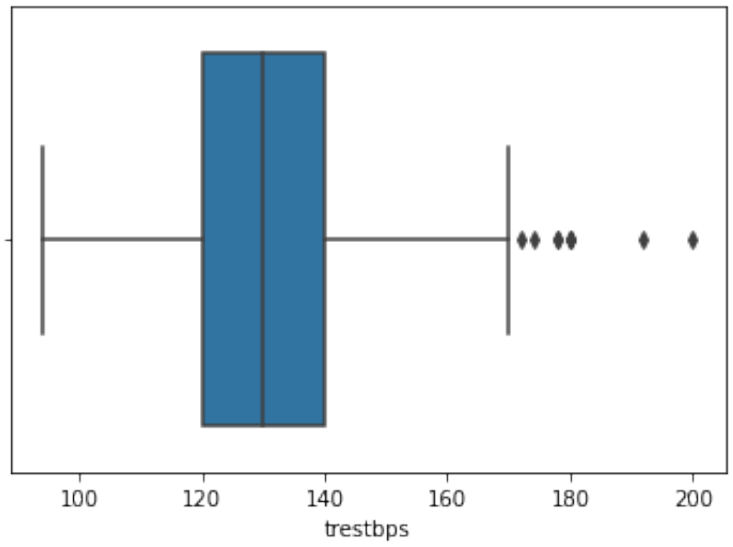
In [15]:
```
sns.boxplot(x=data['chol'])
```

Out[15]: <AxesSubplot:xlabel='chol'>



In [18]:
```
sns.boxplot(x=data['trestbps'])
```

Out[18]: <AxesSubplot:xlabel='trestbps'>



In [ ]: