Аникин Филипп ИУ5-63Б

2 Вариант, задача №1

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель. Доп заданние: для произвольной колонки данных построить график "Ящик с усами (boxplot)".

In [1]:
```
from sklearn.datasets import load_iris
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

In [2]:
```
iris = load_iris()
```

In [3]:
```
iris.data.shape
```

Out[3]:
```
(150, 4)
```

In [4]:
```
data = pd.DataFrame(iris.data)
```

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

Out[5]:
```
<bound method NDFrame.head of         0    1    2    3
0     5.1  3.5  1.4  0.2
1     4.9  3.0  1.4  0.2
2     4.7  3.2  1.3  0.2
3     4.6  3.1  1.5  0.2
4     5.0  3.6  1.4  0.2
..    ...  ...  ...  ...
145   6.7  3.0  5.2  2.3
146   6.3  2.5  5.0  1.9
147   6.5  3.0  5.2  2.0
148   6.2  3.4  5.4  2.3
149   5.9  3.0  5.1  1.8

[150 rows x 4 columns]>
```

In [6]:
```
data.isnull().sum()
```

Out[6]:
```
0    0
1    0
2    0
3    0
dtype: int64
```

In [7]:
```
data.describe()
```

Out[7]:

	0	1	2	3
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

In [8]:
```
data.dtypes
```

Out[8]:
```
0    float64
1    float64
2    float64
3    float64
dtype: object
```

In [9]:
```
data.corr
```

Out[9]:
```
<bound method DataFrame.corr of         0    1    2    3
0     5.1  3.5  1.4  0.2
1     4.9  3.0  1.4  0.2
2     4.7  3.2  1.3  0.2
3     4.6  3.1  1.5  0.2
4     5.0  3.6  1.4  0.2
..    ...  ...  ...  ...
145   6.7  3.0  5.2  2.3
146   6.3  2.5  5.0  1.9
147   6.5  3.0  5.2  2.0
148   6.2  3.4  5.4  2.3
149   5.9  3.0  5.1  1.8

[150 rows x 4 columns]>
```

In [10]:
```
data.corr(method='pearson')
```

Out[10]:

	0	1	2	3
0	1.000000	-0.117570	0.871754	0.817941
1	-0.117570	1.000000	-0.428440	-0.366126
2	0.871754	-0.428440	1.000000	0.962865
3	0.817941	-0.366126	0.962865	1.000000

In [11]:
```
data.corr(method='spearman')
```

Out[11]:

	0	1	2	3
0	1.000000	-0.166778	0.881898	0.834289
1	-0.166778	1.000000	-0.309635	-0.289032
2	0.881898	-0.309635	1.000000	0.937667
3	0.834289	-0.289032	0.937667	1.000000

In [12]:
```
data.corr(method='kendall')
```

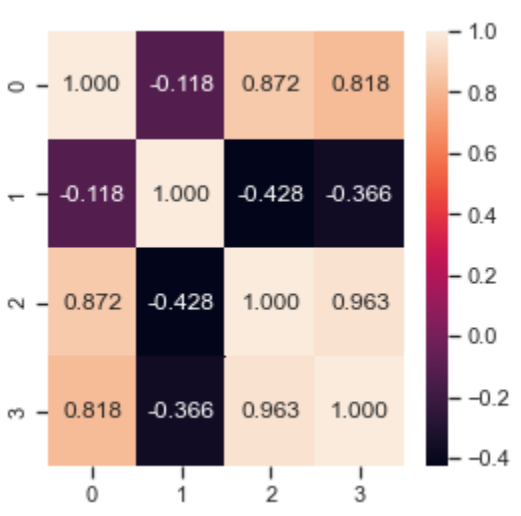
Out[12]:

	0	1	2	3
0	1.000000	-0.076997	0.718516	0.655309
1	-0.076997	1.000000	-0.185994	-0.157126
2	0.718516	-0.185994	1.000000	0.806891
3	0.655309	-0.157126	0.806891	1.000000

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

In [14]:
```
plt.figure(figsize=(4,4))
sns.heatmap(matrix, annot=True, fmt='.3f')
```

Out[14]:
```
<AxesSubplot:>
```

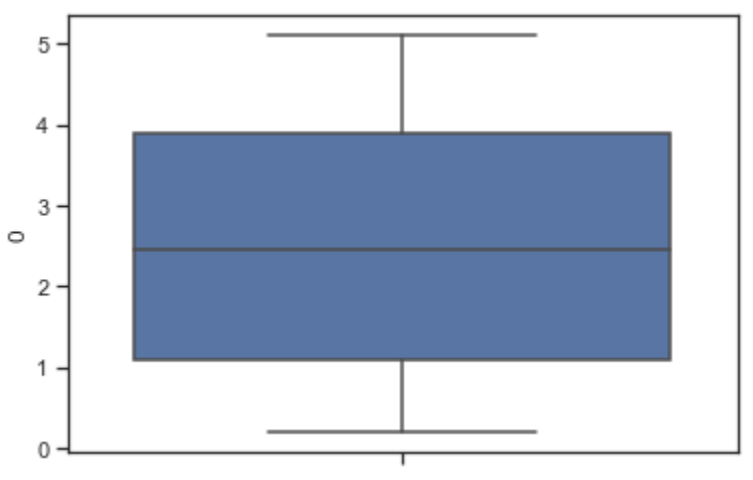


Корреляционная матрица симметрична относительно главной диагонали. На главной диагонали расположены единицы (корреляция признака самого с собой). На основе корреляционной матрицы можно сделать следующие выводы: 1) 1 негативно влияют на корреляционную матрицу, что мешает точной оценке данных. Его стоит удалить. 2) 3 наиболее сильно коррелирует с 2. 3) 0 коррелирует с 3 и 2

Ящик с усами

In [15]:
```
sns.boxplot(y=data.T[0])
```

Out[15]:
```
<AxesSubplot:ylabel='0'>
```



In [16]:
```
sns.boxplot(x=data.T[0])
```

Out[16]:
```
<AxesSubplot:xlabel='0'>
```