FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION
OF HIGHER EDUCATION
ITMO UNIVERSITY

Report
on learning practice №2
Analysis of multivariate random variables

Performed by:
Roman Bezaev
Andrey Getmanov
J4133c

St. Petersburg
2021

1 Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV

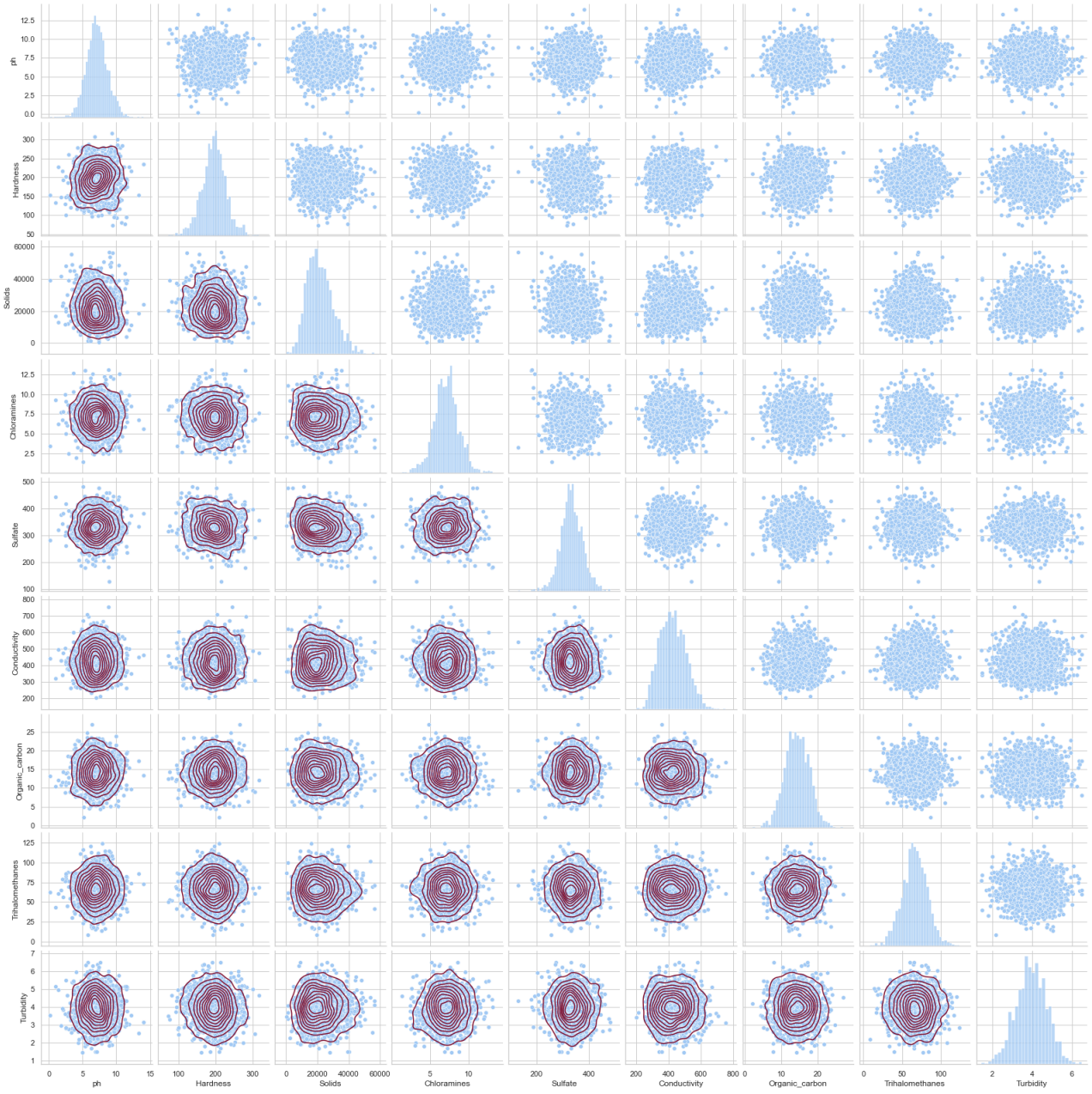
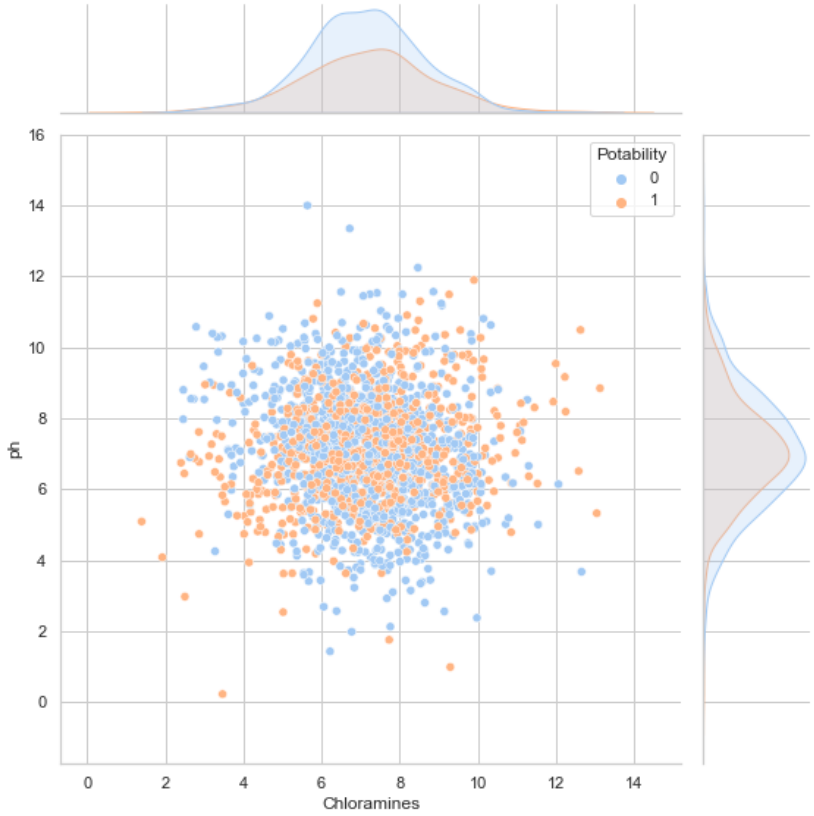
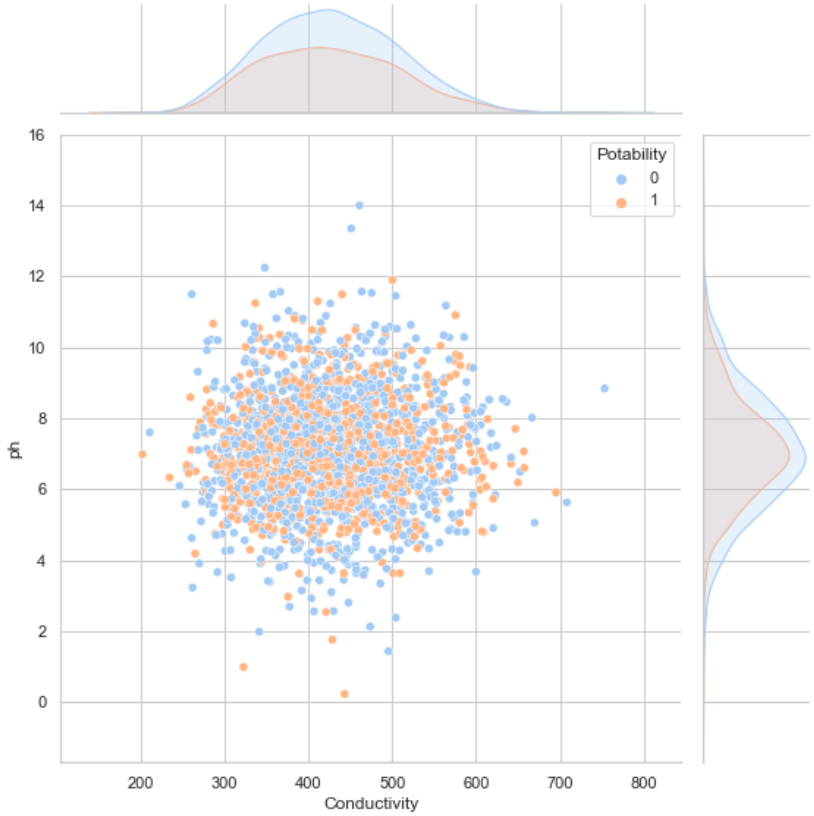
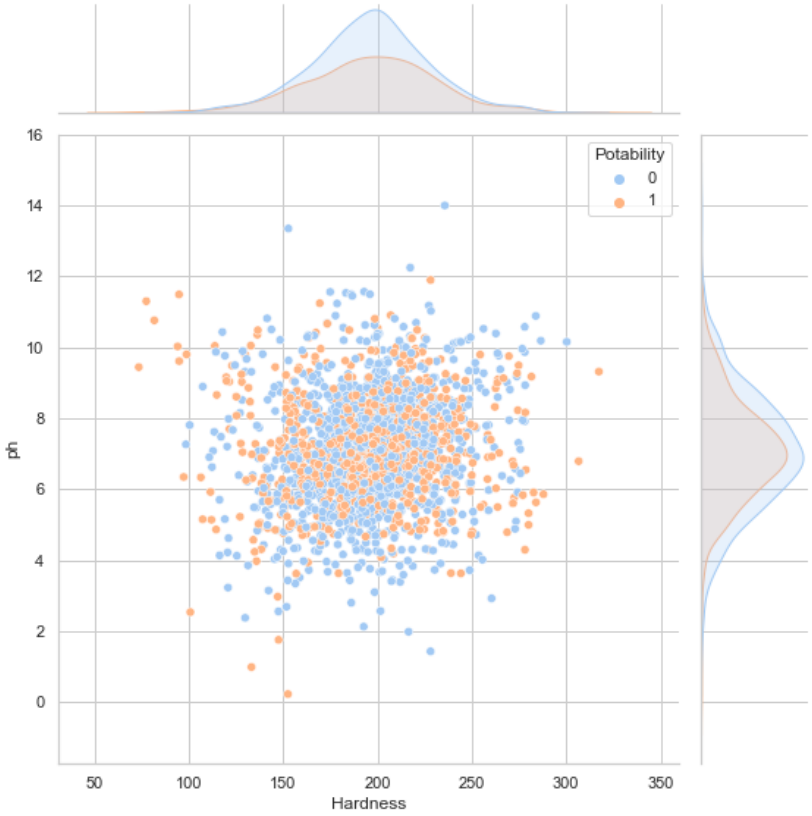
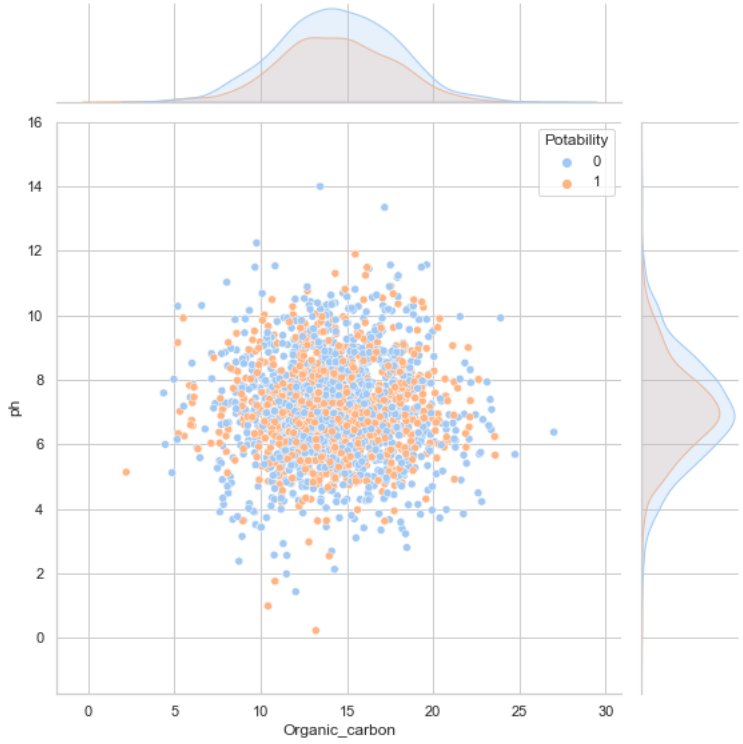


Figure 1: Kde map for dataset

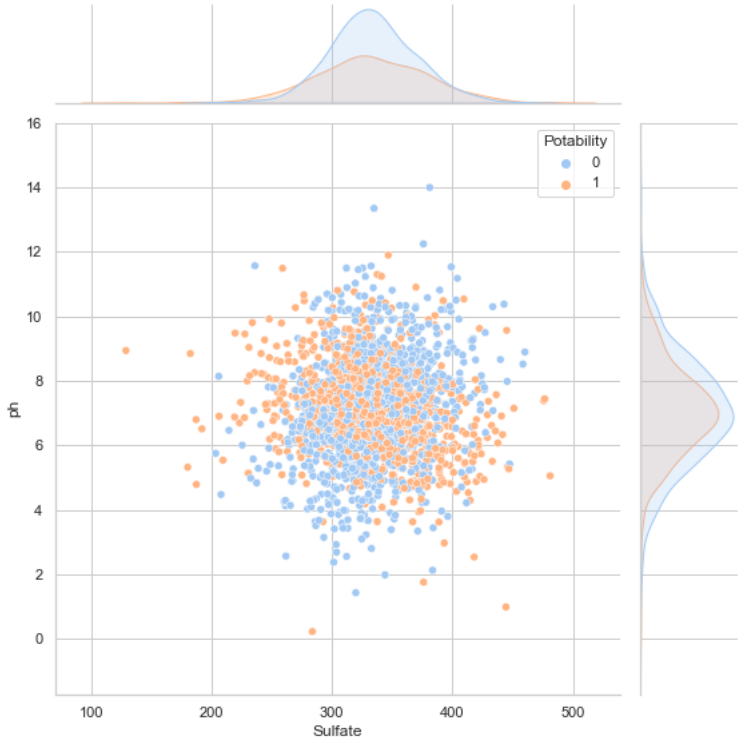

Potability
0
1
ph
Chloramines
0
2
4
6
8
10
12
14
16

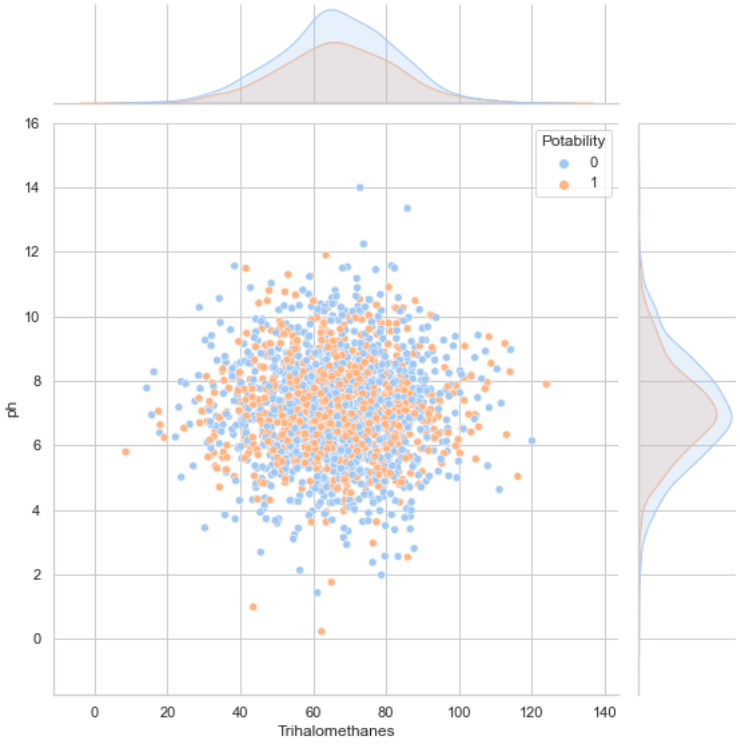

Potability
0
1
ph
Conductivity
200
300
400
500
600
700
800
0
2
4
6
8
10
12
14
16

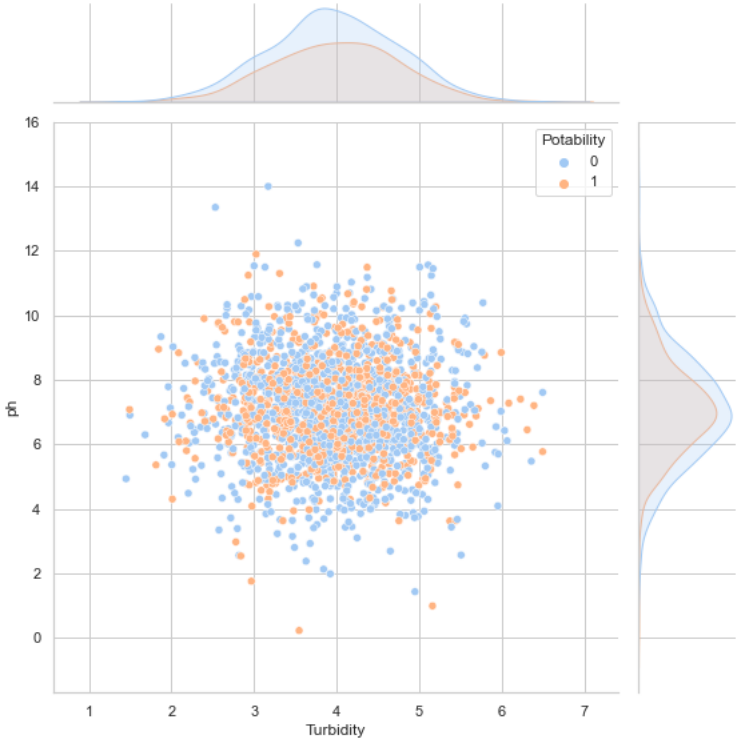

Potability
0
1
ph
Hardness
16
14
12
10
8
6
4
2
0
50
100
150
200
250
300
350


Potability
0
1
ph
Organic_carbon
16
14
12
10
8
6
4
2
0
0
5
10
15
20
25
30


1e−5
Potability
0
1
ph
Solids
0
10000
20000
30000
40000
50000
60000
0
2
4
6
8
10
12
14
16


Potability
0
1
ph
Sulfate
100
200
300
400
500
0
2
4
6
8
10
12
14
16


Potability
0
1
ph
Trihalomethanes
0
20
40
60
80
100
120
140
0
2
4
6
8
10
12
14
16


Potability
0
1
ph
Turbidity
1
2
3
4
5
6
7
0
2
4
6
8
10
12
14
16

2 Estimation of multivariate mathematical expectation and variance

Calculating mathematical expectation and variance with separating by categorial feature.

ph	7.085990
Hardness	195.968072
Solids	21917.441374
Chloramines	7.134338
Sulfate	333.224672
Conductivity	426.526409
Organic_carbon	14.357709
Trihalomethanes	66.400859
Turbidity	3.969729

Figure 2: Mean

ph	2.475388e+00
Hardness	1.065049e+03
Solids	7.468831e+07
Chloramines	2.511654e+00
Sulfate	1.697866e+03
Conductivity	6.514519e+03
Organic_carbon	1.105535e+01
Trihalomethanes	2.584734e+02
Turbidity	6.089401e-01

Figure 3: Variance

3 Non-parametric estimation of conditional distributions, mathematical expectations and variances

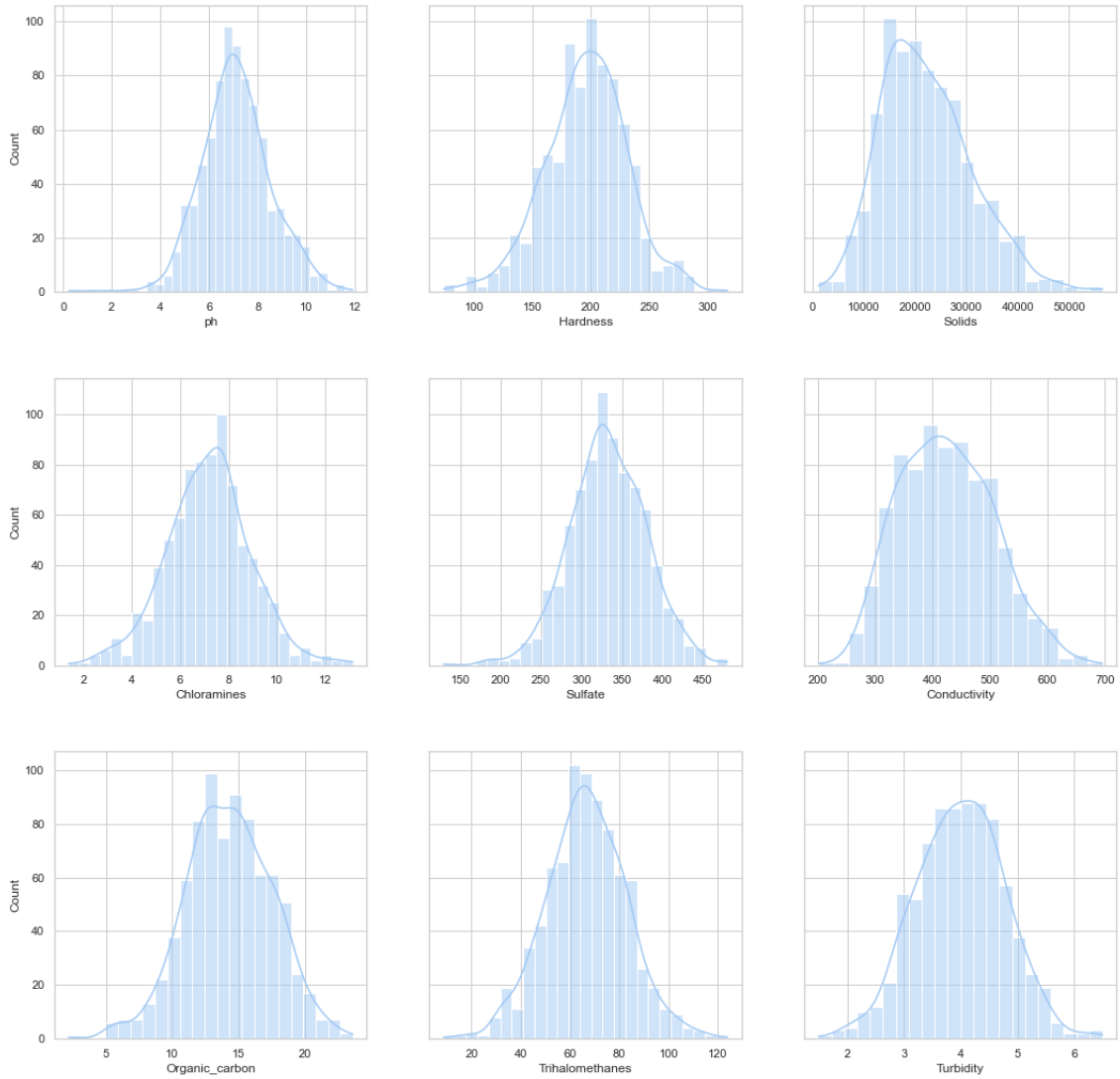


Figure 4: For rows with potability == 1

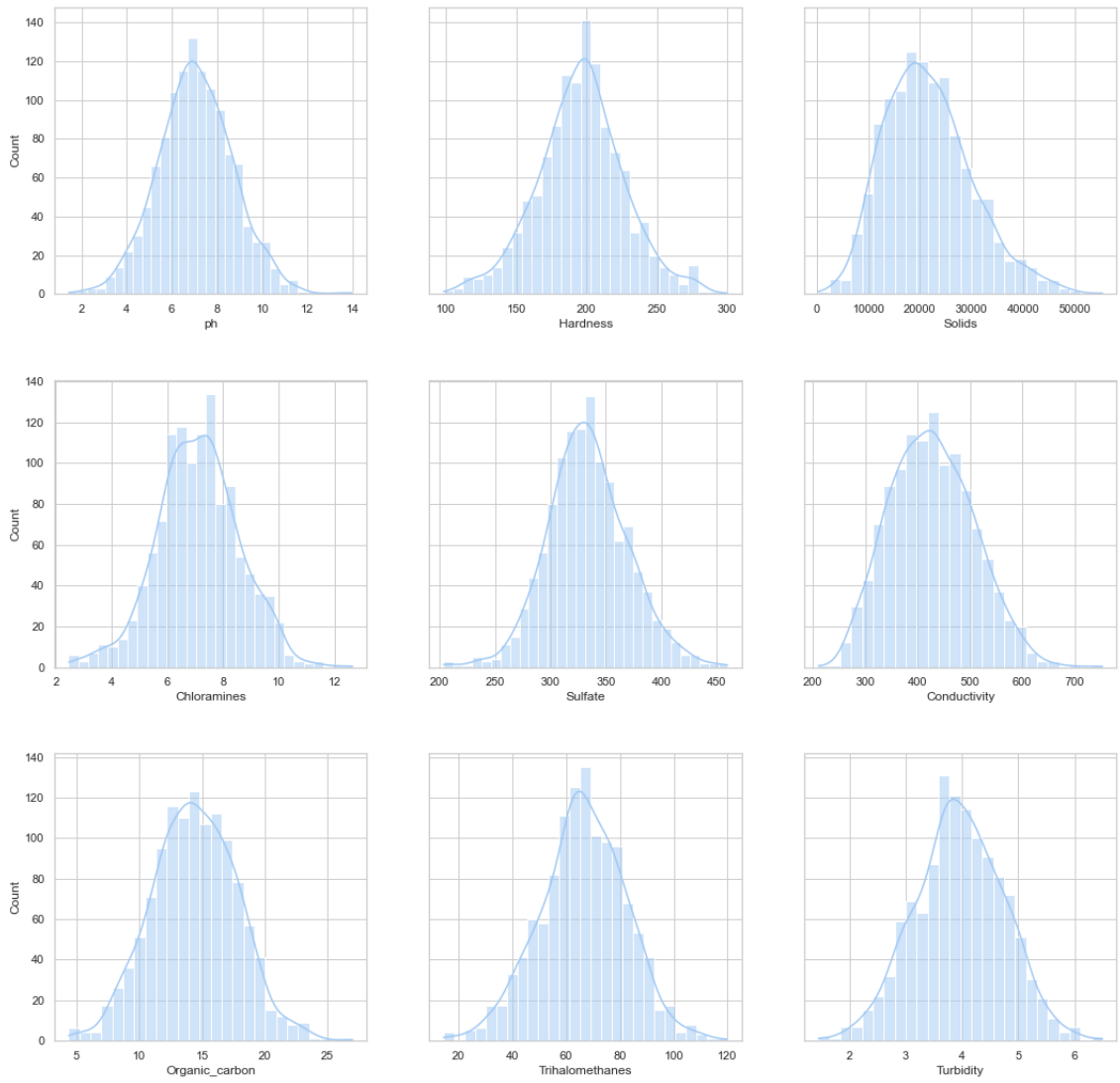


Figure 5: For rows with potability == 0

3.1 For potable water

ph	7.113791
Hardness	195.908341
Solids	22344.922883
Chloramines	7.174395
Sulfate	332.457832
Conductivity	425.005423
Organic_carbon	14.294764
Trihalomethanes	66.581596
Turbidity	3.991254

Figure 6: Mean

ph	2.066759e+00
Hardness	1.246171e+03
Solids	7.905963e+07
Chloramines	3.002580e+00
Sulfate	2.251141e+03
Conductivity	6.715963e+03
Organic_carbon	1.061403e+01
Trihalomethanes	2.656154e+02
Turbidity	6.028091e-01

Figure 7: Variance

3.2 For not potable water

ph	7.067201
Hardness	196.008440
Solids	21628.535122
Chloramines	7.107267
Sulfate	333.742928
Conductivity	427.554342
Organic_carbon	14.400250
Trihalomethanes	66.278712
Turbidity	3.955181

Figure 8: Mean

ph	2.752632e+00
Hardness	9.435735e+02
Solids	7.159036e+07
Chloramines	2.180278e+00
Sulfate	1.324844e+03
Conductivity	6.381242e+03
Organic_carbon	1.135822e+01
Trihalomethanes	2.538271e+02
Turbidity	6.130647e-01

Figure 9: Variance

4 Estimation of pair correlation coefficients, confidence intervals for them and significance levels

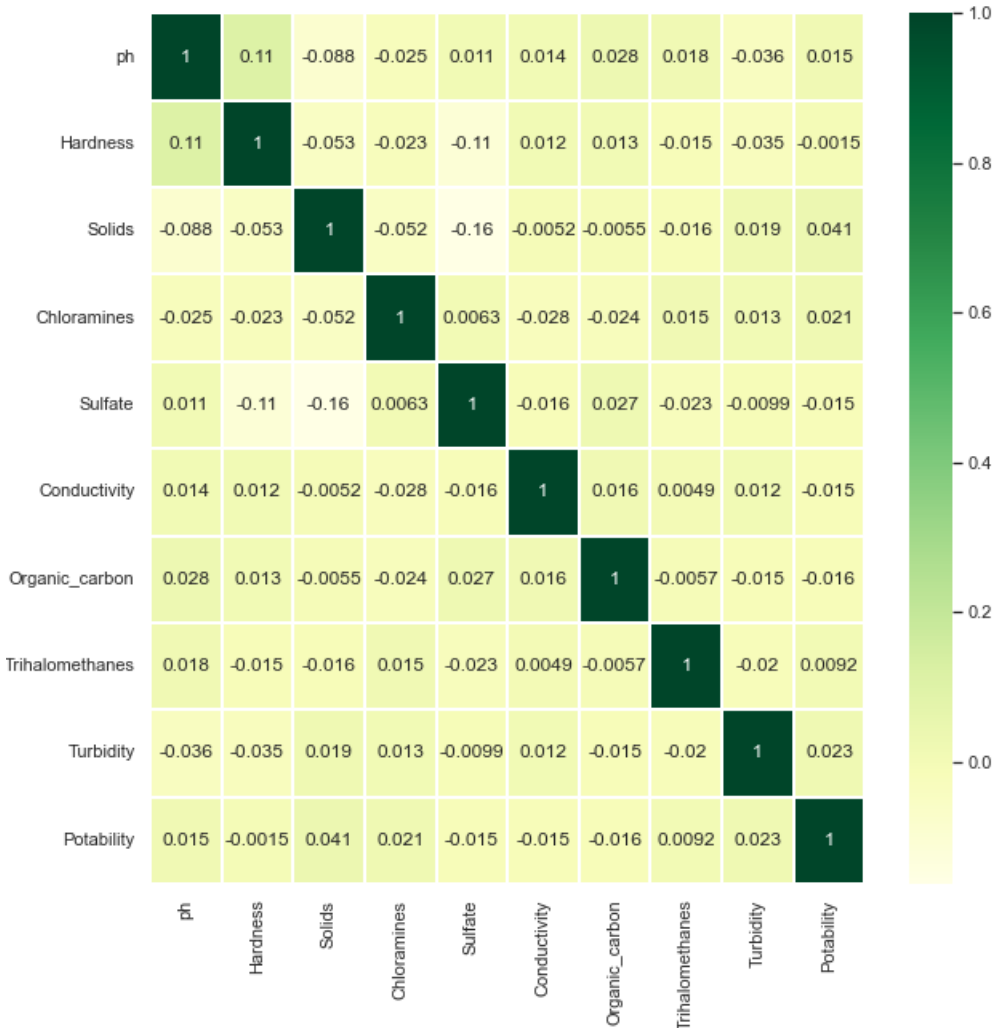


Figure 10: Correlation map

Corr of Hardness and ph is 0.1089, lb-ub 0.0725/0.1451, p-value= 0.0
Corr of Solids and ph is -0.0876, lb-ub -0.1239/-0.0511, p-value= 0.0001
Corr of Chloramines and ph is -0.0248, lb-ub -0.0614/0.0119, p-value= 0.2669
Corr of Sulfate and ph is 0.0105, lb-ub -0.0262/0.0472, p-value= 0.6372
Corr of Conductivity and ph is 0.0141, lb-ub -0.0226/0.0508, p-value= 0.5266
Corr of Organic carbon and ph is 0.0284, lb-ub -0.0083/0.065, p-value= 0.2034
Corr of Trihalomethanes and ph is 0.0183, lb-ub -0.0184/0.0549, p-value= 0.4127

Corr of Turbidity and ph is -0.0358, lb-ub -0.0724/0.0008, p-value= 0.108
Corr of Potability and ph is 0.0145, lb-ub -0.0222/0.0512, p-value= 0.5149
Corr of ph and Hardness is 0.1089, lb-ub 0.0725/0.1451, p-value= 0.0
Corr of Solids and Hardness is -0.0533, lb-ub -0.0898/-0.0166, p-value= 0.0169
Corr of Chloramines and Hardness is -0.0227, lb-ub -0.0593/0.014, p-value= 0.3093
Corr of Sulfate and Hardness is -0.1085, lb-ub -0.1446/-0.0721, p-value= 0.0
Corr of Conductivity and Hardness is 0.0117, lb-ub -0.025/0.0484, p-value= 0.5991
Corr of Organic carbon and Hardness is 0.0132, lb-ub -0.0235/0.0499, p-value= 0.5534
Corr of Trihalomethanes and Hardness is -0.0154, lb-ub -0.0521/0.0213, p-value= 0.4901
Corr of Turbidity and Hardness is -0.0348, lb-ub -0.0714/0.0019, p-value= 0.1184
Corr of Potability and Hardness is -0.0015, lb-ub -0.0382/0.0352, p-value= 0.9462
Corr of ph and Solids is -0.0876, lb-ub -0.1239/-0.0511, p-value= 0.0001
Corr of Hardness and Solids is -0.0533, lb-ub -0.0898/-0.0166, p-value= 0.0169
Corr of Chloramines and Solids is -0.0518, lb-ub -0.0883/-0.0151, p-value= 0.0202
Corr of Sulfate and Solids is -0.1628, lb-ub -0.1983/-0.1268, p-value= 0.0
Corr of Conductivity and Solids is -0.0052, lb-ub -0.0419/0.0315, p-value= 0.8158
Corr of Organic carbon and Solids is -0.0055, lb-ub -0.0422/0.0312, p-value= 0.8059
Corr of Trihalomethanes and Solids is -0.0157, lb-ub -0.0523/0.021, p-value= 0.4825
Corr of Turbidity and Solids is 0.0194, lb-ub -0.0173/0.0561, p-value= 0.3843
Corr of Potability and Solids is 0.0407, lb-ub 0.004/0.0772, p-value= 0.0682

Other values can be found in source code.

5 Task formulation for regression, multivariate correlation

Dependence between each variable and target are very low, that we can not fit a linear regression. R2 score of regression for every feature and target equals 0. For example, there are first four of them:

Figure 11: Sulfate regression

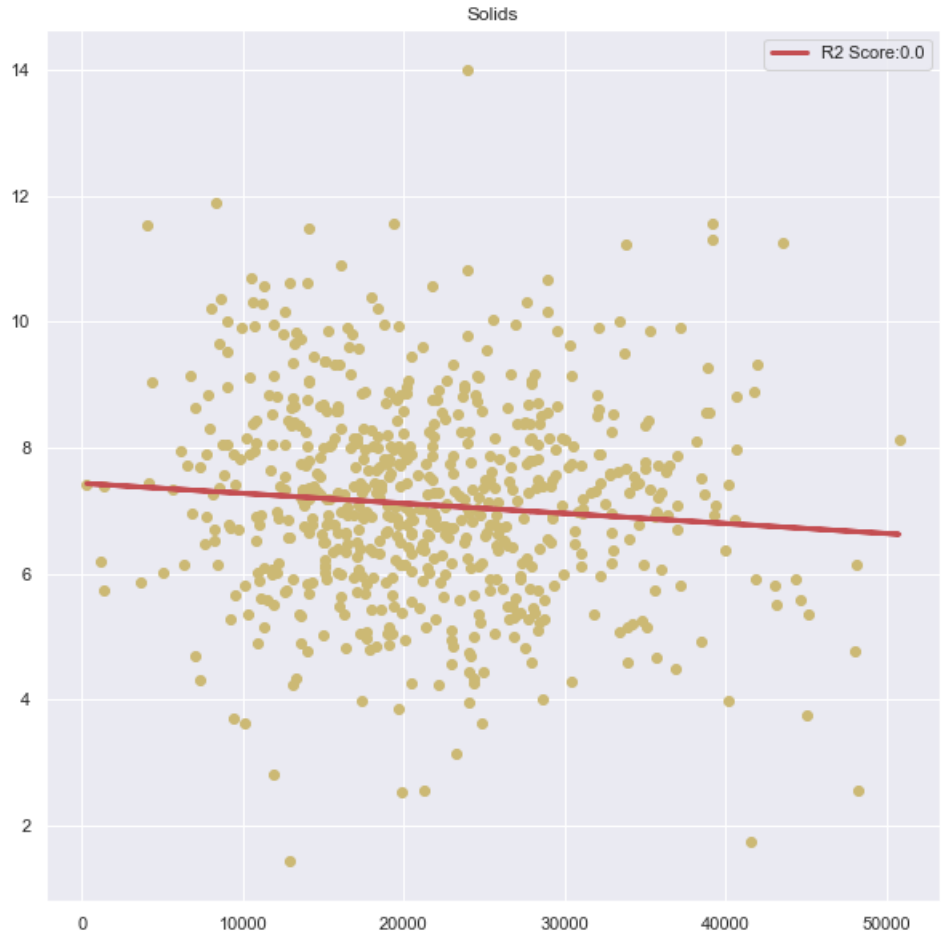


Figure 12: Solids regression

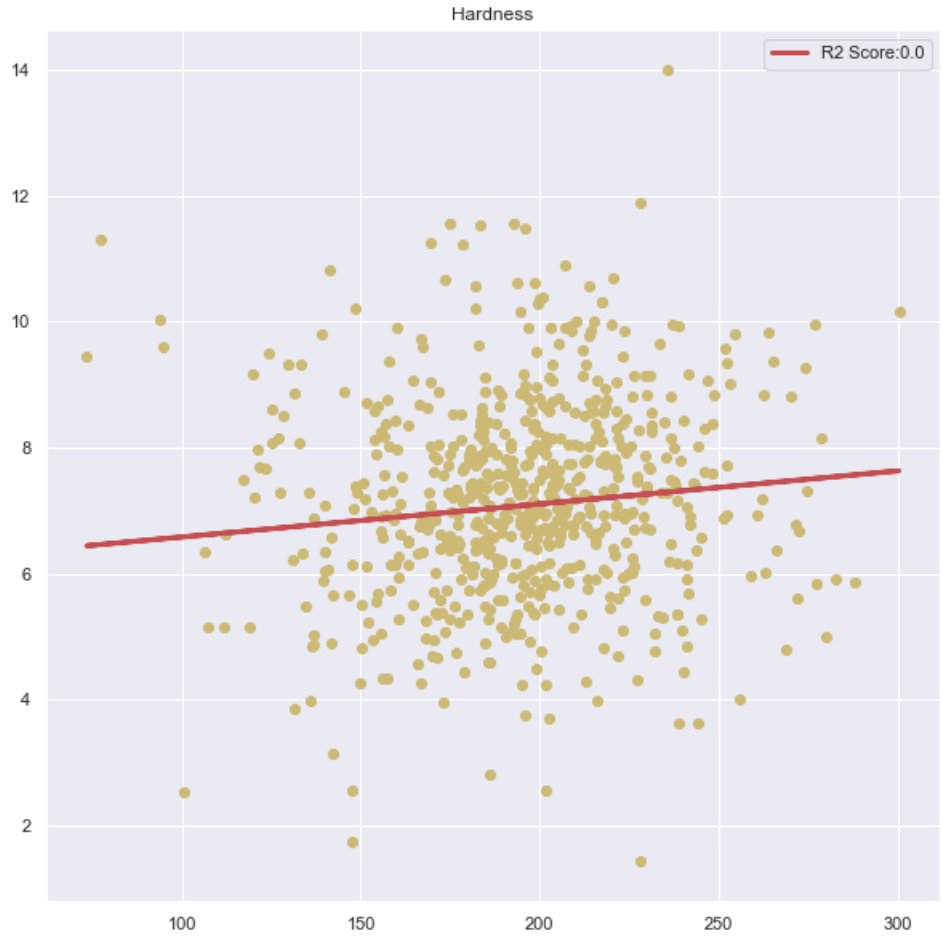


Figure 13: Hardness regression

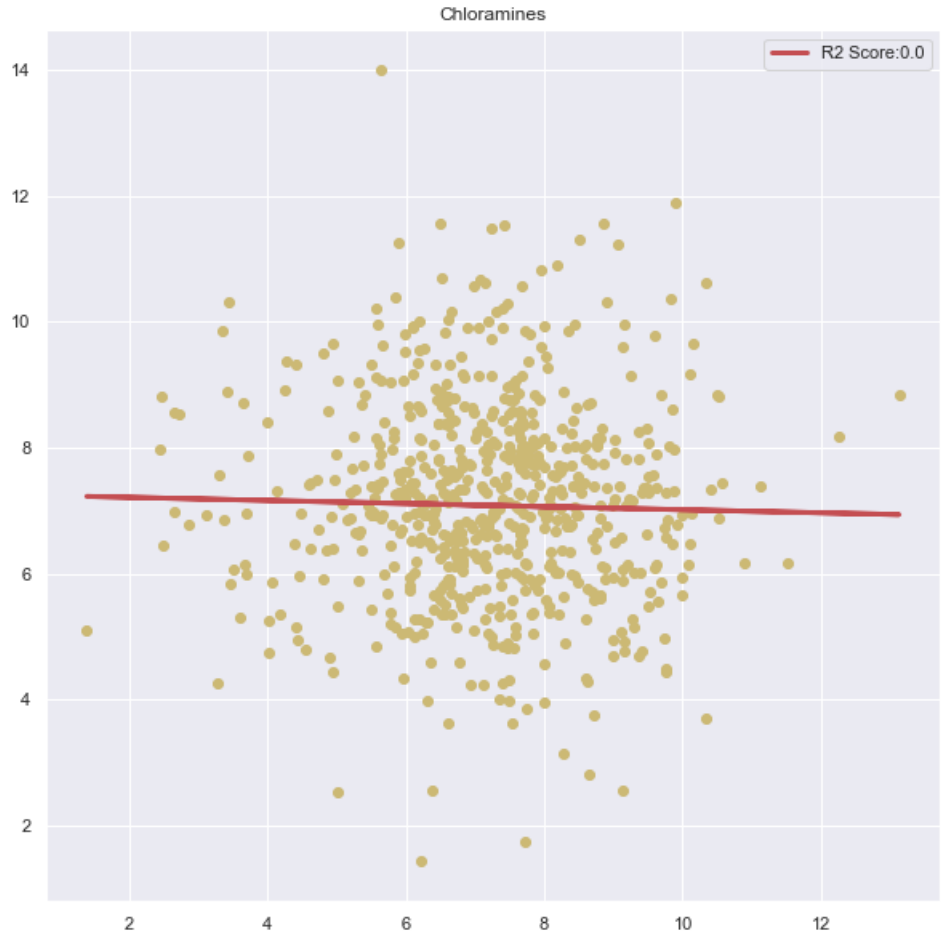


Figure 14: Chloramines regression

As we see on the plots, there is no linear dependency.

Because of low correlation between features and target, we will use all columns of dataset as a features for linear regression.

```
Mean Absolute Error: 1.25
Mean Squared Error: 2.62
R2 Score: 0.0
```

Figure 15: Linear regression

```
Mean Absolute Error: 1.25
Mean Squared Error: 2.62
R2 Score: 0.0
```

Figure 16: Lasso regularization

Obviously, when we try to fit linear regression by using all features it don't make much sense (due to low dependence of all features and target). R2 score for both linear and lasso equals 0.

6 Quality analysis

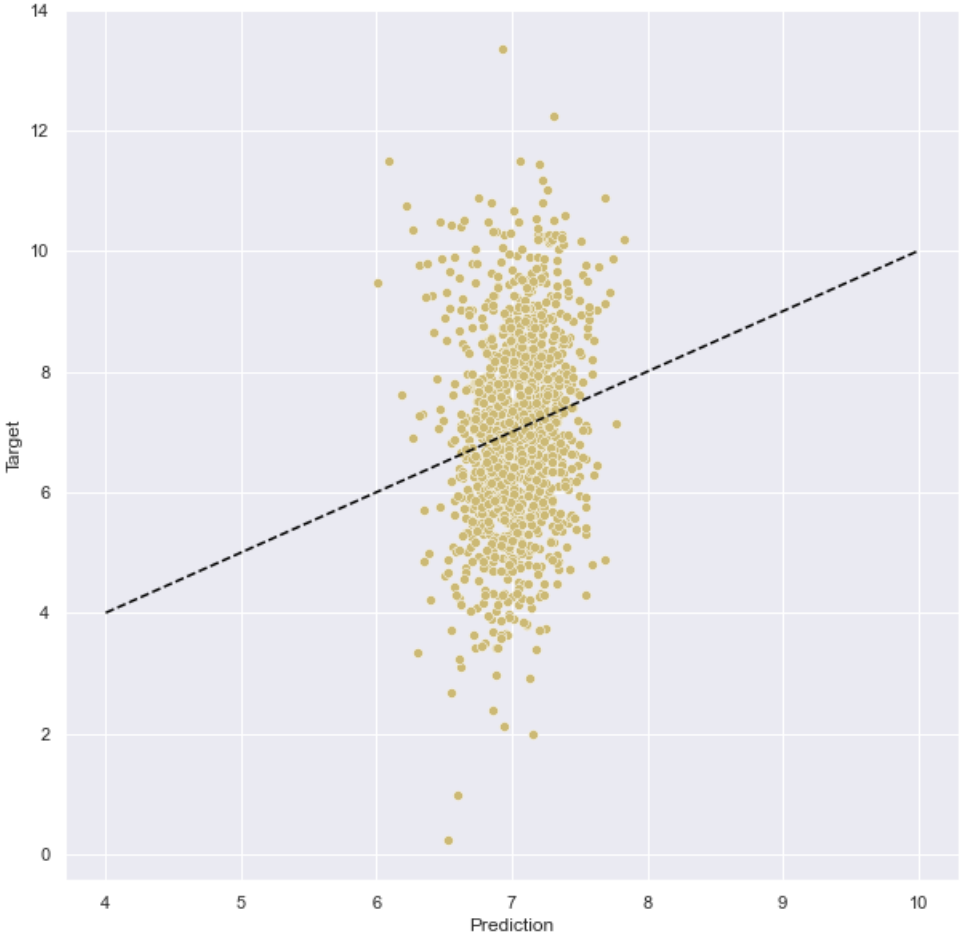
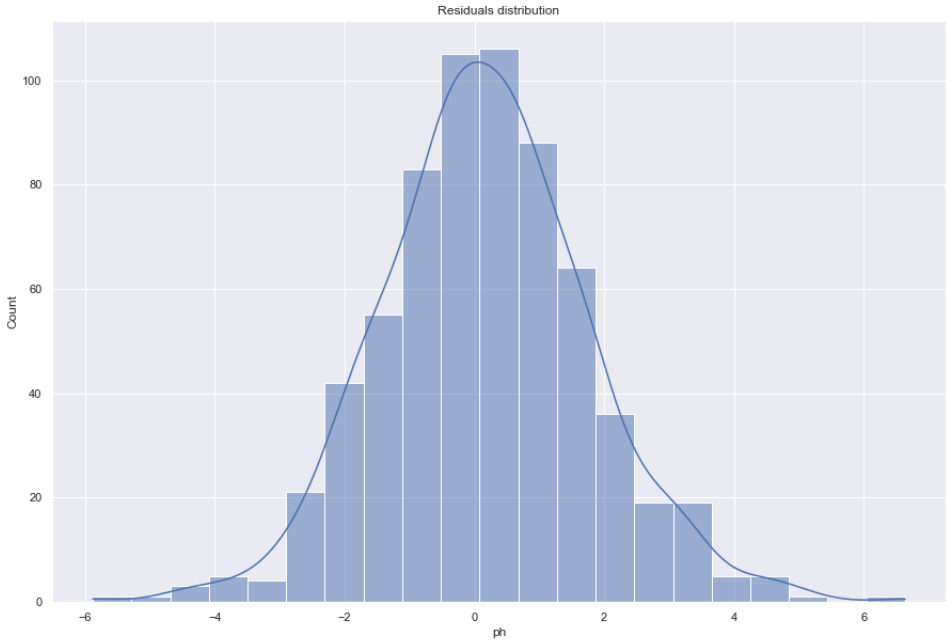
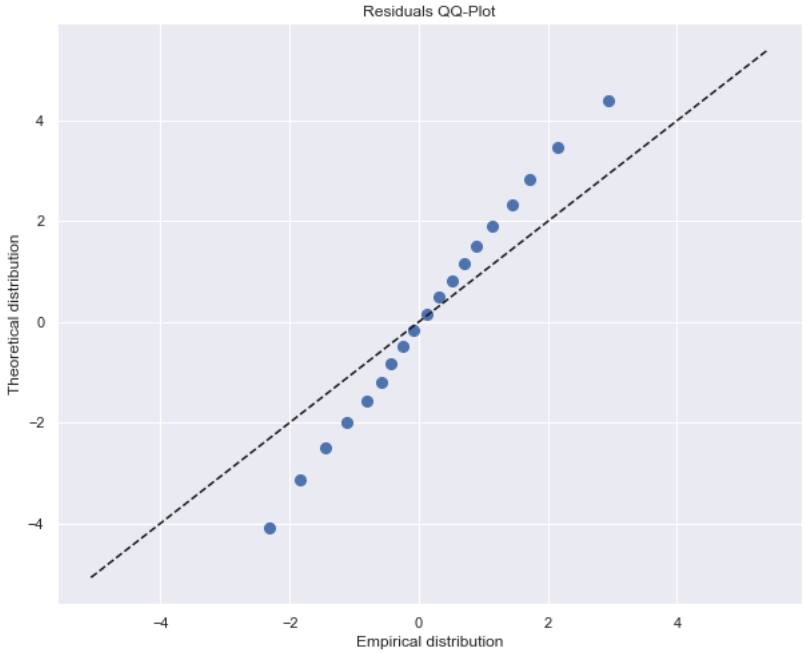


Figure 17: Predictions

Residuals analysis.





count	664.000000
mean	0.166694
std	1.605988
min	-5.880435
25%	-0.812040
50%	0.116388
75%	1.138783
max	6.633515

Figure 18: Residuals statistics

Residuals distribution is not normal according to the QQ plot, KS test, Cramér–von Mises criterion and Shapiro-Wilk test.

```
KstestResult(statistic=0.13362365876770566, pvalue=8.442031946689331e-11)
CramerVonMisesResult(statistic=5.163552840258241, pvalue=8.970213460912646e-11)
ShapiroResult(statistic=0.9943169355392456, pvalue=0.013853596523404121)
```

7 Conclusions

The nature of the data is such, that it is not possible to train an adequate linear regression on it. Residuals analysis shows that the residuals are distributed not normally, i.e. exactly as it was expected according to the regression model.

8 Source code

https://github.com/trixdade/ITMO-Methods-and-models-for-data-analysis/blob/master/Lab2/Lab2.ipynb