Tugas Besar IF2220 Probabilitas dan Statistik

Penarikan Kesimpulan dan Pengujian Hipotesis

Nomor 2

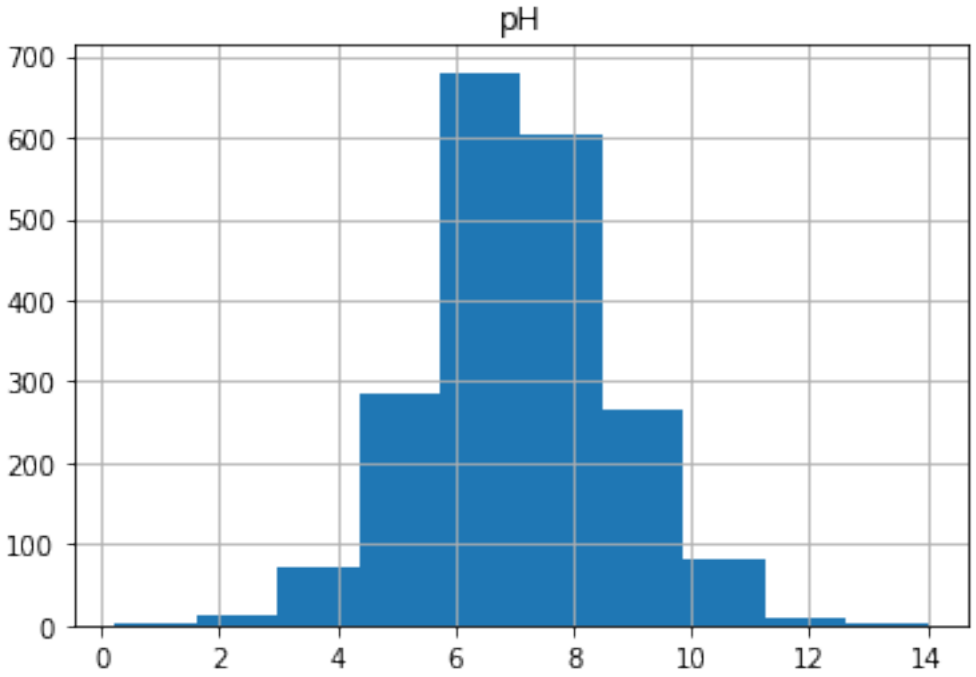
Membuat Visualisasi plot distribusi, dalam bentuk histogram dan boxplot untuk setiap kolom numerik. Berikan uraian penjelasan kondisi setiap kolom berdasarkan kedua plot tersebut.

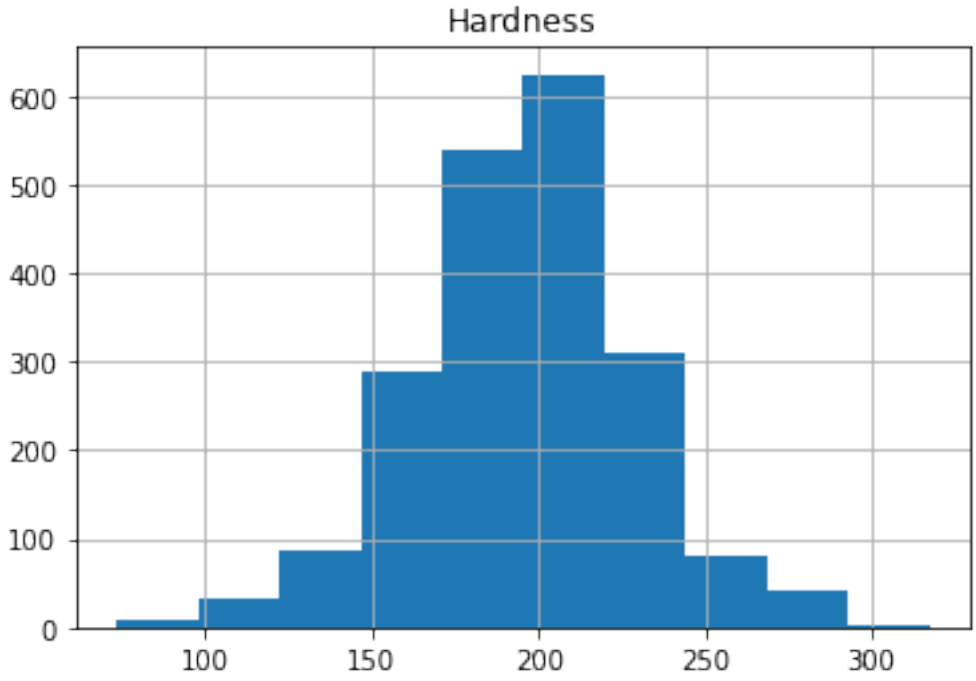
```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('water_potability.csv', index_col=0)
df.head()
```

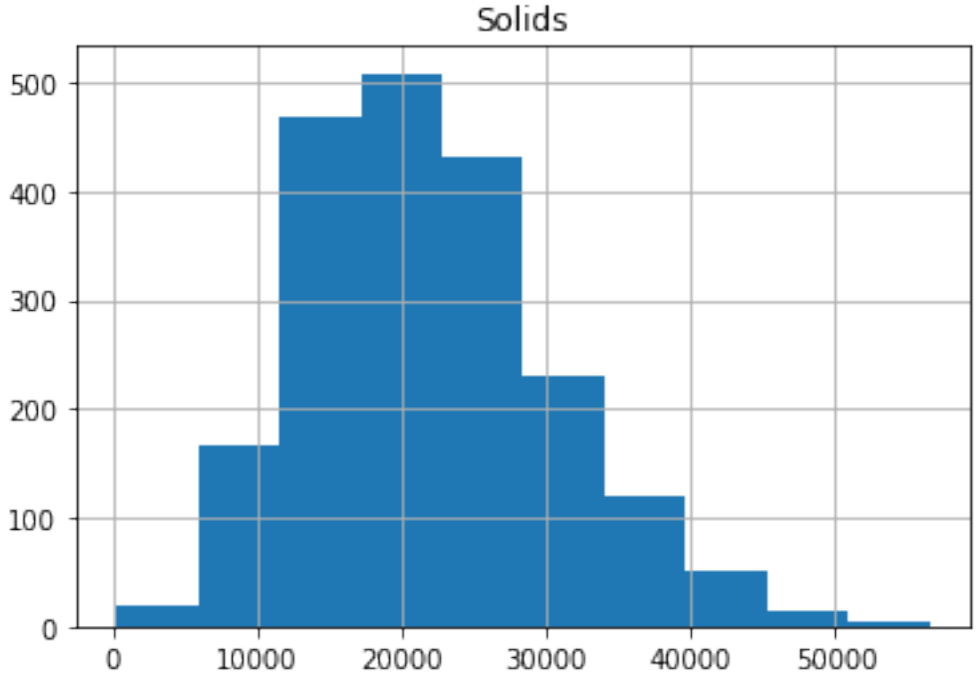
```
            pH    Hardness        Solids  Chloramines     Sulfate  \
id
1    8.316766  214.373394  22018.417441     8.059332  356.886136
2    9.092223  181.101509  17978.986339     6.546600  310.135738
3    5.584087  188.313324  28748.687739     7.544869  326.678363
4   10.223862  248.071735  28749.716544     7.513408  393.663396
5    8.635849  203.361523  13672.091764     4.563009  303.309771

    Conductivity  OrganicCarbon  Trihalomethanes  Turbidity
Potability
id

1     363.266516      18.436524       100.341674   4.628771
0
2     398.410813      11.558279        31.997993   4.075075
0
3     280.467916       8.399735        54.917862   2.559708
0
4     283.651634      13.789695        84.603556   2.672989
0
5     474.607645      12.363817        62.798309   4.401425
0
```

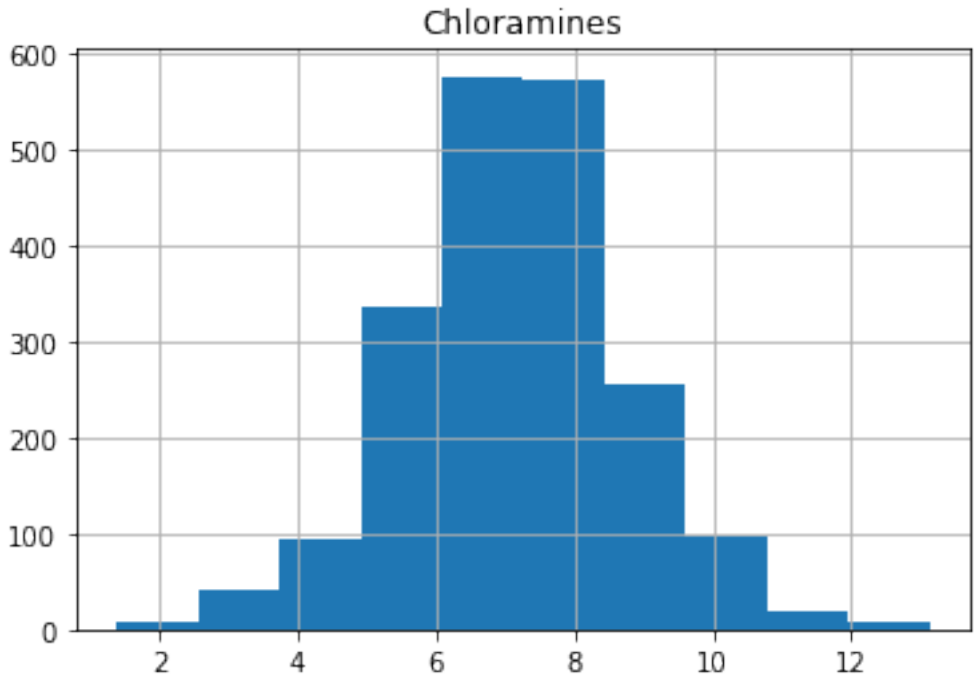
```
# Menyatakan kolom-kolom numerik

col_numeric = [
    'pH',
    'Hardness',
    'Solids',
    'Chloramines',
    'Sulfate',
```

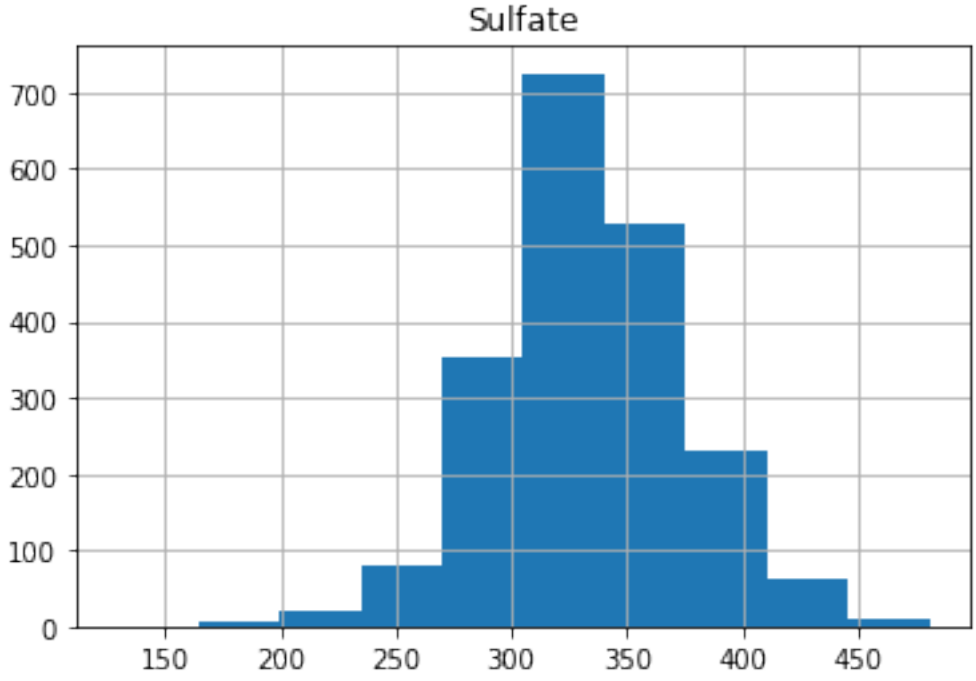
```
    'Conductivity',
    'OrganicCarbon',
    'Trihalomethanes',
    'Turbidity'
]

for col in col_numeric:
    try:
        df[col] = pd.to_numeric(df[col])
        df.hist(column=col)
    except ValueError:
        print('This column can not be represented as a histogram')
```

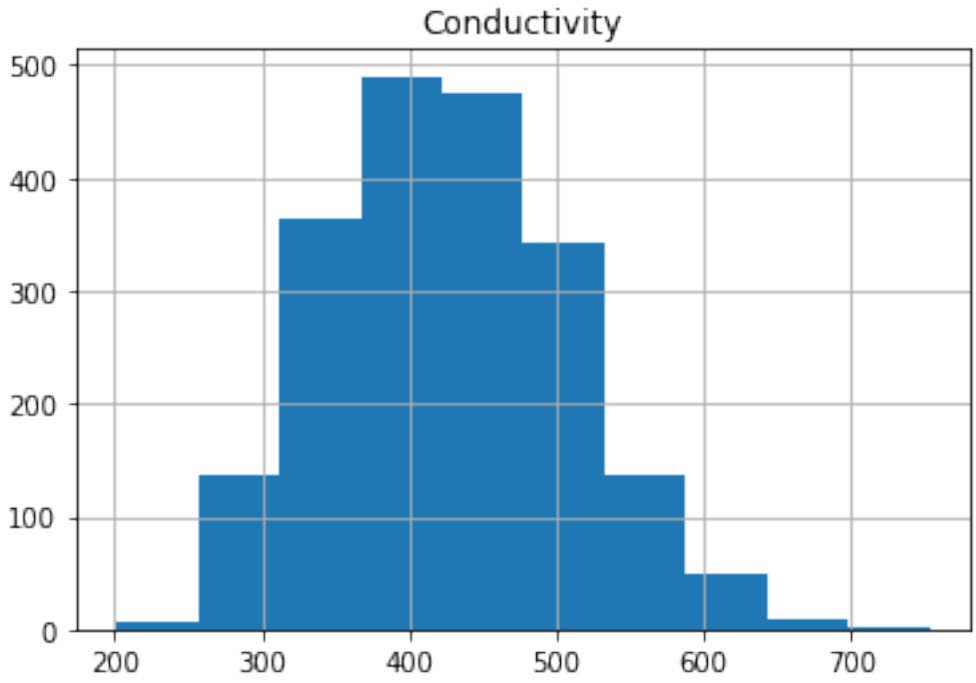


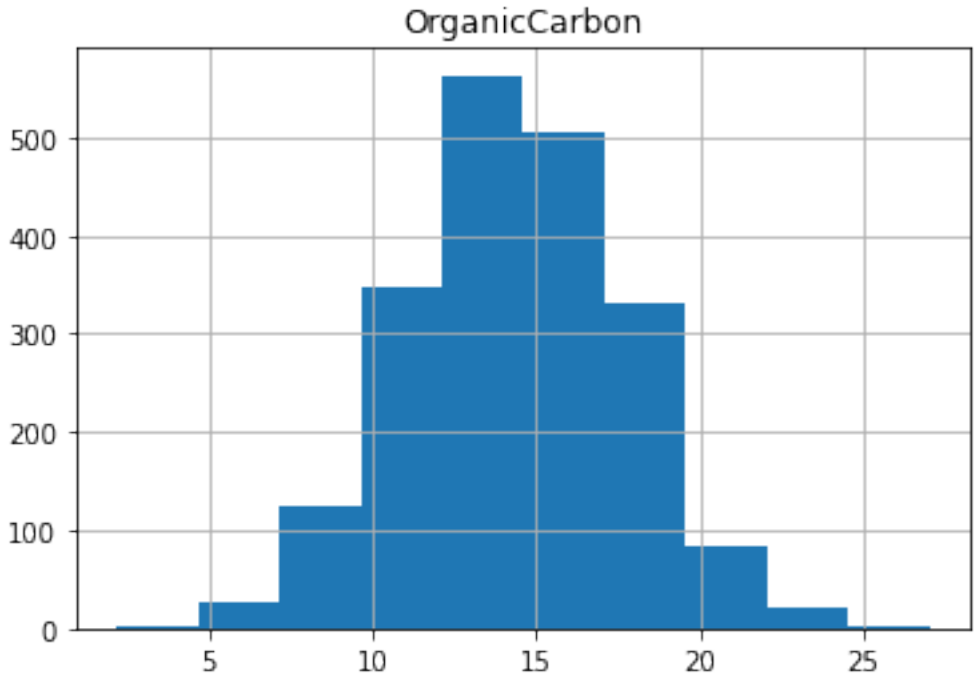

Hardness
600
500
400
300
200
100
0
100
150
200
250
300

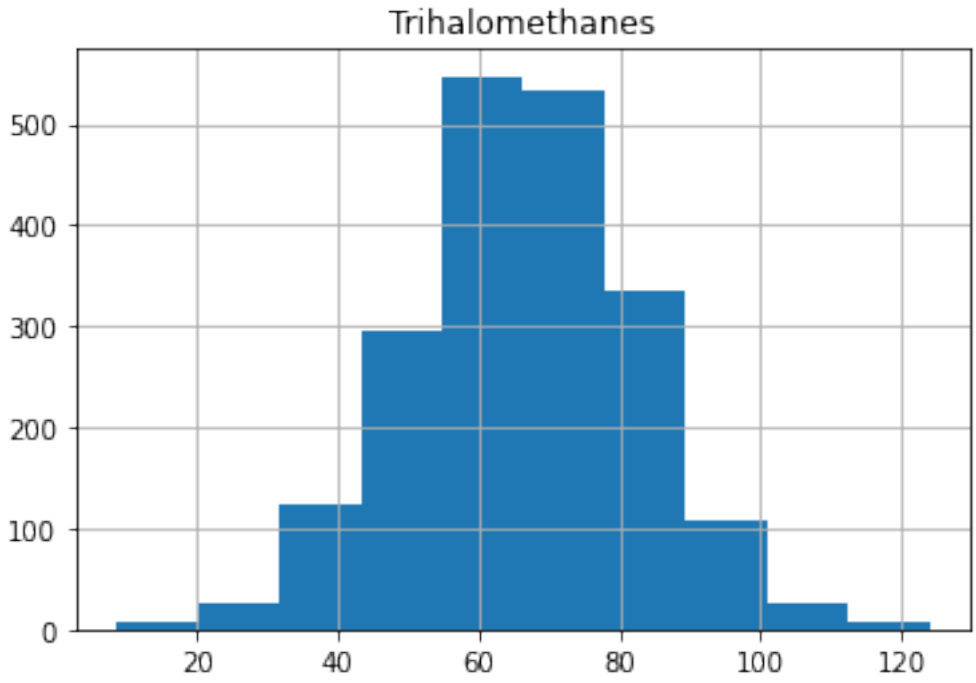

Solids
500
400
300
200
100
0
0
10000
20000
30000
40000
50000

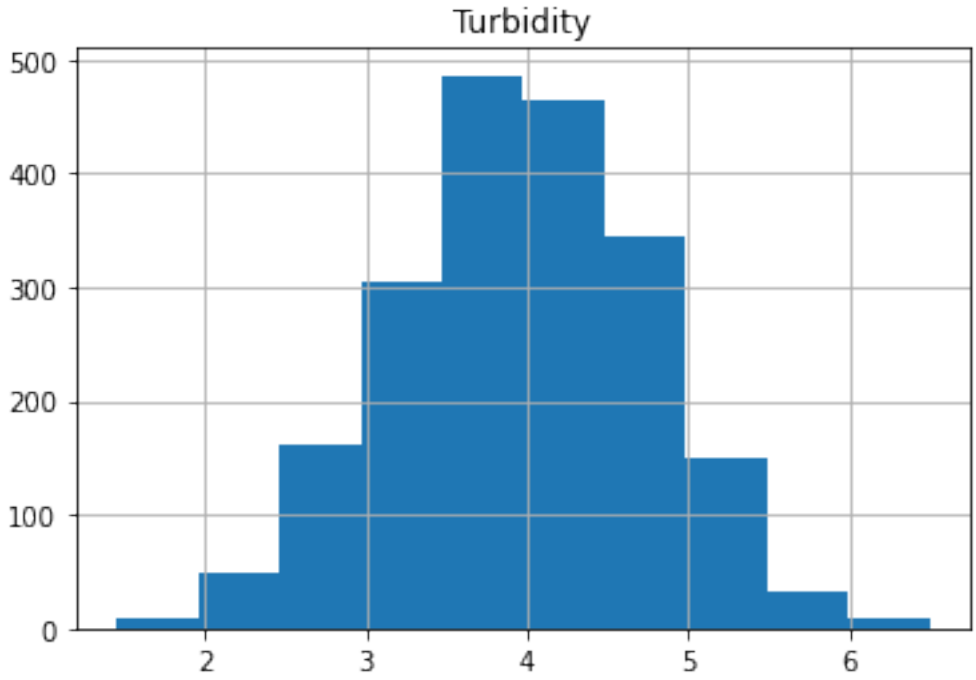

Chloramines
600
500
400
300
200
100
0
2
4
6
8
10
12


Sulfate
700
600
500
400
300
200
100
0
150
200
250
300
350
400
450

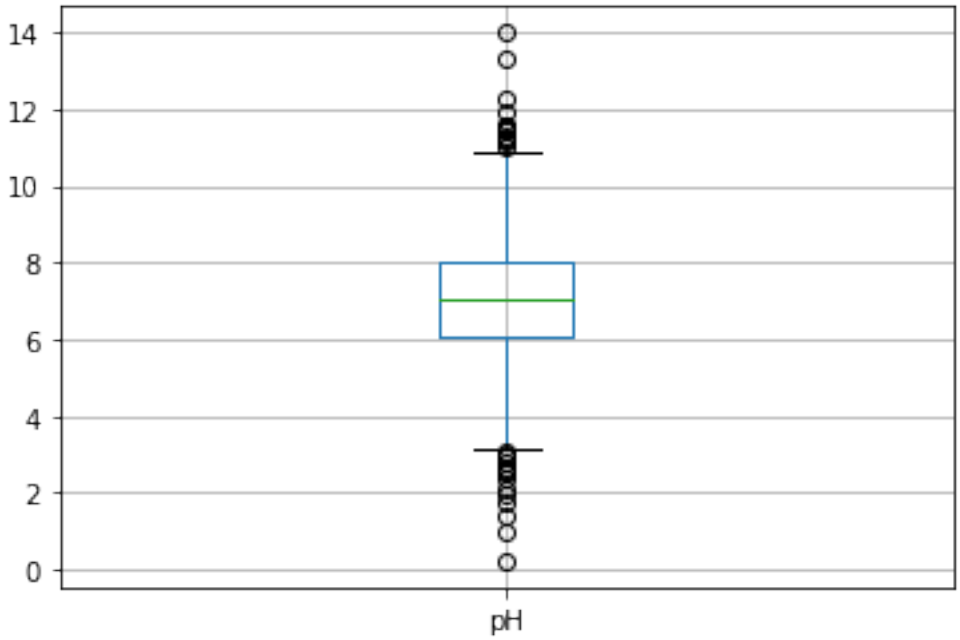

Conductivity
500
400
300
200
100
0
200
300
400
500
600
700

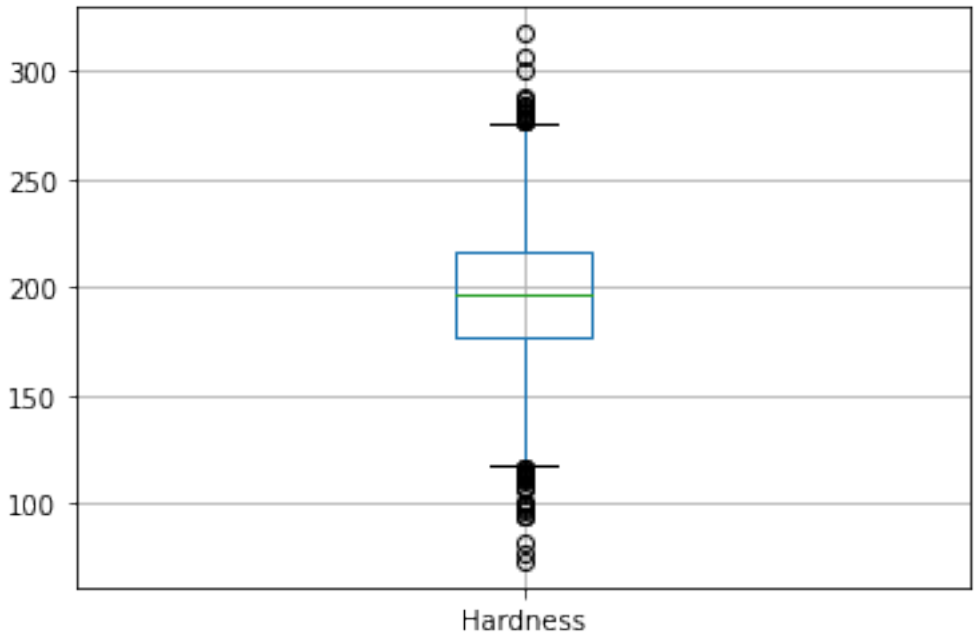

OrganicCarbon
500
400
300
200
100
0
5
10
15
20
25

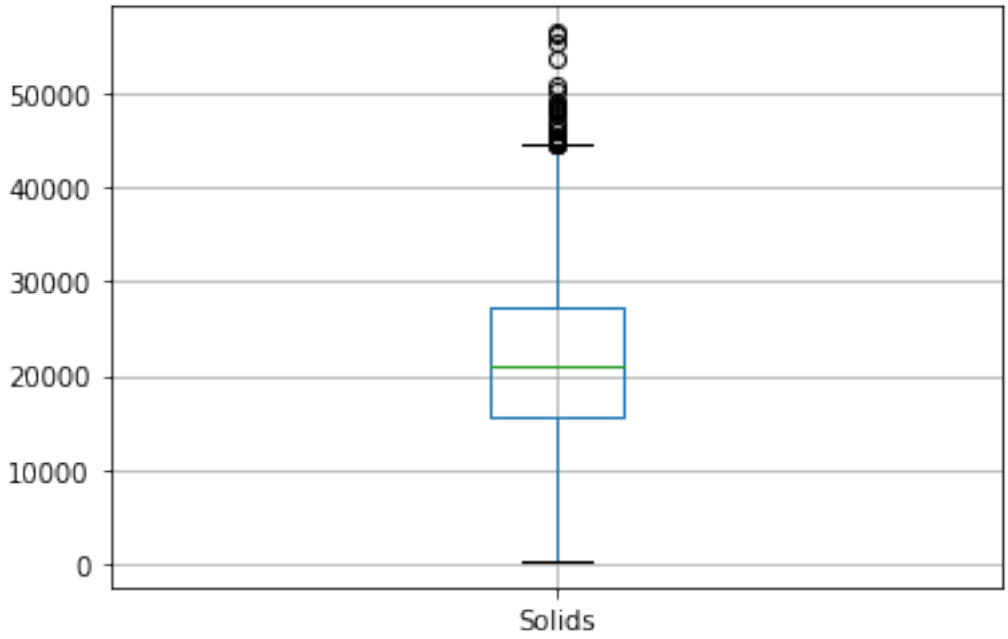


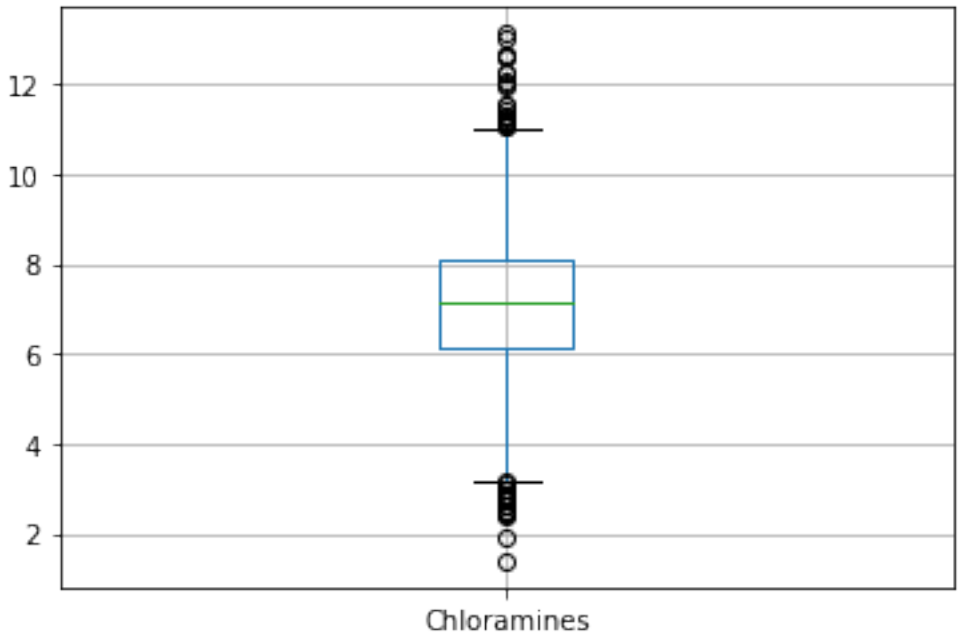


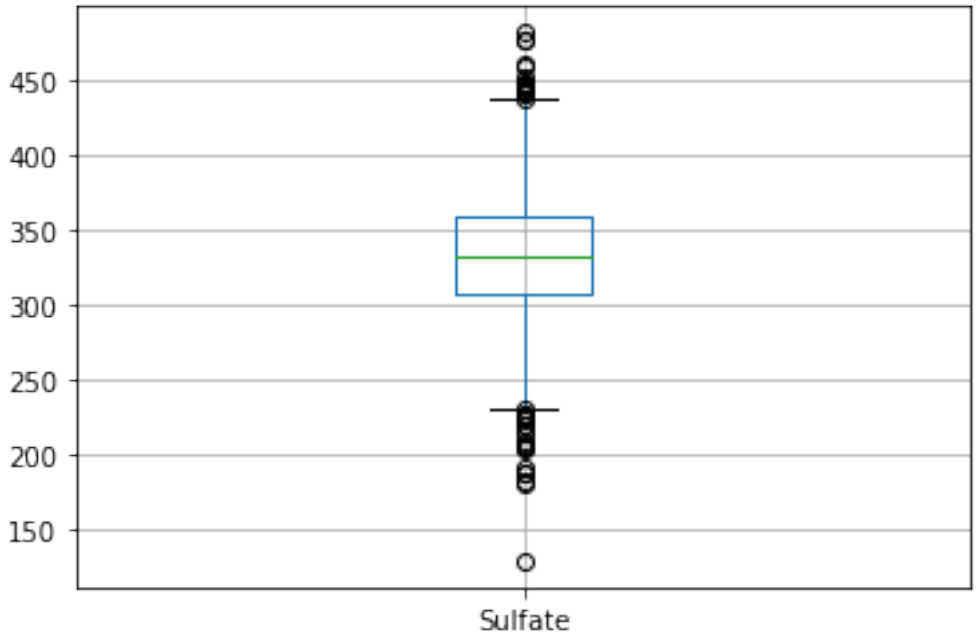

```
for col in col_numeric:
    plt.figure()
    df.boxplot([col])
```

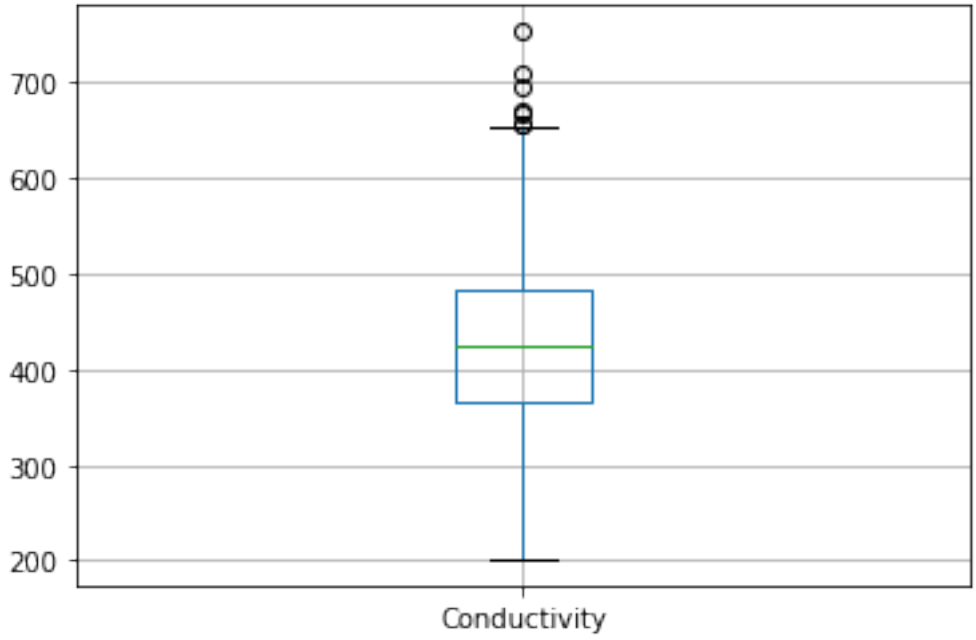

14
12
10
8
6
4
2
0
pH

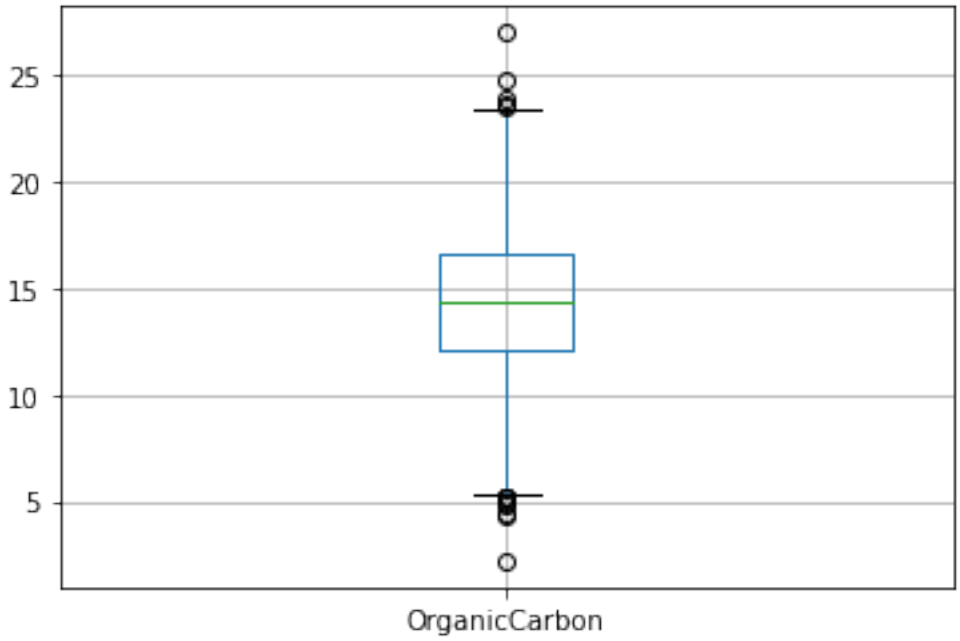

300
250
200
150
100
Hardness

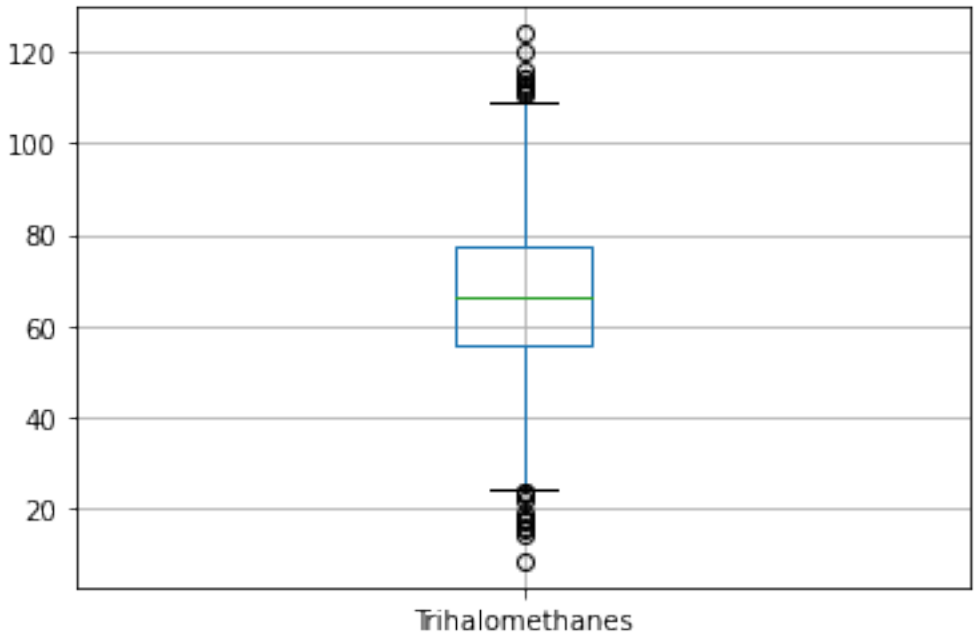

50000
40000
30000
20000
10000
0
Solids

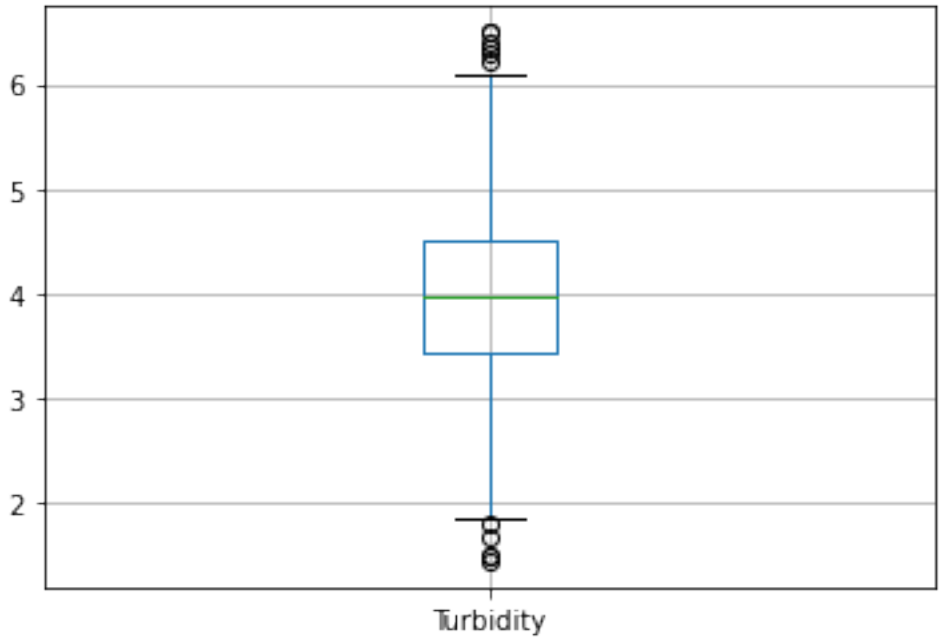

12
10
8
6
4
2
Chloramines


450
400
350
300
250
200
150
Sulfate


700
600
500
400
300
200
Conductivity


25
20
15
10
5
OrganicCarbon


120
100
80
60
40
20
Trihalomethanes


6
5
4
3
2
Turbidity