Tugas Besar IF2220 - Probabilitas dan Statistika

13520070 - Raden Haryosatyo Wisjnunandono

13520118 - Mohamad Daffa Argakoesoemah

In [328…

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as s
import seaborn as sns
from scipy.stats import iqr
import math

# Pembacaan file csv
df = pd.read_csv("water_potability.csv", names = ["id", "pH", "hardness", "solids", "chloramines",
                                                  "sulfate", "conductivity", "organicCarbon", "trihalomethanes",
                                                  "turbidity", "potability"])
df
```

Out[328]:

	id	pH	hardness	solids	chloramines	sulfate	conductivity	organicCarbon	trihalomethanes	turbidity	potability
0	1	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
1	2	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
2	3	5.584087	188.313324	28748.687739	7.544869	326.678363	280.467916	8.399735	54.917862	2.559708	0
3	4	10.223862	248.071735	28749.716544	7.513408	393.663396	283.651634	13.789695	84.603556	2.672989	0
4	5	8.635849	203.361523	13672.091764	4.563009	303.309771	474.607645	12.363817	62.798309	4.401425	0
...	...	...	...	...	...	...	...	...	...	...	...
2005	2006	8.197353	203.105091	27701.794055	6.472914	328.886838	444.612724	14.250875	62.906205	3.361833	1
2006	2007	8.989900	215.047358	15921.412018	6.297312	312.931022	390.410231	9.899115	55.069304	4.613843	1
2007	2008	6.702547	207.321086	17246.920347	7.708117	304.510230	329.266002	16.217303	28.878601	3.442983	1
2008	2009	11.491011	94.812545	37188.826022	9.263166	258.930600	439.893618	16.172755	41.558501	4.369264	1
2009	2010	6.069616	186.659040	26138.780191	7.747547	345.700257	415.886955	12.067620	60.419921	3.669712	1

2010 rows × 11 columns

Nomor 1

Menulis deskripsi statistika (Descriptive Statistics) dari semua kolom pada data yang bersifat numerik, terdiri dari mean, median, modus, standar deviasi, variansi, range, nilai minimum, maksimum, kuartil, IQR, skewness dan kurtosis. Boleh juga ditambahkan deskripsi lain.

In [329…

```
def desc_stat(df):
    print("Mean\t\t:", df.mean())
    print("Median\t\t:", df.median())
    print("Modus\t\t:", df.mode("index")[0])
    print("Standar Deviasi\t:", df.std())
    print("Variansi\t:", df.var())
    print("Range\t\t:", df.max()-df.min())
    print("Nilai Minimum\t:", df.min())
    print("Nilai Maksimum\t:", df.max())
    print("Kuartil pertama\t:", df.quantile(0.25))
    print("Kuartil kedua\t:", df.quantile(0.5))
    print("Kuartil ketiga\t:", df.quantile(0.75))
    print("IQR\t\t:", iqr(df))
    print("Skewness\t:", df.skew())
    print("Kurtosis\t:", df.kurtosis())


    mod = df.mode()
    print("Modus\t\t:", end="")
    if(len(mod) != 2010): #jika tidak semua modus
        for j in range(len(mod)):
            if j != len(mod) -1:
                print(mod[j], end= ", ")
            else:
                print(mod[j])
    else:
        print(" Semua data muncul sekali. ", end="") #semua data unik, semua modus
        print("Oleh karena itu dipilih modus, yaitu", df.mode()[0])
    print()
```

Kolom pH

In [330…

```
desc_stat(df["pH"])
```

```
Mean            : 7.0871927687138285
Median          : 7.029490455474185
Modus           : 0.2274990502021987
Standar Deviasi : 1.5728029470456655
Variansi        : 2.4737091102355304
Range           : 13.7725009497978
```

```
Nilai Minimum   : 0.2274990502021987
Nilai Maksimum  : 13.999999999999998
Kuartil pertama : 6.09078502142353
Kuartil kedua   : 7.029490455474185
Kuartil ketiga  : 8.053006240791538
IQR             : 1.9622212193680078
Skewness        : 0.04853451405270669
Kurtosis        : 0.6269041256617065
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 0.2274990502021987
```

Kolom hardness

In [331…

```
desc_stat(df["hardness"])
```

```
Mean            : 195.96920903783524
Median          : 197.20352491941043
Modus           : 73.4922336890611
Standar Deviasi : 32.643165859429864
Variansi        : 1065.5762773262472
Range           : 243.84589036652147
Nilai Minimum   : 73.4922336890611
Nilai Maksimum  : 317.33812405558257
Kuartil pertama : 176.74065667669896
Kuartil kedua   : 197.20352491941043
Kuartil ketiga  : 216.44758866727156
IQR             : 39.7069319905726
Skewness        : -0.08532104172868622
Kurtosis        : 0.5254804942991402
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 73.4922336890611
```

Kolom solids

In [332…

```
desc_stat(df["solids"])
```

```
Mean            : 21904.673439053095
Median          : 20926.88215534375
Modus           : 320.942611274359
Standar Deviasi : 8625.397911190576
Variansi        : 74397489.12637076
Range           : 56167.72980146483
Nilai Minimum   : 320.942611274359
Nilai Maksimum  : 56488.67241273919
Kuartil pertama : 15614.412961614333
Kuartil kedua   : 20926.88215534375
Kuartil ketiga  : 27170.534648603603
```

```
IQR             : 11556.12168698927
Skewness        : 0.5910113724580447
Kurtosis        : 0.33732026745944976
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 320.942611274359
```

Kolom chloramines

In [333…
```
desc_stat(df["chloramines"])
```

```
Mean            : 7.134322344600104
Median          : 7.1420143046226645
Modus           : 1.3908709048851806
Standar Deviasi : 1.5852140982642102
Variansi        : 2.512903737335613
Range           : 11.736129095114823
Nilai Minimum   : 1.3908709048851806
Nilai Maksimum  : 13.127000000000002
Kuartil pertama : 6.138326387572855
Kuartil kedua   : 7.1420143046226645
Kuartil ketiga  : 8.109933216133502
IQR             : 1.9716068285606472
Skewness        : 0.013003497779569528
Kurtosis        : 0.5497821097667472
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 1.3908709048851806
```

Kolom sulfate

In [334…
```
desc_stat(df["sulfate"])
```

```
Mean            : 333.211376415189
Median          : 332.2141128069568
Modus           : 129.00000000000003
Standar Deviasi : 41.21111102560979
Variansi        : 1698.355671965137
Range           : 352.03064230599716
Nilai Minimum   : 129.00000000000003
Nilai Maksimum  : 481.0306423059972
Kuartil pertama : 307.6269864860709
Kuartil kedua   : 332.2141128069568
Kuartil ketiga  : 359.26814739141554
IQR             : 51.641160905344634
Skewness        : -0.04572780443653543
Kurtosis        : 0.7868544988131605
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 129.00000000000003
```

Kolom conductivity

In [335…
```
desc_stat(df["conductivity"])
```

```
Mean            : 426.47670835257907
Median          : 423.43837202443706
Modus           : 201.6197367551575
Standar Deviasi : 80.70187180729437
Variansi        : 6512.792113200974
Range           : 551.7228828031471
Nilai Minimum   : 201.6197367551575
Nilai Maksimum  : 753.3426195583046
Kuartil pertama : 366.61921929632433
Kuartil kedua   : 423.43837202443706
Kuartil ketiga  : 482.2097724598859
IQR             : 115.5905531635616
Skewness        : 0.26801233302645316
Kurtosis        : -0.23720600574806516
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 201.6197367551575
```

Kolom organic carbon

In [336…
```
desc_stat(df["organicCarbon"])
```

```
Mean            : 14.357939902048074
Median          : 14.323285610653329
Modus           : 2.1999999999999886
Standar Deviasi : 3.3257700016987197
Variansi        : 11.0607461041991
Range           : 24.80670661116602
Nilai Minimum   : 2.1999999999999886
Nilai Maksimum  : 27.00670661116601
Kuartil pertama : 12.122530374047727
Kuartil kedua   : 14.323285610653329
Kuartil ketiga  : 16.683561746173808
IQR             : 4.561031372126081
Skewness        : -0.02021975629181238
Kurtosis        : 0.031018388192253
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 2.1999999999999886
```

Kolom trihalomethanes

In [337…
```
desc_stat(df["trihalomethanes"])
```

```
Mean            : 66.40071666307466
Median          : 66.48204080309809
Modus           : 8.577012932983806
Standar Deviasi : 16.08110898232513
Variansi        : 258.60206610141796
Range           : 115.4229870670162
Nilai Minimum   : 8.577012932983806
Nilai Maksimum  : 124.0
Kuartil pertama : 55.94999302803186
Kuartil kedua   : 66.48204080309809
Kuartil ketiga  : 77.2946128060674
IQR             : 21.344619778035543
Skewness        : -0.05138268451619478
Kurtosis        : 0.2230167810639787
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 8.577012932983806
```

Kolom turbidity

In [338…
```
desc_stat(df["turbidity"])
```

```
Mean            : 3.9694969126303676
Median          : 3.967373963531836
Modus           : 1.45
Standar Deviasi : 0.7804710407083957
Variansi        : 0.6091350453844462
Range           : 5.044748555990993
Nilai Minimum   : 1.45
Nilai Maksimum  : 6.494748555990993
Kuartil pertama : 3.442881623557439
Kuartil kedua   : 3.967373963531836
Kuartil ketiga  : 4.5146627202018825
IQR             : 1.0717810966444437
Skewness        : -0.03226597968019271
Kurtosis        : -0.049830796949249745
Modus           : Semua data muncul sekali. Oleh karena itu dipilih modus, yaitu 1.45
```

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.

In [339…
```
from IPython.display import Markdown, display
```

```
def markdown(input):
    display(Markdown(input))

def printHistBox(col):
    markdown("### Boxplot dan Histogram %s" %(col))
    plt.subplot(1,2,1)
    df[col].plot(kind="hist",figsize=(10,4), color = "green")
    plt.xlabel("Value")
    plt.title('Histogram')
    plt.grid()

    plt.subplot(1,2,2)
    df[col].plot(kind = "box", vert=False)
    plt.title('Boxplot')
    plt.yticks(ticks=[0])
    plt.xlabel("Value")
    plt.grid()
    plt.tight_layout()

def fetch(col):
    skew = df[col].skew()
    kurt = df[col].kurtosis()
    q1 = float(df[col].quantile(0.25))
    median = df[col].quantile(0.50)
    q3 = float(df[col].quantile(0.75))
    mean = df[col].mean()
    iqrange =  iqr(df[col])
    uppperTail = 1.5 * iqrange + q3
    lowerTail = q1 - (1.5 * iqrange)

    Boxplot = "Pada boxplot dapat dilihat bahwa nilai upper tail adalah %f dan nilai lower tail adalah %f." %(uppperTail, lower

    markdown(Boxplot)

    pAtas = ""
    pBawah = ""
    if (df[col].max() > uppperTail):
        pAtas = " Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum da
    if (df[col].min() < lowerTail):
        pBawah = " Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum d

    markdown(pAtas + pBawah)
    text1 = "Jika dilihat pada histogram, diketahui bahwa skewness kolom %s adalah " %(col)

    #skewness
    text2 = "$%f$. Oleh karena itu dapat disimpulkan bahwa kolom bersifat %s " %(skew, col)
```

```
    skewnessCol = "Dari histogram kita bisa lihat bahwa kolom %s bersifat" %(col)
    if (-0.5 < skew < 0.5):
        ket = " ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$"
    elif ((skew <= -0.5) and (mean < median)):
        ket = " ***negatively skewed***, yaitu data dengan nilai yang besar memiliki frekuensi yang tinggi."
    elif ((skew >= 0.5) and (mean > median)):
        ket = " ***positively skewed***, yaitu data dengan nilai yang kecil memiliki frekuensi yang tinggi."

    markdown(text1 + text2 + skewnessCol + ket)

    if (kurt < 0):
        text5 = "Terlihat bahwa kolom tersebut memiliki  ***platykurtic distribution*** karena terdapat banyak niali ekstrem bav
    elif (kurt > 0):
        text5 = "Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas
    else:
        text5 = "Terlihat bahwa kolom tersebut memiliki  ***mesokurtic distribution***. "


    markdown(text5)

def printAll(col):
    printHistBox(col)
    fetch(col)
```

In [340…

```
printAll('pH')
```

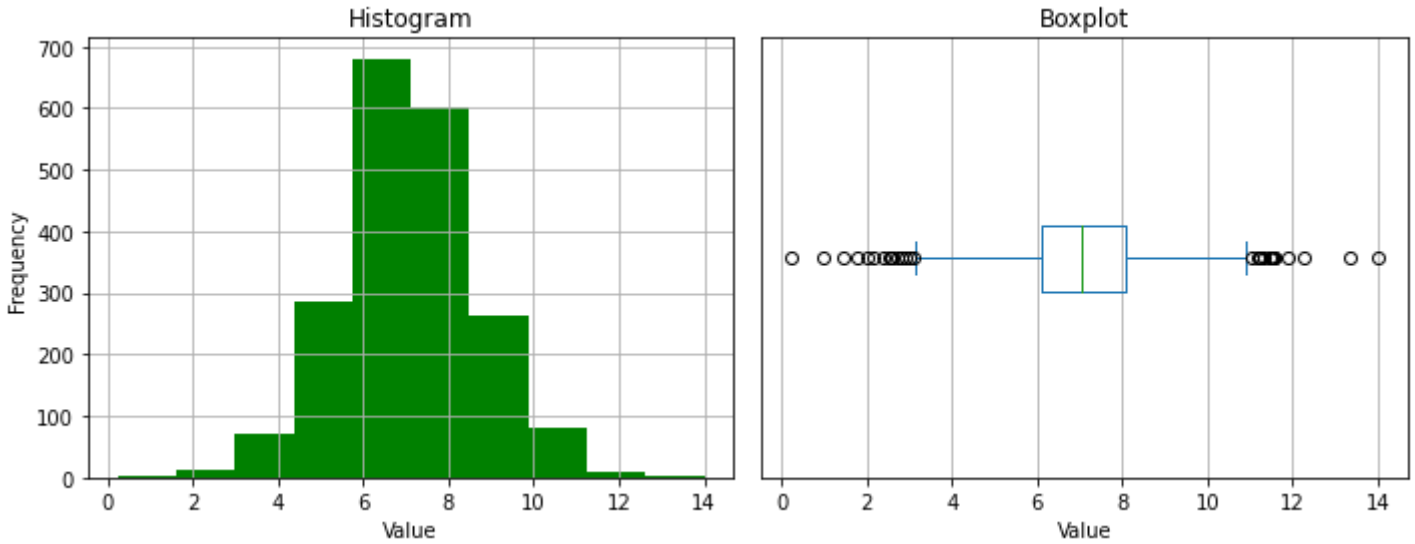
Boxplot dan Histogram pH

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 10.996338 dan nilai lower tail adalah 3.147453.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari pH > upper tail. Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum dari pH < lower tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom pH adalah 0.048535. Oleh karena itu dapat disimpulkan bahwa kolom bersifat pH Dari histogram kita bisa lihat bahwa kolom pH bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas ($kurt > 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai 0.626904.



In [341…

```
printAll('hardness')
```

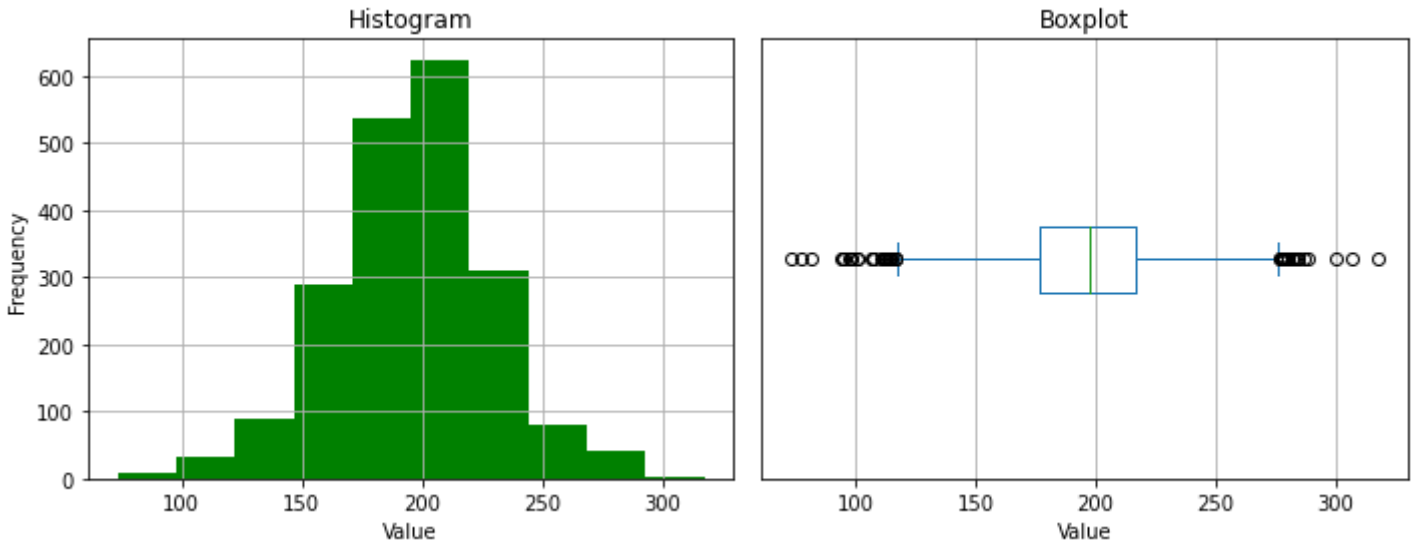
Boxplot dan Histogram hardness

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 276.007987 dan nilai lower tail adalah 117.180259.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari hardness > upper tail. Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum dari hardness < lower tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom hardness adalah -0.085321. Oleh karena itu dapat disimpulkan bahwa kolom bersifat hardness Dari histogram kita bisa lihat bahwa kolom hardness bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas ($kurt > 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai 0.525480.



In [342…
```
printAll('solids')
```

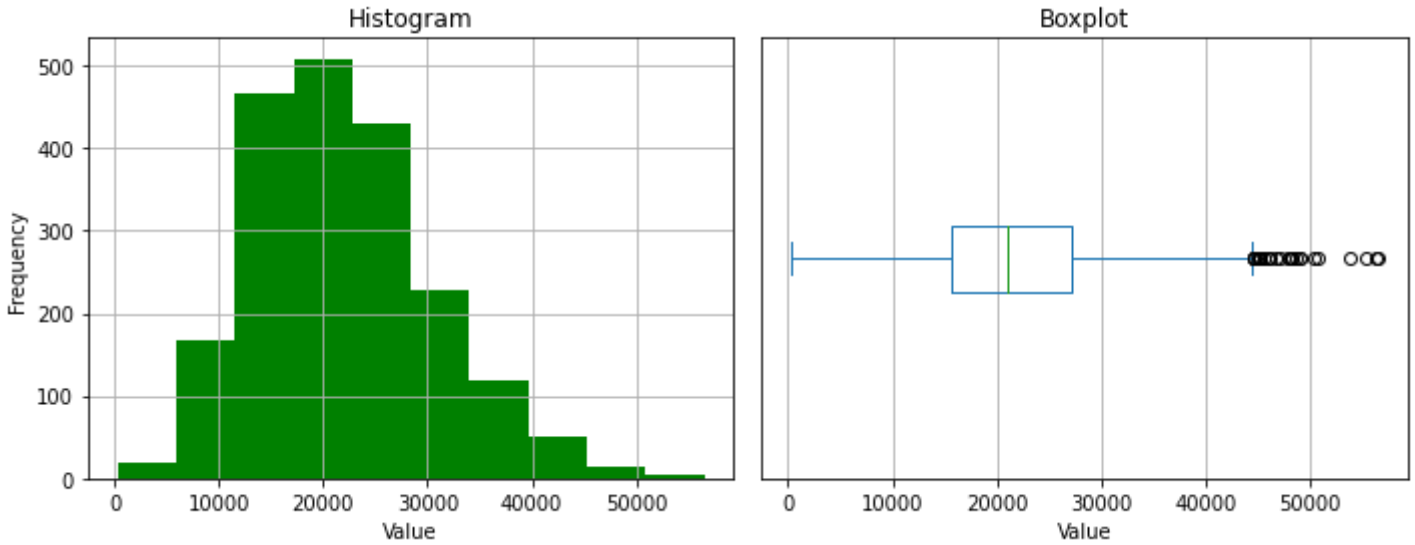
Boxplot dan Histogram solids

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 44504.717179 dan nilai lower tail adalah -1719.769569.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari solids > upper tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom solids adalah 0.591011. Oleh karena itu dapat disimpulkan bahwa kolom bersifat solids Dari histogram kita bisa lihat bahwa kolom solids bersifat ***positively skewed***, yaitu data dengan nilai yang kecil memiliki frekuensi yang tinggi.

Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas ($kurt > 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai 0.337320.



In [343…

```
printAll('chloramines')
```

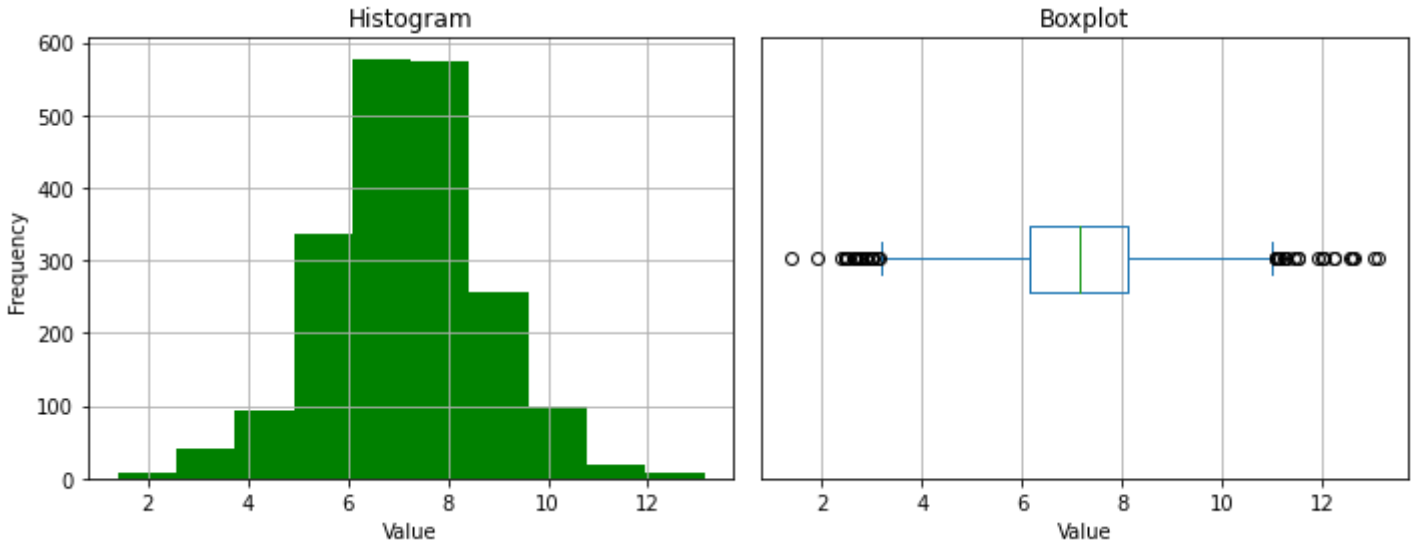
Boxplot dan Histogram chloramines

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 11.067343 dan nilai lower tail adalah 3.180916.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari chloramines > upper tail. Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum dari chloramines < lower tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom chloramines adalah 0.013003. Oleh karena itu dapat disimpulkan bahwa kolom bersifat chloramines Dari histogram kita bisa lihat bahwa kolom chloramines bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas ($kurt > 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai 0.549782.



In [344…

```
printAll('sulfate')
```

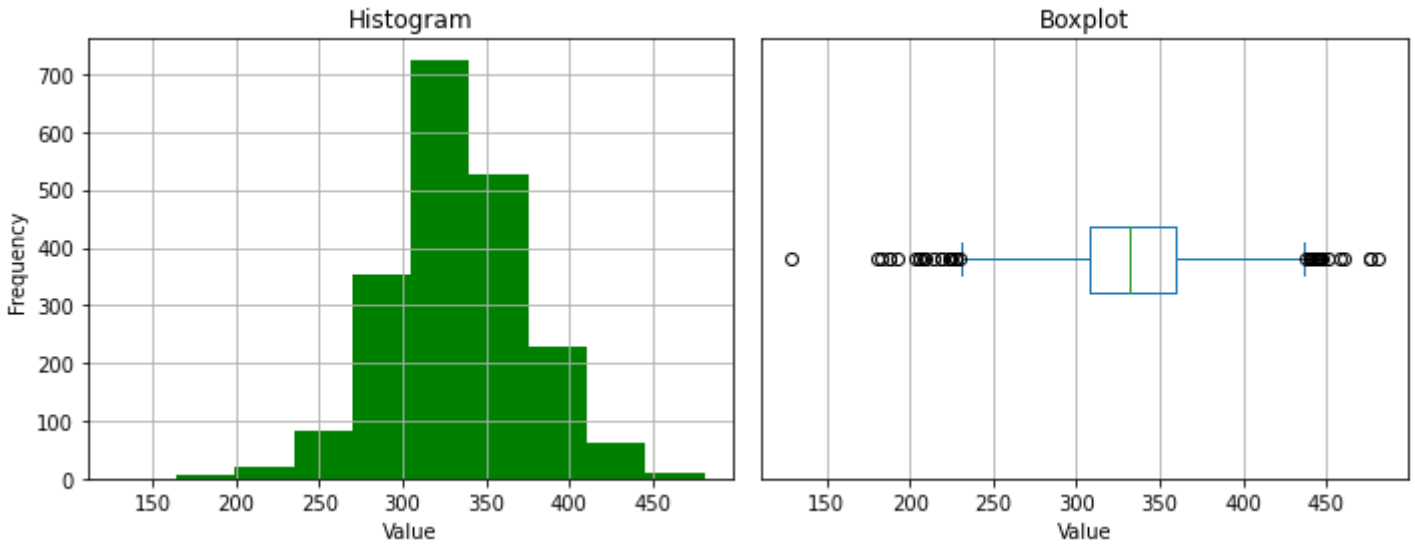
Boxplot dan Histogram sulfate

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 436.729889 dan nilai lower tail adalah 230.165245.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari sulfate > upper tail. Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum dari sulfate < lower tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom sulfate adalah -0.045728. Oleh karena itu dapat disimpulkan bahwa kolom bersifat sulfate Dari histogram kita bisa lihat bahwa kolom sulfate bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas ($kurt > 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai 0.786854.



In [345…

```
printAll('conductivity')
```

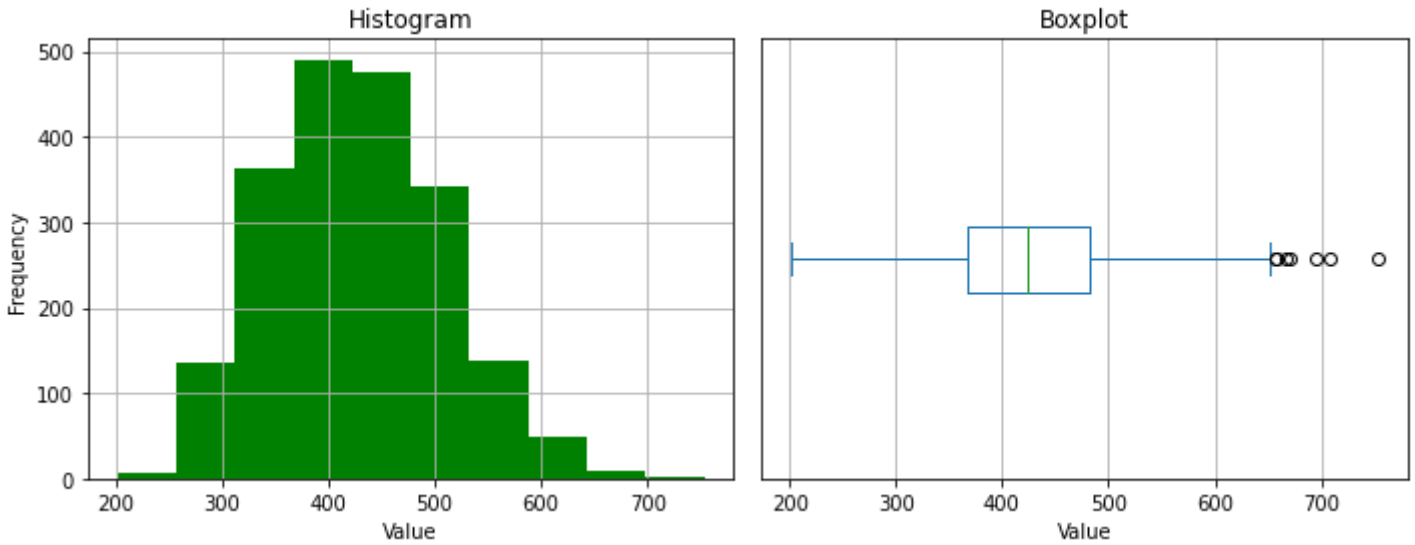
Boxplot dan Histogram conductivity

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 655.595602 dan nilai lower tail adalah 193.233390.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari conductivity > upper tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom conductivity adalah 0.268012. Oleh karena itu dapat disimpulkan bahwa kolom bersifat conductivity Dari histogram kita bisa lihat bahwa kolom conductivity bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***platykurtic distribution*** karena terdapat banyak niali ekstrem bawah ($kurt < 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai -0.237206.



In [346…

```
printAll('organicCarbon')
```

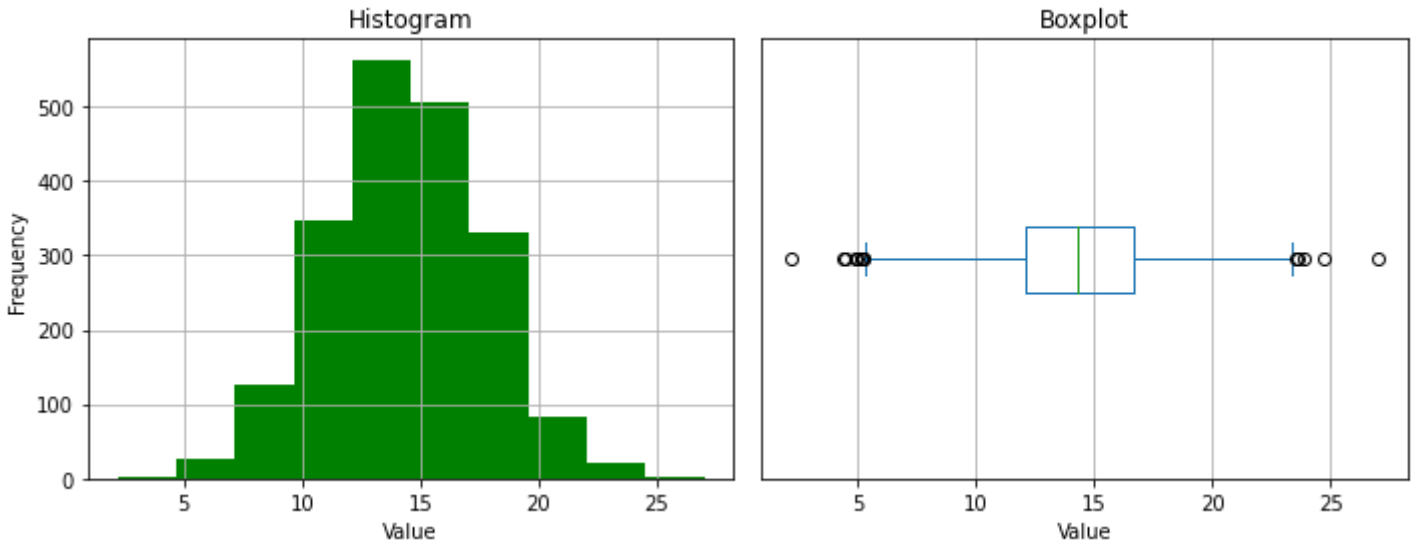
Boxplot dan Histogram organicCarbon

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 23.525109 dan nilai lower tail adalah 5.280983.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari organicCarbon > upper tail. Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum dari organicCarbon < lower tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom organicCarbon adalah -0.020220. Oleh karena itu dapat disimpulkan bahwa kolom bersifat organicCarbon Dari histogram kita bisa lihat bahwa kolom organicCarbon bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas ($kurt > 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai 0.031018.



In [347…

```
printAll('trihalomethanes')
```

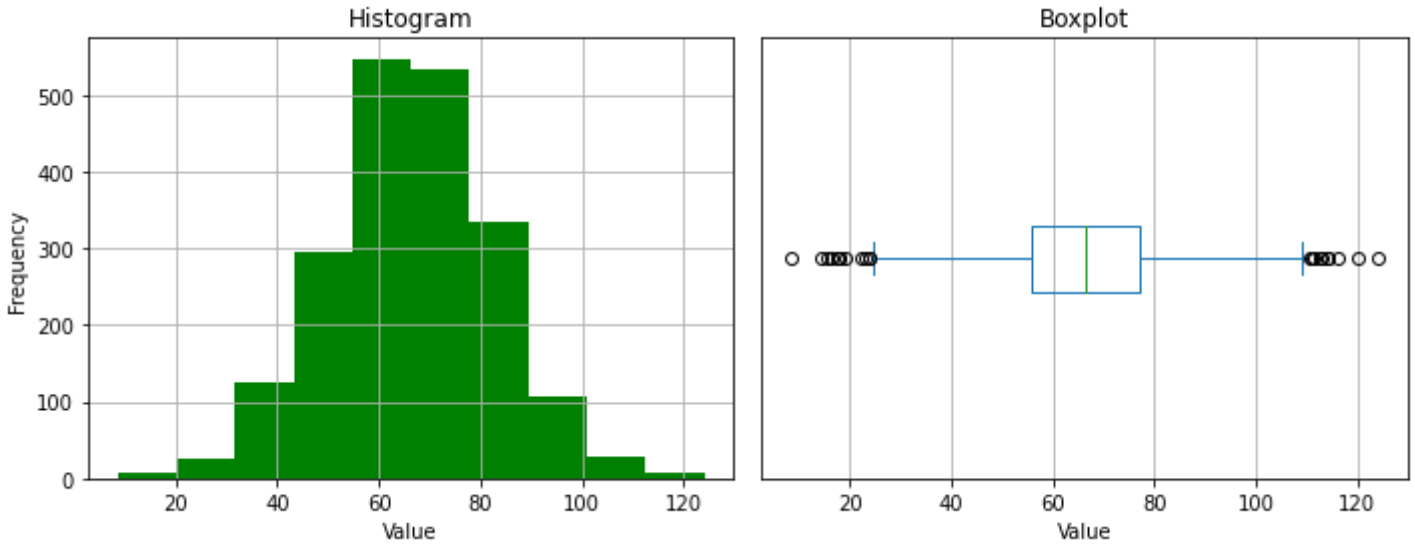
Boxplot dan Histogram trihalomethanes

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 109.311542 dan nilai lower tail adalah 23.933063.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari trihalomethanes > upper tail. Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum dari trihalomethanes < lower tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom trihalomethanes adalah -0.051383. Oleh karena itu dapat disimpulkan bahwa kolom bersifat trihalomethanes Dari histogram kita bisa lihat bahwa kolom trihalomethanes bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***leptokurtic distribution*** karena terdapat banyak nilai ekstrem atas ($kurt > 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai 0.223017.



In [348…

```
printAll('turbidity')
```

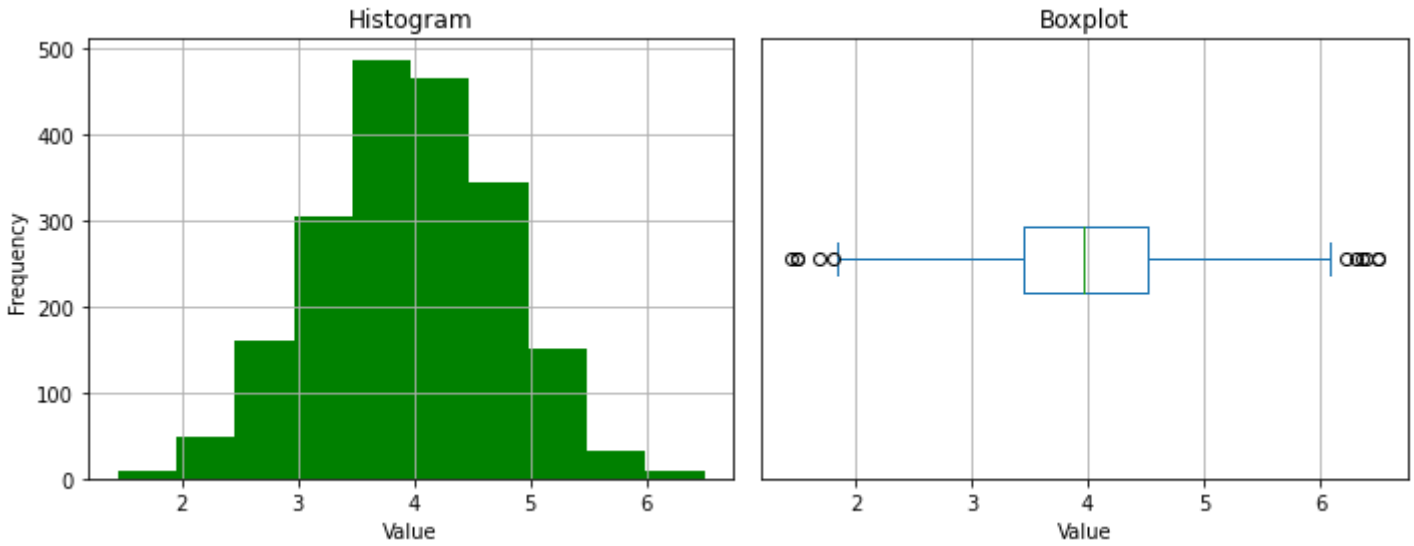
Boxplot dan Histogram turbidity

Pada boxplot dapat dilihat bahwa nilai upper tail adalah 6.122334 dan nilai lower tail adalah 1.835210.

Pada boxplot dapat terlihat ada pencilan atas. Hal ini juga didukung dari perhitungan dimana nilai maksimum dari turbidity > upper tail. Pada boxplot dapat terlihat ada pencilan bawah. Hal ini juga didukung dari perhitungan dimana nilai minimum dari turbidity < lower tail.

Jika dilihat pada histogram, diketahui bahwa skewness kolom turbidity adalah -0.032266. Oleh karena itu dapat disimpulkan bahwa kolom bersifat turbidity Dari histogram kita bisa lihat bahwa kolom turbidity bersifat ***symmetrically distributed***. Hal ini juga didukung oleh data karena $-0.5 < skew < 0.5$

Terlihat bahwa kolom tersebut memiliki ***platykurtic distribution*** karena terdapat banyak niali ekstrem bawah ($kurt < 0$). Hal tersebut juga dibuktikan dengan nilai kurtosis kolom yang bernilai -0.049831.



Nomor 3

Menentukan setiap kolom numerik berdistribusi normal atau tidak. Gunakan normality test yang dikaitkan dengan histogram plot!

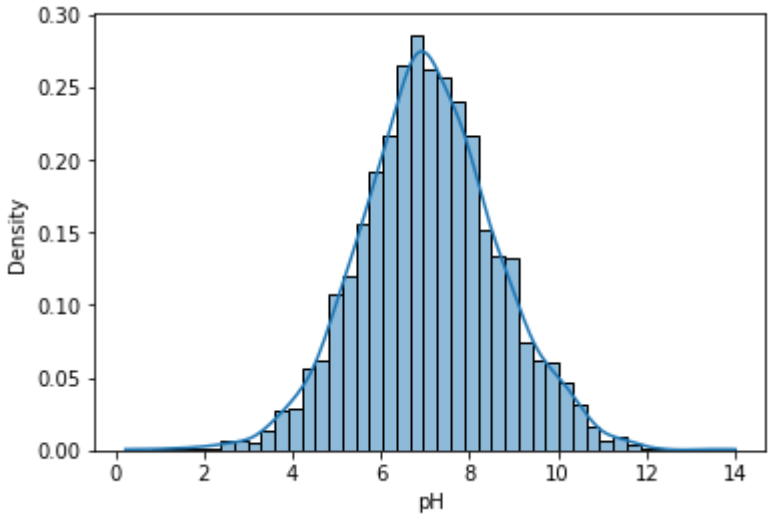
Pada nomor ini akan dilakukan pengujian menggunakan fungsi normaltest dari library scipy yang berdasarkan D'Agostino-Pearson Test. Nilai alpha yang digunakan, yaitu sebesar 0.05. Jika p-value yang didapatkan kurang dari alpha, kolom tersebut tidak berdistribusi normal. Selain itu akan ditampilkan plot histogram untuk dilakukan pengamatan secara visual.

In [349…
```
def normalityTest(column):
    sns.histplot(df[column], kde=True, stat="density", linewidth=1)

    alpha = 0.05
    print("Nilai alpha sebesar " + str(alpha))
    k2, pVal = s.normaltest(df[column])
    print("P-value yang didapatkan sebesar " + str(pVal))
```

Kolom pH

In [350…
```
normalityTest("pH")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 2.6514813346797777e-05
```



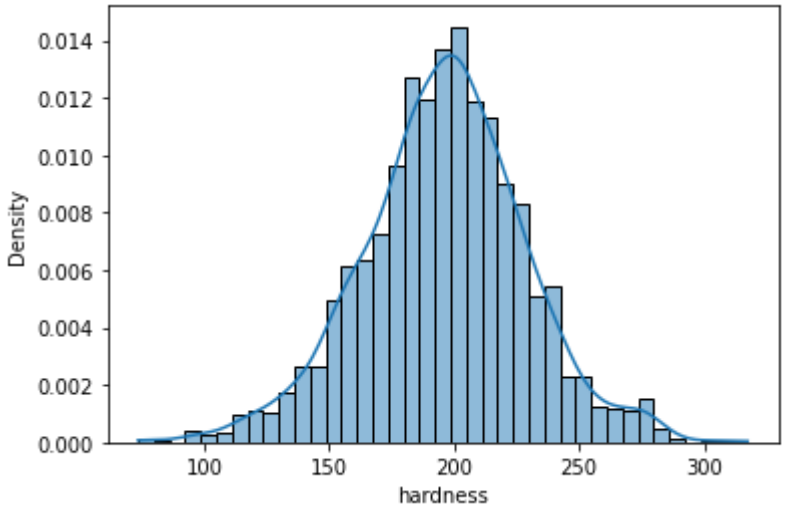
Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 2.6514813346797777e-05 yang kurang dari nilai alpha. Oleh karena itu, kolom pH tidak berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat kurang simetris.

Kolom hardness

In [351…

```
normalityTest("hardness")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 0.00013442428699593753
```



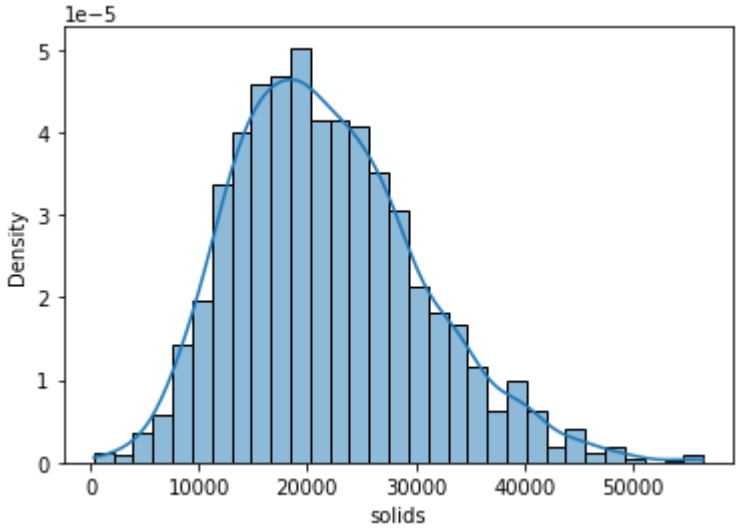
Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 0.00013442428699593753 yang kurang dari nilai alpha. Oleh karena itu, kolom hardness tidak berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat tidak membentuk bell curve. Bagian kanan kurva tidak mendekati

sumbu datar secara asimtotik.

Kolom solids

In [352…
```
normalityTest("solids")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 2.0796613688739523e-24
```

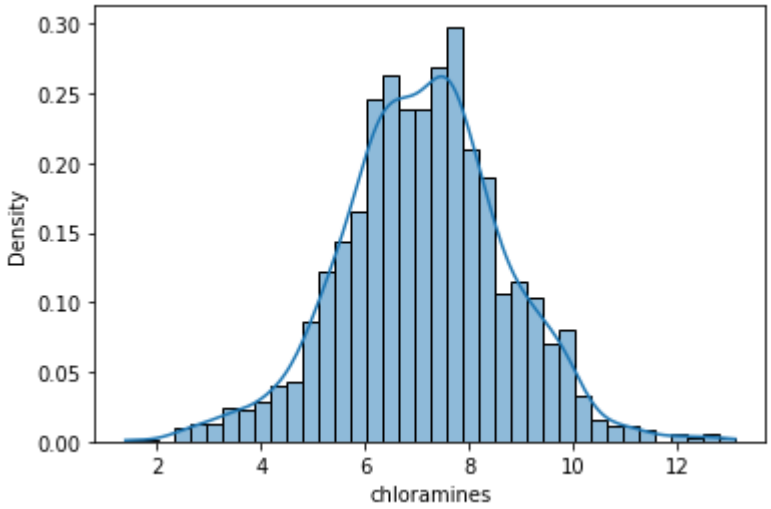


Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 2.0796613688739523e-24 yang kurang dari nilai alpha. Oleh karena itu, kolom solids tidak berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat tidak simetri terhadap sumbu tegak. Kurva terlihat positively skewed.

Kolom chloramines

In [353…
```
normalityTest("chloramines")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 0.0002504831654753917
```



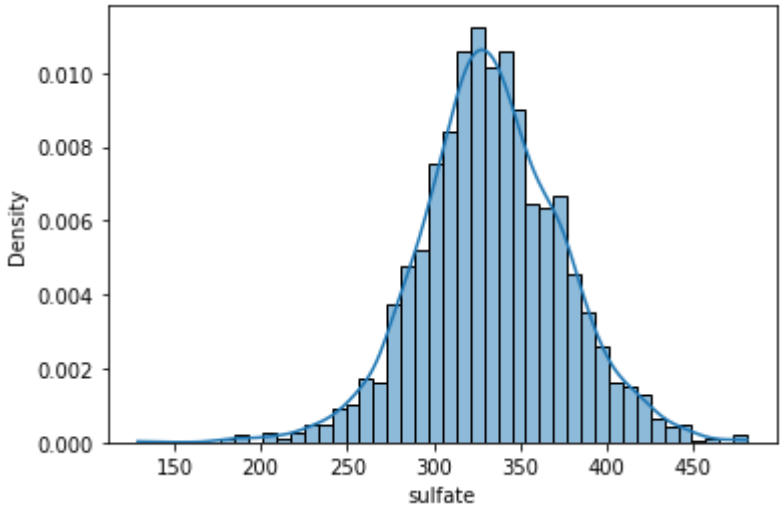
Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 0.0002504831654753917 yang kurang dari nilai alpha. Oleh karena itu, kolom chloramines tidak berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat tidak membentuk bell curve. Bagian kanan dan kiri kurva tidak mendekati sumbu datar secara asimtotik.

Kolom sulfate

In [354…

```
normalityTest("sulfate")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 4.4255936678013136e-07
```



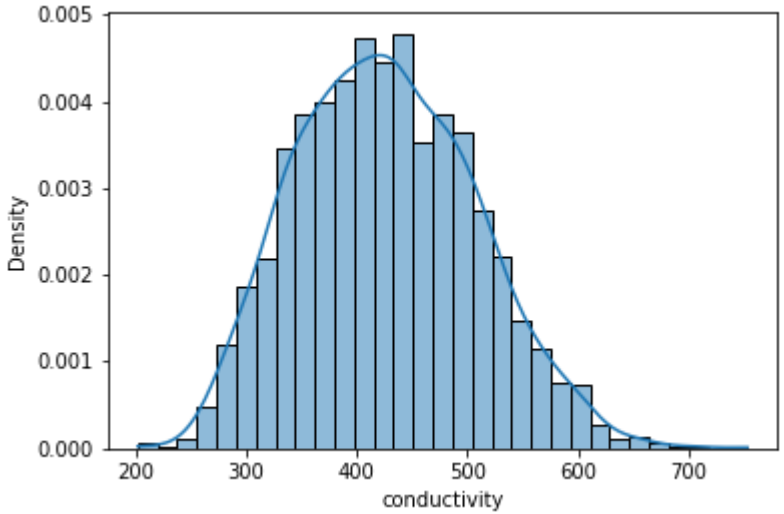
Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 4.4255936678013136e-07 yang kurang dari nilai alpha. Oleh karena itu, kolom sulfate tidak

berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat tidak simetris terhadap sumbu tegak.

Kolom conducitivity

In [355…
```
normalityTest("conductivity")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 4.39018078287845e-07
```

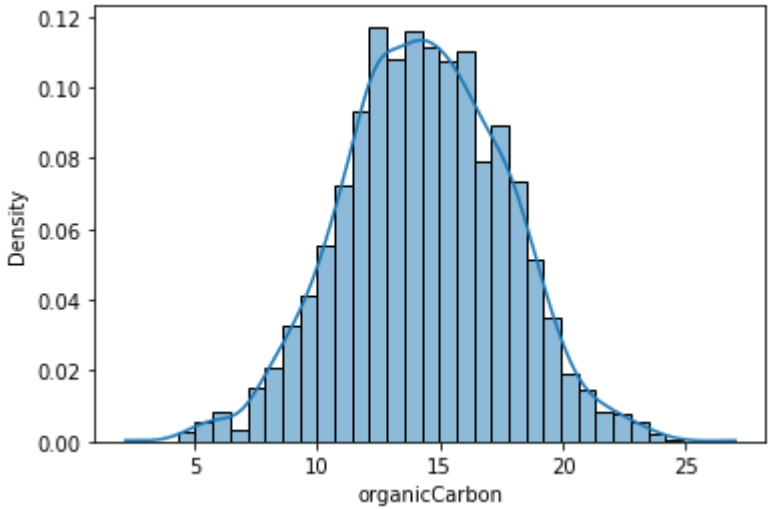


Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 4.39018078287845e-07 yang kurang dari nilai alpha. Oleh karena itu, kolom conductivity tidak berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat tidak simetris terhadap sumbu tegak.

Kolom organic carbon

In [356…
```
normalityTest("organicCarbon")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 0.8825496581408284
```



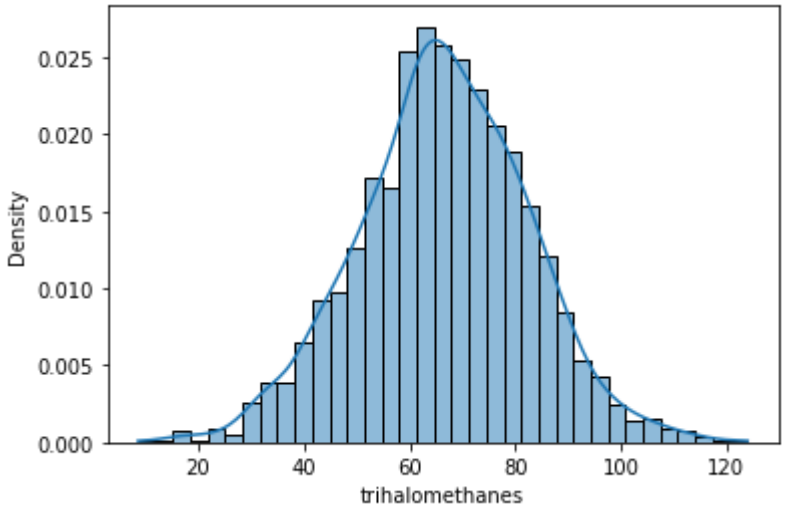
Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 0.8825496581408284 yang lebih besar dari nilai alpha. Oleh karena itu, kolom OrganicCarbon berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat simetris terhadap sumbu tegak. Skewness kurva terlihat normal (no skew).

Kolom trihalomethanes

In [357…

```
normalityTest("trihalomethanes")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 0.1043598441875204
```



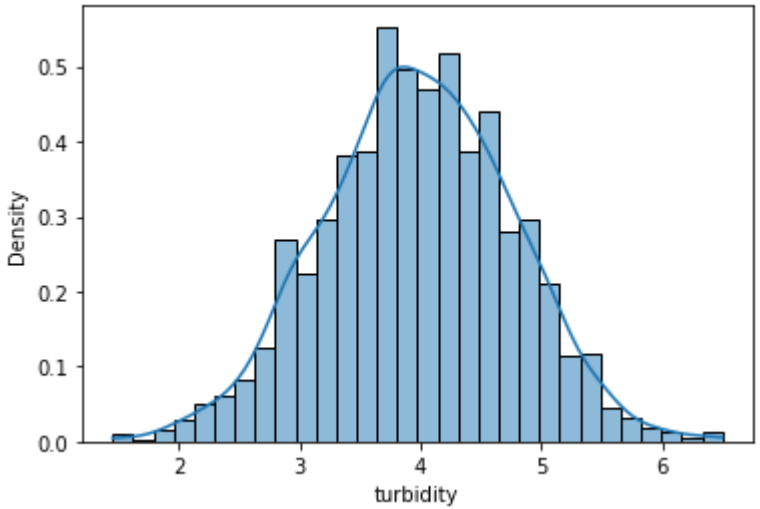
Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 0.1043598441875204 yang lebih besar dari nilai alpha. Oleh karena itu, kolom

trihalomethanes berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat simetris terhadap sumbu tegak. Skewness kurva terlihat normal (no skew).

Kolom turbidity

In [358…]
```
normalityTest("turbidity")
```

```
Nilai alpha sebesar 0.05
P-value yang didapatkan sebesar 0.7694717369961169
```



Berdasarkan hasil pengujian normal, didapatkan p-value sebesar 0.7694717369961169 yang lebih besar dari nilai alpha. Oleh karena itu, kolom turbidity berdistribusi normal. Selain itu, kurva dari histogram plot yang didapatkan terlihat simetris terhadap sumbu tegak. Skewness kurva terlihat normal (no skew).

Nomor 4

Melakukan test hipotesis 1 sampel, dengan menuliskan 6 langkah testing dan menampilkan juga boxplotnya untuk kolom/bagian yang bersesuaian.

In [359…]
```
def z_test(col, mu):
    mean = df[col].mean()
    sigma = df[col].std()
    n  = df[col].count()
    return mean, sigma, n, (mean-mu)/(sigma/(n ** 0.5))
```

a. Nilai Rata-rata pH di atas 7?

Langkah-langkah

1. Tentukan Hipotesis nol (H_0)

   H_0 : μ_{pH} = 7

2. Tentukan Hipotesis alternatif (H_1)

   H_1 : μ_{pH} > 7 (one-tailed test)

3. Tentukan tingkat signifikan (α)

   α = 0.05

4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.

   Dari jawaban nomor 1 diketahui hal-hal berikut:

   A. Jumlah sampel lebih dari 30 $(n \geq 30)$
   B. Nilai rata-rata dan standar deviasi diketahui
   C. Nilai variansi populasi diketahui

   Oleh karena tiga kondisi tersebut maka pada soal ini akan dipilih uji statistik ***Z test***

   Daerah kritis: Test ini merupakan jenis **one-tailed test** karena itu untuk menentukan nilai z sebagai perbandingan, diambil nilai z dari tabel A.3, yang akan dinotasikan sebagai z_α. Nilai *significance level* α adalah 0.05, maka nilai *confidence level* adalah $1 - 0.05 = 0.95$. Untuk mendapat nilai demikian, nilai z_α yang bersesuaian adalah **1.645** .

   Maka, daerah kritis adalah $z_\alpha > 1.645$

1. Hitung nilai uji statistik

In [360…

```
import scipy.stats as stats
alpha = 0.05
mu = 7

mean, sigma, n, z = z_test('pH', mu)
p = 1 - stats.norm.cdf(z)

print("Rata-rata sebenarnya:", mean)
print("Simpangan Baku:", sigma)
print("Jumlah Sampel:", n)
```

```
print("Hasil uji statistik (z):", z)
print("Hasil uji P-value:", p)
```

```
Rata-rata sebenarnya: 7.0871927687138285
Simpangan Baku: 1.5728029470456655
Jumlah Sampel: 2010
Hasil uji statistik (z): 2.485445147379887
Hasil uji P-value: 0.006469476288896492
```

$\bar{x} = 7.0871927687138285$

$\sigma = 1.5728029470456655$

$n = 2010$

$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

Maka:

$z = 2.485445147379887$

$p - value = 0.006469476288896492$

---

1. Mengambil keputusan

Berdasarkan data hasil perhitungan pada poin 5, diperoleh keputusan sebagai berikut:

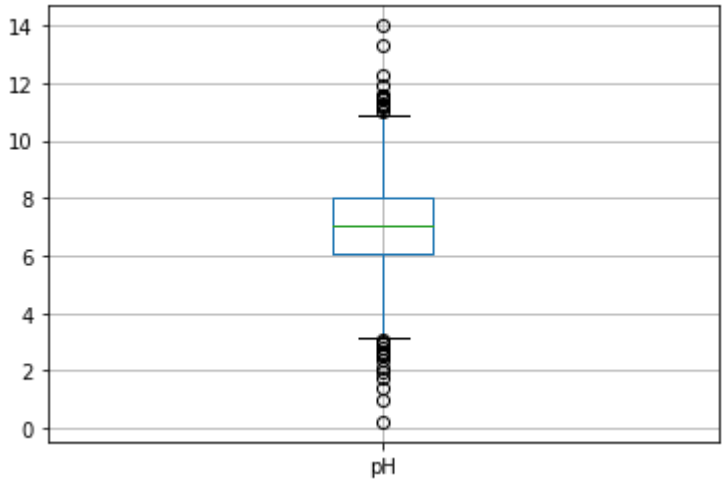
1. H_0 *DITOLAK* karena nilai uji terletak di daerah kritis
2. H_0 *DITOLAK* karena *p-value* lebih kecil dari α

Maka, H_0 *DITOLAK*

Nilai rata-rata pH di atas 7

In [361…

```
df.boxplot(["pH"])
plt.show()
```



b. Nilai Rata-rata Hardness tidak sama dengan 205?

Langkah-langkah

1. Tentukan Hipotesis nol (H_0)

   H_0 : $\mu_{Hardness}$ = 205

2. Tentukan Hipotesis alternatif (H_1)

   H_1 : $\mu_{Hardness} \neq 205$ (two-tailed test)

3. Tentukan tingkat signifikan (α)

   α = 0.05

4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.

   Dari jawaban nomor 1 diketahui hal-hal berikut:

   A. Jumlah sampel lebih dari 30 $(n \geq 30)$
   B. Nilai rata-rata dan standar deviasi diketahui
   C. Nilai variansi populasi diketahui

   Oleh karena tiga kondisi tersebut maka pada soal ini akan dipilih uji statistik ***Z test***

Daerah kritis : Test ini merupakan jenis *two-tailed test* karena daerah kritis untuk hipotesis alternatif $\theta \neq \theta_0$ berada pada *left tail* dan *right tail* dari distribusi.

Karen merupakan *two-tailed test* maka digunakan nilai $\alpha/2 = 0.025$. Untuk mendapat nilai demikian, nilai kritis $z_{\alpha/2}$ yang bersesuaian adalah **1.9600** .

Maka, daerah kritis adalah $z_{\alpha/2} < -1.9600$ dan $z_{\alpha/2} > 1.9600$

5. Hitung nilai uji statistik

In [362…

```
alpha = 0.05
mu = 205

mean, sigma, n, z = z_test("hardness", mu)
p = 2 * stats.norm.sf(z)

print("Rata-rata sebenarnya:", mean)
print("Simpangan Baku:", sigma)
print("Jumlah Sampel:", n)
print("Hasil uji statistik (z):", z)
print("Hasil uji P-value:", p)
```

```
Rata-rata sebenarnya: 195.96920903783524
Simpangan Baku: 32.643165859429864
Jumlah Sampel: 2010
Hasil uji statistik (z): -12.403137170010732
Hasil uji P-value: 2.0
```

$\bar{x} = 195.96920903783524$

$\sigma = 32.643165859429864$

$n = 2010$

$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Maka:

$z = -12.403137170010732$

$p - value = 2.0$

1. Mengambil keputusan

Berdasarkan data hasil perhitungan pada poin 5, diperoleh keputusan sebagai berikut:

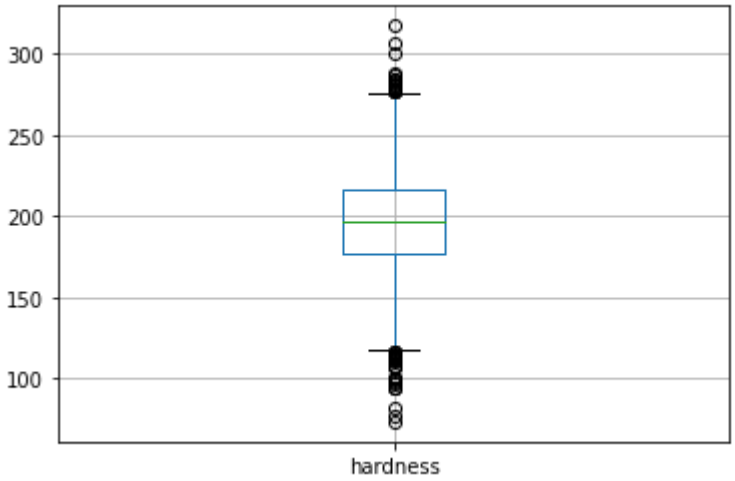
1. H_0 *DITOLAK* karena nilai uji terletak di daerah kritis

Maka, H_0 *DITOLAK*

Nilai rata-rata Hardness tidak sama dengan 205.

In [363…

```
df.boxplot(["hardness"])
plt.show()
```



c. Nilai Rata-rata 100 baris pertama kolom Solids bukan 21900?

Langkah-langkah

1. Tentukan Hipotesis nol (H_0)

   H_0 : $\mu_{100barispertamaSolids}$ = 21900

2. Tentukan Hipotesis alternatif (H_1)

   H_1 : $\mu_{100barispertamaSolids} \neq 21900$ (two-tailed test)

3. Tentukan tingkat signifikan (α)

α = 0.05

4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.

   Dari jawaban nomor 1 diketahui hal-hal berikut:

   A. Jumlah sampel lebih dari 30 $(n \geq 30)$
   B. Nilai rata-rata dan standar deviasi diketahui
   C. Nilai variansi populasi diketahui

      Oleh karena tiga kondisi tersebut maka pada soal ini akan dipilih uji statistik ***Z test***

      Daerah kritis : Test ini merupakan jenis *two-tailed test* karena daerah kritis untuk hipotesis alternatif $\theta \neq \theta_0$ berada pada *left tail* dan *right tail* dari distribusi.

      Karena merupakan *two-tailed test* maka digunakan nilai $\alpha/2 = 0.025$. Untuk mendapat nilai demikian, nilai kritis $z_{\alpha/2}$ yang bersesuaian adalah **1.96** .

      Maka, daerah kritis adalah $z_{\alpha/2} < -1.96$ dan $z_{\alpha/2} > 1.96$

5. Hitung nilai uji statistik

In [364…
```
mean = df.head(100)['solids'].mean()
sigma = df.head(100)['solids'].std()
n  = df.head(100)['solids'].count()
z = (mean-21900)/(sigma/(n ** 0.5))
p = 2 * stats.norm.sf(z)

print("Rata-rata sebenarnya:", mean)
print("Simpangan Baku:", sigma)
print("Jumlah Sampel:", n)
print("Hasil uji statistik (z):", z)
print("Hasil uji P-value:", p)
```

```
Rata-rata sebenarnya: 22347.334446383426
Simpangan Baku: 7935.967706199006
Jumlah Sampel: 100
Hasil uji statistik (z): 0.5636797715721551
Hasil uji P-value: 0.5729720864655174
```

$\bar{x} = 22347.334446383426$

$\sigma = 7935.967706199006$

$n = 100$

$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

Maka:

$z = 0.5636797715721551$

$p - value = 0.5729720864655174$

---

1. Mengambil keputusan

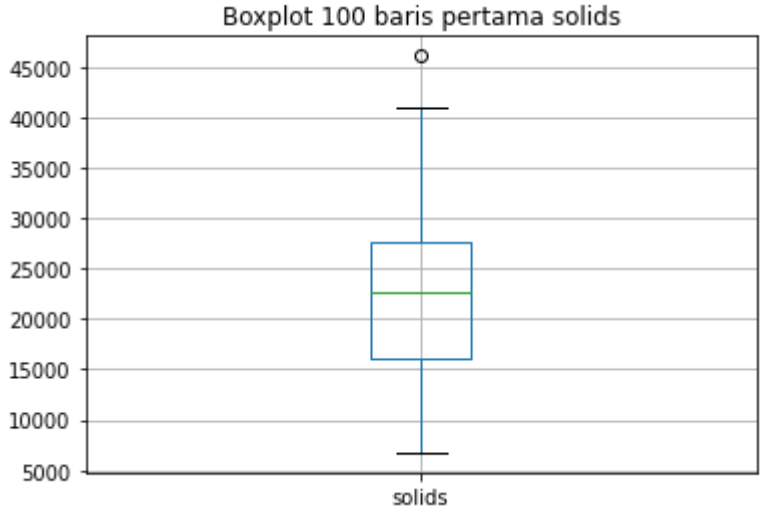
Berdasarkan data hasil perhitungan pada poin 5, diperoleh keputusan sebagai berikut:

1. Nilai uji tidak terletak pada daerah kritis
2. P-value lebih besar dibanding α

Maka: **Pernyataan nilai rata-rata 100 baris pertama kolom solids bukan 21900 tidak terbukti.**

In [365…

```
df[:100].boxplot(["solids"])
plt.title("Boxplot 100 baris pertama solids")
plt.show()
```



d. Proporsi nilai Conductivity yang lebih dari 450, adalah tidak sama dengan 10%?

Langkah-langkah

1. Tentukan Hipotesis nol (H_0)

   $H_0 : p = 0.1$

2. Tentukan Hipotesis alternatif (H_1)

   $H_1 : p \neq 0.1$ (two-tailed test)

1. Tentukan tingkat signifikan (α)

   α = 0.05

2. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.

   Karena jumlah sampel yang cukup besar maka uji statistik yang kami adalah ***distribusi binomial dengan pendekatan distribusi normal***

   Daerah kritis : Test ini merupakan jenis *two-tailed test* karena daerah kritis untuk hipotesis alternatif $\theta \neq \theta_0$ berada pada *left tail* dan *right tail* dari distribusi.

   Karena merupakan *two-tailed test* maka digunakan nilai $\alpha/2 = 0.025$. Untuk mendapat nilai demikian, nilai kritis $z_{\alpha/2}$ yang bersesuaian adalah **1.96** .

   Maka, daerah kritis adalah $z_{\alpha/2} < -1.96$ dan $z_{\alpha/2} > 1.96$

3. Hitung nilai uji statistik

In [366…

```
alpha = 0.05

n = df.shape[0]
temp = df.loc[df['conductivity'] > 450]['conductivity']
x = temp.count()

# propotion = df_conGt450.count()[0]/df.count()
p0 = 0.1
q0 = 1 - p0
# sigma = df['conductivity'].std()


z_alpha = stats.norm.ppf(1-alpha/2)
z = (x-n*p0)/(math.sqrt(n*p0*(1-p0)))
```

```
p = 2 * stats.norm.sf(z)


print("Jumlah data (x):",x)
print("Proportion (p0):", p0)
print("n", n)
print("Hasil uji statistik (z):", z)
print("Hasil uji P-value:", p)
```

```
Jumlah data (x): 745
Proportion (p0): 0.1
n 2010
Hasil uji statistik (z): 40.446376131589325
Hasil uji P-value: 0.0
```

$x = 745$

$p_0 = 0.15$

$n = 2010$

$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$

Maka:

$z = 40.446376131589325$

$p - value = 0.0$

---

1. Mengambil keputusan

Berdasarkan data hasil perhitungan pada poin 5, diperoleh keputusan sebagai berikut:

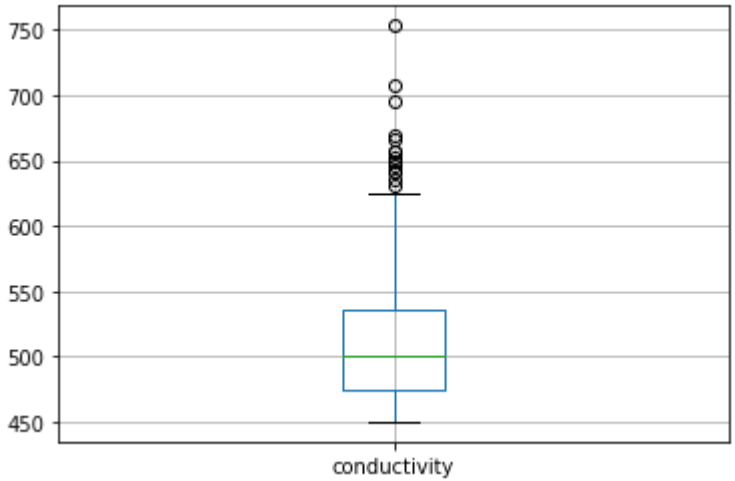
1. H_0 *DITOLAK* karena nilai uji terletak pada daerah kritis
2. H_0 *DITOLAK* karena P-value lebih kecil dibanding α

Maka: **Proporsi nilai Conductivity yang lebih dari 450, adalah tidak sama dengan 10% terbukti**

In [367…

```
temp = df.loc[df['conductivity'] > 450]
temp.boxplot(['conductivity'])
```

```
plt.show()
```



e. Proporsi nilai Trihalomethanes yang kurang dari 40, adalah kurang dari 5%?

Langkah-langkah

1. Tentukan Hipotesis nol (H_0)

   $H_0 : p = 0.05$

2. Tentukan Hipotesis alternatif (H_1)

   $H_1 : p < 0.05$ (one-tailed test)

1. Tentukan tingkat signifikan (α)

   α = 0.05

2. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.

   Karena jumlah sampel yang cukup besar maka uji statistik yang kami adalah ***distribusi binomial dengan pendekatan distribusi normal***

   Daerah kritis: Test ini merupakan jenis **one-tailed test** karena itu untuk menentukan nilai z sebagai perbandingan, diambil nilai z dari tabel A.3, yang akan dinotasikan sebagai z_α. Nilai *significance level* α adalah 0.05, maka nilai *confidence level* adalah $1 - 0.05 = 0.95$. Untuk mendapat nilai demikian, nilai z_α yang bersesuaian adalah **1.645** .

Maka, daerah kritis adalah $z_\alpha < -1.645$

1. Hitung nilai uji statistik

In [368…

```
alpha = 0.05


n = df.shape[0]
temp = df.loc[df['trihalomethanes'] < 40]['trihalomethanes']
x = temp.count()

# propotion = df_conGt450.count()[0]/df.count()
p0 = 0.05
q0 = 1 - p0
# sigma = df['conductivity'].std()

z = (x-n*p0)/(math.sqrt(n*p0*(1-p0)))
p = stats.norm.cdf(z)


print("Jumlah data (x):",x)
print("Proportion (p0):", p0)
print("n", n)
print("Hasil uji statistik (z):", z)
print("Hasil uji P-value:", p)
```

```
Jumlah data (x): 106
Proportion (p0): 0.05
n 2010
Hasil uji statistik (z): 0.5628826416670959
Hasil uji P-value: 0.7132425995092373
```

$x = 106$

$p_0 = 0.05$

$n = 2010$

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

Maka:

$z = 0.5628826416670959$

$p - value = 0.7132425995092373$

---

1. Mengambil keputusan

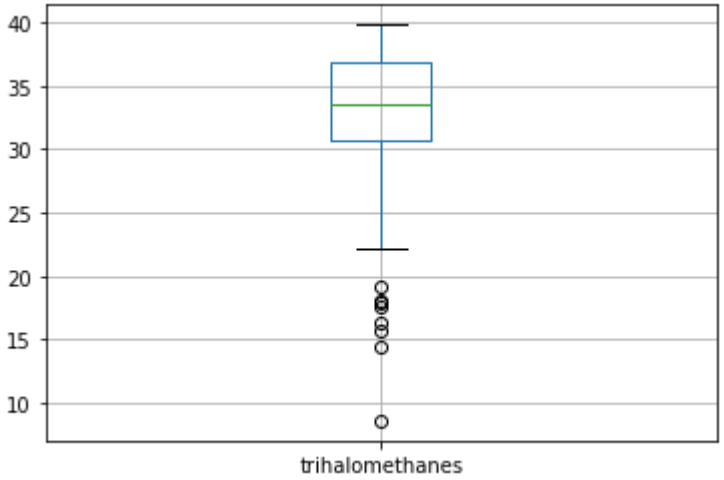
Berdasarkan data hasil perhitungan pada poin 5, diperoleh keputusan sebagai berikut:

1. Nilai uji terletak di luar daerah kritis
2. P-value lebih besar dibanding α

Maka: **Proporsi nilai Trihalomethanes yang kurang dari 40, adalah kurang dari 5% tidak terbukti**

In [369…

```
temp = df.loc[df['trihalomethanes'] < 40]
temp.boxplot(['trihalomethanes'])
# df.boxplot(['trihalomethanes'])

plt.show()
```



Nomor 5

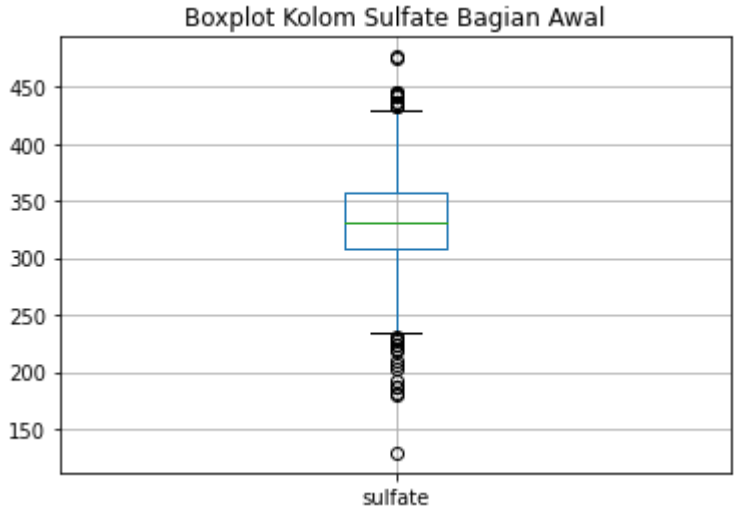
Melakukan test hipotesis 2 sampel, dengan menuliskan 6 langkah testing dan menampilkan juga boxplotnya untuk kolom/bagian yang bersesuaian.

a. Data kolom Sulfate dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata kedua bagian tersebut sama?

In [370…

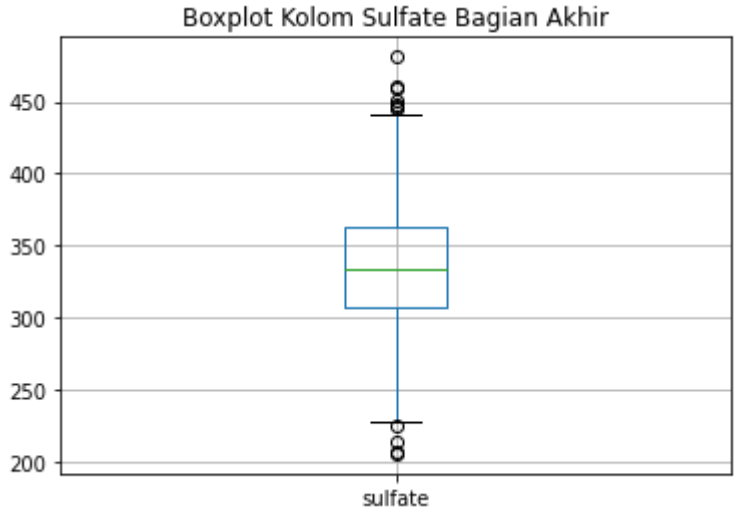
```
# Boxplot data kolom Sulfate bagian awal kolom
```

```
df[:1005].boxplot(["sulfate"])
plt.title("Boxplot Kolom Sulfate Bagian Awal")
plt.show()
```



In [371…

```
# Boxplot data kolom Sulfate bagian akhir kolom

df[1005:].boxplot(["sulfate"])
plt.title("Boxplot Kolom Sulfate Bagian Akhir")
plt.show()
```



μ_1 = rata-rata kolom sulfate bagian awal

μ_2 = rata-rata kolom sulfate bagian akhir

Enam langkah testing:

1. Penentuan H_0: $\mu_1 - \mu_2 = 0$
2. Penentuan H_1: $\mu_1 - \mu_2 \neq 0$
3. Penentuan tingkat signifikan $\alpha = 0.05$
4. Uji statistik yang digunakan adalah two-tailed test dengan kedua variansi diketahui dengan rumus seperti berikut:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

dengan $d_0 = 0$

Penentuan daerah kritis:

$z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$

$z < -z_{0.025}$ atau $z > z_{0.025}$

$z < -1.96$ atau $z > 1.96$ (Menggunakan Tabel Critical Values of the t-Distributon atau Tabel A.4)

1. Perhitungan nilai uji statistik dan p-value:

In [372…

```
# Perhitungan nilai uji statistik

mean1 = df[:1005]["sulfate"].mean()
mean2 = df[1005:]["sulfate"].mean()

var1 = df[:1005]["sulfate"].var()
var2 = df[1005:]["sulfate"].var()


z = (mean1 - mean2) / (math.sqrt(var1/1005 + var2/1005))
print("Nilai uji statistik (z) = " + str(z))

# Perhitungan p-value

'''
    Perhitungan p-value dilakukan dengan bantuan library scipy untuk menghitung
    fungsi distribusi kumulatif
'''
pVal = 2*(1-s.norm.cdf(abs(z)))
print("Nilai p-value = " + str(pVal))
```

```
Nilai uji statistik (z) = -2.0752690696871983
Nilai p-value = 0.03796160438512852
```

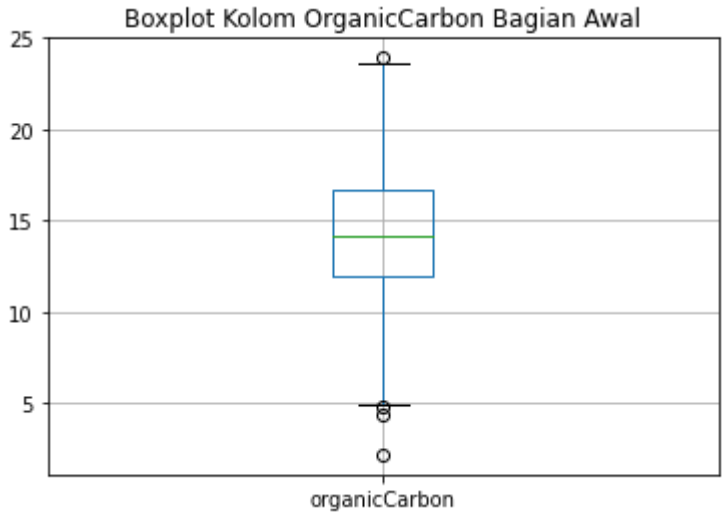
1. Pengambilan keputusan

Berdasarkan nilai uji statistik, yaitu -2.0752690696871983 yang berada di dalam daerah kritis, diambil keputusan bahwa H_0 ditolak. Selain itu, didapatkan nilai p-value sebesar 0.03796160438512852 yang kurang dari nilai alpha, yaitu 0.05. Oleh karena itu, juga diambil keputusan bahwa H_0 ditolak. Dengan kata lain, rata-rata bagian awal kolom Sulfate tidak sama dengan rata-rata bagian akhir kolom Sulfate.

b. Data kolom OrganicCarbon dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata bagian awal lebih besar dari pada bagian akhir sebesar 0.15?

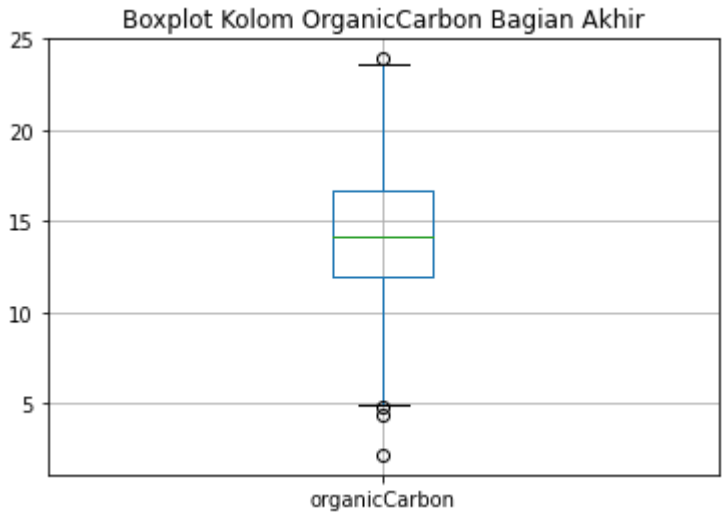
In [373…

```
# Boxplot data kolom OrganicCarbon bagian awal kolom

df[:1005].boxplot(["organicCarbon"])
plt.title("Boxplot Kolom OrganicCarbon Bagian Awal")
plt.show()
```



In [374…

```
# Boxplot data kolom OrganicCarbon bagian akhir kolom

df[:1005].boxplot(["organicCarbon"])
plt.title("Boxplot Kolom OrganicCarbon Bagian Akhir")
plt.show()
```



μ_1 = rata-rata kolom OrganicCarbon bagian awal

μ_2 = rata-rata kolom OrganicCarbon bagian akhir

Enam langkah testing:

1. Penentuan H_0: $\mu_1 - \mu_2 = 0.15$
2. Penentuan H_1: $\mu_1 - \mu_2 \neq 0.15$
3. Penentuan tingkat signifikan $\alpha = 0.05$
4. Uji statistik yang digunakan adalah two-tailed test dengan kedua variansi diketahui dengan rumus seperti berikut:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

dengan $d_0 = 0.15$

Penentuan daerah kritis:

$z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$

$z < -z_{0.025}$ atau $z > z_{0.025}$

$z < -1.96$ atau $z > 1.96$ (Menggunakan Tabel Critical Values of the t-Distributon atau Tabel A.4)

1. Perhitungan nilai uji statistik dan p-value:

In [375…
```
# Perhitungan nilai uji statistik

mean1 = df[:1005]["organicCarbon"].mean()
mean2 = df[1005:]["organicCarbon"].mean()

var1 = df[:1005]["organicCarbon"].var()
var2 = df[1005:]["organicCarbon"].var()

z = (mean1 - mean2 - 0.15) / (math.sqrt(var1/1005 + var2/1005))
print("Nilai uji statistik (z) = " + str(z))

# Perhitungan p-value

'''
    Perhitungan p-value dilakukan dengan bantuan library scipy untuk menghitung
    fungsi distribusi kumulatif
'''
pVal = 2*(1-s.norm.cdf(abs(z)))
print("Nilai p-value = " + str(pVal))
```

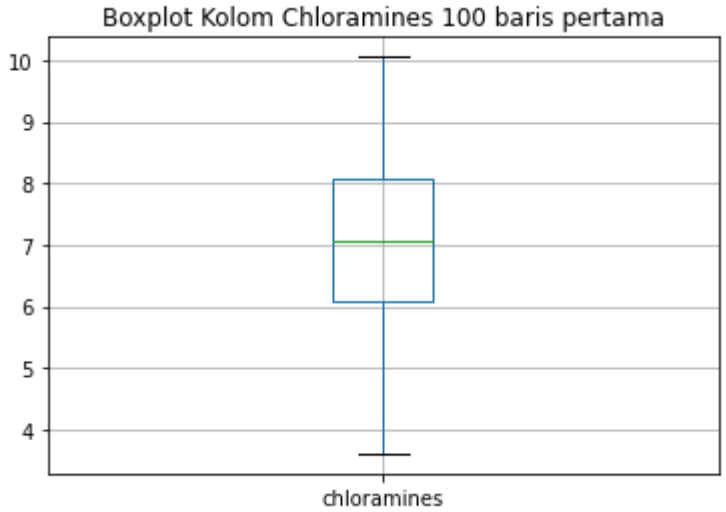
```
Nilai uji statistik (z) = -2.413145517798807
Nilai p-value = 0.015815503817599996
```

1. Pengambilan keputusan

Berdasarkan nilai uji statistik, yaitu -2.413145517798807 yang berada di dalam daerah kritis, diambil keputusan bahwa H_0 ditolak. Selain itu, didapatkan nilai p-value sebesar 0.015815503817599996 yang kurang dari nilai alpha, yaitu 0.05. Oleh karena itu, juga diambil keputusan bahwa H_0 ditolak. Dengan kata lain, rata-rata bagian awal kolom OrganicCarbon tidak lebih besar dari rata-rata bagian akhir kolom OrganicCarbon sebesar 0.15.

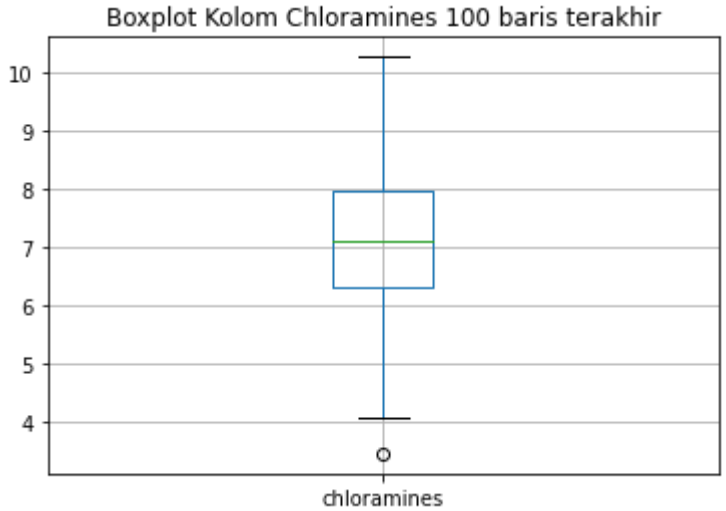
c. Rata-rata 100 baris pertama kolom Chloramines sama dengan 100 baris terakhirnya?

In [376…
```
# Boxplot data kolom Chloramines 100 baris pertama

df[:100].boxplot(["chloramines"])
plt.title("Boxplot Kolom Chloramines 100 baris pertama")
plt.show()
```



In [377…

```
# Boxplot data kolom Chloramines 100 baris terakhir

df[1910:].boxplot(["chloramines"])
plt.title("Boxplot Kolom Chloramines 100 baris terakhir")
plt.show()
```



μ_1 = rata-rata kolom Chloramines 100 baris pertama

μ_2 = rata-rata kolom Chloramines 100 baris terakhir

Enam langkah testing:

1. Penentuan H_0: $\mu_1 - \mu_2 = 0$

2. Penentuan H_1: $\mu_1 - \mu_2 \neq 0$
3. Penentuan tingkat signifikan $\alpha = 0.05$
4. Uji statistik yang digunakan adalah two-tailed test dengan kedua variansi diketahui dengan rumus seperti berikut:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

dengan $d_0 = 0$

Penentuan daerah kritis:

$z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$

$z < -z_{0.025}$ atau $z > z_{0.025}$

$z < -1.96$ atau $z > 1.96$ (Menggunakan Tabel Critical Values of the t-Distributon atau Tabel A.4)

1. Perhitungan nilai uji statistik dan p-value:

In [378…

```
# Perhitungan nilai uji statistik

mean1 = df[:100]["chloramines"].mean()
mean2 = df[1910:]["chloramines"].mean()

var1 = df[:100]["chloramines"].var()
var2 = df[1910:]["chloramines"].var()

z = (mean1 - mean2) / (math.sqrt(var1/1005 + var2/1005))
print("Nilai uji statistik (z) = " + str(z))


# Perhitungan p-value

'''
    Perhitungan p-value dilakukan dengan bantuan library scipy untuk menghitung
    fungsi distribusi kumulatif
'''
pVal = 2*(1-s.norm.cdf(abs(z)))
print("Nilai p-value = " + str(pVal))
```

```
Nilai uji statistik (z) = -2.2379601537718905
Nilai p-value = 0.0252236536904078
```

1. Pengambilan keputusan

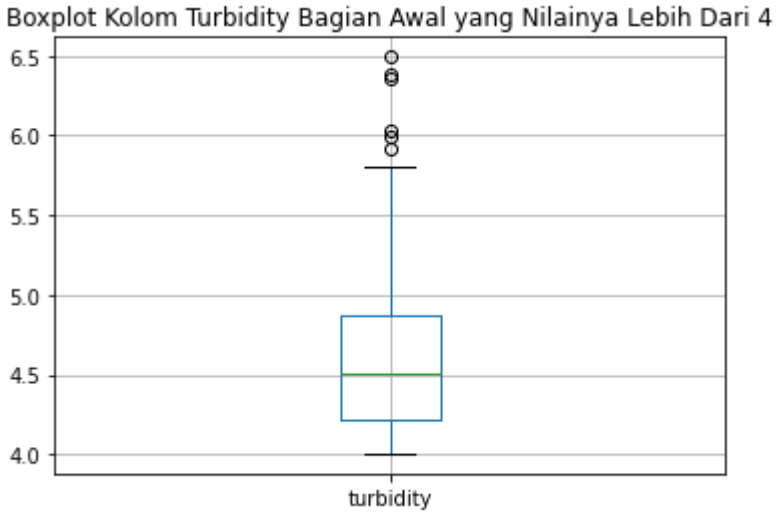
Berdasarkan nilai uji statistik, yaitu -2.2379601537718905 yang berada di dalam daerah kritis, diambil keputusan bahwa H_0 ditolak. Selain itu, didapatkan nilai p-value sebesar 0.0252236536904078 yang kurang dari nilai alpha, yaitu 0.05. Oleh karena itu, juga diambil keputusan bahwa H_0 ditolak. Dengan kata lain, rata-rata 100 baris pertama kolom Chloramines tidak sama dengan rata-rata 100 baris terakhir kolom Chloramines.

d. Proporsi nilai bagian awal Turbidity yang lebih dari 4, adalah lebih besar daripada,

proporsi nilai yang sama di bagian akhir Turbidity ?

In [379…

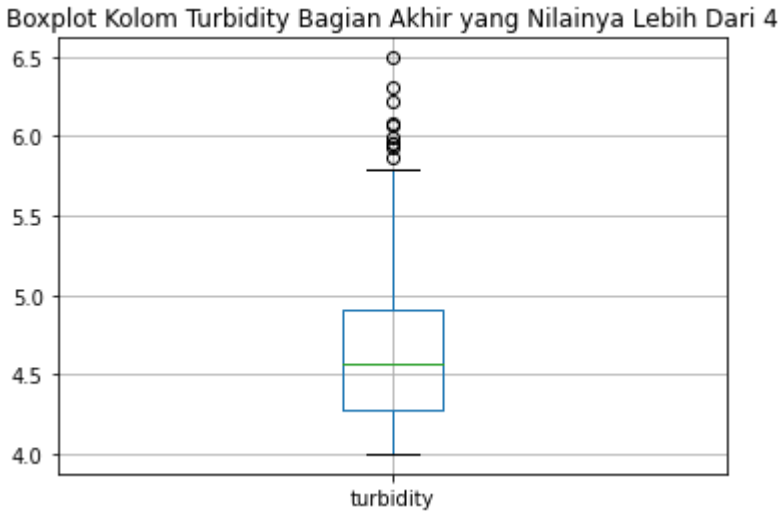
```
# Boxplot data kolom Turbidity bagian awal yang nilainya lebih dari 4

awal = df[:1005]
awal = awal[awal.turbidity > 4]

awal.boxplot(["turbidity"])

plt.title("Boxplot Kolom Turbidity Bagian Awal yang Nilainya Lebih Dari 4")
plt.show()
```



In [380…

```
# Boxplot data kolom Turbidity bagian akhir yang nilainya lebih dari 4

akhir = df[1005:]
akhir = akhir[akhir.turbidity > 4]

akhir.boxplot(["turbidity"])
```

```
plt.title("Boxplot Kolom Turbidity Bagian Akhir yang Nilainya Lebih Dari 4")
plt.show()
```



p_1 = proporsi kolom Turbidity bagian awal yang nilainya lebih dari 4

p_2 = proporsi kolom Turbidity bagian akhir yang nilainya lebih dari 4

Enam langkah testing:

1. Penentuan H_0: $p_1 - p_2 = 0$
2. Penentuan H_1: $p_1 - p_2 > 0$
3. Penentuan tingkat signifikan $\alpha = 0.05$
4. Uji statistik yang digunakan adalah one tailed-test pada dua proporsi dengan sampel yang berdistribusi normal dengan rumus seperti berikut:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Penentuan daerah kritis:

$z > z_\alpha$

$z > z_{0.05}$

$z > 1.645$ (Menggunakan Tabel Critical Values of the t-Distributon atau Tabel A.4)

1. Perhitungan nilai uji statistik dan p-value:

In [381…]
```
# bagian awal kolom
awal = df[:1005]
awal = awal[awal.turbidity > 4]["turbidity"]

# bagian akhir kolom
akhir = df[1005:]
akhir = akhir[akhir.turbidity > 4]["turbidity"]

p1 = awal.count() / 1005
p2 = akhir.count() / 1005

p = (awal.count() + akhir.count()) / (1005 + 1005)

z = (p1 - p2) / (math.sqrt(p * (1-p) * (1/1005 + 1/1005)))
print("Nilai uji statistik (z) = " + str(z))

# Perhitungan p-value

'''
    Perhitungan p-value dilakukan dengan bantuan library scipy untuk menghitung
    fungsi distribusi kumulatif
'''

pVal = 1-(s.norm.cdf(abs(z)))
print("Nilai p-value = " + str(pVal))
```

```
Nilai uji statistik (z) = -0.13388958661778735
Nilai p-value = 0.4467449424088169
```

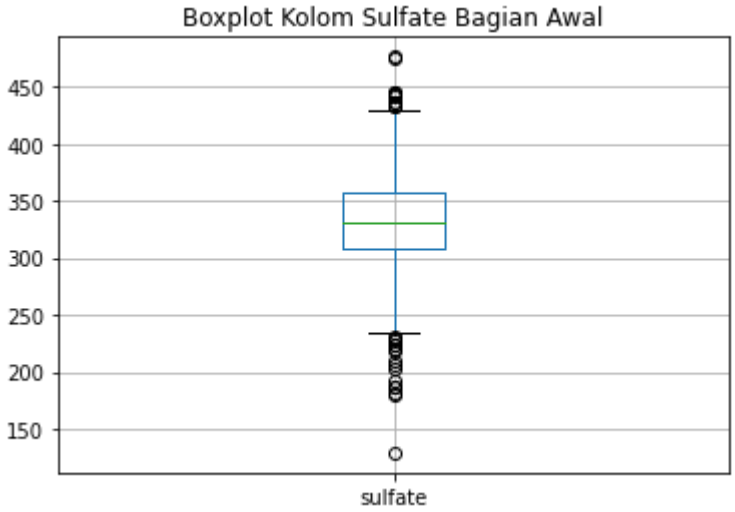
1. Pengambilan keputusan

Berdasarkan nilai uji statistik, yaitu -0.13388958661778735 yang berada di luar daerah kritis, diambil keputusan bahwa H_0 diterima. Selain itu, didapatkan nilai p-value sebesar 0.4467449424088169 yang lebih besar dari nilai alpha, yaitu 0.05. Oleh karena itu, juga diambil keputusan bahwa H_0 diterima. Dengan kata lain, proporsi nilai bagian awal Turbidity yang lebih dari 4 sama dengan proporsi nilai bagian akhir Turbidity yang lebih dari 4.

e. Bagian awal kolom Sulfate memiliki variansi yang sama dengan bagian akhirnya?

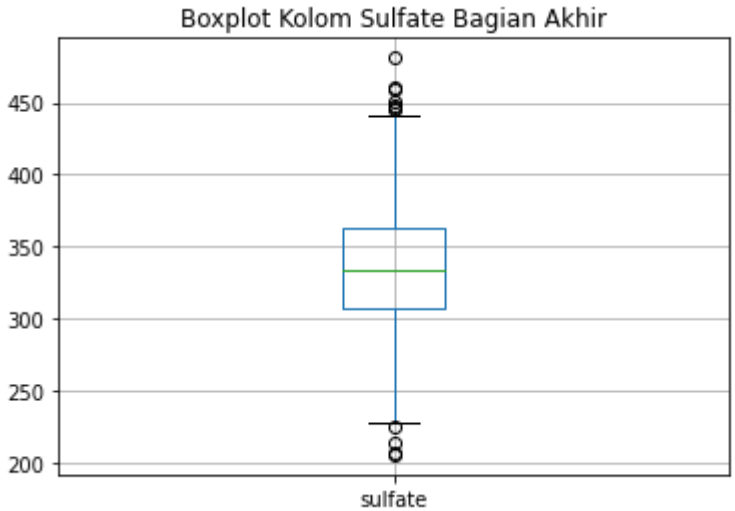
In [382…]
```
# Boxplot data kolom Sulfate bagian awal kolom

df[:1005].boxplot(["sulfate"])
```

```
plt.title("Boxplot Kolom Sulfate Bagian Awal")
plt.show()
```



In [383…

```
# Boxplot data kolom Sulfate bagian akhir kolom

df[1005:].boxplot(["sulfate"])
plt.title("Boxplot Kolom Sulfate Bagian Akhir")
plt.show()
```



σ_1^2 = variansi kolom Sulfate bagian awal

σ_2^2 = variansi kolom Sulfate bagian akhir

Enam langkah testing:

1. Penentuan H_0: $\sigma_1^2 - \sigma_2^2 = 0$
2. Penentuan H_1: $\sigma_1^2 - \sigma_2^2 \neq 0$
3. Penentuan tingkat signifikan $\alpha = 0.05$
4. Uji statistik yang digunakan adalah two tailed-test pada dua variansi dengan sampel yang berdistribusi F dengan rumus seperti berikut:

$$f = \frac{S_2^2}{S_1^2}$$

Penentuan daerah kritis, yaitu:

$f < f_{1-\alpha/2}(v_1, v_2)$ atau $f > f_{\alpha/2}(v_1, v_2)$

dengan $v_1 = n_1 - 1$ dan $v_2 = n_2 - 1$

$v_1 = 1005 - 1 = 1004$

$v_2 = 1005 - 1 = 1004$

In [384…
```
# penentuan daerah kritis menggunakan library scipy

upper = s.f.ppf(1 - 0.05/2, 1005 - 1, 1005 - 1)
lower = s.f.ppf(0.05/2, 1005 - 1, 1005 - 1)

print("Nilai batas bawah daerah kritis = " + str(lower))
print("Nilai batas atas daerah kritis = " + str(upper))
```

```
Nilai batas bawah daerah kritis = 0.883572344355818
Nilai batas atas daerah kritis = 1.1317692392568777
```

Didapatkan daerah kritis sebagai berikut:

f < 0.883572344355818 atau f > 1.1317692392568777

1. Perhitungan nilai uji statistik dan p-value:

In [385…
```
var1 = df[:1005]["sulfate"].var()
var2 = df[1005:]["sulfate"].var()

f = var1 / var2
print("Nilai uji statistik (f) = " + str(f))
```

```
# Perhitungan p-value

'''
    Perhitungan p-value dilakukan dengan bantuan library scipy untuk menghitung
    fungsi distribusi kumulatif
'''

pVal = 2*(1-s.norm.cdf(abs(f)))
print("Nilai p-value = " + str(pVal))
```

```
Nilai uji statistik (f) = 1.0152511043950063
Nilai p-value = 0.30998614559492665
```

1. Pengambilan keputusan

Berdasarkan nilai uji statistik, yaitu 1.0152511043950063 yang berada di luar daerah kritis, diambil keputusan bahwa H_0 diterima. Selain itu, didapatkan nilai p-value sebesar 0.30998614559492665 yang lebih besar dari nilai alpha, yaitu 0.05. Oleh karena itu, juga diambil keputusan bahwa H_0 diterima. Dengan kata lain, variansi kolom Sulfate bagian awal sama dengan variansi kolom Sulfate bagian akhir.

Nomor 6

Test korelasi: tentukan apakah setiap kolom non-target berkorelasi dengan kolom target, dengan menggambarkan juga scatter plot nya. Gunakan correlation test!

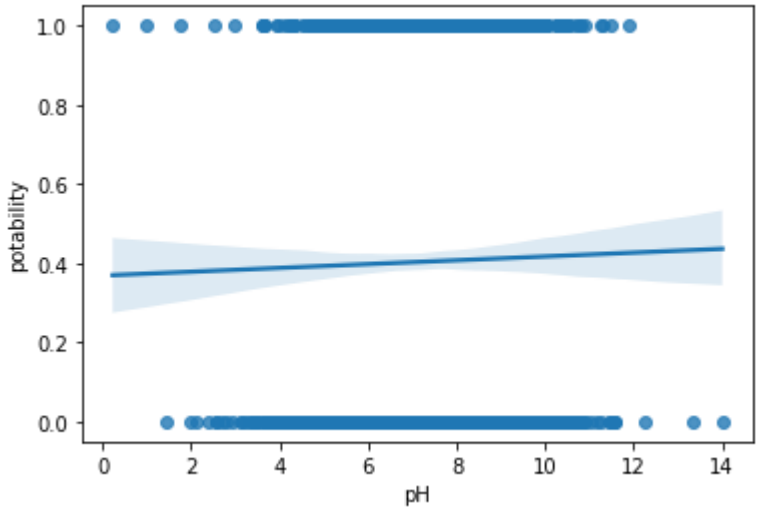
Tes korelasi yang dilakukan menggunakan bantuan fungsi pearsonr dari library scipy yang berdasarkan Person Correlation Coefficient atau Koefisien Korelasi Pearson.

In [386…
```
def correlation(column1, column2):
    cor, pVal = s.pearsonr(df[column1], df[column2])
    sns.regplot(x = df[column1], y = df[column2])
    print("Nilai korelasi antara kolom " + column1 + " dan " + column2 + " sebesar " + str(cor))
```

Kolom pH dan potability

In [387…
```
correlation("pH", "potability")
```

```
Nilai korelasi antara kolom pH dan potability sebesar 0.015475094408433499
```



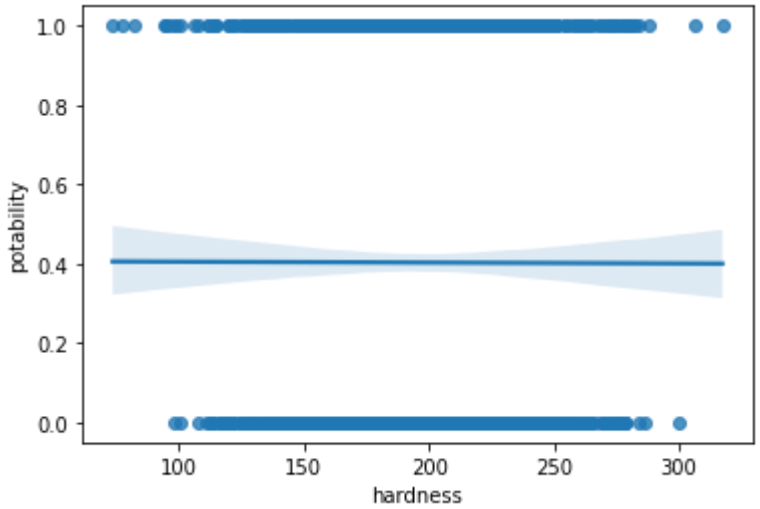
Nilai korelasi antara kolom pH dan potability sebesar 0.015475094408433499 lebih besar dari nol. Artinya, kolom pH dan potability berkorelasi secara positif atau berbanding lurus. Hal itu juga terlihat dari grafik regresi yang memiliki gradien positif.

Kolom hardness dan potability

In [388…

```
correlation("hardness", "potability")
```

```
Nilai korelasi antara kolom hardness dan potability sebesar -0.0014631528959479546
```

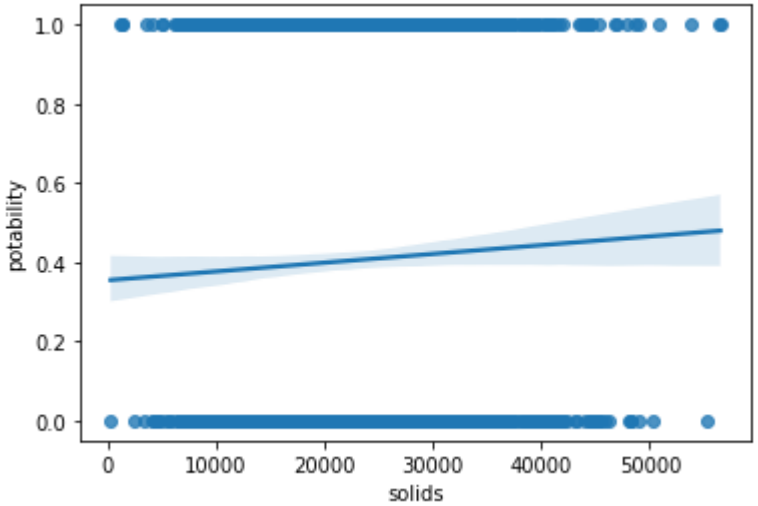


Nilai korelasi antara kolom hardness dan potability sebesar -0.0014631528959479546 lebih kecil dari nol. Artinya, kolom hardness dan potability berkorelasi secara negatif atau berbanding terbalik. Hal itu juga terlihat dari grafik regresi yang memiliki gradien negatif.

Kolom solids dan potability

In [389…
```
correlation("solids", "potability")
```

```
Nilai korelasi antara kolom solids dan potability sebesar 0.03897657818173474
```

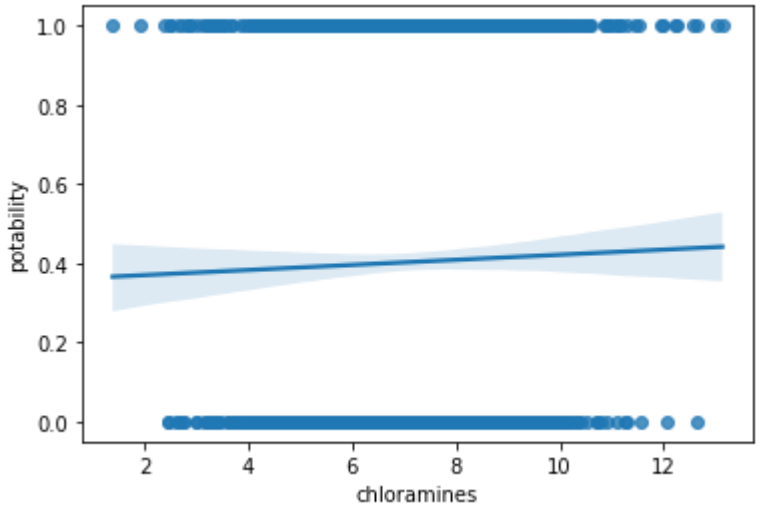


Nilai korelasi antara kolom solids dan potability sebesar 0.03897657818173474 lebih besar dari nol. Artinya, kolom solids dan potability berkorelasi secara positif atau berbanding lurus. Hal itu juga terlihat dari grafik regresi yang memiliki gradien positif.

Kolom chloramines dan potability

In [390…
```
correlation("chloramines", "potability")
```

```
Nilai korelasi antara kolom chloramines dan potability sebesar 0.020778921840524135
```



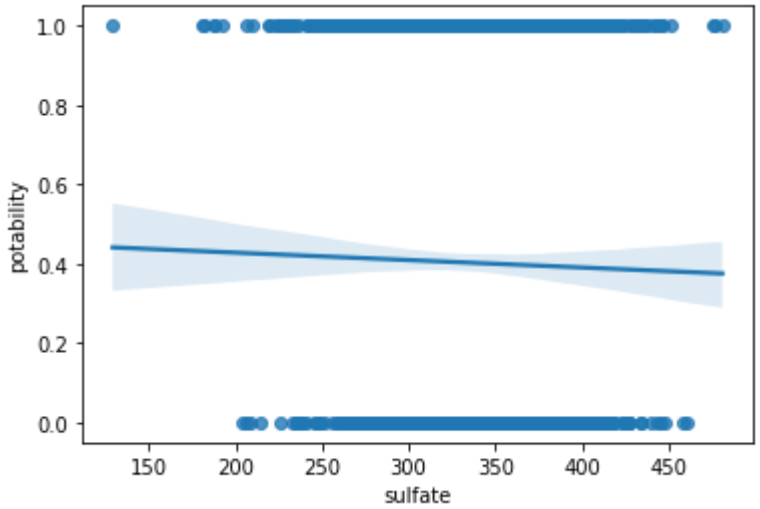
Nilai korelasi antara kolom chloramines dan potability sebesar 0.020778921840524135 lebih besar dari nol. Artinya, kolom chloramines dan potability berkorelasi secara positif atau berbanding lurus. Hal itu juga terlihat dari grafik regresi yang memiliki gradien positif.

Kolom sulfate dan potability

In [391…

```
correlation("sulfate", "potability")
```

```
Nilai korelasi antara kolom sulfate dan potability sebesar -0.01570316441927381
```

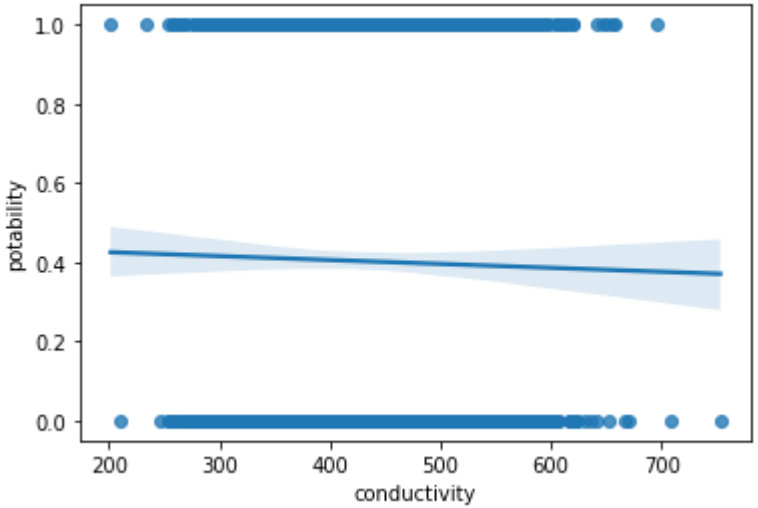


Nilai korelasi antara kolom sulfate dan potability sebesar -0.01570316441927381 lebih kecil dari nol. Artinya, kolom sulfate dan potability berkorelasi secara negatif atau berbanding terbalik. Hal itu juga terlihat dari grafik regresi yang memiliki gradien negatif.

Kolom conductivity dan potability

In [392…
```
correlation("conductivity", "potability")
```

```
Nilai korelasi antara kolom conductivity dan potability sebesar -0.01625712011137709
```

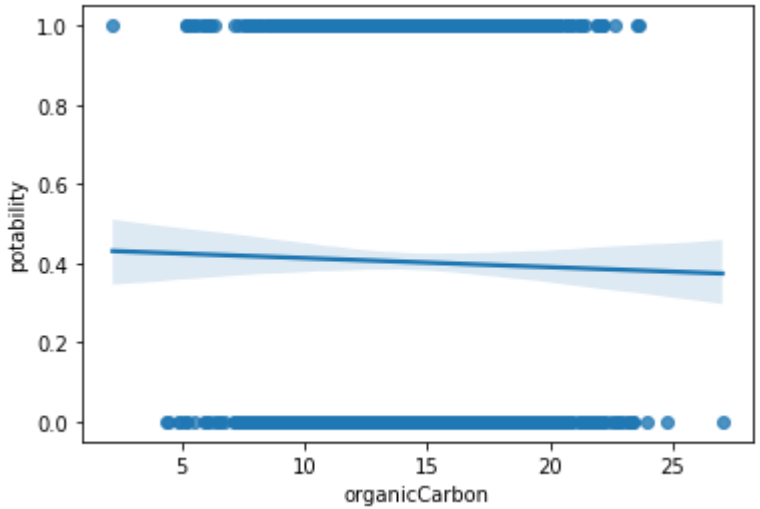


Nilai korelasi antara kolom conductivity dan potability sebesar -0.01625712011137709 lebih kecil dari nol. Artinya, kolom conductivity dan potability berkorelasi secara negatif atau berbanding terbalik. Hal itu juga terlihat dari grafik regresi yang memiliki gradien negatif.

Kolom organic carbon dan potability

In [393…
```
correlation("organicCarbon", "potability")
```

```
Nilai korelasi antara kolom organicCarbon dan potability sebesar -0.015488461910747308
```



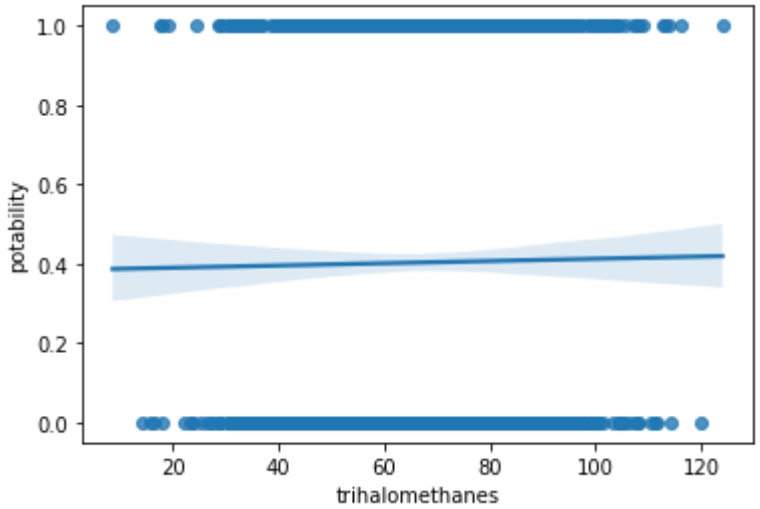
Nilai korelasi antara kolom OrganicCarbon dan potability sebesar -0.015488461910747308 lebih kecil dari nol. Artinya, kolom OrganicCarbon dan potability berkorelasi secara negatif atau berbanding terbalik. Hal itu juga terlihat dari grafik regresi yang memiliki gradien negatif.

Kolom trihalomethanes dan potability

In [394…

```
correlation("trihalomethanes", "potability")
```

```
Nilai korelasi antara kolom trihalomethanes dan potability sebesar 0.009236711064713032
```



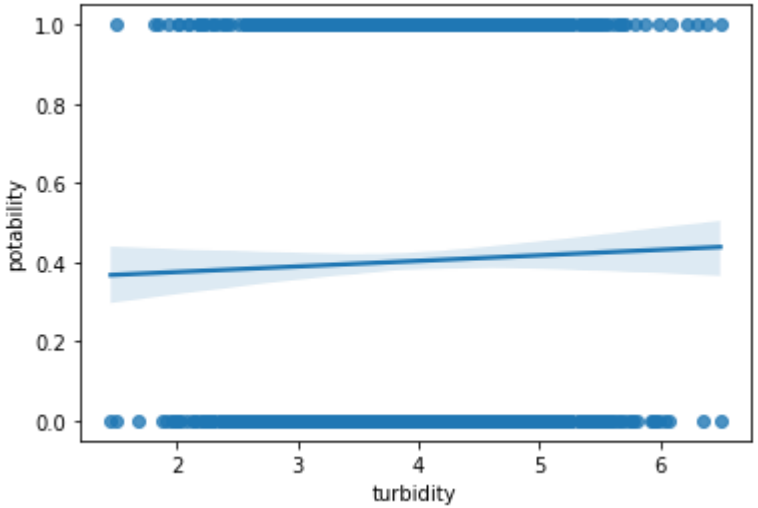
Nilai korelasi antara kolom trihalomethanes dan potability sebesar 0.009236711064713032 lebih besar dari nol. Artinya, kolom trihalomethanes dan potability berkorelasi secara positif atau berbanding lurus. Hal itu juga terlihat dari grafik regresi yang memiliki gradien positif.

Kolom turbidity dan potability

In [395…

```
correlation("turbidity", "potability")
```

```
Nilai korelasi antara kolom turbidity dan potability sebesar 0.0223310426406227
```



Nilai korelasi antara kolom turbidity dan potability sebesar 0.0223310426406227 lebih besar dari nol. Artinya, kolom turbidity dan potability berkorelasi secara positif atau berbanding lurus. Hal itu juga terlihat dari grafik regresi yang memiliki gradien positif.