notebook

April 11, 2022

1 Penarikan Kesimpulan dan Pengujian Hipotesis Data Matriks Kualitas Air

Tugas Besar IF2220 Probabilitas dan Statistika

Disusun oleh: 1. 13520047 Hana Fathiyah 2. 13520128 Bayu Samudra

---

1.1 Requirement Modul Analisis

Pada tugas besar ini, kami menggunakan modul-modul sebagai berikut. 1. Numpy versi 1.22.3 2. Pandas versi 1.4.1 3. Seaborn versi 0.11.2 4. Matplotlib versi 3.5.1 5. Jupyterlab versi 3.3.2

Modul-modul tersebut dapat di-*install* dengan perintah sebagai berikut.

```
pip install -r requirements.txt
```

Berikut ini kami mencoba untuk melakukan *import library* (pustaka) tersebut.

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

sns.set_theme()
```

1.2 Persiapan Data

Diberikan suatu *dataset* dengan nama `water_potability.csv`. Pada bagian ini, dataset tersebut akan di-*import* ke dalam sebuah variabel yang diberi nama `data`

[ ]:
```
data = pd.read_csv("water_potability.csv")
data.head()
```

[ ]:

	id	pH	Hardness	Solids	Chloramines	Sulfate	\
0	1	8.316766	214.373394	22018.417441	8.059332	356.886136	
1	2	9.092223	181.101509	17978.986339	6.546600	310.135738	
2	3	5.584087	188.313324	28748.687739	7.544869	326.678363	
3	4	10.223862	248.071735	28749.716544	7.513408	393.663396	

```
4   5   8.635849  203.361523  13672.091764     4.563009  303.309771
```

	Conductivity	OrganicCarbon	Trihalomethanes	Turbidity	Potability
0	363.266516	18.436524	100.341674	4.628771	0
1	398.410813	11.558279	31.997993	4.075075	0
2	280.467916	8.399735	54.917862	2.559708	0
3	283.651634	13.789695	84.603556	2.672989	0
4	474.607645	12.363817	62.798309	4.401425	0

Berikut ini adalah metadata dari dataset yang telah diimport

[ ]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2010 entries, 0 to 2009
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               2010 non-null   int64
 1   pH               2010 non-null   float64
 2   Hardness         2010 non-null   float64
 3   Solids           2010 non-null   float64
 4   Chloramines      2010 non-null   float64
 5   Sulfate          2010 non-null   float64
 6   Conductivity     2010 non-null   float64
 7   OrganicCarbon    2010 non-null   float64
 8   Trihalomethanes  2010 non-null   float64
 9   Turbidity        2010 non-null   float64
 10  Potability       2010 non-null   int64
dtypes: float64(9), int64(2)
memory usage: 172.9 KB
```

1.3 Nomor 1: Deskripsi Statistika

Pada nomor 1 ini, kami mencari 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.

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

[ ]:

	id	pH	Hardness	Solids	Chloramines	\
count	2010.00000	2010.000000	2010.000000	2010.000000	2010.000000	
mean	1005.50000	7.087193	195.969209	21904.673439	7.134322	
std	580.38134	1.572803	32.643166	8625.397911	1.585214	
min	1.00000	0.227499	73.492234	320.942611	1.390871	
25%	503.25000	6.090785	176.740657	15614.412962	6.138326	
50%	1005.50000	7.029490	197.203525	20926.882155	7.142014	
75%	1507.75000	8.053006	216.447589	27170.534649	8.109933	

```
max    2010.00000    14.000000   317.338124  56488.672413    13.127000

           Sulfate  Conductivity  OrganicCarbon  Trihalomethanes    Turbidity  \
count  2010.000000   2010.000000    2010.000000      2010.000000  2010.000000
mean    333.211376    426.476708      14.357940        66.400717     3.969497
std      41.211111     80.701872       3.325770        16.081109     0.780471
min     129.000000    201.619737       2.200000         8.577013     1.450000
25%     307.626986    366.619219      12.122530        55.949993     3.442882
50%     332.214113    423.438372      14.323286        66.482041     3.967374
75%     359.268147    482.209772      16.683562        77.294613     4.514663
max     481.030642    753.342620      27.006707       124.000000     6.494749

        Potability
count  2010.000000
mean      0.402985
std       0.490620
min       0.000000
25%       0.000000
50%       0.000000
75%       1.000000
max       1.000000
```

Data di atas menampilkan rata-rata (ditunjukkan dengan mean), median (ditunjukkan dengan baris 50%), standar deviasi (ditunjukkan dengan std), nilai minimum (ditunjukkan dengan min), nilai maksimum (ditunjukkan dengan max), dan kuartil (ditunjukkan dengan 25% (Q1), 50% (Q2), dan 75% (Q3)).

Selanjutnya akan dicari nilai variansi untuk setiap kolom pada dataset `water_potability.csv` tersebut

[ ]:
```
data.var()
```

[ ]:
```
id                3.368425e+05
pH                2.473709e+00
Hardness          1.065576e+03
Solids            7.439749e+07
Chloramines       2.512904e+00
Sulfate           1.698356e+03
Conductivity      6.512792e+03
OrganicCarbon     1.106075e+01
Trihalomethanes   2.586021e+02
Turbidity         6.091350e-01
Potability        2.407079e-01
dtype: float64
```

Selanjutnya, akan dicari nilai range untuk setiap kolom pada dataset `water_potability.csv` tersebut

[ ]:
```
data.max() - data.min()
```

[ ]:
```
id                 2009.000000
pH                   13.772501
Hardness            243.845890
Solids            56167.729801
Chloramines          11.736129
Sulfate             352.030642
Conductivity        551.722883
OrganicCarbon        24.806707
Trihalomethanes     115.422987
Turbidity             5.044749
Potability            1.000000
dtype: float64
```

Selanjutnya akan dicari nilai IQR untuk setiap kolom pada dataset `water_potability.csv` tersebut

[ ]:
```
q1 = data.quantile(0.25)
q3 = data.quantile(0.75)
q3 - q1
```

[ ]:
```
id                 1004.500000
pH                    1.962221
Hardness             39.706932
Solids            11556.121687
Chloramines           1.971607
Sulfate              51.641161
Conductivity        115.590553
OrganicCarbon         4.561031
Trihalomethanes      21.344620
Turbidity             1.071781
Potability            1.000000
dtype: float64
```

Selanjutnya akan dicari nilai skewness untuk setiap kolom pada dataset `water_potability.csv` tersebut

[ ]:
```
data.skew()
```

[ ]:
```
id                0.000000
pH                0.048535
Hardness         -0.085321
Solids            0.591011
Chloramines       0.013003
Sulfate          -0.045728
Conductivity      0.268012
OrganicCarbon    -0.020220
```

```
Trihalomethanes   -0.051383
Turbidity         -0.032266
Potability         0.395873
dtype: float64
```

Selanjutnya ditentukan nilai kurtosis untuk setiap kolom pada dataset `water_potability.csv` tersebut

[ ]:
```
data.kurtosis()
```

[ ]:
```
id                -1.200000
pH                 0.626904
Hardness           0.525480
Solids             0.337320
Chloramines        0.549782
Sulfate            0.786854
Conductivity      -0.237206
OrganicCarbon      0.031018
Trihalomethanes    0.223017
Turbidity         -0.049831
Potability        -1.845122
dtype: float64
```

Selanjutnya akan dicari nilai modus untuk setiap kolom pada dataset `water_potability.csv` tersebut

[ ]:
```
data.mode()
```

[ ]:
```
         id         pH    Hardness        Solids  Chloramines     Sulfate  \
0         1   0.227499   73.492234    320.942611     1.390871  129.000000
1         2   0.989912   77.459586   1198.943699     1.920271  180.206746
2         3   1.431782   81.710895   1351.906979     2.397985  182.397370
3         4   1.757037   94.091307   1372.091043     2.456014  187.170714
4         5   1.985383   94.812545   2552.962804     2.458609  187.424131
…     …         …          …           …            …         …
2005  2006  11.568768  286.567991  50793.898917    12.580026  458.441072
2006  2007  11.898078  287.975540  53735.899194    12.626900  460.107069
2007  2008  12.246928  300.292476  55334.702799    12.653362  475.737460
2008  2009  13.349889  306.627481  56351.396304    13.043806  476.539717
2009  2010  14.000000  317.338124  56488.672413    13.127000  481.030642

      Conductivity  OrganicCarbon  Trihalomethanes  Turbidity  Potability
0       201.619737       2.200000         8.577013   1.450000         0.0
1       210.319182       4.371899        14.343161   1.492207         NaN
2       233.907965       4.466772        15.684877   1.496101         NaN
3       245.859632       4.861631        16.291505   1.680554         NaN
4       252.968328       4.966862        17.527765   1.812529         NaN
…            …            …              …        …         …
```

```
2005    666.690618    23.569645    114.034946    6.307678        NaN
2006    669.725086    23.604298    114.208671    6.357439        NaN
2007    695.369528    23.917601    116.161622    6.389161        NaN
2008    708.226364    24.755392    120.030077    6.494249        NaN
2009    753.342620    27.006707    124.000000    6.494749        NaN

[2010 rows x 11 columns]
```

[ ]:
```
data.shape
```

[ ]:
```
(2010, 11)
```

Pada data di atas, terlihat bahwa nilai modus pada kolom selain kolom *portability* memiliki nilai lebih dari satu. Lebih jauh lagi, setiap kolom numerik selain kolom *portability* memiliki data yang unik sehingga semua nilai merupakan nilai modus.

1.4 Nomor 2: Visualisasi

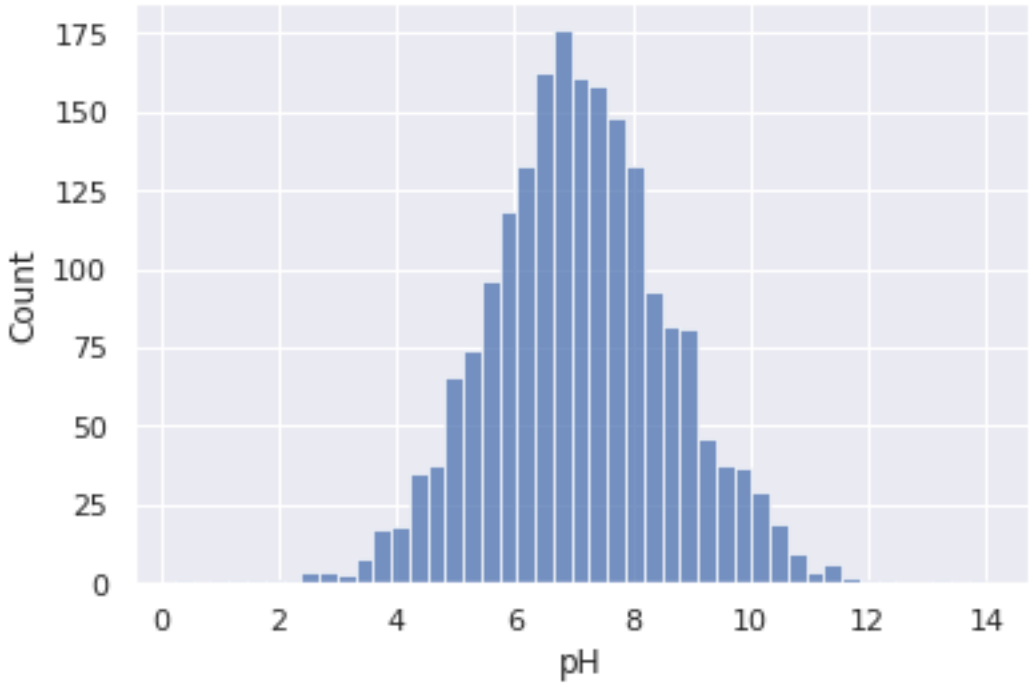
Pada nomor ini, akan ditampilkan visualisasi distribusi plot untuk setiap kolom numerik

1.4.1 Data pH

Berikut ini adalah histogram untuk data pH pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="pH")
```

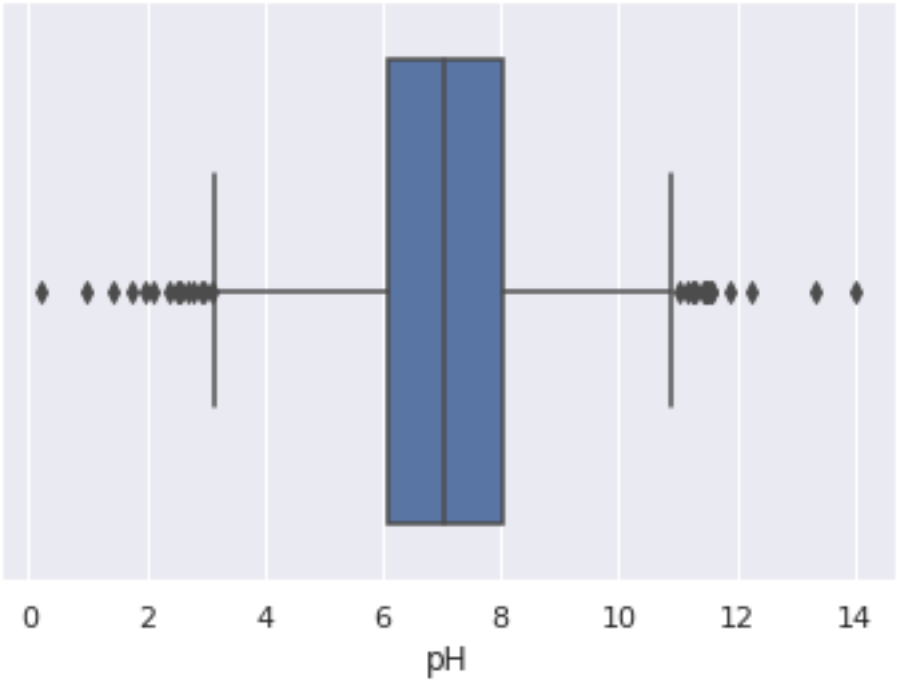
[ ]:
```
<AxesSubplot:xlabel='pH', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data pH pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "pH")
```

[ ]:
```
<AxesSubplot:xlabel='pH'>
```

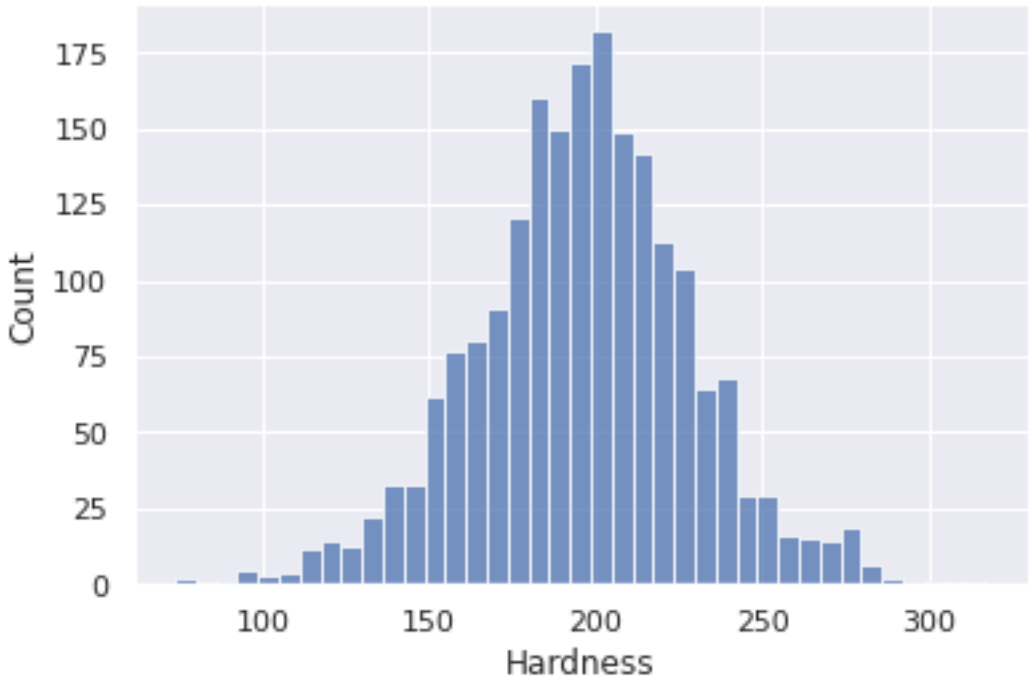


1.4.2 Data Hardness

Berikut ini adalah histogram untuk data Hardness pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="Hardness")
```

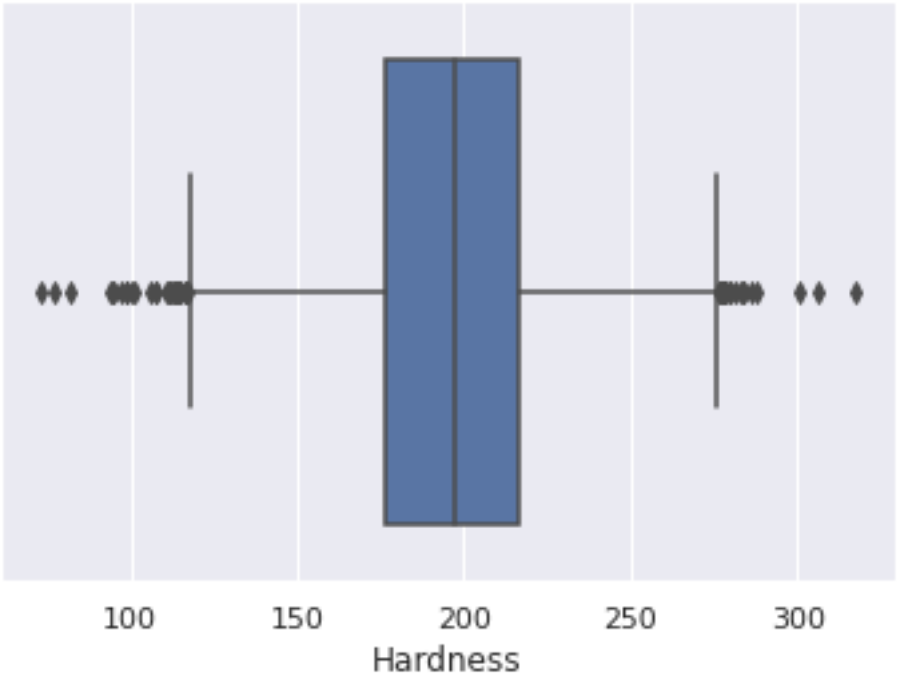
[ ]:
```
<AxesSubplot:xlabel='Hardness', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data Hardness pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "Hardness")
```

[ ]:
```
<AxesSubplot:xlabel='Hardness'>
```

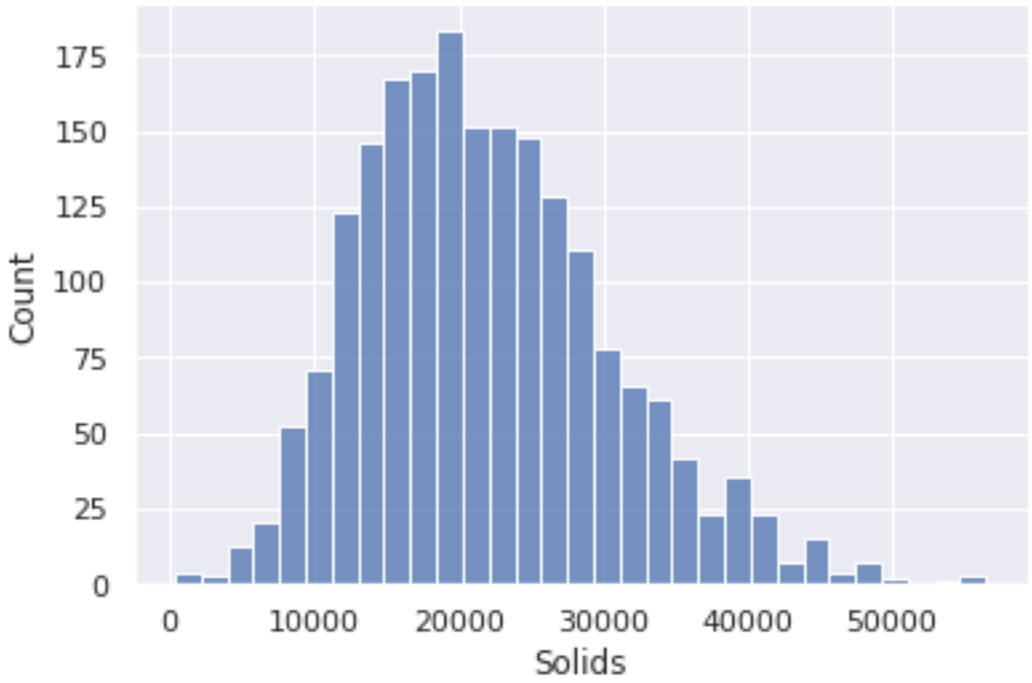


1.4.3 Data Solids

Berikut ini adalah histogram untuk data Solids pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="Solids")
```

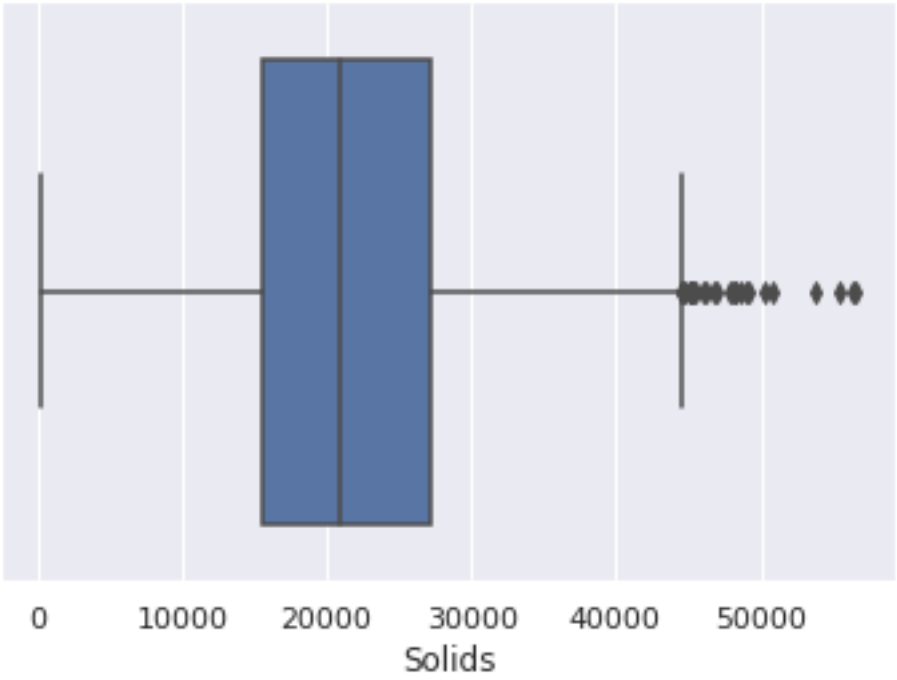
[ ]:
```
<AxesSubplot:xlabel='Solids', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data Solids pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "Solids")
```

[ ]:
```
<AxesSubplot:xlabel='Solids'>
```

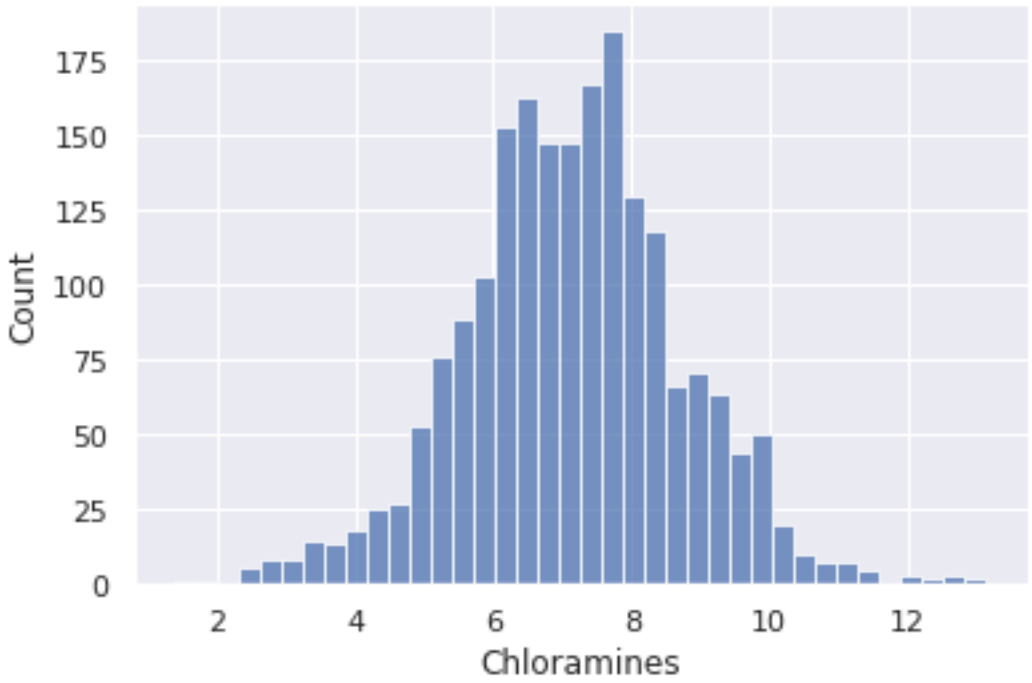


1.4.4 Data Chloramines

Berikut ini adalah histogram untuk data Chloramines pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="Chloramines")
```

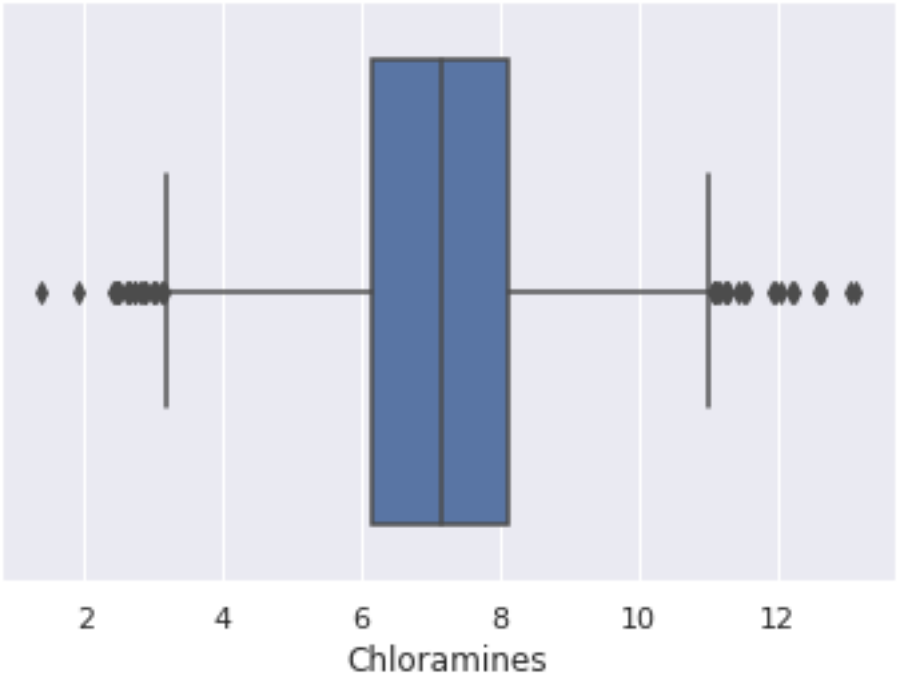
[ ]:
```
<AxesSubplot:xlabel='Chloramines', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data Chloramines pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "Chloramines")
```

[ ]:
```
<AxesSubplot:xlabel='Chloramines'>
```

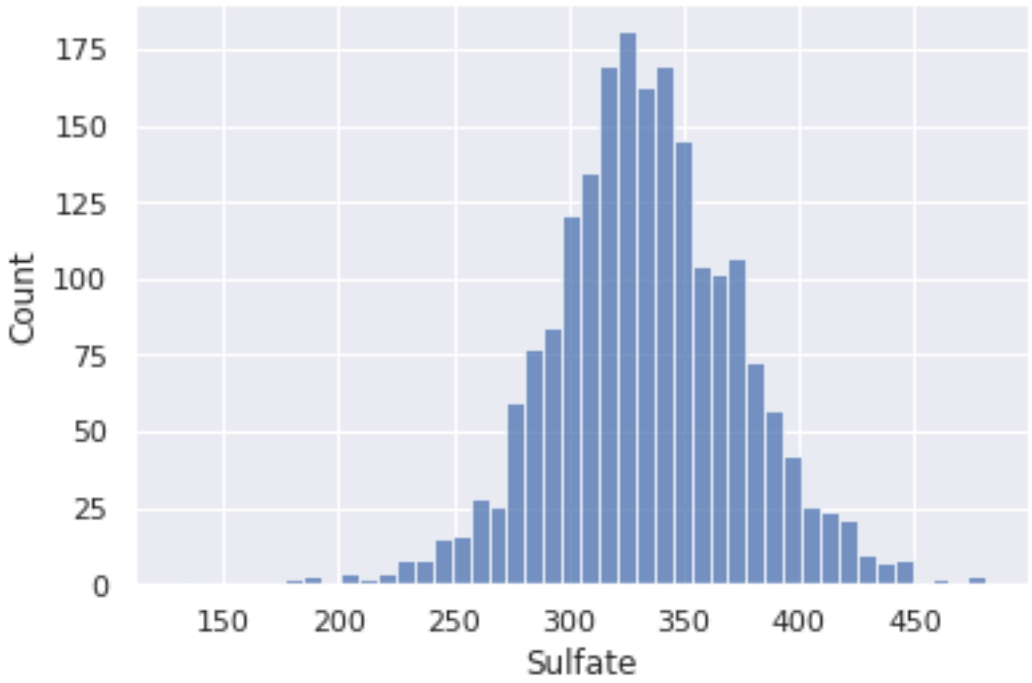


1.4.5 Data Sulfate

Berikut ini adalah histogram untuk data Sulfate pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="Sulfate")
```

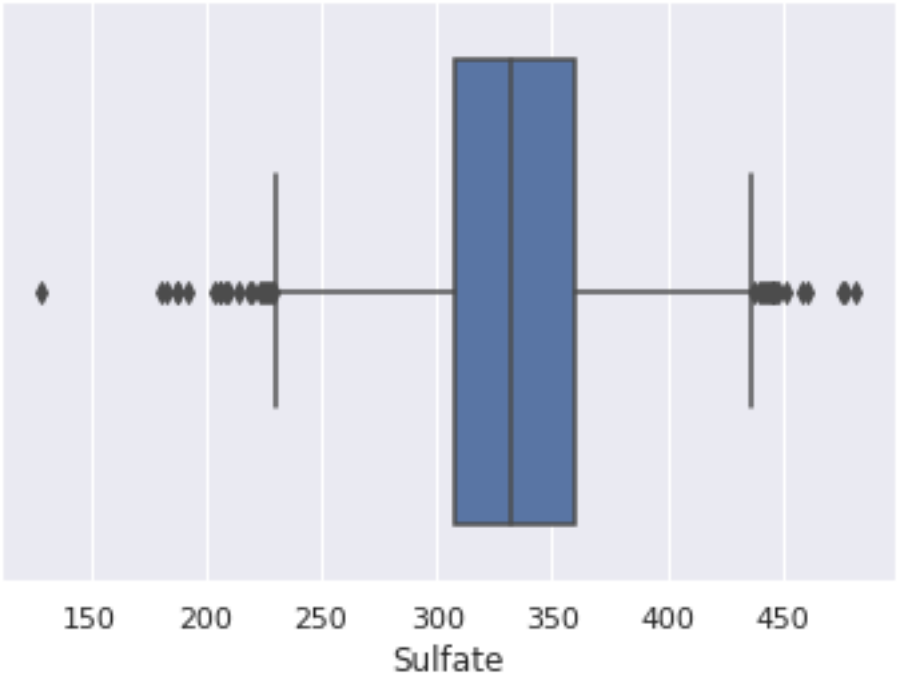
[ ]:
```
<AxesSubplot:xlabel='Sulfate', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data Sulfate pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "Sulfate")
```

[ ]:
```
<AxesSubplot:xlabel='Sulfate'>
```

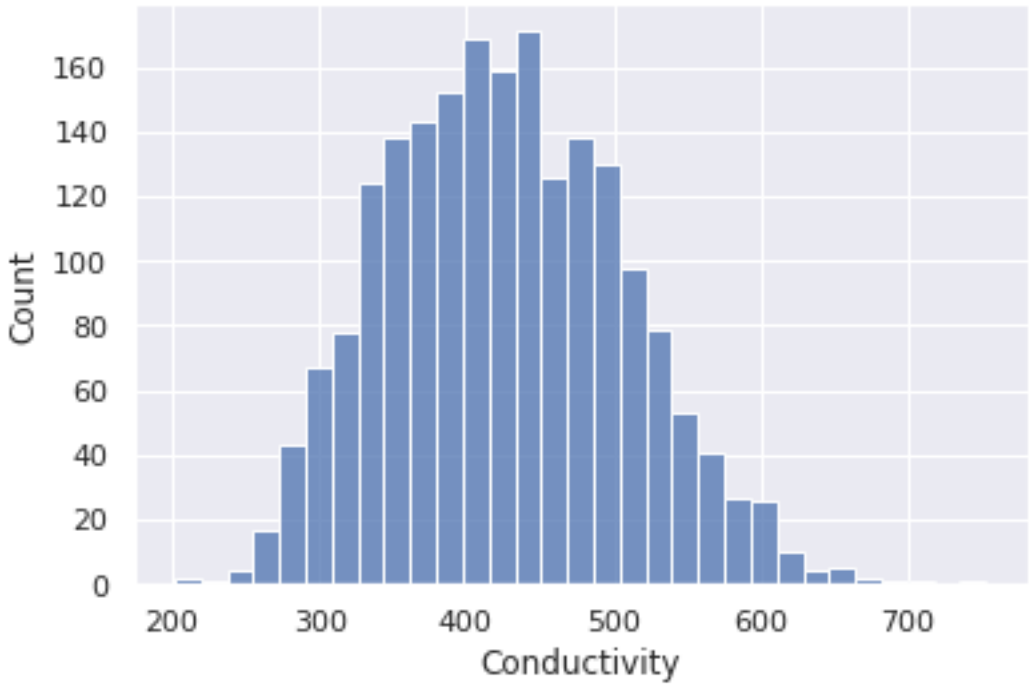


1.4.6 Data Conductivity

Berikut ini adalah histogram untuk data Conductivity pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="Conductivity")
```

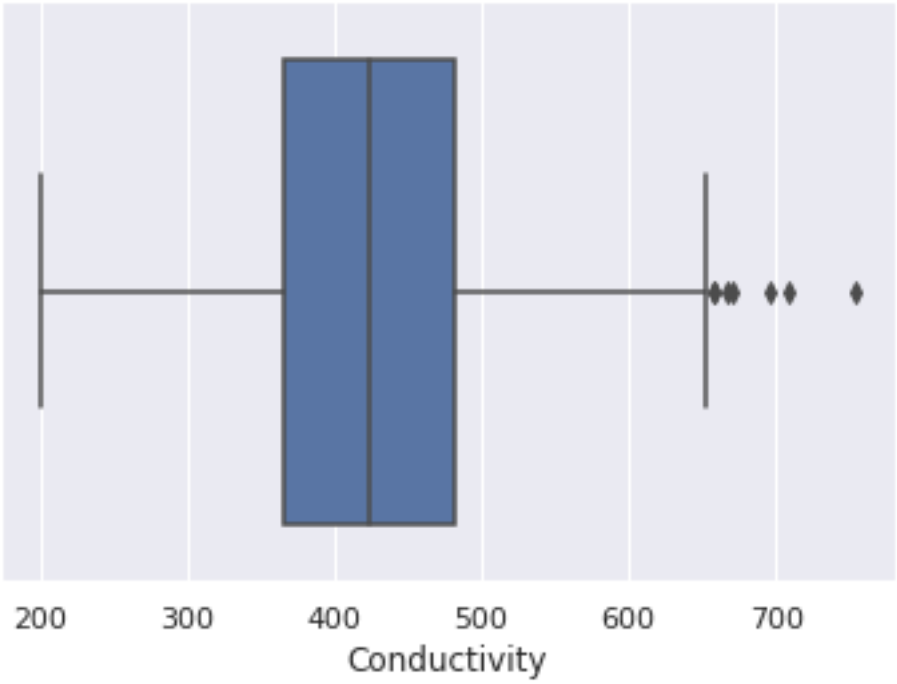
[ ]:
```
<AxesSubplot:xlabel='Conductivity', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data Conductivity pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "Conductivity")
```

[ ]:
```
<AxesSubplot:xlabel='Conductivity'>
```

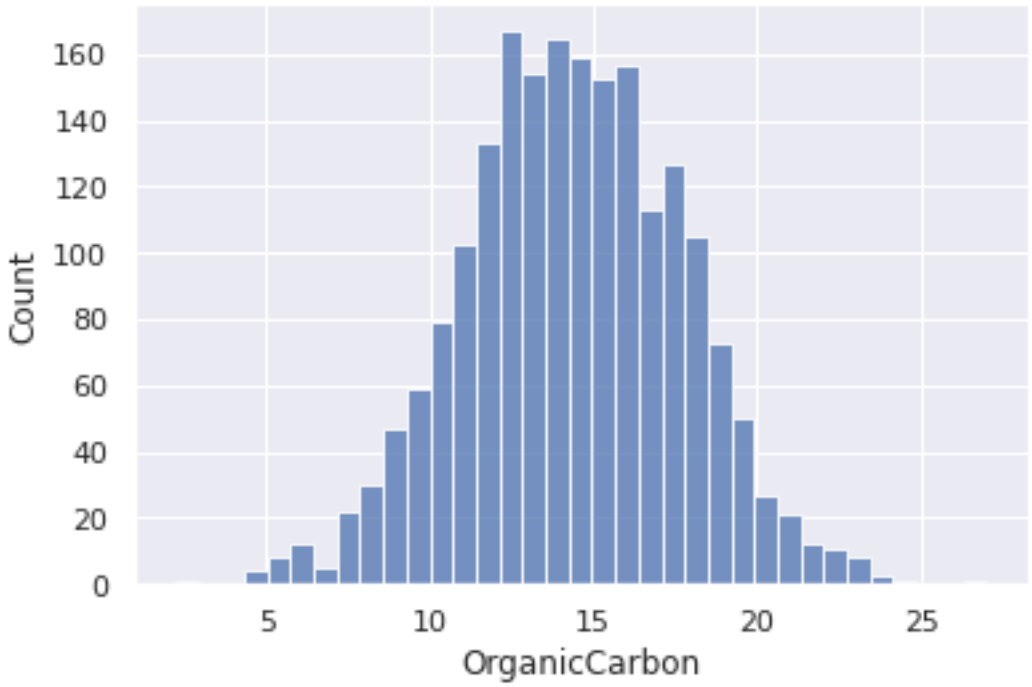


1.4.7 Data OrganicCarbon

Berikut ini adalah histogram untuk data OrganicCarbon pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="OrganicCarbon")
```

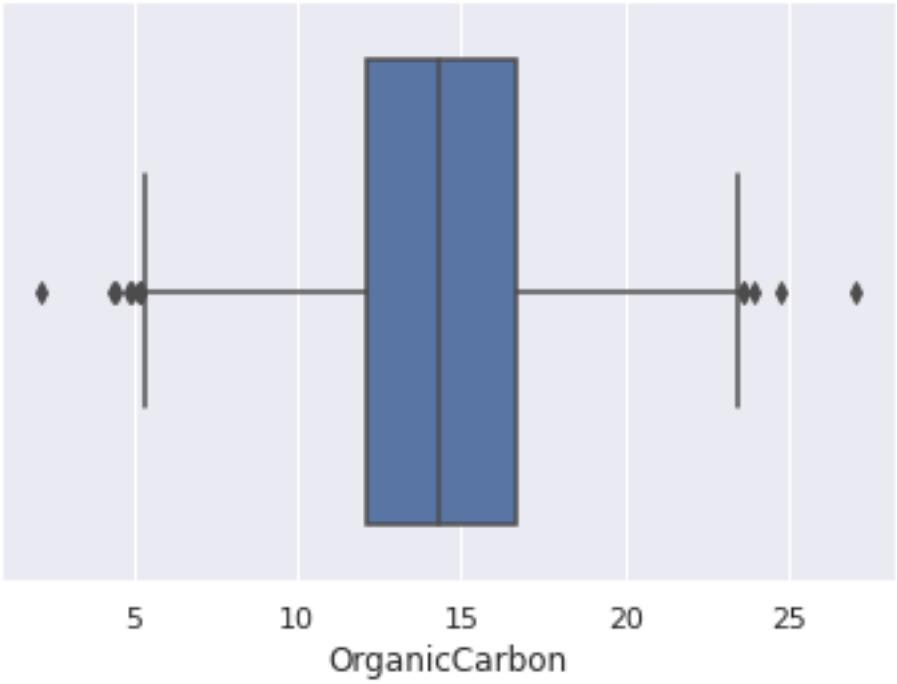
[ ]:
```
<AxesSubplot:xlabel='OrganicCarbon', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data OrganicCarbon pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "OrganicCarbon")
```

[ ]:
```
<AxesSubplot:xlabel='OrganicCarbon'>
```

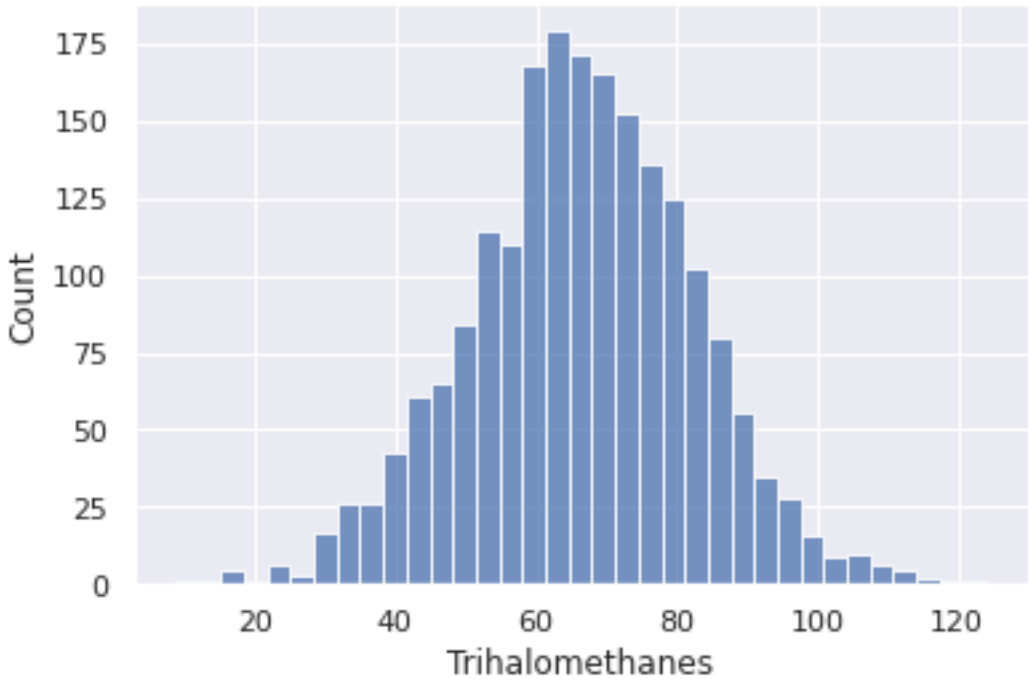


1.4.8 Data Trihalomethanes

Berikut ini adalah histogram untuk data Trihalomethanes pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="Trihalomethanes")
```

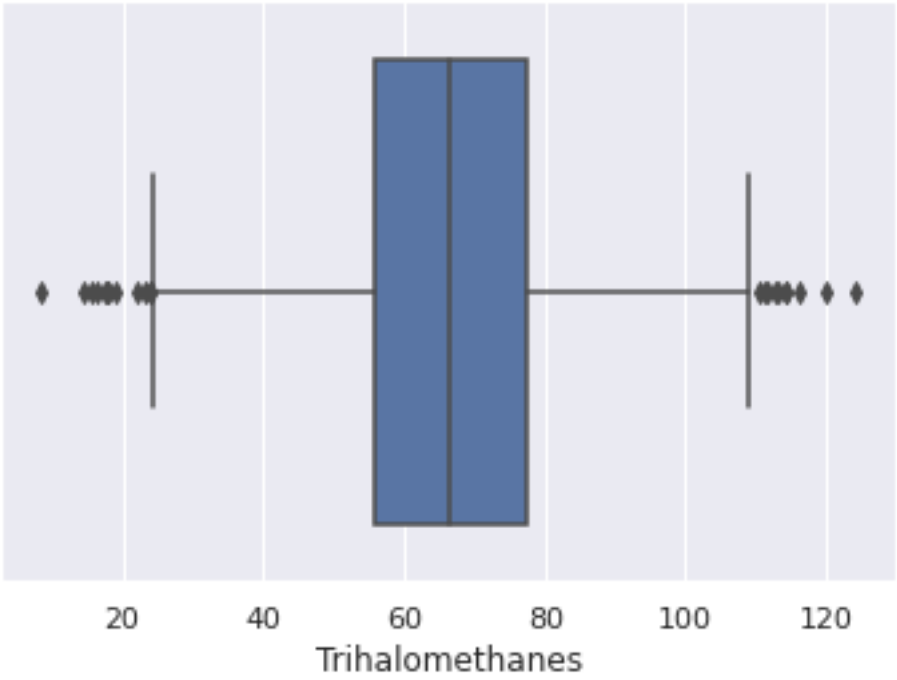
[ ]:
```
<AxesSubplot:xlabel='Trihalomethanes', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data Trihalomethanes pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "Trihalomethanes")
```

[ ]:
```
<AxesSubplot:xlabel='Trihalomethanes'>
```

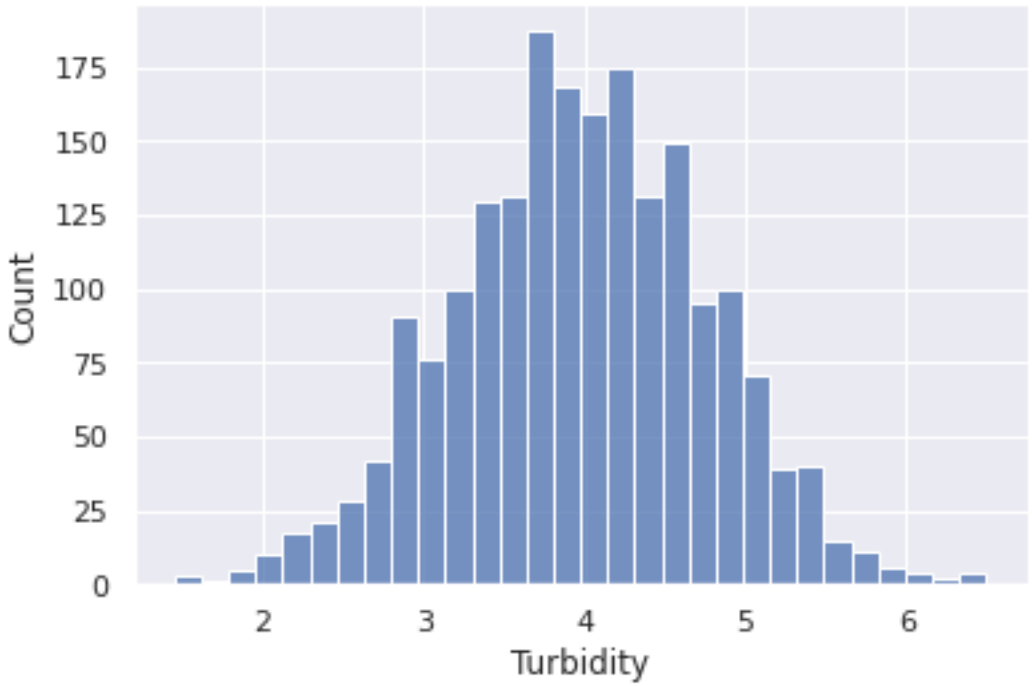


1.4.9 Data Turbidity

Berikut ini adalah histogram untuk data Turbidity pada dataset `water_portability.csv`

[ ]:
```
sns.histplot(data,x="Turbidity")
```

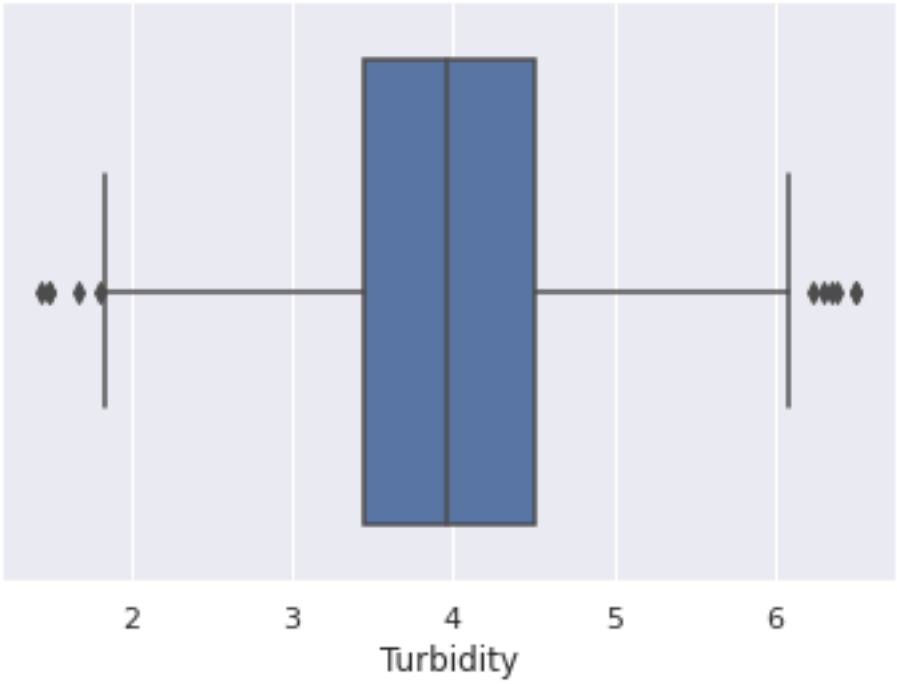
[ ]:
```
<AxesSubplot:xlabel='Turbidity', ylabel='Count'>
```



Berikut ini adalah boxplot untuk data Turbidity pada dataset `water_portability.csv`

[ ]:
```
sns.boxplot(data = data, x = "Turbidity")
```

[ ]:
```
<AxesSubplot:xlabel='Turbidity'>
```



1.5 Nomor 3: Tes Distribusi Normal

Pada bagian ini, akan dites apakah setiap kolom berdistribusi normal atau tidak. Kolom yang akan dianalisis adalah kolom numerik, yaiur kolom 2 sampai dengan kolom 10.

1.5.1 Metode Tes

Metode pengetesan akan dilakukan dengan dua cara, yaitu metode grafik dan statistik.

Metode Grafik Pada metode grafik, kami akan menggunakan QQ Plot dengan histogram. Pada tahap ini kami hanya mengamati seberapa dekat suatu kolom dengan normalnya.

Pembuatan grafik QQ dapat dilakukan dengan menjadikan setiap data merupakan quantiles dari semua data. Setelah itu, setiap quantiles dihitung korespondensinya terhadap tabel normal. Setelah itu akan dilakukan plotting menggunakan scatter plot dan dibuat regresinya. Apabila kebanyakan titik berada pada garis, maka data berdistribusi normal.

Berikut ini adalah fungsi yang akan membantu membuat QQ Plot

[ ]:
```
def QQ_Plot(data):
  dataset = np.sort(data)
  norm = scipy.stats.norm()
  normalDataset = np.array([
    norm.ppf((i+0.5)/len(dataset)) for i in range(len(dataset))
  ])
```

```
sns.regplot(x=normalDataset, y=dataset)
plt.xlabel("Normal Quantiles")
plt.ylabel("Data Quantiles")
```

Metode Statistik Pada metode statistik, kami menggunakan D'Agostino-Pearson Omnibus test untuk pengujian statistik. Pengetesan akan dilakukan dengan menggunakan pengujian hipotestis.

Berikut ini adalah hipotesinya: 1. Hipotesis nol (H_0) dari pengetesan ini adalah kolom berdistribusi normal. 2. Hipotesis slternatif (H_1) dari pengetesan ini adalah kolom tidak berdistribusi normal.

Tingkat signifikansi yang digunakan adalah $\alpha = 0.05$

[ ]:
```
alpha = 0.05
```

Berikut ini adalah langkah pengujian statistik yang dilakukan: 1. Kurtosis dan juga skewness dari sebuah kolom perlu dihitung terlebih dahulu. 2. Menghitung error standard untuk skewness. Rumus untuk perhitungan skewness standard error adalah sebagai berikut:

$$s.e = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

3. Menghitung error standar untuk kurtosis. Rumus untuk melakukan perhitungan ini adalah sebagai berikut:

$$k.e = 2 \cdot (s.e) \cdot \sqrt{\frac{n^2-1}{(n-3)(n+5)}}$$

4. Perlu dihtung standar score untu skewness. Berikut ini adalah rumusnya:

$$z_s = \frac{Sk}{s.e}$$

5. Perlu dihitung standar error untuk kurtosis. Berikut ini adalah rumusnya:

$$z_k = \frac{Kur}{k.e}$$

6. Jumlah kuadrat dari Nilai dari standar skor untuk skewness dan kurtosis dapat didekatkan dengan distribusi chi-square derajat dua.

$$z_x^2 + z_k^2 \approx \chi_\alpha^2$$

Oleh karena itu, nilai p dapat dihitung dengan mencari distribusi dari chi-square berderajat 2.

Proses diatas dapat dilakukan dengan menggunakan library dari scipy, yaitu `scipy.stat.normaltest`.

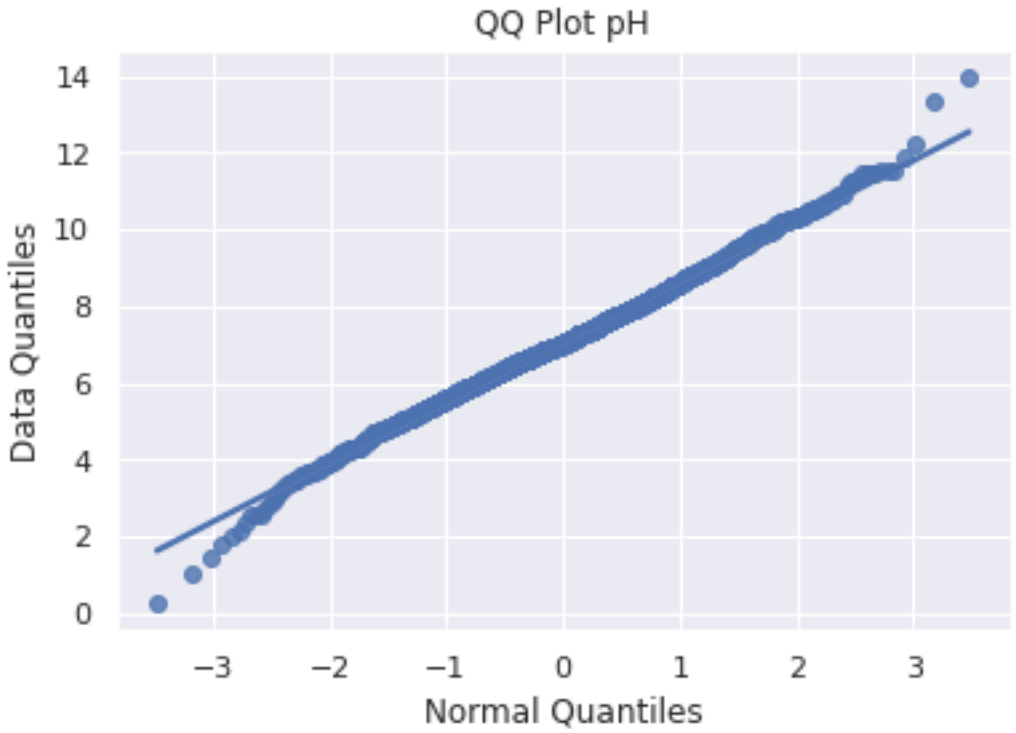
Pada langkah terakhir, akan diperiksa apakah nilai p kurang dari level signifikansi. Bila kurang, maka hipotesis H_0 dapat ditolak.

1.5.2 Data pH

Pada bagian ini, akan dicoba untuk melakukan test normal pada data pH. Berikut ini adalah histogram dan juga QQ plot dari data pH.

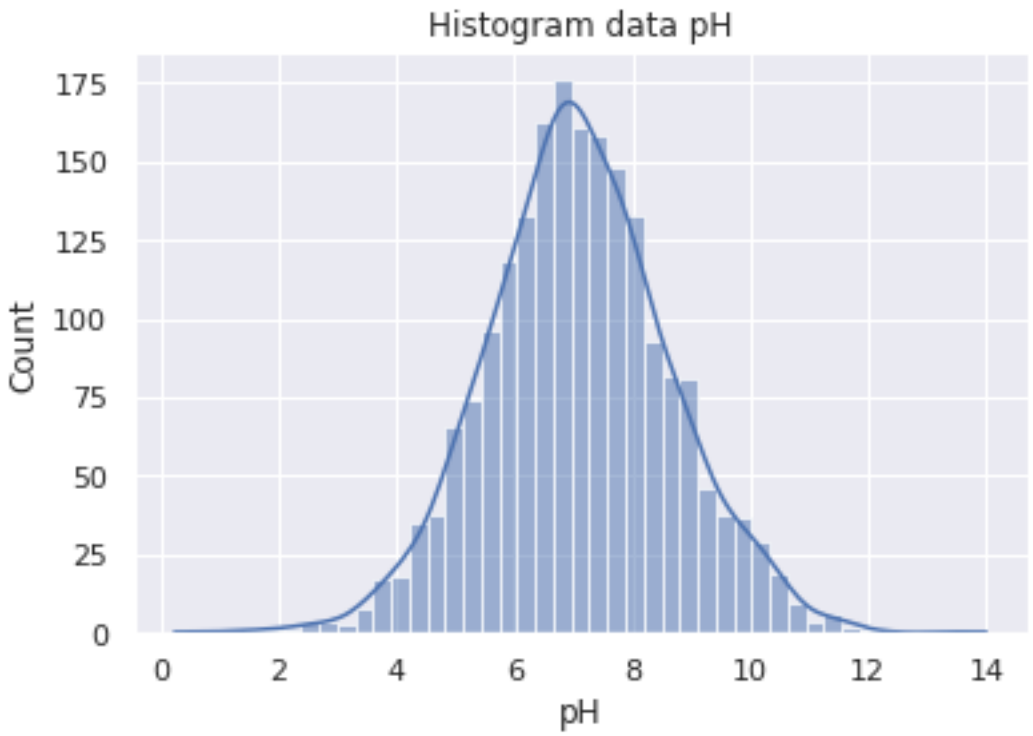
[ ]:
```
QQ_Plot(data["pH"])
plt.title("QQ Plot pH")
```

[ ]:
```
Text(0.5, 1.0, 'QQ Plot pH')
```



[ ]:
```
sns.histplot(data=data, x="pH", kde=True)
plt.title("Histogram data pH")
```

[ ]:
```
Text(0.5, 1.0, 'Histogram data pH')
```



Dari kedua grafik diatas, data pH terlihat data bisa jadi tidak berdistribusi normal. Hal ini terlihat pada ujung kiri dan ujung kanan QQ Plot yang menjauh dari garis.

Pada bagian selanjutnya, data akan diuji menggunakan pendekatan statistik.

[ ]:
```
_, p = scipy.stats.normaltest(data["pH"])
print(f"p = {p}")

if p < alpha:
  print("Data tidak berdistribusi normal")
else:
  print("Data berdistribusi normal")
```

```
p = 2.6514813346797777e-05
Data tidak berdistribusi normal
```

Berdasarkan pengujian statistik, terlihat bahwa data tidak berdistribusi normal. Hal ini dikarenakan nilai $p < 0.05$ sehingga hipotesis H_0 dapat ditolak.

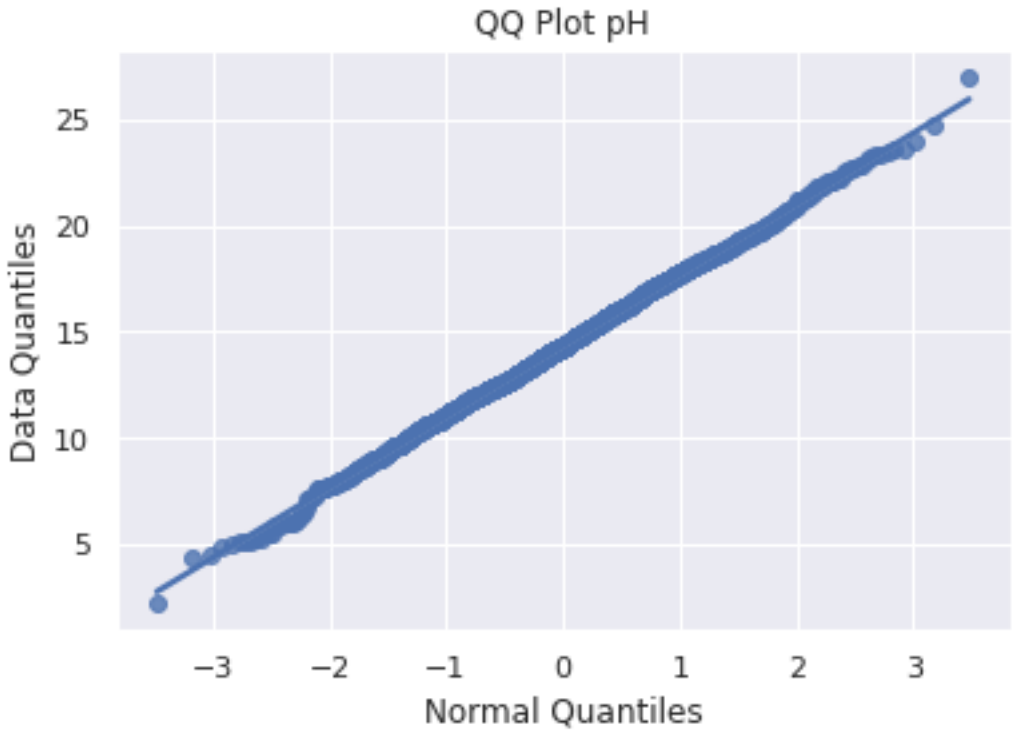
Kesimpulan dari pengujian ini adalah data pH bukan merupakan data yang berdistribusi normal

1.5.3 Data OrganicCarbon

Pada bagian ini, akan dicoba untuk melakukan test normal pada data OrganicCarbon. Berikut ini adalah histogram dan juga QQ plot dari data pH.

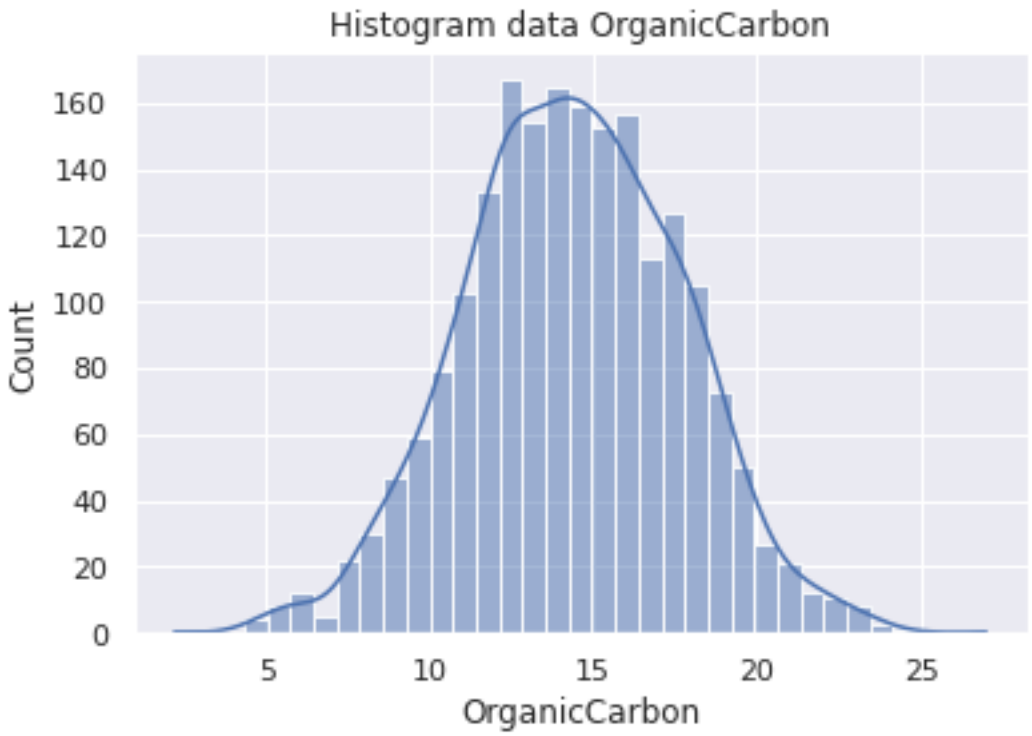
[ ]:
```
QQ_Plot(data["OrganicCarbon"])
plt.title("QQ Plot pH")
```

[ ]:
```
Text(0.5, 1.0, 'QQ Plot pH')
```



[ ]:
```
sns.histplot(data=data, x="OrganicCarbon", kde=True)
plt.title("Histogram data OrganicCarbon")
```

[ ]:
```
Text(0.5, 1.0, 'Histogram data OrganicCarbon')
```



Dari kedua grafik diatas, data OrganicCarbon terlihat mendekati bentuk normal. Hal ini dapat terlihat bahwa pada QQ plot, sebagian besar titik berada pada garis. Oleh karena itu, dapat disimpulkan bahwa pH merupakan data yang berkemungkinan berdistribusi normal.

Pada bagian selanjutnya, data akan diuji menggunakan pendekatan statistik.

[ ]:
```
_, p = scipy.stats.normaltest(data["OrganicCarbon"])
print(f"p = {p}")

if p < alpha:
  print("Data tidak berdistribusi normal")
else:
  print("Data berdistribusi normal")
```

```
p = 0.8825496581408284
Data berdistribusi normal
```

Berdasarkan pengujian statistik, terlihat bahwa berdistribusi normal. Hal ini ditunjukan bahwa nilai $p > 0.05$. Oleh karena itu, hipotesis H_0 tidak dapat ditolak.

Kesimpulan dari pengujian ini adalah data OrganicCarbon bukan merupakan data yang berdistribusi normal

1.6 Nomor 6: Korelasi

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

[ ]:
```
                      id        pH  Hardness    Solids  Chloramines  \
id              1.000000 -0.031175 -0.014818 -0.021336     0.004946
pH             -0.031175  1.000000  0.108959 -0.085582    -0.024767
Hardness       -0.014818  0.108959  1.000000 -0.053282    -0.022684
Solids         -0.021336 -0.085582 -0.053282  1.000000    -0.051933
Chloramines     0.004946 -0.024767 -0.022684 -0.051933     1.000000
Sulfate         0.052322  0.011028 -0.108509 -0.164106     0.006248
Conductivity   -0.034291  0.015089  0.011778 -0.007045    -0.028300
OrganicCarbon   0.035022  0.028285  0.013219 -0.005290    -0.023806
Trihalomethanes -0.026509  0.018302 -0.015400 -0.015729     0.014990
Turbidity       0.024003 -0.035416 -0.034813  0.018569     0.013132
Potability      0.122027  0.015475 -0.001463  0.038977     0.020779

                    Sulfate  Conductivity  OrganicCarbon  Trihalomethanes  \
id                 0.052322     -0.034291       0.035022        -0.026509
pH                 0.011028      0.015089       0.028285         0.018302
Hardness          -0.108509      0.011778       0.013219        -0.015400
Solids            -0.164106     -0.007045      -0.005290        -0.015729
Chloramines        0.006248     -0.028300      -0.023806         0.014990
Sulfate            1.000000     -0.016600       0.026823        -0.023355
Conductivity      -0.016600      1.000000       0.015739         0.004879
OrganicCarbon      0.026823      0.015739       1.000000        -0.005666
Trihalomethanes   -0.023355      0.004879      -0.005666         1.000000
Turbidity         -0.010129      0.012133      -0.015388        -0.020504
Potability        -0.015703     -0.016257      -0.015488         0.009237

                  Turbidity  Potability
id                 0.024003    0.122027
pH                -0.035416    0.015475
Hardness          -0.034813   -0.001463
Solids             0.018569    0.038977
Chloramines        0.013132    0.020779
Sulfate           -0.010129   -0.015703
Conductivity       0.012133   -0.016257
OrganicCarbon     -0.015388   -0.015488
Trihalomethanes   -0.020504    0.009237
Turbidity          1.000000    0.022331
Potability         0.022331    1.000000
```

[ ]: