Tugas Besar IF2220 Probabilitas dan Statistika: Penarikan Kesimpulan dan Pengujian Hipotesis

Nomor 3

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

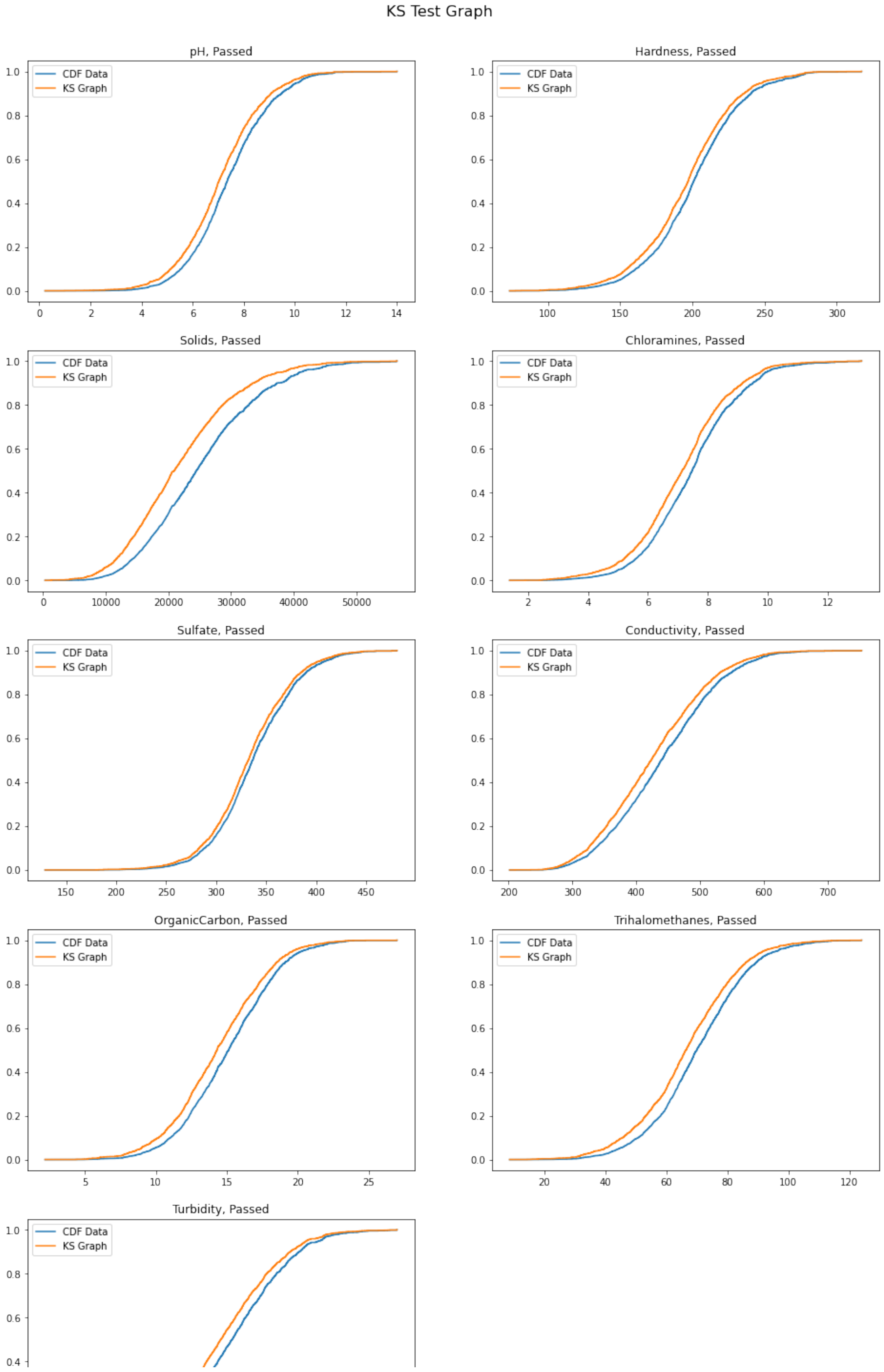
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm, kstest
import math

# Membaca data dan melihat beberapa data pertama
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'
]
```

Testing menggunakan Kolmogorov-Smirnov Test

```
def graph_ks_test(ax, data):

    sorted_data = sorted(data)
    data_num = len(data)
    data_sum = data.sum()
    s = [(x + 1) / data_num for x in range(data_num)]

    #fig, ax = plt.subplots()
    ax.step(
        sorted_data,
        [sum(sorted_data[:(i + 1)]) / data_sum for i in 
range(data_num)],
        where='post',
        label='CDF Data'
    )
    ax.step(
        sorted_data,
        s,
        where='post',
        label='KS Graph'
)

def graph_all_ks_test(df, col_names):
    # Hitung banyak kolom data dan bagi subplot sehingga memiliki 2 
kolom
    col_num = len(col_names)
    fig, axs = plt.subplots(col_num // 2 + 1, 2, figsize=(16, 3 * 
col_num))
    fig.suptitle('KS Test Graph', y=0.91, fontsize=16)
    # Isi tiap subplot
    for idx, col_name in enumerate(col_names):
        _col = idx // 2
        _row = idx % 2
        # Hasil KS Test
        _, p_val = kstest(df[col_name], 'norm')
        graph_ks_test(
            axs[_col][_row],
            df[col_name]
        )
        # Di judulnya, tulis passed jika p_val kurang dari 0.05
        axs[_col][_row].set_title(f"{col_name}, {'Passed' if p_val < 
0.05 else 'Failed'}")
        axs[_col][_row].legend()
    # Jika jumlah kolomnya ganjil, perlu dihapus plot terakhir
    if col_num % 2 == 1:
        fig.delaxes(axs[col_num // 2][1])

    return fig, axs

fig, ax, = graph_all_ks_test(df[col_numeric], col_numeric)
```



KS Test Graph

pH, Passed

Hardness, Passed

Solids, Passed

Chloramines, Passed

Sulfate, Passed

Conductivity, Passed

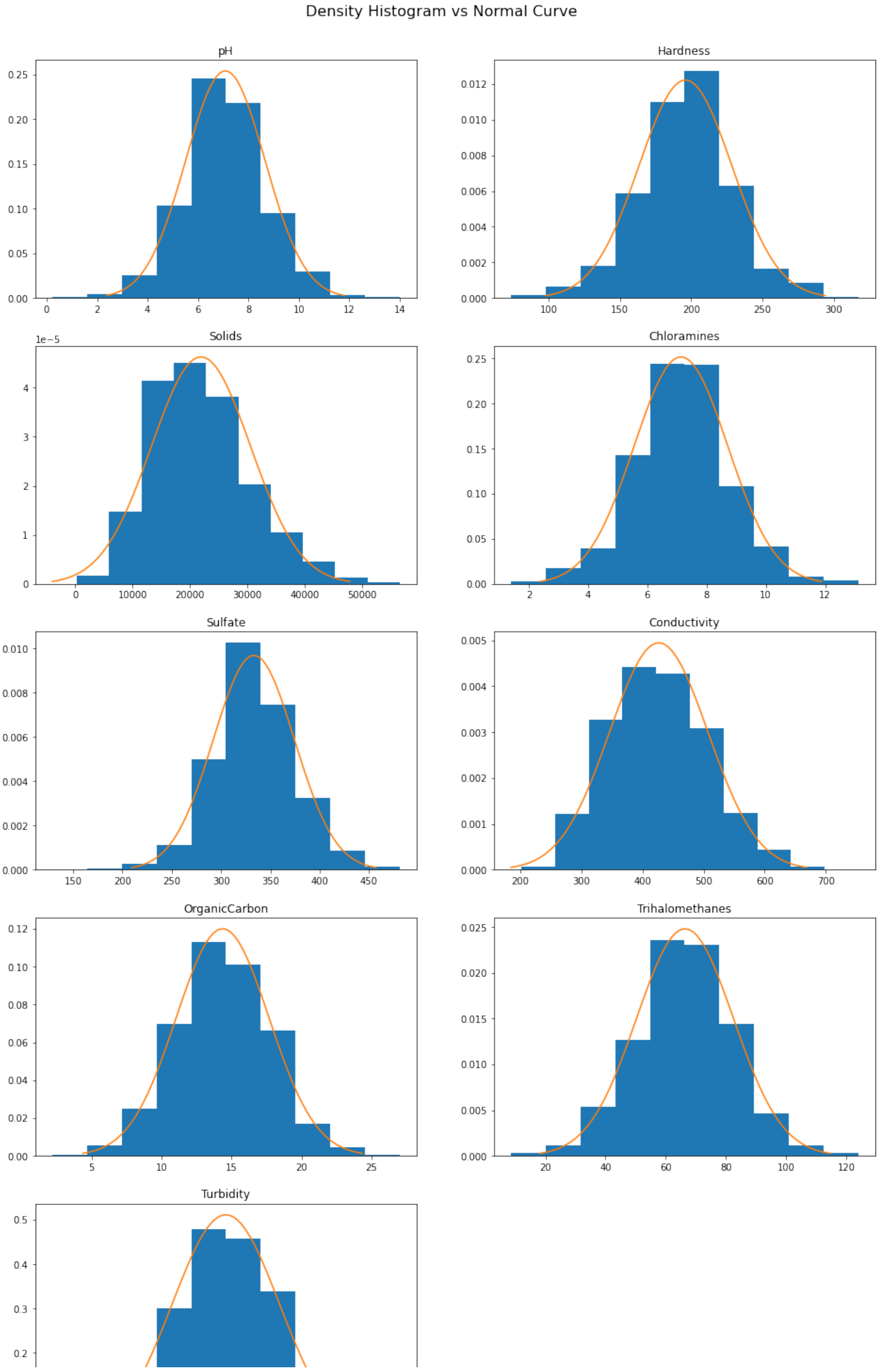
OrganicCarbon, Passed

Trihalomethanes, Passed

Turbidity, Passed

Testing Secara Visual dengan Membandingkan *Graph*

```
def histogram_vs_normal(axs, data):
    # Density Histogram
    axs.hist(data, density=True)
    _mean, _std = norm.fit(data)
    x_axis = np.linspace(_mean - 3 * _std, _mean + 3 * _std, 100)
    # Normal Curve
    axs.plot(x_axis, norm.pdf(x_axis, _mean, _std))

def histogram_vs_normal_col_name(col_names, df):
    # Hitung banyak kolom data dan bagi subplot sehingga memiliki 2 
kolom
    col_num = len(col_names)
    fig, axs = plt.subplots(col_num // 2 + 1, 2, figsize=(16, 3 * 
col_num))
    #fig.suptitle('Density Histogram vs Normal Curve', fontsize=16)
    fig.suptitle('Density Histogram vs Normal Curve', y=0.91, 
fontsize=16)
    # Isi tiap subplot
    for idx, col_name in enumerate(col_names):
        _col = idx // 2
        _row = idx % 2
        histogram_vs_normal(
            axs[_col][_row],
            df[col_name]
        )
        axs[_col][_row].set_title(col_name)
    # Jika jumlah kolomnya ganjil, perlu dihapus plot terakhir
    if col_num % 2 == 1:
        fig.delaxes(axs[col_num // 2][1])
    return fig, axs

fig, axs = histogram_vs_normal_col_name(col_numeric, df[col_numeric])
```


Density Histogram vs Normal Curve
pH
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
14
Hardness
0.000
0.002
0.004
0.006
0.008
0.010
0.012
100
150
200
250
300
Solids
1e−5
0
1
2
3
4
0
10000
20000
30000
40000
50000
Chloramines
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
12
Sulfate
0.000
0.002
0.004
0.006
0.008
0.010
150
200
250
300
350
400
450
Conductivity
0.000
0.001
0.002
0.003
0.004
0.005
200
300
400
500
600
700
OrganicCarbon
0.00
0.02
0.04
0.06
0.08
0.10
0.12
5
10
15
20
25
Trihalomethanes
0.000
0.005
0.010
0.015
0.020
0.025
20
40
60
80
100
120
Turbidity
0.2
0.3
0.4
0.5

Kesimpulan

Dapat dilihat dari hasil KS Test dan visualisasi *plot-plot* diatas bahwa kesembilan kolom dari data mengikuti distribusi normal.