P1_Descriptive_Statistics

April 17, 2023

1 DESKRIPSI STATISTIKA

Deskripsi: mean, median, modus, standar deviasi, variansi, range, nilai minimum, maksimum, kuartil, IQR, skewness dan kurtosis.

[1]:
```
# Import Library Pandas
import pandas as pd

# Read csv file
df = pd.read_csv("../data/anggur.csv")

# Print descriptive statistics function
def print_descriptive_statistics(dataframe):
    # Mean
    print("Mean:", dataframe.mean())
    print("----------------------------------------")

    # Median
    print("Median:", dataframe.median())
    print("----------------------------------------")

    # Modus
    print("Modus:")
    all_modes = dataframe.mode().values.tolist()
    if (len(all_modes) == dataframe.count()):
        print("Ada", dataframe.count(), "modus pada kolom ini. Jumlah tersebut␣
 ↪sama dengan jumlah nilai pada kolom ini.")
        print("Hal ini menandakan kolom ini memiliki nilai-nilai yang berbeda␣
 ↪satu sama lain.")
    else:
        for mode in all_modes:
            print(mode)
    print("----------------------------------------")

    # Standar Deviasi
    print("Standar Deviasi:", dataframe.std())
    print("----------------------------------------")
```

```
    # Variansi
    print("Variansi:", dataframe.var())
    print("----------------------------------------")

    # Range
    print("Range:", dataframe.max() - dataframe.min())
    print("----------------------------------------")

    # Minimum
    print("Nilai Minimum:", dataframe.min())
    print("----------------------------------------")

    # Maximum
    print("Nilai Maksimum:", dataframe.max())
    print("----------------------------------------")

    # Kuartil
    print("Kuartil Bawah:", dataframe.quantile(0.25))
    print("Kuartil Tengah:", dataframe.quantile(0.50))
    print("Kuartil Atas:", dataframe.quantile(0.75))
    print("----------------------------------------")

    # IQR
    print("IQR:", dataframe.quantile(0.75) - dataframe.quantile(0.25))
    print("----------------------------------------")

    # Skewness
    print("Skewness:", dataframe.skew())
    print("----------------------------------------")

    # Kurtosis
    print("Kurtosis:", dataframe.kurtosis())
    print("========================================")
```

[2]:
```
display(df)
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	5.90	0.4451	0.1813	2.049401	0.070574	
1	8.40	0.5768	0.2099	3.109590	0.101681	
2	7.54	0.5918	0.3248	3.673744	0.072416	
3	5.39	0.4201	0.3131	3.371815	0.072755	
4	6.51	0.5675	0.1940	4.404723	0.066379	
..	…	…	…	…	…	
995	7.96	0.6046	0.2662	1.592048	0.057555	
996	8.48	0.4080	0.2227	0.681955	0.051627	
997	6.11	0.4841	0.3720	2.377267	0.042806	
998	7.76	0.3590	0.3208	4.294486	0.098276	
999	5.87	0.5214	0.1883	2.179490	0.052923	

```
     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0              16.593818                 42.27   0.9982  3.27       0.71
1              22.555519                 16.01   0.9960  3.35       0.57
2               9.316866                 35.52   0.9990  3.31       0.64
3              18.212300                 41.97   0.9945  3.34       0.55
4               9.360591                 46.27   0.9925  3.27       0.45
..                   …                    …       …   …        …
995            14.892445                 44.61   0.9975  3.35       0.54
996            23.548965                 25.83   0.9972  3.41       0.46
997            21.624585                 48.75   0.9928  3.23       0.55
998            12.746186                 44.53   0.9952  3.30       0.66
999            16.203864                 24.37   0.9983  3.29       0.70

     alcohol  quality
0       8.64        7
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …      …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

1.1 Kolom Fixed Acidity

[3]:
```
# Fixed Acidity
print("====================o====================")
print("Deskripsi Statistika Kolom Fixed Acidity")
print("====================o====================")

df_fixed_acidity = df["fixed acidity"]

# Print Descriptive Statistics
print_descriptive_statistics(df_fixed_acidity)
```

```
====================o====================
Deskripsi Statistika Kolom Fixed Acidity
====================o====================
Mean: 7.152530000000006
----------------------------------------
Median: 7.15
----------------------------------------
```

```
Modus:
6.54
----------------------------------------
Standar Deviasi: 1.2015975764938276
----------------------------------------
Variansi: 1.4438367358358397
----------------------------------------
Range: 8.17
----------------------------------------
Nilai Minimum: 3.32
----------------------------------------
Nilai Maksimum: 11.49
----------------------------------------
Kuartil Bawah: 6.3774999999999995
Kuartil Tengah: 7.15
Kuartil Atas: 8.0
----------------------------------------
IQR: 1.6225000000000005
----------------------------------------
Skewness: -0.028878575532660055
----------------------------------------
Kurtosis: -0.019292120932933532
========================================
```

1.2 Kolom Volatile Acidity

[4]:
```python
# Volatile Acidity
print("====================o====================")
print("Deskripsi Statistika Kolom Volatile Acidity")
print("====================o====================")

df_volatile_acidity = df["volatile acidity"]

# Print Descriptive Statistics
print_descriptive_statistics(df_volatile_acidity)
```

```
====================o====================
Deskripsi Statistika Kolom Volatile Acidity
====================o====================
Mean: 0.5208384999999999
----------------------------------------
Median: 0.52485
----------------------------------------
Modus:
0.5546
----------------------------------------
Standar Deviasi: 0.09584827405534954
----------------------------------------
```

```
Variansi: 0.009186891639389393
----------------------------------------
Range: 0.6652
----------------------------------------
Nilai Minimum: 0.1399
----------------------------------------
Nilai Maksimum: 0.8051
----------------------------------------
Kuartil Bawah: 0.4561
Kuartil Tengah: 0.52485
Kuartil Atas: 0.585375
----------------------------------------
IQR: 0.12927499999999997
----------------------------------------
Skewness: -0.1976986986092083
----------------------------------------
Kurtosis: 0.16185290336961788
========================================
```

1.3 Kolom Citric Acid

[5]:
```python
# Citric Acid
print("====================o====================")
print("Deskripsi Statistika Kolom Citric Acid")
print("====================o====================")

df_citric_acid = df["citric acid"]

# Print Descriptive Statistics
print_descriptive_statistics(df_citric_acid)
```

```
====================o====================
Deskripsi Statistika Kolom Citric Acid
====================o====================
Mean: 0.27051699999999995
----------------------------------------
Median: 0.2722
----------------------------------------
Modus:
0.3019
----------------------------------------
Standar Deviasi: 0.04909837147076352
----------------------------------------
Variansi: 0.0024106500810810853
----------------------------------------
Range: 0.29290000000000005
----------------------------------------
Nilai Minimum: 0.1167
```

```
----------------------------------------
Nilai Maksimum: 0.4096
----------------------------------------
Kuartil Bawah: 0.2378
Kuartil Tengah: 0.2722
Kuartil Atas: 0.302325
----------------------------------------
IQR: 0.064525
----------------------------------------
Skewness: -0.045576058685017296
----------------------------------------
Kurtosis: -0.1046792495951605
========================================
```

1.4 Kolom Residual Sugar

[6]:
```python
# Residual Sugar
print("====================o====================")
print("Deskripsi Statistika Kolom Residual Sugar")
print("====================o====================")

df_residual_sugar = df["residual sugar"]

# Print Descriptive Statistics
print_descriptive_statistics(df_residual_sugar)
```

```
====================o====================
Deskripsi Statistika Kolom Residual Sugar
====================o====================
Mean: 2.5671036825067572
----------------------------------------
Median: 2.519430272865794
----------------------------------------
Modus:
Ada 1000 modus pada kolom ini. Jumlah tersebut sama dengan jumlah nilai pada
kolom ini.
Hal ini menandakan kolom ini memiliki nilai-nilai yang berbeda satu sama lain.
----------------------------------------
Standar Deviasi: 0.9879154365046932
----------------------------------------
Variansi: 0.9759769096842584
----------------------------------------
Range: 5.5182004097078625
----------------------------------------
Nilai Minimum: 0.032554525015195
----------------------------------------
Nilai Maksimum: 5.550754934723058
----------------------------------------
```

```
Kuartil Bawah: 1.896329943488683
Kuartil Tengah: 2.519430272865794
Kuartil Atas: 3.220873482829786
----------------------------------------
IQR: 1.3245435393411031
----------------------------------------
Skewness: 0.13263808618992312
----------------------------------------
Kurtosis: -0.04298003436476261
========================================
```

1.5 Kolom Chlorides

[7]:
```python
# Chlorides
print("====================o====================")
print("Deskripsi Statistika Kolom Chlorides")
print("====================o====================")

df_chlorides = df["chlorides"]

# Print Descriptive Statistics
print_descriptive_statistics(df_chlorides)
```

```
====================o====================
Deskripsi Statistika Kolom Chlorides
====================o====================
Mean: 0.08119515250784973
----------------------------------------
Median: 0.0821669021645236
----------------------------------------
Modus:
Ada 1000 modus pada kolom ini. Jumlah tersebut sama dengan jumlah nilai pada
kolom ini.
Hal ini menandakan kolom ini memiliki nilai-nilai yang berbeda satu sama lain.
----------------------------------------
Standar Deviasi: 0.020110647243996742
----------------------------------------
Variansi: 0.0004044381325724738
----------------------------------------
Range: 0.1256351302653488
----------------------------------------
Nilai Minimum: 0.0151224391657095
----------------------------------------
Nilai Maksimum: 0.1407575694310583
----------------------------------------
Kuartil Bawah: 0.06657363190977357
Kuartil Tengah: 0.0821669021645236
Kuartil Atas: 0.09531150148556258
```

```
----------------------------------------
IQR: 0.028737869575789013
----------------------------------------
Skewness: -0.05131929742072573
----------------------------------------
Kurtosis: -0.2465081359240382
========================================
```

1.6 Kolom Free Sulfur Dioxide

[8]:
```
# Free Sulfur Dioxide
print("====================o====================")
print("Deskripsi Statistika Kolom Free Sulfur Dioxide")
print("====================o====================")

df_free_sulfur_dioxide = df["free sulfur dioxide"]

# Print Descriptive Statistics
print_descriptive_statistics(df_free_sulfur_dioxide)
```

```
====================o====================
Deskripsi Statistika Kolom Free Sulfur Dioxide
====================o====================
Mean: 14.907679251029792
----------------------------------------
Median: 14.860346236568924
----------------------------------------
Modus:
Ada 1000 modus pada kolom ini. Jumlah tersebut sama dengan jumlah nilai pada
kolom ini.
Hal ini menandakan kolom ini memiliki nilai-nilai yang berbeda satu sama lain.
----------------------------------------
Standar Deviasi: 4.888099705756564
----------------------------------------
Variansi: 23.89351873341741
----------------------------------------
Range: 27.26784690109891
----------------------------------------
Nilai Minimum: 0.194678523326937
----------------------------------------
Nilai Maksimum: 27.462525424425845
----------------------------------------
Kuartil Bawah: 11.426716949457617
Kuartil Tengah: 14.860346236568924
Kuartil Atas: 18.313097915395005
----------------------------------------
IQR: 6.886380965937388
----------------------------------------
```

```
Skewness: 0.007130415991143398
----------------------------------------
Kurtosis: -0.36496364342685306
========================================
```

1.7 Kolom Total Sulfur Dioxide

[9]:
```
# Total Sulfur Dioxide
print("====================o====================")
print("Deskripsi Statistika Kolom Total Sulfur Dioxide")
print("====================o====================")

df_total_sulfur_dioxide = df["total sulfur dioxide"]

# Print Descriptive Statistics
print_descriptive_statistics(df_total_sulfur_dioxide)
```

```
====================o====================
Deskripsi Statistika Kolom Total Sulfur Dioxide
====================o====================
Mean: 40.290150000000075
----------------------------------------
Median: 40.19
----------------------------------------
Modus:
35.2
37.25
39.64
40.61
41.05
41.59
44.51
----------------------------------------
Standar Deviasi: 9.965767376218295
----------------------------------------
Variansi: 99.3165193968969
----------------------------------------
Range: 66.80999999999999
----------------------------------------
Nilai Minimum: 3.15
----------------------------------------
Nilai Maksimum: 69.96
----------------------------------------
Kuartil Bawah: 33.785
Kuartil Tengah: 40.19
Kuartil Atas: 47.0225
----------------------------------------
IQR: 13.237500000000004
```

```
----------------------------------------
Skewness: -0.024060026812269975
----------------------------------------
Kurtosis: 0.06394978916172311
========================================
```

1.8 Kolom Density

[10]:
```
# Density
print("====================o====================")
print("Deskripsi Statistika Kolom Density")
print("====================o====================")

df_density = df["density"]

# Print Descriptive Statistics
print_descriptive_statistics(df_density)
```

```
====================o====================
Deskripsi Statistika Kolom Density
====================o====================
Mean: 0.9959253000000002
----------------------------------------
Median: 0.996
----------------------------------------
Modus:
0.9959
0.9961
0.9965
0.997
----------------------------------------
Standar Deviasi: 0.0020201809426487133
----------------------------------------
Variansi: 4.081131041041044e-06
----------------------------------------
Range: 0.013799999999999923
----------------------------------------
Nilai Minimum: 0.9888
----------------------------------------
Nilai Maksimum: 1.0026
----------------------------------------
Kuartil Bawah: 0.9946
Kuartil Tengah: 0.996
Kuartil Atas: 0.9972
----------------------------------------
IQR: 0.0025999999999999357
----------------------------------------
Skewness: -0.07688278915513917
```

```
----------------------------------------
Kurtosis: 0.01636562128503849
========================================
```

1.9 Kolom pH

[11]:
```
# pH
print("====================o====================")
print("Deskripsi Statistika Kolom pH")
print("====================o====================")

df_pH = df["pH"]

# Print Descriptive Statistics
print_descriptive_statistics(df_pH)
```

```
====================o====================
Deskripsi Statistika Kolom pH
====================o====================
Mean: 3.303610000000003
----------------------------------------
Median: 3.3
----------------------------------------
Modus:
3.34
----------------------------------------
Standar Deviasi: 0.10487548220040155
----------------------------------------
Variansi: 0.010998866766766742
----------------------------------------
Range: 0.7399999999999998
----------------------------------------
Nilai Minimum: 2.97
----------------------------------------
Nilai Maksimum: 3.71
----------------------------------------
Kuartil Bawah: 3.23
Kuartil Tengah: 3.3
Kuartil Atas: 3.37
----------------------------------------
IQR: 0.14000000000000012
----------------------------------------
Skewness: 0.14767259510827038
----------------------------------------
Kurtosis: 0.0809095518741838
========================================
```

1.10 Kolom Sulphates

[12]:
```
# Sulphates
print("====================o====================")
print("Deskripsi Statistika Kolom Sulphates")
print("====================o====================")

df_sulphates = df["sulphates"]

# Print Descriptive Statistics
print_descriptive_statistics(df_sulphates)
```

```
====================o====================
Deskripsi Statistika Kolom Sulphates
====================o====================
Mean: 0.5983899999999999
----------------------------------------
Median: 0.595
----------------------------------------
Modus:
0.59
----------------------------------------
Standar Deviasi: 0.10081900799141184
----------------------------------------
Variansi: 0.010164472372372365
----------------------------------------
Range: 0.6699999999999999
----------------------------------------
Nilai Minimum: 0.29
----------------------------------------
Nilai Maksimum: 0.96
----------------------------------------
Kuartil Bawah: 0.53
Kuartil Tengah: 0.595
Kuartil Atas: 0.67
----------------------------------------
IQR: 0.14
----------------------------------------
Skewness: 0.1491989008699043
----------------------------------------
Kurtosis: 0.06481928180859686
========================================
```

1.11 Kolom Alcohol

[13]:
```
# Alcohol
print("====================o====================")
print("Deskripsi Statistika Kolom Alcohol")
print("====================o====================")

df_alcohol = df["alcohol"]

# Print Descriptive Statistics
print_descriptive_statistics(df_alcohol)
```

```
====================o====================
Deskripsi Statistika Kolom Alcohol
====================o====================
Mean: 10.592279999999985
----------------------------------------
Median: 10.61
----------------------------------------
Modus:
9.86
10.31
----------------------------------------
Standar Deviasi: 1.5107060052287598
----------------------------------------
Variansi: 2.282232634234237
----------------------------------------
Range: 8.989999999999998
----------------------------------------
Nilai Minimum: 6.03
----------------------------------------
Nilai Maksimum: 15.02
----------------------------------------
Kuartil Bawah: 9.56
Kuartil Tengah: 10.61
Kuartil Atas: 11.622499999999999
----------------------------------------
IQR: 2.0624999999999982
----------------------------------------
Skewness: -0.01899140432111647
----------------------------------------
Kurtosis: -0.13173155932281988
========================================
```

1.12 Kolom Quality

[14]:
```
# Quality
print("====================o====================")
print("Deskripsi Statistika Kolom Quality")
print("====================o====================")

df_quality = df["quality"]

# Print Descriptive Statistics
print_descriptive_statistics(df_quality)
```

```
====================o====================
Deskripsi Statistika Kolom Quality
====================o====================
Mean: 7.958
----------------------------------------
Median: 8.0
----------------------------------------
Modus:
8
----------------------------------------
Standar Deviasi: 0.9028017783827452
----------------------------------------
Variansi: 0.8150510510510475
----------------------------------------
Range: 5
----------------------------------------
Nilai Minimum: 5
----------------------------------------
Nilai Maksimum: 10
----------------------------------------
Kuartil Bawah: 7.0
Kuartil Tengah: 8.0
Kuartil Atas: 9.0
----------------------------------------
IQR: 2.0
----------------------------------------
Skewness: -0.08905409122491781
----------------------------------------
Kurtosis: 0.10829100232871003
========================================
```

P2_Visualization

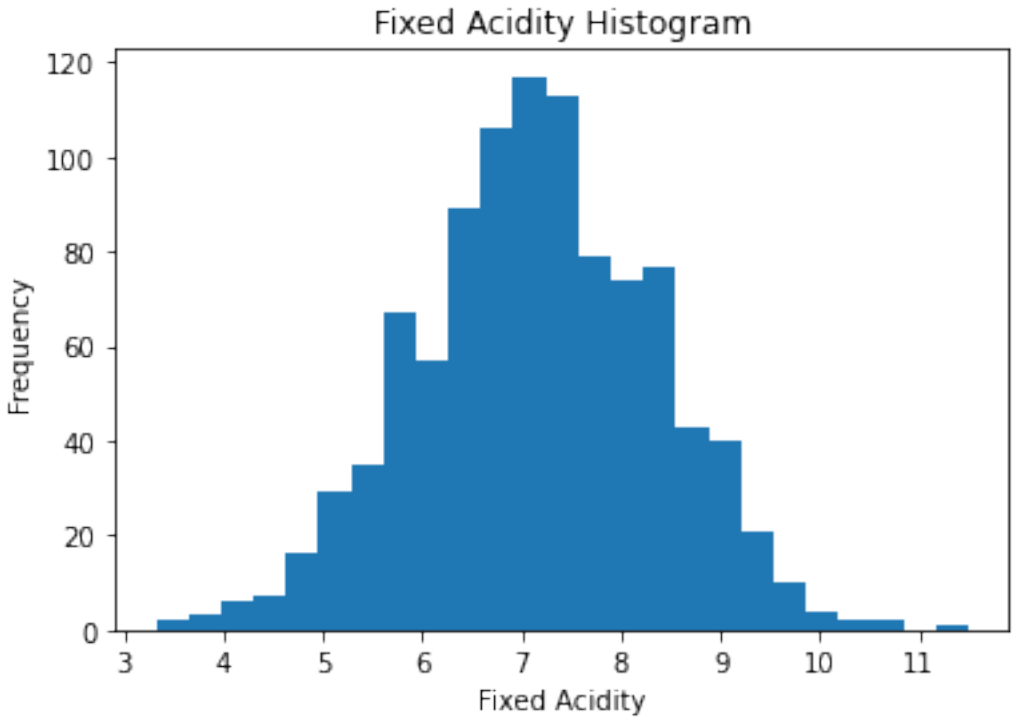
April 17, 2023

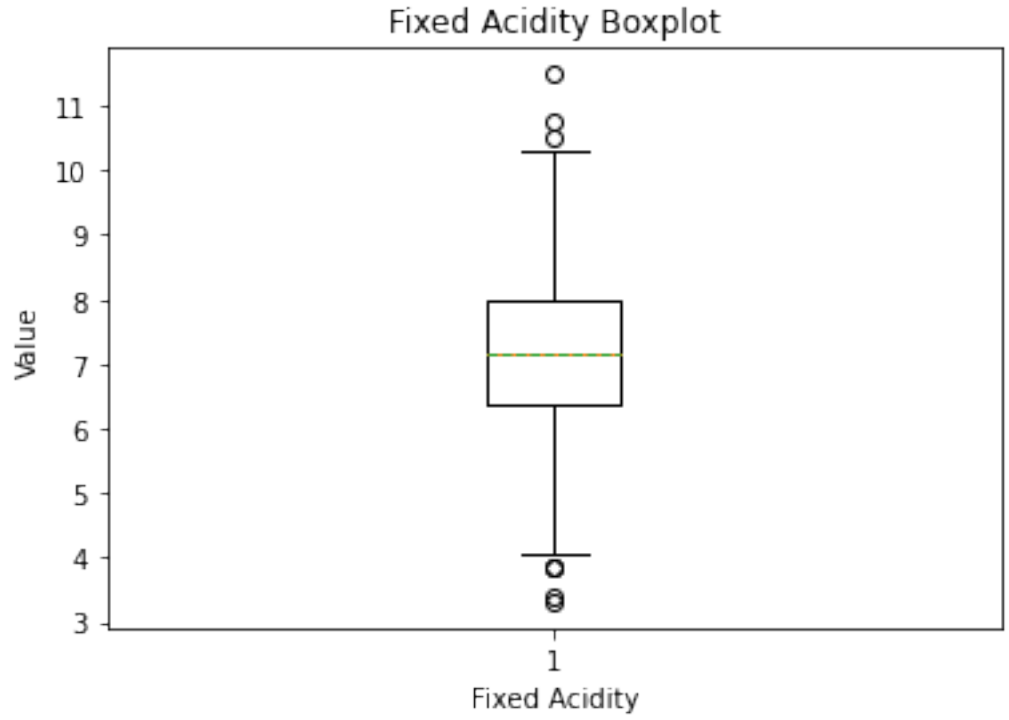
1 VISUALIZATION

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt

# Read csv file
df = pd.read_csv("../data/anggur.csv")
```

1.1 Kolom Fixed Acidity

[2]:
```
df_fixed_acidity = df["fixed acidity"]

# Histogram
plt.hist(df_fixed_acidity, bins = 25)
plt.title('Fixed Acidity Histogram')
plt.xlabel('Fixed Acidity')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_fixed_acidity, showmeans = True, meanline = True)
plt.title('Fixed Acidity Boxplot')
plt.xlabel('Fixed Acidity')
plt.ylabel('Value')
plt.show()
```

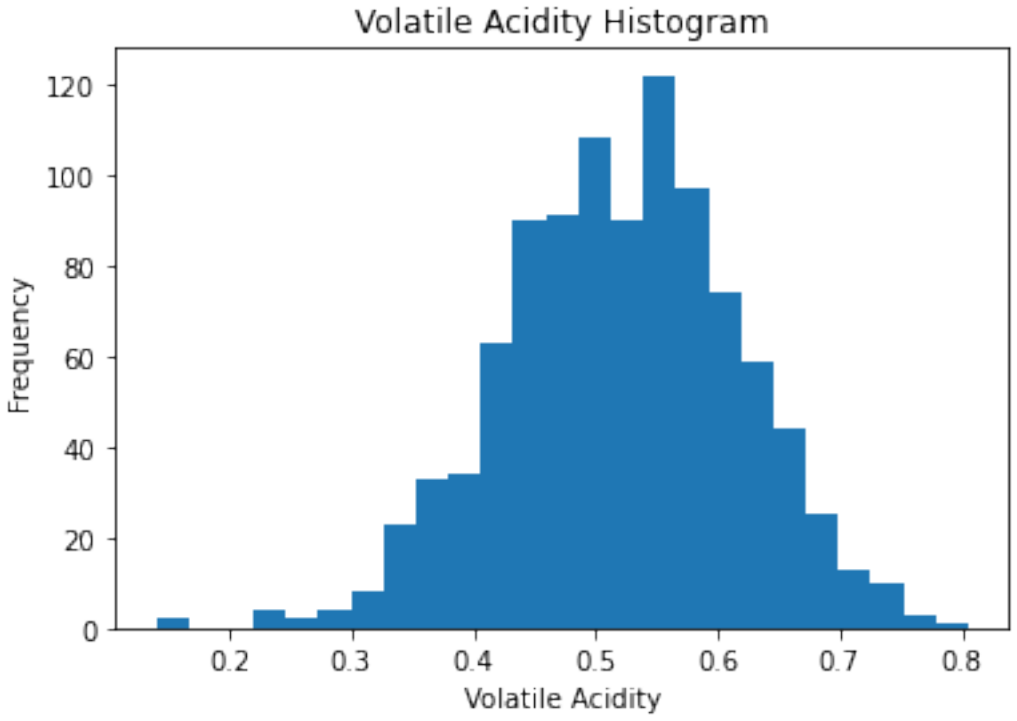

Fixed Acidity Histogram
Frequency
Fixed Acidity
0
20
40
60
80
100
120
3
4
5
6
7
8
9
10
11

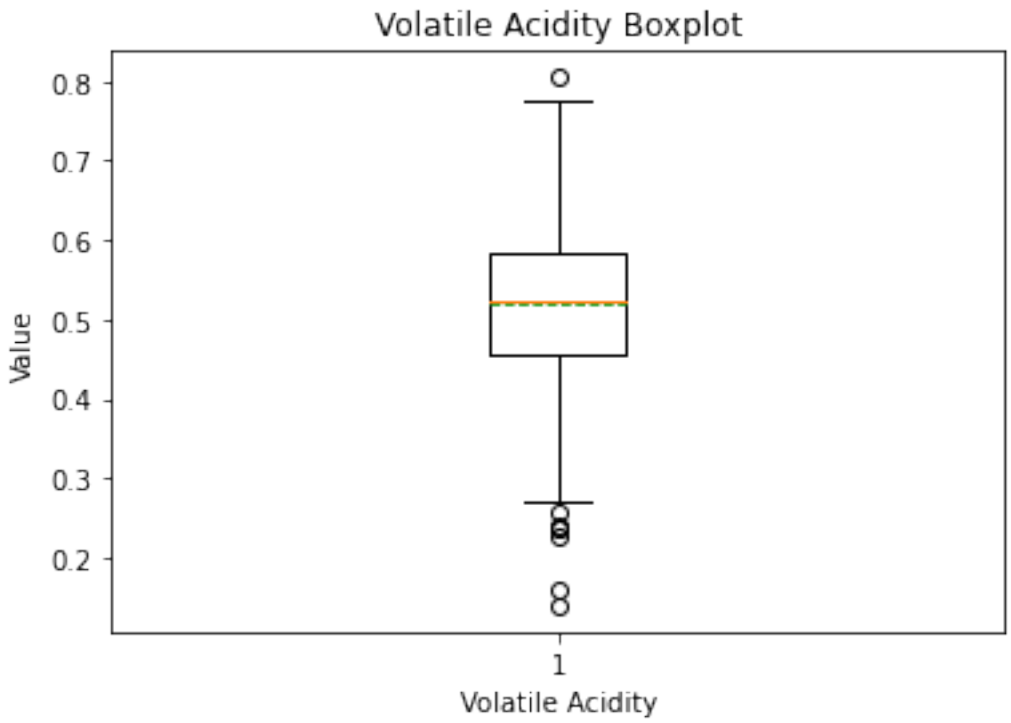

Fixed Acidity Boxplot
Value
Fixed Acidity
1
3
4
5
6
7
8
9
10
11

Berdasarkan histogram tersebut, terlihat bahwa distribusi Fixed Acidity cenderung condong ke kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 3-4 dan 10-12. Terlihat juga bahwa kuartil bawahnya berada di sekitar 6.2, kuartil tengahnya berada di sekitar 7.1, dan kuartil atasnya berada di sekitar 8. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 7.1 (sedikit di atas median).

1.2 Kolom Volatile Acidity

[3]:
```
df_volatile_acidity = df["volatile acidity"]

# Histogram
plt.hist(df_volatile_acidity, bins = 25)
plt.title('Volatile Acidity Histogram')
plt.xlabel('Volatile Acidity')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_volatile_acidity, showmeans = True, meanline = True)
plt.title('Volatile Acidity Boxplot')
plt.xlabel('Volatile Acidity')
plt.ylabel('Value')
plt.show()
```

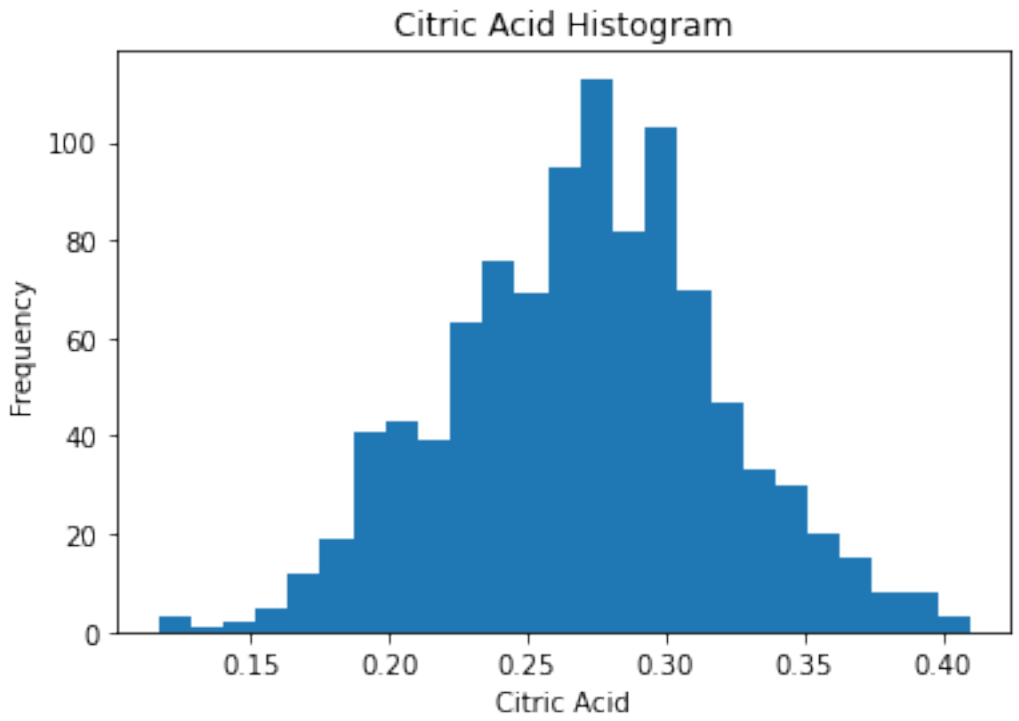


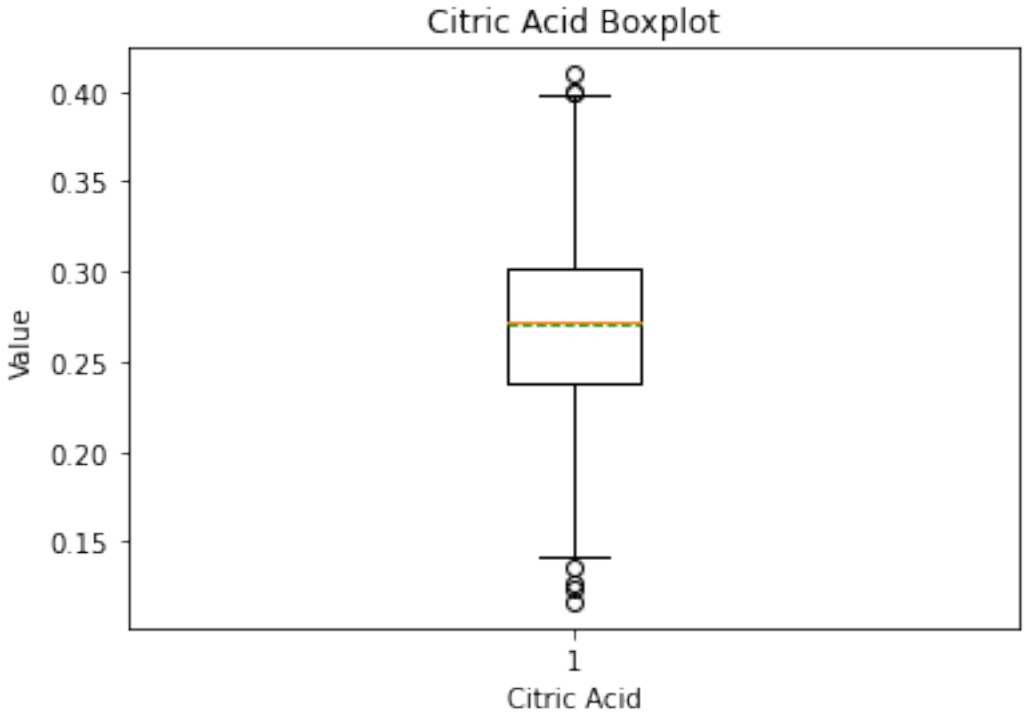


Berdasarkan histogram tersebut, terlihat bahwa distribusi Volatile Acidity cenderung condong ke kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 0.1-0.28 dan 0.78-0.83. Terlihat juga bahwa kuartil bawahnya berada di sekitar 0.46, kuartil tengahnya berada di sekitar 0.53, dan kuartil atasnya berada di sekitar 0.58. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 0.53 (sedikit di bawah median).

1.3 Kolom Citric Acid

[4]:
```
df_citric_acid = df["citric acid"]

# Histogram
plt.hist(df_citric_acid, bins = 25)
plt.title('Citric Acid Histogram')
plt.xlabel('Citric Acid')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_citric_acid, showmeans = True, meanline = True)
plt.title('Citric Acid Boxplot')
plt.xlabel('Citric Acid')
plt.ylabel('Value')
plt.show()
```

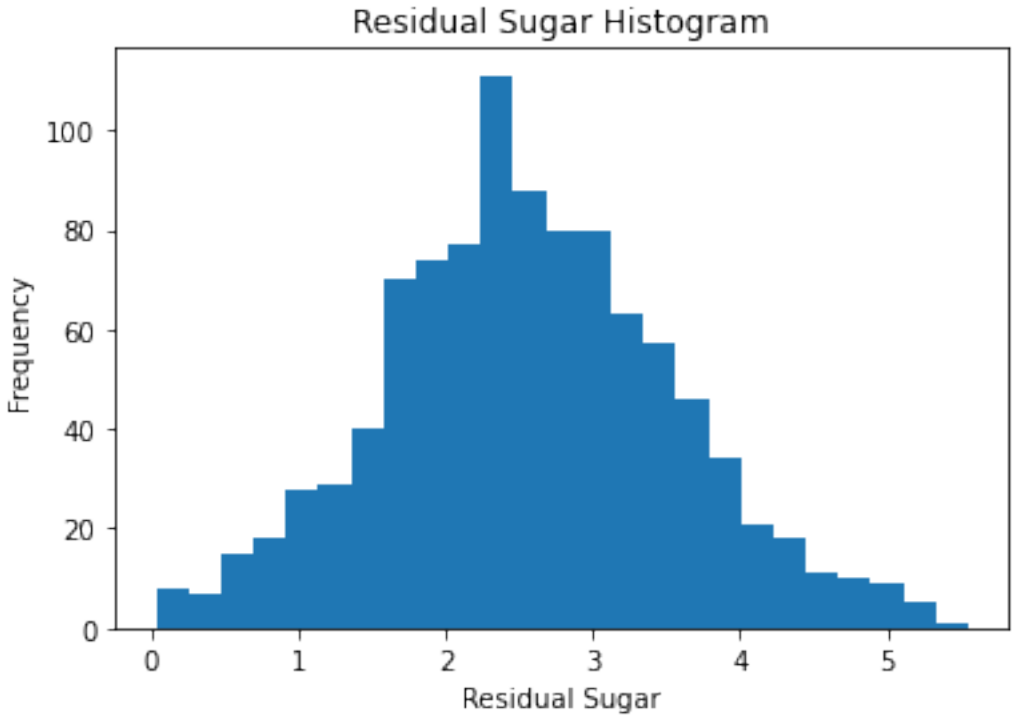

Citric Acid Histogram
Frequency
0
20
40
60
80
100
0.15
0.20
0.25
0.30
0.35
0.40
Citric Acid

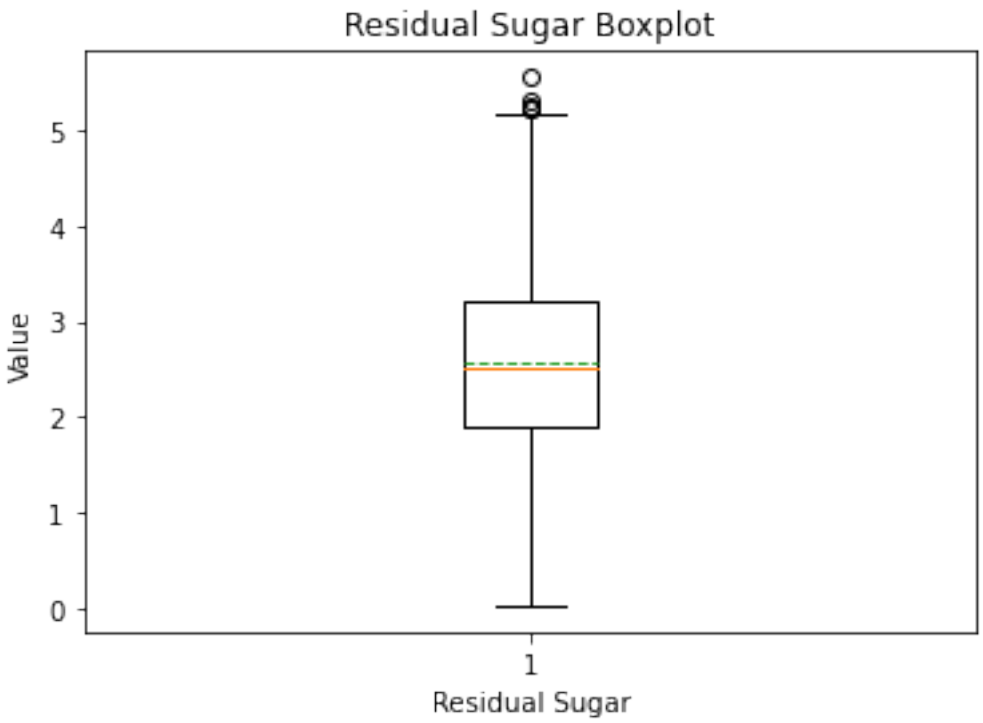

Citric Acid Boxplot
Value
0.15
0.20
0.25
0.30
0.35
0.40
1
Citric Acid

Berdasarkan histogram tersebut, terlihat bahwa distribusi Citric Acid cenderung condong ke arah kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 0.11-0.15 dan 0.4-0.43. Terlihat juga bahwa kuartil bawahnya berada di sekitar 0.24, kuartil tengahnya berada di sekitar 0.27, dan kuartil atasnya berada di sekitar 0.30. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 0.27 (sedikit di bawah median).

1.4 Kolom Residual Sugar

[5]:
```
df_residual_sugar = df["residual sugar"]

# Histogram
plt.hist(df_residual_sugar, bins = 25)
plt.title('Residual Sugar Histogram')
plt.xlabel('Residual Sugar')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_residual_sugar, showmeans = True, meanline = True)
plt.title('Residual Sugar Boxplot')
plt.xlabel('Residual Sugar')
plt.ylabel('Value')
plt.show()
```

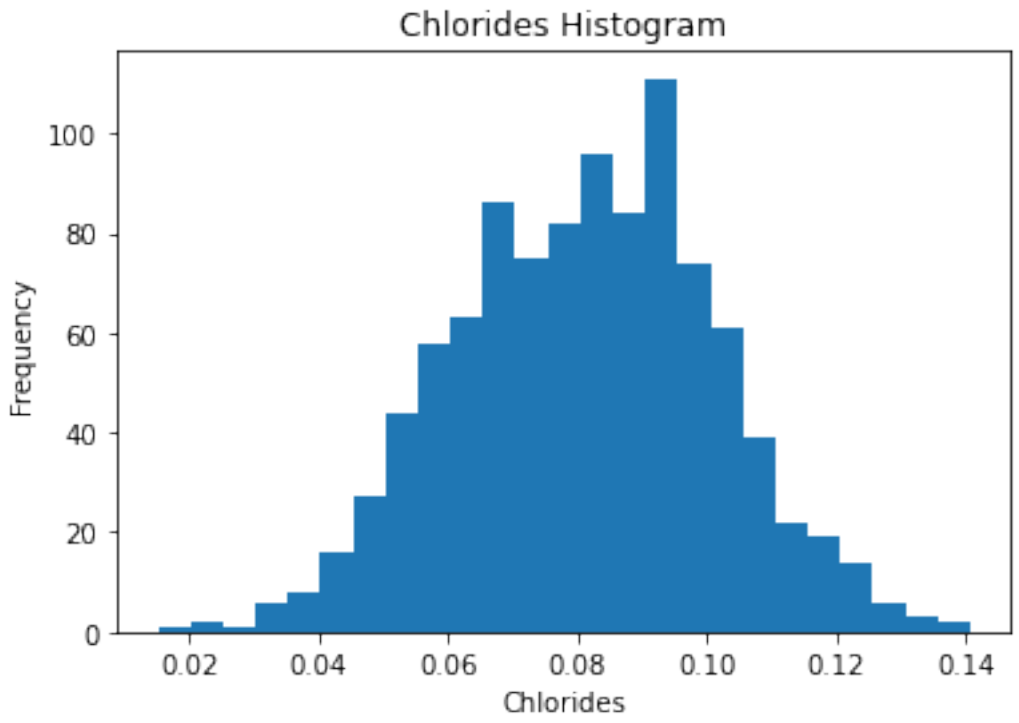


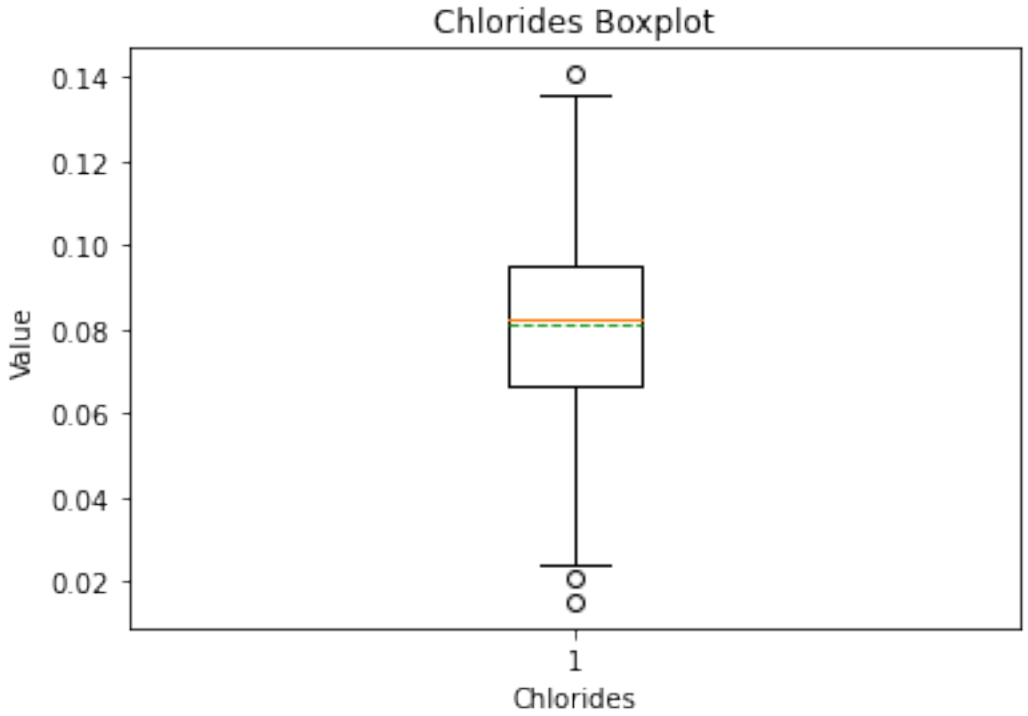


Berdasarkan histogram tersebut, terlihat bahwa distribusi Residual Sugar cenderung condong ke arah kanan(positively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 5-6. Terlihat juga bahwa kuartil bawahnya berada di sekitar 1.9, kuartil tengahnya berada di sekitar 2.51, dan kuartil atasnya berada di sekitar 3.2. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 2.55 (sedikit di atas median).

1.5 Kolom Chlorides

[6]:
```
df_chlorides = df["chlorides"]

# Histogram
plt.hist(df_chlorides, bins = 25)
plt.title('Chlorides Histogram')
plt.xlabel('Chlorides')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_chlorides, showmeans = True, meanline = True)
plt.title('Chlorides Boxplot')
plt.xlabel('Chlorides')
plt.ylabel('Value')
plt.show()
```

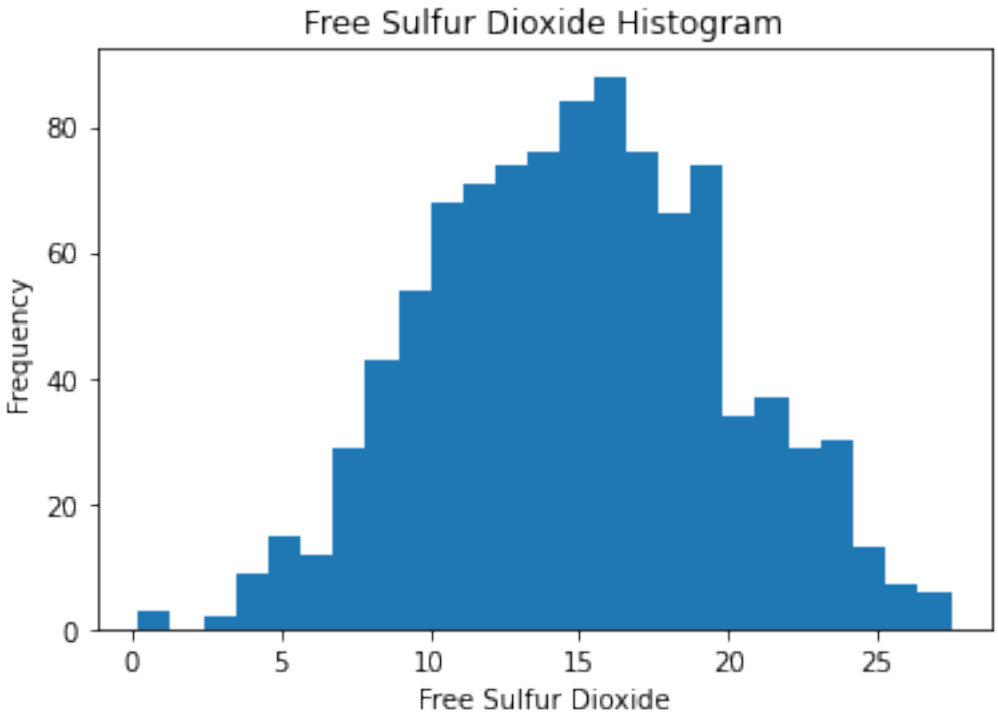

Chlorides Histogram
Frequency
100
80
60
40
20
0
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Chlorides

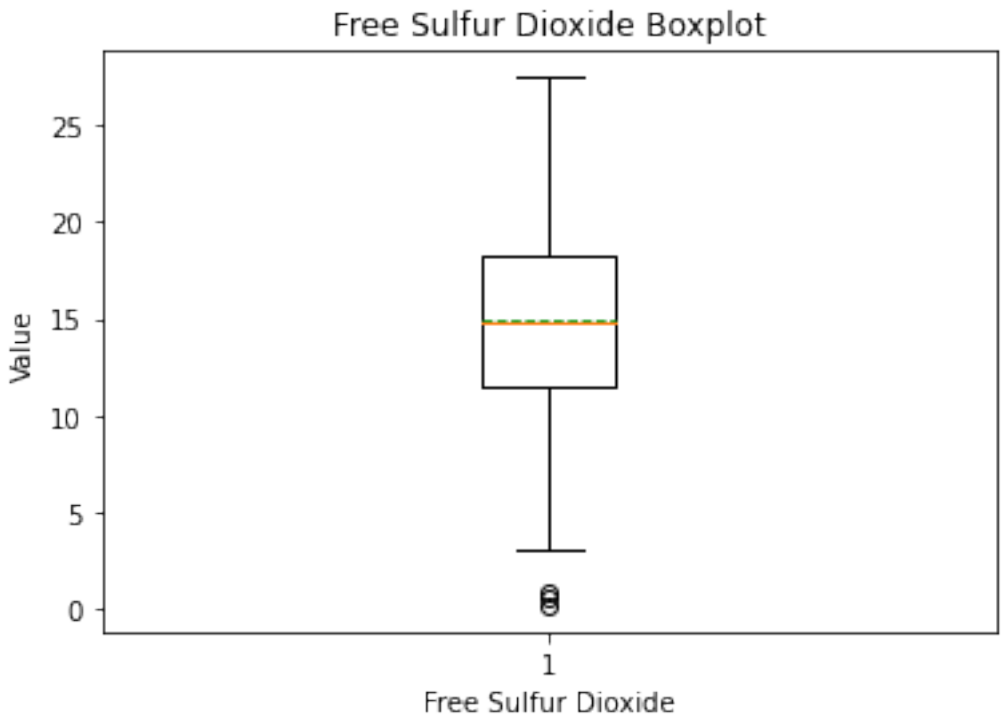

Chlorides Boxplot
Value
0.14
0.12
0.10
0.08
0.06
0.04
0.02
1
Chlorides

Berdasarkan histogram tersebut, terlihat bahwa distribusi Chlorides cenderung condong ke arah kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 0.01-0.03 dan 0.13-0.141. Terlihat juga bahwa kuartil bawahnya berada di sekitar 0.067, kuartil tengahnya berada di sekitar 0.082, dan kuartil atasnya berada di sekitar 0.096. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 0.081 (sedikit di bawah median).

1.6 Kolom Free Sulfur Dioxide

[7]:
```
df_free_sulfur_dioxide = df["free sulfur dioxide"]

# Histogram
plt.hist(df_free_sulfur_dioxide, bins = 25)
plt.title('Free Sulfur Dioxide Histogram')
plt.xlabel('Free Sulfur Dioxide')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_free_sulfur_dioxide, showmeans = True, meanline = True)
plt.title('Free Sulfur Dioxide Boxplot')
plt.xlabel('Free Sulfur Dioxide')
plt.ylabel('Value')
plt.show()
```

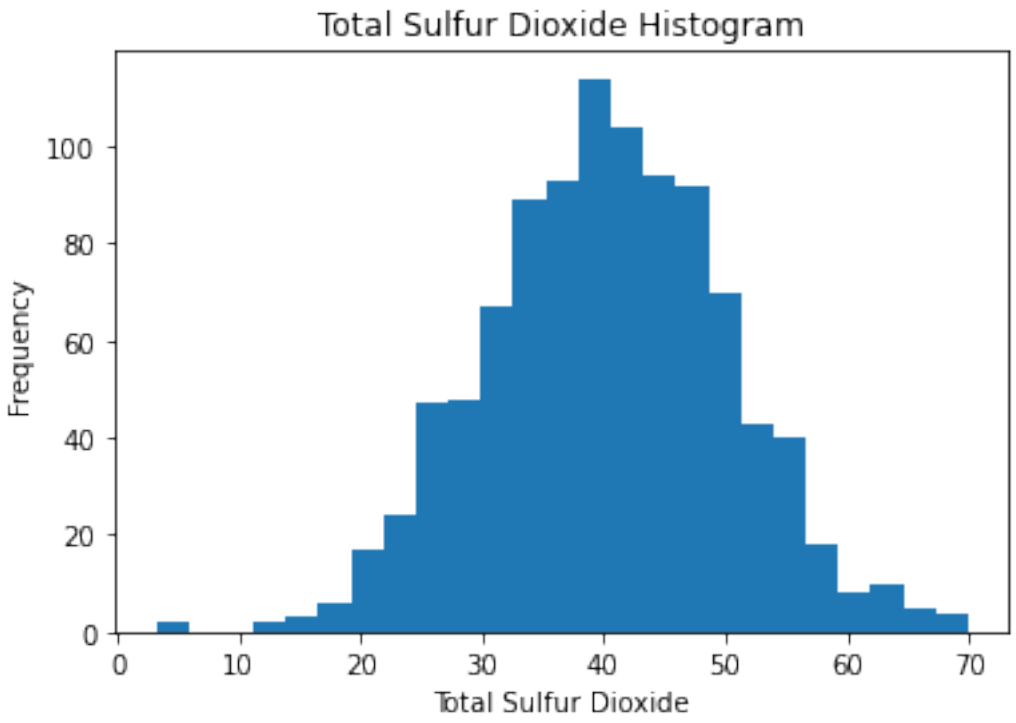


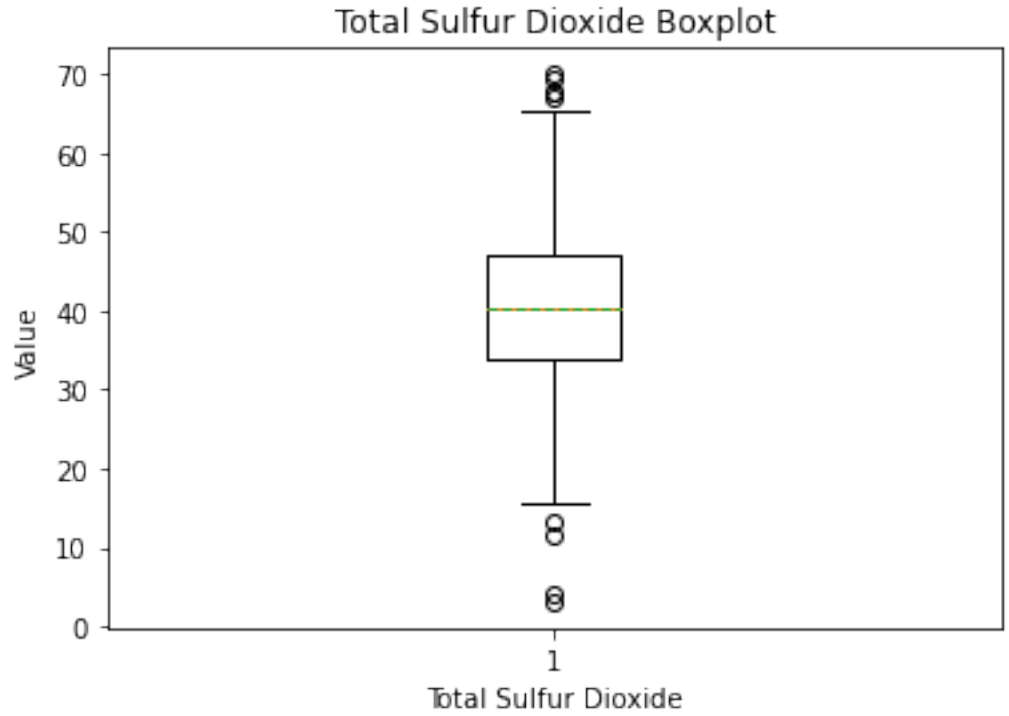


Berdasarkan histogram tersebut, terlihat bahwa distribusi Free Sulfur Dioxide cenderung mendekati distribusi normal karena bentuknya hampir simetris. Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 0-2.5. Terlihat juga bahwa kuartil bawahnya berada di sekitar 11.4, kuartil tengahnya berada di sekitar 14.85, dan kuartil atasnya berada di sekitar 18.3. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 14.9 (sedikit di atas median).

1.7 Kolom Total Sulfur Dioxide

[8]:
```
df_total_sulfur_dioxide = df["total sulfur dioxide"]

# Histogram
plt.hist(df_total_sulfur_dioxide, bins = 25)
plt.title('Total Sulfur Dioxide Histogram')
plt.xlabel('Total Sulfur Dioxide')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_total_sulfur_dioxide, showmeans = True, meanline = True)
plt.title('Total Sulfur Dioxide Boxplot')
plt.xlabel('Total Sulfur Dioxide')
plt.ylabel('Value')
plt.show()
```

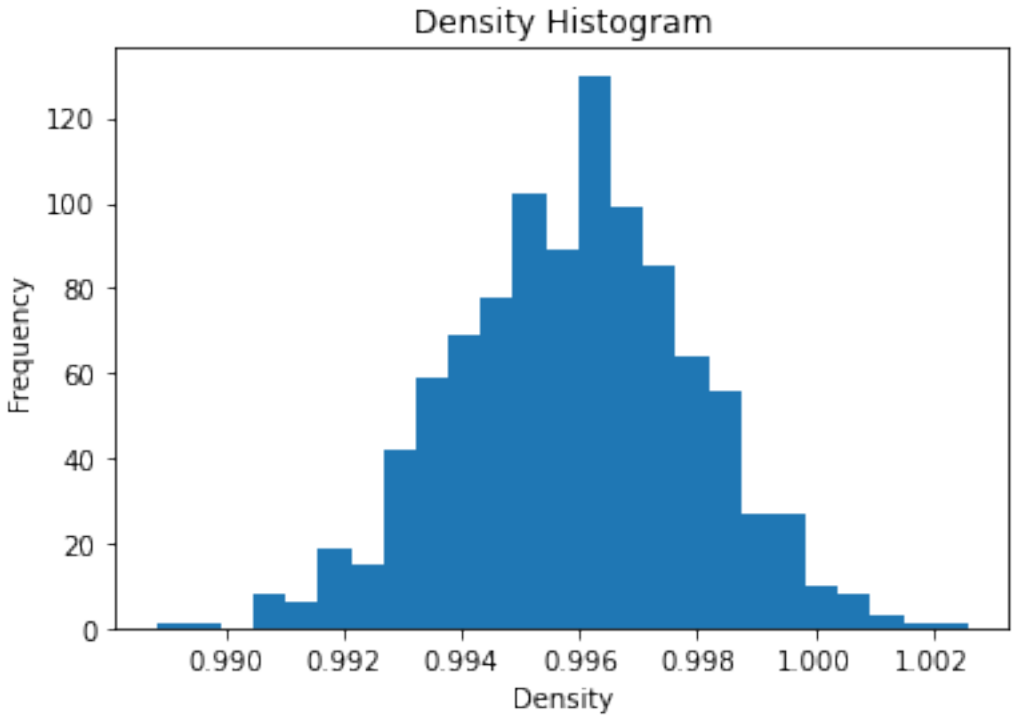

Total Sulfur Dioxide Histogram
Frequency
Total Sulfur Dioxide
0
20
40
60
80
100
0
10
20
30
40
50
60
70

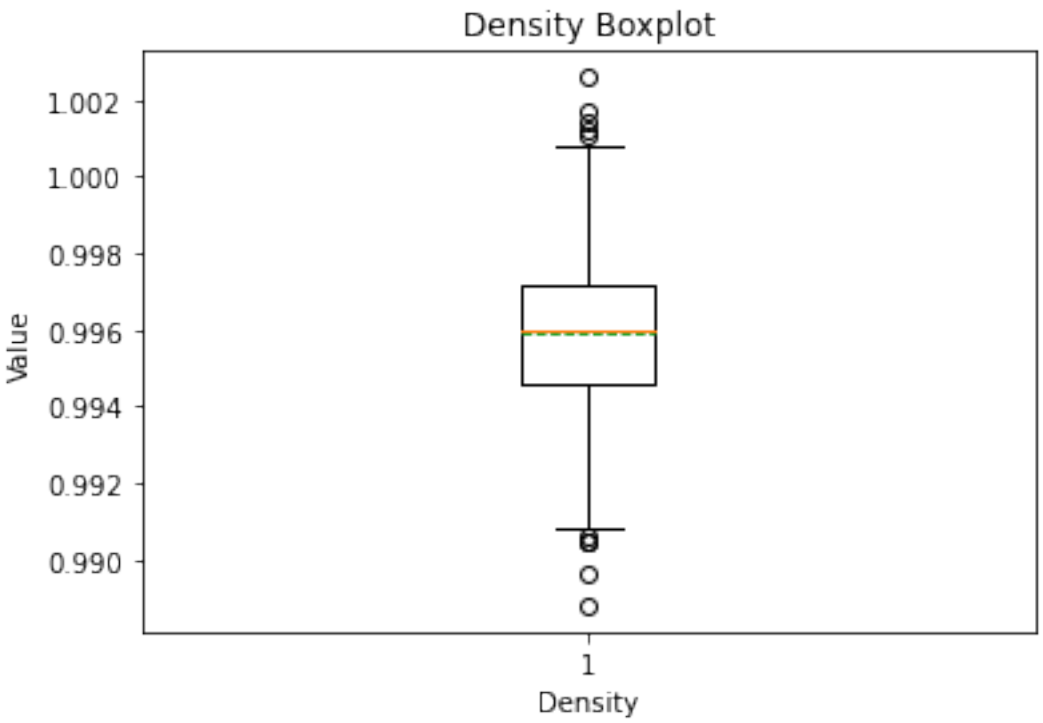

Total Sulfur Dioxide Boxplot
Value
Total Sulfur Dioxide
0
10
20
30
40
50
60
70
1

Berdasarkan histogram tersebut, terlihat bahwa distribusi Total Sulfur Dioxide cenderung condong ke kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 3-15 dan 65-70. Terlihat juga bahwa kuartil bawahnya berada di sekitar 33.5, kuartil tengahnya berada di sekitar 40.2, dan kuartil atasnya berada di sekitar 47. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 40.2 (sedikit di atas median).

1.8 Kolom Density

[9]:
```
df_density = df["density"]

# Histogram
plt.hist(df_density, bins = 25)
plt.title('Density Histogram')
plt.xlabel('Density')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_density, showmeans = True, meanline = True)
plt.title('Density Boxplot')
plt.xlabel('Density')
plt.ylabel('Value')
plt.show()
```

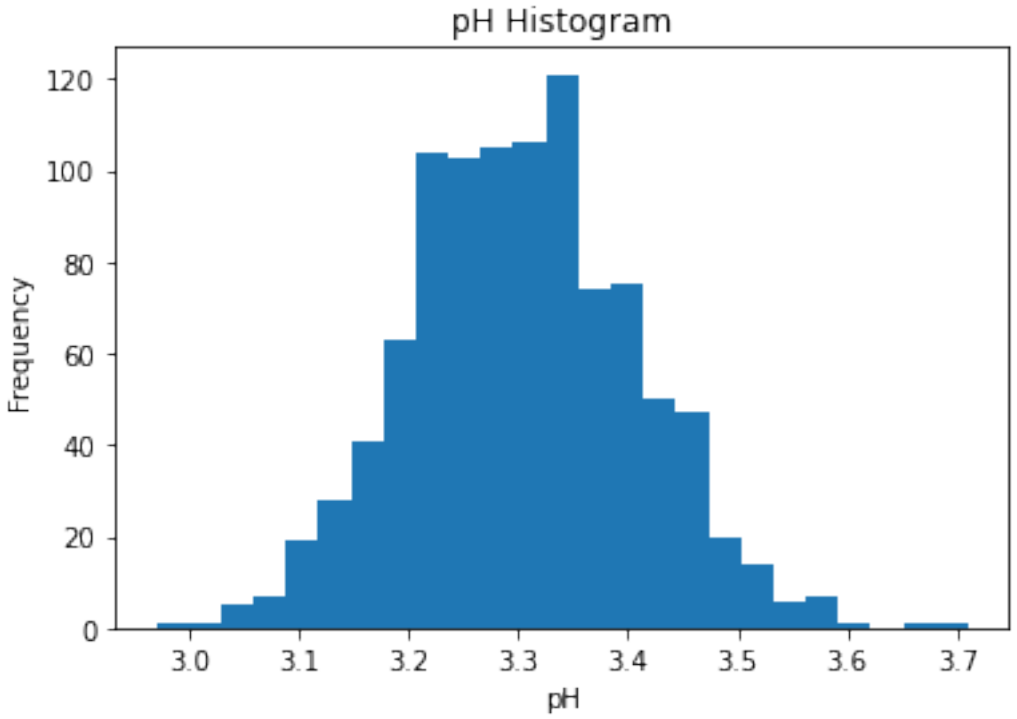


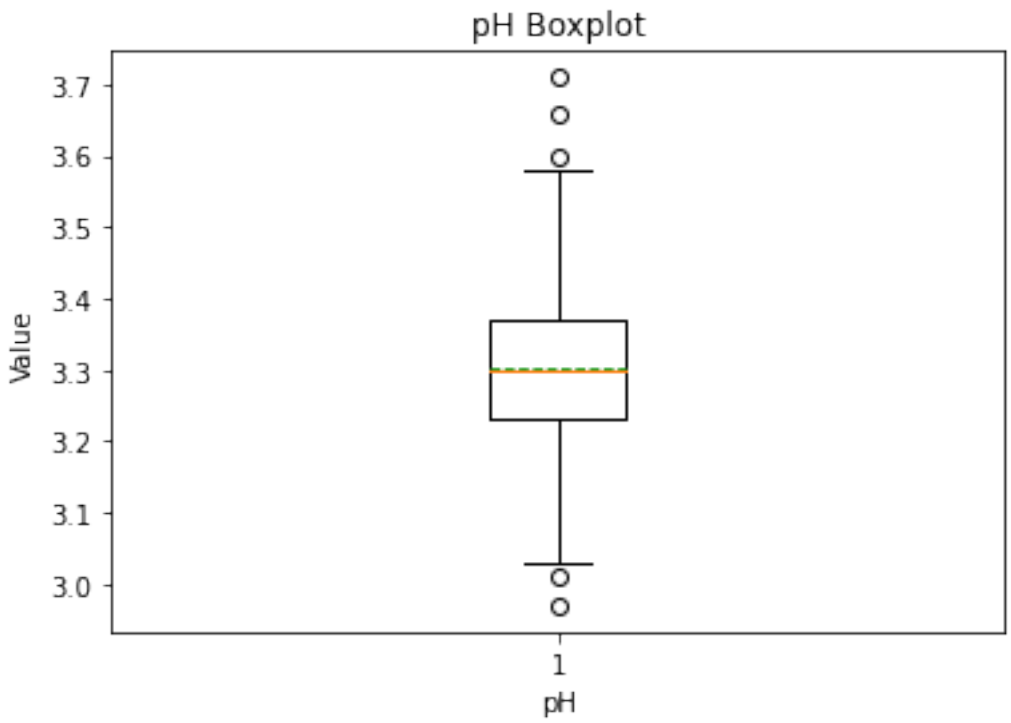


Berdasarkan histogram tersebut, terlihat bahwa distribusi Density cenderung condong ke kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 0.989-0.991 dan 1.001-1.003. Terlihat juga bahwa kuartil bawahnya berada di sekitar 0.995, kuartil tengahnya berada di sekitar 0.996, dan kuartil atasnya berada di sekitar 0.997. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 0.996 (sedikit di bawah median).

1.9 Kolom pH

[10]:
```
df_pH = df["pH"]

# Histogram
plt.hist(df_pH, bins = 25)
plt.title('pH Histogram')
plt.xlabel('pH')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_pH, showmeans = True, meanline = True)
plt.title('pH Boxplot')
plt.xlabel('pH')
plt.ylabel('Value')
plt.show()
```

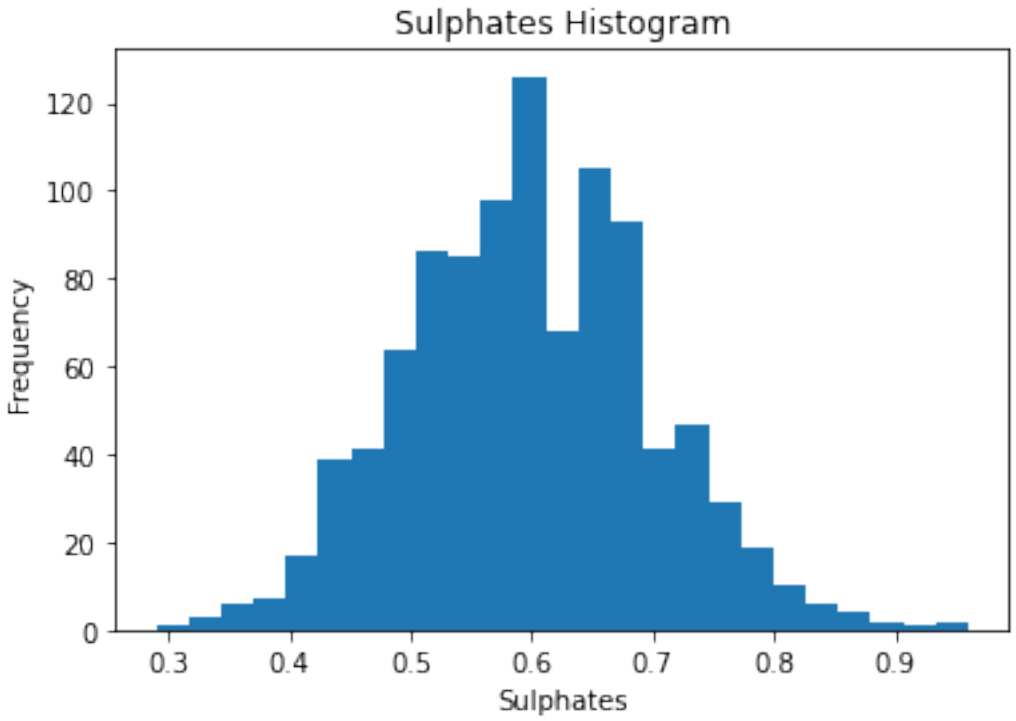


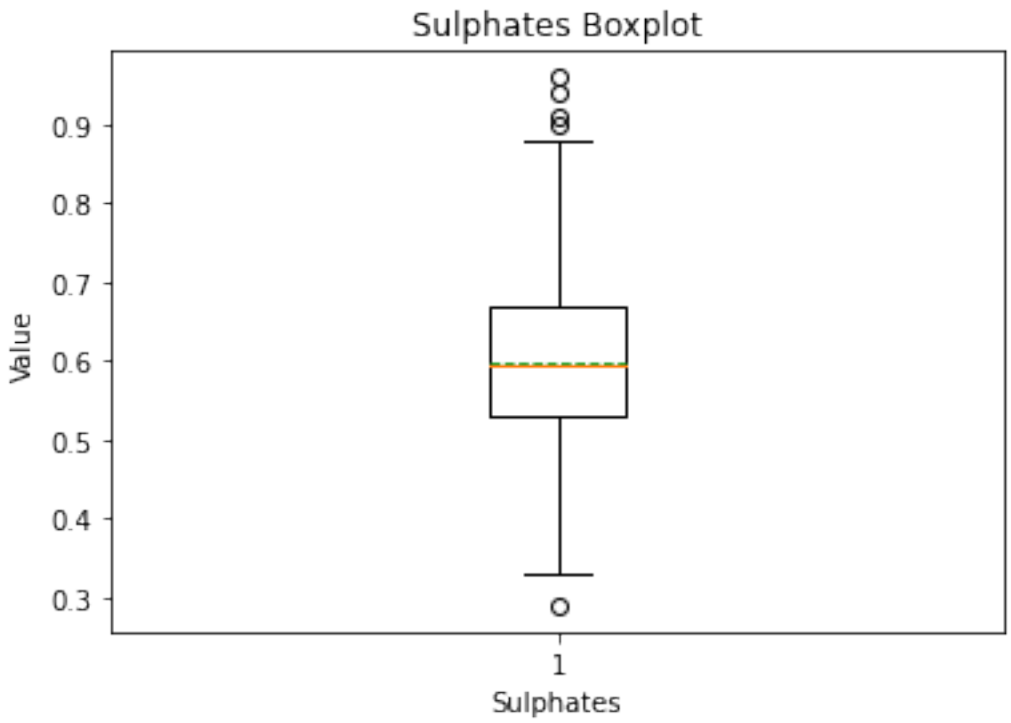


Berdasarkan histogram tersebut, terlihat bahwa distribusi pH cenderung condong ke kanan (positively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 2.95-3.05 dan 3.58-3.73. Terlihat juga bahwa kuartil bawahnya berada di sekitar 3.24, kuartil tengahnya berada di sekitar 3.3, dan kuartil atasnya berada di sekitar 3.37. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 3.3 (sedikit di atas median).

1.10 Kolom Sulphates

[11]:
```
df_sulphates = df["sulphates"]

# Histogram
plt.hist(df_sulphates, bins = 25)
plt.title('Sulphates Histogram')
plt.xlabel('Sulphates')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_sulphates, showmeans = True, meanline = True)
plt.title('Sulphates Boxplot')
plt.xlabel('Sulphates')
plt.ylabel('Value')
plt.show()
```

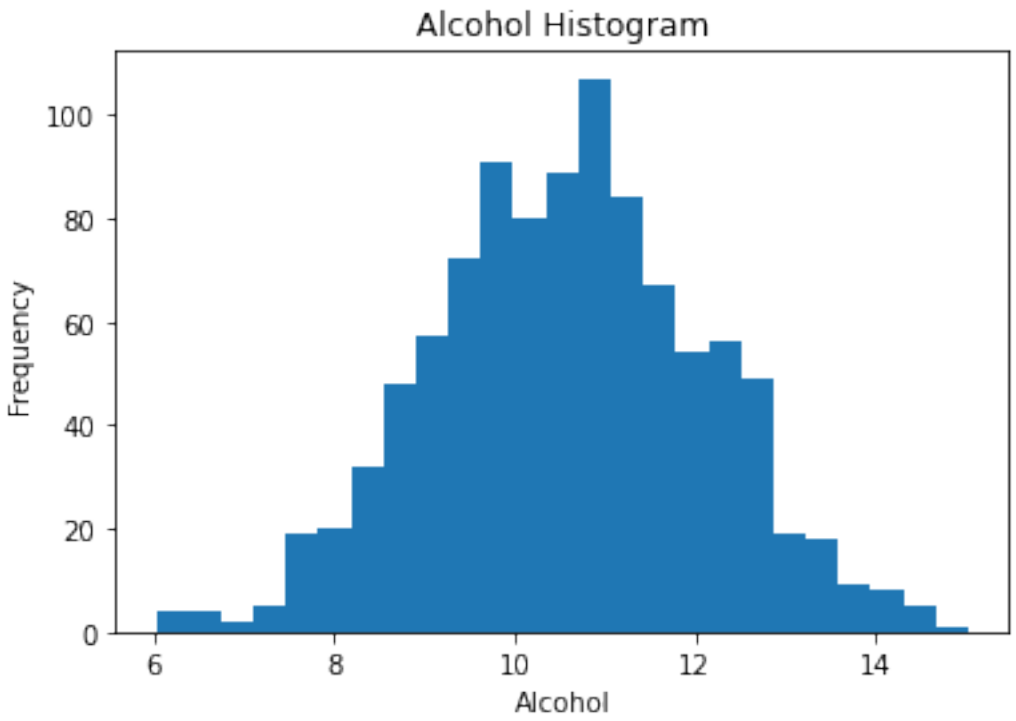


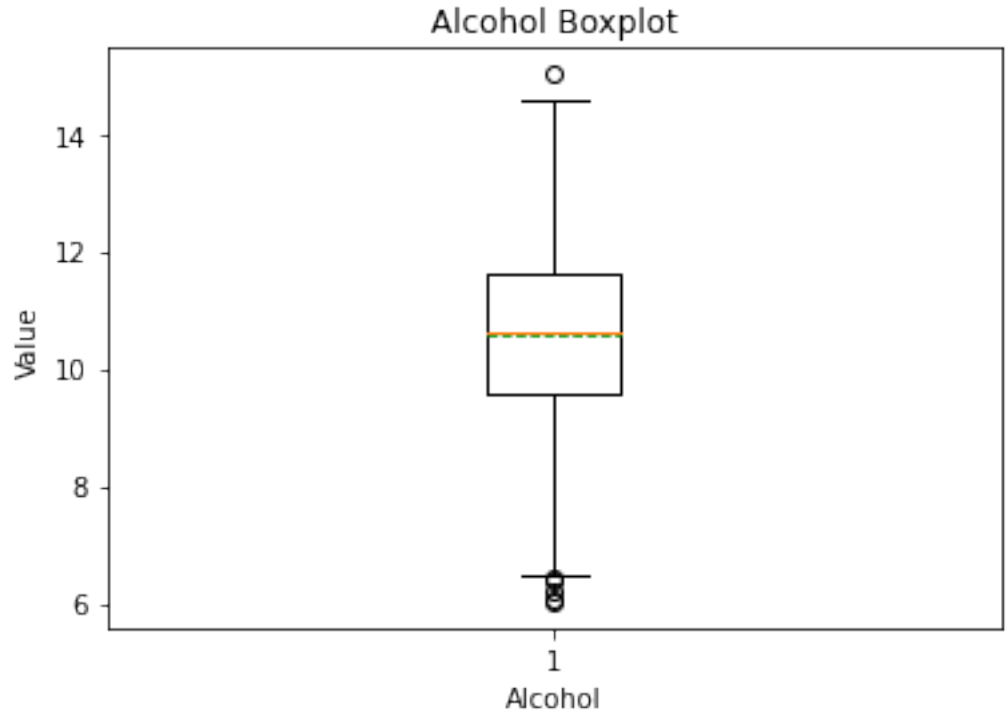


Berdasarkan histogram tersebut, terlihat bahwa distribusi Sulphates cenderung condong ke kanan (positively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 0.29-0.31 dan 0.87-0.98. Terlihat juga bahwa kuartil bawahnya berada di sekitar 0.53, kuartil tengahnya berada di sekitar 0.6, dan kuartil atasnya berada di sekitar 0.67. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 0.6 (sedikit di atas median).

1.11 Kolom Alcohol

[12]:
```
df_alcohol = df["alcohol"]

# Histogram
plt.hist(df_alcohol, bins = 25)
plt.title('Alcohol Histogram')
plt.xlabel('Alcohol')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_alcohol, showmeans = True, meanline = True)
plt.title('Alcohol Boxplot')
plt.xlabel('Alcohol')
plt.ylabel('Value')
plt.show()
```

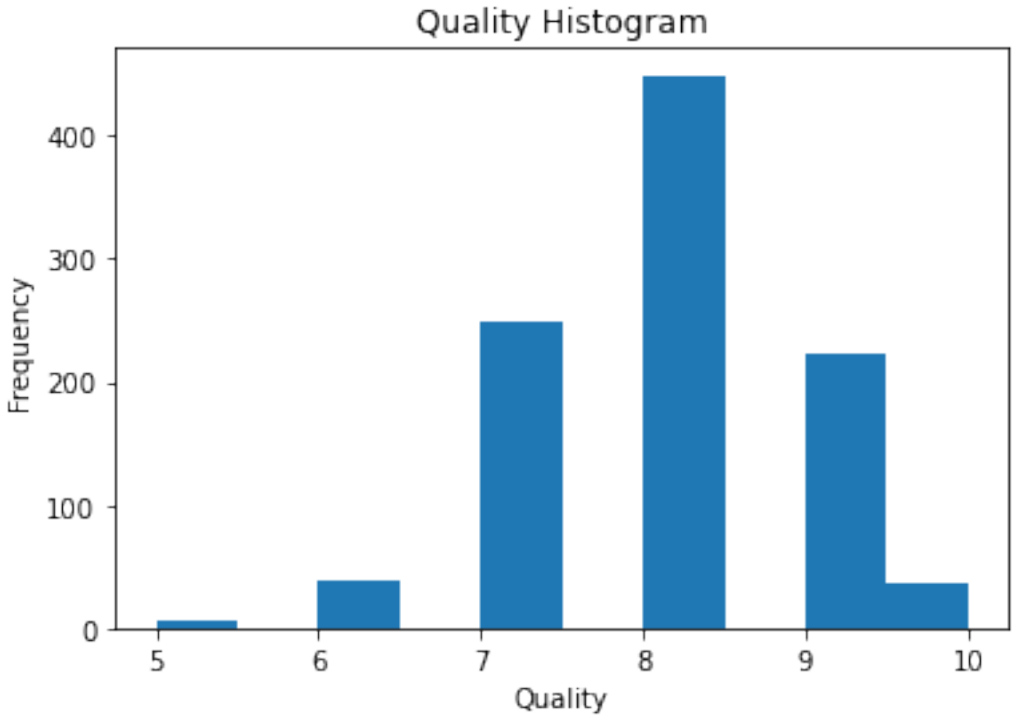

Alcohol Histogram
Frequency
0
20
40
60
80
100
6
8
10
12
14
Alcohol

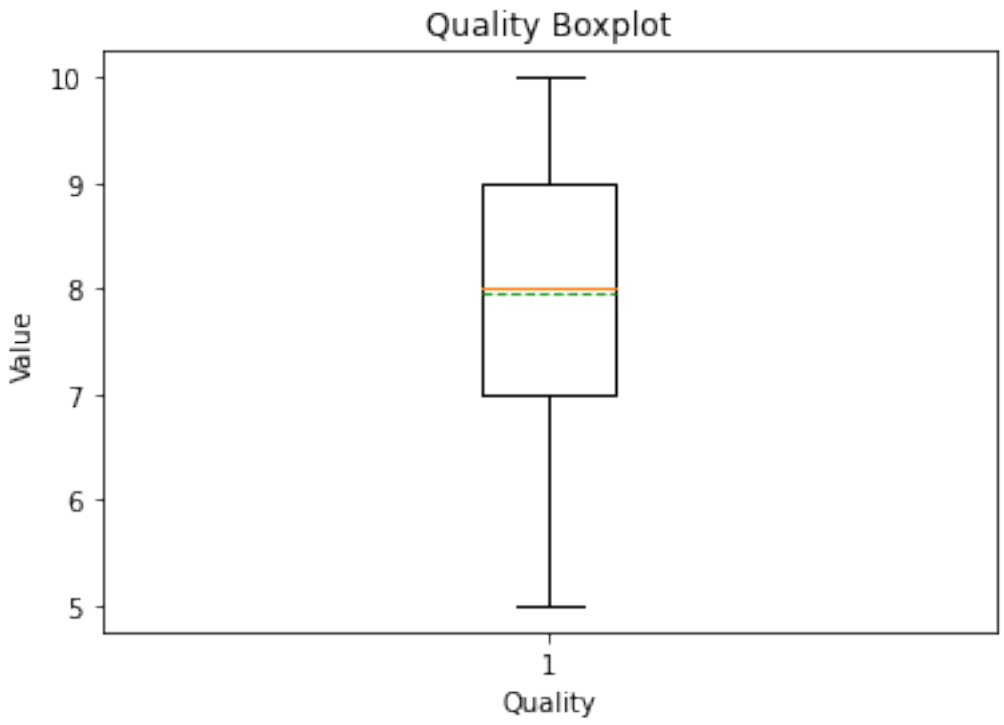

Alcohol Boxplot
Value
6
8
10
12
14
1
Alcohol

Berdasarkan histogram tersebut, terlihat bahwa distribusi Alcohol cenderung condong ke kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa beberapa data outlier berada dalam rentang sekitar 6-7 dan 14.5-15.1. Terlihat juga bahwa kuartil bawahnya berada di sekitar 9.6, kuartil tengahnya berada di sekitar 10.6, dan kuartil atasnya berada di sekitar 11.7. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 10.6 (sedikit di bawah median).

1.12 Kolom Quality

[13]:
```
df_quality = df["quality"]

# Histogram
plt.hist(df_quality, bins = 10)
plt.title('Quality Histogram')
plt.xlabel('Quality')
plt.ylabel('Frequency')
plt.show()

# Boxplot
plt.boxplot(df_quality, showmeans = True, meanline = True)
plt.title('Quality Boxplot')
plt.xlabel('Quality')
plt.ylabel('Value')
plt.show()
```





Berdasarkan histogram tersebut, terlihat bahwa distribusi Quality cenderung condong ke kiri (negatively skewed). Selain itu, berdasarkan boxplot tersebut, terlihat bahwa tidak ada data outlier. Terlihat juga bahwa kuartil bawahnya berada di 7, kuartil tengahnya berada di 8, dan kuartil atasnya berada di 9. Meannya juga terlihat mendekati mediannya (kuartil tengah), yaitu sekitar 8 (sedikit di bawah median).

P3_Normality_Test

April 17, 2023

1 Normality Test

Langkah-langkah: - Menguji normalitas dari setiap kolom A, dengan hipotesis pengujian sebagai berikut. - H_0 = kolom A berdistribusi normal - H_1 = kolom A tidak berdistribusi normal - Tingkat signifikan yang digunakan adalah $\alpha = 0.05$ - Uji statistik yang digunakan adalah normal-test (D'Agostino's K^2 test) - Pengambilan keputusan: - Tolak H_0 jika pvalue $< \alpha$ - H_0 tidak ditolak jika pvalue $\geq \alpha$

[1]:
```
# Import Libraries
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as st
import seaborn as sns

significance = 0.05

# Read csv file
df = pd.read_csv("../data/anggur.csv")
```

[2]:
```
# Print df
display(df)
```

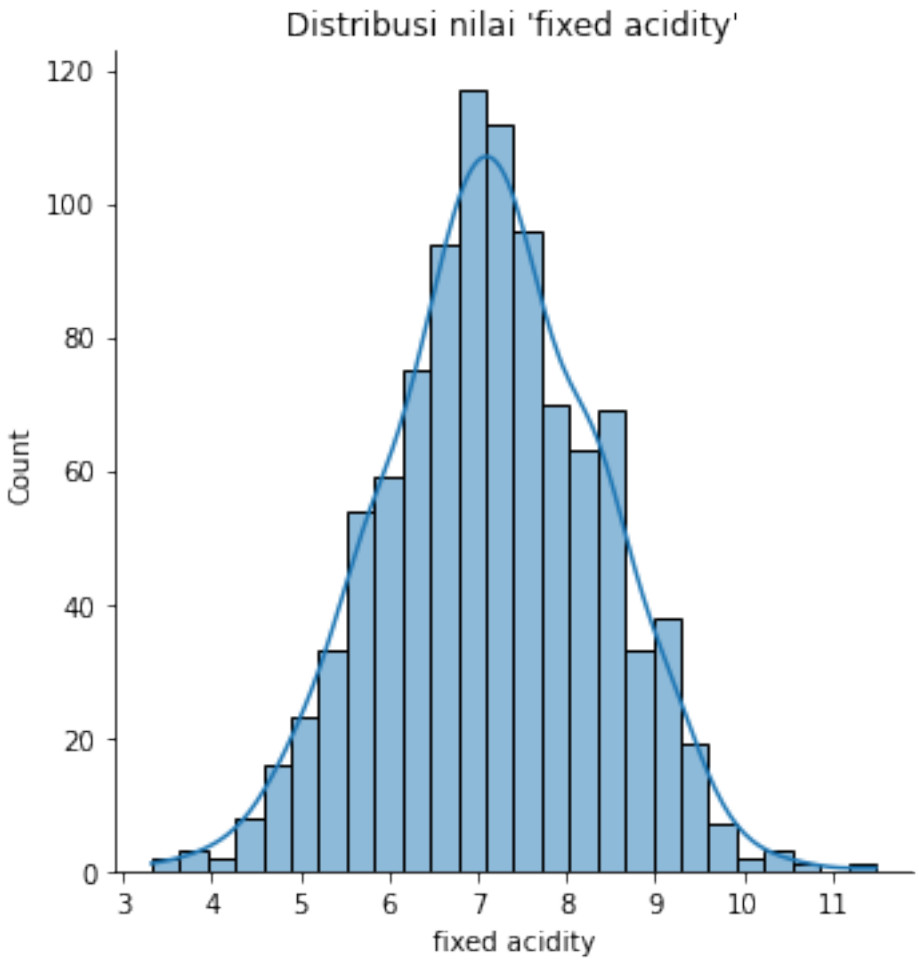
```
     fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0             5.90            0.4451       0.1813        2.049401   0.070574
1             8.40            0.5768       0.2099        3.109590   0.101681
2             7.54            0.5918       0.3248        3.673744   0.072416
3             5.39            0.4201       0.3131        3.371815   0.072755
4             6.51            0.5675       0.1940        4.404723   0.066379
..             …               …            …             …          …
995           7.96            0.6046       0.2662        1.592048   0.057555
996           8.48            0.4080       0.2227        0.681955   0.051627
997           6.11            0.4841       0.3720        2.377267   0.042806
998           7.76            0.3590       0.3208        4.294486   0.098276
999           5.87            0.5214       0.1883        2.179490   0.052923

     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0              16.593818                 42.27   0.9982  3.27       0.71
1              22.555519                 16.01   0.9960  3.35       0.57
2               9.316866                 35.52   0.9990  3.31       0.64
```

```
3              18.212300                41.97   0.9945  3.34       0.55
4               9.360591                46.27   0.9925  3.27       0.45
..                   …                   …      …   …        …
995            14.892445                44.61   0.9975  3.35       0.54
996            23.548965                25.83   0.9972  3.41       0.46
997            21.624585                48.75   0.9928  3.23       0.55
998            12.746186                44.53   0.9952  3.30       0.66
999            16.203864                24.37   0.9983  3.29       0.70

     alcohol  quality
0       8.64        7
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …      …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

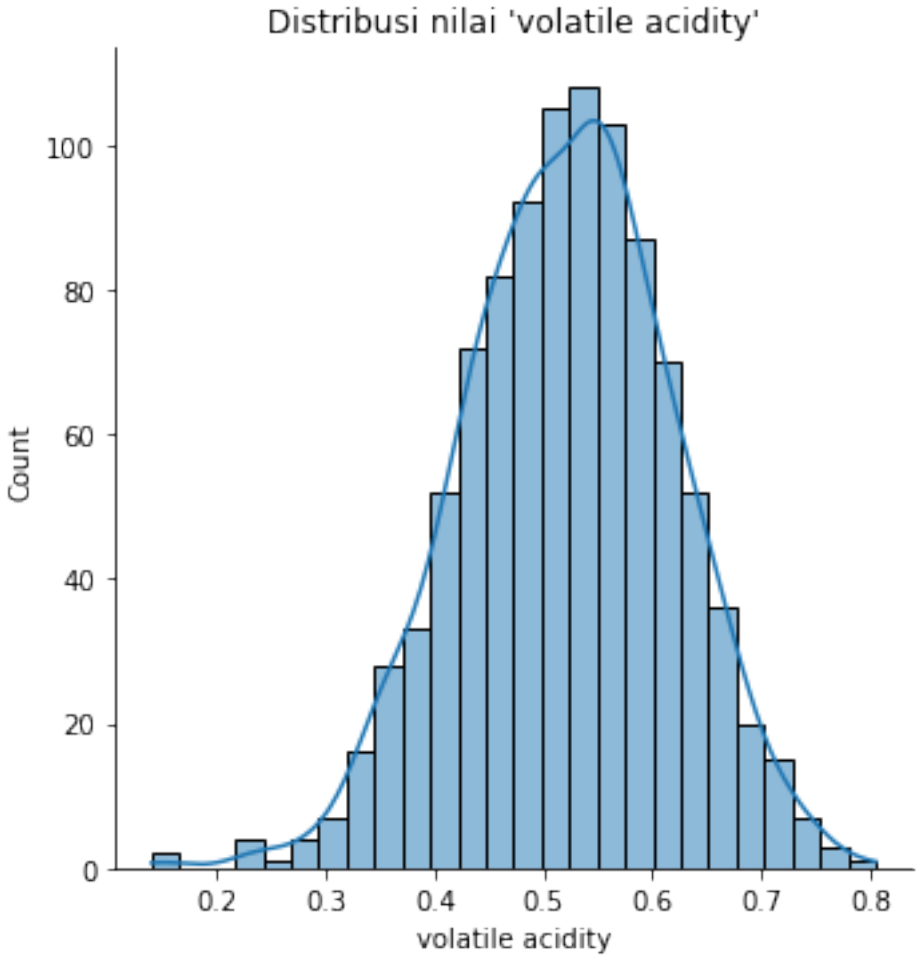
1.1 Kode Pengujian Hipotesis untuk Setiap Kolom

[3]:
```python
for column in df.columns:
    # D'Agostino's K^2 test
    stat, pvalue = st.normaltest(df[column])

    # Plot data and distribution curve
    if (column == "quality"):
        sns.histplot(df[column], discrete=True)
    else:
        sns.displot(df[column], kde=True)
    plt.title(f"Distribusi nilai '{column}'")
    plt.show()

    # K^2 Test Result
    print(f"Statistic:\n K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = {stat}")
    print(f"Two-sided Chi-Squared Probability Test:\n pvalue = {pvalue}")
    print(f"Significance:\n alpha = {significance}")

    # Hypothesis testing
    if pvalue >= significance:       # H0 not rejected
        print("\npvalue >= alpha")
```

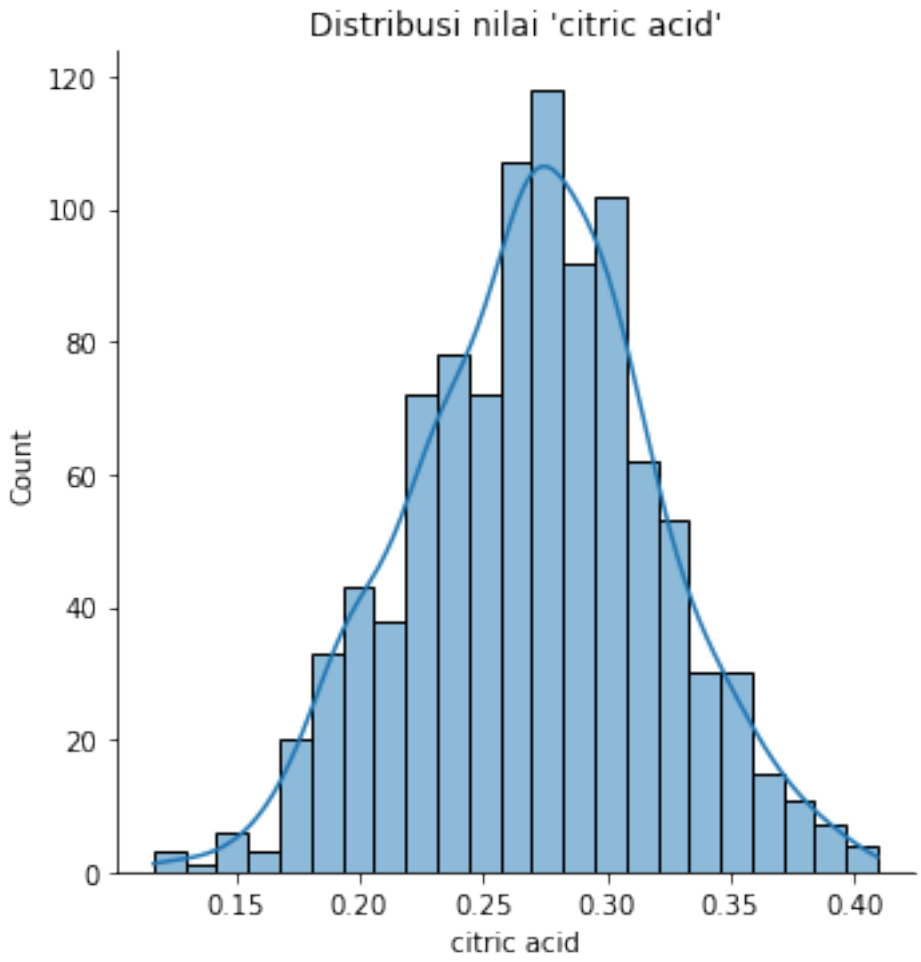
```
        verdict = f"Kesimpulan: H0 tidak ditolak, '{column}' berdistribusi␣
↪normal\n"
    else:
        print("\npvalue < alpha")
        verdict = f"Kesimpulan: H0 ditolak, '{column}' tidak berdistribusi␣
↪normal\n"
    print(verdict)
```



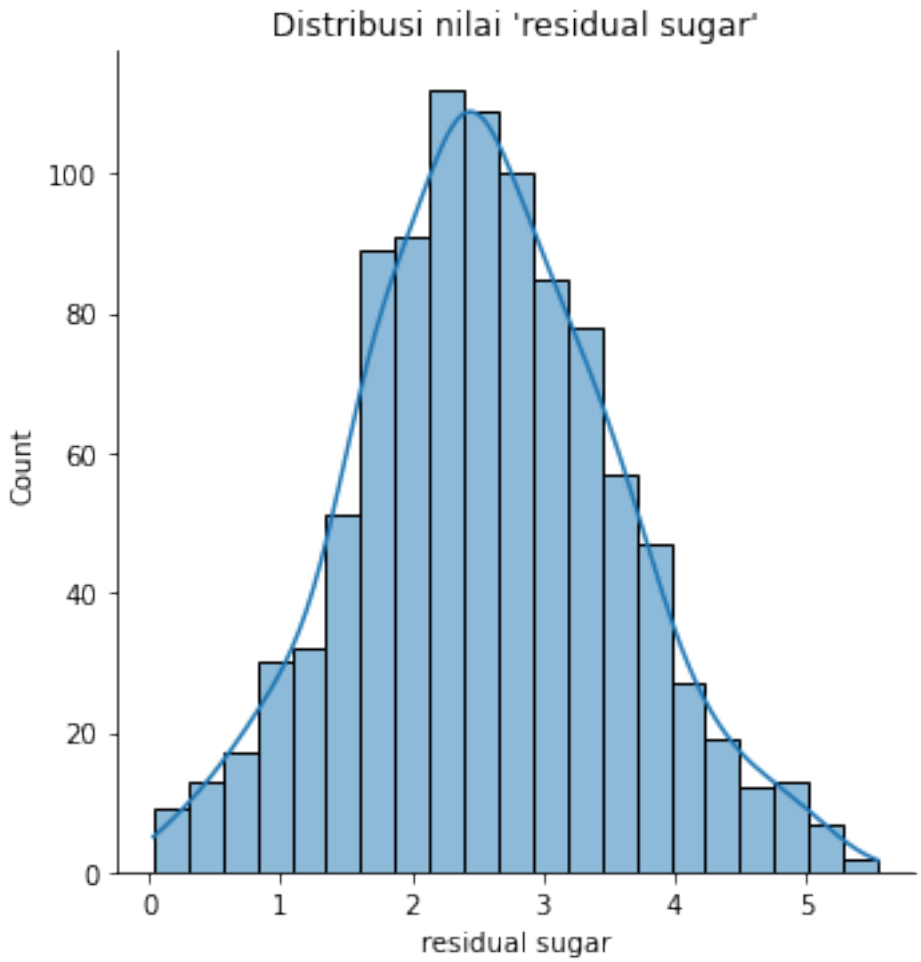

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 0.14329615661430725
Two-sided Chi-Squared Probability Test:
 pvalue = 0.9308584274486692
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'fixed acidity' berdistribusi normal
```



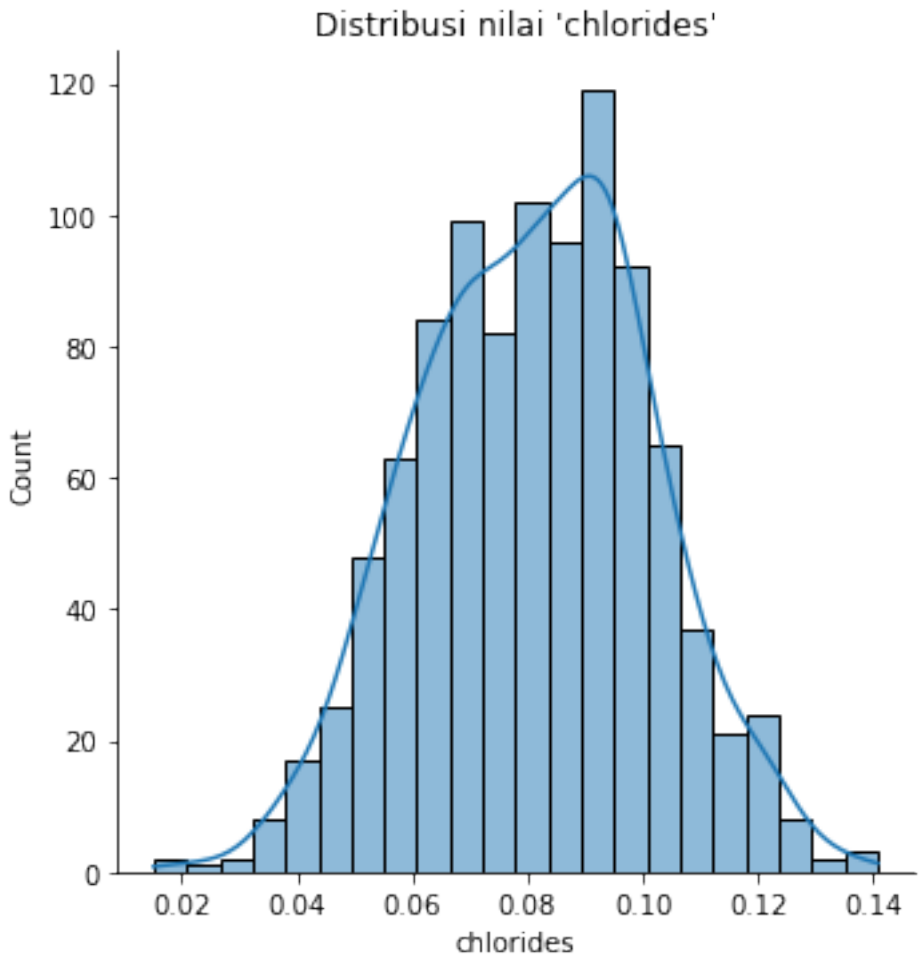

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 7.581251985533493
Two-sided Chi-Squared Probability Test:
 pvalue = 0.022581461594113835
Significance:
 alpha = 0.05

pvalue < alpha
Kesimpulan: H0 ditolak, 'volatile acidity' tidak berdistribusi normal
```



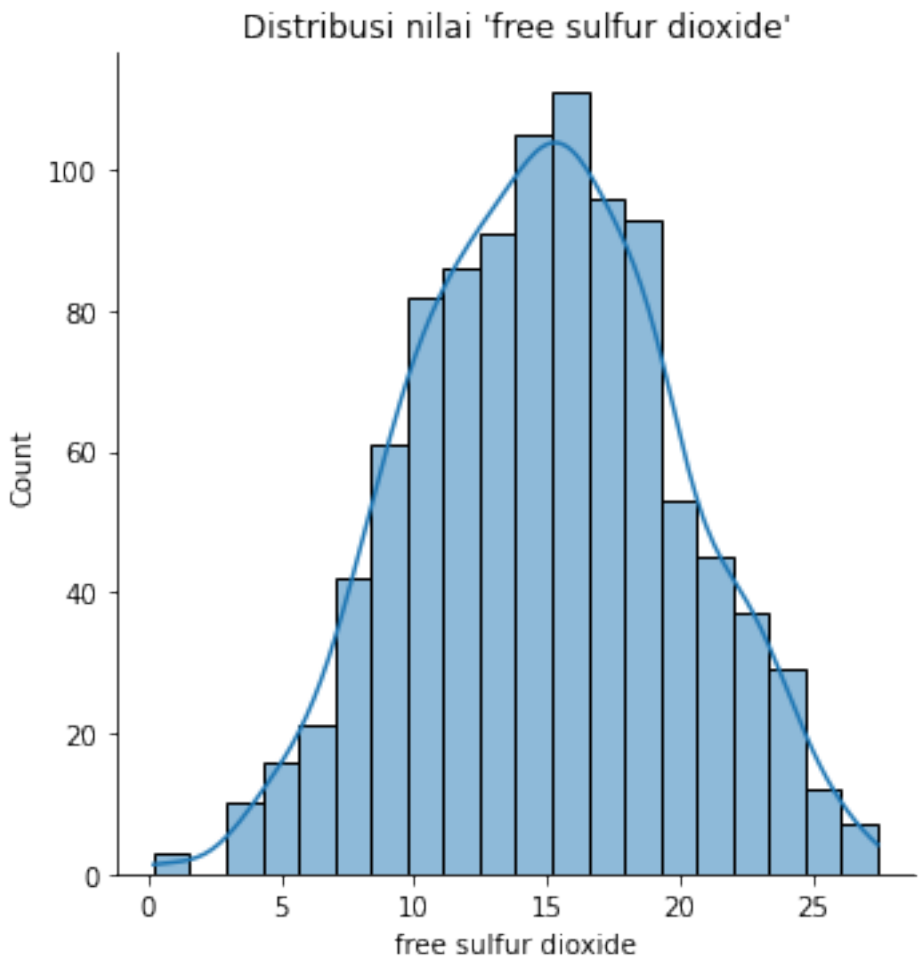

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 0.7663607229418252
Two-sided Chi-Squared Probability Test:
 pvalue = 0.6816899375976969
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'citric acid' berdistribusi normal
```



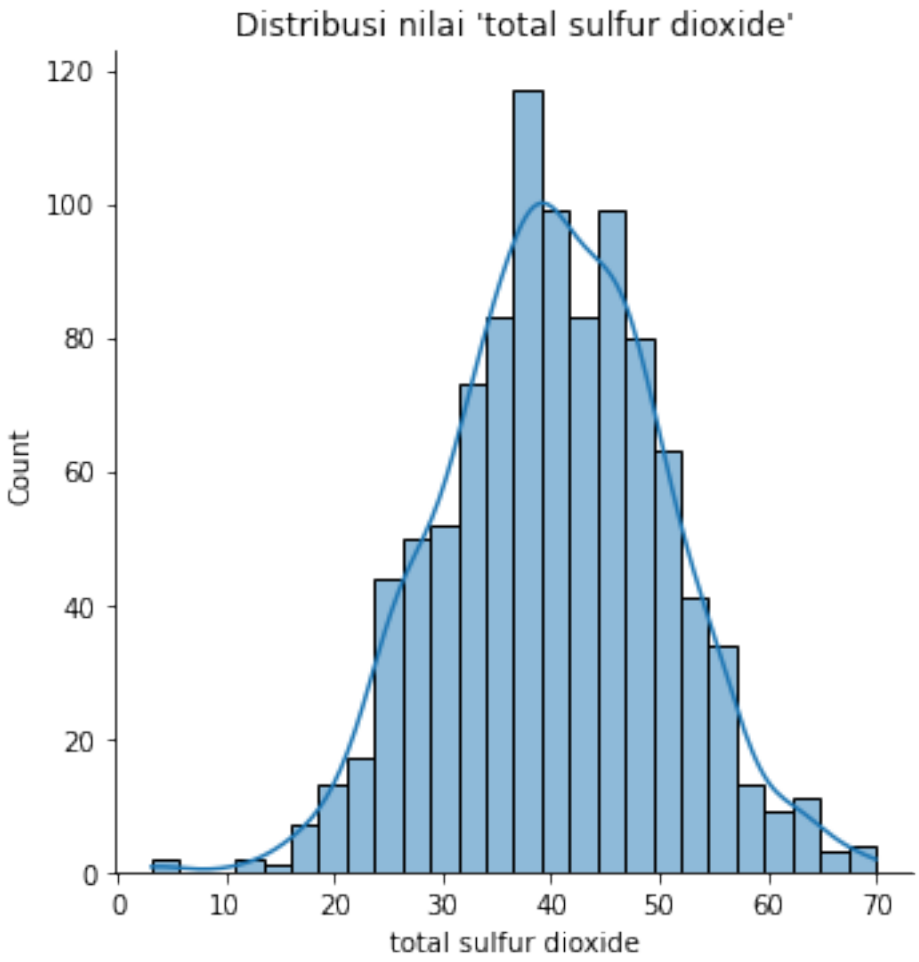

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 2.9862716504538622
Two-sided Chi-Squared Probability Test:
 pvalue = 0.22466703321310558
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'residual sugar' berdistribusi normal
```



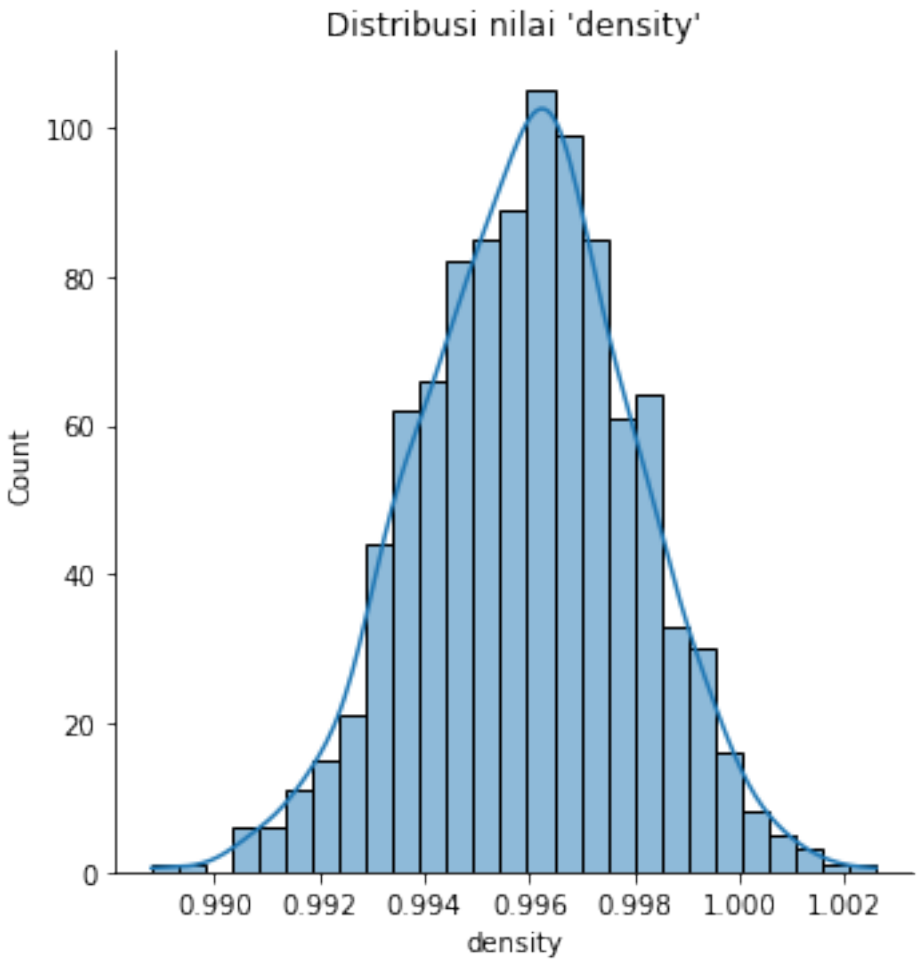

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 3.538242355484952
Two-sided Chi-Squared Probability Test:
 pvalue = 0.17048274704296862
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'chlorides' berdistribusi normal
```



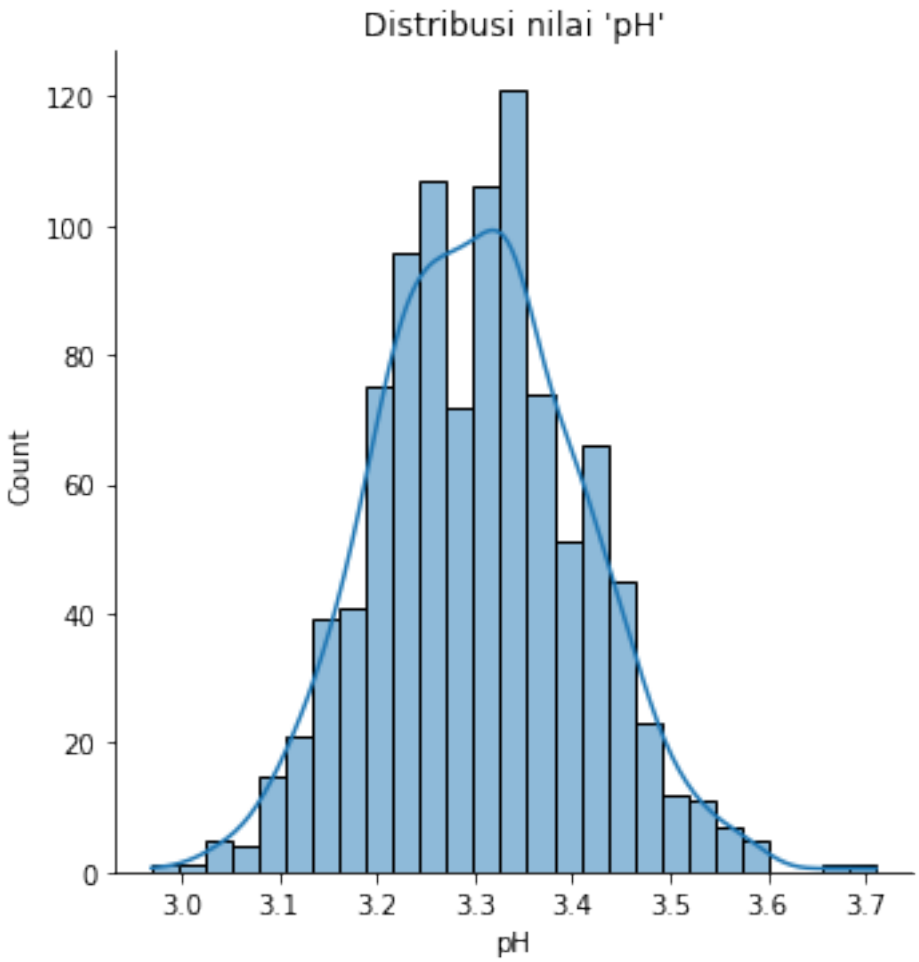

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 8.099074980855514
Two-sided Chi-Squared Probability Test:
 pvalue = 0.01743043451827735
Significance:
 alpha = 0.05

pvalue < alpha
Kesimpulan: H0 ditolak, 'free sulfur dioxide' tidak berdistribusi normal
```



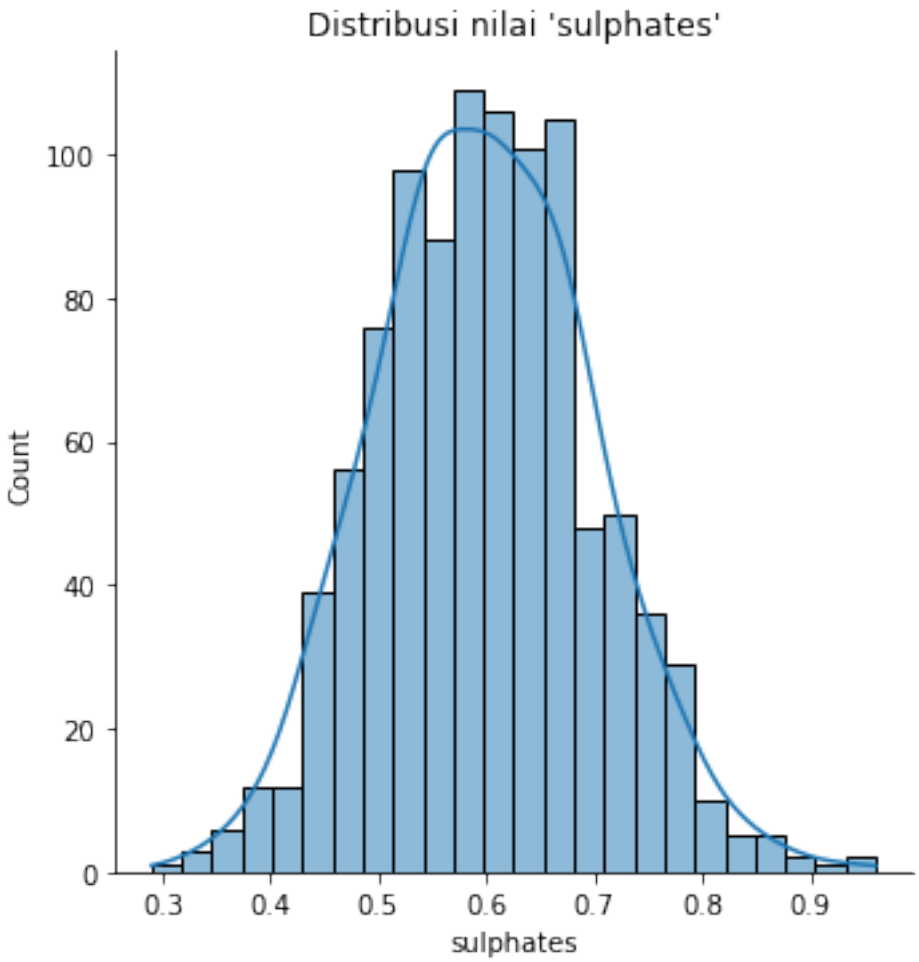

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 0.3276640291639825
Two-sided Chi-Squared Probability Test:
 pvalue = 0.8488846101395726
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'total sulfur dioxide' berdistribusi normal
```



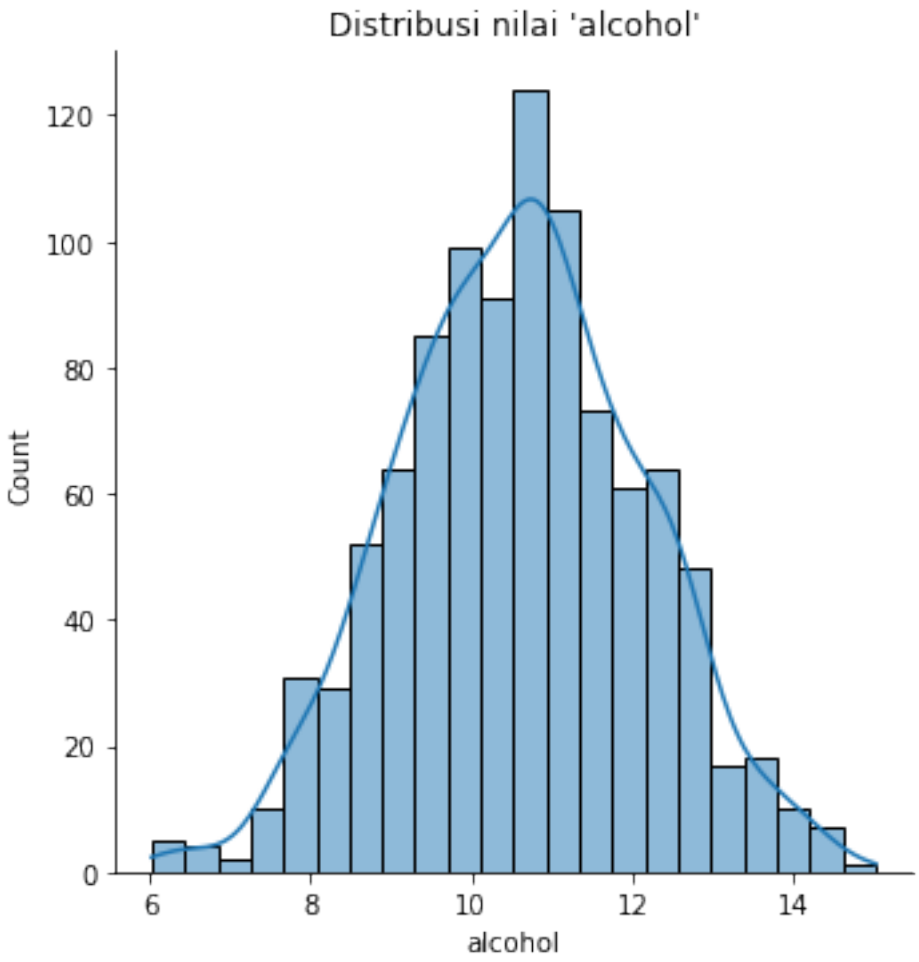

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 1.026581544320803
Two-sided Chi-Squared Probability Test:
 pvalue = 0.5985227325531981
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'density' berdistribusi normal
```



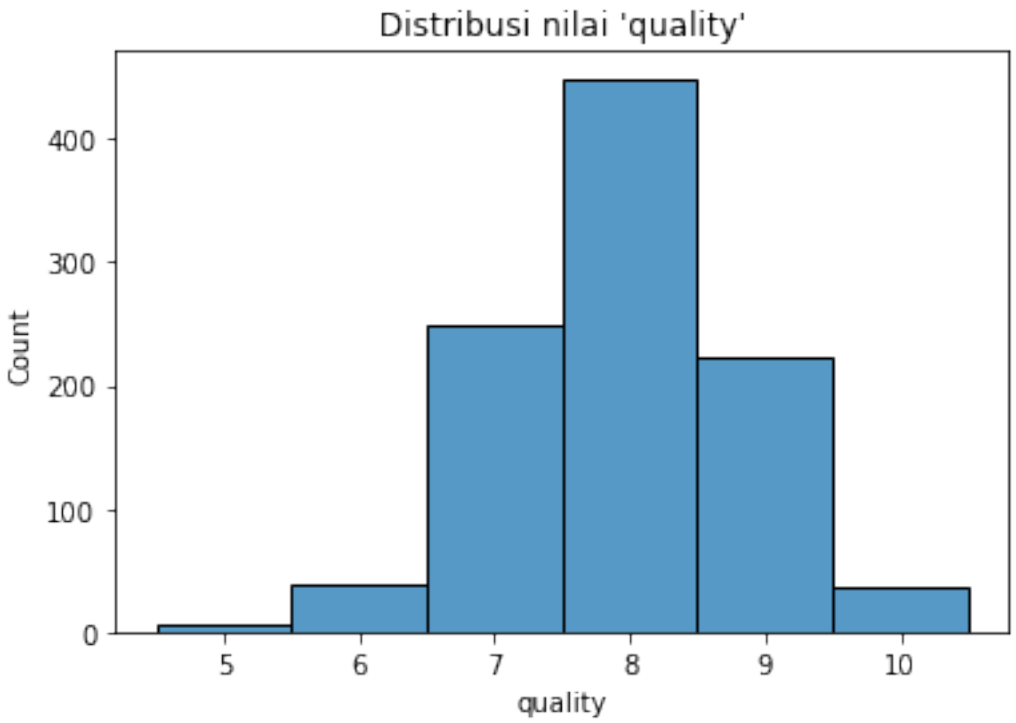

```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 3.9786546459928545
Two-sided Chi-Squared Probability Test:
 pvalue = 0.13678740824860436
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'pH' berdistribusi normal
```




```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 3.948820277859041
Two-sided Chi-Squared Probability Test:
 pvalue = 0.13884318628391681
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'sulphates' berdistribusi normal
```




```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 0.7740076714171271
Two-sided Chi-Squared Probability Test:
 pvalue = 0.6790884901361043
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'alcohol' berdistribusi normal
```




```
Statistic:
 K^2 = (Z_skew)^2 + (Z_kurtosis)^2 = 1.8893087092494893
Two-sided Chi-Squared Probability Test:
 pvalue = 0.3888139394184818
Significance:
 alpha = 0.05

pvalue >= alpha
Kesimpulan: H0 tidak ditolak, 'quality' berdistribusi normal
```

P4_One_Sample_Hypothesis_Test

April 17, 2023

1 Pengujian Hipotesis Terhadap Satu Sampel

[1]:
```
# Import Libraries
import pandas as pd
import scipy.stats as s
from statsmodels.stats.weightstats import ztest
from statsmodels.stats.proportion import proportions_ztest

# Read csv file
df = pd.read_csv("../data/anggur.csv")

display(df)
```

```
     fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0             5.90            0.4451       0.1813        2.049401   0.070574
1             8.40            0.5768       0.2099        3.109590   0.101681
2             7.54            0.5918       0.3248        3.673744   0.072416
3             5.39            0.4201       0.3131        3.371815   0.072755
4             6.51            0.5675       0.1940        4.404723   0.066379
..             …               …            …             …          …
995           7.96            0.6046       0.2662        1.592048   0.057555
996           8.48            0.4080       0.2227        0.681955   0.051627
997           6.11            0.4841       0.3720        2.377267   0.042806
998           7.76            0.3590       0.3208        4.294486   0.098276
999           5.87            0.5214       0.1883        2.179490   0.052923

     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0              16.593818                 42.27   0.9982  3.27       0.71
1              22.555519                 16.01   0.9960  3.35       0.57
2               9.316866                 35.52   0.9990  3.31       0.64
3              18.212300                 41.97   0.9945  3.34       0.55
4               9.360591                 46.27   0.9925  3.27       0.45
..                   …                     …      …   …       …
995            14.892445                 44.61   0.9975  3.35       0.54
996            23.548965                 25.83   0.9972  3.41       0.46
997            21.624585                 48.75   0.9928  3.23       0.55
998            12.746186                 44.53   0.9952  3.30       0.66
999            16.203864                 24.37   0.9983  3.29       0.70
```

```
     alcohol  quality
0       8.64        7
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …       …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

1.0.1 Langkah-Langkah Pembuktian Hipotesis:

1. Tentukan hipotesis nol H_0.
2. Tentukan hipotesis alternatif H_1.
3. Tentukan tingkat signifikan α.
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
5. Hitung nilai uji statistik dari data sample. Hitung *p-value* sesuai dengan uji statistik yang digunakan.
6. Ambil keputusan "Tolak H_0" jika nilai uji statistik terletak di daerah kritis, atau dengan tes signifikan, "Tolak H_0" jika *p-value* lebih kecil dibanding tingkat signifikansi α yang diinginkan.

1.1 Q1: Nilai rata-rata pH di atas 3.29?

Langkah-langkah: 1. $H_0 : \mu = 3.29$ 2. $H_1 : \mu > 3.29$ 3. Significance Level : $\alpha = 0.05$ 4. Uji Statistik: One-Tailed Test

```
Daerah Kritis: 1 - $\alpha$ = 0.95 dan P(z < 1.645) = 0.95 sehingga daerah kritisnya adalah z >

Perhitungannya juga ada di kode di bawah ini.
```

5. Test Statistik:

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

   Perhitungan z dan p-value ada pada kode di bawah ini.

6. Tolak H_0 jika nilai uji terletak di daerah kritis atau dengan dengan tes signifikan ($z > z_\alpha$) atau tolak H_0 jika p-value lebih kecil dibandingkan tingkat signifikansi α yang diinginkan. Jika di luar kondisi tersebut, terima H_0.

   Pengambilan keputusan tersebut ada pada kode di bawah ini.

[2]:
```
df_pH = df["pH"]

# Significance Level
```

```
alpha = 0.05

# z value and p value
z_val_pH, p_val_pH = ztest(df_pH, value = 3.29, alternative = 'larger')
print("z =", z_val_pH)

# z-alpha value
z_alpha_val_pH = s.norm.ppf(1 - alpha)
print("z-alpha =", z_alpha_val_pH)

# Pengambilan Keputusan
if (z_val_pH > z_alpha_val_pH):
    print("Nilai z lebih besar dari z-alpha sehingga nilai uji terletak di␣
 ↪daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
else:
    print("Nilai z tidak lebih besar dari z-alpha sehingga nilai uji tidak␣
 ↪terletak di daerah kritis.")
    print("Keputusan dari uji ini adalah tidak tolak H0.\n")

# p value
p_val_pH = s.norm.sf(z_val_pH)
print("p =", p_val_pH)

# Pengambilan Keputusan
if (p_val_pH < alpha):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")
```

```
z = 4.1037807933651145
z-alpha = 1.6448536269514722
Nilai z lebih besar dari z-alpha sehingga nilai uji terletak di daerah kritis.
Keputusan dari uji ini adalah tolak H0.

p = 2.0322630043302333e-05
Nilai p lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tolak H0
```

Kesimpulan:

Nilai rata-rata pH di atas 3.29.

1.2 Q2: Nilai rata-rata Residual Sugar tidak sama dengan 2.50?

Langkah-langkah: 1. $H_0 : \mu = 2.50$ 2. $H_1 : \mu \neq 2.50$ 3. Significance Level : $\alpha = 0.05$ 4. Uji Statistik: Two-Tailed Test

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

Perhitungannya juga ada di kode di bawah ini.

5. Test Statistik:

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

   Perhitungan z-value dan p-value ada pada kode di bawah ini.

6. Tolak H_0 jika ($z > z_{\alpha/2}$ atau $z < -z_{\alpha/2}$) atau tolak H_0 jika p-value lebih kecil dibandingkan tingkat signifikansi α yang diinginkan. Jika di luar kondisi tersebut, terima H_0.

   Pengambilan keputusan tersebut ada pada kode di bawah ini.

[3]:
```
df_residual_sugar = df["residual sugar"]

# Significance Level
alpha = 0.05

# z value and p value
z_val_residual_sugar, p_val_residual_sugar = ztest(df_residual_sugar, value = 2.
 ↪50, alternative = 'two-sided')

print("z =", z_val_residual_sugar)

# z-alpha value
z_alpha_val_residual_sugar = s.norm.ppf(1 - (alpha/2))
print("z-alpha =", z_alpha_val_residual_sugar)

# Pengambilan Keputusan
if (z_val_residual_sugar > z_alpha_val_residual_sugar):
    print("Nilai z lebih besar dari z-alpha/2 sehingga nilai uji terletak di␣
 ↪daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
elif (z_val_residual_sugar < z_alpha_val_residual_sugar*(-1)):
    print("Nilai z lebih kecil dari minus z-alpha/2 sehingga nilai uji terletak␣
 ↪di daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
else:
    print("Nilai z berada diantara dari minus z-alpha/2 dan z-alpha/2 sehingga␣
 ↪nilai uji tidak terletak di daerah kritis.")
    print("Keputusan dari uji ini adalah tidak tolak H0.\n")

print("p =", p_val_residual_sugar)

# Pengambilan Keputusan
if (p_val_residual_sugar < alpha):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0.")
```

```
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0.")
```

```
z = 2.1479619435539523
z-alpha = 1.959963984540054
Nilai z lebih besar dari z-alpha/2 sehingga nilai uji terletak di daerah kritis.
Keputusan dari uji ini adalah tolak H0.

p = 0.031716778818727434
Nilai p lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tolak H0.
```

Kesimpulan:

Nilai rata-rata Residual Sugar tidak sama dengan 2.50.

1.3 Q3: Nilai rata-rata 150 baris pertama kolom sulphates bukan 0.65?

Langkah-langkah: 1. $H_0 : \mu = 0.65$ 2. $H_1 : \mu \neq 0.65$ 3. Significance Level : $\alpha = 0.05$ 4. Uji Statistik: Two-Tailed Test

```
Daerah Kritis: z > $z_{\alpha/2}$ atau z < -$z_{\alpha/2}$

Perhitungannya juga ada di kode di bawah ini.
```

5. Test Statistik:

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

   Perhitungan z-value dan p-value ada pada kode di bawah ini.

6. Tolak H_0 jika ($z > z_{\alpha/2}$ atau $z < -z_{\alpha/2}$) atau tolak H_0 jika p-value lebih kecil dibandingkan tingkat signifikansi α yang diinginkan. Jika di luar kondisi tersebut, terima H_0.

   Pengambilan keputusan tersebut ada pada kode di bawah ini.

[4]:
```
df_sulphates = df["sulphates"].head(150)

# Significance Level
alpha = 0.05

# z value and p value
z_val_sulphates, p_val_sulphates = ztest(df_sulphates, value = 0.65,␣
 ↪alternative = 'two-sided')

print("z =", z_val_sulphates)

# z-alpha value
z_alpha_sulphates = s.norm.ppf(1 - (alpha/2))
print("z-alpha =", z_alpha_sulphates)
```

```
# Pengambilan Keputusan
if (z_val_sulphates > z_alpha_sulphates):
    print("Nilai z lebih besar dari z-alpha/2 sehingga nilai uji terletak di␣
 ↪daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
elif (z_val_sulphates < z_alpha_sulphates*(-1)):
    print("Nilai z lebih kecil dari minus z-alpha/2 sehingga nilai uji terletak␣
 ↪di daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
else:
    print("Nilai z berada diantara dari minus z-alpha/2 dan z-alpha/2 sehingga␣
 ↪nilai uji tidak terletak di daerah kritis.")
    print("Keputusan dari uji ini adalah tidak tolak H0.\n")

print("p =", p_val_sulphates)

# Pengambilan Keputusan
if (p_val_sulphates < alpha):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0.")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0.")
```

```
z = -4.964843393315918
z-alpha = 1.959963984540054
Nilai z lebih kecil dari minus z-alpha/2 sehingga nilai uji terletak di daerah
kritis.
Keputusan dari uji ini adalah tolak H0.

p = 6.875652918327359e-07
Nilai p lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tolak H0.
```

Kesimpulan:

Nilai rata-rata 150 baris pertama kolom sulphates bukan 0.65.

1.4 Q4: Nilai rata-rata total sulfur dioxide di bawah 35?

Langkah-langkah: 1. $H_0 : \mu = 35$ 2. $H_1 : \mu < 35$ 3. Significance Level : $\alpha = 0.05$ 4. Uji Statistik: One-Tailed Test

```
Daerah Kritis: 1 - $\alpha$ = 0.95 dan P(z > -1.645) = 0.95 sehingga daerah kritisnya adalah z

Perhitungannya juga ada di kode di bawah ini.
```

5. Test Statistik:

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

   Perhitungan z-value dan p-value ada pada kode di bawah ini.

6. Tolak H_0 jika nilai uji terletak di daerah kritis atau dengan dengan tes signifikan ($z < -z_\alpha$) atau tolak H_0 jika p-value lebih kecil dibandingkan tingkat signifikansi α yang diinginkan. Jika di luar kondisi tersebut, terima H_0.

   Pengambilan keputusan tersebut ada pada kode di bawah ini.

[5]:
```
df_total_sulfur_dioxide_1 = df["total sulfur dioxide"]

# Significance Level
alpha = 0.05

# z value and p value
z_val_total_sulfur_dioxide_1, p_val_total_sulfur_dioxide_1 =␣
 ↪ztest(df_total_sulfur_dioxide_1, value = 35, alternative = 'smaller')
print("z =", z_val_total_sulfur_dioxide_1)

# z-alpha value
z_alpha_total_sulfur_dioxide_1 = -s.norm.ppf(1 - alpha)
print("z-alpha =", z_alpha_total_sulfur_dioxide_1)

# Pengambilan Keputusan
if (z_val_total_sulfur_dioxide_1 < z_alpha_total_sulfur_dioxide_1):
    print("Nilai z lebih kecil dari minus z-alpha sehingga nilai uji terletak␣
 ↪di daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
else:
    print("Nilai z tidak lebih kecil dari minus z-alpha sehingga nilai uji␣
 ↪tidak terletak di daerah kritis.")
    print("Keputusan dari uji ini adalah tidak tolak H0.\n")

print("p =", p_val_total_sulfur_dioxide_1)

# Pengambilan Keputusan
if (p_val_total_sulfur_dioxide_1 < alpha):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")
```

```
z = 16.786387372296744
z-alpha = -1.6448536269514722
Nilai z tidak lebih kecil dari minus z-alpha sehingga nilai uji tidak terletak
```

```
di daerah kritis.
Keputusan dari uji ini adalah tidak tolak H0.

p = 1.0
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0
```

Kesimpulan:

Nilai rata-rata total sulfur dioxide tidak di bawah 35.

1.5 Q5: Proporsi nilai total Sulfat Dioxide yang lebih dari 40, adalah tidak sama dengan 50% ?

Langkah-langkah: 1. H_0: Proporsi nilai total Sulfat Dioxide yang lebih dari 40 sama dengan 50% (p = 0.5) 2. H_1: Proporsi nilai total Sulfat Dioxide yang lebih dari 40 tidak sama dengan 50% (p $\neq$ 0.5) 3. Significance Level : $\alpha = 0.05$ 4. Uji Statistik: Two-Tailed Test

```
Daerah Kritis: 1 - $\alpha$ = 0.95 dan P(z < 1.645) = 0.95 sehingga daerah kritisnya adalah z >

Perhitungannya juga ada di kode di bawah ini.
```

5. Test Statistik:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

   Perhitungan z-value dan p-value ada pada kode di bawah ini.

6. Tolak H_0 jika ($z > z_{\alpha/2}$ atau $z < -z_{\alpha/2}$) atau tolak H_0 jika p-value lebih kecil dibandingkan tingkat signifikansi α yang diinginkan. Jika di luar kondisi tersebut, terima H_0.

   Pengambilan keputusan tersebut ada pada kode di bawah ini.

[6]:
```
df_total_sulfur_dioxide_2 = df[df["total sulfur dioxide"] > 40]

# Significance Level
alpha = 0.05

# z value and p value
z_val_total_sulfur_dioxide_2, p_val_total_sulfur_dioxide_2 =␣
 ↪proportions_ztest(len(df_total_sulfur_dioxide_2),len(df), value = 0.5,␣
 ↪prop_var = 0.5, alternative = 'two-sided')

print("z =", z_val_total_sulfur_dioxide_2)

# z-alpha value
z_alpha_total_sulfur_dioxide_2 = s.norm.ppf(1 - (alpha/2))
print("z-alpha =", z_alpha_total_sulfur_dioxide_2)

# Pengambilan Keputusan
if (z_val_total_sulfur_dioxide_2 > z_alpha_total_sulfur_dioxide_2):
```

```
    print("Nilai z lebih besar dari z-alpha/2 sehingga nilai uji terletak di␣
 ↪daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
elif (z_val_total_sulfur_dioxide_2 < z_alpha_total_sulfur_dioxide_2*(-1)):
    print("Nilai z lebih kecil dari minus z-alpha/2 sehingga nilai uji terletak␣
 ↪di daerah kritis.")
    print("Keputusan dari uji ini adalah tolak H0.\n")
else:
    print("Nilai z berada diantara dari minus z-alpha/2 dan z-alpha/2 sehingga␣
 ↪nilai uji tidak terletak di daerah kritis.")
    print("Keputusan dari uji ini adalah tidak tolak H0.\n")

print("p =", p_val_total_sulfur_dioxide_2)

# Pengambilan Keputusan
if (p_val_total_sulfur_dioxide_2 < alpha):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0.")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0.")
```

```
z = 0.7589466384404118
z-alpha = 1.959963984540054
Nilai z berada diantara dari minus z-alpha/2 dan z-alpha/2 sehingga nilai uji
tidak terletak di daerah kritis.
Keputusan dari uji ini adalah tidak tolak H0.

p = 0.4478844782641115
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0.
```

Kesimpulan:

Proporsi nilai total Sulfat Dioxide yang lebih dari 40, adalah tidak berbeda dengan 50%.

P5_Two_Sample_Hypothesis_Test

April 17, 2023

1 Pengujian Hipotesis Terhadap Dua Sampel

[1]:
```
# Import Libraries
import pandas as pd
import scipy.stats as st
import statsmodels.stats.weightstats as ws
from statsmodels.stats.proportion import proportions_ztest

# Read csv file
df = pd.read_csv("../data/anggur.csv")

display(df)
```

```
     fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0             5.90            0.4451       0.1813        2.049401   0.070574
1             8.40            0.5768       0.2099        3.109590   0.101681
2             7.54            0.5918       0.3248        3.673744   0.072416
3             5.39            0.4201       0.3131        3.371815   0.072755
4             6.51            0.5675       0.1940        4.404723   0.066379
..             …                 …            …               …          …
995           7.96            0.6046       0.2662        1.592048   0.057555
996           8.48            0.4080       0.2227        0.681955   0.051627
997           6.11            0.4841       0.3720        2.377267   0.042806
998           7.76            0.3590       0.3208        4.294486   0.098276
999           5.87            0.5214       0.1883        2.179490   0.052923

     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0              16.593818                 42.27   0.9982  3.27       0.71
1              22.555519                 16.01   0.9960  3.35       0.57
2               9.316866                 35.52   0.9990  3.31       0.64
3              18.212300                 41.97   0.9945  3.34       0.55
4               9.360591                 46.27   0.9925  3.27       0.45
..                   …                     …        …     …          …
995            14.892445                 44.61   0.9975  3.35       0.54
996            23.548965                 25.83   0.9972  3.41       0.46
997            21.624585                 48.75   0.9928  3.23       0.55
998            12.746186                 44.53   0.9952  3.30       0.66
999            16.203864                 24.37   0.9983  3.29       0.70
```

```
     alcohol  quality
0       8.64        7
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …       …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

1.0.1 Langkah-Langkah Pembuktian Hipotesis:

1. Tentukan hipotesis nol H_0.
2. Tentukan hipotesis alternatif H_1.
3. Tentukan tingkat signifikan α.
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
5. Hitung nilai uji statistik dari data sample. Hitung *p-value* sesuai dengan uji statistik yang digunakan.
6. Ambil keputusan "Tolak H_0" jika nilai uji statistik terletak di daerah kritis, atau dengan tes signifikan, "Tolak H_0" jika *p-value* lebih kecil dibanding tingkat signifikansi α yang diinginkan.

1.1 Q1: Data kolom fixed acidity dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata kedua bagian tersebut sama?

Sampel pengujian: - sampel_1: bagian awal kolom 'fixed acidity' - sampel_2: bagian akhir kolom 'fixed acidity'

Langkah-langkah: 1. H_0: $\mu_1 - \mu_2 = 0$ (rata-rata kedua sampel sama) 2. H_1: $\mu_1 - \mu_2 \neq 0$ (rata-rata kedua sampel berbeda) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Standar deviasi populasi (σ) dari kedua sampel diketahui sama karena diambil dari populasi yang sama - Uji hipotesis adalah *two-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$z = \frac{(\bar{x_1} - \bar{x_2}) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

- Daerah kritis adalah $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ 5. Perhitungan nilai uji statistik z ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ - H_0 tidak ditolak jika $-z_{\alpha/2} \leq z \leq z_{\alpha/2}$

[2]:
```
# Sample setup
fixed_acidity = df['fixed acidity']
fixed_acidity_sample_1 = fixed_acidity[:len(fixed_acidity)//2]
fixed_acidity_sample_2 = fixed_acidity[len(fixed_acidity)//2:]
```

```
# Test statistic calculation
diff = 0
significance = 0.05

z_value_1, ztest_pvalue_1 = ws.ztest(fixed_acidity_sample_1,␣
 ↪fixed_acidity_sample_2, value=diff)

z_alpha_over_2 = st.norm.ppf(1 - significance/2)

# Drawing a conclusion
print(f"Critical region: z < {-z_alpha_over_2} or z > {z_alpha_over_2}")
print(f"Test statistic: z = {z_value_1}")
print(f"p-value = {ztest_pvalue_1}")
print()
if (z_value_1 < -z_alpha_over_2 or z_value_1 > z_alpha_over_2):
    print("Nilai z berada dalam critical region")
    verdict = "H0 ditolak, rata-rata sampel 1 tidak sama dengan rata-rata␣
 ↪sampel 2"
else:
    print("Nilai z berada di luar critical region")
    verdict = "H0 tidak ditolak, rata-rata sampel 1 sama dengan rata-rata␣
 ↪sampel 2"

if (ztest_pvalue_1 < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")

print("\nKesimpulan: " + verdict)
```

```
Critical region: z < -1.959963984540054 or z > 1.959963984540054
Test statistic: z = 0.02604106999906379
p-value = 0.9792245804254097

Nilai z berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, rata-rata sampel 1 sama dengan rata-rata sampel 2
```

1.2 Q2: Data kolom chlorides dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata bagian awal lebih besar daripada bagian akhir sebesar 0.001?

Sampel pengujian: - sampel_1: bagian awal kolom 'chlorides' - sampel_2: bagian akhir kolom 'chlorides'

Langkah-langkah: 1. H_0: $\mu_1 - \mu_2 = 0.001$ (rata-rata bagian awal lebih besar daripada bagian akhir sebesar 0.001) 2. H_1: $\mu_1 - \mu_2 \neq 0.001$ (selisih rata-rata bagian awal dengan bagian akhir bukan 0.001) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Standar deviasi populasi (σ) dari kedua sampel diketahui sama karena diambil dari populasi yang sama - Uji hipotesis adalah *two-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$z = \frac{(\bar{x_1} - \bar{x_2}) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

- Daerah kritis adalah $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ 5. Perhitungan nilai uji statistik z ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ - H_0 tidak ditolak jika $-z_{\alpha/2} \leq z \leq z_{\alpha/2}$

[3]:
```
# Sample setup
chlorides = df['chlorides']
chlorides_sample_1 = chlorides[:len(chlorides)//2]
chlorides_sample_2 = chlorides[len(chlorides)//2:]

# Test statistic calculation
diff = 0.001
significance = 0.05

z_value_2, ztest_pvalue_2 = ws.ztest(chlorides_sample_1, chlorides_sample_2,␣
 ↪value=diff)

z_alpha_over_2 = st.norm.ppf(1 - significance/2)

# Drawing a conclusion
print(f"Critical region: z < {-z_alpha_over_2} or z > {z_alpha_over_2}")
print(f"Test statistic: z = {z_value_2}")
print(f"p-value = {ztest_pvalue_2}")
print()
if (z_value_2 < -z_alpha_over_2 or z_value_2 > z_alpha_over_2):
    print("Nilai z berada dalam critical region")
    verdict = "H0 ditolak, selisih rata-rata sampel 1 dan sampel 2 tidak sama␣
 ↪dengan 0.001"
else:
    print("Nilai z berada di luar critical region")
    verdict = "H0 tidak ditolak, rata-rata sampel 1 lebih besar dari rata-rata␣
 ↪sampel 2 sebanyak 0.001"

if (ztest_pvalue_2 < significance):
```

```
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
Critical region: z < -1.959963984540054 or z > 1.959963984540054
Test statistic: z = -0.467317122852132
p-value = 0.640273007581107

Nilai z berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, rata-rata sampel 1 lebih besar dari rata-rata
sampel 2 sebanyak 0.001
```

1.3 Q3: Benarkah rata-rata sampel 25 baris pertama kolom Volatile Acidity sama dengan rata-rata 25 baris pertama kolom Sulphates ?

Sampel pengujian: - sampel_1: 25 baris pertama kolom 'volatile acidity' - sampel_2: 25 baris pertama kolom 'sulphates'

Langkah-langkah: 1. H_0: $\mu_1 - \mu_2 = 0$ (rata-rata kedua sampel sama) 2. H_1: $\mu_1 - \mu_2 \neq 0$ (rata-rata kedua sampel berbeda) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Standar deviasi populasi (σ) dari kedua sampel diketahui berbeda - Uji hipotesis adalah *two-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$t = \frac{(\bar{x_1} - \bar{x_2}) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- Daerah kritis adalah $t < -t_{\alpha/2}$ atau $t > t_{\alpha/2}$ 5. Perhitungan nilai uji statistik t ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $t < -t_{\alpha/2}$ atau $t > t_{\alpha/2}$ - H_0 tidak ditolak jika $-t_{\alpha/2} \leq t \leq t_{\alpha/2}$

[4]:
```
# Sample setup
volatile_acidity = df['volatile acidity']
sample_1_volatile_acidity = volatile_acidity[:25]

sulphates = df['sulphates']
sample_2_sulphates = sulphates[:25]

# Test statistic calculation
diff = 0
```

```
significance = 0.05

t_value, ttest_pvalue, dof = ws.ttest_ind(sample_1_volatile_acidity,␣
 ↪sample_2_sulphates, value=diff)

t_alpha_over_2 = st.t.ppf(1 - significance/2, dof)

# Drawing a conclusion
print(f"Critical region: t < {-t_alpha_over_2} or t > {t_alpha_over_2}")
print(f"Degree of Freedom: v = {dof}")
print(f"Test statistic: t = {t_value}")
print(f"p-value = {ttest_pvalue}")
print()
if (t_value < -t_alpha_over_2 or t_value > t_alpha_over_2):
    print("Nilai t berada dalam critical region")
    verdict = "H0 ditolak, rata-rata sampel 1 tidak sama dengan rata-rata␣
 ↪sampel 2"
else:
    print("Nilai t berada di luar critical region")
    verdict = "H0 tidak ditolak, rata-rata sampel 1 sama dengan rata-rata␣
 ↪sampel 2"

if (ttest_pvalue < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
Critical region: t < -2.0106347546964454 or t > 2.0106347546964454
Degree of Freedom: v = 48.0
Test statistic: t = -2.6374821676748703
p-value = 0.011223058174680032

Nilai t berada dalam critical region
Nilai p lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tolak H0

Kesimpulan: H0 ditolak, rata-rata sampel 1 tidak sama dengan rata-rata sampel 2
```

1.4 Q4: Bagian awal kolom residual sugar memiliki variansi yang sama dengan bagian akhirnya?

Sampel pengujian: - sampel_1: bagian awal dari kolom ‘residual sugar’ - sampel_2: bagian akhir dari kolom ‘residual sugar’

Langkah-langkah: 1. H_0: $\sigma_1^2 = \sigma_2^2$ (variansi kedua sampel sama) 2. H_1: $\sigma_1^2 \neq \sigma_2^2$ (variansi kedua sampel berbeda) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Uji hipotesis adalah *two-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$f = \frac{s_1^2}{s_2^2}$$

- Daerah kritis adalah $f < f_{1-\alpha/2}(v_1, v_2)$ atau $f > f_{\alpha/2}(v_1, v_2)$ 5. Perhitungan nilai uji statistik f ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $f < f_{1-\alpha/2}(v_1, v_2)$ atau $f > f_{\alpha/2}(v_1, v_2)$ - H_0 tidak ditolak jika $f_{1-\alpha/2}(v_1, v_2) \leq f \leq f_{\alpha/2}(v_1, v_2)$

[5]:
```
# Sample setup
residual_sugar = df['chlorides']
residual_sugar_sample_1 = residual_sugar[:len(residual_sugar)//2]
residual_sugar_sample_2 = residual_sugar[len(residual_sugar)//2:]

# Hypothesis testing setup
sample_1_variance = residual_sugar_sample_1.var(ddof=1)
sample_2_variance = residual_sugar_sample_2.var(ddof=1)
print(f"Sample_1 variance: s1^2 = {sample_1_variance}")
print(f"Sample_2 variance: s2^2 = {sample_2_variance}")
print()

# Test statistic calculation
diff = 0
significance = 0.05

f_value = sample_1_variance / sample_2_variance


# f-distribution test critical points, note: ppf accepts left-side percentage
f_left_tail = st.f.ppf(1-(1 - significance/2), len(residual_sugar_sample_1)-1,␣
 ↪len(residual_sugar_sample_2)-1)
f_right_tail = st.f.ppf(1-(significance/2), len(residual_sugar_sample_1)-1,␣
 ↪len(residual_sugar_sample_2)-1)
f_test_pvalue = st.f.cdf(f_value, len(residual_sugar_sample_1)-1,␣
 ↪len(residual_sugar_sample_2)-1)

# Drawing a conclusion
print(f"Critical region: f < {f_left_tail} or f > {f_right_tail}")
print(f"Test statistic: f = {f_value}")
print(f"p-value = {f_test_pvalue}")
print()
if (f_value < f_left_tail or f_value > f_right_tail):
    print("Nilai f berada dalam critical region")
    verdict = "H0 ditolak, variansi kedua sampel berbeda"
else:
    print("Nilai f berada di luar critical region")
```

```
    verdict = "H0 tidak ditolak, variansi kedua sampel sama"

if (f_test_pvalue < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
Sample_1 variance: s1^2 = 0.00040667352898471836
Sample_2 variance: s2^2 = 0.00040293091542206646

Critical region: f < 0.8388857772763105 or f > 1.1920574017201653
Test statistic: f = 1.0092884745731947
p-value = 0.5411032946184126

Nilai f berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, variansi kedua sampel sama
```

1.5 Q5: Proporsi nilai setengah bagian awal alcohol yang lebih dari 7, adalah lebih besar daripada, proporsi nilai yang sama di setengah bagian akhir alcohol?

Sampel pengujian: - sampel_1: bagian awal dari kolom 'alcohol' yang bernilai lebih dari 7 - sampel_2: bagian akhir dari kolom 'alcohol' yang bernilai lebih dari 7

Langkah-langkah: 1. H_0: $p_1 - p_2 = 0$ (proporsi kedua sampel sama) 2. H_1: $p_1 - p_2 > 0$ (proporsi sampel pertama lebih besar dari proporsi sampel kedua) 3. Penentuan tingkat signifikan: $\alpha = 0.05$ 4. Penentuan uji statistik dan daerah kritis: - Uji hipotesis adalah *one-tailed test* - Oleh karena itu, rumus pengujian yang digunakan adalah sebagai berikut

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

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

- Daerah kritis adalah $z > z_\alpha$ 5. Perhitungan nilai uji statistik z ada pada kode di bawah ini. 6. Pengambilan keputusan: - Tolak H_0 jika $z > z_\alpha$ - H_0 tidak ditolak jika $z \leq z_\alpha$

[6]:
```
# Sample setup
alcohol = df['alcohol']
alcohol_sample_1 = alcohol[:len(alcohol)//2]
alcohol_sample_2 = alcohol[len(alcohol)//2:]
```

```python
# Filter sample to greater than 7
alcohol_sample_1_gt7 = alcohol_sample_1[alcohol_sample_1 > 7]
alcohol_sample_2_gt7 = alcohol_sample_2[alcohol_sample_2 > 7]

# Hypothesis testing setup
x1_x2 = [len(alcohol_sample_1_gt7), len(alcohol_sample_2_gt7)]
n1_n2 = [len(alcohol_sample_1), len(alcohol_sample_2)]
print(f"x1, x2 = {x1_x2}")
print(f"n1, n2 = {n1_n2}")

# Test statistic calculation
diff = 0
significance = 0.05
z_value_5, proportion_ztest_pvalue = proportions_ztest(x1_x2, n1_n2,␣
 ↪value=diff, alternative="larger")

z_alpha = st.norm.ppf(1 - significance)


# Drawing a conclusion
print(f"Critical region: z > {z_alpha}")
print(f"Test statistic: z = {z_value_5}")
print(f"p-value = {proportion_ztest_pvalue}")
print()
if (z_value_5 > z_alpha):
    print("Nilai z berada dalam critical region")
    verdict = "H0 ditolak, proporsi sampel 1 lebih besar dari proporsi sampel 2"
else:
    print("Nilai z berada di luar critical region")
    verdict = "H0 tidak ditolak, proporsi sampel 1 sama dengan proporsi sampel␣
 ↪2"

if (proportion_ztest_pvalue < significance):
    print("Nilai p lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tolak H0")
else:
    print("Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan")
    print("Keputusan dari uji ini adalah tidak tolak H0")


print("\nKesimpulan: " + verdict)
```

```
x1, x2 = [495, 495]
n1, n2 = [500, 500]
Critical region: z > 1.6448536269514722
Test statistic: z = 0.0
```

```
p-value = 0.5

Nilai z berada di luar critical region
Nilai p tidak lebih kecil dari tingkat signifikansi yang diinginkan
Keputusan dari uji ini adalah tidak tolak H0

Kesimpulan: H0 tidak ditolak, proporsi sampel 1 sama dengan proporsi sampel 2
```