K02-T1-IF2220-13521082

April 18, 2023

1 Tugas Besar IF2220 Probabilitas dan Statistika

Farizki Kurniawan - 13521082

Frankie Huang - 13521092

1.1 Table of Contents

[1]:
```
library(moments)
```

[2]:
```
# Read and set data

data <- read.csv("anggur.csv")

columns <- colnames(data)
```

[3]:
```
# Data frame to hold descriptive statistics

statistics <- data.frame (
    Value = c(
        "Mean",
        "Median",
        "Mode",
        "Standard Deviation",
        "Variance", "Range",
        "Minimum", "Maximum",
        "First Quartile",
        "Second Quartile",
        "Third Quartile",
        "Interquartile Range",
        "Skewness",
        "Kurtosis"
```

```
    )
)
```

1.2 Soal 1

Menulis deskripsi statistika (*Descriptive Statistics*) dari semua kolom pada data yang bersifat numerik, terdiri *mean*, *median*, *modus*, *standar deviasi*, *variansi*, *range*, *minimum value*, *maximum value*, *quartile*, *IQR*, *skewness*, dan *kurtosis*

[4]:
```r
# Insert descriptive statistics value from column into data_frame

insert_values <- function(column, data_frame) {
    # Column
    column_value <- names(column)

    column <- sort(as.numeric(unlist(column)))

    # Mean
    mean_value <- mean(column)

    # Median
    median_value <- median(column)

    # Mode
    mode_value <- names(sort(-table(column)))[1]

    # Standard Deviation
    standard_deviation_value <- sd(column)

    # Variance
    variance_value <- var(column)

    # Range
    range_value <- column[length(column)] - column[1]

    # Minimum
    minimum_value <- column[1]

    # Maximum
    maximum_value <- column[length(column)]

    # Quartile
    first_quartile_value <- quantile(column, 0.25)
    second_quartile_value <- quantile(column, 0.5)
    third_quartile_value <- quantile(column, 0.75)

    # IQR
```

```
    iqr_value <- third_quartile_value - first_quartile_value

    # Skewness
    skewness_value <- skewness(column)

    # Kurtosis
    kurtosis_value <- kurtosis(column)

    # Insert value
    data_frame[column_value] <- c(
        mean_value,
        median_value,
        mode_value,
        standard_deviation_value,
        variance_value,
        range_value,
        minimum_value,
        maximum_value,
        first_quartile_value,
        second_quartile_value,
        third_quartile_value,
        iqr_value,
        skewness_value,
        kurtosis_value
    )

    return (data_frame)
}
```

[60]:
```
for (column in columns) {
    statistics <- insert_values(data[column], statistics)
}

statistics[1:3]
statistics[4:6]
statistics[6:8]
statistics[8:10]
statistics[10:11]
statistics[12:13]
```

A data.frame: 14 × 3

Value <chr>	fixed.acidity <chr>	volatile.acidity <chr>
Mean	7.15253	0.5208385
Median	7.15	0.52485
Mode	6.54	0.5546
Standard Deviation	1.20159757649383	0.0958482740553495
Variance	1.44383673583584	0.00918689163938939
Range	8.17	0.6652
Minimum	3.32	0.1399
Maximum	11.49	0.8051
First Quartile	6.3775	0.4561
Second Quartile	7.15	0.52485
Third Quartile	8	0.585375
Interquartile Range	1.6225	0.129275
Skewness	-0.0288352396076068	-0.197402026913042
Kurtosis	2.9748101987292	3.15505076502095

A data.frame: 14 × 3

citric.acid <chr>	residual.sugar <chr>	chlorides <chr>
0.270517	2.56710368250676	0.0811951525078498
0.2722	2.51943027286579	0.0821669021645236
0.3019	0.032554525015195	0.0151224391657095
0.0490983714707635	0.987915436504693	0.0201106472439967
0.00241065008108108	0.975976909684258	0.000404438132572474
0.2929	5.51820040970786	0.125635130265349
0.1167	0.032554525015195	0.0151224391657095
0.4096	5.55075493472306	0.140757569431058
0.2378	1.89632994348868	0.0665736319097736
0.2722	2.51943027286579	0.0821669021645236
0.302325	3.22087348282979	0.0953115014855626
0.064525	1.3245435393411	0.028737869575789
-0.0455076660920009	0.132439046103759	-0.0512422863775635
2.88984940842673	2.95124055916741	2.74872867443544

A data.frame: 14 × 3

chlorides <chr>	free.sulfur.dioxide <chr>	total.sulfur.dioxide <chr>
0.0811951525078498	14.9076792510298	40.29015
0.0821669021645236	14.8603462365689	40.19
0.0151224391657095	0.194678523326937	35.2
0.0201106472439967	4.88809970575656	9.9657673762183
0.000404438132572474	23.8935187334174	99.3165193968969
0.125635130265349	27.2678469010989	66.81
0.0151224391657095	0.194678523326937	3.15
0.140757569431058	27.4625254244258	69.96
0.0665736319097736	11.4267169494576	33.785
0.0821669021645236	14.8603462365689	40.19
0.0953115014855626	18.313097915395	47.0225
0.028737869575789	6.88638096593739	13.2375
-0.0512422863775635	0.00711971590752342	-0.024023921723998
2.74872867443544	2.63086461587304	3.05763648155113

A data.frame: 14 × 3

total.sulfur.dioxide <chr>	density <chr>	pH <chr>
40.29015	0.9959253	3.30361
40.19	0.996	3.3
35.2	0.9959	3.34
9.9657673762183	0.00202018094264871	0.104875482200402
99.3165193968969	4.08113104104104e-06	0.0109988667667668
66.81	0.0137999999999999	0.74
3.15	0.9888	2.97
69.96	1.0026	3.71
33.785	0.9946	3.23
40.19	0.996	3.3
47.0225	0.9972	3.37
13.2375	0.00259999999999994	0.14
-0.024023921723998	-0.076767416885835	0.147450993855577
3.05763648155113	3.01028990166223	3.07451156408369

A data.frame: 14 × 2

pH <chr>	sulphates <chr>
3.30361	0.59839
3.3	0.595
3.34	0.59
0.104875482200402	0.100819007991412
0.0109988667667668	0.0101644723723724
0.74	0.67
2.97	0.29
3.71	0.96
3.23	0.53
3.3	0.595
3.37	0.67
0.14	0.14
0.147450993855577	0.148975009203956
3.07451156408369	3.05850163281688

A data.frame: 14 × 2

alcohol <chr>	quality <chr>
10.59228	7.958
10.61	8
9.86	8
1.51070600522876	0.902801778382747
2.28223263423423	0.815051051051051
8.99	5
6.03	5
15.02	10
9.56	7
10.61	8
11.6225	9
2.0625	2
-0.0189629053366952	-0.088920454390281
2.86293217101661	3.10175629881938

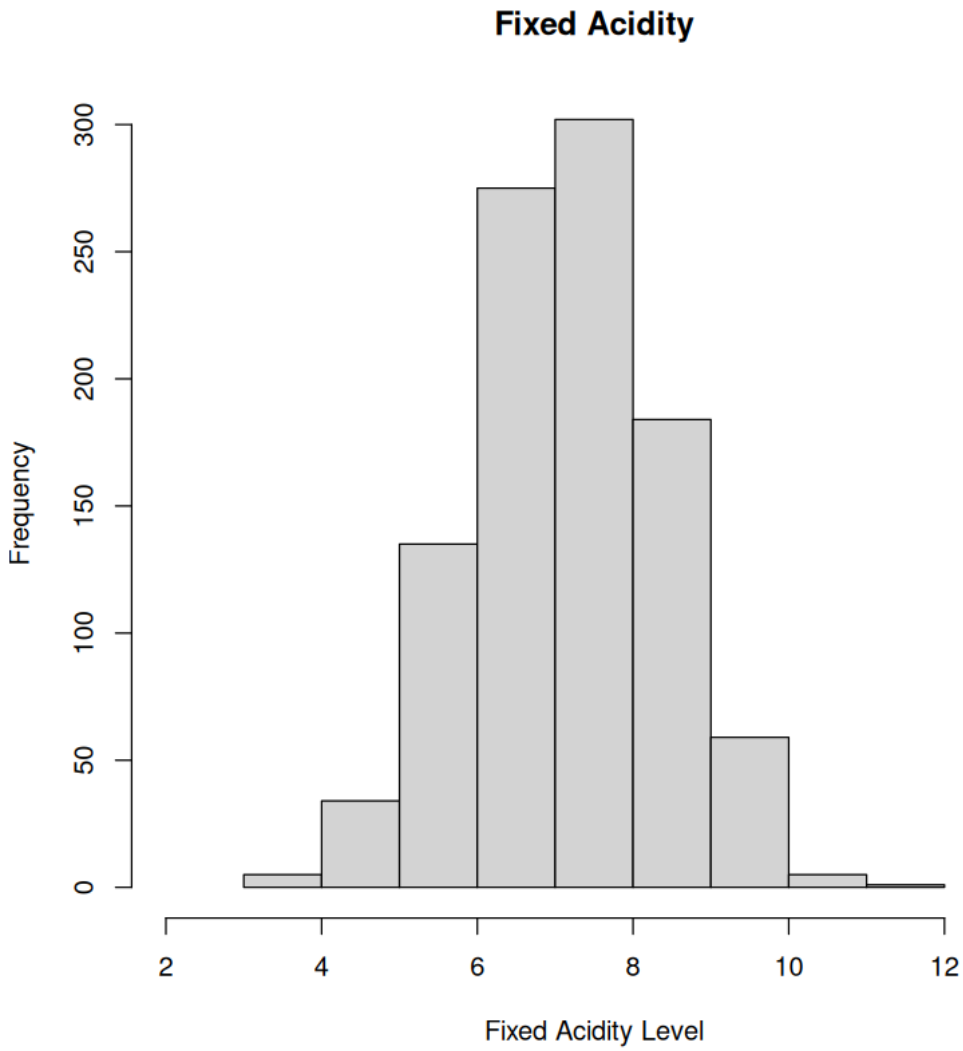
1.3 Soal 2

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

[6]:
```r
# Visualization function

histogram_plot <- function(column, main_title, x_title, x_limit = NULL) {
    column <- sort(as.numeric(unlist(column)))

    if (is.null(x_limit)) {
        min_value <- floor(column[1])
        max_value <- ceiling(column[length(column)])
        x_limit <- c(min_value, max_value)
```

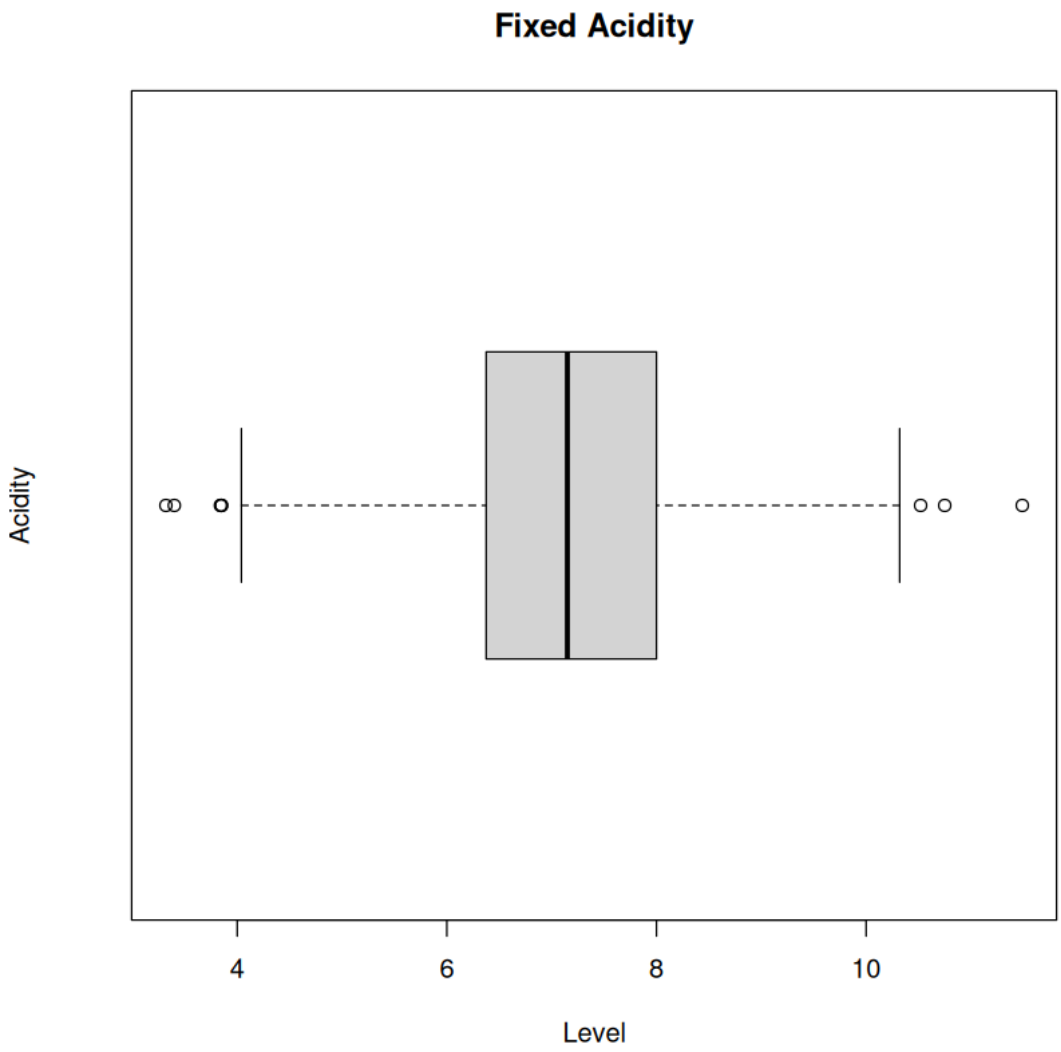
```
    }

    hist (
        x = column,
        main = main_title,
        xlab = x_title,
        xlim = x_limit
    )
}

box_plot <- function(column, main_title, x_title, y_title) {
    boxplot (
        column,
        main = main_title,
        xlab = x_title,
        ylab = y_title,
        horizontal = TRUE
    )
}
```

[7]:
```
# Fixed Acidity

histogram_plot(data["fixed.acidity"], "Fixed Acidity", "Fixed Acidity Level",␣
 ↪c(2, 13))
```



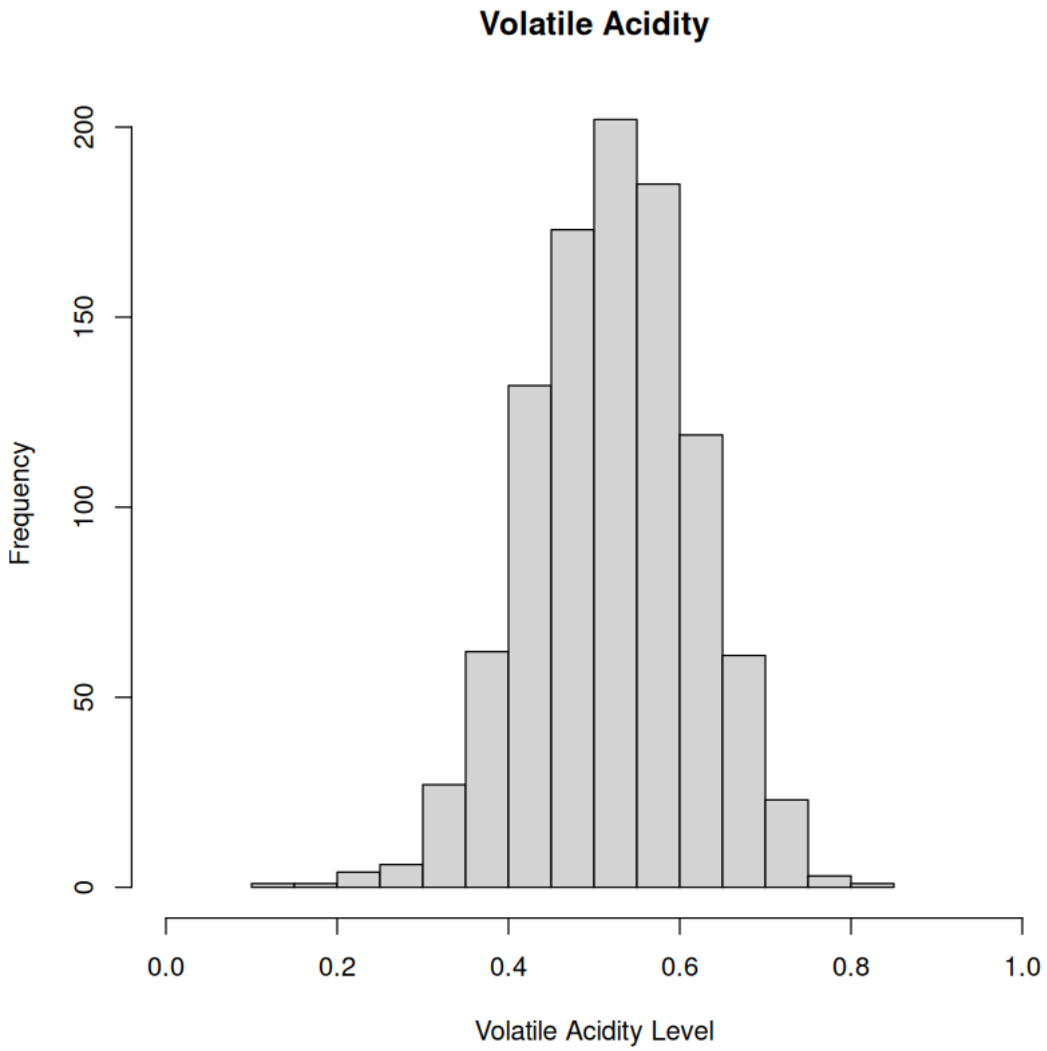
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 7-8, nilai minimum di rentang 3-4 dan nilai maksimum di rentang 11-12. Selain itu, plot menunjukkan bahwa grafik memiliki skewness hampir mendekati nol.

[8]:
```
box_plot(data["fixed.acidity"], "Fixed Acidity", "Level", "Acidity")
```



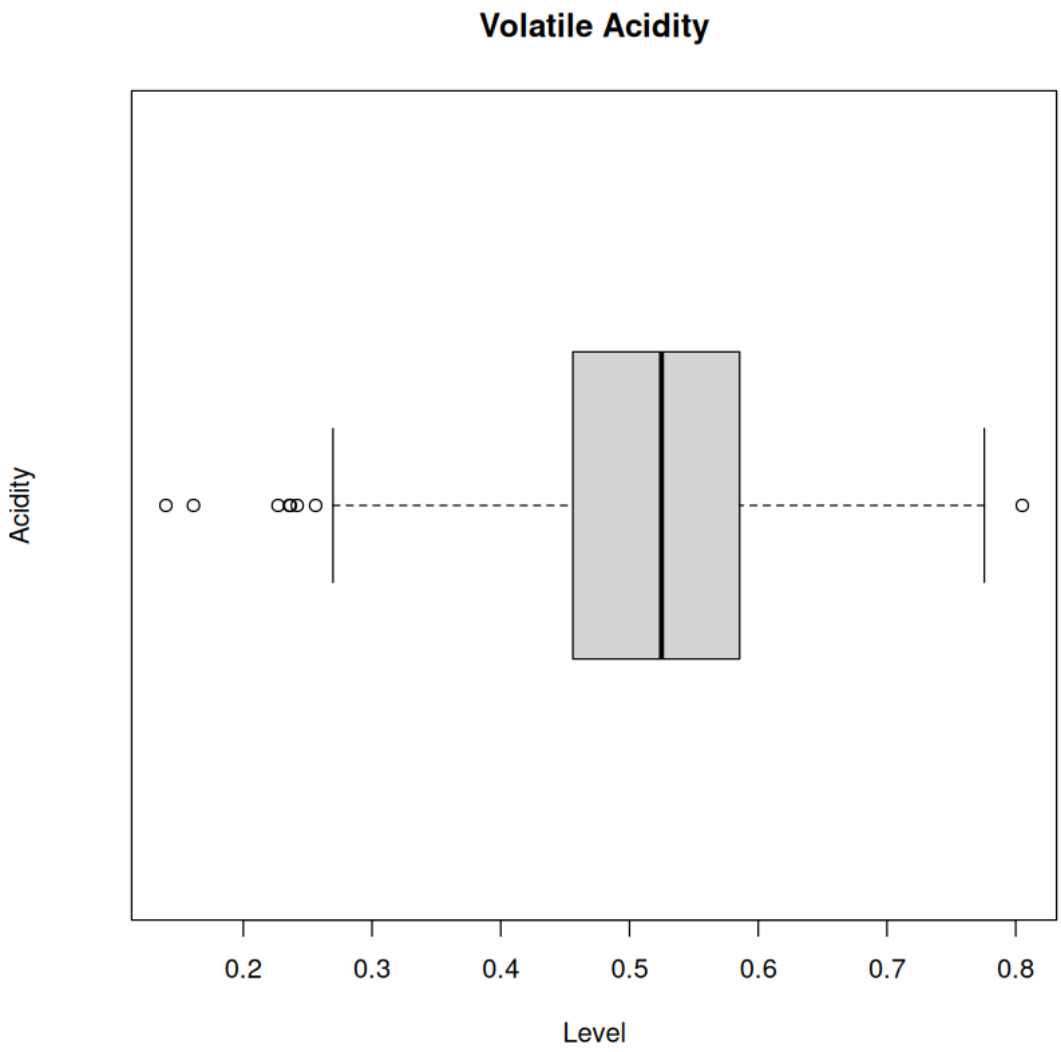
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 6-7, Q2 pada rentang 6-7, dan Q3 pada rentang 7-8. Selain itu, terdapat 3 nilai outlier pada bagian kiri dan 3 nilai outlier pada bagian kanan plot.

[9]:
```
# Volatile Acidity

histogram_plot(data["volatile.acidity"], "Volatile Acidity", "Volatile Acidity␣
↪Level", c(0, 1))
```



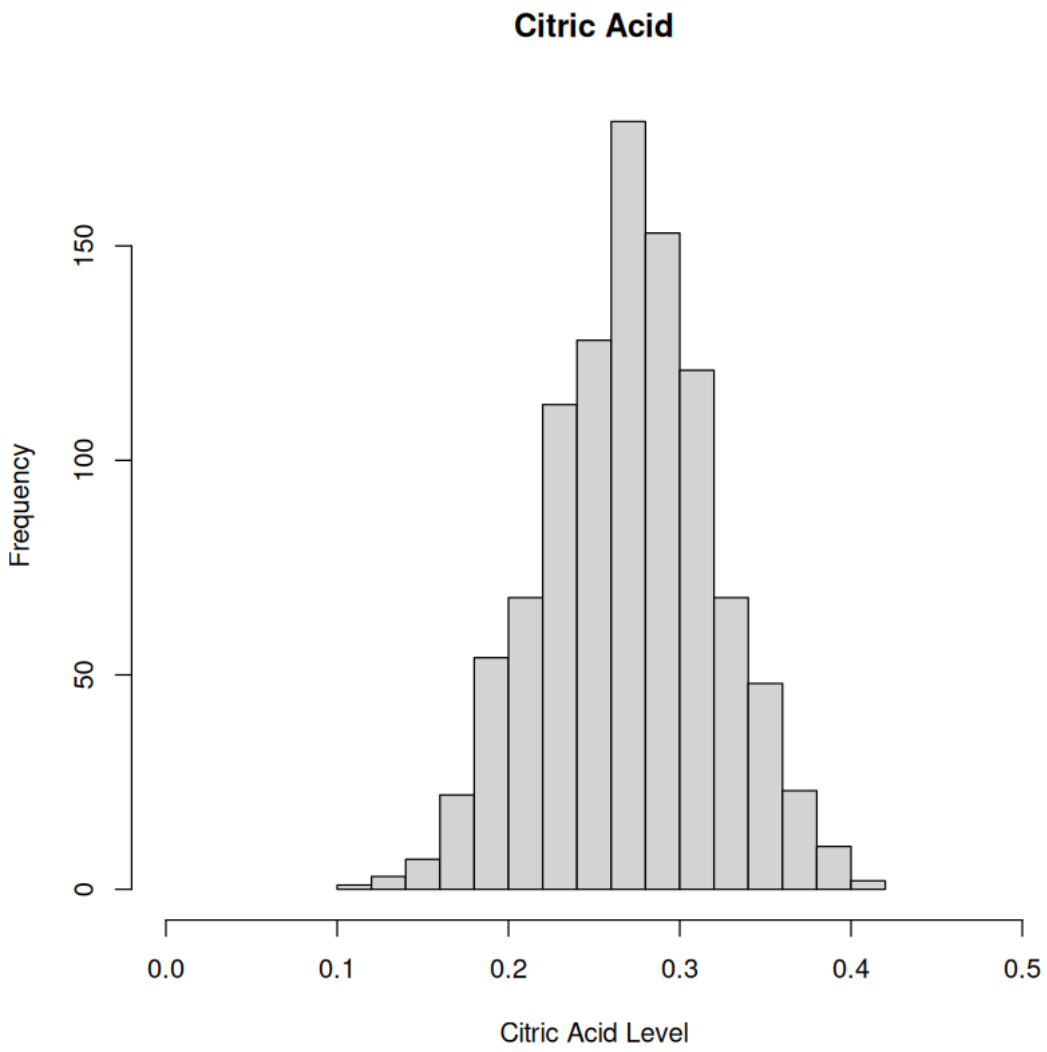
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 0.50 - 0.55, nilai minimum di rentang 0.10-0.15 dan nilai maksimum di rentang 0.80-0.85. Selain itu, plot menunjukkan bahwa grafik memiliki skewness negatif.

[10]:
```
box_plot(data["volatile.acidity"], "Volatile Acidity", "Level", "Acidity")
```



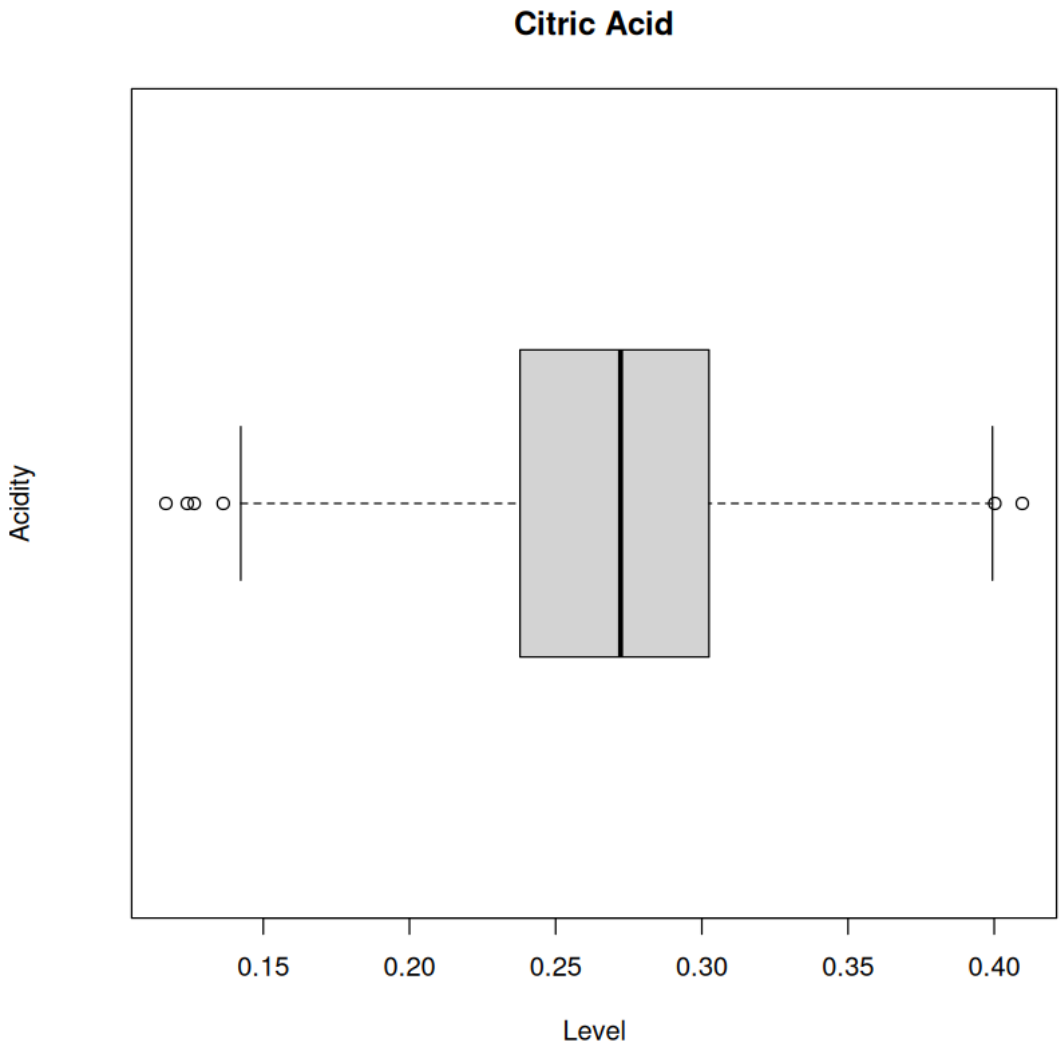
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 0.4-0.5, Q2 pada rentang 0.5-0.6, dan Q3 pada rentang 0.5-0.6. Selain itu, terdapat 6 nilai outlier pada bagian kiri dan 1 nilai outlier pada bagian kanan plot.

[11]:
```
# Citric Acid

histogram_plot(data["citric.acid"], "Citric Acid", "Citric Acid Level", c(0, 0.
↪5))
```



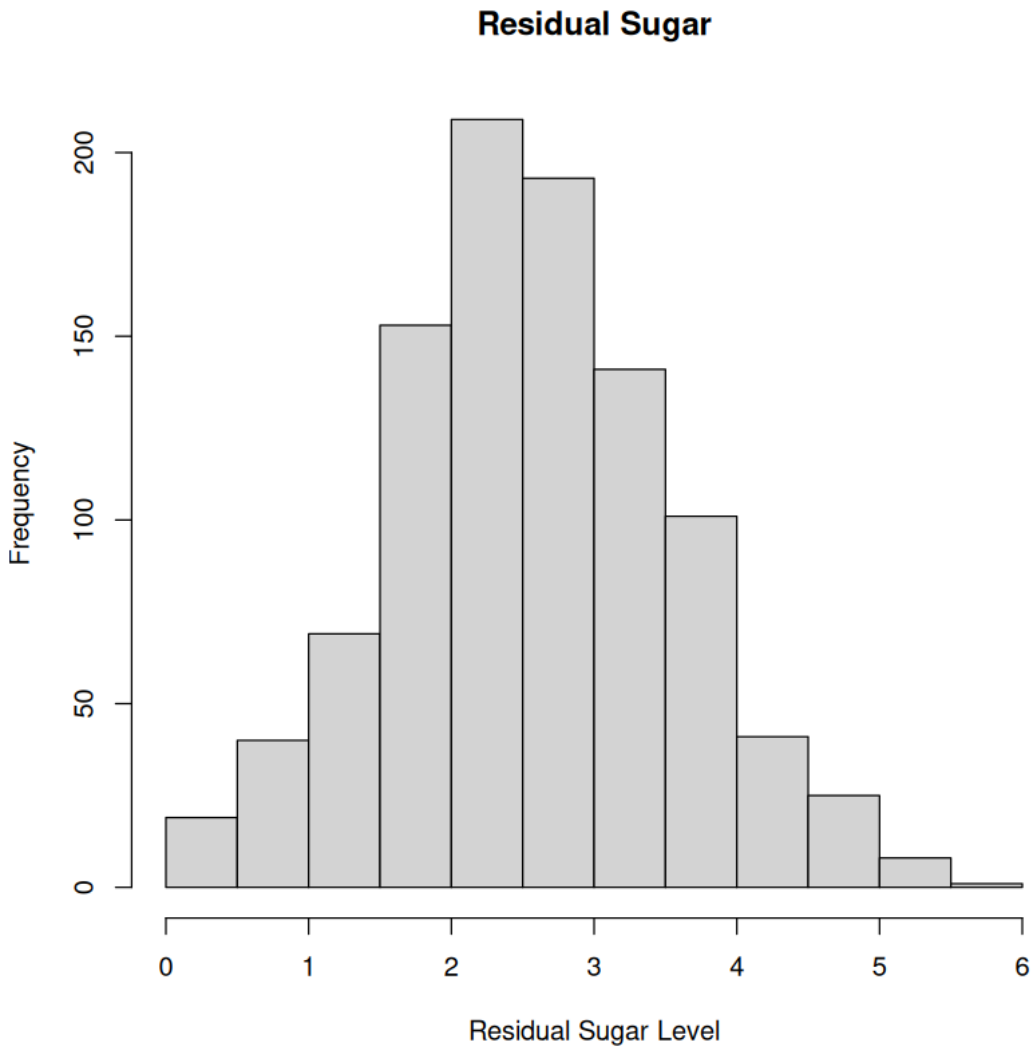
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 0.26 - 0.28, nilai minimum di rentang 0.10-0.12 dan nilai maksimum di rentang 0.40-0.42. Selain itu, plot menunjukkan bahwa grafik memiliki skewness negatif.

[12]:
```
box_plot(data["citric.acid"], "Citric Acid", "Level", "Acidity")
```



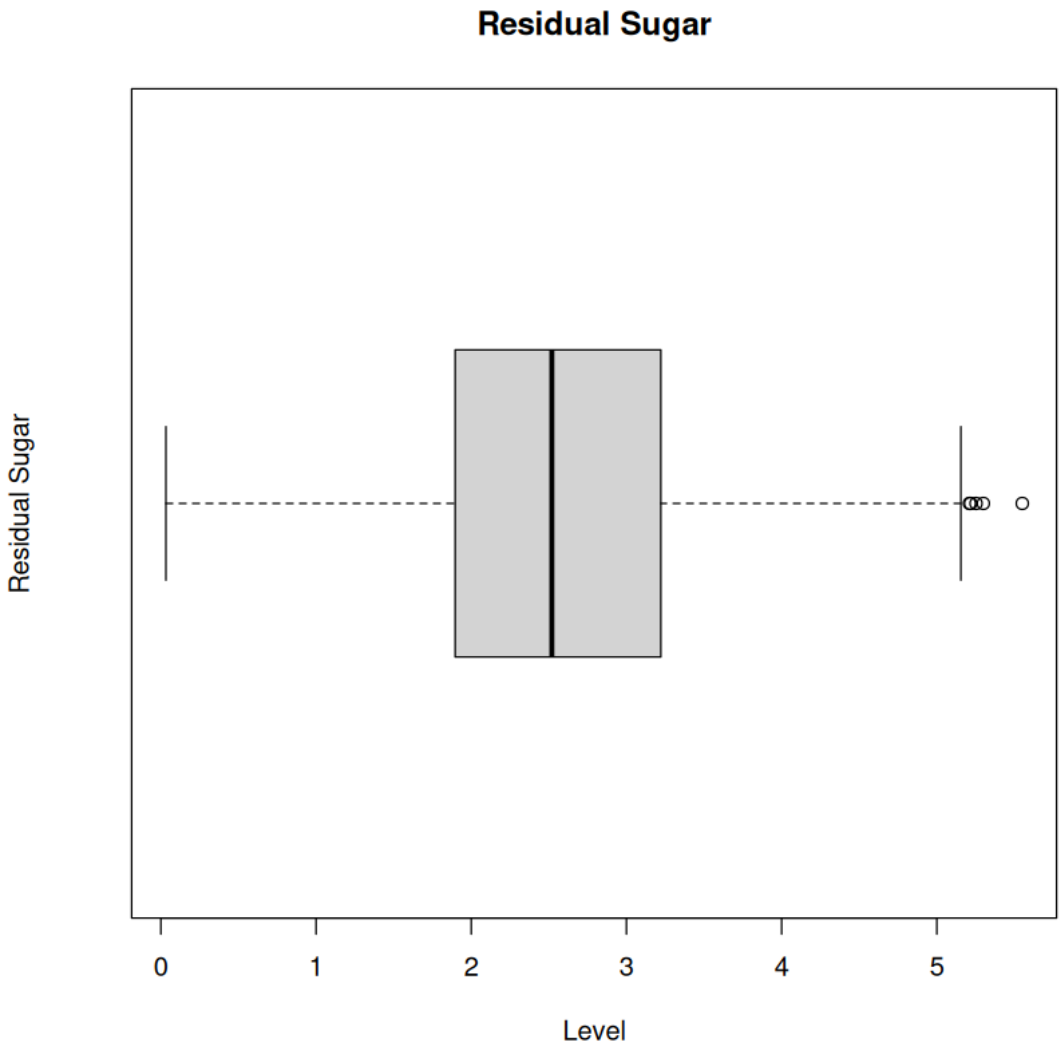
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 0.20-0.25, Q2 pada rentang 0.25-0.30, dan Q3 pada rentang 0.30-0.35. Selain itu, terdapat 4 nilai outlier pada bagian kiri dan 2 nilai outlier pada bagian kanan plot.

[13]:
```
# Residual Sugar

histogram_plot(data["residual.sugar"], "Residual Sugar", "Residual Sugar Level")
```



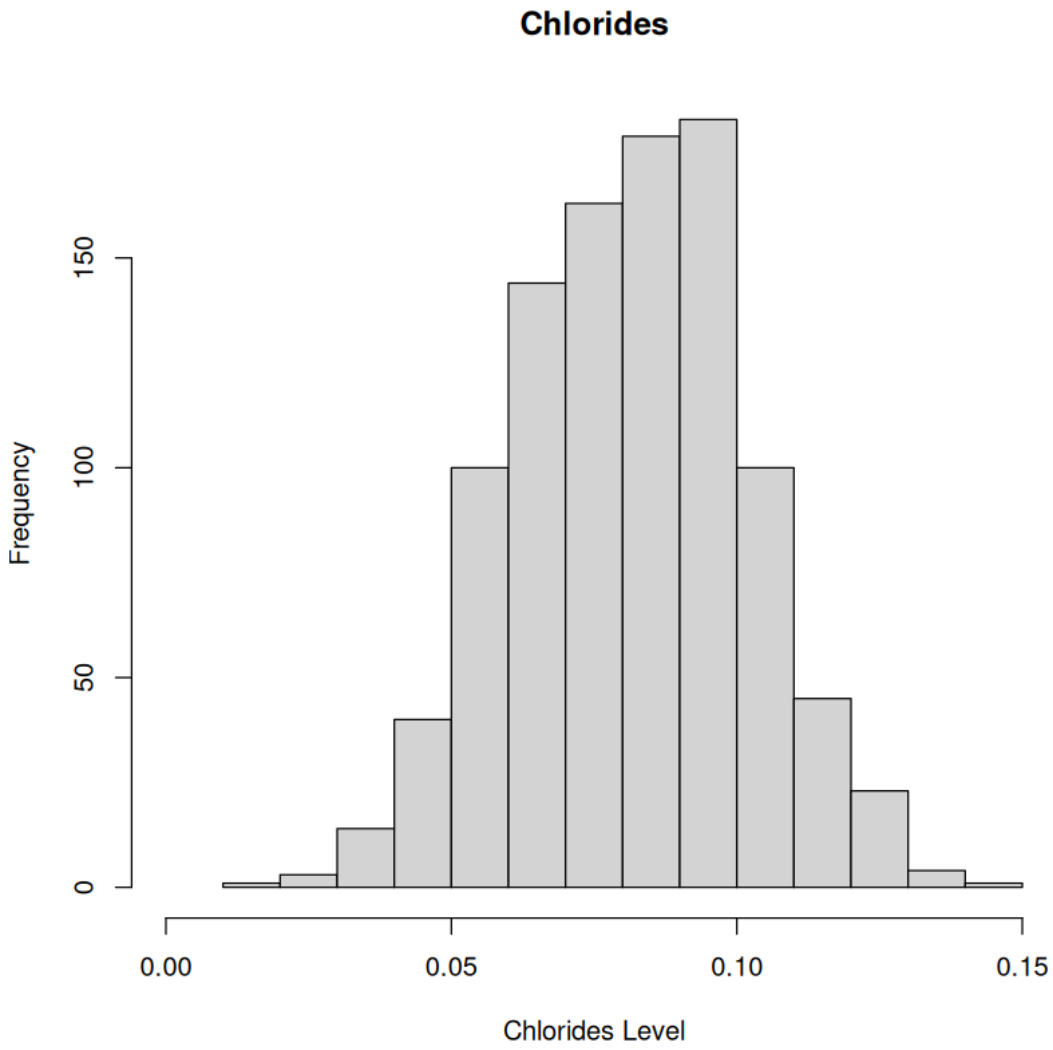
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 2.0-2.5, nilai minimum di rentang 0-0.5 dan nilai maksimum di rentang 5.5-6.0. Selain itu, plot menunjukkan bahwa grafik memiliki skewness positif.

[14]:
```
box_plot(data["residual.sugar"], "Residual Sugar", "Level", "Residual Sugar")
```



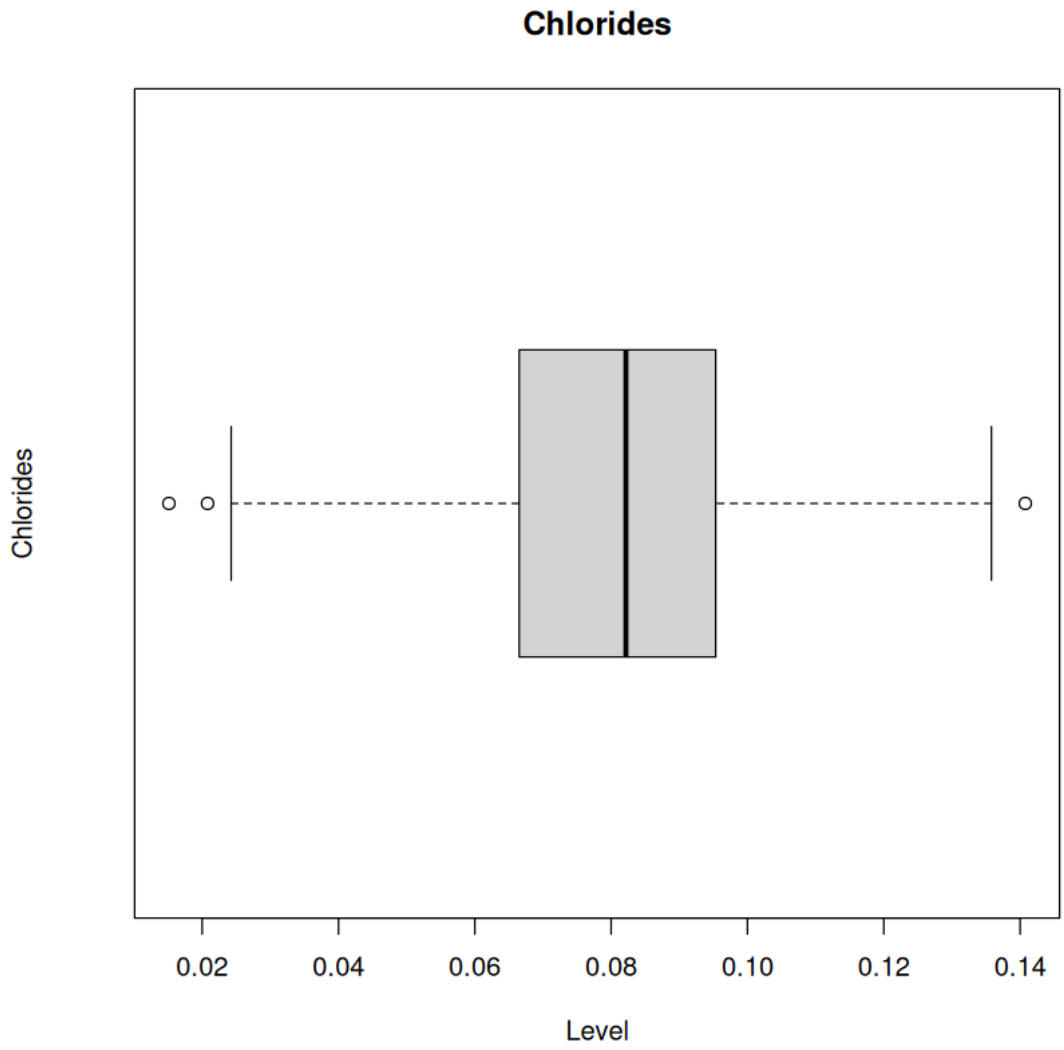
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 2-3, Q2 pada rentang 2-3, dan Q3 pada rentang 3-4. Selain itu, tidak terdapat nilai outlier pada bagian kiri dan 4 nilai outlier pada bagian kanan plot.

[15]:
```
# Chlorides

histogram_plot(data["chlorides"], "Chlorides", "Chlorides Level", c(0, 0.15))
```



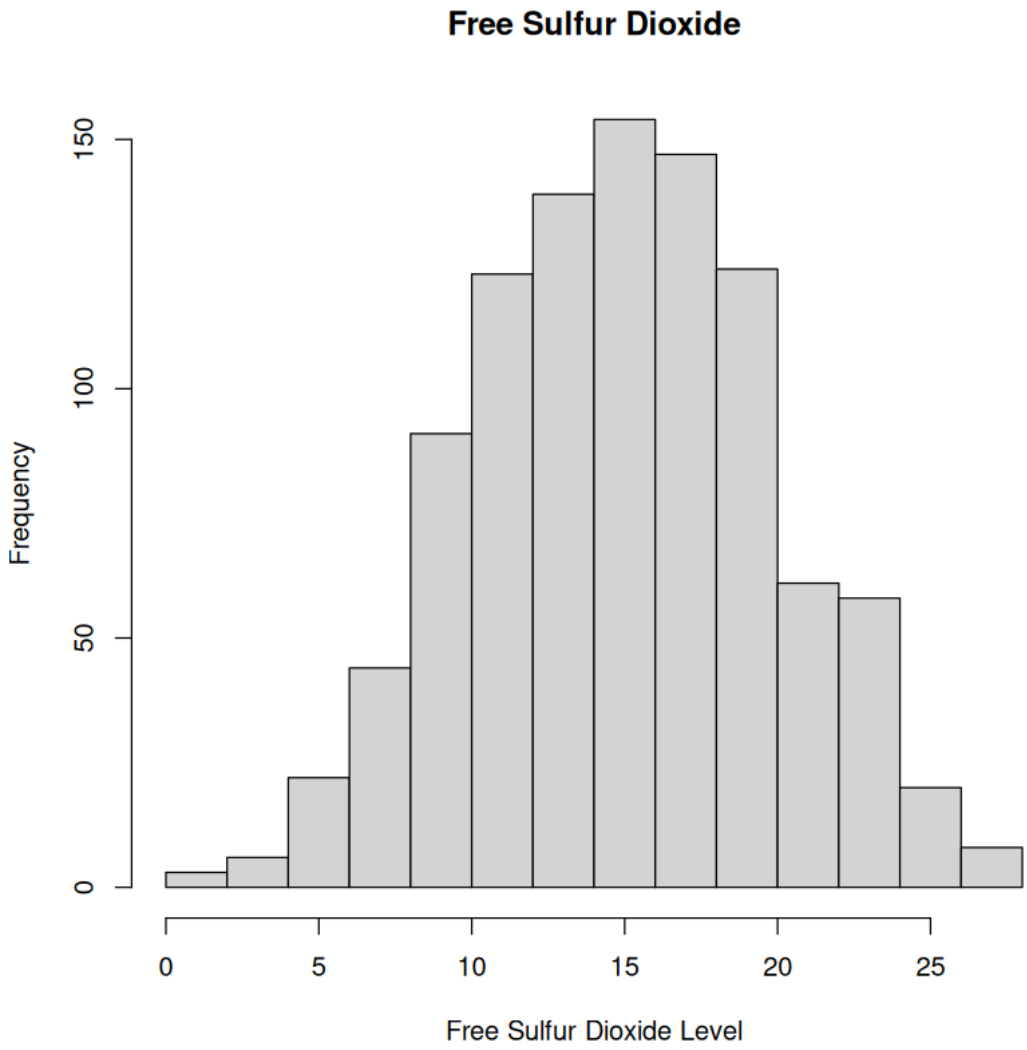
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 0.09 - 0.10, nilai minimum di rentang 0.01-0.02 dan nilai maksimum di rentang 0.14-0.15. Selain itu, plot menunjukkan bahwa grafik memiliki skewness negatif.

[16]:
```
box_plot(data["chlorides"], "Chlorides", "Level", "Chlorides")
```



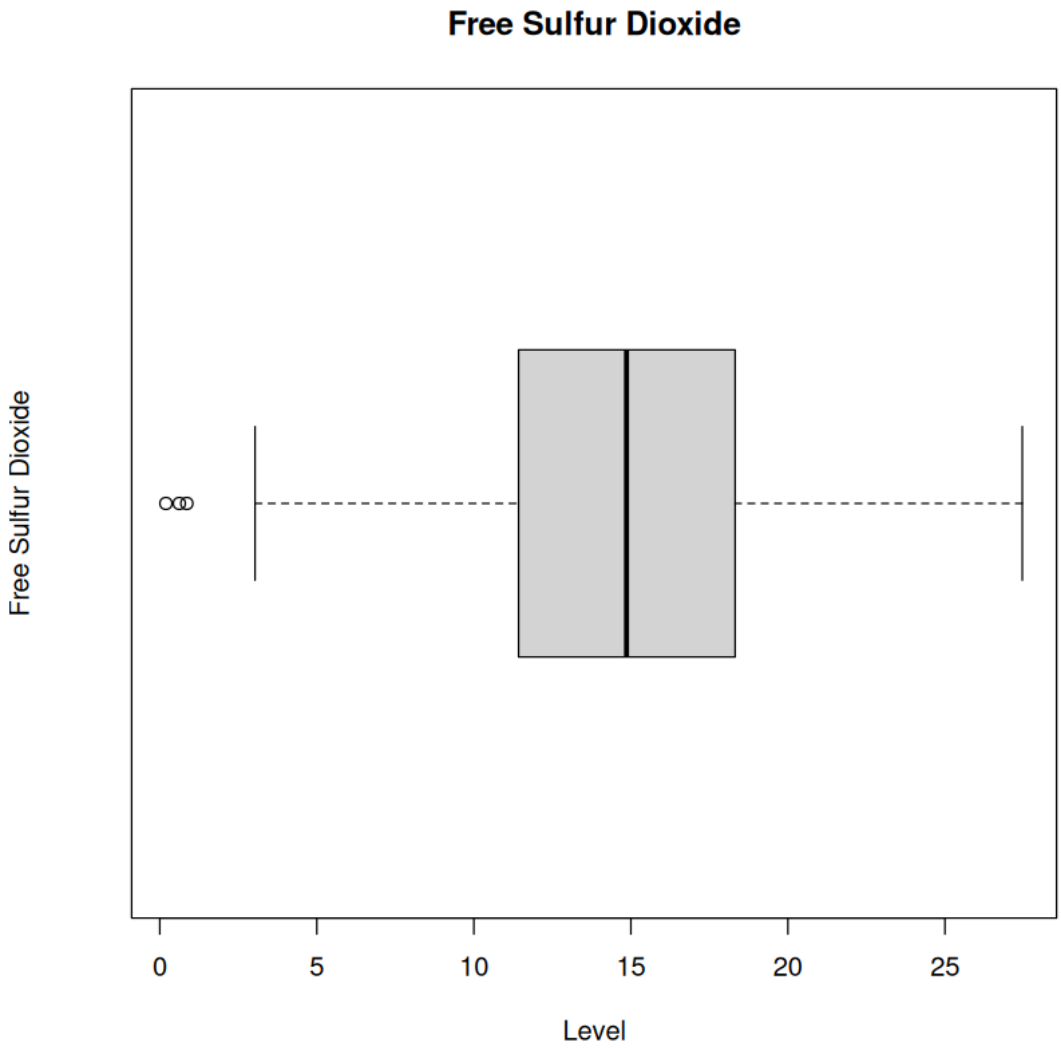
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 0.06-0.08, Q2 pada rentang 0.08-0.10, dan Q3 pada rentang 0.08-0.10. Selain itu, terdapat 2 nilai outlier pada bagian kiri dan 1 nilai outlier pada bagian kanan plot.

[17]:
```
# Free Sulfur Dioxide

histogram_plot(data["free.sulfur.dioxide"], "Free Sulfur Dioxide", "Free Sulfur␣
↪Dioxide Level")
```



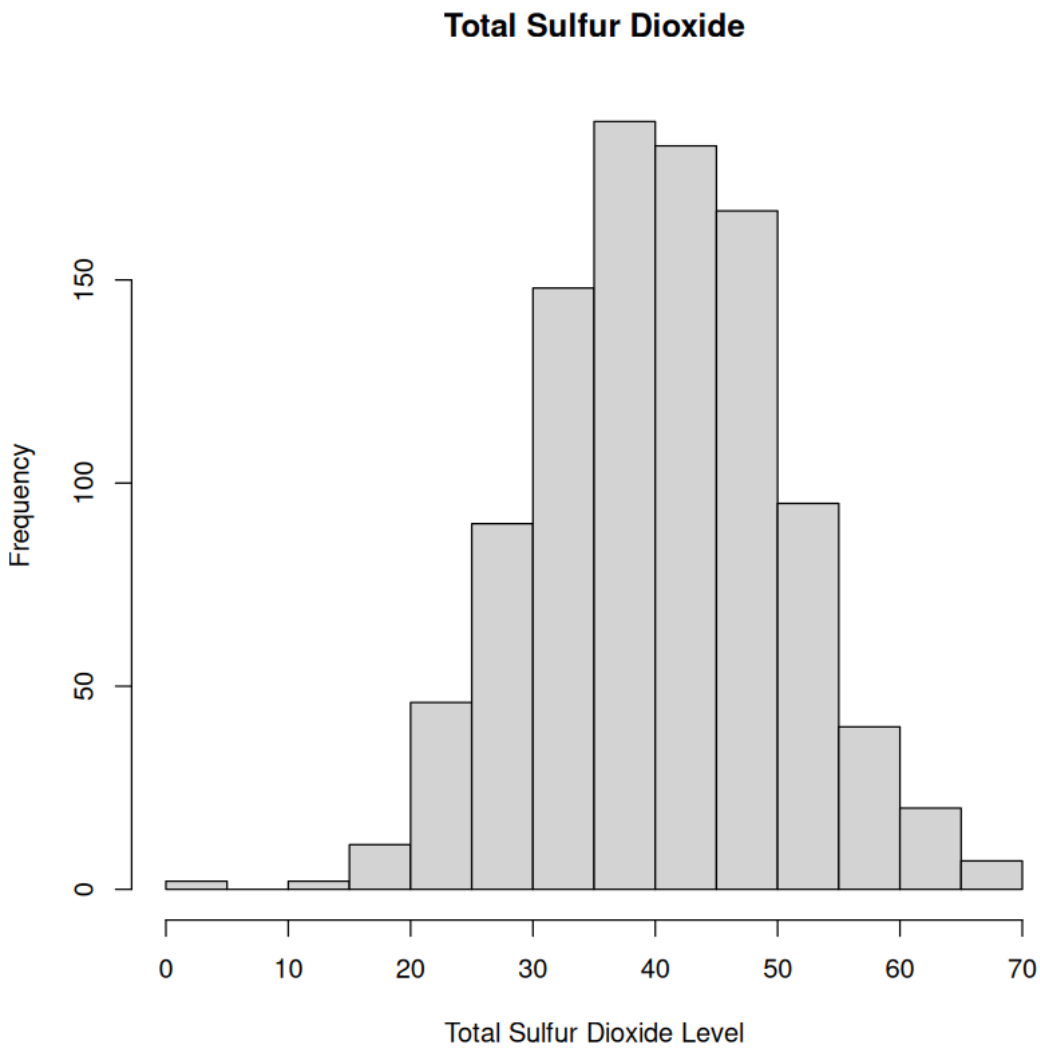
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 14-16, nilai minimum di rentang 0-2 dan nilai maksimum di rentang 0.40-0.42. Selain itu, plot menunjukkan bahwa grafik memiliki skewness hampir mendekati nol.

[18]:
```
box_plot(data["free.sulfur.dioxide"], "Free Sulfur Dioxide", "Level", "Free␣
↪Sulfur Dioxide")
```



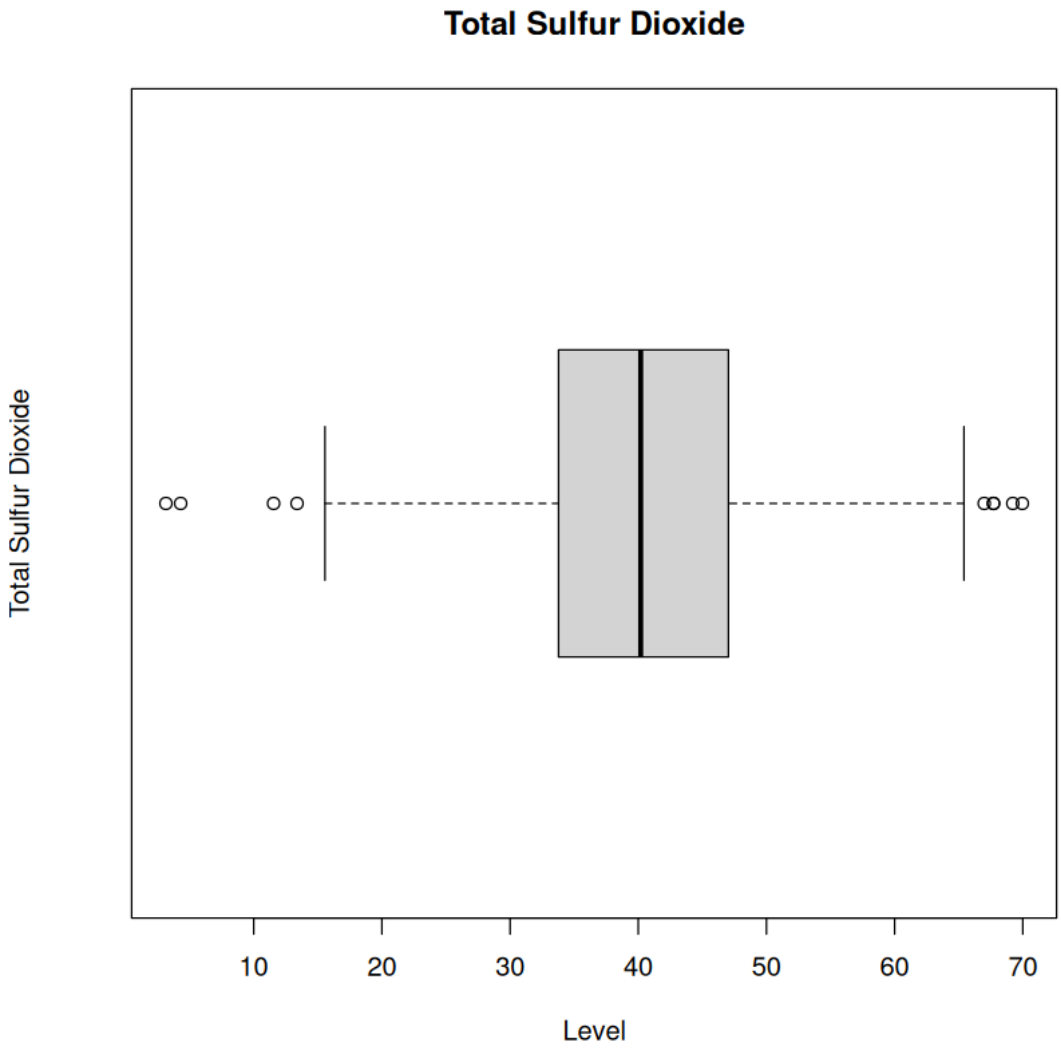
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 10-15, Q2 pada rentang mendekati 15, dan Q3 pada rentang 15-20. Selain itu, terdapat 3 nilai outlier pada bagian kiri dan tidak terdapat nilai outlier pada bagian kanan plot.

[19]:
```
# Total Sulfur Dioxide

histogram_plot(data["total.sulfur.dioxide"], "Total Sulfur Dioxide", "Total␣
↪Sulfur Dioxide Level", c(0, 70))
```



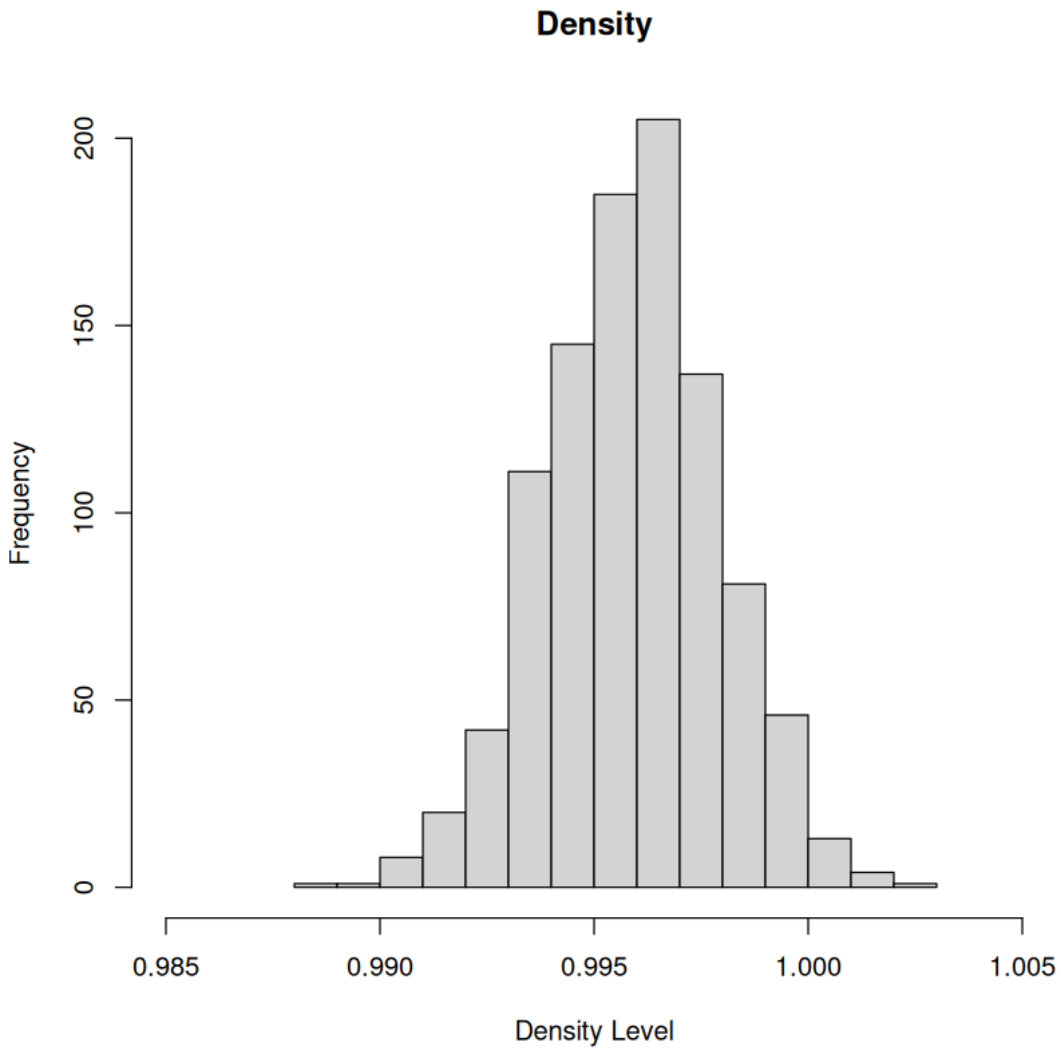
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 35-40, nilai minimum di rentang 0-5 dan nilai maksimum di rentang 75-70. Selain itu, plot menunjukkan bahwa grafik memiliki skewness negatif.

[20]:
```
box_plot(data["total.sulfur.dioxide"], "Total Sulfur Dioxide", "Level", "Total␣
↪Sulfur Dioxide")
```



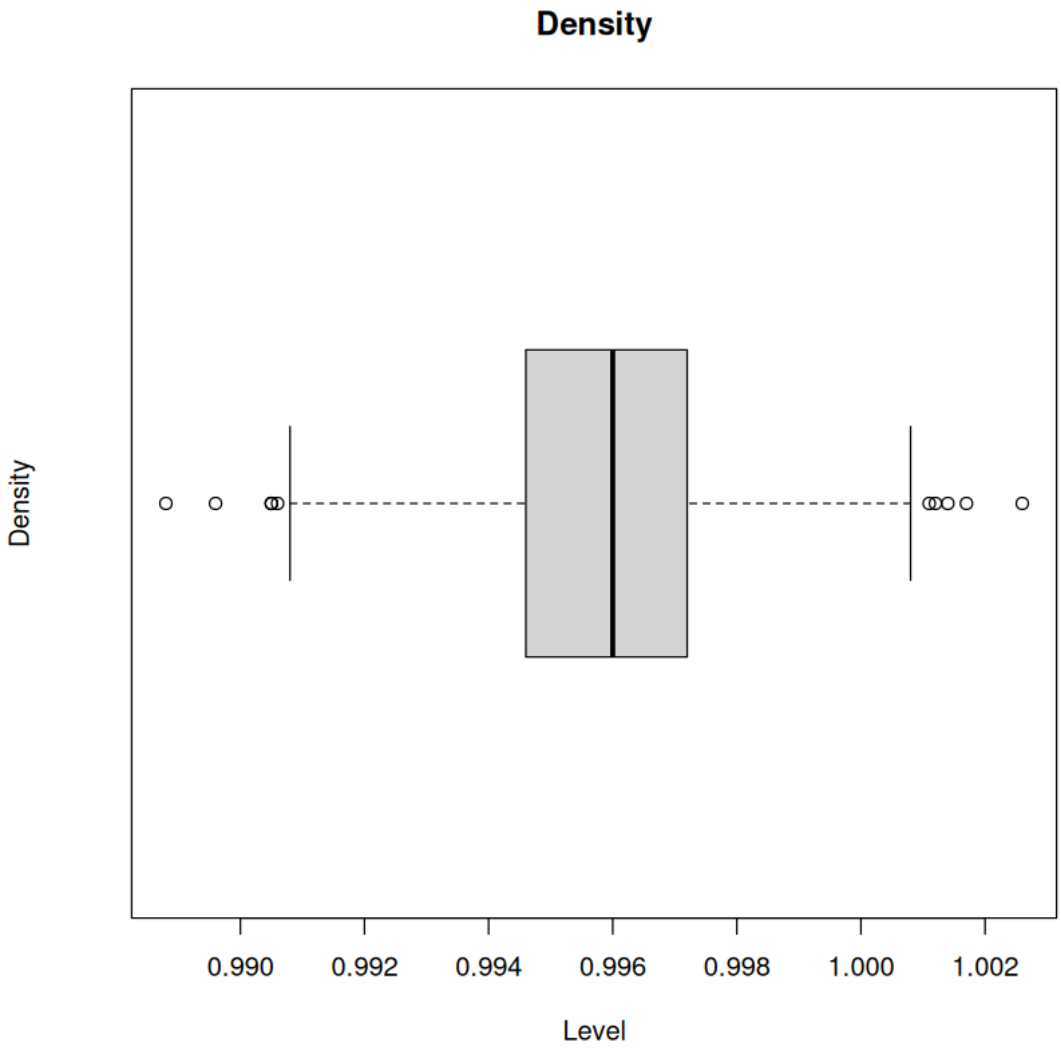
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 30-40, Q2 pada rentang 40-50, dan Q3 pada rentang 40-50. Selain itu, terdapat 4 nilai outlier pada bagian kiri dan 4 nilai outlier pada bagian kanan plot.

[21]:
```
# Density

histogram_plot(data["density"], "Density", "Density Level", c(0.985, 1.005))
```



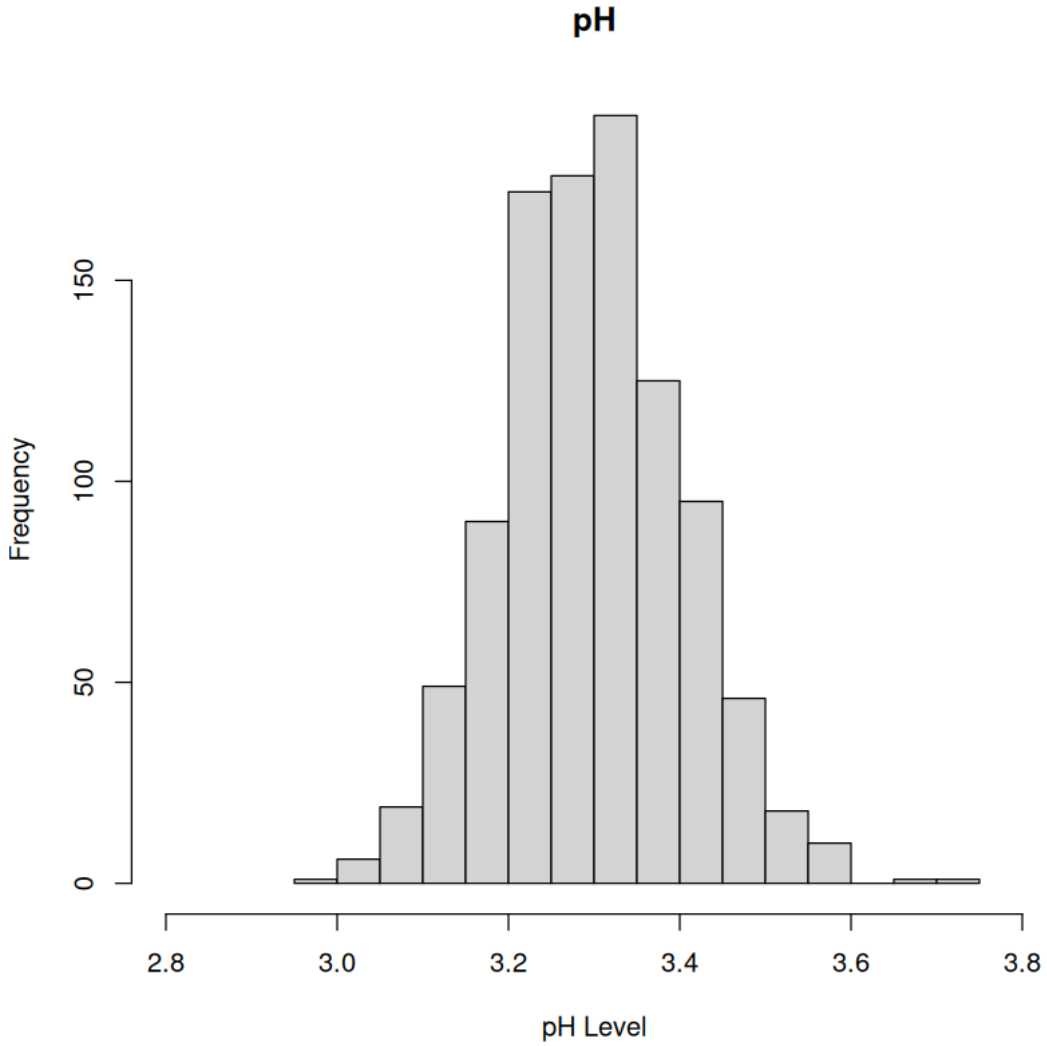
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 0.996-0.997, nilai minimum di rentang 0.988 - 0.989 dan nilai maksimum di rentang 1.001-1.002. Selain itu, plot menunjukkan bahwa grafik memiliki skewness negatif.

[22]:
```
box_plot(data["density"], "Density", "Level", "Density")
```



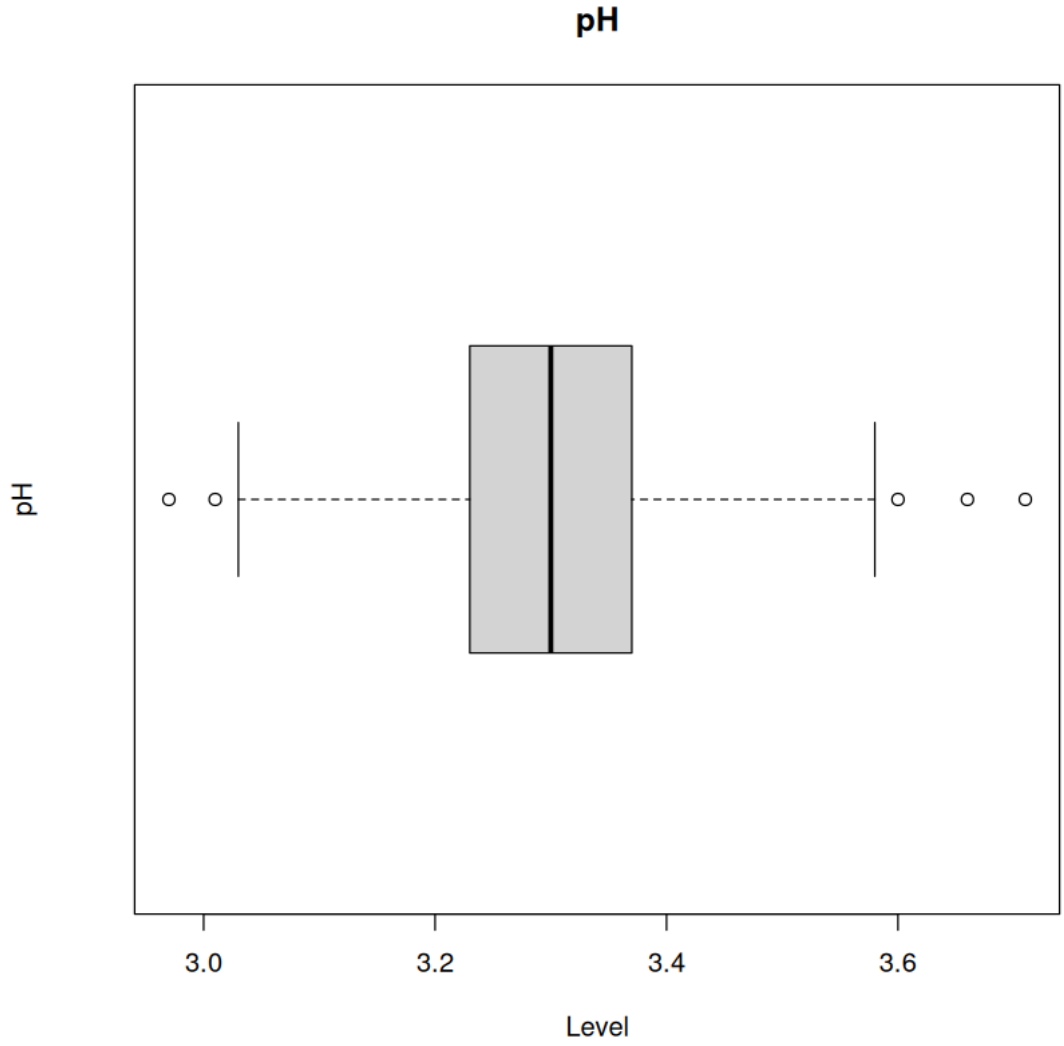
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 0.994-0.996, Q2 pada rentang mendekati 0.996, dan Q3 pada rentang 0.996-0.998. Selain itu, terdapat 4 nilai outlier pada bagian kiri dan 5 nilai outlier pada bagian kanan plot.

[23]:
```
# pH

histogram_plot(data["pH"], "pH", "pH Level", c(2.8, 3.8))
```



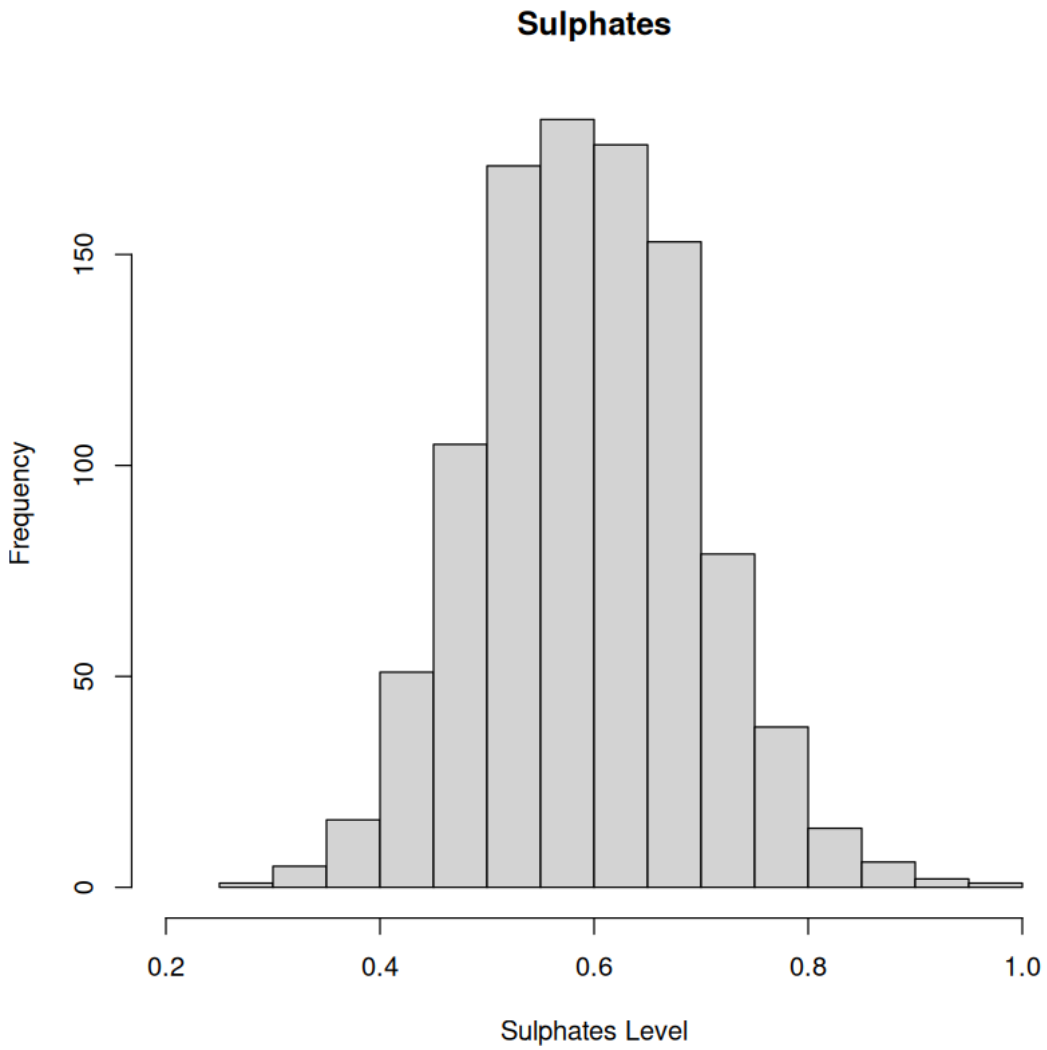
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 3.30-3.35, nilai minimum di rentang 2.95-3.00 dan nilai maksimum di rentang 3.70-3.75. Selain itu, plot menunjukkan bahwa grafik memiliki skewness positif.

[24]:
```
box_plot(data["pH"], "pH", "Level", "pH")
```



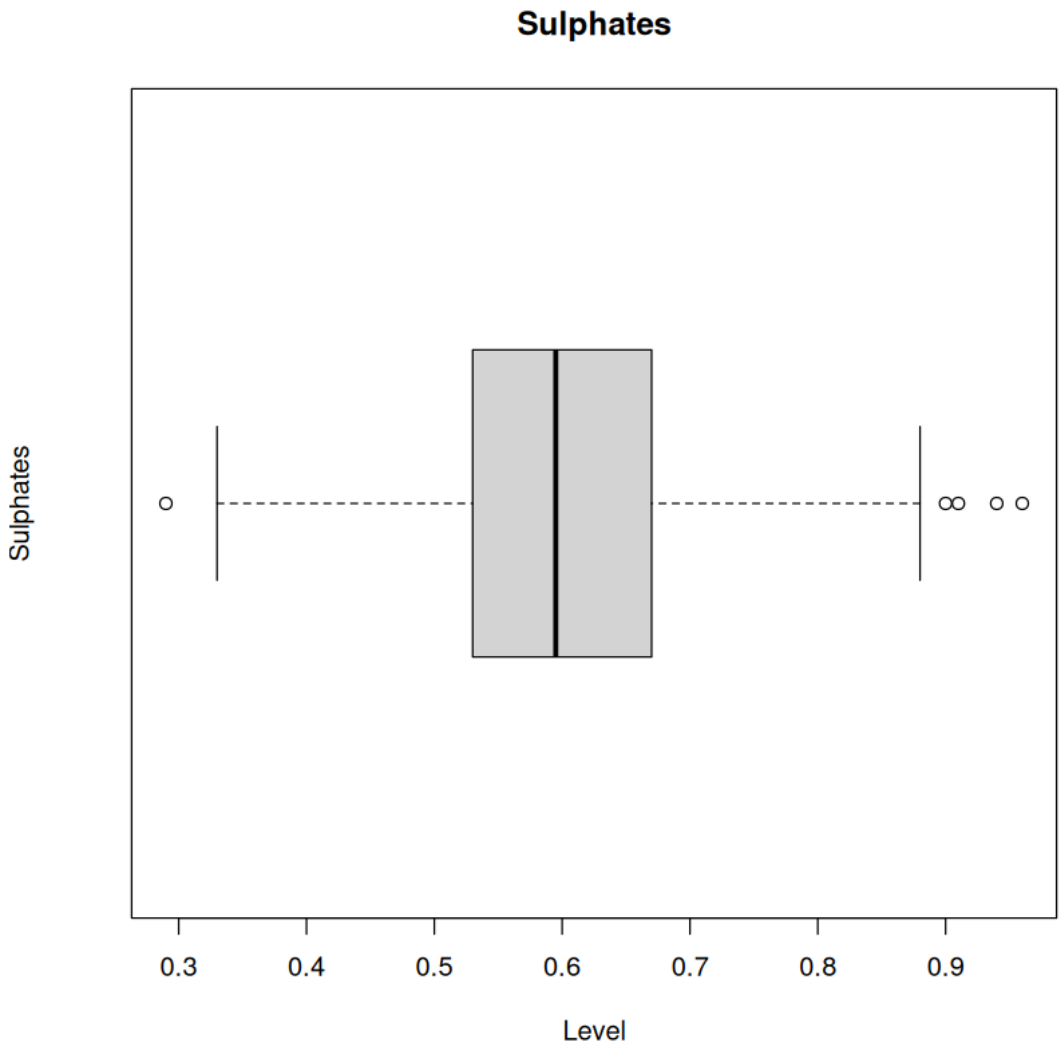
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 3.2-3.4, Q2 pada rentang 3.2-3.4, dan Q3 pada rentang 3.2-3.4. Selain itu, terdapat 2 nilai outlier pada bagian kiri dan 3 nilai outlier pada bagian kanan plot.

[25]:
```
# Sulphates

histogram_plot(data["sulphates"], "Sulphates", "Sulphates Level", c(0.2, 1))
```



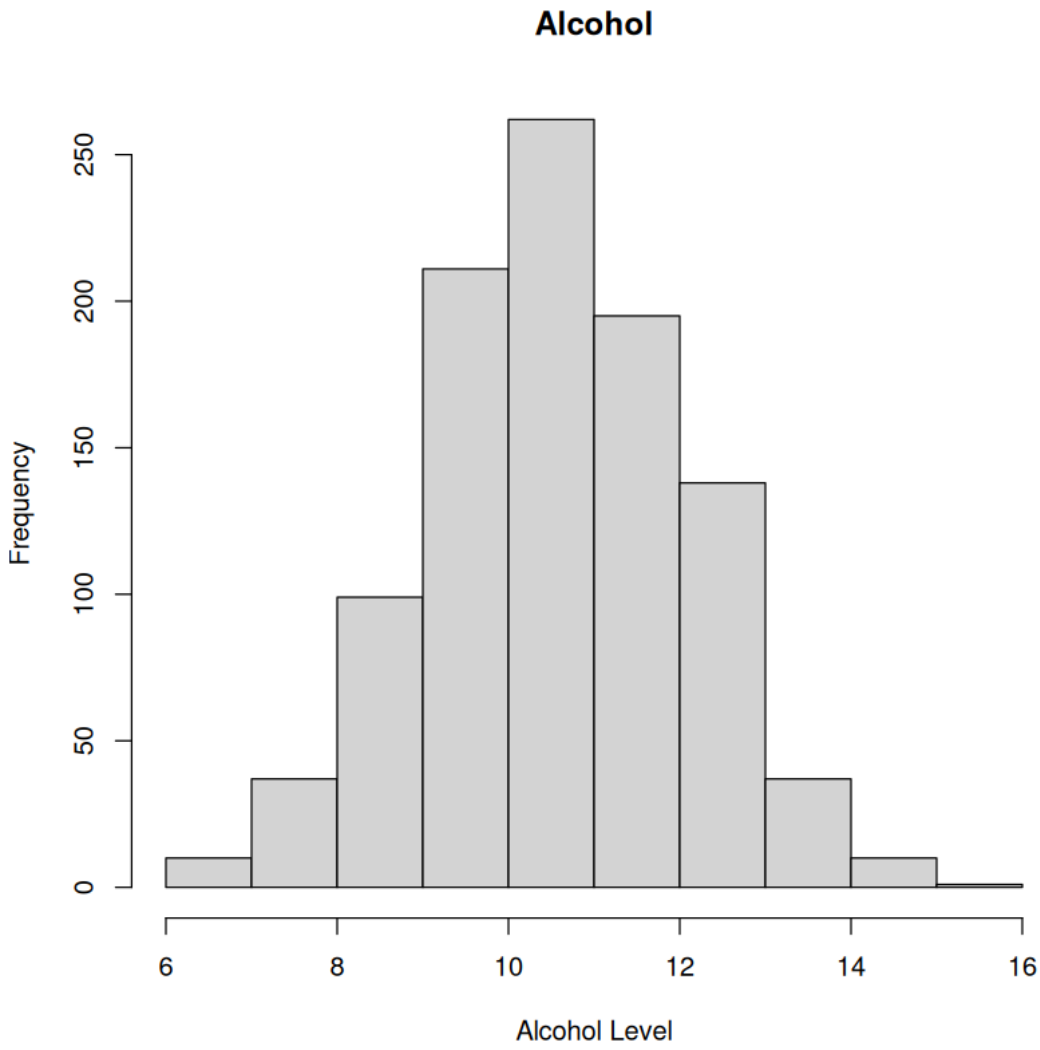
Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 0.55-0.6, nilai minimum di rentang 0.25-0.30 dan nilai maksimum di rentang 0.95-1.00. Selain itu, plot menunjukkan bahwa grafik memiliki skewness positif.

[26]:
```
box_plot(data["sulphates"], "Sulphates", "Level", "Sulphates")
```



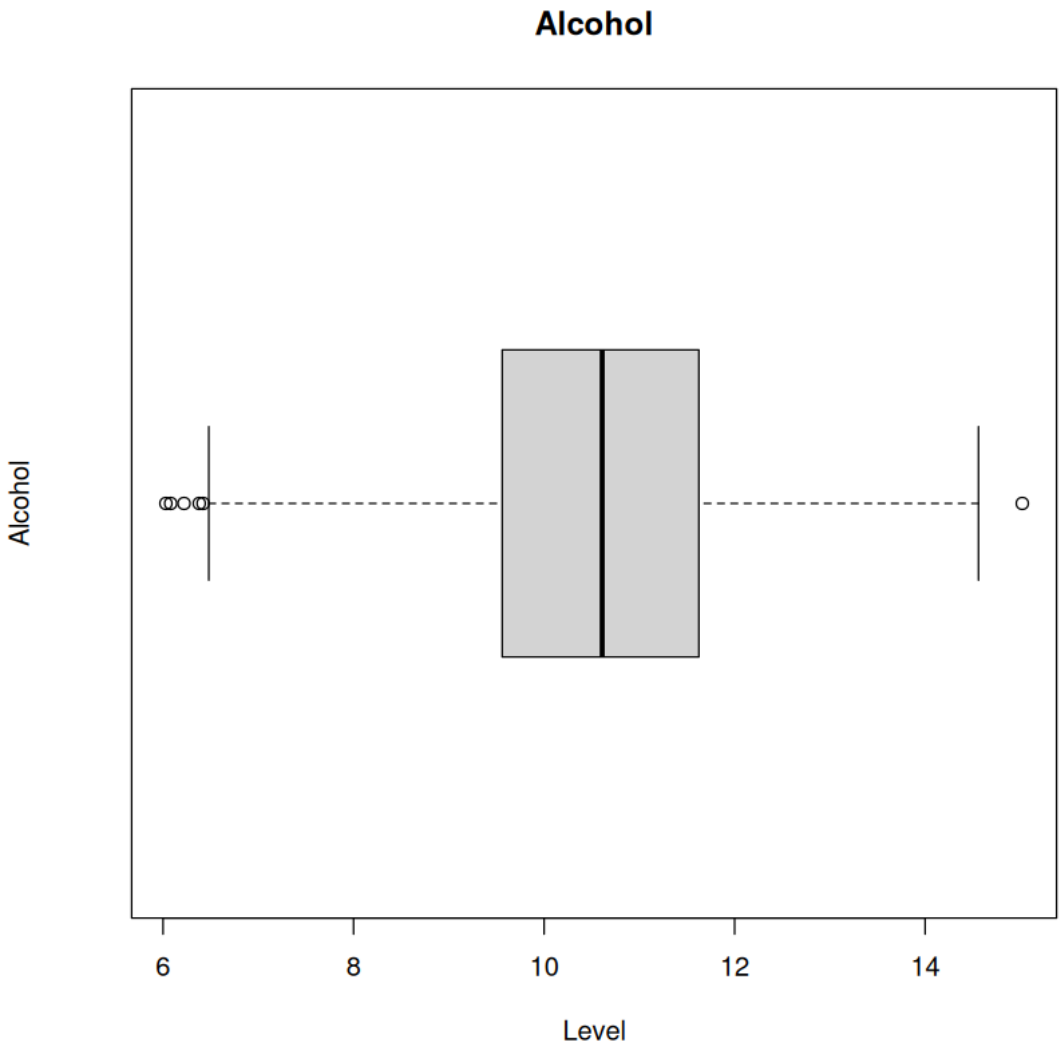
Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 0.5-0.6, Q2 pada rentang mendekati 0.6, dan Q3 pada rentang 0.6-0.7. Selain itu, terdapat 1 nilai outlier pada bagian kiri dan 4 nilai outlier pada bagian kanan plot.

[27]:
```
# Alcohol

histogram_plot(data["alcohol"], "Alcohol", "Alcohol Level")
```



Dari plot histogram di atas, dapat dilihat bahwa nilai modus berada pada rentang 10-11, nilai minimum di rentang 6-7 dan nilai maksimum di rentang 15-16. Selain itu, plot menunjukkan bahwa grafik memiliki skewness negatif.

[28]:
```
box_plot(data["alcohol"], "Alcohol", "Level", "Alcohol")
```



Dari boxplot di atas, dapat dilihat bahwa nilai Q1 berada pada rentang 8-10, Q2 pada rentang 10-12, dan Q3 pada rentang 10-12. Selain itu, terdapat 6 nilai outlier pada bagian kiri dan 1 nilai outlier pada bagian kanan plot.

1.4 Soal 3

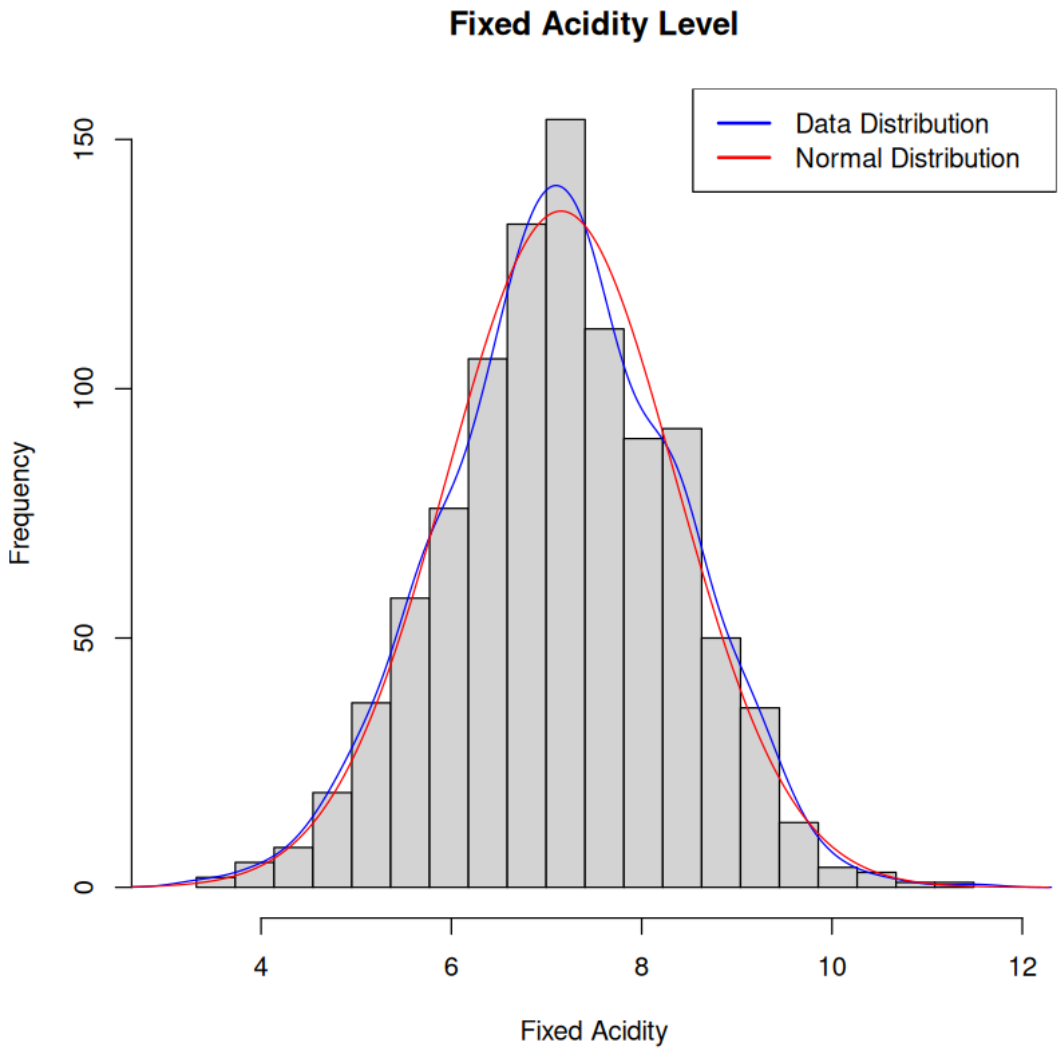
Menentukan setiap kolom numerik berdistribusi normal atau tidak. Gunakan normality test yang dikaitkan dengan histogram plot

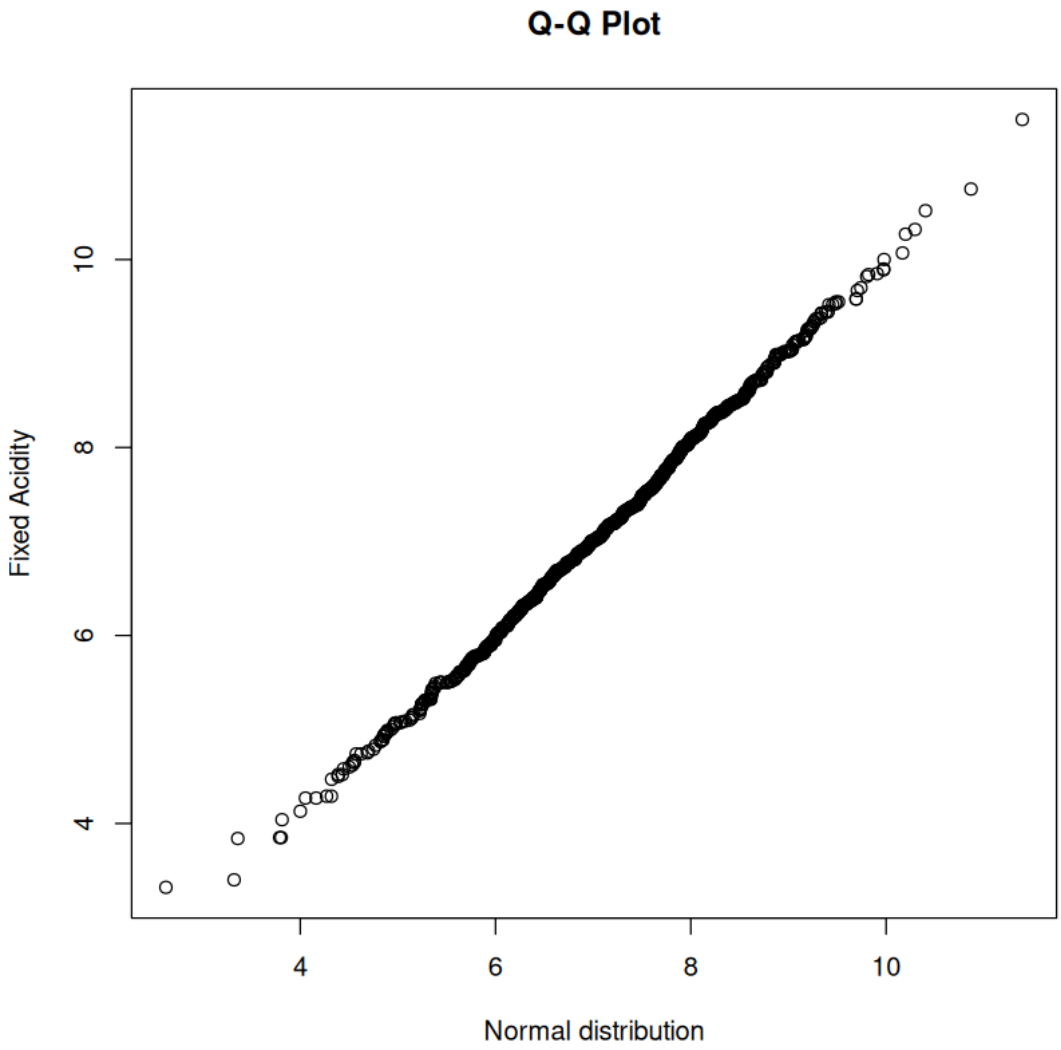
[29]:
```
plot <- function(column, main_title, x_title, x_limit = NULL) {
    column <- sort(as.numeric(unlist(column)))

    if (is.null(x_limit)) {
```

```
        min_value <- floor(column[1])
        max_value <- ceiling(column[length(column)])
        x_limit <- c(min_value, max_value)
    }

    mean_value <- mean(column)
    standard_deviation_value <- sd(column)
    n_value <- length(column)

    binwidth <- (column[length(column)] - column[1]) / 20
    breaks <- seq(
        min(column),
        max(column),
        binwidth
    )

    hist (
        x = column,
        breaks = breaks,
        main = main_title,
        xlab = x_title,
        xlim = x_limit,
    )

    x_line <- unlist(density(column)["x"])
    y_line <- unlist(density(column)["y"]) * n_value * binwidth

    lines(
        x = x_line,
        y = y_line,
        col = "blue"
    )

    line <- seq(
        qnorm(0.00001, mean_value, standard_deviation_value),
        qnorm(0.99999, mean_value, standard_deviation_value),
        length.out = n_value
    )

    lines(
        line,
        n_value * dnorm(line, mean_value, standard_deviation_value) * binwidth,
        col = "red"
    )

    legend(
        x = "topright",
```

```
        legend = c("Data Distribution", "Normal Distribution"),
        col = c("blue", "red"),
        lwd = 2
    )

    qqplot(
        x = rnorm(n_value, mean_value, standard_deviation_value),
        y = column,
        xlab = "Normal distribution",
        ylab = x_title,
        main = "Q-Q Plot"
    )
}
```

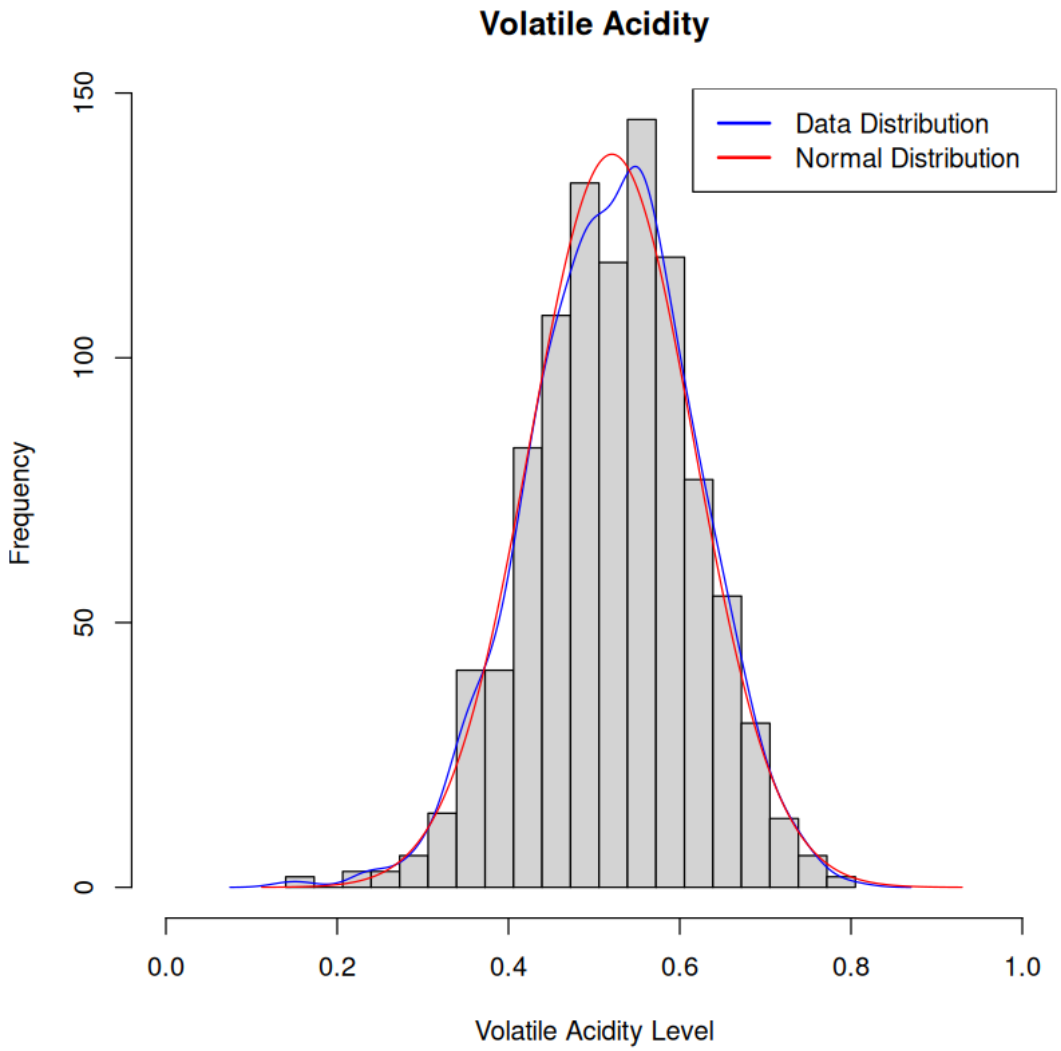
[30]:
```
# Fixed Acidity

plot(data["fixed.acidity"], "Fixed Acidity Level", "Fixed Acidity")
```

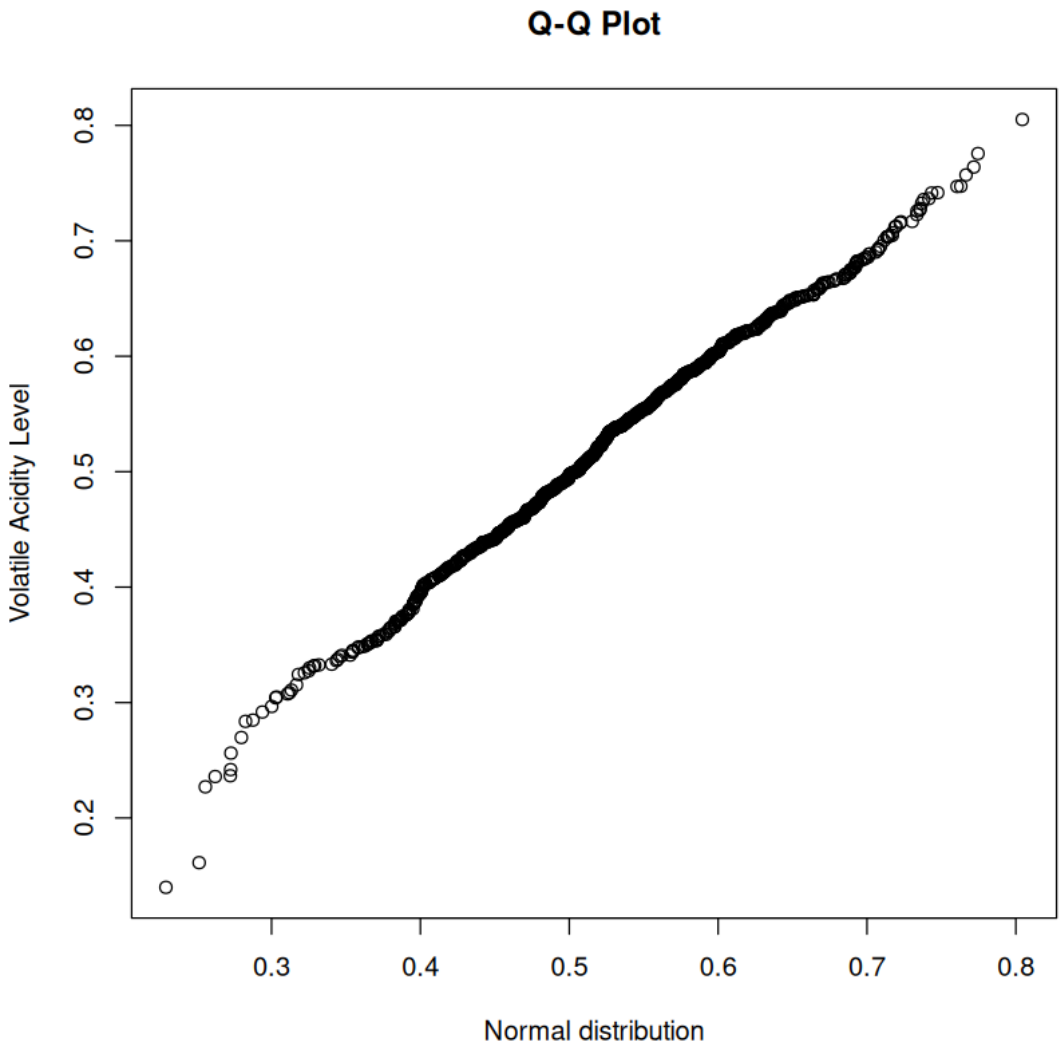

Fixed Acidity Level
Data Distribution
Normal Distribution
Frequency
0
50
100
150
4
6
8
10
12
Fixed Acidity



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Fixed Acidity** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

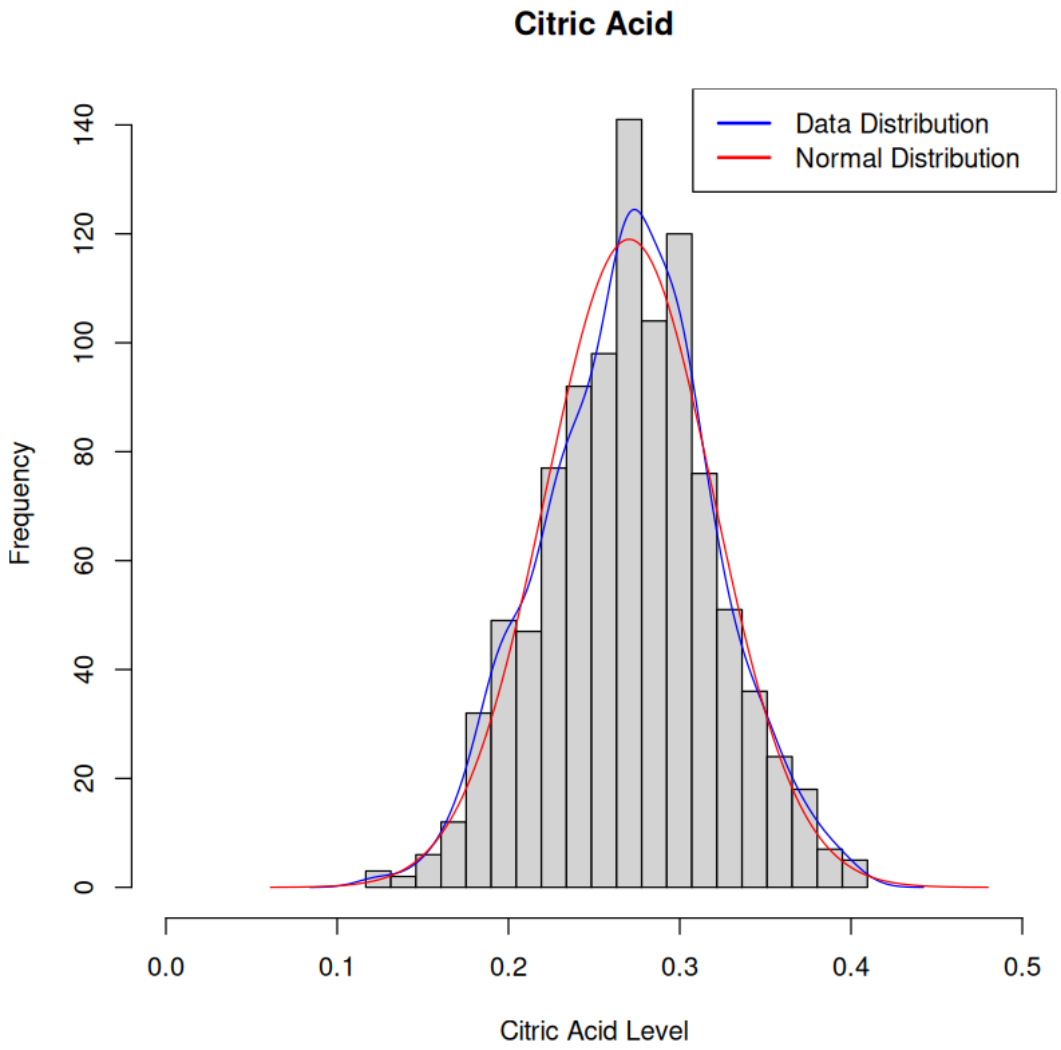
[31]:
```
# Volatile Acidity

plot(data["volatile.acidity"], "Volatile Acidity", "Volatile Acidity Level")
```

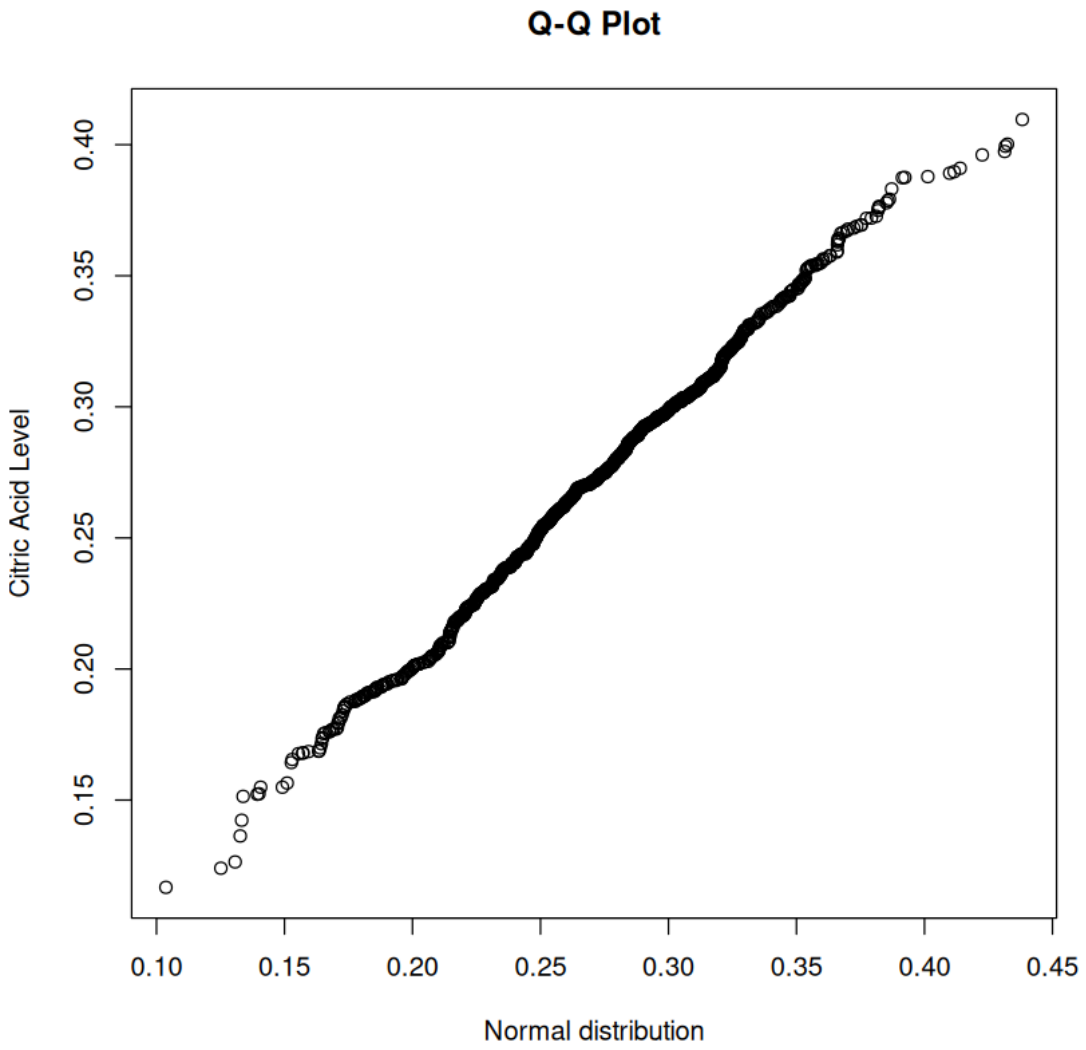

Volatile Acidity
Data Distribution
Normal Distribution
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
Volatile Acidity Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Volatile Acidity** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

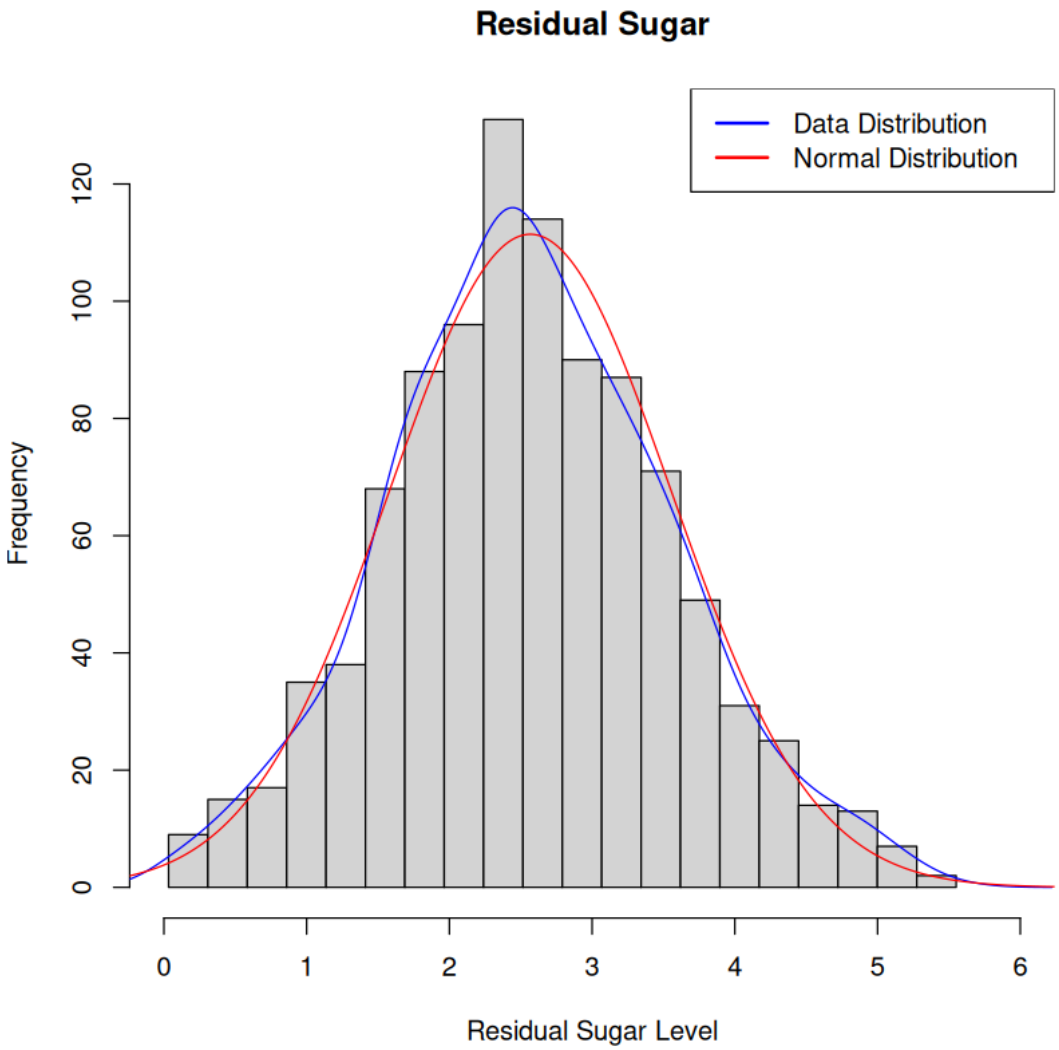
[32]:
```
# Citric Acid

plot(data["citric.acid"], "Citric Acid", "Citric Acid Level", c(0, 0.5))
```

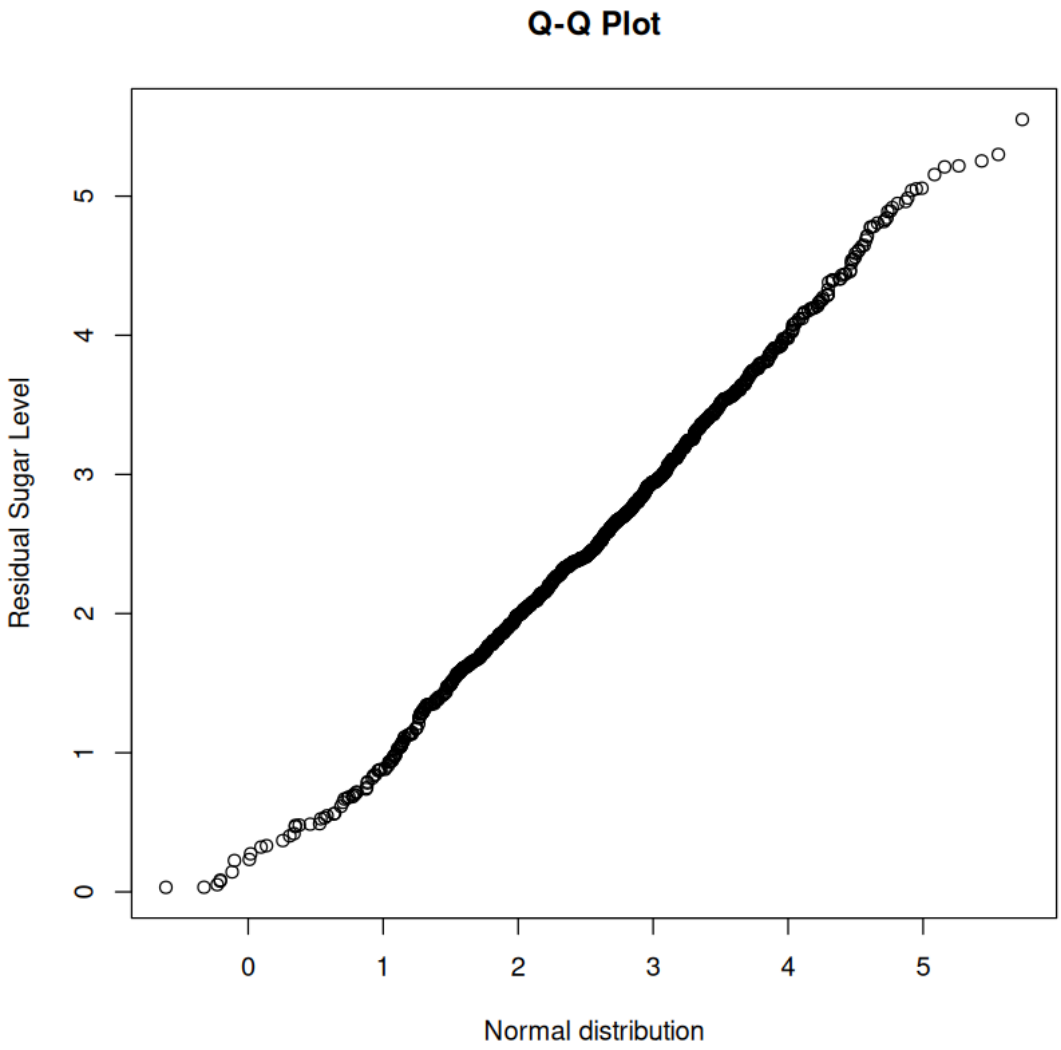

Citric Acid
Data Distribution
Normal Distribution
Frequency
0
20
40
60
80
100
120
140
0.0
0.1
0.2
0.3
0.4
0.5
Citric Acid Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Citric Acid** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

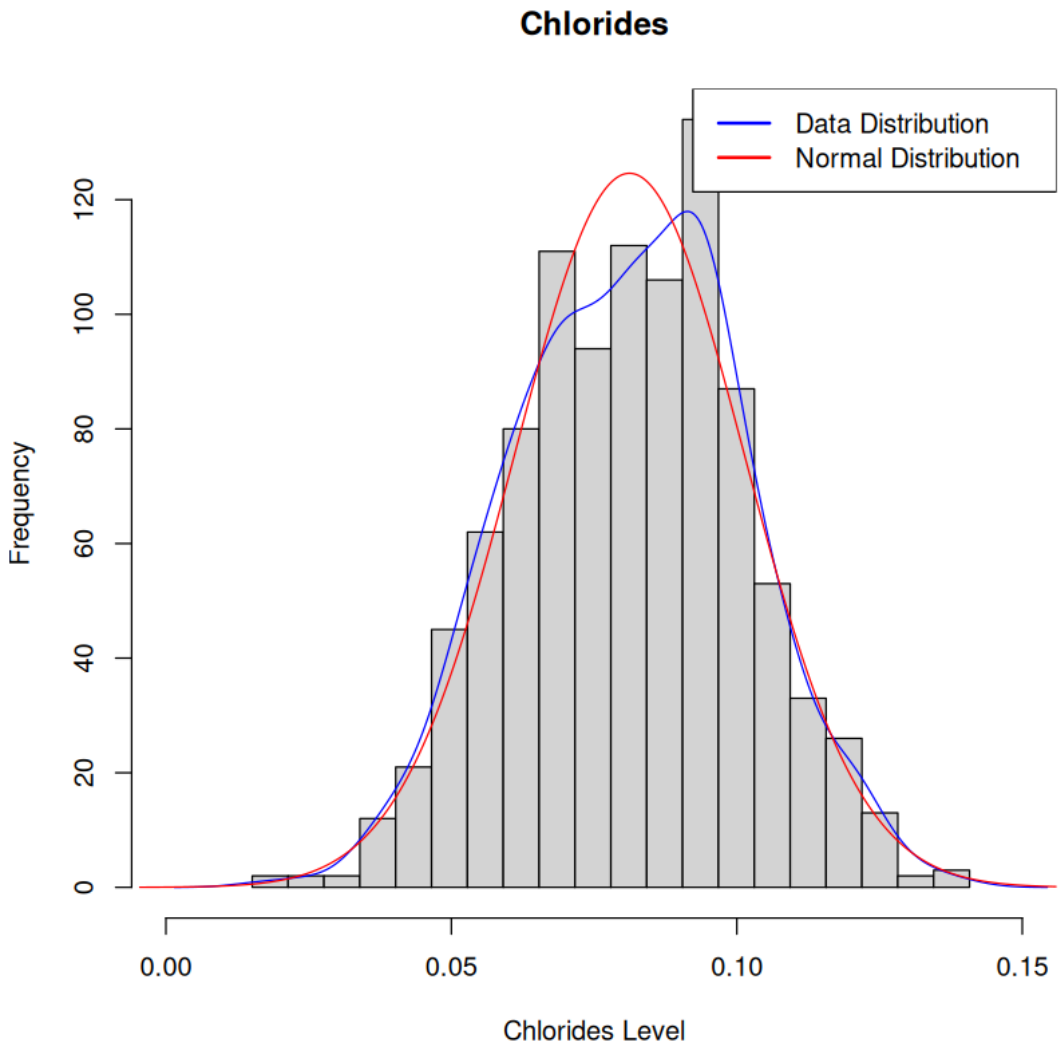
[33]:
```
# Residual Sugar

plot(data["residual.sugar"], "Residual Sugar", "Residual Sugar Level")
```

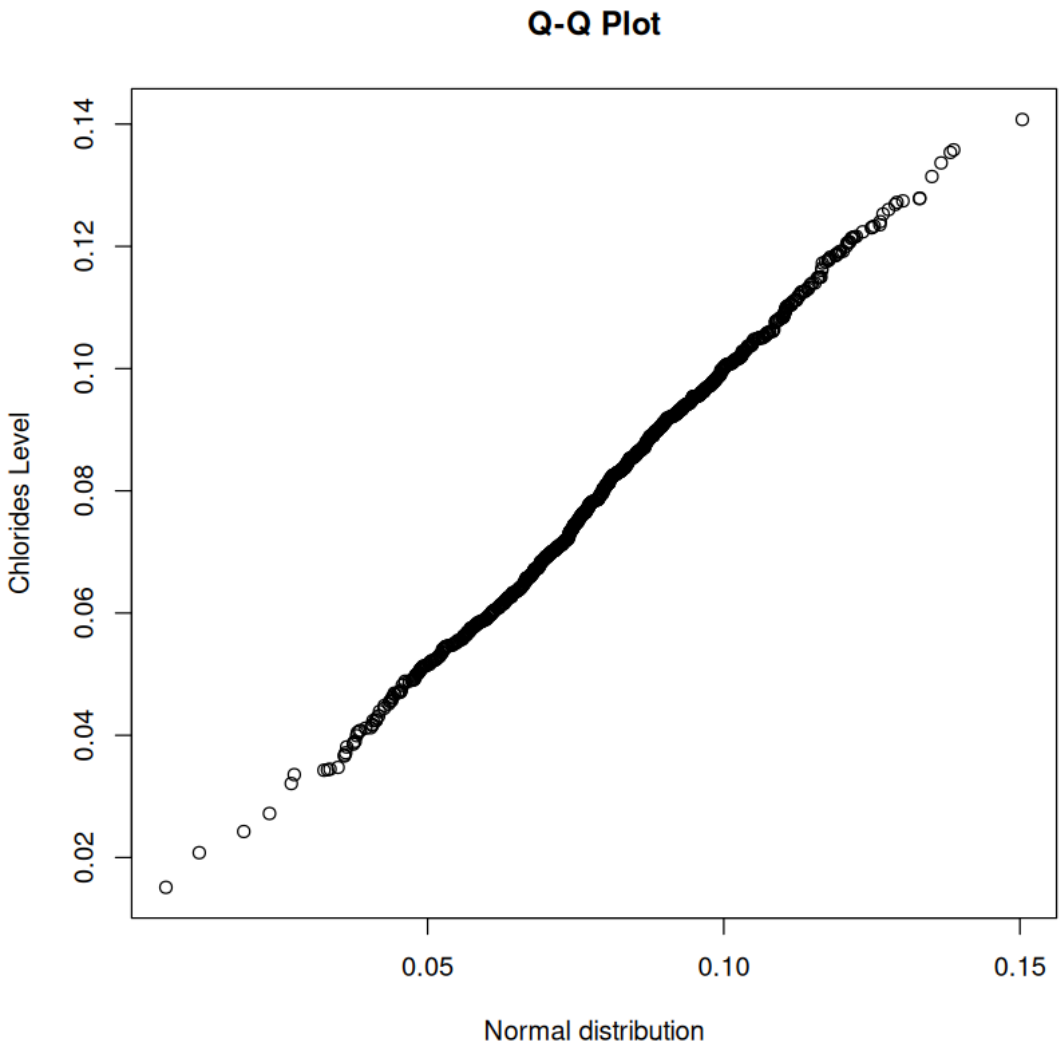

Residual Sugar
Data Distribution
Normal Distribution
Frequency
0
20
40
60
80
100
120
0
1
2
3
4
5
6
Residual Sugar Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Residual Sugar** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

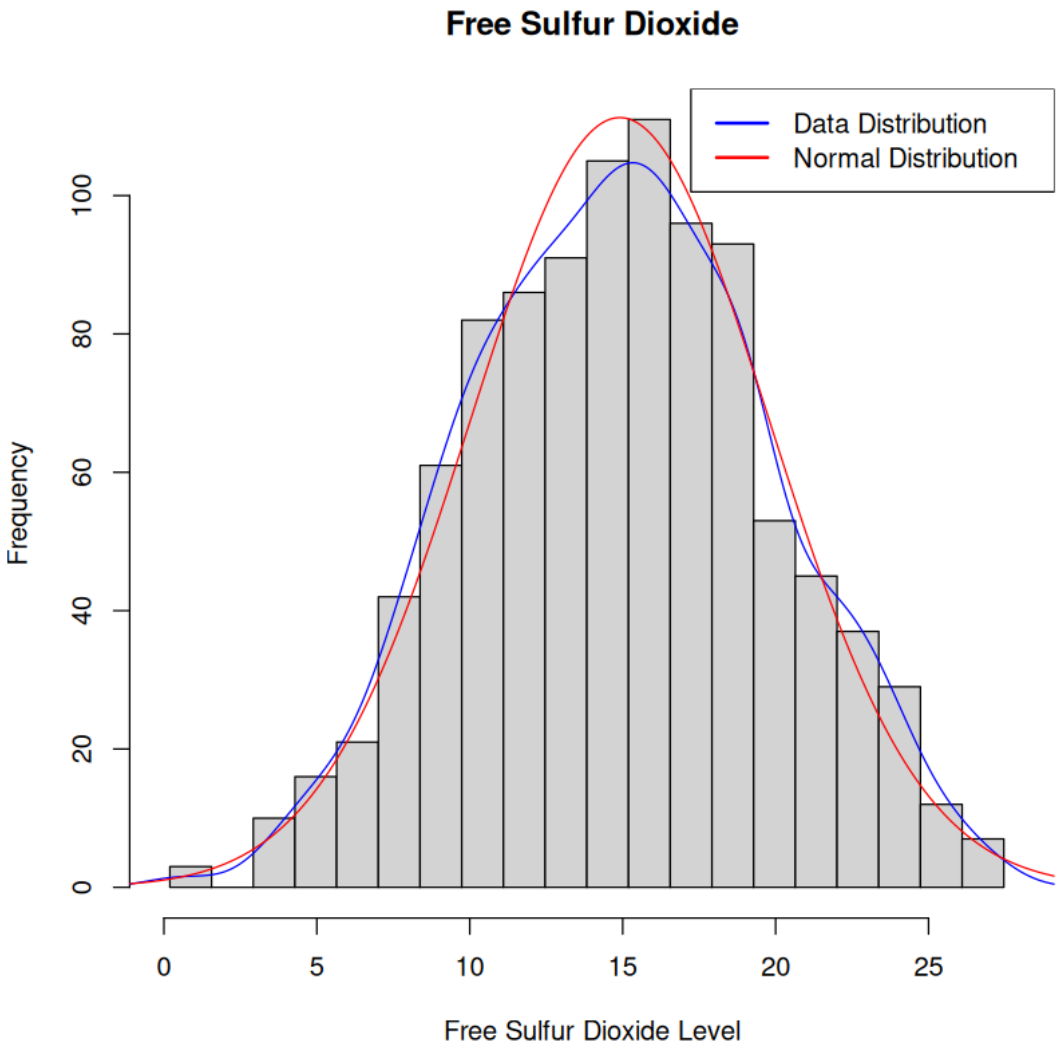
[34]:
```
# Chlorides

plot(data["chlorides"], "Chlorides", "Chlorides Level", c(0, 0.15))
```

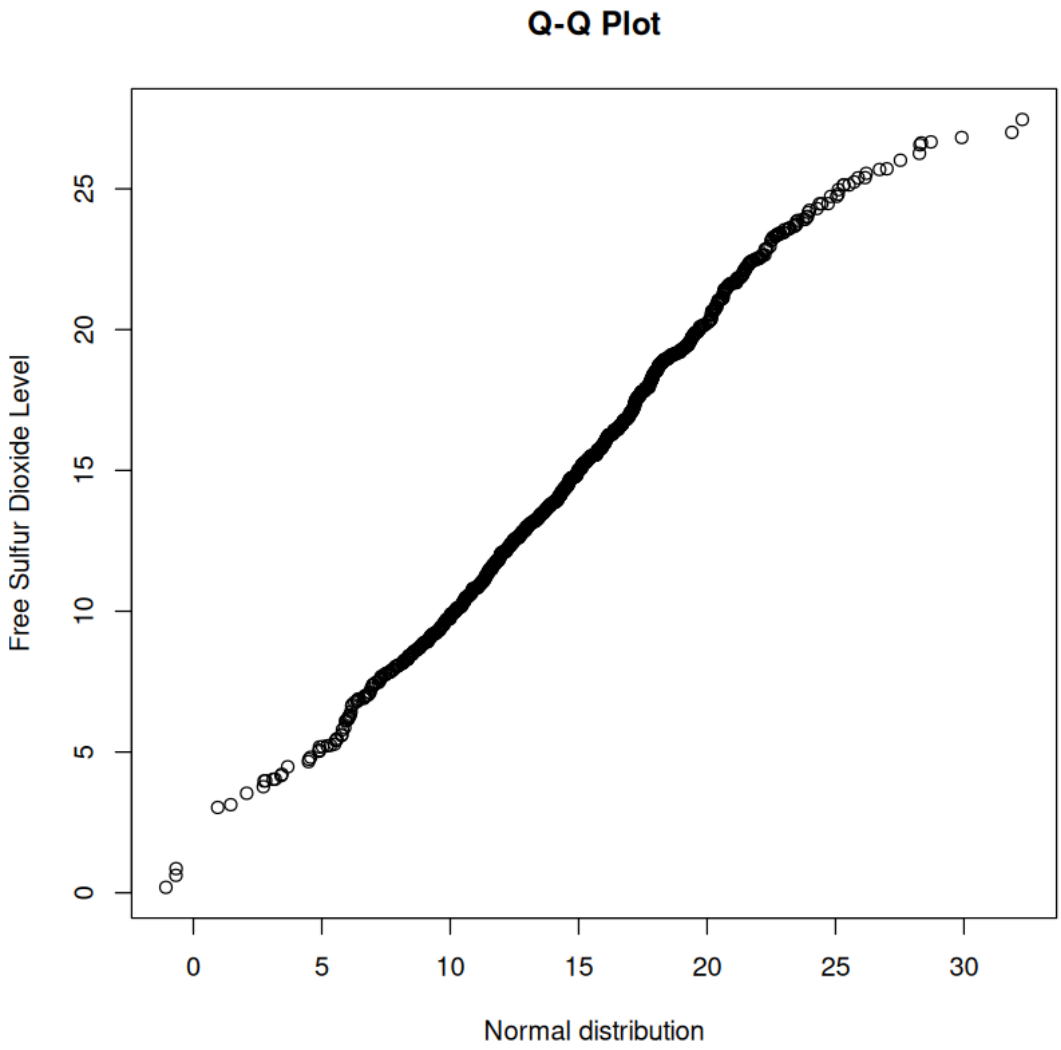

Chlorides
Data Distribution
Normal Distribution
Frequency
0
20
40
60
80
100
120
0.00
0.05
0.10
0.15
Chlorides Level



Karena plot garis pada histogram tidak mendekati distribusi normal, maka data **Chlorides** dapat dikatakan tidak berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik kiri tidak berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

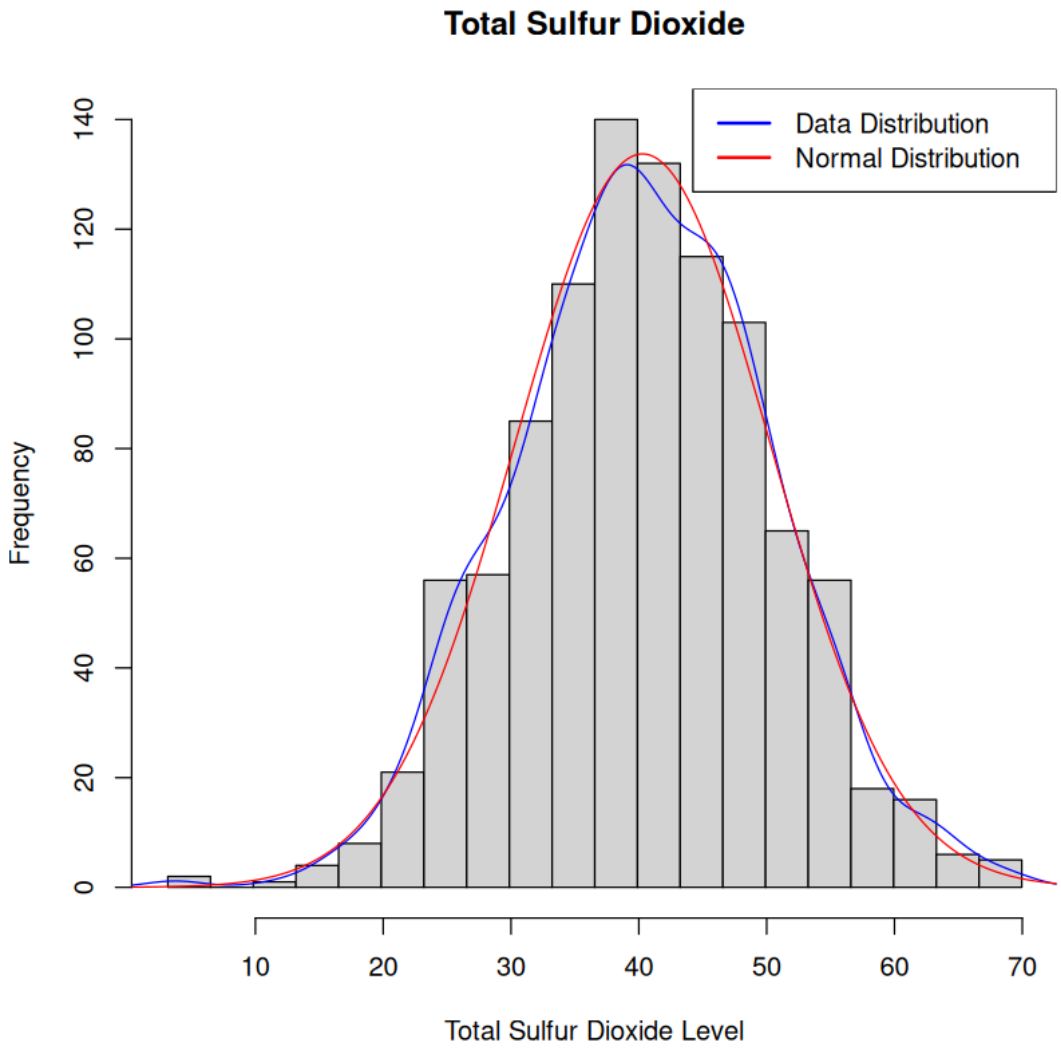
[35]:
```
# Free Sulfur Dioxide

plot(data["free.sulfur.dioxide"], "Free Sulfur Dioxide", "Free Sulfur Dioxide␣
↪Level")
```

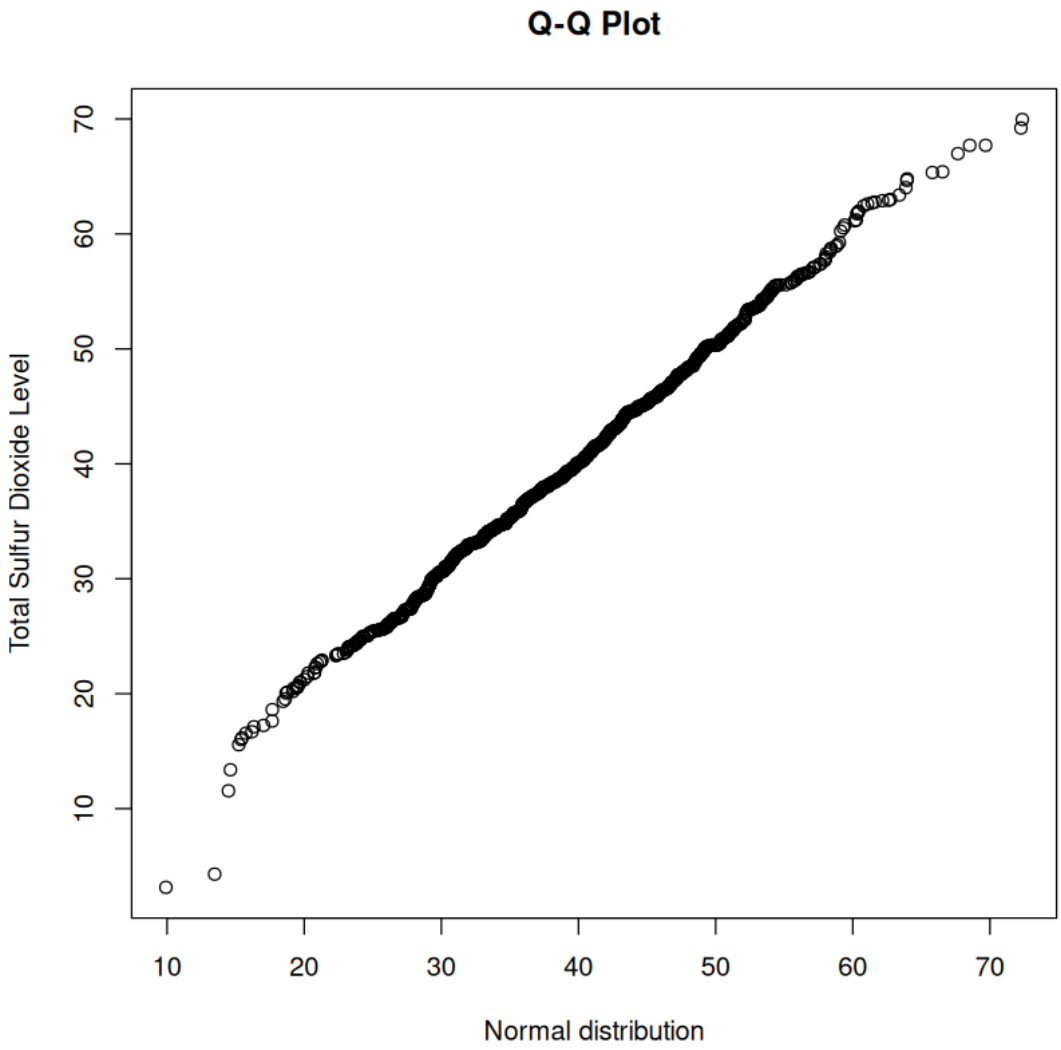

Free Sulfur Dioxide
Data Distribution
Normal Distribution
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
Free Sulfur Dioxide Level



Karena plot garis pada histogram tidak menyerupai distribusi normal, maka data **Free Sulfur Dioxide** dapat dikatakan tidak berdistribusi normal. Namun, Q-Q plot menunjukkan bahwa titik berada pada garis lurus yang menyimpulkan bahwa data berdistribusi normal.

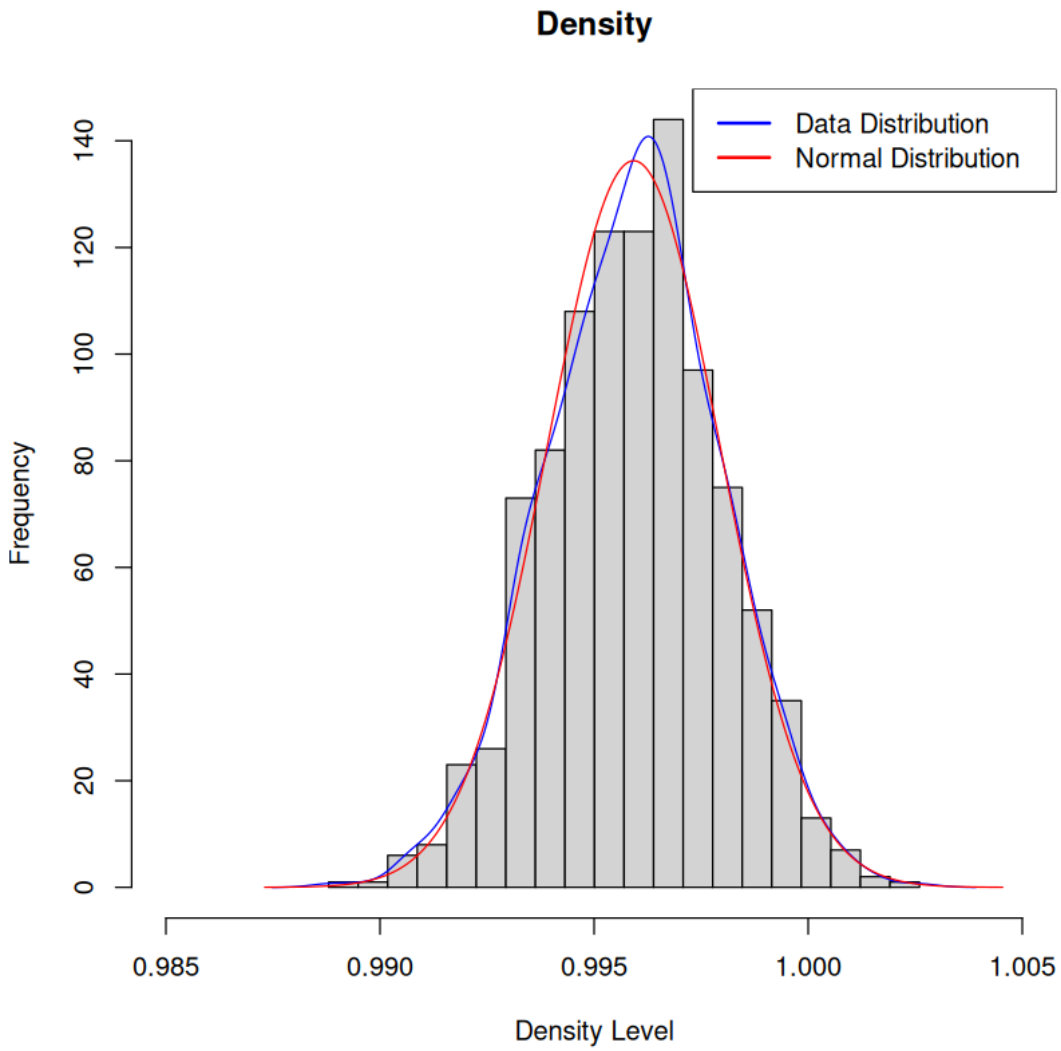
[36]:
```
# Total Sulfur Dioxide

plot(data["total.sulfur.dioxide"], "Total Sulfur Dioxide", "Total Sulfur␣
↪Dioxide Level")
```

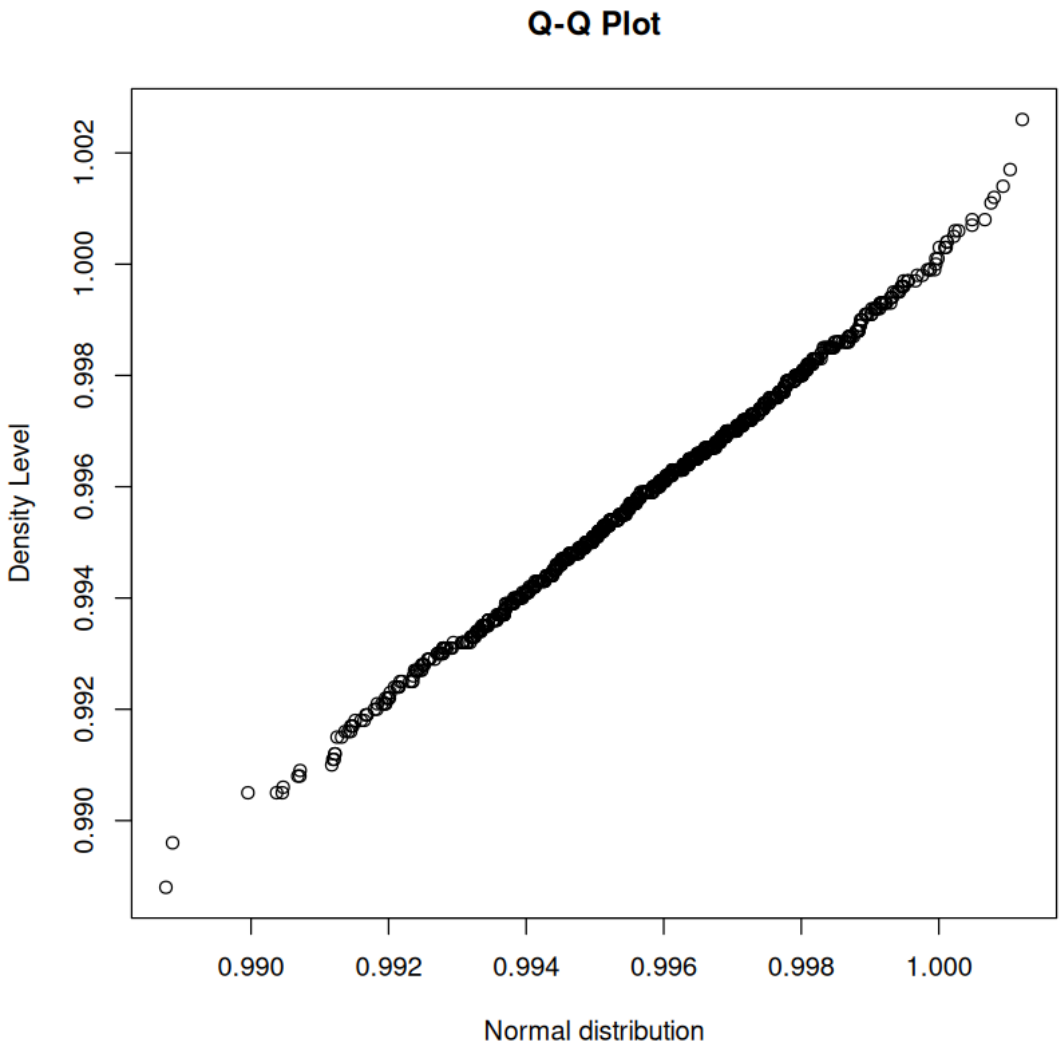

Total Sulfur Dioxide
Data Distribution
Normal Distribution
Frequency
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70
Total Sulfur Dioxide Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Total Sulfur Dioxide** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

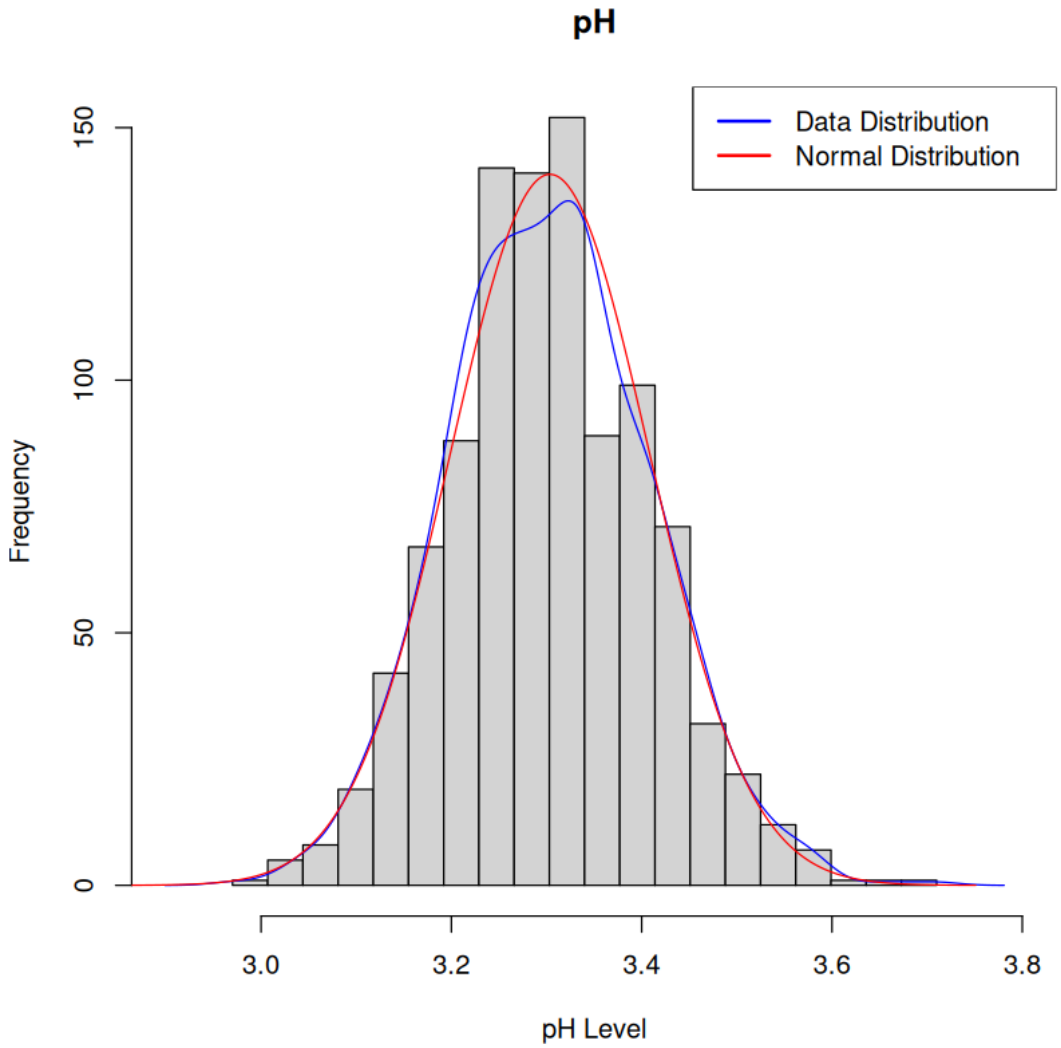
[37]:
```
# Density

plot(data["density"], "Density", "Density Level", c(0.985, 1.005))
```

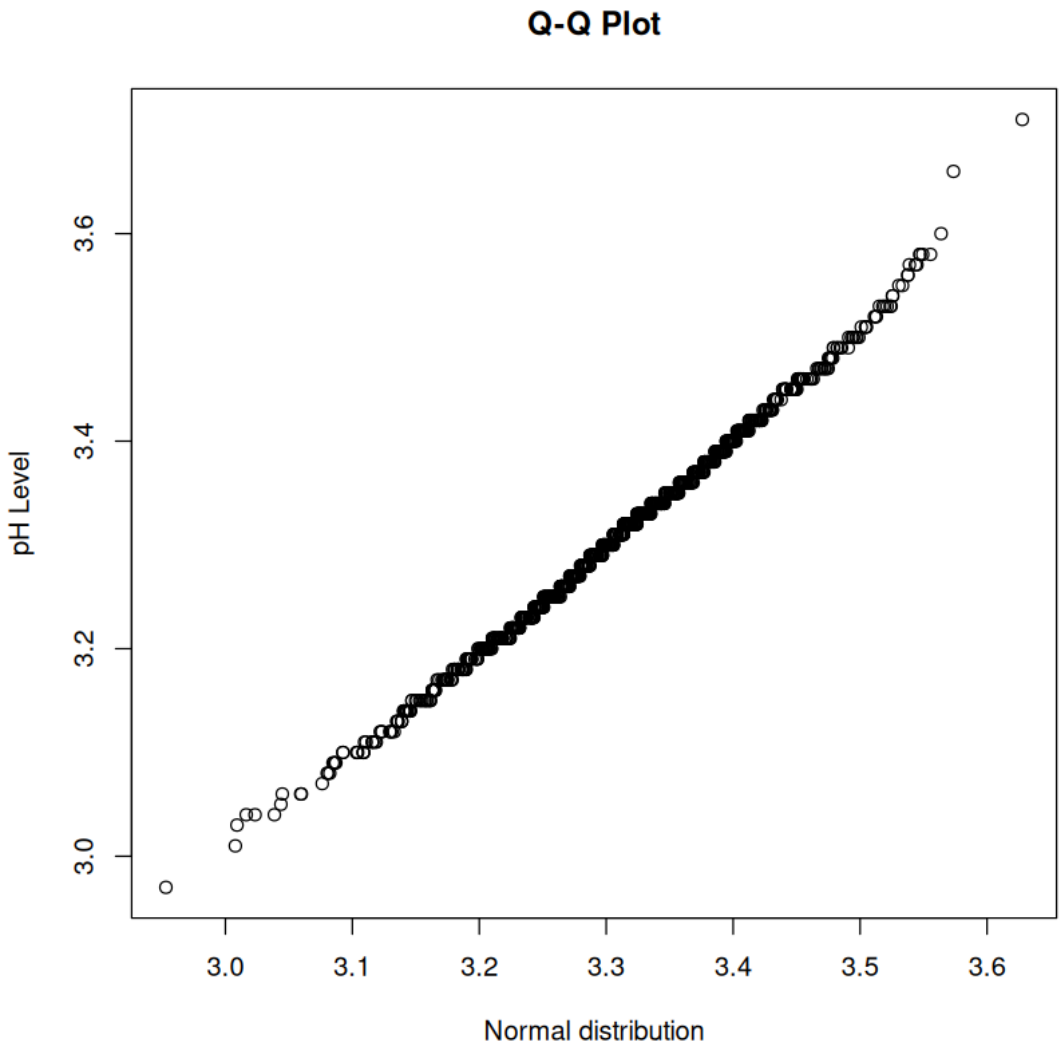

Density
Data Distribution
Normal Distribution
Frequency
0
20
40
60
80
100
120
140
0.985
0.990
0.995
1.000
1.005
Density Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Density** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

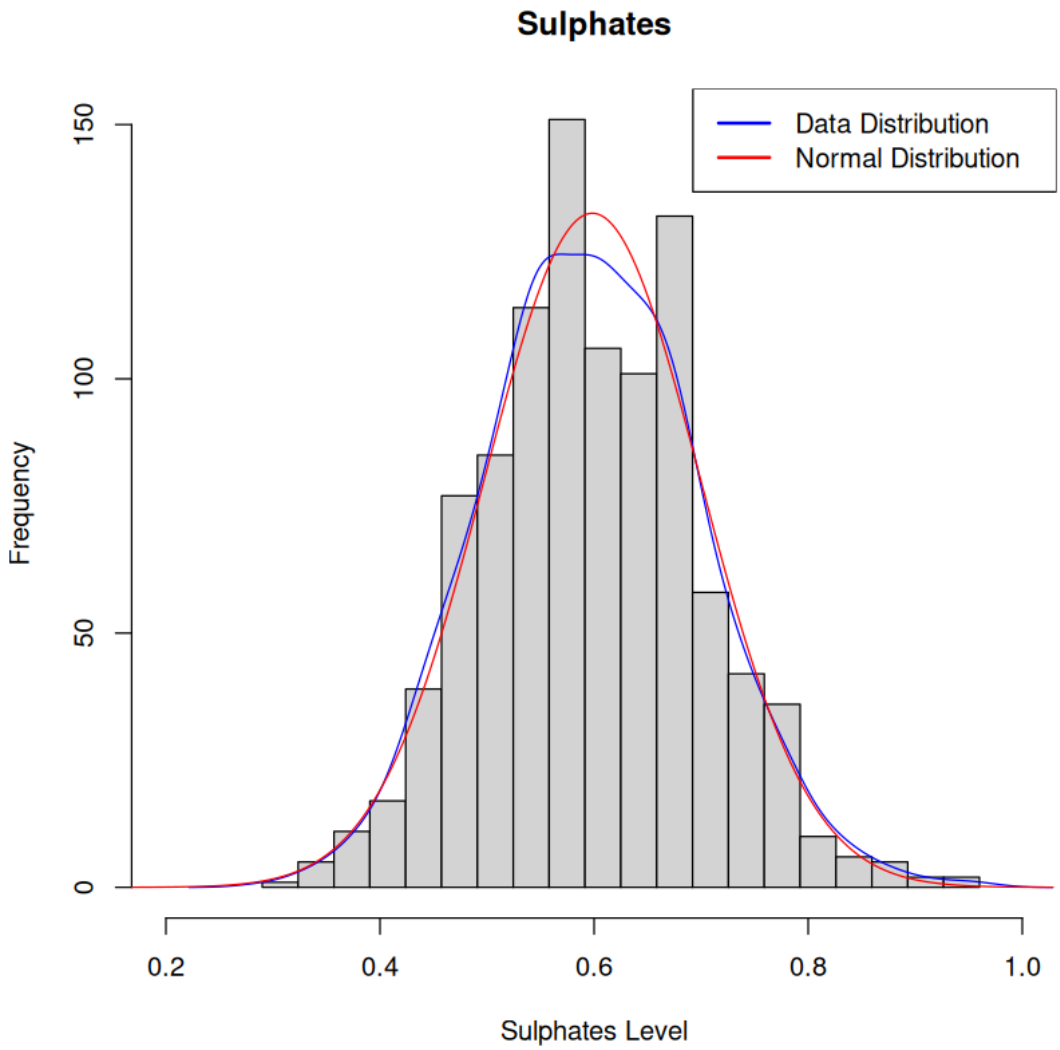
[38]:
```
# pH

plot(data["pH"], "pH", "pH Level", c(2.9, 3.8))
```

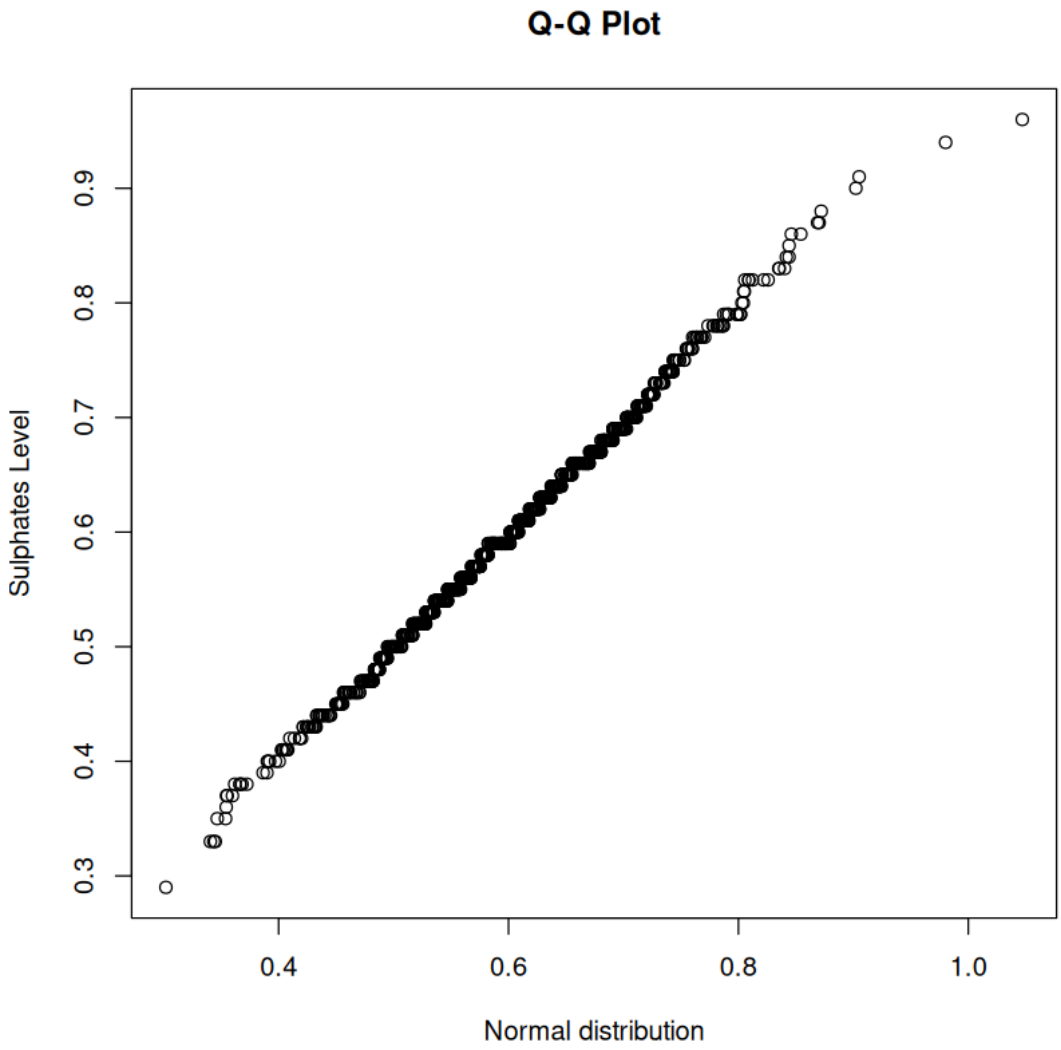

pH
Data Distribution
Normal Distribution
Frequency
0
50
100
150
3.0
3.2
3.4
3.6
3.8
pH Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **pH** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

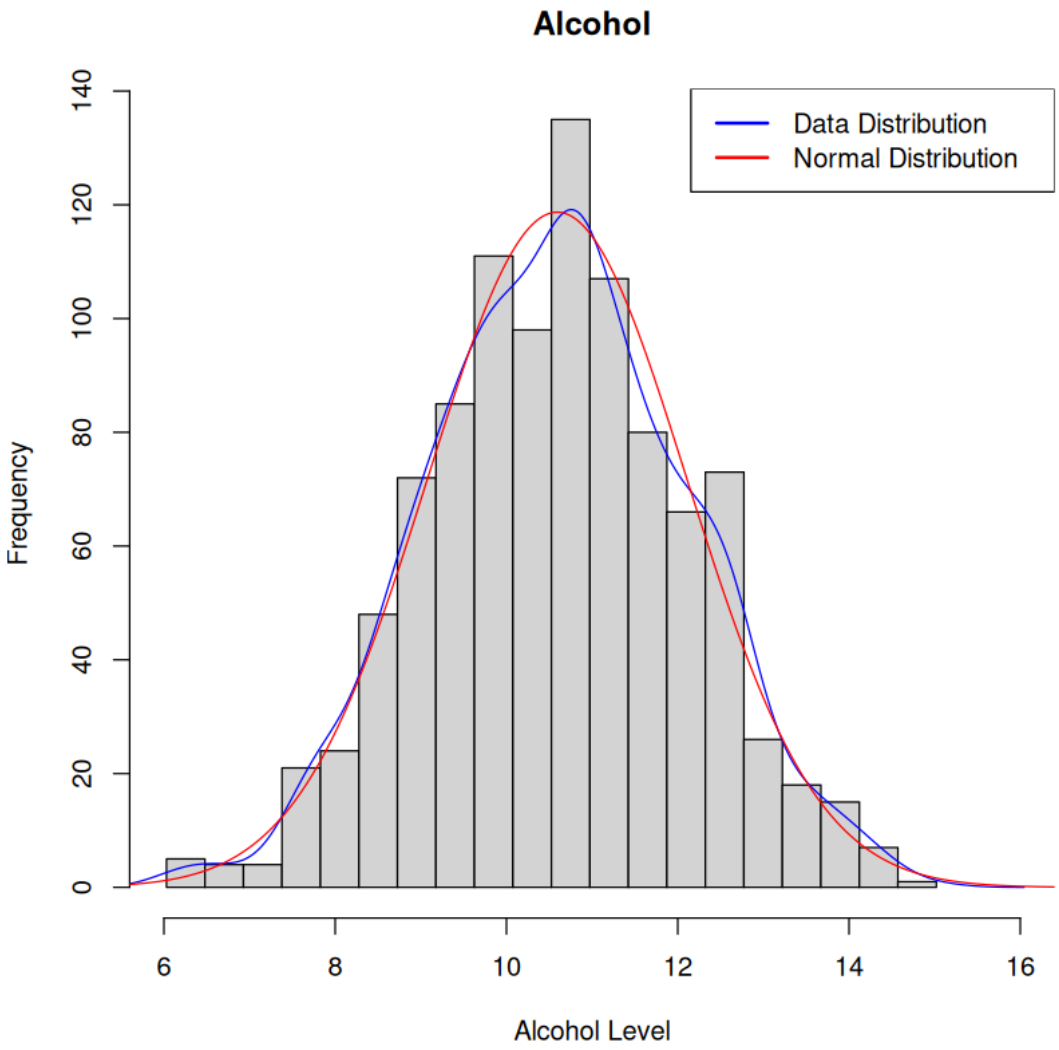
[39]:
```
# Sulphates

plot(data["sulphates"], "Sulphates", "Sulphates Level", c(0.2, 1))
```

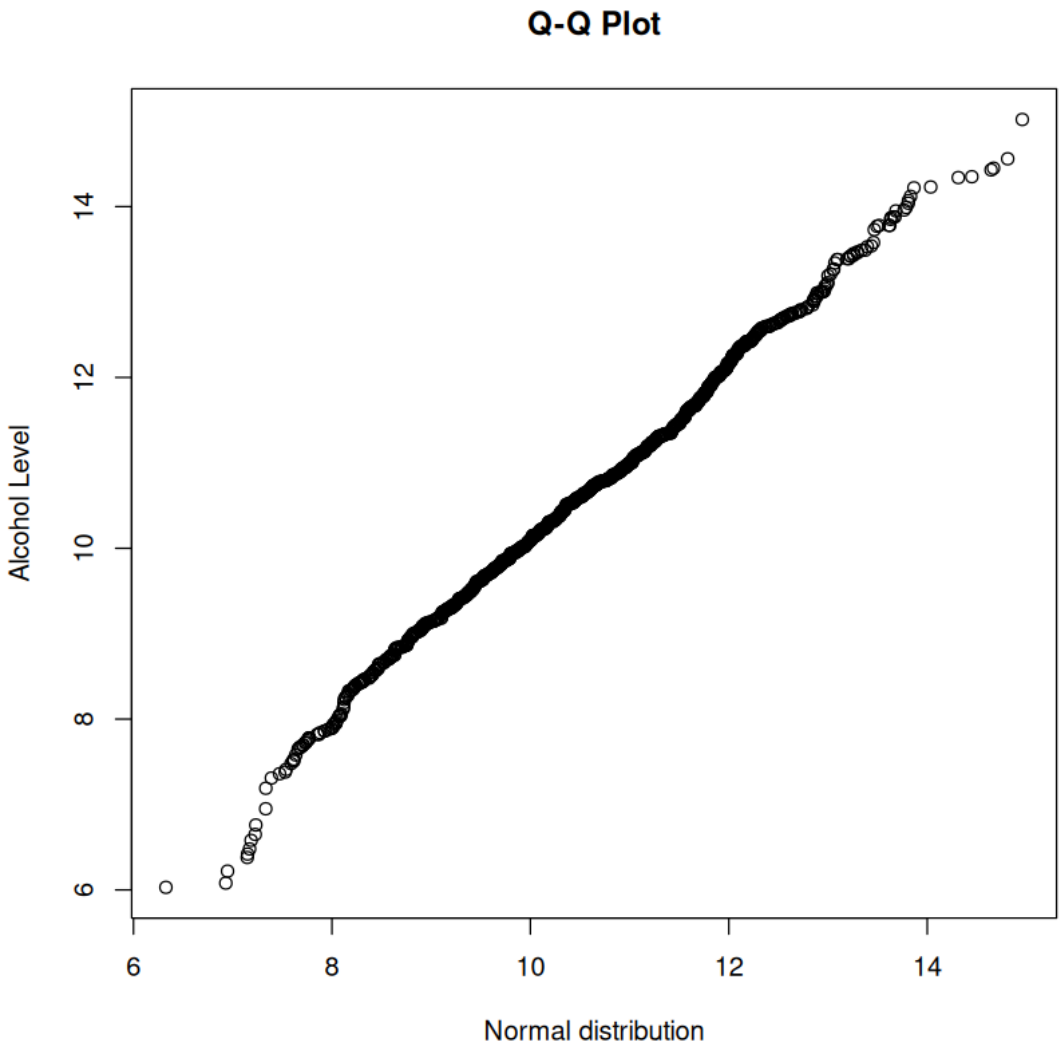

Sulphates
Data Distribution
Normal Distribution
Frequency
0
50
100
150
0.2
0.4
0.6
0.8
1.0
Sulphates Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Sulphates** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

[40]:
```
# Alcohol

plot(data["alcohol"], "Alcohol", "Alcohol Level")
```


Alcohol
Data Distribution
Normal Distribution
Frequency
0
20
40
60
80
100
120
140
6
8
10
12
14
16
Alcohol Level



Karena plot garis pada histogram hampir mendekati distribusi normal, maka data **Alcohol** dapat dikatakan berdistribusi normal. Selain itu, Q-Q plot juga menunjukkan bahwa titik berada pada garis lurus, sehingga memperkuat kesimpulan plot histogram.

1.5 Enam Langkah Tes Hipotesis

1. Tentukan hipotesis nol (H_0: $\theta = \theta_0$), dimana θ bisa berupa μ, σ^2, p, atau data lain berdistribusi tertentu (normal, binomial, dll.).
2. Pilih hipotesis alternatif H_1, salah satu dari $\theta < \theta_0$, $\theta > \theta_0$, $\theta \neq \theta_0$.
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 salah satu keputusan, yaitu **TOLAK** H_0 jika nilai uji terletak di daerah kritis atau

dengan tes signifikan, **TOLAK** H_0 jika p-value lebih kecil dibandingkan dengan tingkat signifikan α yang diinginkan.

[41]:
```r
# Fungsi pengecekan hipotesis

one_sample_t_value_test <- function(H0, column, alt) {
    column <- as.numeric(unlist(column))

    mean <- mean(column)
    standard_distribution <- sd(column)
    length <- length(column)

    t_value <- (mean - H0) / (standard_distribution / sqrt(length))

    if (alt == "greater") {
        p_value <- pt(t_value, length - 1, lower.tail = FALSE)
    }
    else if (alt == "less") {
        p_value <- pt(t_value, length - 1, lower.tail = TRUE)
    }
    else {
        p_value <- 2 * pt(t_value, length - 1, lower.tail = FALSE)
        if (p_value > 1) {
            p_value <- 2 - p_value
        }
    }

    cat(paste("mean       : ", mean, "\n"))
    cat(paste("t-value    : ", t_value, "\n"))
    cat(paste("p-value    : ", p_value, "\n"))
}

two_sample_t_value_test <- function(H0, X, Y, alt) {
    X <- as.numeric(unlist(X))
    x_mean <- mean(X)
    x_standard_deviation <- sd(X)
    x_length <- length(X)

    Y <- as.numeric(unlist(Y))
    y_mean <- mean(Y)
    y_standard_deviation <- sd(Y)
    y_length <- length(Y)

    t_value <- ((x_mean - y_mean) - H0) /
                sqrt(
                    ((x_standard_deviation^2)/x_length) +
                    ((y_standard_deviation^2)/y_length)
```

```
                )

    v <- ((x_standard_deviation^2 / x_length) + (y_standard_deviation^2 /␣
 ↪y_length))^2 /
            (
                ((x_standard_deviation^2 / x_length)^2 / (x_length - 1)) +
                ((y_standard_deviation^2 / y_length)^2 / (y_length - 1))
            )

    if (alt == "greater") {
        p_value <- pt(t_value, v, lower.tail = FALSE)
    }
    else if (alt == "less") {
        p_value <- pt(t_value, v, lower.tail = TRUE)
    }
    else {
        p_value <- 2 * pt(t_value, v, lower.tail = FALSE)
        if (p_value > 1) {
            p_value <- 2 - p_value
        }
    }

    cat(paste("x-mean     : ", x_mean, "\n"))
    cat(paste("y-mean     : ", y_mean, "\n"))
    cat(paste("difference : ", abs(x_mean - y_mean), "\n"))
    cat(paste("t-value    : ", t_value, "\n"))
    cat(paste("p-value    : ", p_value, "\n"))
}

one_sample_proportion_test <- function(x_value, n_value, H0, alt) {
    p_bar <- x_value / n_value
    q_bar <- 1 - p_bar

    z_value <- (p_bar - H0) / sqrt(p_bar * q_bar / n_value)

    if (alt == "greater") {
        p_value <- pnorm(z_value, lower.tail = FALSE)
    }
    else if (alt == "less") {
        p_value <- pnorm(z_value, lower.tail = TRUE)
    }
    else {
        p_value <- 2 * pnorm(z_value, lower.tail = FALSE)
    }

    cat(paste("p       : ", p_bar, "\n"))
    cat(paste("z-value : ", z_value, "\n"))
```

```
    cat(paste("p-value : ", p_value, "\n"))
}

two_sample_proportion_test <- function(x_value, n_value, alt) {
    p_one <- x_value[1] / n_value[1]

    p_two <- x_value[2] / n_value[2]

    p_bar <- (x_value[1] + x_value[2]) / (n_value[1] + n_value[2])
    q_bar <- 1 - p_bar

    z_value <- (p_one - p_two) / (sqrt(p_bar * q_bar * ((1 / n_value[1]) + (1 /␣
 ↪n_value[2]))))

    if (alt == "greater") {
        p_value <- pnorm(z_value, lower.tail = FALSE)
    }
    else if (alt == "less") {
        p_value <- pnorm(z_value, lower.tail = TRUE)
    }
    else {
        p_value <- 2 * pnorm(z_value, lower.tail = FALSE)
    }

    cat(paste("p1      : ", p_one, "\n"))
    cat(paste("p2      : ", p_two, "\n"))
    cat(paste("z-value : ", z_value, "\n"))
    cat(paste("p-value : ", p_value, "\n"))
}

variance_test <- function(X, Y, ratio_value, alt) {
    X <- as.numeric(unlist(X))
    x_standard_deviation <- sd(X)
    x_length <- length(X)

    Y <- as.numeric(unlist(Y))
    y_standard_deviation <- sd(Y)
    y_length <- length(Y)


    f_value <- x_standard_deviation^2 / y_standard_deviation^2

    if (alt == "greater") {
        p_value <- pf(f_value, x_length - 1, y_length - 1, lower.tail = FALSE)
    }
    else if (alt == "less") {
        p_value <- pf(f_value, x_length - 1, y_length - 1, lower.tail = TRUE)
```

```
    }
    else {
        p_value <- 2 * pf(f_value, x_length - 1, y_length - 1, lower.tail =␣
↪FALSE)
        if (p_value > 1) {
            p_value <- 2 - p_value
        }
    }

    cat(paste("x-sd    : ", x_standard_deviation, "\n"))
    cat(paste("y-sd    : ", y_standard_deviation, "\n"))
    cat(paste("f-value : ", f_value, "\n"))
    cat(paste("p-value : ", p_value, "\n"))
}
```

[ ]:

1.6 Soal 4

Melakukan test hipotesis 1 sampel

1.6.1 a. Nilai rata-rata pH di atas 3.29

1. Tentukan hipotesis nol
$$H_0 : \mu = 3.29$$
2. Pilih hipotesis alternatif
$$H_1 : \mu > 3.29$$
3. Tentukan tingkat signifikan
$$\alpha = 0.05$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis
Akan digunakan uji statistik tes mean untuk menghitung nilai t.
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
$$v = n - 1$$
Daerah kritis yang diperoleh berada pada rentang
$$t > t_{0.05} = 1.646$$

[42]:
```
pH <- data["pH"]

one_sample_t_value_test(3.29, pH, "greater")
```

```
mean        :  3.30361
t-value     :  4.10378079336511
p-value     :  2.19795830638601e-05
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan p-value $= 0.00002198$

6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $P(t > 4.1038) = 0.00002 < 0.05$, maka hipotesis nol ditolak

1.6.2 b. Nilai rata-rata residual sugar tidak sama dengan 2.50

1. Tentukan hipotesis nol
$$H_0 : \mu = 2.50$$

2. Pilih hipotesis alternatif
$$H_1 : \mu \neq 2.50$$

3. Tentukan tingkat signifikan
$$\alpha = 0.05$$

4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
   Akan digunakan uji statistik tes mean untuk menghitung nilai t.
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
$$v = n - 1$$
   Daerah kritis yang diperoleh berada pada rentang
$$t < -t_{0.025} = -1.962 \quad or \quad t > t_{0.025} = 1.962$$

[43]:
```
residual_sugar = data["residual.sugar"]

one_sample_t_value_test(2.50, residual_sugar, "two.sided")
```

```
mean        :  2.56710368250676
t-value     :  2.14796194355395
p-value     :  0.0319567267086168
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan nilai p-value $= 0.03196$

6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $2P(t > 2.1479) = 0.0319 < 0.05$, maka hipotesis nol ditolak

1.6.3 c. Nilai rata-rata 150 baris pertama dari kolom sulphates bukan 0.65

1. Tentukan hipotesis nol
$$H_0 : \mu = 0.65$$
2. Pilih hipotesis alternatif
$$H_1 : \mu \neq 0.65$$
3. Tentukan tingkat signifikan
$$\alpha = 0.05$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
Akan digunakan uji statistik tes mean untuk menghitung nilai t.
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
$$v = n - 1$$
Daerah kritis yang diperoleh berada pada rentang
$$t < -t_{0.025} = -1.962 \quad or \quad t > t_{0.025} = 1.962$$

[44]:
```
sulphates <- as.numeric(unlist(data["sulphates"][1:150, 1]))

one_sample_t_value_test(0.65, sulphates, "two.sided")
```

```
mean        :  0.605866666666667
t-value     :  -4.96484339331592
p-value     :  1.85901512139708e-06
```

5. Hitung nilai uji statistik dan p-value.
Didapatkan p-value $= 0.00000186$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
Dikarenakan nilai $2P(t < -4.9648) = 0.00000186 < 0.05$, maka hipotesis nol ditolak

1.6.4 d. Nilai rata-rata total sulfur dioxide di bawah 35

1. Tentukan hipotesis nol
$$H_0 : \mu = 35$$
2. Pilih hipotesis alternatif
$$H_1 : \mu < 35$$
3. Tentukan tingkat signifikan
$$\alpha = 0.05$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
Akan digunakan uji statistik tes mean untuk menghitung nilai t.
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$v = n - 1$$

Daerah kritis yang diperoleh berada pada rentang

$$t < -t_{0.05} = -1.646$$

[45]:
```
total_sulfur_dioxide <- data["total.sulfur.dioxide"]

one_sample_t_value_test(35, total_sulfur_dioxide, "less")
```

```
mean        :  40.29015
t-value     :  16.7863873722967
p-value     :  1
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan p-value $= 1$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $P(t < 16.7863) = 1 > 0.05$, maka hipotesis nol diterima

1.6.5 e. Proporsi nilai total sulfur dioxide yang lebih dari 40 adalah tidak sama dengan 50%

1. Tentukan hipotesis nol
$$H_0 : p = 0.5$$
2. Pilih hipotesis alternatif
$$H_1 : p \neq 0.5$$
3. Tentukan tingkat signifikan
$$\alpha = 0.05$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
   Akan digunakan uji statistik tes proporsi untuk menghitung nilai Z.
$$Z = \frac{\bar{p} - p}{\sqrt{pq/n}}$$
   Daerah kritis yang diperoleh berada pada rentang
$$Z < -Z_{0.025} = -1.9599 \quad or \quad Z > Z_{0.025} = 1.9599$$

[46]:
```
total_sulfur_dioxide <- data["total.sulfur.dioxide"]

x_value <- length(total_sulfur_dioxide[total_sulfur_dioxide > 40])

n_value <- length(total_sulfur_dioxide[,])

one_sample_proportion_test(x_value, n_value, 0.5, "two.sided")
```

```
p       :  0.512
z-value :  0.759165309542734
p-value :  0.447753674993189
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan p-value $= 0.4477$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $2P(Z > 0.7591) = 0.4477 > 0.05$, maka hipotesis nol diterima

[ ]:

1.7 Soal 5

Melakukan tes hipotesis 2 sampel

1.7.1 a. Data kolom *fixed acidity* dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata kedua bagian tersebut sama

1. Tentukan hipotesis nol
$$H_0 : \mu_{p_0} = \mu_{p_1}$$
2. Pilih hipotesis alternatif
$$H_1 : \mu_{p_0} \neq \mu_{p_1}$$
3. Tentukan tingkat signifikan
$$\alpha = 0.05$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
   Akan digunakan uji statistik tes mean untuk menghitung nilai t.
$$t = \frac{(\bar{x_0} - \bar{x_1}) - \mu}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Daerah kritis yang diperoleh berada pada rentang

$$t < -t_{0.025} = -1.962 \quad or \quad t > t_{0.025} = 1.962$$

[47]:
```
fixed_acidity <- data["fixed.acidity"]

first_half <- fixed_acidity[1 : (length(fixed_acidity[, ])/2), ]
second_half <- fixed_acidity[(length(fixed_acidity[, ])/2 + 1) :␣
 ↪length(fixed_acidity[, ]), ]

two_sample_t_value_test(0, first_half, second_half, "two.sided")
```

```
x-mean      :  7.15352
y-mean      :  7.15154
difference  :  0.00198000000000054
t-value     :  0.0260410699990871
p-value     :  0.979229786496273
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan p-value $= 0.97922$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $2P(t > 0.026041) = 0.97922 > 0.05$, maka hipotesis nol diterima

1.7.2 b. Data kolom *chlorides* dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata bagian awal lebih besar daripada rata-rata bagian akhir sebesar 0.001

1. Tentukan hipotesis nol
$$H_0 : \mu_{p_0} = \mu_{p_1}$$
2. Pilih hipotesis alternatif
$$H_1 : \mu_{p_0} - \mu_{p_1} > 0.001$$
3. Tentukan tingkat signifikan
$$\alpha = 0.05$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
   Akan digunakan uji statistik tes mean untuk menghitung nilai t.
$$t = \frac{(\bar{x_0} - \bar{x_1}) - \mu}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Daerah kritis yang diperoleh berada pada rentang

$$t > t_{0.05} = 1.646$$

[48]:
```
chlorides <- data["chlorides"]

first_half <- chlorides[1: (length(chlorides[, ]) / 2), ]
second_half <- chlorides[(length(chlorides[, ])/2 + 1) : length(chlorides[, ]),␣
↪]

two_sample_t_value_test(0.001, first_half, second_half, "greater")
```

```
x-mean      :  0.0813978263367367
y-mean      :  0.0809924786789628
difference  :  0.000405347657773877
t-value     :  -0.467317122852143
p-value     :  0.679812488252313
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan p-value $= 0.67981$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $P(t > -0.46731) = 0.67981 > 0.05$, maka hipotesis nol diterima

1.7.3 c. Benarkah rata-rata sampel 25 baris pertama kolom *volatile acidity* sama dengan rata-rata sampel 25 baris pertama kolom *sulphates*

1. Tentukan hipotesis nol
$$H_0 : \mu_{p_0} = \mu_{p_1}$$
2. Pilih hipotesis alternatif
$$H_1 : \mu_{p_0} \neq \mu_{p_1}$$
3. Tentukan tingkat signifikan
$$\alpha = 0.05$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
   Akan digunakan uji statistik tes mean untuk menghitung nilai t.
$$t = \frac{(\bar{x_0} - \bar{x_1}) - \mu}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Daerah kritis yang diperoleh berada pada rentang

$$t < -t_{0.025} = -1.962 \quad or \quad t > t_{0.025} = 1.962$$

[49]:
```
volatile_acidity <- data["volatile.acidity"][1:25, ]
sulphates <- data["sulphates"][1:25, ]

two_sample_t_value_test(0, volatile_acidity, sulphates, "two.sided")
```

```
x-mean     :  0.501424
y-mean     :  0.5768
difference :  0.075376
t-value    :  -2.63748216767487
p-value    :  0.0115340886236583
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan p-value $= 0.01153$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $2P(t > -2.63748) = 0.011534 < 0.05$, maka hipotesis nol ditolak

1.7.4 d. Bagian awal kolom *residual sugar* memiliki variansi yang sama dengan bagian akhirnya

1. Tentukan hipotesis nol
$$H_0 : \sigma_0^2 = \sigma_1^2$$
2. Pilih hipotesis alternatif
$$H_1 : \sigma_0^2 \neq \sigma_1^2$$
3. Tentukan tingkat signifikan
$$\alpha = 0.1$$
4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
Akan digunakan uji statistik tes variansi untuk menghitung nilai F.
$$f = \frac{s_1^2}{s_2^2}$$

Daerah kritis yang diperoleh berada pada rentang

$$f < f_{0.95} = 0.8629 \quad or \quad f > f_{0.05} = 1.1588$$

[50]:
```
residual_sugar <- data["residual.sugar"]

first_half <- residual_sugar[1: (length(residual_sugar[, ]) / 2), ]
second_half <- residual_sugar[(length(residual_sugar[, ])/2 + 1) :␣
 ↪length(residual_sugar[, ]), ]

variance_test(first_half, second_half, 1, "two.sided")
```

```
x-sd    :  0.973535424878982
y-sd    :  1.00305641814463
f-value :  0.942004106694162
p-value :  0.504820359524758
```

5. Hitung nilai uji statistik dan f-value.
Didapatkan f-value $= 0.942$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
Dikarenakan nilai $2P(f > 0.942) = 0.5048 > 0.05$, maka hipotesis nol diterima

1.7.5 e. Proporsi nilai setengah bagian awal *alcohol* yang lebih dari 7 adalah lebih besar daripada proporsi nilai yang sama di setengah bagian akhir *alcohol*

1. Tentukan hipotesis nol
$$H_0 : p_0 = p_1$$
2. Pilih hipotesis alternatif
$$H_1 : p_0 > p_1$$

3. Tentukan tingkat signifikan

$$\alpha = 0.05$$

4. Tentukan uji statistik yang sesuai dan tentukan daerah kritis.
   Akan digunakan uji statistik tes proporsi untuk menghitung nilai Z.

$$Z = \frac{p_0 - p_1}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$p_i = \frac{x_i}{n_i}$$

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

$$q = 1 - p$$

Daerah kritis yang diperoleh berada pada rentang

$$Z > Z_{0.05} = 1.645$$

[51]:
```
alcohol <- data["alcohol"]

first_half <- alcohol[1: (length(alcohol[, ])/2), ]

second_half <- alcohol[(length(alcohol[, ])/2 + 1) : length(alcohol[, ]), ]

x_value <- c(length(na.omit(first_half[alcohol > 7])), length(na.
 ↪omit(second_half[alcohol > 7])))

n_value <- c(length(first_half), length(second_half))

two_sample_proportion_test(x_value, n_value, "greater")
```

```
p1      :  0.99
p2      :  0.99
z-value :  0
p-value :  0.5
```

5. Hitung nilai uji statistik dan p-value.
   Didapatkan p-value $= 0.5$
6. Ambil keputusan berdasarkan nilai yang diperoleh.
   Dikarenakan nilai $P(Z > 0) = 0.5 > 0.05$, maka hipotesis nol diterima