## Descriptive statistics

| | fixed acidity | volatile acidity | citric acid | residual sugar | quality | chlorides | free $SO_2$ | total $SO_2$ | density | pH | sulphates | alcohol |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 8.319 | 0.527 | 0.27 | 2.538 | 5.636023 | 0.0874 | 15.874 | 46.467 | 0.001 | 331 | 0.6581 | 10.4229 |
| median | 7.900 | 0.5200 | 0.26 | 2.200 | 6.000 | 0.0790 | 14.000 | 38.000 | 0.996 | 3.31 | 0.620 | 10.200 |
| mode | 7.200 | 0.600 | 0.00 | 2.00 | 5.00 | 0.080 | 6.000 | 28.00 | 0.997 | 3.30 | 0.600 | 9.500 |
| standard dev. | 1.741 | 0.179 | 0.194 | 1.409 | 0.807 | 0.047 | 10.460 | 32.895 | 0.001 | 0.154 | 0.169 | 1.065 |
| $Q_1$ | 7.10 | 0.39 | 0.09 | 1.90 | 5.00 | 0.070 | 7.000 | 22.000 | 0.995 | 3.210 | 0.55 | 9.5 |
| $Q_3$ | 9.20 | 0.64 | 0.42 | 2.60 | 6.00 | 0.0900 | 21.00 | 62.00 | 0.997 | 3.40 | 0.73 | 11.10 |

Interpretation

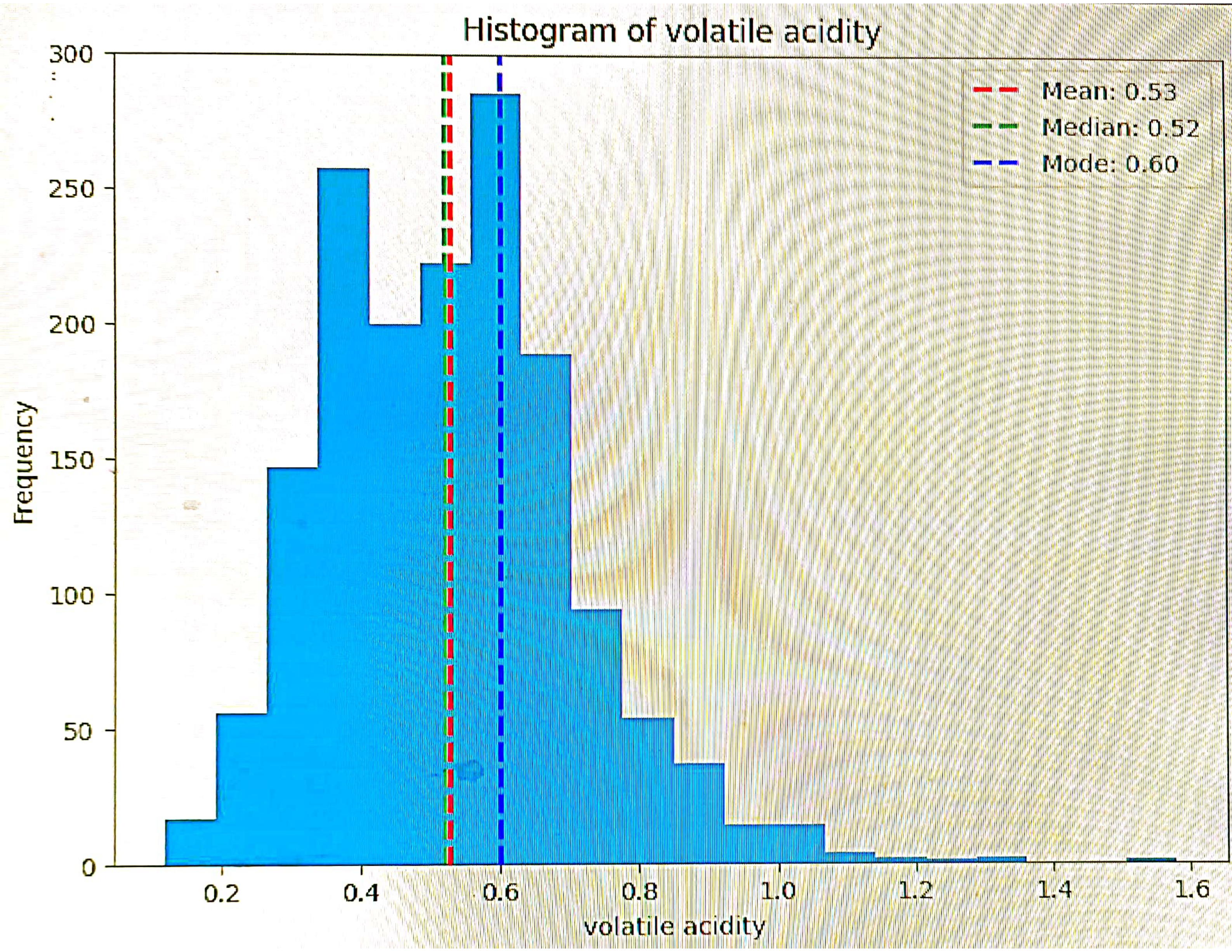
- Mean (Quality) / Median (Quality) / Mode (Quality)
  = 5.63, 6.00, 5.00 (respectively)

⇒ most of wines have quality of 5.

- Most of the wines (pH) are around (3.30) with very little standard deviation (0.154)

• Histogram of volatile acidity

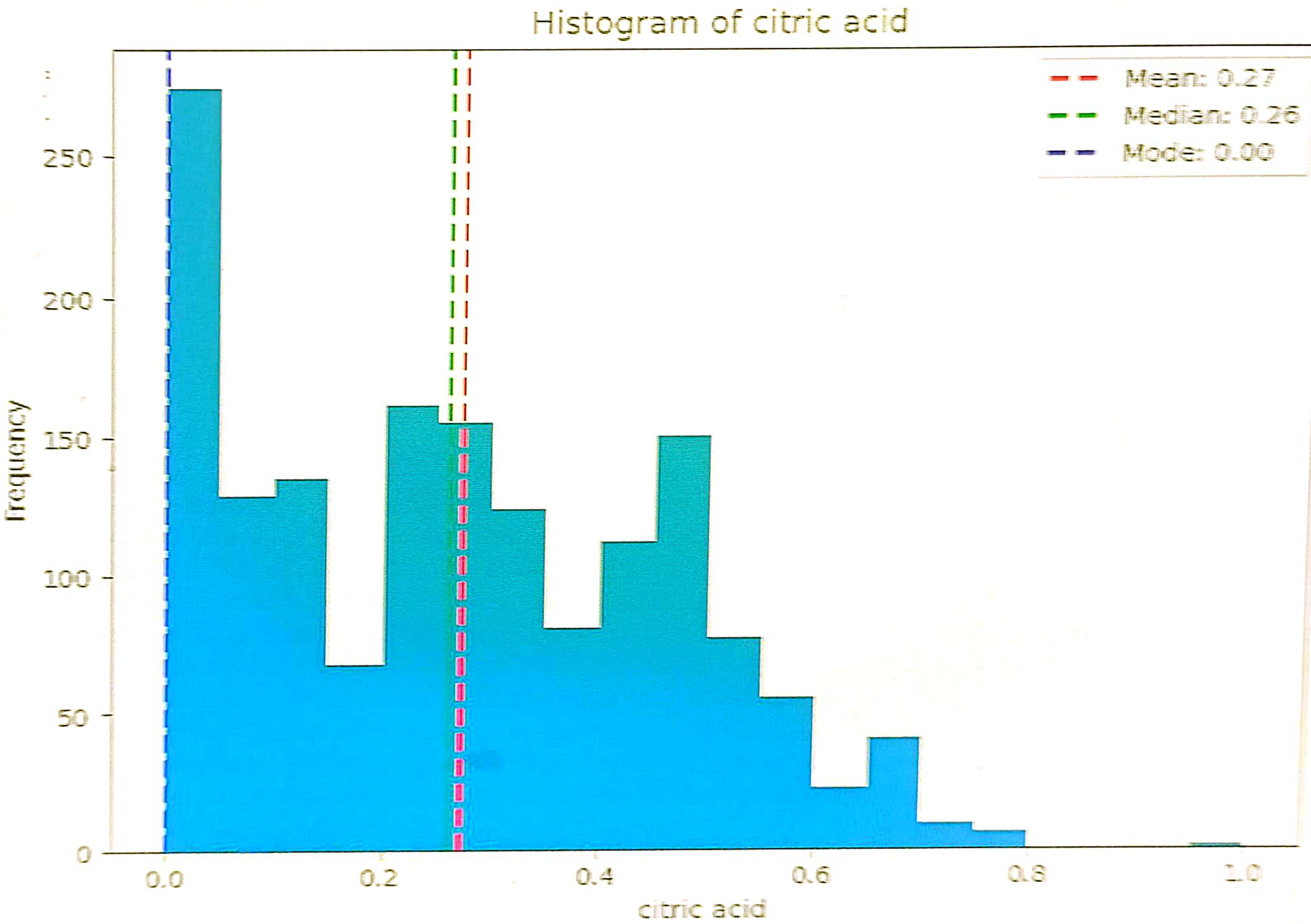
⇒ potential outliers outside 1.4 range.



Histogram of volatile acidity
Mean: 0.53
Median: 0.52
Mode: 0.60
Frequency
volatile acidity
0
50
100
150
200
250
300
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6

• citric acid

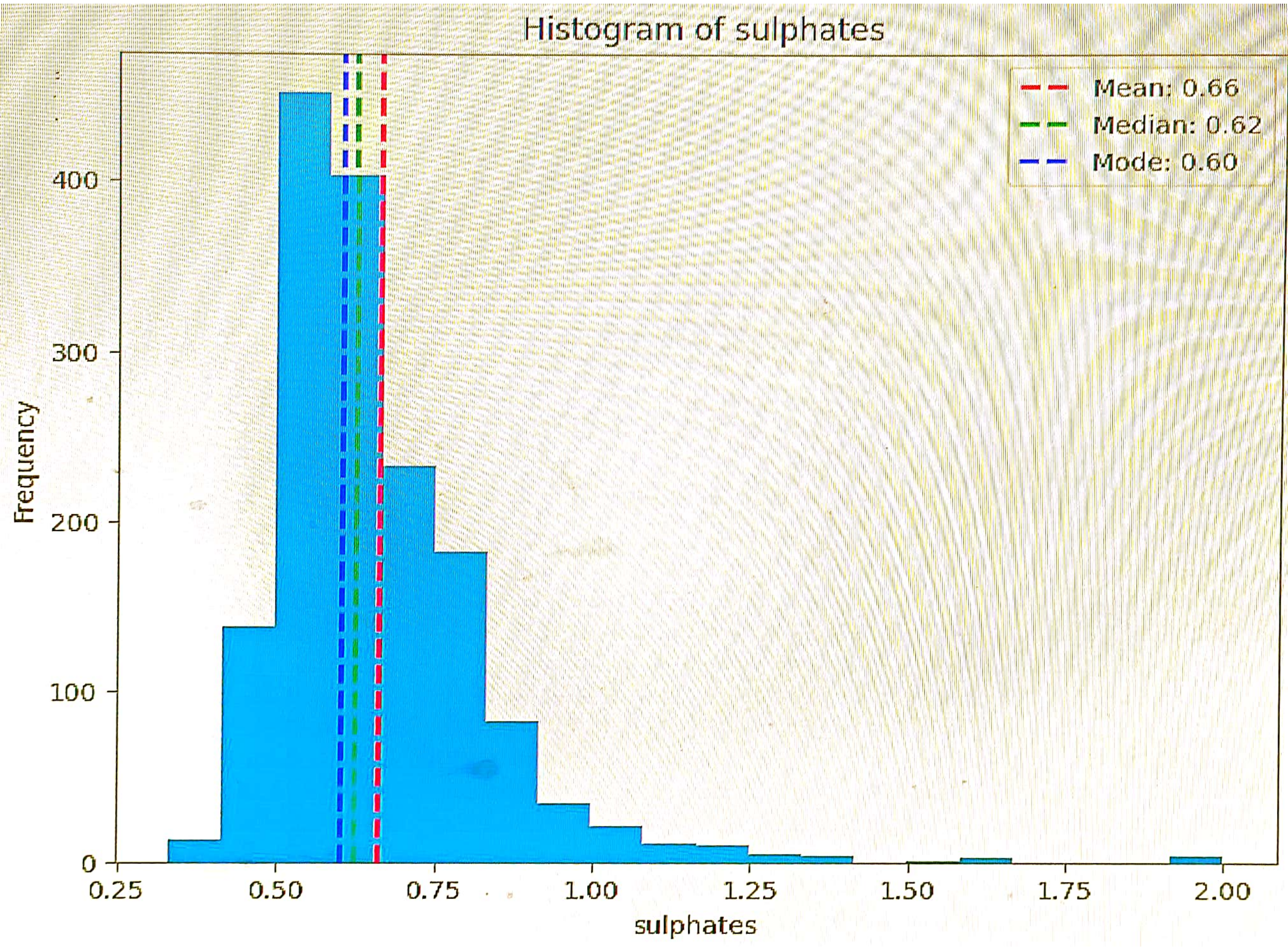
⇒ potential ~~acid~~ outliers outside 0.8 range
⇒ most of wines have no citric acid content.



Histogram of citric acid
Mean: 0.27
Median: 0.26
Mode: 0.00
frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
citric acid

- sulphates

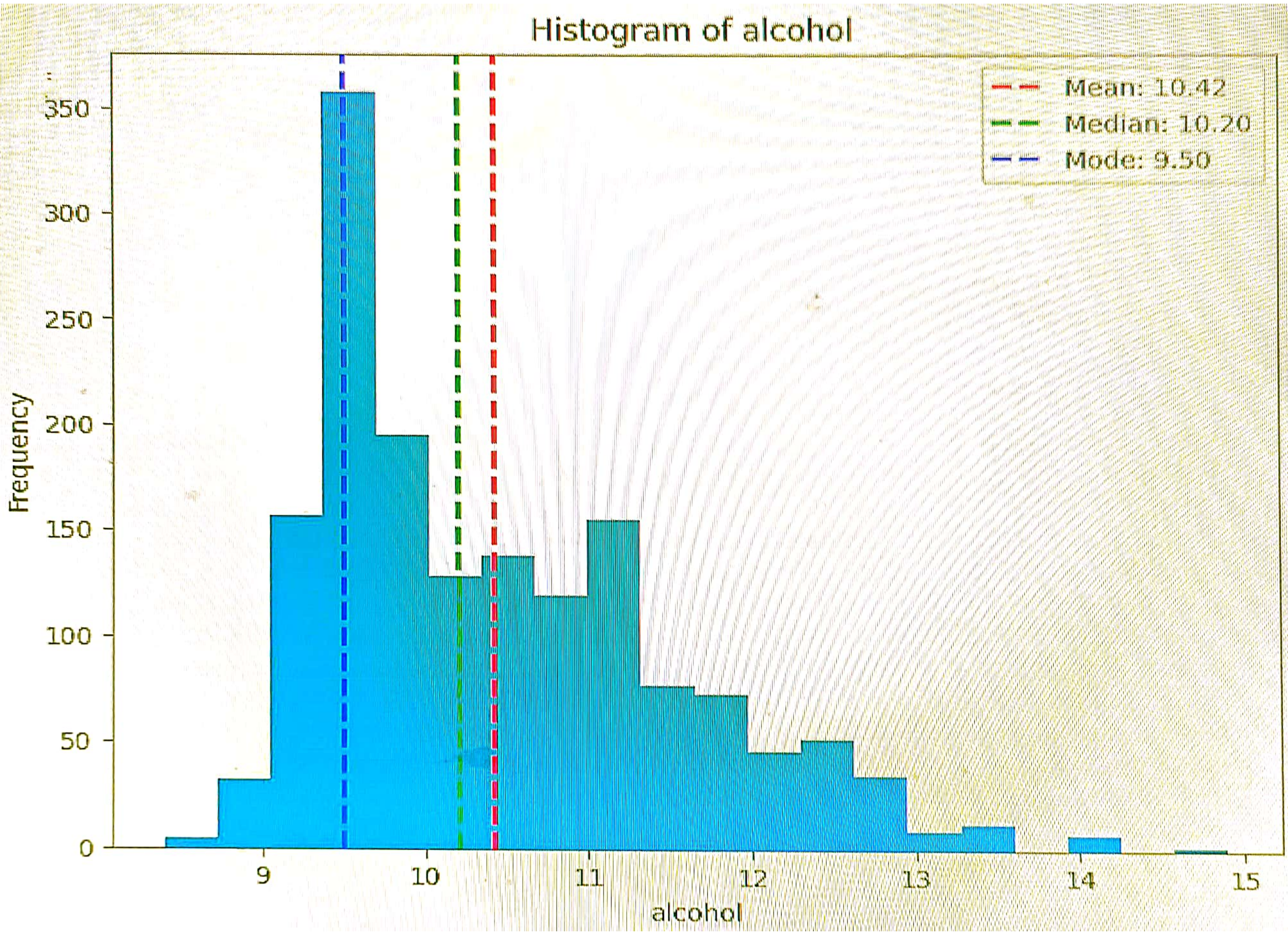
⇒ it shows outliers in the extremeties



Histogram of sulphates
Mean: 0.66
Median: 0.62
Mode: 0.60
Frequency
0
100
200
300
400
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
sulphates

- alcohol -

⇒ potential outliers in range (14, 15)


Histogram of alcohol
Mean: 10.42
Median: 10.20
Mode: 9.50
Frequency
0
50
100
150
200
250
300
350
9
10
11
12
13
14
15
alcohol

# Box plot analysis

↳ Citric acid

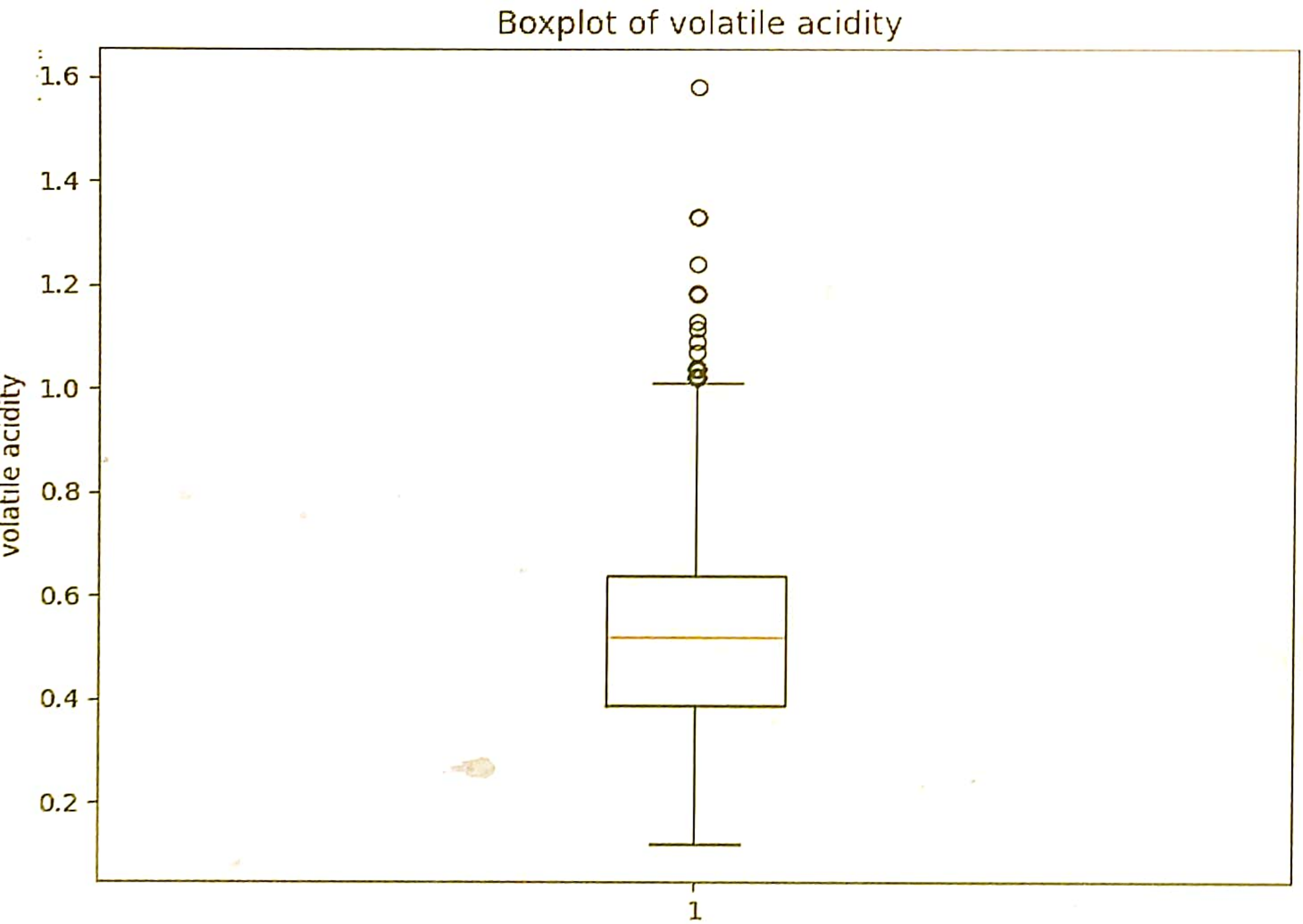
⟹ only 1 wine has citric acid content of 1 (can be an outlier)

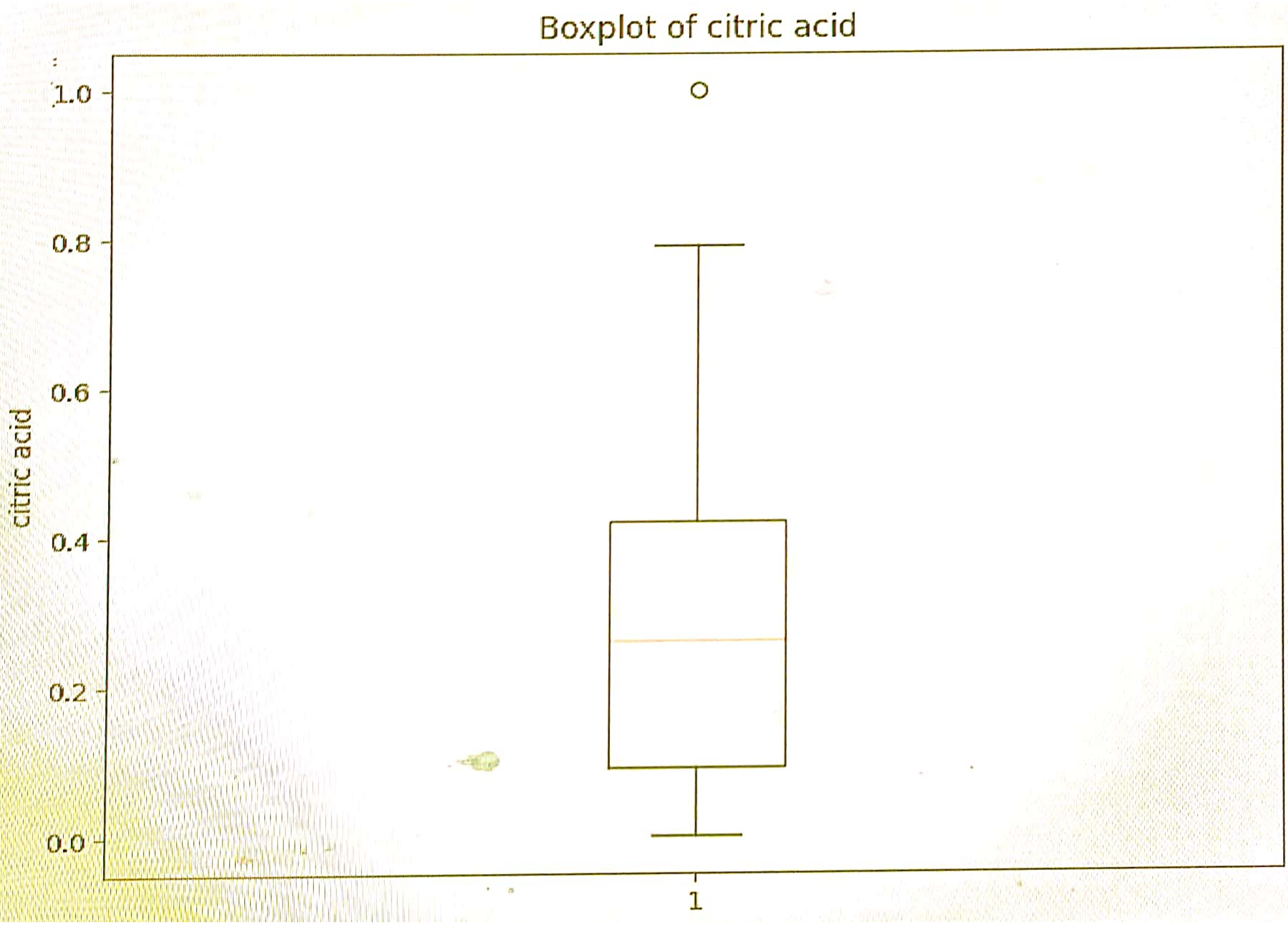
↳ Sulphates

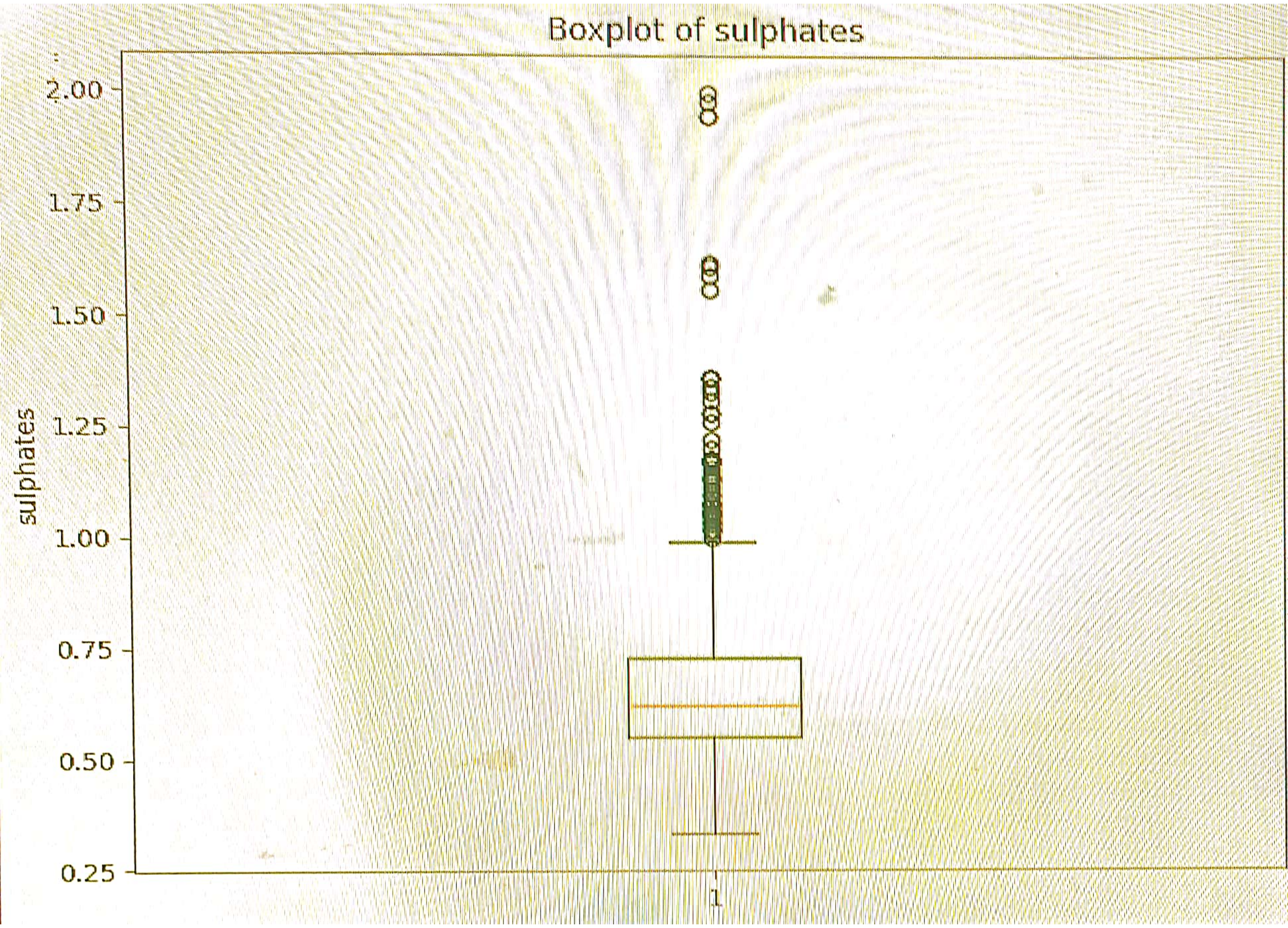
⟹ ∵ many wines have sulphate levels above the whiskers we can't consider them outliers.

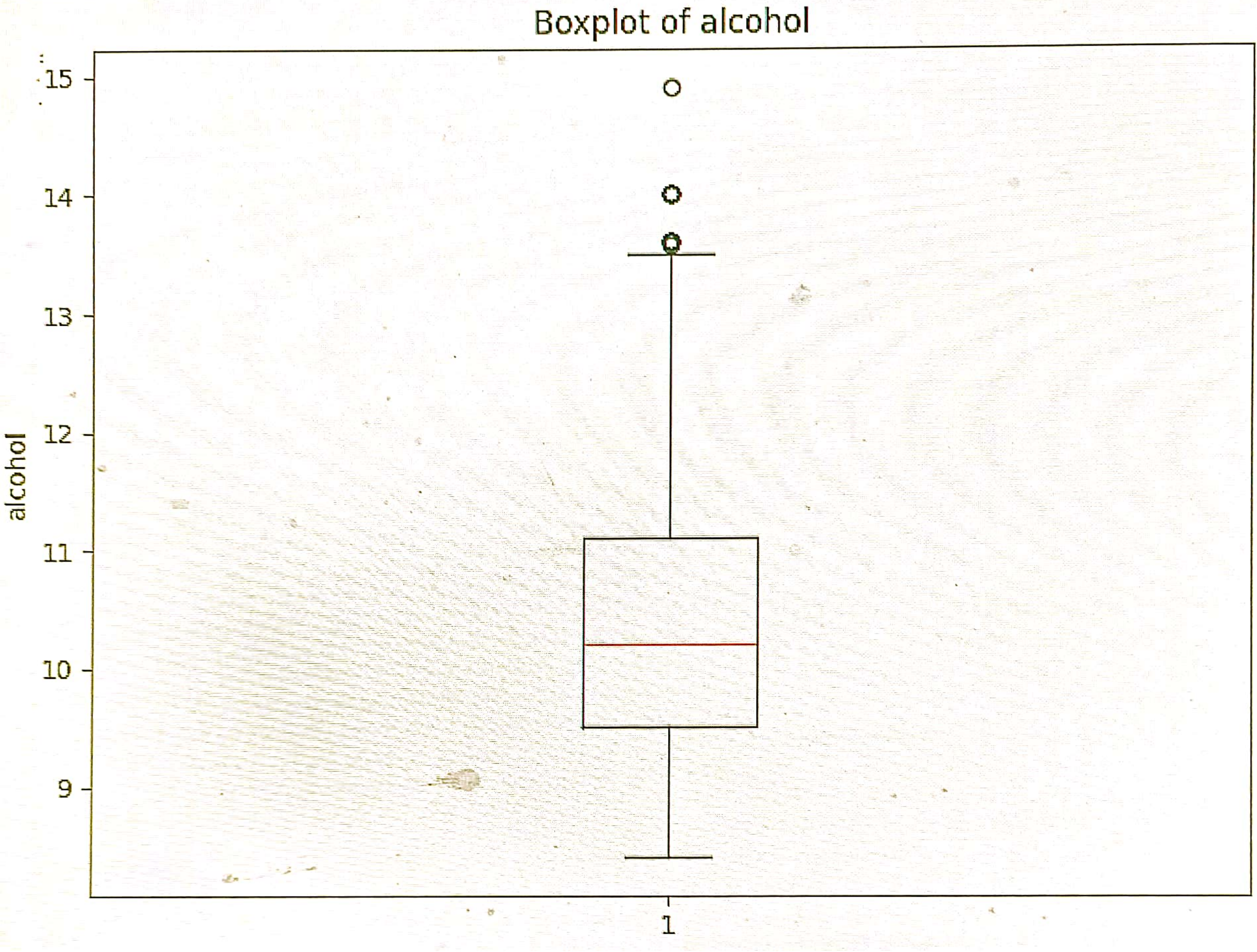
↳ alcohol

↳ by observing dataset we see high alcohol levels are truly impacting the quality so we won't drop them.


Boxplot of volatile acidity
volatile acidity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
1


Boxplot of citric acid
citric acid
1.0
0.8
0.6
0.4
0.2
0.0
1


Boxplot of sulphates
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
1


Boxplot of alcohol
alcohol
15
14
13
12
11
10
9
1

Date : / /
Page :

## Correlation analysis:

- $\rho$(Quality, alcohol) = 0.48 is highest it shows moderately strong +ve correlation.
- $\rho$(Quality, sulphates) = 0.25 (shows weak +ve correlation).
- $\rho$(Quality, citric acid) = 0.23 (shows weak +ve correlation).
- $\rho$(Quality, volatile acidity) = −0.39 (shows −ve correlation).

## Model Selection

we will choose features with highest absolute value of correlation for model selection
so,

- y (target variable) = Quality

- $x_1$ (alcohol)
- $x_2$ (Sulphate)
- $x_3$ (Citric acid)
- $x_4$ (Volatile acidity)

## Model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

## Loss function :

$$L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

## Initialize parameters:

$\beta_0 = 0$ , $\beta_1 = 0$ , $\beta_2 = 0$, $\beta_3 = 0$, $\beta_4 = 0$

$\alpha$ (Learning rate) $= 0.01$

## Partial derivatives

1. $$\frac{\partial L}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^{n} \left( y^{(i)} - \hat{y}^{(i)} \right)$$

2. $$\frac{\partial L}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^{n} x_1^{(i)} \left( y^{(i)} - \hat{y}^{(i)} \right)$$

3. $$\frac{\partial L}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^{n} x_2^{(i)} \left( y^{(i)} - \hat{y}^{(i)} \right)$$

4. $$\frac{\partial L}{\partial \beta_2} = -\frac{2}{n} \sum_{i=1}^{n} x_2^{(i)} \left( y^{(i)} - \hat{y}^{(i)} \right)$$

5. $$\frac{\partial L}{\partial \beta_3} = -\frac{2}{n} \sum_{i=1}^{n} x_3^{(i)} \left( y^{(i)} - \hat{y}^{(i)} \right)$$

6. $$\frac{\partial L}{\partial \beta_4} = -\frac{2}{n} \sum_{i=1}^{n} x_4^{(i)} \left( y^{(i)} - \hat{y}^{(i)} \right)$$

then update each parameter

$$\beta_j^{(t+1)} = \beta_j^{(t)} - \alpha \frac{\partial L}{\partial \beta_j}$$

## Example

| y (Quality) | $X_1$ (Alcohol) | $X_2$ (Sulphates) | $X_3$ (Citric Acid) | $X_4$ |
|---|---|---|---|---|
| 6.0 | 9.4 | 0.56 | 0.31 | 0.28 |
| 5.0 | 9.8 | 0.68 | 0.35 | 0.31 |
| 5.0 | 9.8 | 0.65 | 0.33 | 0.29 |

$\beta_0 = 0, \; \beta_1 = 0, \; \beta_2 = 0, \; \beta_3 = 0, \; \beta_4 = 0$

$\alpha = 0.01$

## First iteration

$\because$ all coefficients are zero,

$\hat{y} = 0$

2. Errors : $(y - \hat{y})$

Row (1) = 6
Row (2) = 5
Row (3) = 5

3. Compute Gradients:

$$\frac{\partial L}{\partial \beta_0} = -\frac{2}{3}(6+5+5) = -10.67$$

$$\frac{\partial L}{\partial \beta_1} = \frac{-2}{3}(9.4\times 6+9.8\times 5+9.8\times 5) = -102.93$$

$$\frac{\partial L}{\partial \beta_2} = \frac{-2}{3}(0.56\times 6+0.68\times 5+0.65\times 5) = -6.67$$

$$\frac{\partial L}{\partial \beta_3} = \frac{-2}{3}(0.31\times 6+0.35\times 5+0.33\times 5) = -3.507$$

$$\frac{\partial L}{\partial \beta_4} = -3.12$$

4. Update parameters:

$$\beta_0 = \beta_0 - \alpha \frac{\partial L}{\partial \beta_0} = 0-0.01\times(-10.67) = 0.1067$$

$\beta_1 = 1.0293$

$\beta_2 = 0.0667$

$\beta_3 = 0.03507$

$\beta_4 = 0.0312.$

$\beta_1$ has the largest $\therefore$ alcohol has greatest effect so far.

Ranking the features

$$\text{Importance}_j = \beta_j = \begin{bmatrix} \beta_1 \\ \beta_0 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

$$|\beta_0| = 0.1067$$
$$|\beta_1| = 1.0293$$
$$|\beta_2| = 0.0667$$
$$|\beta_3| = 0.03567$$
$$|\beta_4| = 0.0312$$

so, we will select alcohol for top 1 feature.

X—————o—————X

$X_1$ : Alcohol
$X_2$ : sulphates

## Model training -

$$y = \beta_0 + \beta_1(\text{Alcohol}) + \beta_2(\text{Sulphates})$$

**Initialize parameters:**

$$\beta_0 = 0, \quad \beta_1 = 0, \quad \beta_2 = 0$$

**Loss function:**

$$L(\beta_0, \beta_1, \beta_2) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

**Example dataset**

| y | $X_1$ | $X_2$ |
|---|---|---|
| 6 | 9.4 | 0.56 |
| 5 | 9.8 | 0.68 |
| 5 | 9.8 | 0.65 |

$n = 3$

$\alpha = 0.01$

$$e^{(i)} = y^{(i)} - \hat{y}^{(i)}$$

- $$\frac{\partial L}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^{n} e^{(i)}$$

- $$\frac{\partial L}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^{n} x_1^{(i)} e^{(i)}$$

- $\frac{\partial L}{\partial \beta_2} = \frac{-2}{n} \sum_{i=1}^{n} x_2^{(i)} e^{(i)}$

$$\beta_j \leftarrow \beta_j - \alpha \frac{\partial L}{\partial \beta_j}$$

<u>First iteration</u>

$e^{(i)} = 6-0,\ 5-0,\ 5-0$
$= 6, 5, 5$

- <u>Gradients</u>

1. $\frac{\partial L}{\partial \beta_0} = \frac{-2}{3} \sum e^{(i)} = \frac{-2}{3}(6+5+5) = \frac{-2}{3} \times 16$

$= -10.67$

2. $\frac{\partial L}{\partial \beta_1} = \frac{-2}{3} \sum x_1^{(i)} e^{(i)} = \frac{-2}{3}(9.4\times 6 + 9.8\times 5 + 9.8\times 5)$
$= \frac{-2}{3} \times 154.4 = -102.93$

3. $\frac{\partial L}{\partial \beta_2} = \frac{-2}{3} \sum x_2^{(i)} e^{(i)} = \frac{-2}{3} \times 10.01 = -6.67$

Date : / /
Page :

**Update Parameters:**

$\beta_0 = 0 - 0.01 \times (-10.67) = 0.1067$

$\beta_1 = 0 - 0.01 \times (-102.93) = 1.0293$

$\beta_2 = 0 - 0.01 \times (-6.67) = 0.0667$

**2nd Iteration**

$\beta_0^{(1)} = 0.1067, \quad \beta_1^{(1)} = 1.0293, \quad \beta_2^{(1)} = 0.0667$

**Predictions** $(\hat{y})$

- $\hat{y}_1^{(1)} = 0.1067 + 1.0293 \times 9.4 + 0.0667 \times 0.5$

$\hat{y}_1^{(1)} = 9.8185$

similarly,

- $\hat{y}_2^{(2)} = 10.2452$
- $\hat{y}_3^{(1)} = 10.2432$

errors $\begin{cases} e_1 = -3.81852 \\ e_2 = -5.2452 \\ e_3 = -5.2432 \end{cases}$

## Compute Gradients

- $\frac{\partial L}{\partial \beta_0} = \frac{-2}{3} \sum e_i = \frac{-2}{3}(-14.3069) = 9.538$

- $\frac{\partial L}{\partial \beta_1} = \frac{-2}{3} \sum x_{i,1} e_i = \frac{-2}{3}\big(9.4 \times (-3.8185) + 9.8 \times (-5.2452)$
  $+ 9.8 \times (-5.2432)\big)$

  $= 92.511$

- $\frac{\partial L}{\partial \beta_2} = 6.0754$

## Update Parameters

$\beta_0^{(2)} = \beta_0^{(1)} - \alpha \frac{\partial L}{\partial \beta_0} = 0.01132$

$\beta_1^{(2)} = \beta_1^{(1)} - \alpha \frac{\partial L}{\partial \beta_1} = 0.10419$

$\beta_2^{(2)} = \beta_2^{(1)} - \alpha \frac{\partial L}{\partial \beta_2} = 0.00595$

Interpretation

- intercept decreased from 0.1067 to 0.01132
- coefficient for alcohol ($\beta_1$) - drastic drop

$1.0293 \longrightarrow 0.1042$,

model now predicting smaller changes in quality from alcohol content.

- coefficient for sulphates ($\beta$) -

0.0667 to 0.00595

Model Evaluation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$\beta_0 = 0.01132$
$\beta_1 = 0.10412$
$\beta_2 = 0.00595$

$\hat{y_1} = 0.9940$

$e_1 = y_1 - \hat{y_1} = 6 - 0.9940 = 5.006$

$\hat{y_2} = \cancel{\beta}\ 1.0364$

$e_2 = y_2 - \hat{y_2} = 3.9636$

$\hat{y_3} = 1.03625$

$e_3 = y_3 - \hat{y_3} = 5 - 1.03625 = 3.96375$

Compute MAE -

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$

$n = 3$

$$MAE = \frac{|5.006| + |3.9636| + |3.96375|}{3}$$

$$= \cancel{12}.\ 4.31112$$

Date : / /
Page :

5. <u>RMSE</u> $= \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$

$e_1^2 = 25.06$
$e_2^2 = 15.71$
$e_3^2 = 15.71$

$RMSE = \sqrt{18.4933} \approx 4.3019$

for show fairly large errors but with more data the model should converge better.