# PROBABILITY AND STATISTICS PROJECT REPORT

### Explain

This document is the report of the project of probability and statistics course. The project is written in java programming language (only histogram and boxplot written in python) and it also includes a GUI.

The dataset I chose is red wine quality. There are columns with red wine-related properties in it. The column I chose is about the total sulfur dioxide of the wine. The reason I chose it is because it contains float numbers and calculations can be done easily with java programming language.

### Mean

It is found by summing all data and dividing it by the number of data.

### Median

If the number of data is odd after the data is sorted from small to large, the number in the middle of the data is the media. If the number of data is even, the average of the 2 elements in the middle is the media.

### Variance

It is found by subtracting each data from the mean, squaring it and dividing the number of data by one less.

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}.$$

### Standart Deviation

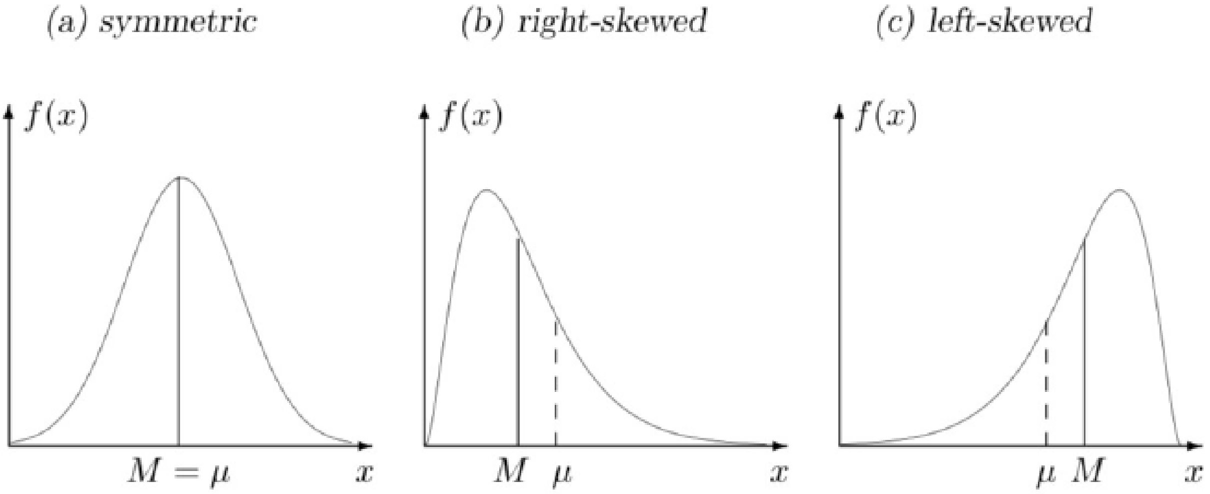
It is calculated by taking the square root of the variance.

**Standart Error**

It is calculated by finding the standard deviation to the square root of the number of data.

**Shape of distribution**

The mean of the data is 46 and the median is 38.Mean > median so right skewed.



(a) symmetric (b) right-skewed (c) left-skewed

FIGURE 8.2: A mean $\mu$ and a median $M$ for distributions of different shapes.

**Outliers**

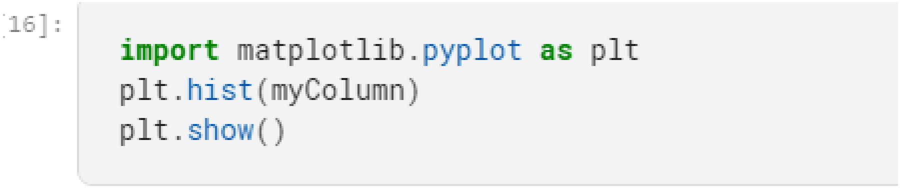
Q1 = First quartile.

Q3 = Third quartile.

Interquartile = Q3 – Q1

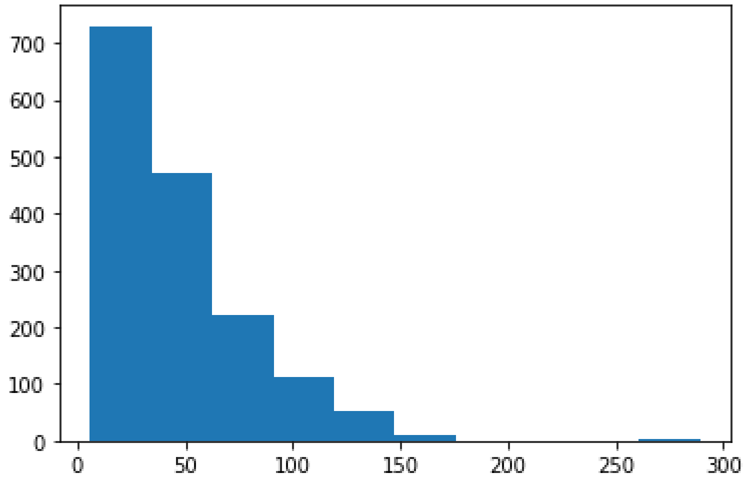
[ Q1 – 1.5 x Interquartile , Q3 + 1.5 x Interquartile ] => numbers outside this range are outliers.

## Histogram

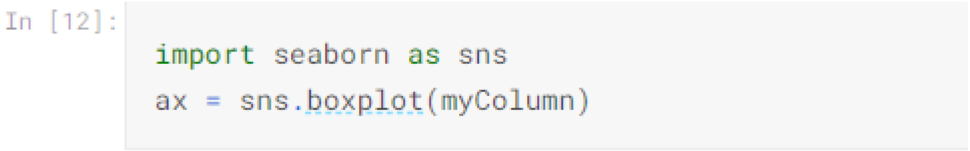


16]:

```
import matplotlib.pyplot as plt
plt.hist(myColumn)
plt.show()
```

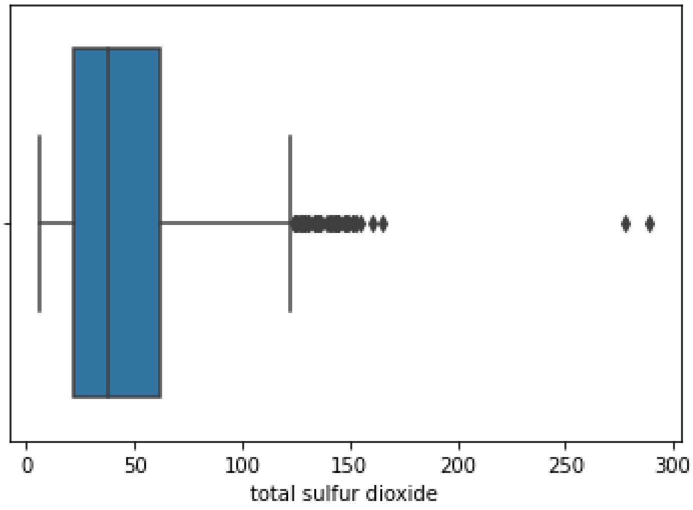


## Boxplot



In [12]:

```
import seaborn as sns
ax = sns.boxplot(myColumn)
```



## %95 Confidence Interval for the Mean and Variance

z alpha divided by 2 is calculated by reading the Z table. The e value is z alpha divided by the product of the standard deviation divided by the square root of the data number and the confidence interval is calculated using the formula below.

$$P\left(\overline{X}-\frac{z_{\alpha/2}\sigma}{\sqrt{n}}\leq\mu\leq\overline{X}+\frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right)=1-\alpha$$

## How large a sample with a margin at most 0.1 units with confidence 90%.

z is the square of the product of the alpha value and the standard deviation.

$$n\geq\left(\frac{z_{\alpha/2}\cdot\sigma}{\Delta}\right)^2$$