# Wine Project - Part A

Samarth Sathe

2023-09-19

## Red Wine

### Introduction

Through statistical analysis and visualization, we embark on a journey to decipher the secrets hidden within the wineglass, revealing insights that may help winemakers, enthusiasts, and sommeliers alike appreciate the art and science behind the world of wine.

```r
#Importing the required libraries
library(tidyverse)
library(here)
```

#### *Importing the dataset*

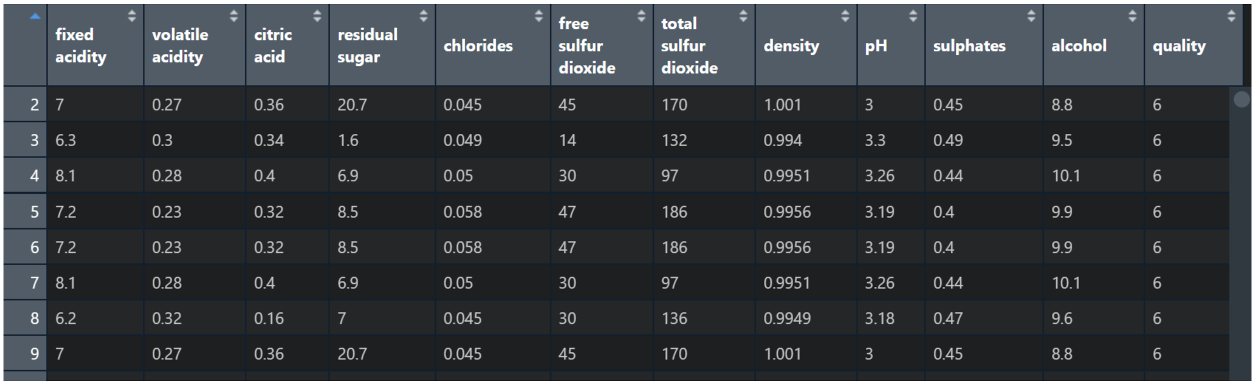
```r
#Setting the file path using "here" library for importing the csv file
file_path1 <- here("Data", "winequality-red.csv")

#reading the csv file using read.csv function and using "sep=;"
#to specific the delimiter in the csv file.
data_red_wine <- read.csv(file_path1, sep = ";", header = TRUE, stringsAsFactors = FALSE)
#The data provided has separation in semicolon instead of colon as in any normal
#dataset and hence we had to specify the seperator.

#To view the data
view(data_red_wine)

attach(data_red_wine)
```

#### *Snapshot of seperated Dataset -*



| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 7 | 0.27 | 0.36 | 20.7 | 0.045 | 45 | 170 | 1.001 | 3 | 0.45 | 8.8 | 6 |
| 3 | 6.3 | 0.3 | 0.34 | 1.6 | 0.049 | 14 | 132 | 0.994 | 3.3 | 0.49 | 9.5 | 6 |
| 4 | 8.1 | 0.28 | 0.4 | 6.9 | 0.05 | 30 | 97 | 0.9951 | 3.26 | 0.44 | 10.1 | 6 |
| 5 | 7.2 | 0.23 | 0.32 | 8.5 | 0.058 | 47 | 186 | 0.9956 | 3.19 | 0.4 | 9.9 | 6 |
| 6 | 7.2 | 0.23 | 0.32 | 8.5 | 0.058 | 47 | 186 | 0.9956 | 3.19 | 0.4 | 9.9 | 6 |
| 7 | 8.1 | 0.28 | 0.4 | 6.9 | 0.05 | 30 | 97 | 0.9951 | 3.26 | 0.44 | 10.1 | 6 |
| 8 | 6.2 | 0.32 | 0.16 | 7 | 0.045 | 30 | 136 | 0.9949 | 3.18 | 0.47 | 9.6 | 6 |
| 9 | 7 | 0.27 | 0.36 | 20.7 | 0.045 | 45 | 170 | 1.001 | 3 | 0.45 | 8.8 | 6 |

### Question 5a. Calculating Sample size

```
sample_size_red_wine <- nrow(data_red_wine)
cat("Sample size for red wine:", sample_size_red_wine, "\n")
```

```
## Sample size for red wine: 1599
```

1. ***Fixed.Acidity***

- We will first check the summary of fixed.acidity column

```
summary(fixed.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.60    7.10    7.90    8.32    9.20   15.90
```

### Question 5b. Any outliers? Do you have any concerns about the data quality?

Observing the summary of fixed.acidity we can see that the range between the 3rd Quartile and Max is greater, which tells us that 75% of the data is till 9.20 and rest 25% is till 15.90. This indicates that there could be outliers in the data.

```
sum(is.na(fixed.acidity))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data and the statistics shows a typical distribution. The mean and median are relatively close to each other which suggests the skewness is not extreme. Using this data we can suggest that the data does not seem to have any quality issues.

### Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(fixed.acidity)
```

```
## [1] 3.031416
```

```
sd(fixed.acidity)
```

```
## [1] 1.741096
```

```
mean(fixed.acidity)-2*sd(fixed.acidity)
```

```
## [1] 4.837445
```

```
mean(fixed.acidity)+2*sd(fixed.acidity)
```

```
## [1] 11.80183
```

```
sum((fixed.acidity>4.8 & fixed.acidity<11.8)==TRUE)/1599
```

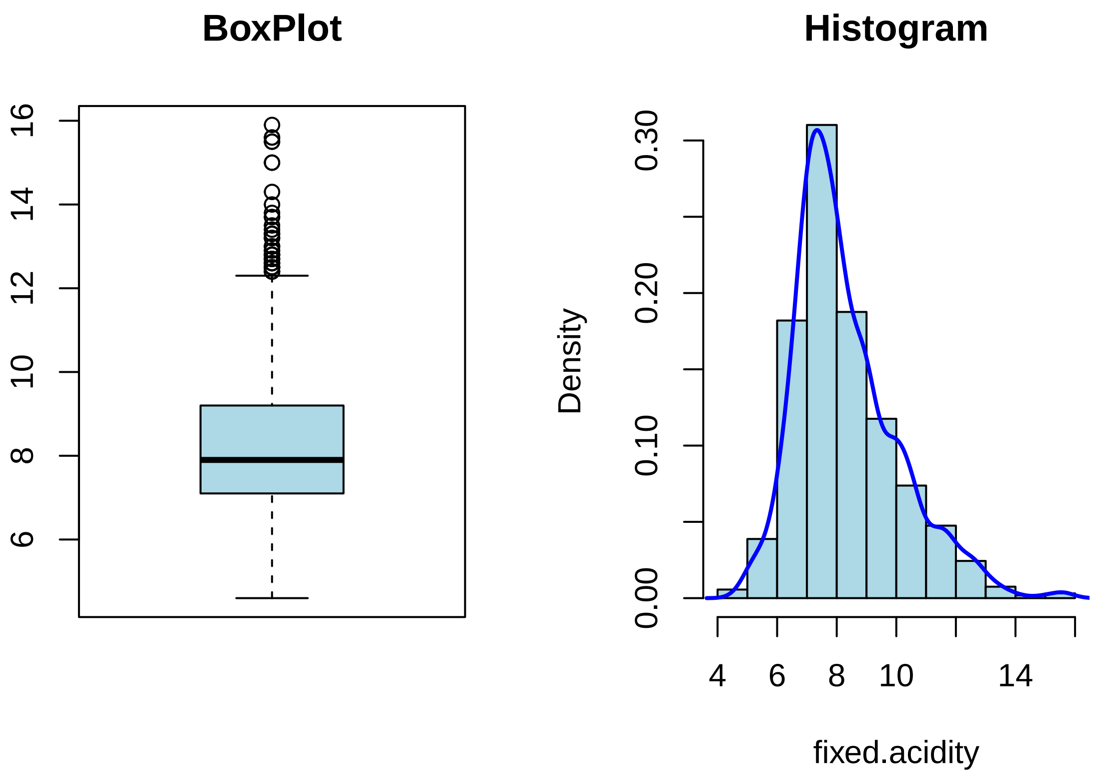
```
## [1] 0.9480926
```

Using the formula above we can see that the spread of data is ~94%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(fixed.acidity, main = "BoxPlot",col = "lightblue")
hist(fixed.acidity, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(fixed.acidity), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Looking at the histogram above we can see that the data is skewed to the right.

2. ***Volatile.Acidity***

- We will first check the summary of volatile.acidity column

```
summary(volatile.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of volatile.acidity we can see that the range between the 3rd Quartile and Max is greater, which tells us that 75% of the data is till 0.64 and rest 25% is till 1.58. This indicates that there could be outliers in the data.

```
sum(is.na(volatile.acidity))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data and the statistics shows a typical distribution. The mean and median are almost same to each other which suggests the skewness is not extreme. Using this data we can suggest that the data does not seem to have any quality issues.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(volatile.acidity)
```

```
## [1] 0.03206238
```

```
sd(volatile.acidity)
```

```
## [1] 0.1790597
```

```
mean(volatile.acidity)-2*sd(volatile.acidity)
```

```
## [1] 0.1697011
```

```
mean(volatile.acidity)+2*sd(volatile.acidity)
```

```
## [1] 0.8859399
```

```
sum((volatile.acidity>0.16 & volatile.acidity<0.88)==TRUE)/1599
```

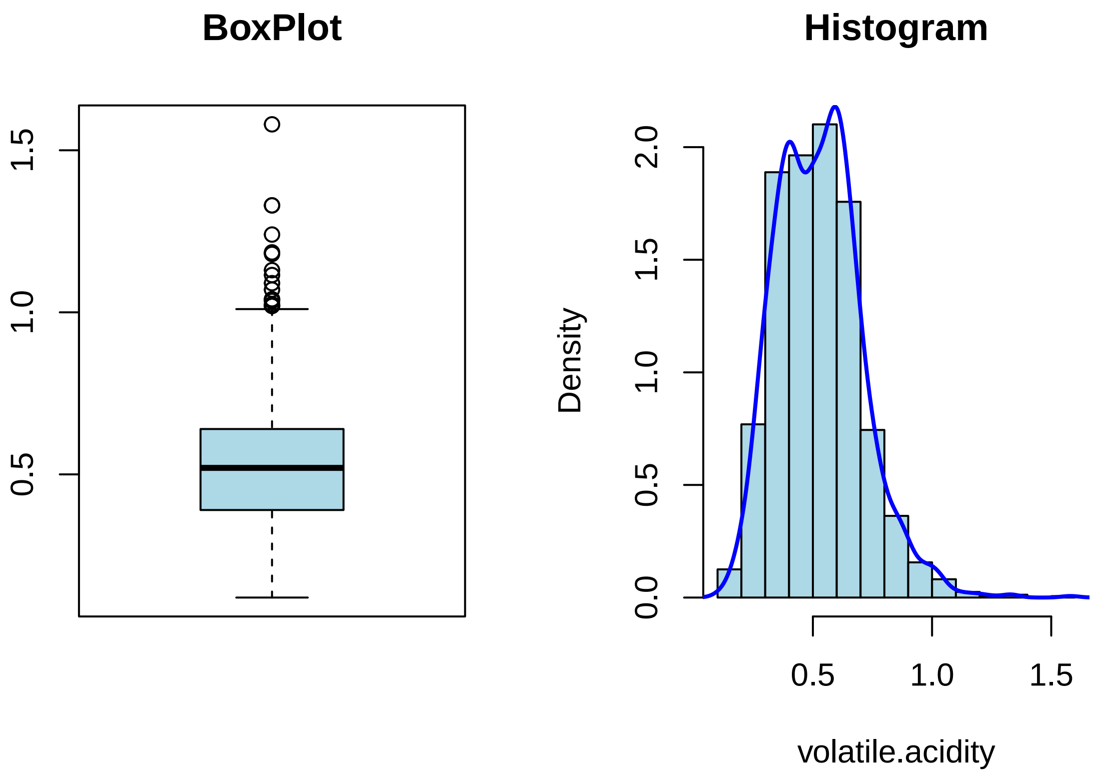
```
## [1] 0.9587242
```

Using the formula above we can see that the spread of data is ~96%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```r
par(mfrow = c(1, 2))
boxplot(volatile.acidity, main = "BoxPlot",col="lightblue")
hist(volatile.acidity, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(volatile.acidity), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Looking at the histogram above we can see that the data is slightly skewed to the right.

3. ***Citric.Acid***

- We will first check the summary of citric.acid column

```r
summary(citric.acid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of citric.acid we can see that the range between the 3rd Quartile and Max is on higher side, which tells us that 75% of the data is till 0.42 and rest 25% is till 1. This indicates that there could be outliers in the data.

```
sum(is.na(citric.acid))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data and the statistics shows a typical distribution. The mean and median are almost same to each other which suggests the skewness is not extreme. Using this data we can suggest that the data does not seem to have any quality issues.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(citric.acid)
```

```
## [1] 0.03794748
```

```
sd(citric.acid)
```

```
## [1] 0.1948011
```

```
mean(citric.acid)-2*sd(citric.acid)
```

```
## [1] -0.1186267
```

```
mean(citric.acid)+2*sd(citric.acid)
```

```
## [1] 0.6605779
```

```
sum((citric.acid>-0.11 & citric.acid<0.66)==TRUE)/1599
```

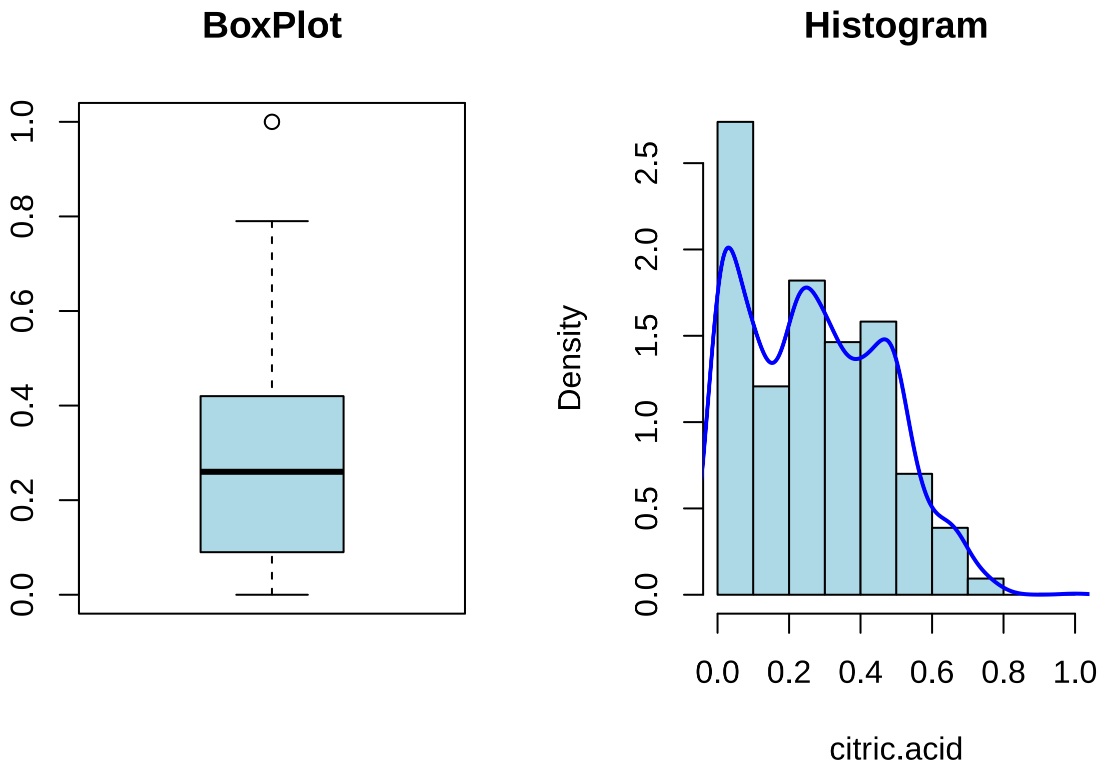
```
## [1] 0.9693558
```

Using the formula above we can see that the spread of data is ~97%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(citric.acid, main = "BoxPlot",col="lightblue")
hist(citric.acid, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(citric.acid), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Here mean>median and by looking at the histogram above we can see that the data is skewed to the right.

4. ***Residual.sugar***

- We will first check the summary of residual.sugar column

```
summary(residual.sugar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.900   1.900   2.200   2.539   2.600  15.500
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of residual.sugar we can see that the range between min and 1st quartile is high and the range for 3rd quartile and max is significantly greater which tells that the 75% of data is till 2.6 and 25% is till 15.5.Also, the mean and median are not close to each other. These variations tells us that there are outliers in the data.

```
sum(is.na(residual.sugar))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data however, the outliers in the data could be a concerning data quality issue.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```r
var(residual.sugar)
```

```
## [1] 1.987897
```

```r
sd(residual.sugar)
```

```
## [1] 1.409928
```

```r
mean(residual.sugar)-2*sd(residual.sugar)
```

```
## [1] -0.2810506
```

```r
mean(residual.sugar)+2*sd(residual.sugar)
```

```
## [1] 5.358662
```

```r
sum((residual.sugar>-0.28 & residual.sugar<5.35)==TRUE)/1599
```

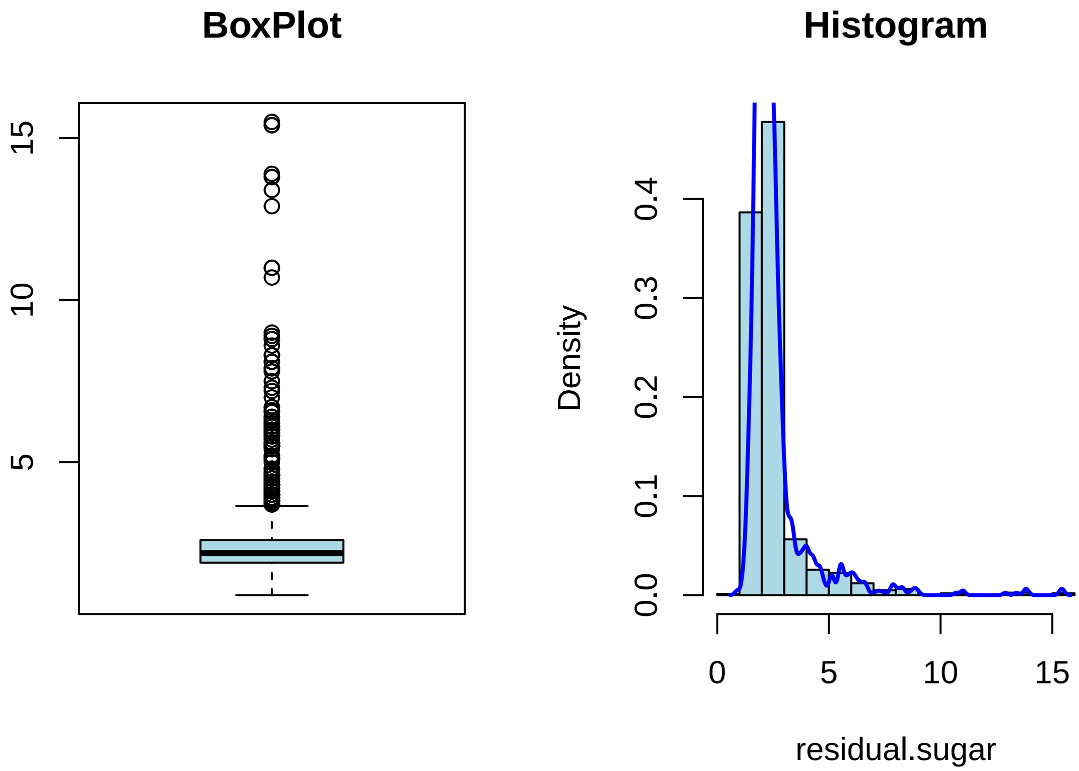
```
## [1] 0.9530957
```

Using the formula above we can see that the spread of data is ~95%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```r
par(mfrow = c(1, 2))
boxplot(residual.sugar, main = "BoxPlot",col="lightblue")
hist(residual.sugar, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(residual.sugar), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Here mean>median and also by looking at the histogram above we can see that the data is skewed to the right.

5. ***Chlorides***

- We will first check the summary of chlorides column

```
summary(chlorides)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of chlorides we can see that the range between min 1st quartile, 3rd quartile and max is significantly greater which suggests that 75% of the data is till 0.0900 and 25% of the data is till 0.611. These variations tells us that there are outliers in the data.

```
sum(is.na(chlorides))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data however, the outliers in the data could be a concerning data quality issue.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(chlorides)
```

```
## [1] 0.002215143
```

```
sd(chlorides)
```

```
## [1] 0.0470653
```

```
mean(chlorides)-2*sd(chlorides)
```

```
## [1] -0.006664062
```

```
mean(chlorides)+2*sd(chlorides)
```

```
## [1] 0.1815971
```

```
sum((chlorides>-0.006 & chlorides<0.18)==TRUE)/1599
```

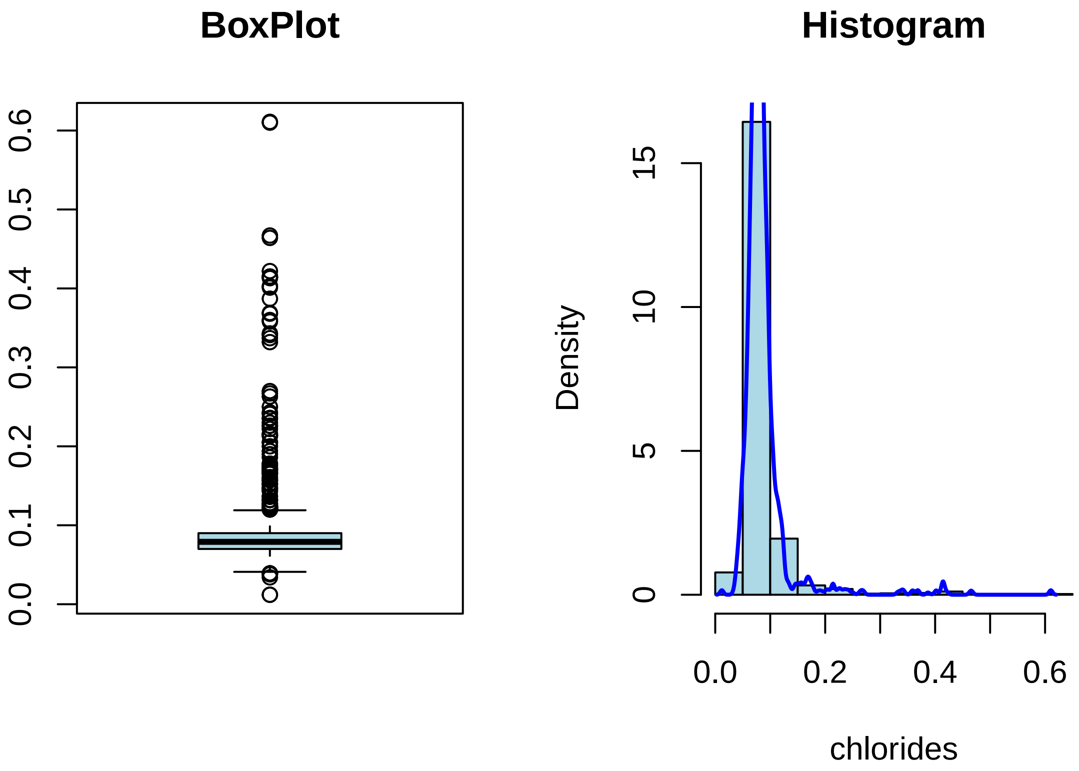
```
## [1] 0.9718574
```

Using the formula above we can see that the spread of data is ~97%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(chlorides, main = "BoxPlot",col="lightblue")
hist(chlorides, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(chlorides), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Here mean>median and also by looking at the histogram above we can see that the data is skewed to the right.

6. ***free.sulfur.dioxide***

- We will first check the summary of free.sulfur.dioxide column

```
summary(free.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    7.00   14.00   15.87   21.00   72.00
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of free.sulfur.dioxide we can see that the range between min 1st quartile, 3rd quartile and max is significantly greater. This points to variations in the data provided which points that there are outliers in the data.

```
sum(is.na(free.sulfur.dioxide))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data however, the outliers in the data could be a concerning data quality issue.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```r
var(free.sulfur.dioxide)
```

```
## [1] 109.4149
```

```r
sd(free.sulfur.dioxide)
```

```
## [1] 10.46016
```

```r
mean(free.sulfur.dioxide)-2*sd(free.sulfur.dioxide)
```

```
## [1] -5.045392
```

```r
mean(free.sulfur.dioxide)+2*sd(free.sulfur.dioxide)
```

```
## [1] 36.79524
```

```r
sum((free.sulfur.dioxide>-5.04 & free.sulfur.dioxide<36.79)==TRUE)/1599
```

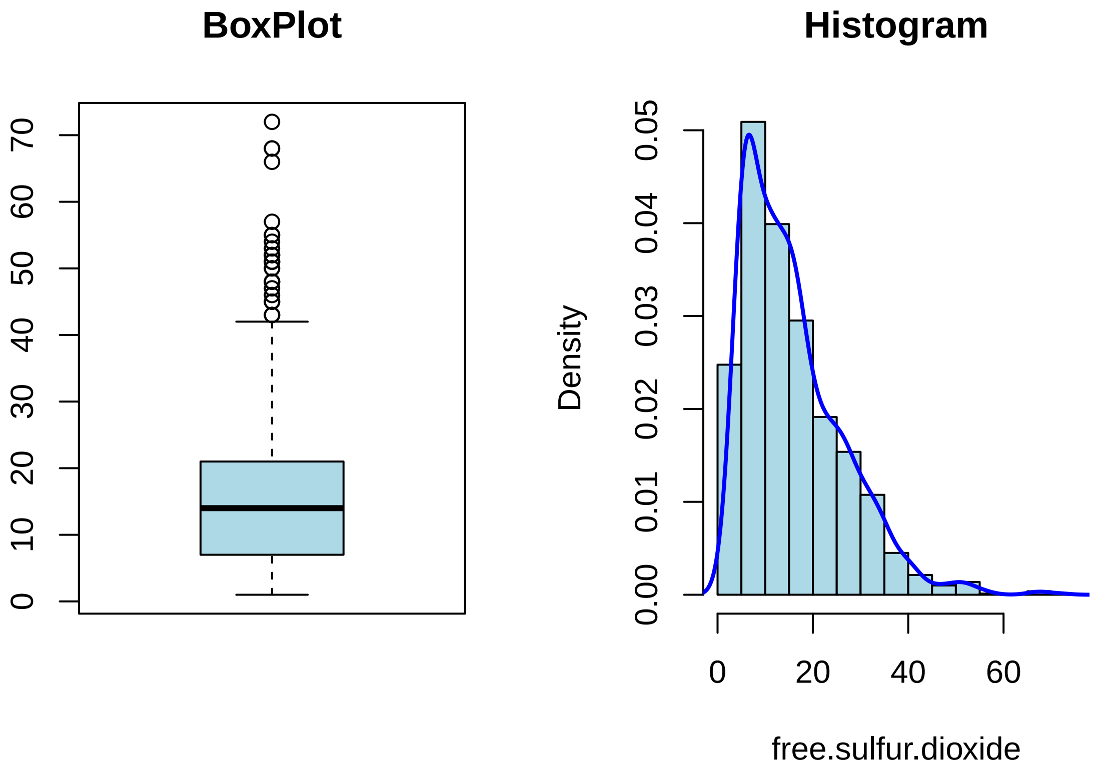
```
## [1] 0.9587242
```

Using the formula above we can see that the spread of data is ~96%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```r
par(mfrow = c(1, 2))
boxplot(free.sulfur.dioxide, main = "BoxPlot",col="lightblue")
hist(free.sulfur.dioxide, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(free.sulfur.dioxide), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Here Mean>Median and also by looking at the histogram above we can see that the data is skewed to the right.

7. ***total.sulfur.dioxide***

- We will first check the summary of total.sulfur.dioxide column

```
summary(total.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   22.00   38.00   46.47   62.00  289.00
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of total.sulfur.dioxide we can see that the range between min 1st quartile, 3rd quartile and max is significantly greater also, the mean and median values differ by a lot which tells us that there are many outliers in the data.

```
sum(is.na(total.sulfur.dioxide))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data but, the difference in median and median values is high and also, there are many outliers in the data which points to concerning data quality issues.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(total.sulfur.dioxide)
```

```
## [1] 1082.102
```

```
sd(total.sulfur.dioxide)
```

```
## [1] 32.89532
```

```
mean(total.sulfur.dioxide)-2*sd(total.sulfur.dioxide)
```

```
## [1] -19.32286
```

```
mean(total.sulfur.dioxide)+2*sd(total.sulfur.dioxide)
```

```
## [1] 112.2584
```

```
sum((total.sulfur.dioxide>-19.32 & total.sulfur.dioxide<112.25)==TRUE)/1599
```

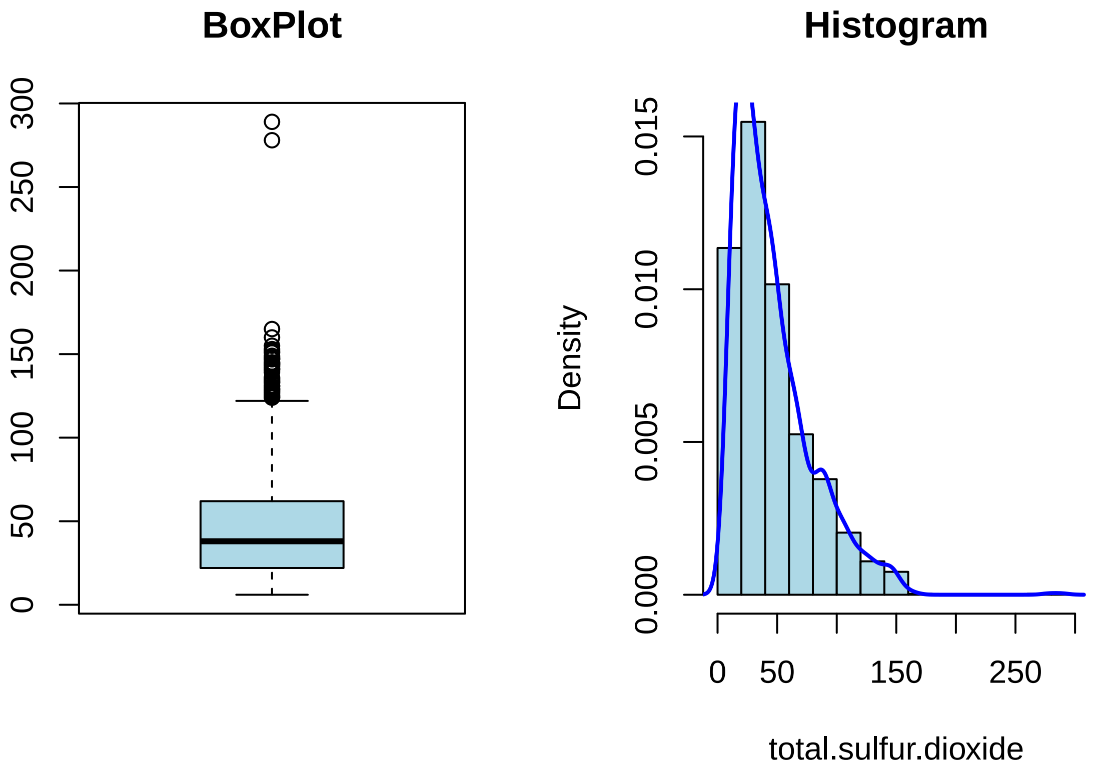
```
## [1] 0.9499687
```

Using the formula above we can see that the spread of data is ~95%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(total.sulfur.dioxide, main = "BoxPlot",col="lightblue")
hist(total.sulfur.dioxide, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(total.sulfur.dioxide), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Here, mean>median and also by looking at the histogram above we can see that the data is skewed to the right.

8. ***density***

- We will first check the summary of density column

```
summary(density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

The summary statistics of density suggests that there are no outliers in the dataset as the max and min values are close to the quartiles.

```
sum(is.na(density))
```

```
## [1] 0
```

**Data Quality concerns-** There are no immediate data quality concerns, the mean and median are very close to each other and there are no outliers in the data. The data is consistent.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(density)
```

```
## [1] 3.562029e-06
```

```
sd(density)
```

```
## [1] 0.001887334
```

```
mean(density)-2*sd(density)
```

```
## [1] 0.992972
```

```
mean(density)+2*sd(density)
```

```
## [1] 1.000521
```

```
sum((density>0.99 & density<1)==TRUE)/1599
```

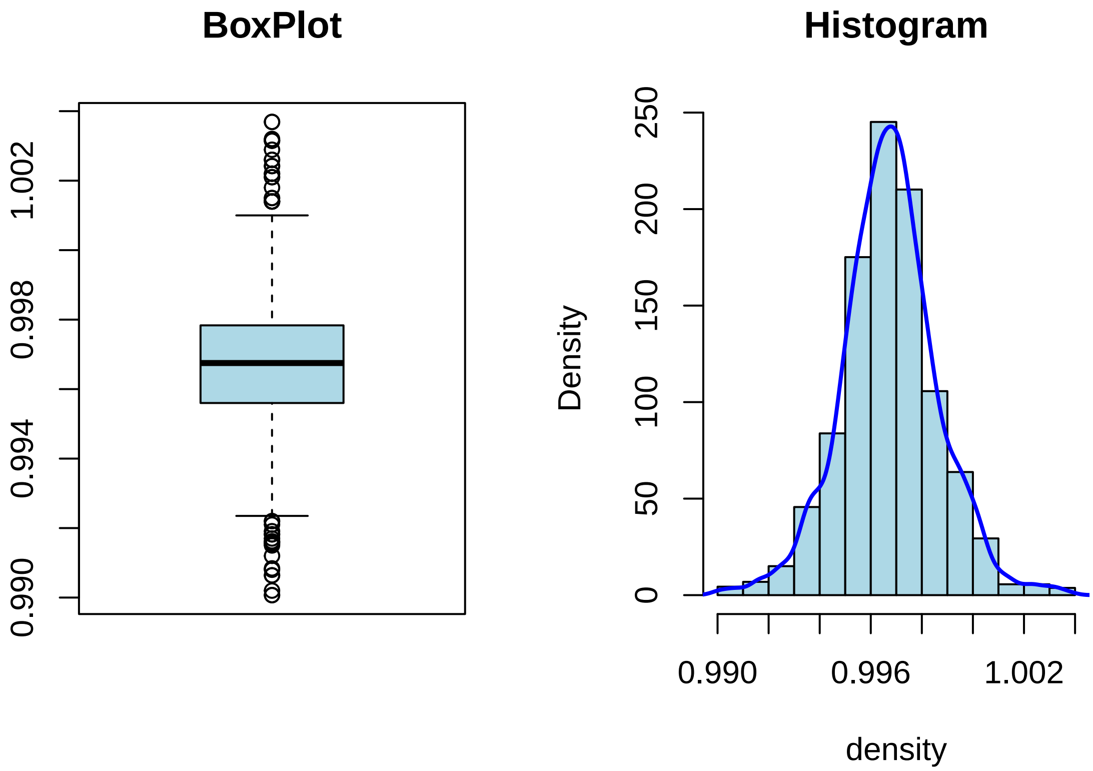
```
## [1] 0.9493433
```

Using the formula above we can see that the spread of data is ~95%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(density, main = "BoxPlot",col="lightblue")
hist(density, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(density), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

The mean and median are very close to each other which indicates that there is no skewness.

9. ***pH***

- We will first check the summary of pH column

```
summary(pH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.740   3.210   3.310   3.311   3.400   4.010
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of pH we can see that the range between min and 1st Quartile and the 3rd Quartile and Max is greater, which tells us that there are a few outliers in the data.

```
sum(is.na(pH))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data and the statistics shows a typical distribution. The mean and median are almost similar to each other which suggests the skewness is not extreme and distribution is approximately symmetric. Using this data we can suggest that the data does not seem to have any concerning quality issues.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(pH)
```

```
## [1] 0.02383518
```

```
sd(pH)
```

```
## [1] 0.1543865
```

```
mean(pH)-2*sd(pH)
```

```
## [1] 3.00234
```

```
mean(pH)+2*sd(pH)
```

```
## [1] 3.619886
```

```
sum((pH>3.0 & pH<3.6)==TRUE)/1599
```

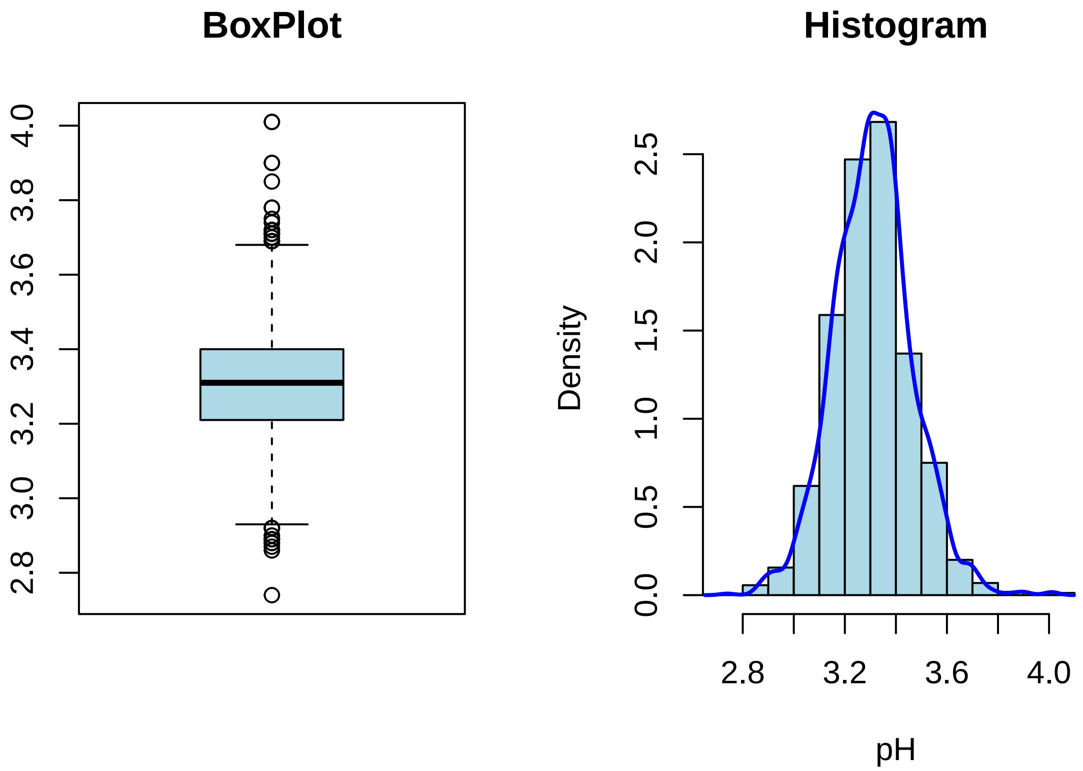
```
## [1] 0.9437148
```

Using the formula above we can see that the spread of data is ~94%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(pH, main = "BoxPlot",col="lightblue")
hist(pH, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(pH), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Looking at the histogram above we can see that the data is slightly skewed to the right.

10. ***sulphates***

- We will first check the summary of sulphates column

```
summary(sulphates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

Observing the summary of sulphates we can see that the range between the 3rd Quartile and Max is greater, which tells us that 75% of the data is till 0.73 and rest 25% is till 2. This indicates that are many outliers in the data.

```
sum(is.na(sulphates))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data and the statistics shows a typical distribution. As the range of 3rd Qua. and max is on the greater side which suggests that there are outliers and this could cause data quality concerns.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(sulphates)
```

```
## [1] 0.02873262
```

```
sd(sulphates)
```

```
## [1] 0.169507
```

```
mean(sulphates)-2*sd(sulphates)
```

```
## [1] 0.3191349
```

```
mean(sulphates)+2*sd(sulphates)
```

```
## [1] 0.9971628
```

```
sum((sulphates>0.31 & sulphates<0.99)==TRUE)/1599
```

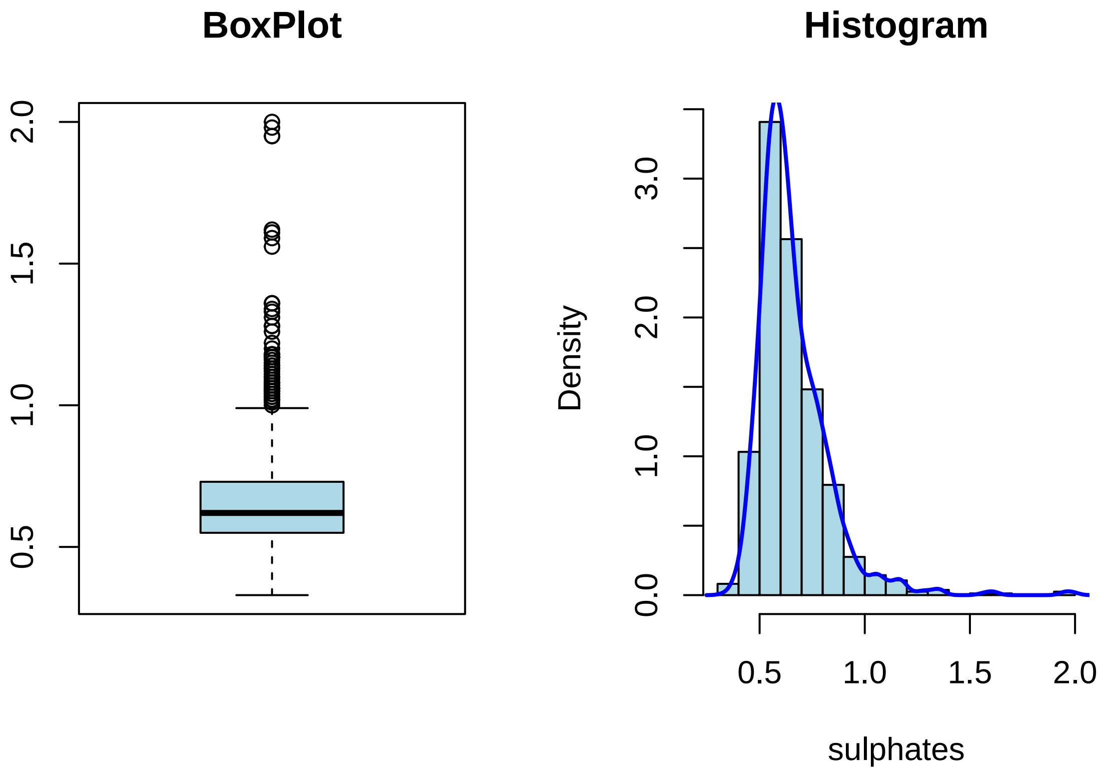
```
## [1] 0.9612258
```

Using the formula above we can see that the spread of data is ~96%

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(sulphates, main = "BoxPlot",col="lightblue")
hist(sulphates, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(sulphates), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

Here mean>median and also by looking at the histogram above we can see that the data is skewed to the right.

11. ***alcohol***

- We will first check the summary of alcohol column

```
summary(alcohol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.40    9.50   10.20   10.42   11.10   14.90
```

**Question 5b. Any outliers? Do you have any concerns about the data quality?**

By observing the summary of the data, we can see that the difference in the 3rd Quartile and max. is greater which suggests that 75% of the data is till 11.10 and rest 25% data is till 14.90. This tells us that there are a few outliers in the data.

```
sum(is.na(alcohol))
```

```
## [1] 0
```

**Data quality concerns -** There are no missing values in the data and the statistics shows a typical distribution. The mean and median are almost close to each other which suggests the skewness is not extreme. Using this data we can suggest that the data does not seem to have any quality issues.

**Question 5c. How can you summarize the data of each variable in a concise way? What statistics are you going to present?**

Using the summary function we were able to get the min,1st quadrant,median, mean, 3rd quadrant and max values. Additional, we can calculate the variance, standard deviation.

```
var(alcohol)
```

```
## [1] 1.135647
```

```
sd(alcohol)
```

```
## [1] 1.065668
```

```
mean(alcohol)-2*sd(alcohol)
```

```
## [1] 8.291648
```

```
mean(alcohol)+2*sd(alcohol)
```

```
## [1] 12.55432
```

```
sum((alcohol>8.29 & alcohol<12.55)==TRUE)/1599
```

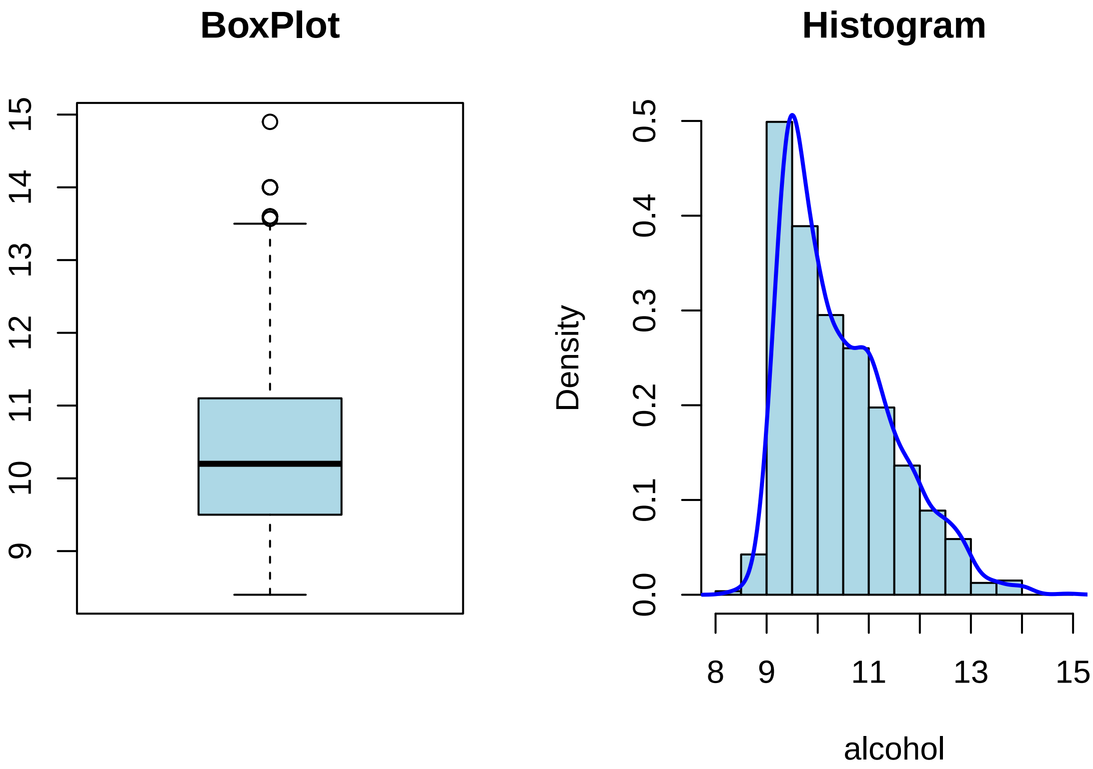
```
## [1] 0.9562226
```

Using the formula above we can see that the spread of data is 95.6% ~96%.

**Question 5d. How can you visualize the distribution of each variable?**

We can use boxplot and histogram to visualize the distribution.

```
par(mfrow = c(1, 2))
boxplot(alcohol, main = "BoxPlot",col="lightblue")
hist(alcohol, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(alcohol), lwd=2, col='blue')
```



**Question 5e. Do you see any skewed distributions?**

The mean is greater than median and also by looking at the histogram above we can see that the data is skewed to the right.

12. ***quality***

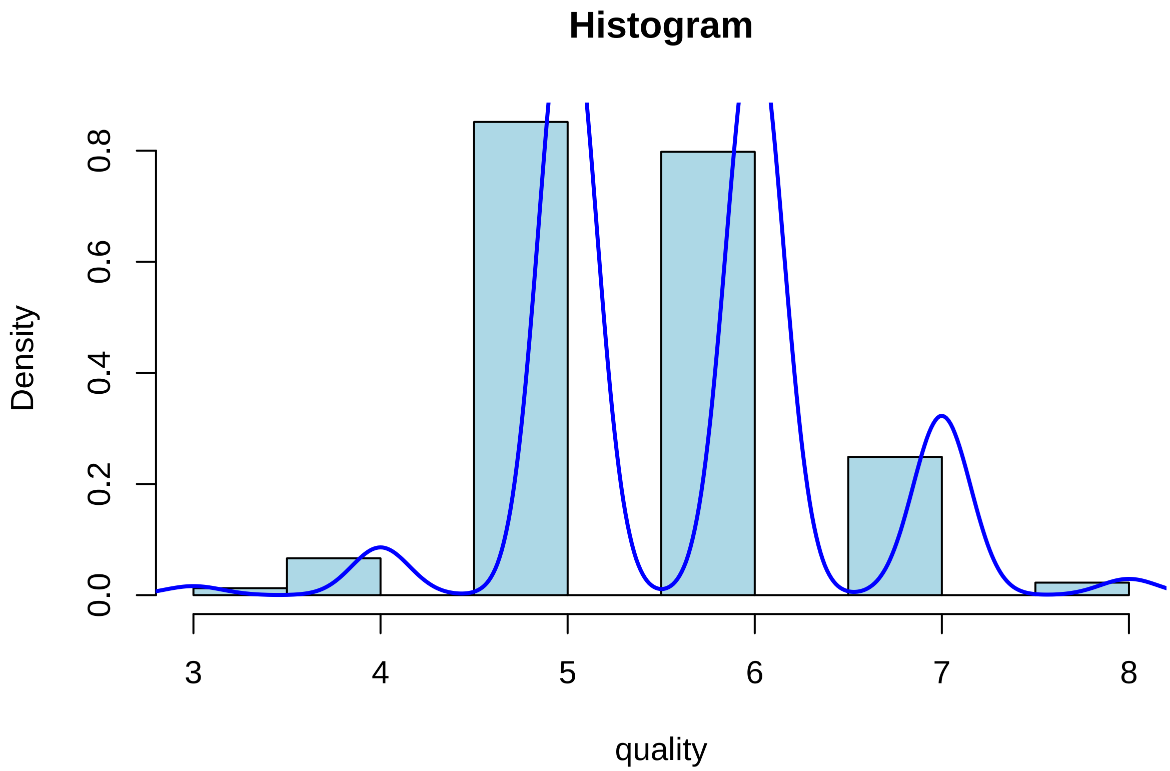
- We will first check the summary of quality column

```
summary(quality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.000   5.000   6.000   5.636   6.000   8.000
```

Quality is the target variable whose value depends on the other physiochemical variables in the dataset. The summary function will provide us with the basic statistics. We can visualize it to calculate the mode of the data.

```
hist(quality, freq = FALSE, main = "Histogram",col="lightblue")
lines(density(quality), lwd=2, col='blue')
```



The histogram tells us that the mode of the quality column is 5, which tells us that maximum of the wine quality lies in the range of 5.