BANA7051 Assignment 1

Ang Zhang

2024-08-30

A. Sample size of the data.

first load the data, and then take the sample size with nrow() function.

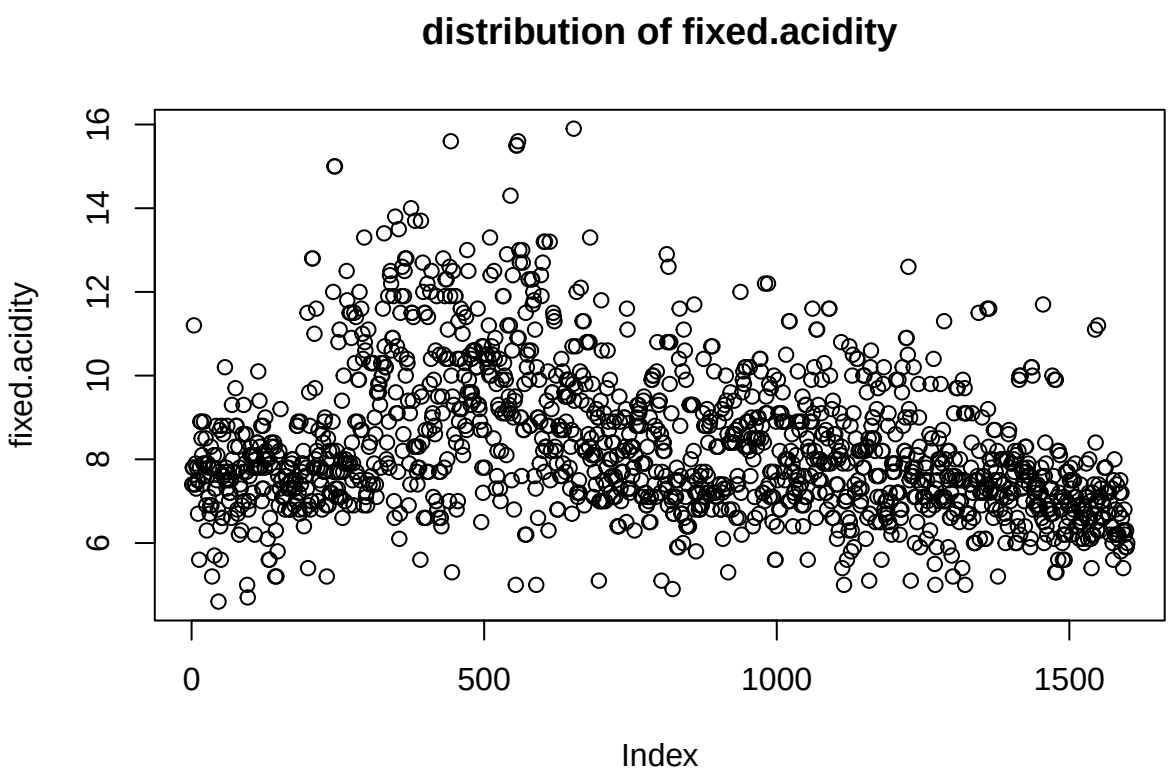
```
wine <- read.csv("data/winequality-red.csv", sep = ";")
wine <- select(wine, c("fixed.acidity", "volatile.acidity", "citric.acid"))
sample_size <- nrow(wine)
print(paste('sample size is ', sample_size))
```

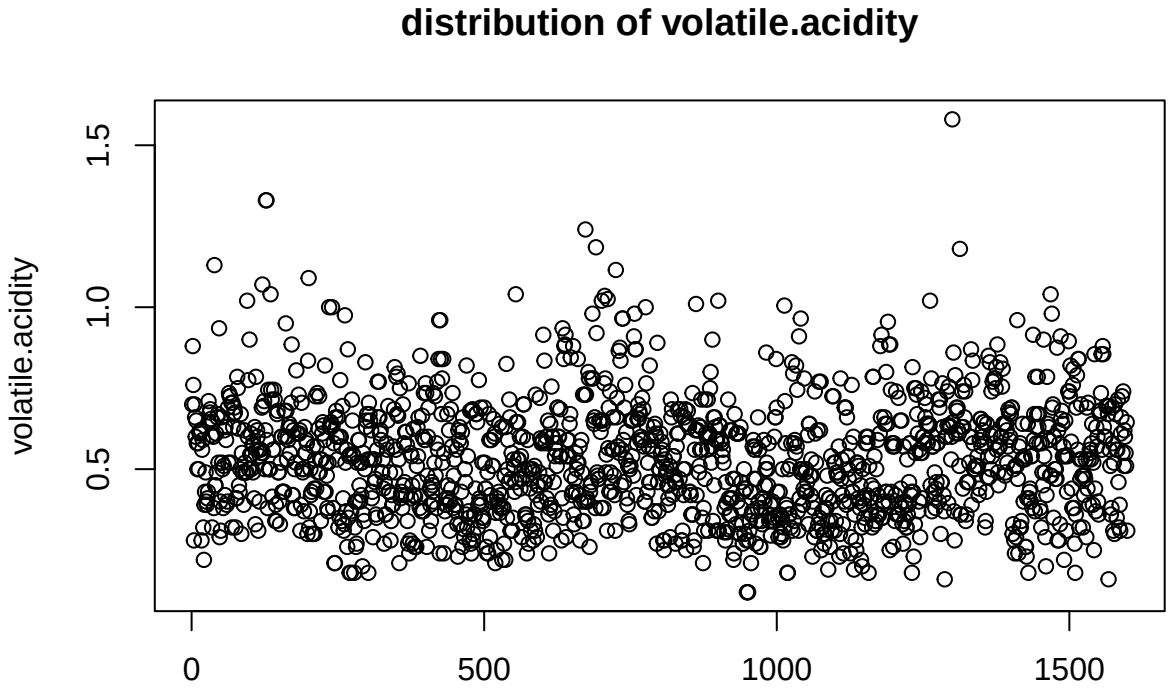
```
## [1] "sample size is  1599"
```

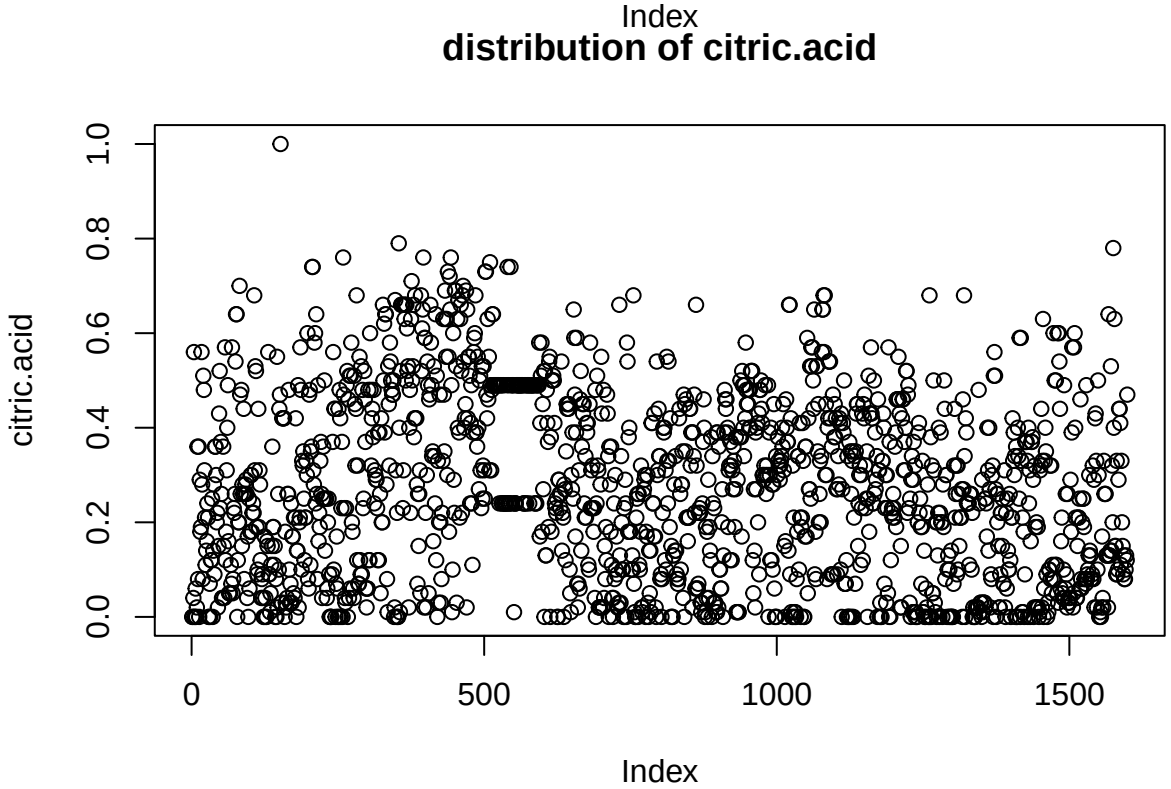
B. Identify outliers

Draw a plot for each of the variables:

```
for (col_name in colnames(wine))
  plot(wine[[col_name]], main = paste("distribution of", col_name), ylab = col_name)
```







One outlier can be observed in the citric.acid variable.

C. Summarize of data.

The summary() function provides a basic summary of Min, 1st Quantile, median, third quntile, max:

```
summary(wine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000
```

Moreover, I would like to include standard deviation to give a little bit more insight:

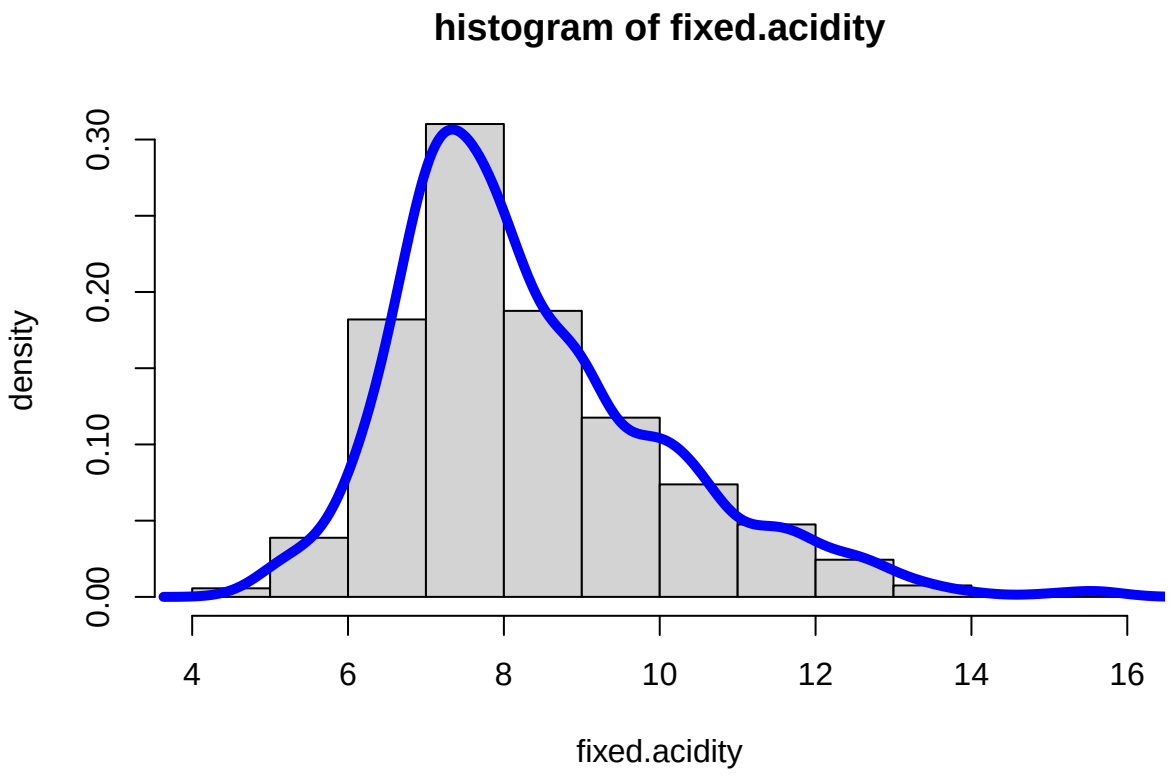
```
for (col_name in colnames(wine)) {
  sd = sd(wine[[col_name]])
  print(paste("standard deviation of ", col_name, ": ", round(sd, 2)))
}
```

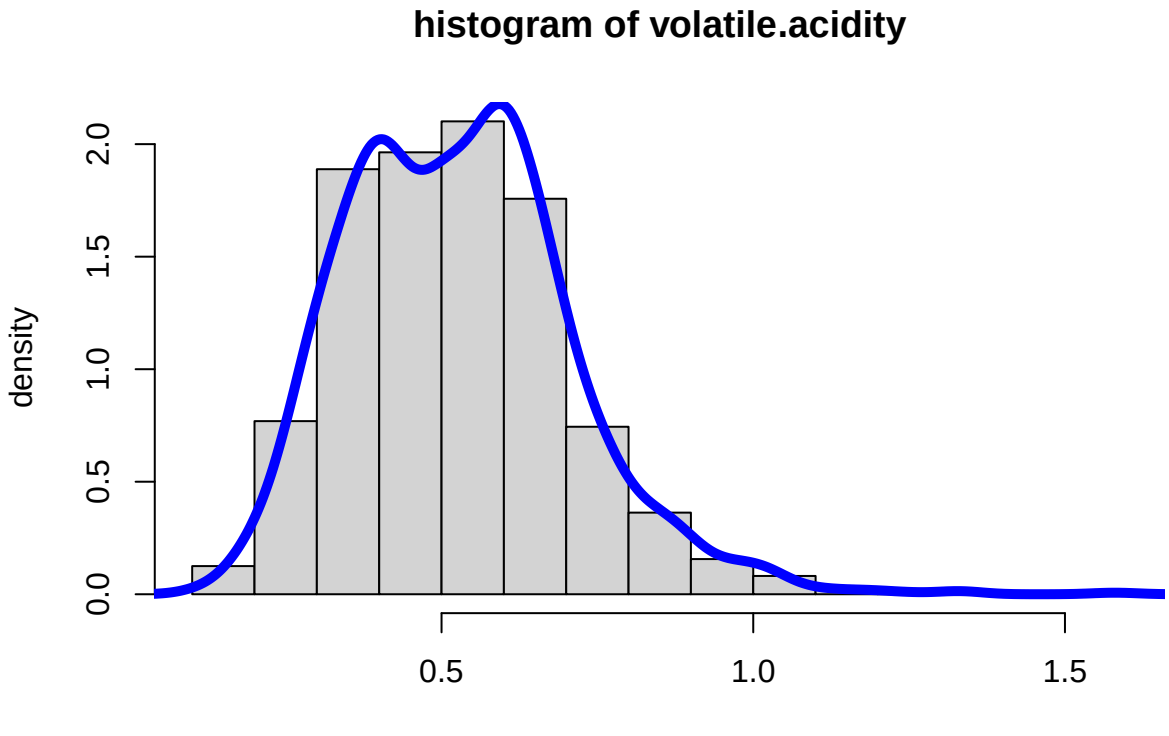
```
## [1] "standard deviation of  fixed.acidity :  1.74"
## [1] "standard deviation of  volatile.acidity :  0.18"
## [1] "standard deviation of  citric.acid :  0.19"
```

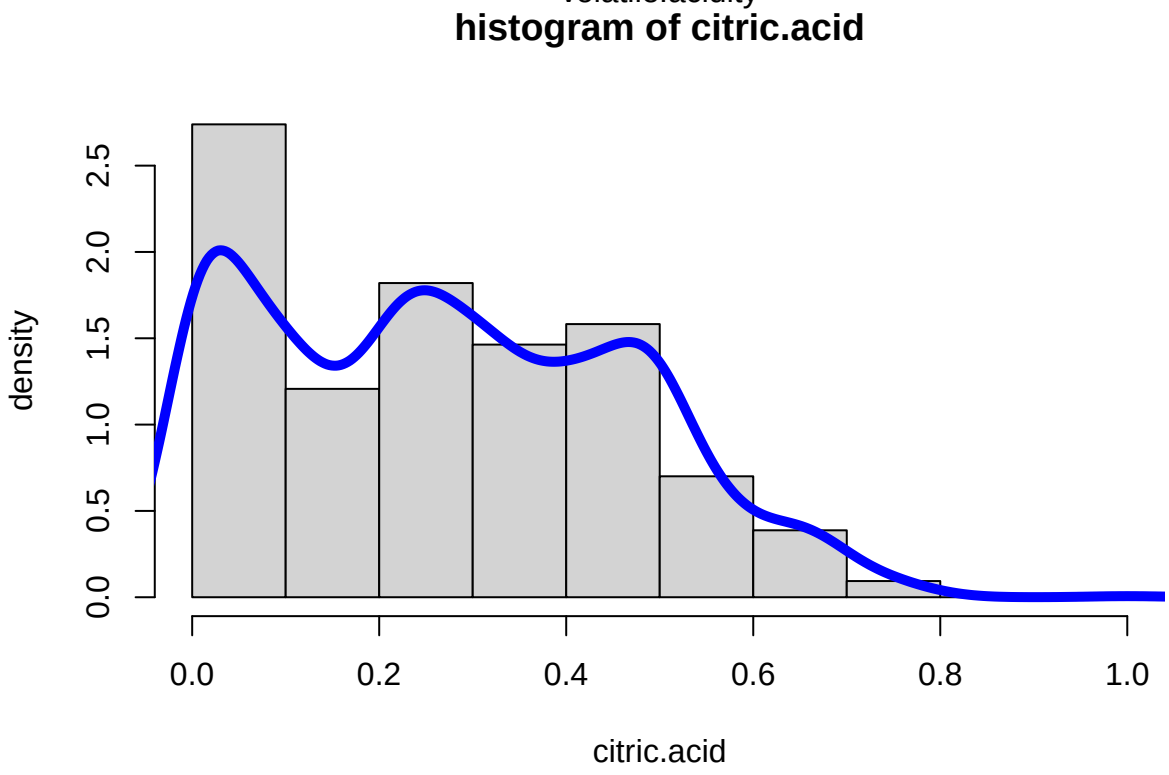
D. Visualize the distribution of each variable.

Draw a histogram of each variable with hist() function, and draw a density curve on top of it.

```
for (col_name in colnames(wine)) {
  hist(wine[[col_name]], main = paste("histogram of", col_name), freq = F, xlab = col_name, ylab = "dens
  lines(density(wine[[col_name]]), lwd = 5, col = "blue")
}
```







E. Any skewed distribution in D?

The fixed.acidity variable appears to be right skewed. So does the citric.acid variable.

F. What data mining methods are used in this paper?

The author discussed linear/multiple regression (MR), neural networks (NN), and support vector machines (SVM). MR can be seen as a reduced form of NN when there's no layer of hidden node. Empirical results shows that SVM outperformed NN (and also MR) in this study case, especially for white wine.