Tipología de Datos

Alejandro Torres, Cristina Rodríguez

20 de mayo de 2019

Table of Contents

—-load_libraries, include=FALSE————————————————

```
if(!require(knitr)){install.packages("knitr")}
```

```
## Loading required package: knitr
```

```
if(!require(lubridate)){install.packages("lubridate")}
```

```
## Loading required package: lubridate
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
if(!require(VIM)){install.packages("VIM")}
```

```
## Loading required package: VIM
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at:
https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep

if(!require(stringr)){install.packages("stringr")}

## Loading required package: stringr

if(!require(psych)){install.packages("psych")}

## Loading required package: psych

if(!require(pROC)){install.packages("pROC")}

## Loading required package: pROC

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:colorspace':
##
##     coords

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

if(!require(dplyr)){install.packages("dploYr")}

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
if(!require(ggplot2)){install.packages("ggplot2")}
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
if(!require(nortest)){install.packages("nortest")}
```

```
## Loading required package: nortest
```

```r
if(!require(Kendall)){install.packages("Kendall")}
```

```
## Loading required package: Kendall
```

```r
if(!require(BSDA)){install.packages("BSDA")}
```

```
## Loading required package: BSDA
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

1. Descripción del dataset. Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos objeto de análisis se ha obtenido a partir de este enlace en Kaggle y está constituido por 12 características (columnas) que presentan 1599 muestras de vino portigués (filas o registros). Las variables que se disponen son fisicoquímicas (entradas) y sensoriales (la salida) (por ejemplo, no hay datos sobre tipos de uva, marca de vino, precio de venta del vino, etc.).

Entre los campos de este conjunto de datos, encontramos los siguientes:

fixed.acidity: Acidez fija, la mayoría de los ácidos relacionados con el vino. Fijos o no volátiles (no se evaporan fácilmente).

volatile.acidity: Acidez volátil, la cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre.

citric.acid: El ácido cítrico se encuentra en pequeñas cantidades, el ácido cítrico puede agregar 'frescura' y sabor a los vinos.

residual.sugar: azúcar residual, la cantidad de azúcar restante después de que se detiene la fermentación, es raro encontrar vinos con menos de 1 gramo / litro y vinos con más de 45 gramos / litro se consideran dulces.

chlorides: Cloruro, la cantidad de sal en el vino.

free.sulfur.dioxide: sin dioxido de azufre, la forma sin SO2 existe en equilibrio entre el SO2 molecular (como un gas disuelto) y el ión bisulfito; Previene el crecimiento microbiano y la oxidación del vino.

total.sulfur.dioxide: Cantidad total de dióxido de azufre de formas libres y unidas de SO2; en bajas concentraciones, el SO2 es mayormente indetectable en el vino, pero a concentraciones de SO2 libres de más de 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino.

density: La densidad del agua es cercana a la del agua dependiendo del porcentaje de alcohol y de contenido de azúcar.

pH: El pH describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); La mayoría de los vinos están entre 3-4 en la escala de pH.

sulphates: Aditivo de vino sulphatesa que puede contribuir a los niveles de dióxido de azufre (S02), que actúa como un antimicrobiano y antioxidante.

alcohol: alcohol el porcentaje de alcohol del vino .

quality: Variable de calidad de salida (basada en datos sensoriales, puntuación entre 0 y 10).

En esta práctica unificaremos los datos, los limpiaremos y trataremos de estipueda predecir cuáles serán los de moejor calidad. Se compararán las características de los vinos calificados como de calidad media/baja y de calidad alta, para tratar de establecer las diferencias en cuanto a cualidades de los mismos.

2. Integración y selección de los datos de interés a analizar.

fixed.acidity: Si a la acidez total le quitamos la acidez volátil debida al acético, la diferencia se llama acidez fija. Con lo que si hacemos la operación a la inversa podemos obtener la acidez total.

volatile.acidity: La acidez volátil puede oscilar entre 0,2 - 1 gr/L hasta un gramo por litro, una larga crianza en roble pueden situarse alrededor de 0,8 gr/L de acidez volátil sin manifestar sensaciones desagradables.

citric.acid: Este ácido desaparece lentamente al ser fermentado por las bacterias. No es muy abundante en la uva.

residual.sugar: Se considera vino seco o sin azúcar, cuando ese contenido es inferior a cinco gramos por litro. En función de esta cantidad de azúcares residuales, el vino puede ser: seco, semiseco, semidulce y dulce. Variando de menor a mayor cantidad de azúcar,

pudiendo ir desde 1 gramo hasta 200 gramos por litro de azúcar. En los vinos secos en muchas ocasiones no suele sobrepasarse los 2 gramos.

chlorides: 1 a 50 mg/L: potasio (agente diurético, efecto que se potencia en los vinos espumosos debido al alto contenido en dióxido de carbono), sodio, hierro, manganeso, boro, nitratos, cloruros y silicio

free.sulfur.dioxide :

total.sulfur.dioxide: En Europa su valor máximo no suele sobrepasar los 50 miligramos por litro pero en E.E.U.U. puede alcanzar los 350 miligramos por litro con lo que se puede saber si este vino es de buena calidad y se puede exportar.

density:

- Vino blanco seco: 0,9880-0,9930 g/mL.
- Vinos tinto seco: 0,9910-0,9950 g/mL.
- Vino espumoso: 0,9890-1,0080 g/mL.
- Vino de licor (moscatel): 1,0500-1,0700 g/mL.
- Mosto: 1,0590-1,1150 g/mL.

pH: El pH de la mayoría de los vinos se encuentra en el intervalo de 2,5 a 4,5 , lo que lógicamente recae en el lado ácido de la escala. Un vino con un pH de 2,8 es extremadamente ácido mientras que uno con un pH en torno a 4 es plano, carente de acidez. Los vinos blancos suelen estar entre 3 y 3,3 y la mayoría de los tintos entre 3,3 y 3,6

sulphates: Cualquier vino que contiene más de 10 mg por litro de dióxido de azufre deberá indicar en su etiqueta la expresión “Contiene sulfitos”

alcohol: La graduación alcohólica que oscila entre los 3,5 y los 15 grados

quality:

```
library(knitr)
library(lubridate)
library(VIM)
library(stringr)
library(psych)
library(pROC)
library(dplyr)

#myfile <- "C:\\Users\\prosy\\Documents\\UOC\\Master Ciencia de
#datos\\Tipolog?a y ciclo de vida de los datos\\Practica #02\\winequality-
red.csv"

practica02 <-read.csv("winequality-red.csv", header = TRUE,sep=",")
length(practica02)
```

```
## [1] 12

str(practica02)

## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58
0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069
0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36
3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...

#En este resumen inicial
summary(practica02)

##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
```

```
##  3rd Qu.:6.000
##  Max.   :8.000
```

```
head(practica02)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
# Esta variable no hace falta, no se esta usando para nada
#practica02$total.acidity<-
practica02$fixed.acidity+practica02$volatile.acidity
```

```
# read data
res <- sapply(practica02,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
fixed.acidity	numeric
volatile.acidity	numeric
citric.acid	numeric
residual.sugar	numeric
chlorides	numeric
free.sulfur.dioxide	numeric
total.sulfur.dioxide	numeric
density	numeric
pH	numeric

sulphates	numeric
alcohol	numeric
quality	integer
****Eliminar****	
Se crea una variable nu	eva llamada total.acidity (acidez total que es la suma de la acidez fija+acidez volatil).

Creamos una variable para quality en forma categórica donde los vinos calificados con una nota de 3 a 4 serán de calidad baja, de 4 a 6 calidad normal y de 6 a 8 calidad alta.

```
practica02[12] <- lapply(practica02[12], as.numeric)
#res <- sapply(practica02,class)
#kable(data.frame(variables=names(res),clase=as.vector(res)))
practica02$quality_categoric <- cut(practica02$quality, breaks=c(3, 5, 8),
labels=c("Baja/Media", "Alta"))

# Separo los datos en subsets para poder comparar las características de los
vinos de calidad baja/media con los de calidad alta

baja <- subset(practica02, quality_categoric == "Baja/Media")
alta <- subset(practica02, quality_categoric == "Alta")
```

3. Limpieza de los datos.

3.1. Los datos contienen ceros o elementos vacíos? Cómo gestionarías cada uno de estos casos?

Se comprueba si hay valores NA y ese caso se sustituyen por un campo vacío. En este dataset ha resultado no haber valores NA ni vacíos

```
colSums(is.na(practica02))

##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
##    quality_categoric
##                   10

colSums(practica02=="")

##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
```

```
##                    0                   0                    0
## total.sulfur.dioxide             density                   pH
##                    0                   0                    0
##            sulphates             alcohol              quality
##                    0                   0                    0
##    quality_categoric
##                   NA
```

```
colSums(practica02==0)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                   0                  132
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                   0                    0
## total.sulfur.dioxide             density                   pH
##                    0                   0                    0
##            sulphates             alcohol              quality
##                    0                   0                    0
##    quality_categoric
##                   NA
```

```
# Imputar los valores a partir de los k-vecinos más cercanos usando la
distancia de Gower con la
# información de todas las variables

practica02$quality_categoric[is.na(practica02$quality_categoric)] <-
c("Alta")
# Hay valores NA en la variable quality_categoric, así que las sustituyo y
compruebo
colSums(is.na(practica02))
```

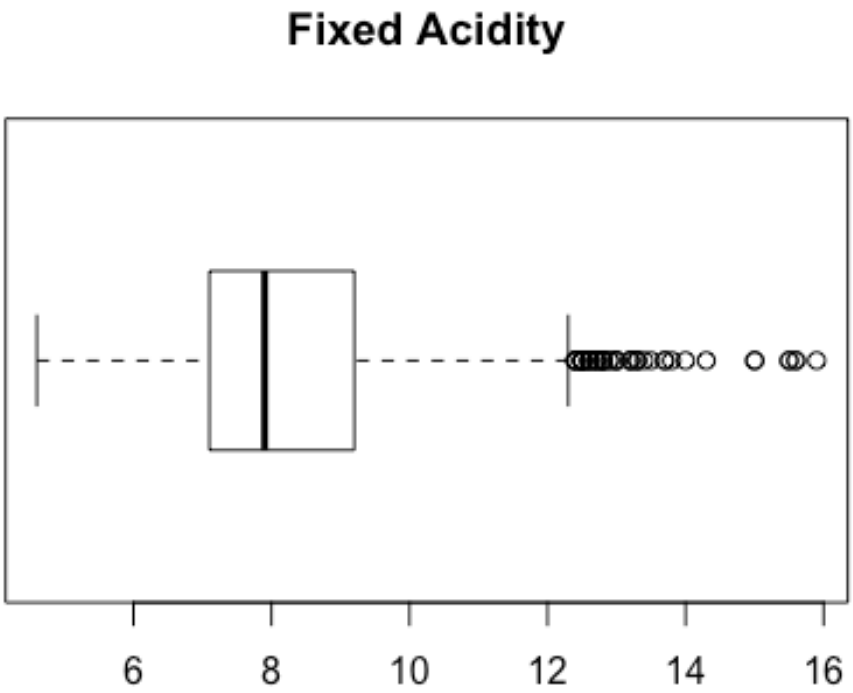
```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                   0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                   0                    0
## total.sulfur.dioxide             density                   pH
##                    0                   0                    0
##            sulphates             alcohol              quality
##                    0                   0                    0
##    quality_categoric
##                    0
```

Se comprueba si hay valores que en lugar de tener valores vacios o NA tuvieran valor 0.

También se comprueba si hay valores que en lugar de nulos se hubiesen puesto con valor 0, se ve que el ácido cítrico posee 132 registros con valor 0, pero este es su valor real de acidez por lo que es correcto.

3.2. Identificación y tratamiento de valores extremos.

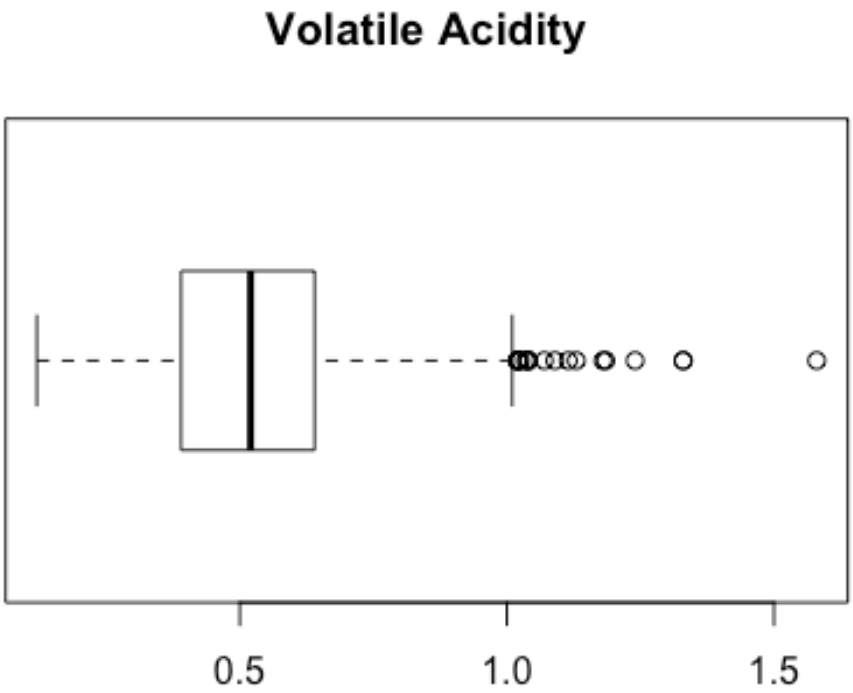
```
wine.fixed<-boxplot(practica02$fixed.acidity,horizontal = TRUE,main="Fixed
Acidity")
```




```
wine.fixed$out
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8
## [15] 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4
## [29] 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2
## [43] 13.2 13.2 15.9 13.3 12.9 12.6 12.6
```

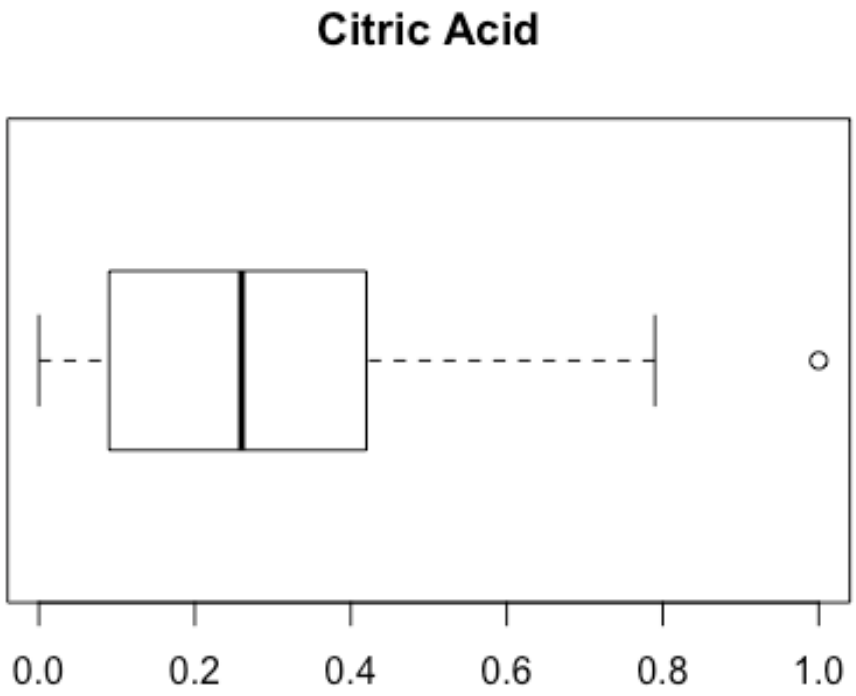
```
wine.volatile<-boxplot(practica02$volatile.acidity,horizontal =
TRUE,main="Volatile Acidity")
```




```
wine.volatile$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
## [12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

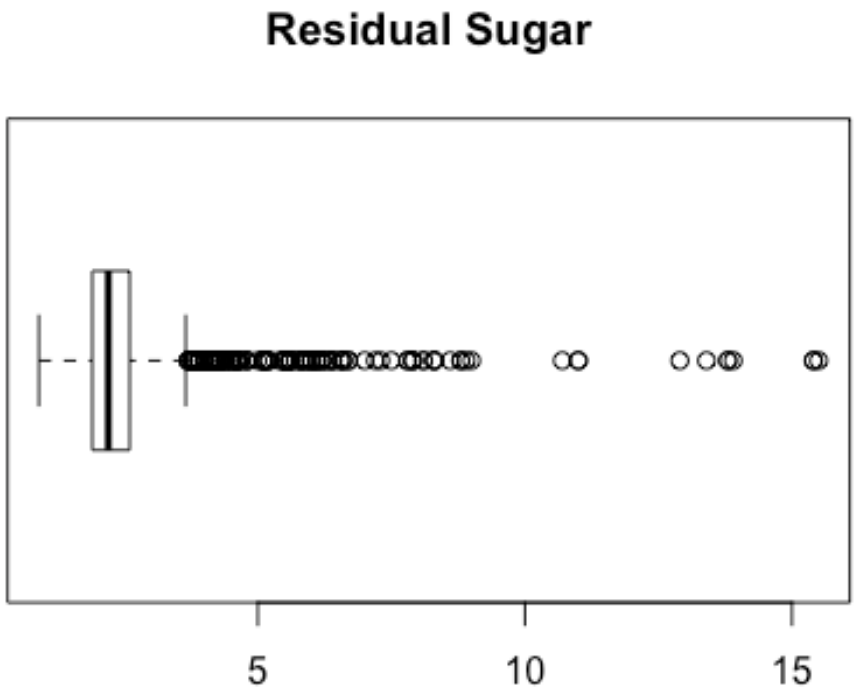
```
wine.citric<-boxplot(practica02$citric.acid,horizontal = TRUE,main="Citric
Acid")
```




```
wine.citric$out
```

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

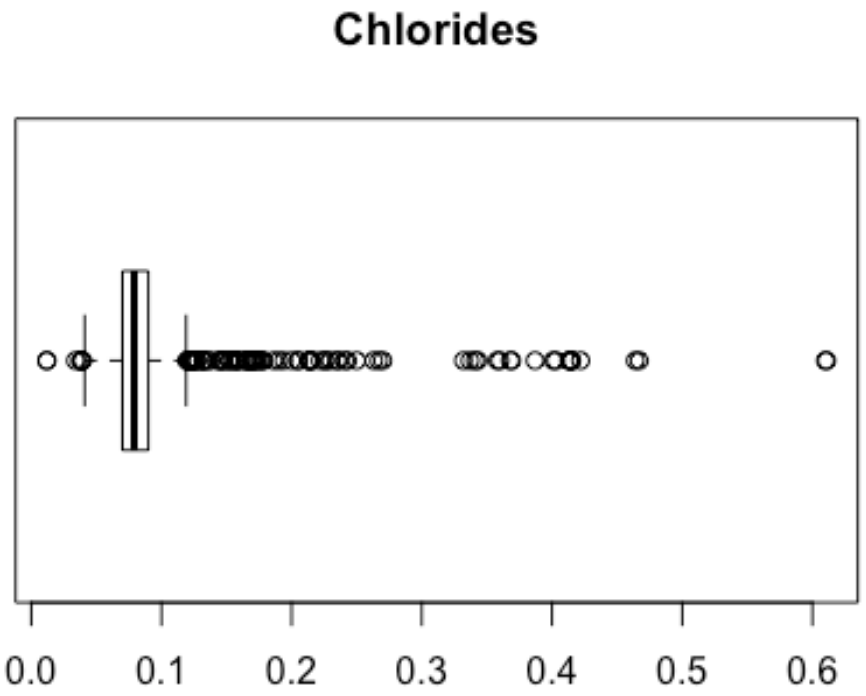
```
wine.sugar<-boxplot(practica02$residual.sugar,horizontal =
TRUE,main="Residual Sugar")
```




```
wine.sugar$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25
##  [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00
##  [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00
## [100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50
## [111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30
## [122] 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75
## [144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10
## [155]  7.80
```

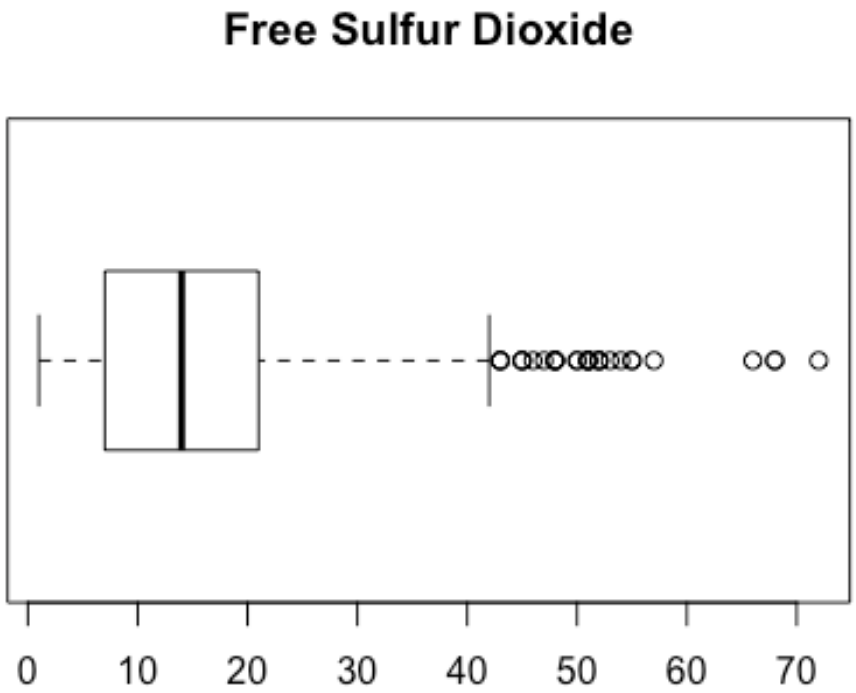
```
wine.chlorides<-boxplot(practica02$chlorides,horizontal =
TRUE,main="Chlorides")
```




```
wine.chlorides$out
```

```
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178
##  [12] 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343
##  [23] 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122
##  [34] 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171
##  [45] 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157
##  [56] 0.422 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126
##  [67] 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120
##  [78] 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136
##  [89] 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415
## [100] 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235
## [111] 0.230 0.038
```

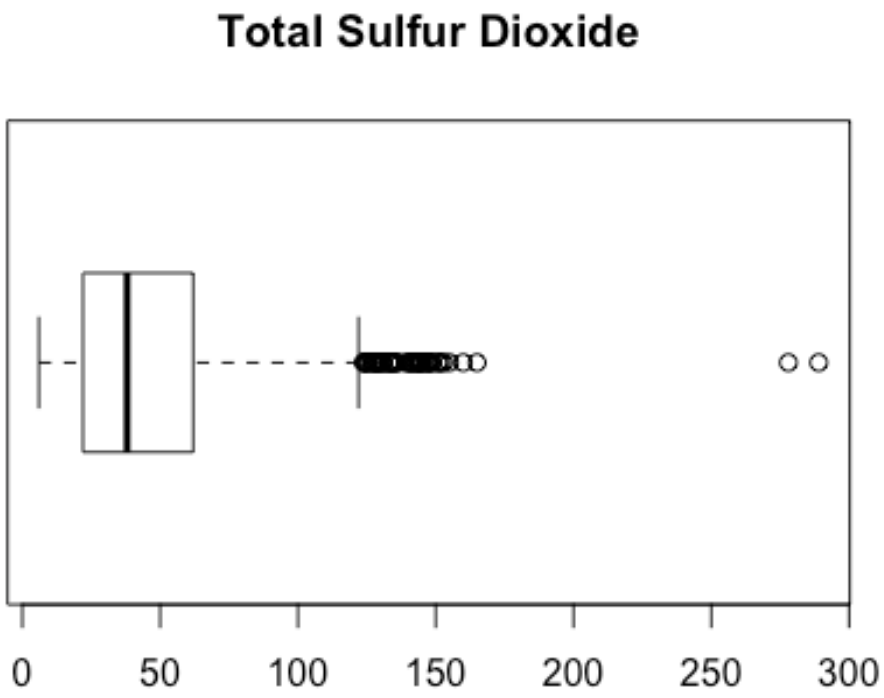
```
wine.freesulfurdioxide<-boxplot(practica02$free.sulfur.dioxide,horizontal =
TRUE,main="Free Sulfur Dioxide")
```




```
wine.freesulfurdioxide$out
```

```
##  [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51
## [24] 51 52 55 55 48 48 66
```

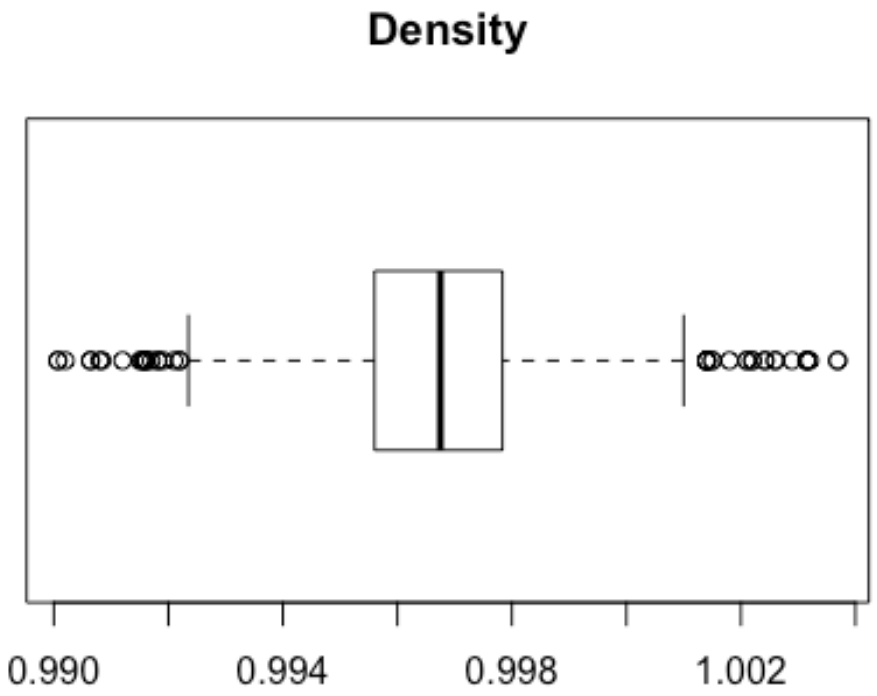
```
wine.totalsulfurdioxide<-boxplot(practica02$total.sulfur.dioxide,horizontal =
TRUE,main="Total Sulfur Dioxide")
```




```
wine.totalsulfurdioxide$out
```

```
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127
## [18] 126 145 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148
## [35] 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147
## [52] 147 131 131 131
```

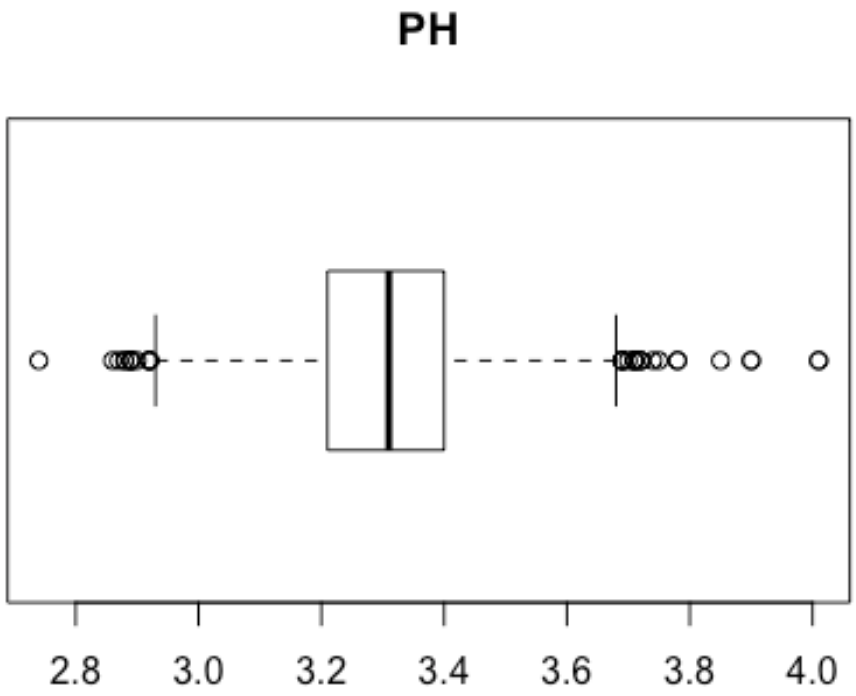
```
wine.density<-boxplot(practica02$density,horizontal = TRUE,main="Density")
```




```
wine.density$out
```

```
##  [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220
##  [9] 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140
## [17] 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260
## [25] 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162 0.99007 0.99007
## [33] 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369
## [41] 1.00369 1.00242 0.99182 1.00242 0.99182
```

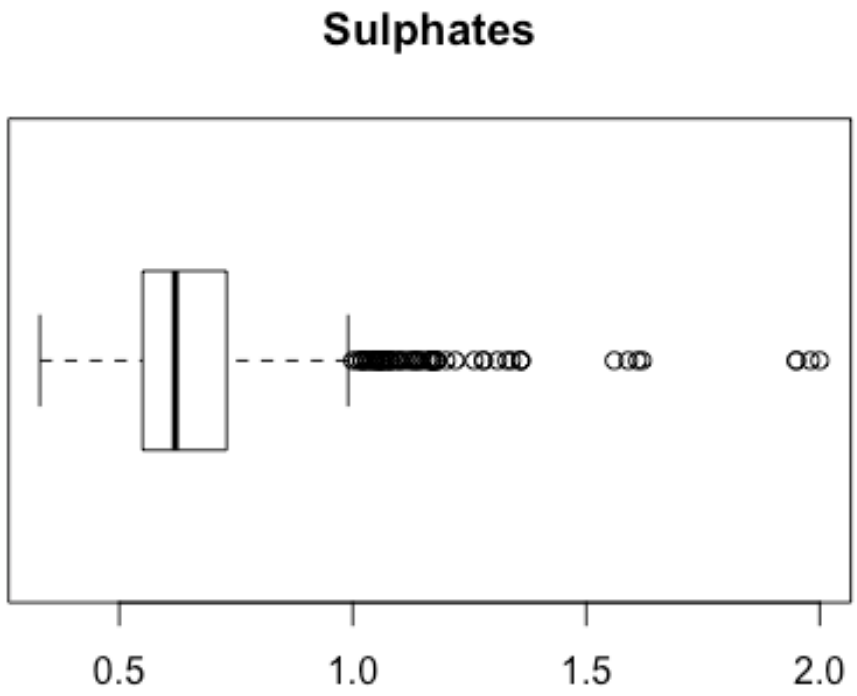
```
wine.pH<-boxplot(practica02$pH,horizontal = TRUE,main="PH")
```




```
wine.pH$out
```

```
##  [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87
## [15] 2.89 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78
## [29] 4.01 2.90 4.01 3.71 2.88 3.72 3.72
```

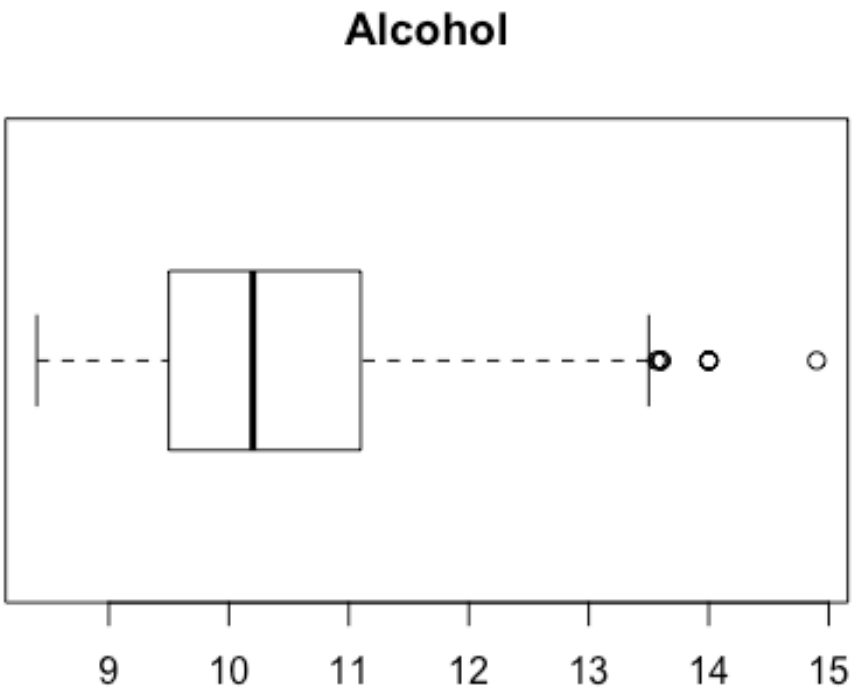
```
wine.sulphates<-boxplot(practica02$sulphates,horizontal =
TRUE,main="Sulphates")
```




```
wine.sulphates$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08
## [15] 1.59 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13
## [29] 1.07 1.06 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17
## [43] 1.62 1.06 1.18 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03
## [57] 1.17 1.10 1.01
```

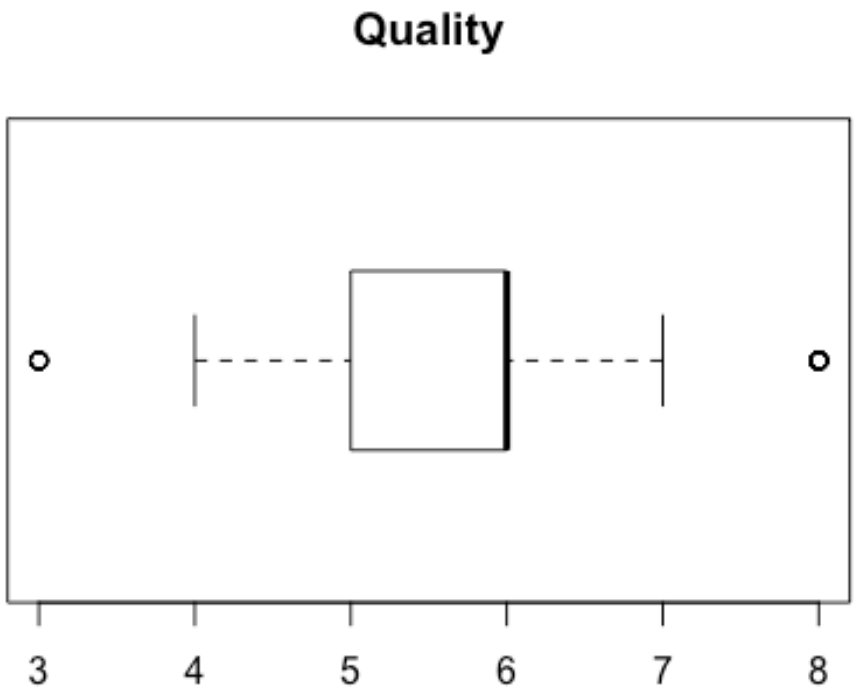
```
wine.alcohol<-boxplot(practica02$alcohol,horizontal = TRUE,main="Alcohol")
```




```
wine.alcohol$out
```

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000
##  [8] 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
```

```
wine.quality<-boxplot(practica02$quality,horizontal = TRUE,main="Quality")
```




```
wine.quality$out
```

```
##  [1] 8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 3 3 8 8 3 3 3 8
```

Tras analizar los boxplot se ve que sí que hay algunos valores que están fuera del rango habitual de las muestras. Sin embargo, al analizar el significado de dichos valores se ve que aunque atípicos, siguen siendo posibles. Por esta razón los valores no se han eliminado del análisis.

4. Análisis de los datos.

4.1. Selección de El Ph interviene principalmente en la sensación ácida del vino, pero también afecta al color y conservación del vino. Los valores normales en los vinos oscilan entre 2,5 y 4,5. con lo que no se ven valores no coherentes con los datos.los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. No obstante, como se verá en el apartado consistente en la realización de pruebas estadísticas, no todos se utilizarán.

- alcohol vs. density
- fixed.acidity vs. density
- residual.sugar vs total.sulfur.dioxide
- residual.sugar vs. density
- residual.sugar vs. alcohol
- chlorides vs. density
- chlorides vs. sulphates
- quality vs. alcohol

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

```
# Comprobación para el total de los datos
library(nortest)
alpha = 0.05
col.names = colnames(practica02)
for (i in 1:ncol(practica02)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(practica02[,i]) | is.numeric(practica02[,i])) {
    p_val = ad.test(practica02[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
# Format output
    if (i < ncol(practica02) - 1) cat(", ")
    if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## fixed.acidity, volatile.acidity, citric.acid,
## residual.sugar, chlorides, free.sulfur.dioxide,
## total.sulfur.dioxide, density, pH,
## sulphates, alcohol, quality
```

```r
# Comprobación para los vinos de calidad baja/media
col.names = colnames(baja)
for (i in 1:ncol(baja)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(baja[,i]) | is.numeric(baja[,i])) {
    p_val = ad.test(baja[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
# Format output
    if (i < ncol(baja) - 1) cat(", ")
    if (i %% 3 == 0) cat("\n")
    }
  }
}
```

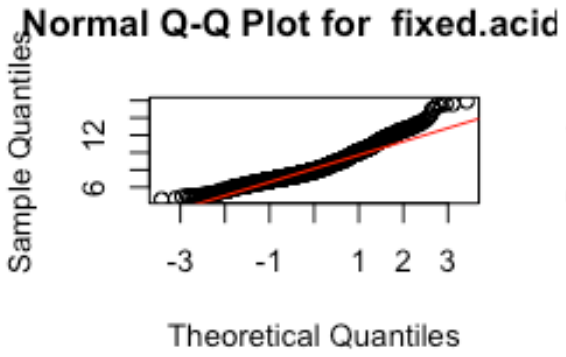
```
## Variables que no siguen una distribución normal:
## fixed.acidity, volatile.acidity, citric.acid,
## residual.sugar, chlorides, free.sulfur.dioxide,
## total.sulfur.dioxide, density, sulphates, alcohol, quality
```

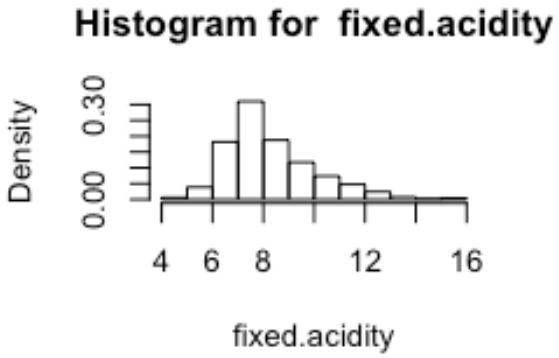
```r
# Comprobación para los vinos de calidad alta
col.names = colnames(alta)
for (i in 1:ncol(alta)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(alta[,i]) | is.numeric(alta[,i])) {
    p_val = ad.test(alta[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
# Format output
    if (i < ncol(alta) - 1) cat(", ")
    if (i %% 3 == 0) cat("\n")
    }
  }
}
```

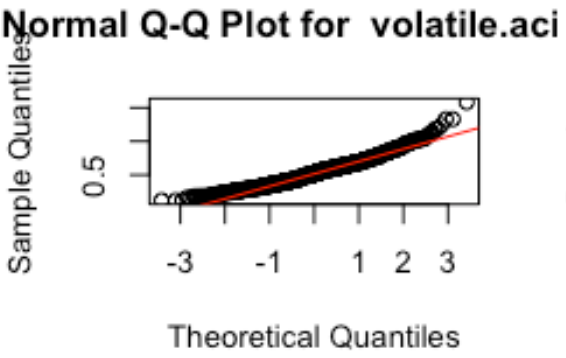
```
## Variables que no siguen una distribución normal:
## fixed.acidity, volatile.acidity, citric.acid,
## residual.sugar, chlorides, free.sulfur.dioxide,
## total.sulfur.dioxide, density, pH,
## sulphates, alcohol, quality
```

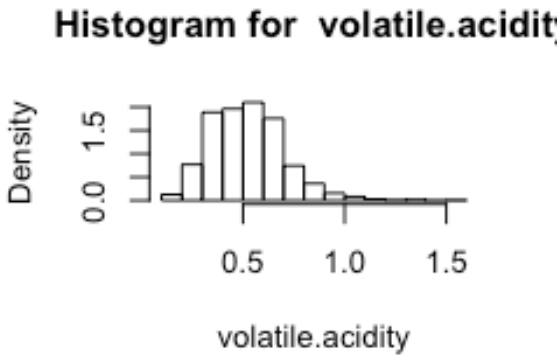
```r
# Visualización de la distribución de las variables para el conjunto de todos
los datos
par(mfrow=c(2,2))
for(i in 1:ncol(practica02)) {
if (is.numeric(practica02[,i])){
qqnorm(practica02[,i],main = paste("Normal Q-Q Plot for
",colnames(practica02)[i]))
qqline(practica02[,i],col="red")
hist(practica02[,i],
main=paste("Histogram for ", colnames(practica02)[i]),
```

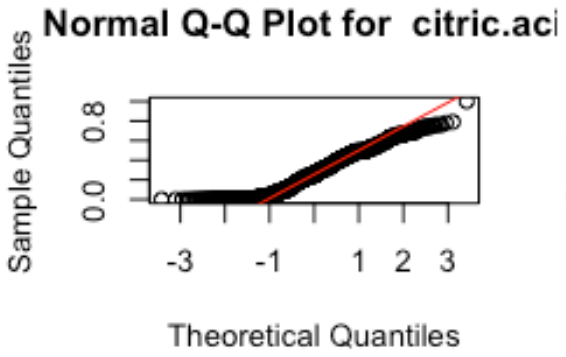
```
xlab=colnames(practica02)[i], freq = FALSE)
}
}
```

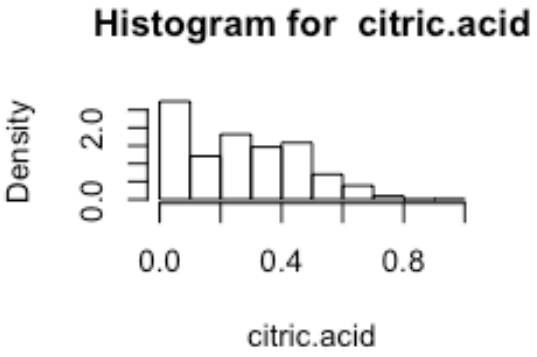

Normal Q-Q Plot for fixed.acid
Sample Quantiles
6
12
-3
-1
1
2
3
Theoretical Quantiles

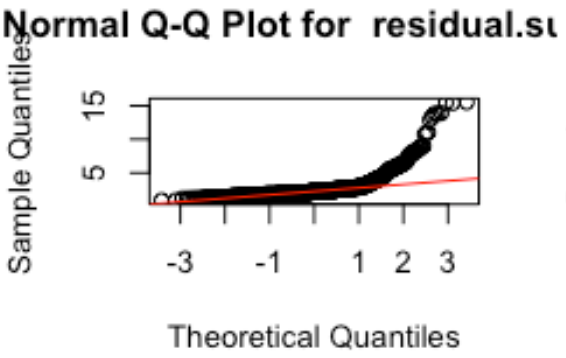

Histogram for fixed.acidity
Density
0.00
0.30
4
6
8
12
16
fixed.acidity

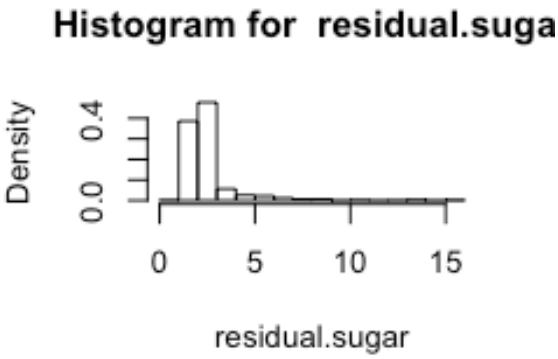

Normal Q-Q Plot for volatile.aci
Sample Quantiles
0.5
-3
-1
1
2
3
Theoretical Quantiles

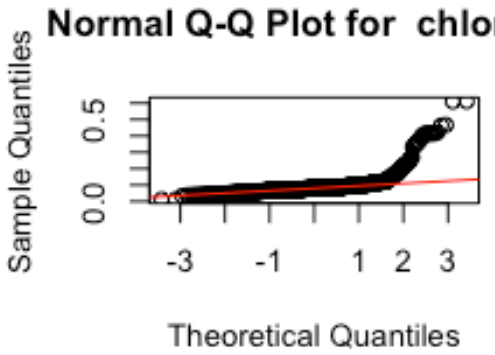

Histogram for volatile.acidit
Density
0.0
1.5
0.5
1.0
1.5
volatile.acidity

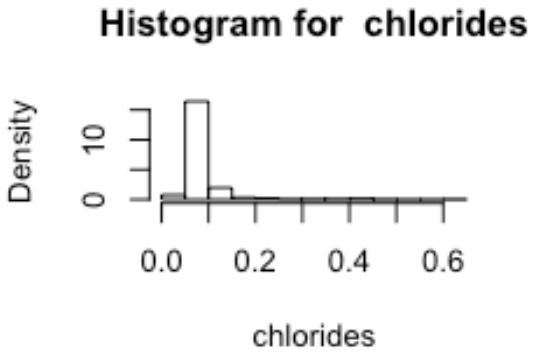

Normal Q-Q Plot for citric.aci
Sample Quantiles
0.0
0.8
-3
-1
1
2
3
Theoretical Quantiles

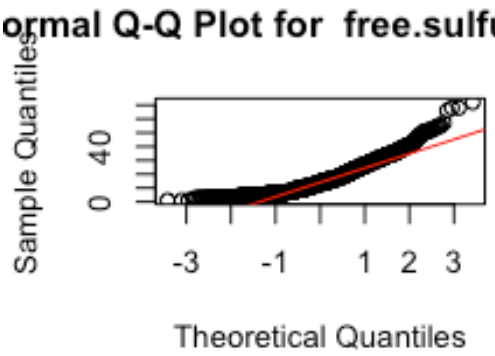

Histogram for citric.acid
Density
0.0
2.0
0.0
0.4
0.8
citric.acid

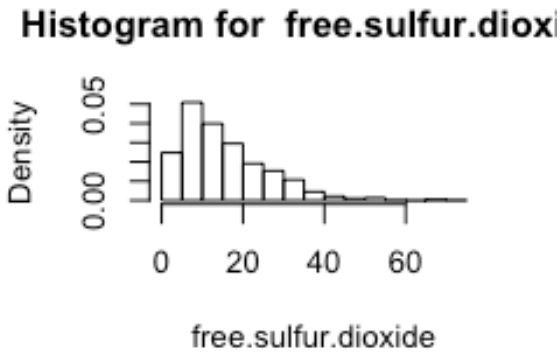

Normal Q-Q Plot for residual.su
Sample Quantiles
5
15
-3
-1
1
2
3
Theoretical Quantiles

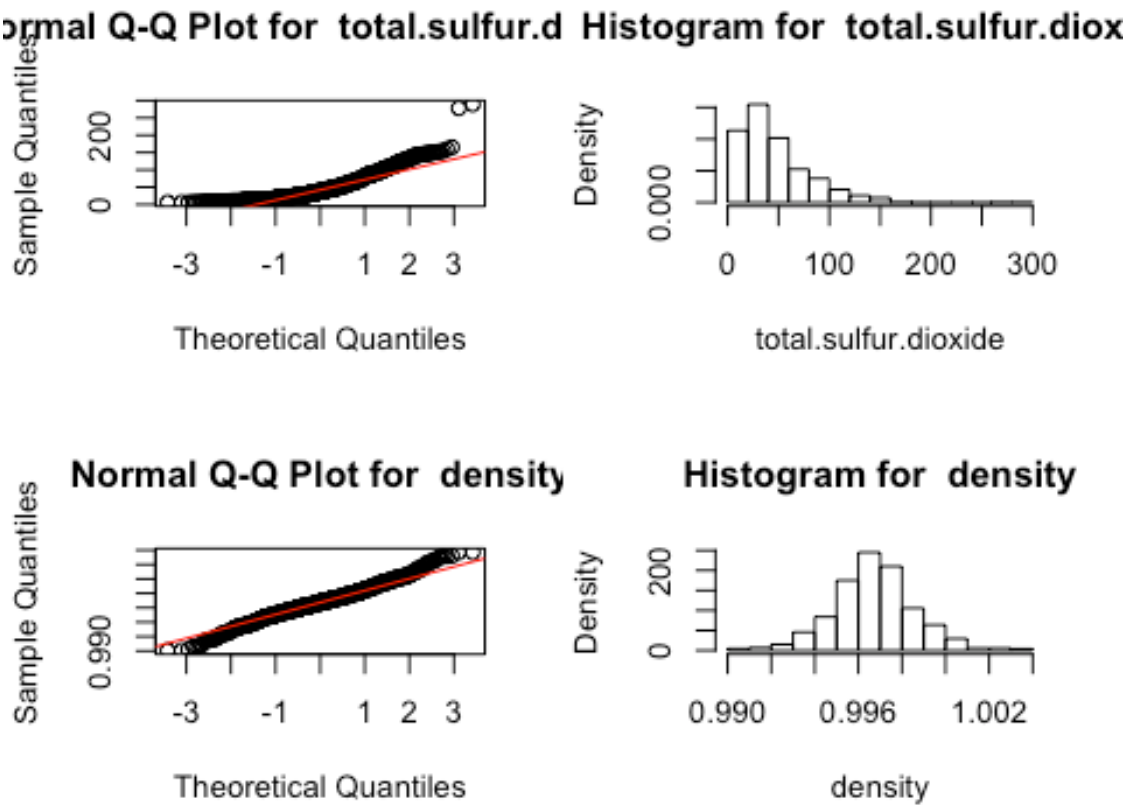

Histogram for residual.suga
Density
0.0
0.4
0
5
10
15
residual.sugar

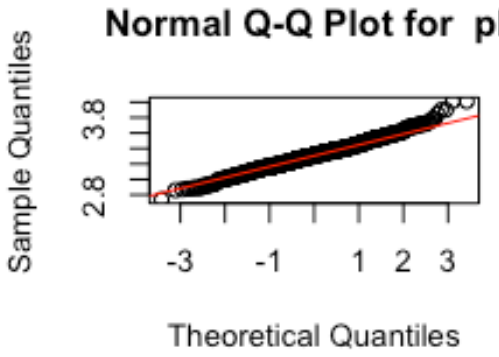

Normal Q-Q Plot for chloride
Sample Quantiles
0.0
0.5
-3
-1
1
2
3
Theoretical Quantiles

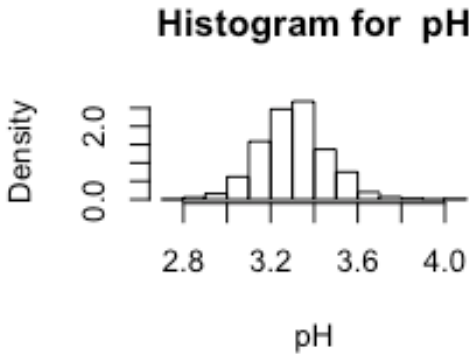

Histogram for chlorides
Density
0
10
0.0
0.2
0.4
0.6
chlorides

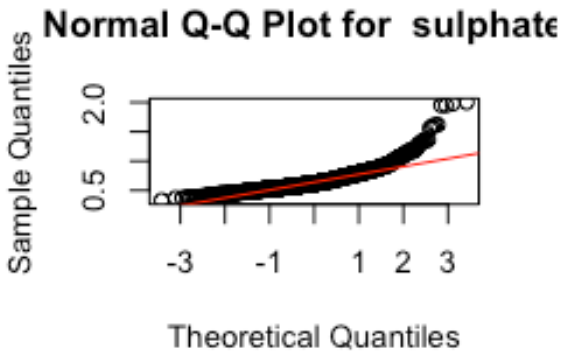

ormal Q-Q Plot for free.sulfur.di
Sample Quantiles
0
40
-3
-1
1
2
3
Theoretical Quantiles

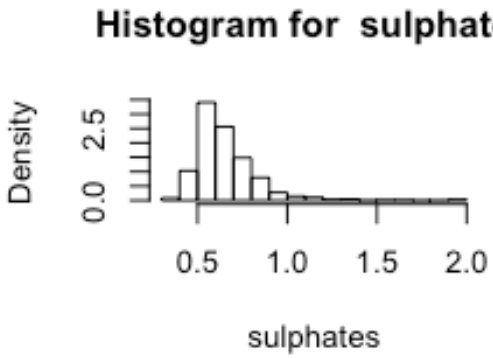

Histogram for free.sulfur.dioxi
Density
0.00
0.05
0
20
40
60
free.sulfur.dioxide

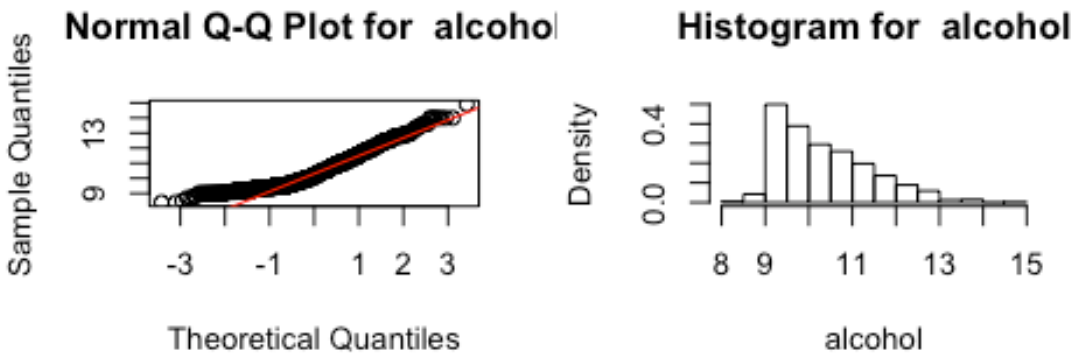

ormal Q-Q Plot for total.sulfur.d
Sample Quantiles
0
200
-3
-1
1
2
3
Theoretical Quantiles
Histogram for total.sulfur.diox
Density
0.000
0
100
200
300
total.sulfur.dioxide
Normal Q-Q Plot for density
Sample Quantiles
0.990
-3
-1
1
2
3
Theoretical Quantiles
Histogram for density
Density
0
200
0.990
0.996
1.002
density

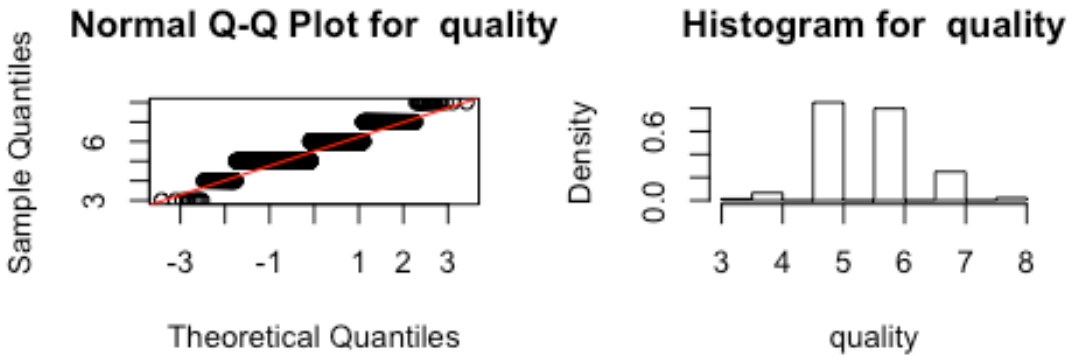

Normal Q-Q Plot for pH
Sample Quantiles
2.8
3.8
-3
-1
1
2
3
Theoretical Quantiles


Histogram for pH
Density
0.0
2.0
2.8
3.2
3.6
4.0
pH

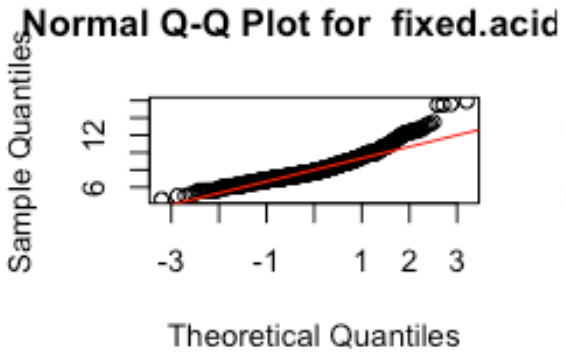

Normal Q-Q Plot for sulphate
Sample Quantiles
0.5
2.0
-3
-1
1
2
3
Theoretical Quantiles

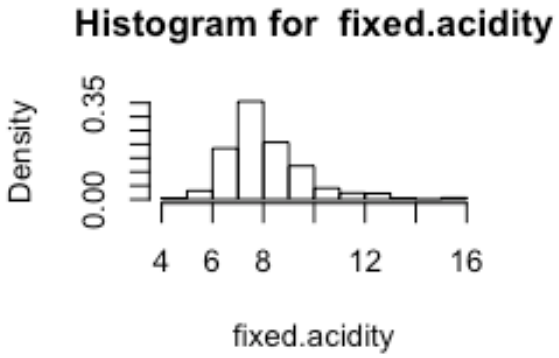

Histogram for sulphates
Density
0.0
2.5
0.5
1.0
1.5
2.0
sulphates

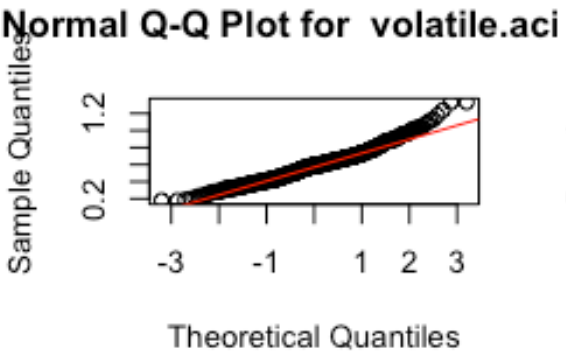


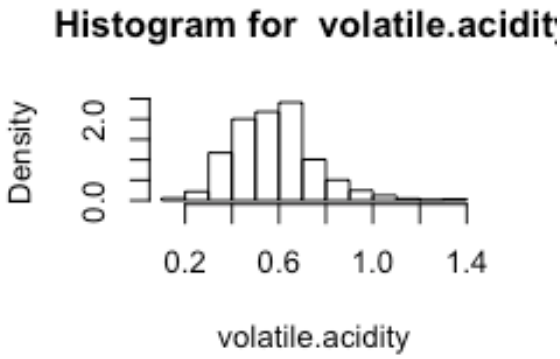


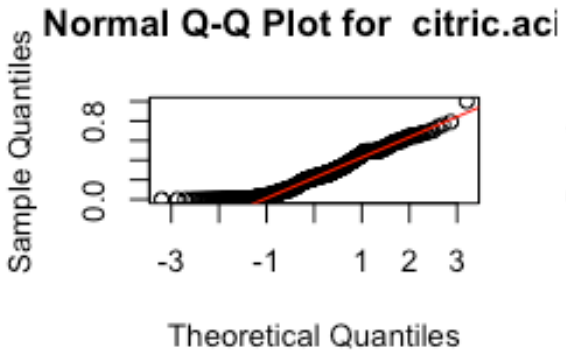

```r
# Visualización de la distribución de las variables para los vinos de calidad baja
par(mfrow=c(2,2))
for(i in 1:ncol(baja)) {
if (is.numeric(baja[,i])){
qqnorm(baja[,i],main = paste("Normal Q-Q Plot for ",colnames(baja)[i]))
qqline(baja[,i],col="red")
hist(baja[,i],
main=paste("Histogram for ", colnames(baja)[i]),
xlab=colnames(baja)[i], freq = FALSE)
}
}
```

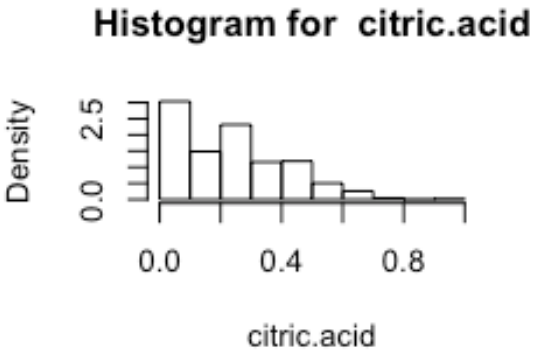

Normal Q-Q Plot for fixed.acid
Sample Quantiles
6
12
-3
-1
1
2
3
Theoretical Quantiles

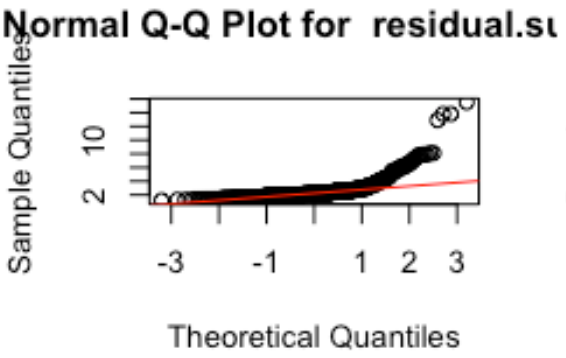

Histogram for fixed.acidity
Density
0.00
0.35
4
6
8
12
16
fixed.acidity

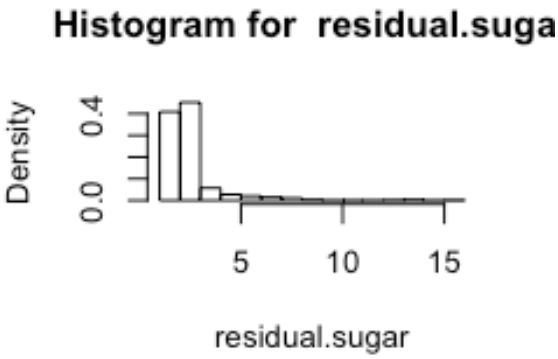

Normal Q-Q Plot for volatile.aci
Sample Quantiles
0.2
1.2
-3
-1
1
2
3
Theoretical Quantiles

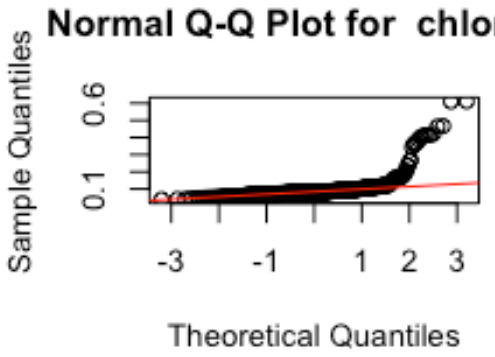

Histogram for volatile.acidit
Density
0.0
2.0
0.2
0.6
1.0
1.4
volatile.acidity

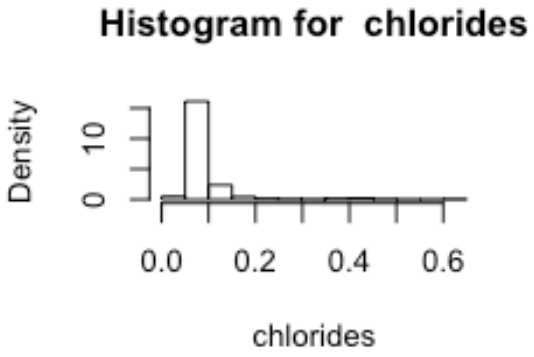

Normal Q-Q Plot for citric.aci
Sample Quantiles
0.0
0.8
-3
-1
1
2
3
Theoretical Quantiles

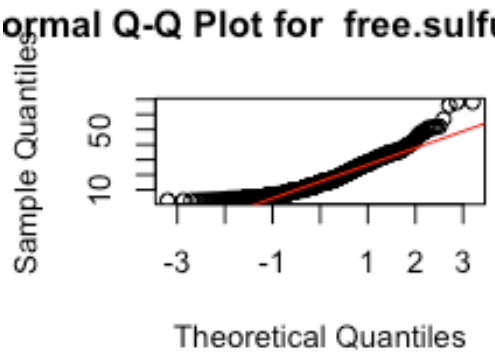

Histogram for citric.acid
Density
0.0
2.5
0.0
0.4
0.8
citric.acid

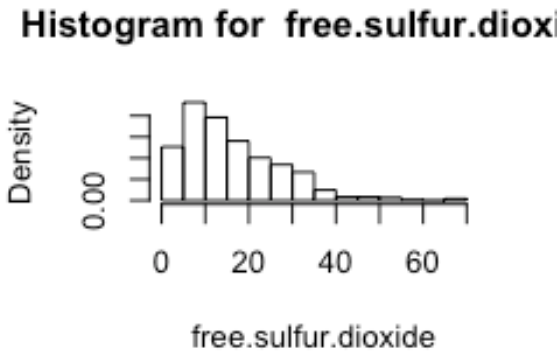

Normal Q-Q Plot for residual.su
Sample Quantiles
2
10
-3
-1
1
2
3
Theoretical Quantiles

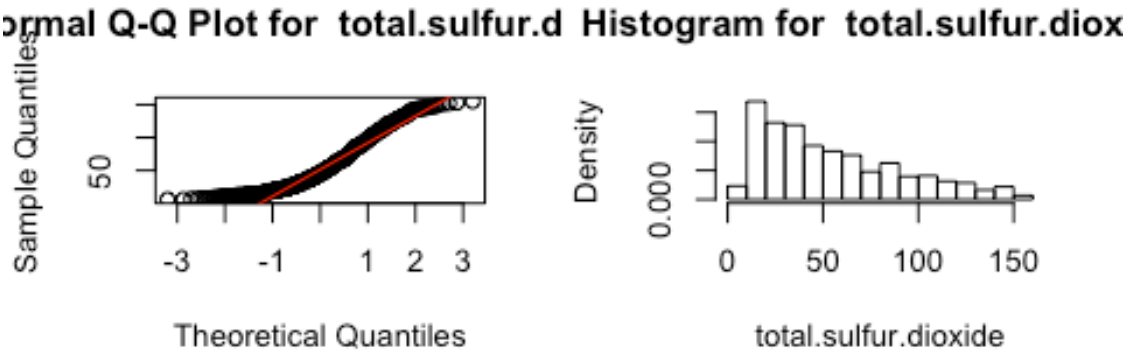

Histogram for residual.suga
Density
0.0
0.4
5
10
15
residual.sugar

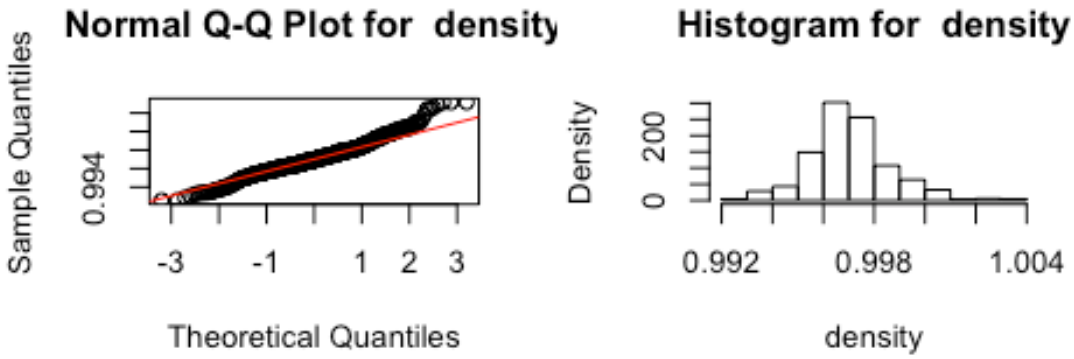

Normal Q-Q Plot for chloride
Sample Quantiles
0.1
0.6
-3
-1
1
2
3
Theoretical Quantiles

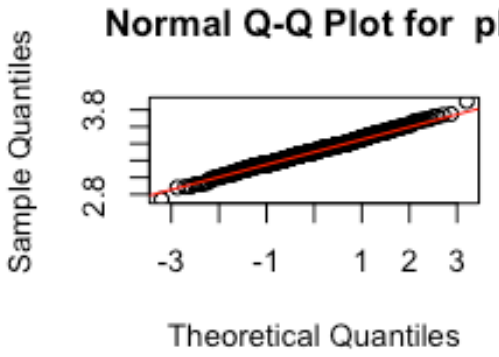

Histogram for chlorides
Density
0
10
0.0
0.2
0.4
0.6
chlorides

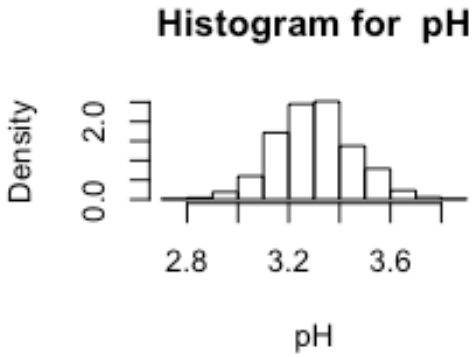

ormal Q-Q Plot for free.sulfur.di
Sample Quantiles
10
50
-3
-1
1
2
3
Theoretical Quantiles

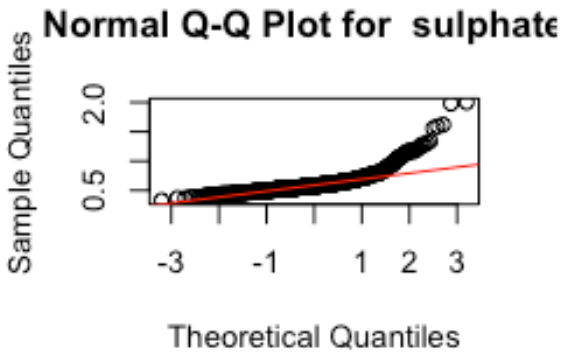

Histogram for free.sulfur.dioxi
Density
0.00
0
20
40
60
free.sulfur.dioxide

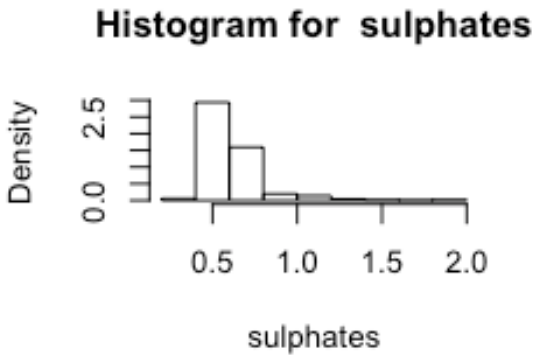

ormal Q-Q Plot for total.sulfur.d
Sample Quantiles
50
-3
-1
1
2
3
Theoretical Quantiles
Histogram for total.sulfur.diox
Density
0.000
0
50
100
150
total.sulfur.dioxide

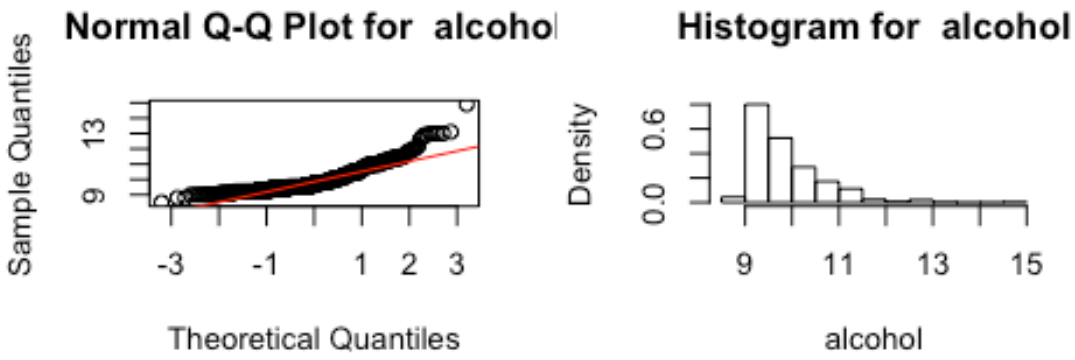

Normal Q-Q Plot for density
Sample Quantiles
0.994
-3
-1
1
2
3
Theoretical Quantiles
Histogram for density
Density
0
200
0.992
0.998
1.004
density

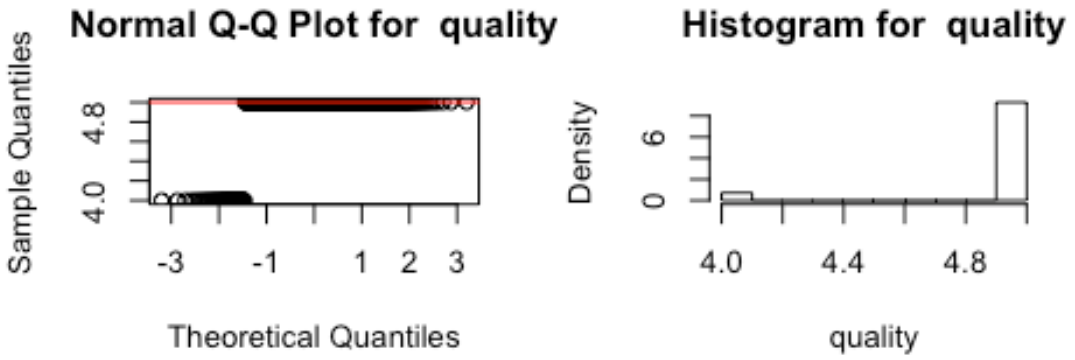

Normal Q-Q Plot for pH
Sample Quantiles
2.8
3.8
-3
-1
1
2
3
Theoretical Quantiles


Histogram for pH
Density
0.0
2.0
2.8
3.2
3.6
pH

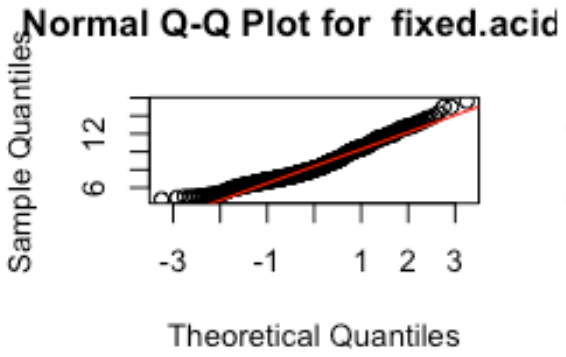

Normal Q-Q Plot for sulphate
Sample Quantiles
0.5
2.0
-3
-1
1
2
3
Theoretical Quantiles

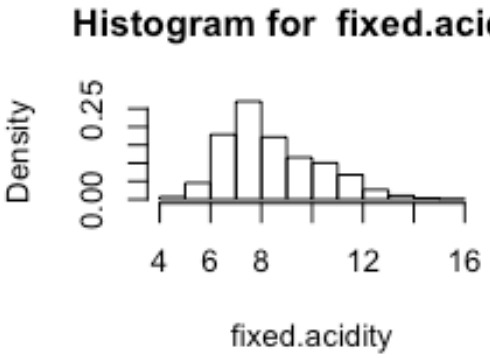

Histogram for sulphates
Density
0.0
2.5
0.5
1.0
1.5
2.0
sulphates

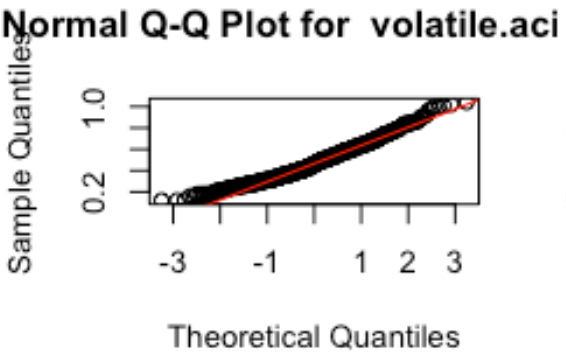


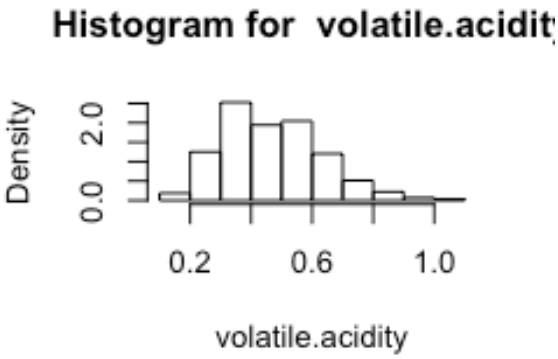


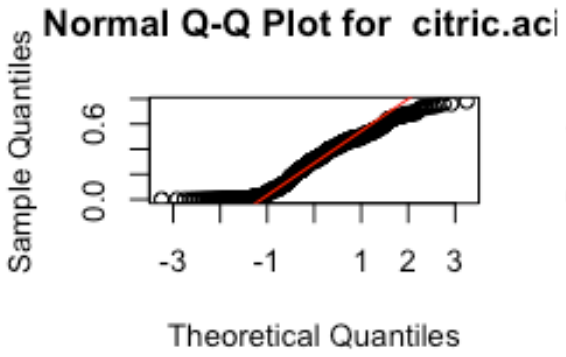

```r
# Visualización de la distribución de las variables para los vinos de calidad alta
par(mfrow=c(2,2))
for(i in 1:ncol(alta)) {
if (is.numeric(alta[,i])){
qqnorm(alta[,i],main = paste("Normal Q-Q Plot for ",colnames(alta)[i]))
qqline(alta[,i],col="red")
hist(alta[,i],
main=paste("Histogram for ", colnames(alta)[i]),
xlab=colnames(alta)[i], freq = FALSE)
}
}
```

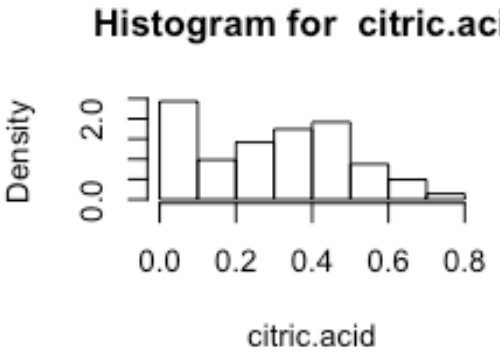

Normal Q-Q Plot for fixed.acid
Sample Quantiles
6
12
-3
-1
1
2
3
Theoretical Quantiles

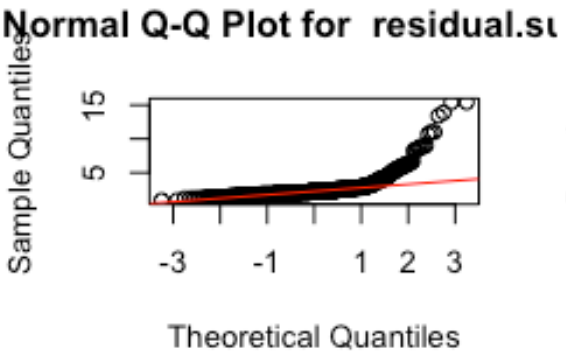

Histogram for fixed.acidity
Density
0.00
0.25
4
6
8
12
16
fixed.acidity

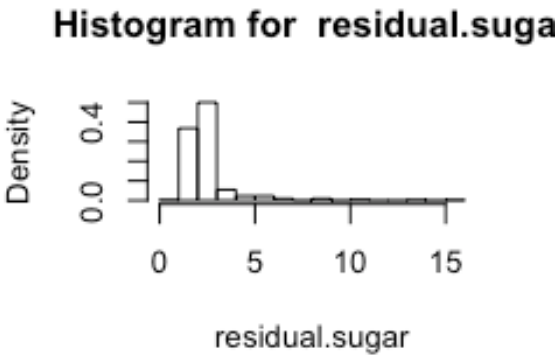

Normal Q-Q Plot for volatile.aci
Sample Quantiles
0.2
1.0
-3
-1
1
2
3
Theoretical Quantiles

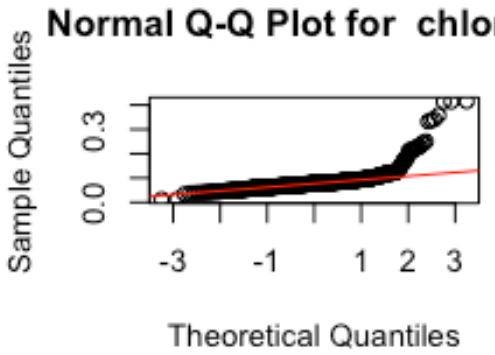

Histogram for volatile.acidit
Density
0.0
2.0
0.2
0.6
1.0
volatile.acidity

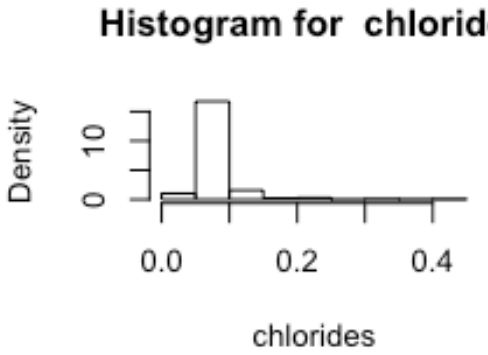

Normal Q-Q Plot for citric.aci
Sample Quantiles
0.0
0.6
-3
-1
1
2
3
Theoretical Quantiles

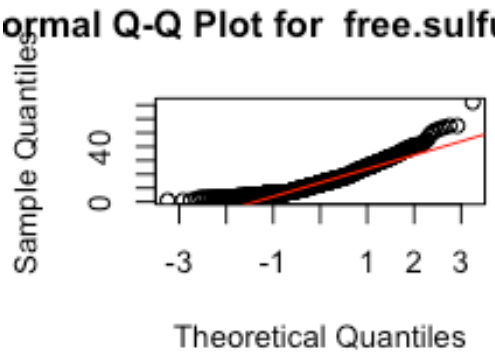

Histogram for citric.acid
Density
0.0
2.0
0.0
0.2
0.4
0.6
0.8
citric.acid

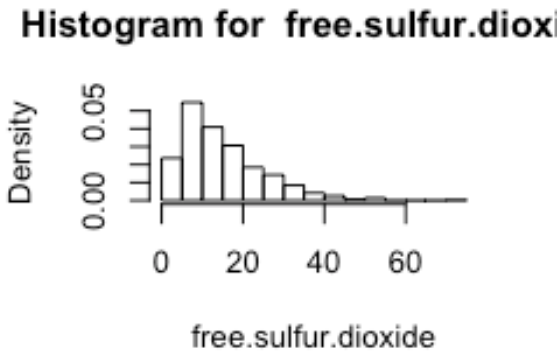

Normal Q-Q Plot for residual.su
Sample Quantiles
5
15
-3
-1
1
2
3
Theoretical Quantiles

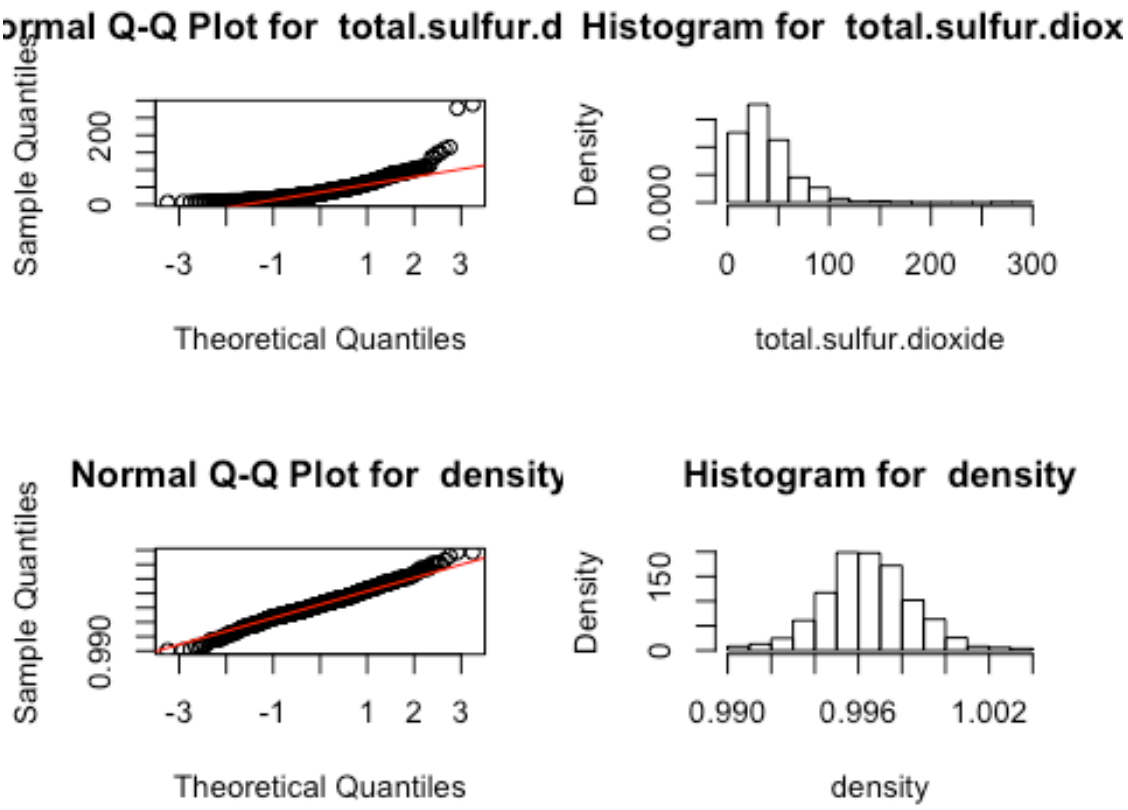

Histogram for residual.suga
Density
0.0
0.4
0
5
10
15
residual.sugar

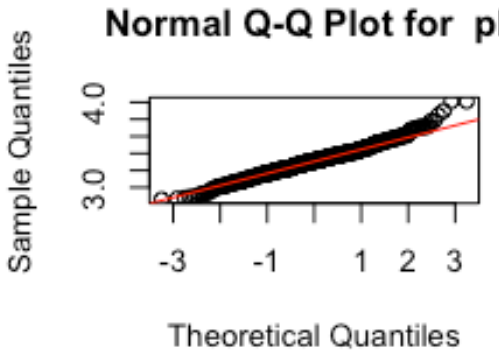

Normal Q-Q Plot for chloride
Sample Quantiles
0.0
0.3
-3
-1
1
2
3
Theoretical Quantiles

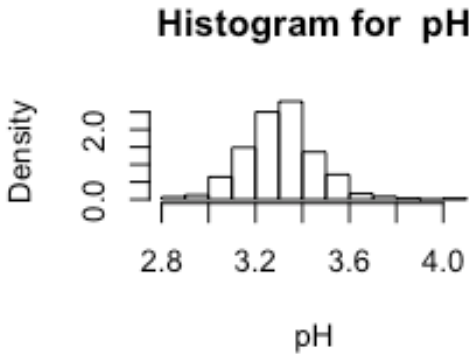

Histogram for chlorides
Density
0
10
0.0
0.2
0.4
chlorides

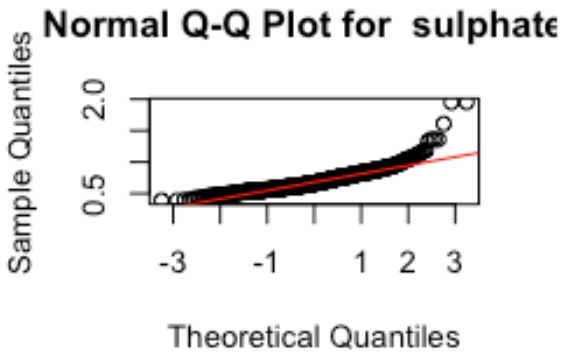

ormal Q-Q Plot for free.sulfur.di
Sample Quantiles
0
40
-3
-1
1
2
3
Theoretical Quantiles

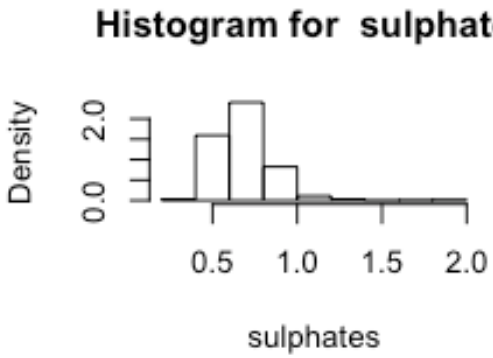

Histogram for free.sulfur.dioxi
Density
0.00
0.05
0
20
40
60
free.sulfur.dioxide

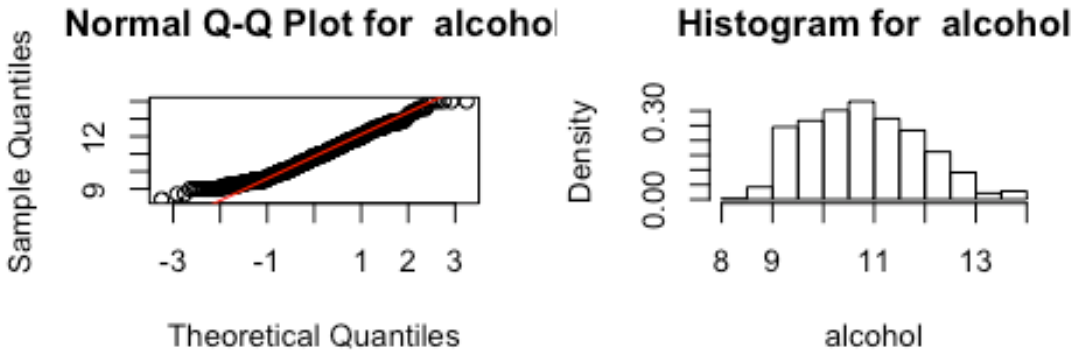

ormal Q-Q Plot for total.sulfur.d
Sample Quantiles
0
200
-3
-1
1
2
3
Theoretical Quantiles
Histogram for total.sulfur.diox
Density
0.000
0
100
200
300
total.sulfur.dioxide
Normal Q-Q Plot for density
Sample Quantiles
0.990
-3
-1
1
2
3
Theoretical Quantiles
Histogram for density
Density
0
150
0.990
0.996
1.002
density

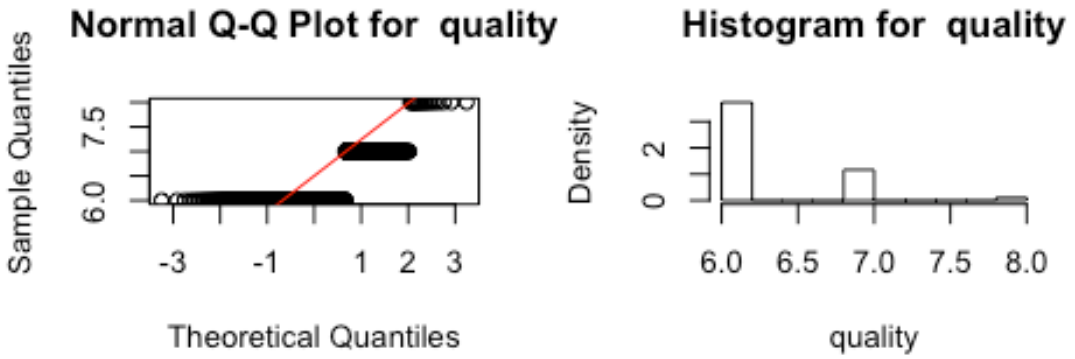

Normal Q-Q Plot for pH
Sample Quantiles
3.0
4.0
-3
-1
1
2
3
Theoretical Quantiles


Histogram for pH
Density
0.0
2.0
2.8
3.2
3.6
4.0
pH


Normal Q-Q Plot for sulphate
Sample Quantiles
0.5
2.0
-3
-1
1
2
3
Theoretical Quantiles


Histogram for sulphates
Density
0.0
2.0
0.5
1.0
1.5
2.0
sulphates





El test nos indica que ninguna variable, excepto el pH para los vinos de calidad baja, sigue una distribución normal, ya que el p-valor es inferior al coeficiente 0.05, por lo que se puede rechazar la hipotesis nula y concluir que la distribución no es normal. En la representación gráfica sin embargo, no se detecta una gran diferencia en la distribución del pH para los vinos de calidad baja.

Seguidamente, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por alcohol vs. density, fixed.acidity vs. density,residual.sugar vs total.sulfur.dioxide, residual.sugar vs. density,residual.sugar vs. alcohol, chlorides vs. density, chlorides vs. sulphates, quality vs. alcohol. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

Homogeneidad de varianzas

En el caso de que el p-valor sea inferior a 0.05 se rechazará la hipótesis nula y por tanto las varianzas serán distintas. Esto significa que para aquellas variables cuyas varianzas sean distintas no serán aplicables tests paramétricos tales como el ANOVA.

alcohol vs. quality

```
fligner.test(alcohol ~ quality, data = baja)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  alcohol by quality
## Fligner-Killeen:med chi-squared = 15.125, df = 1, p-value =
## 0.0001006
```

```
fligner.test(alcohol ~ quality, data = alta)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  alcohol by quality
## Fligner-Killeen:med chi-squared = 2.9208, df = 2, p-value = 0.2321
```

Dado que ambas pruebas resultan en un p-valor inferior al nivel de significancia (<0.05), se rechaza la hipótesis nula de homocedasticidad y se concluye que la variable quality presenta varianzas estadísticamente diferentes para los diferentes grupos de alcohol Esto se cumple tanto para los vinos de calidad baja/media, como para los de calidad alta.

fixed.acidity vs. quality

```
fligner.test(fixed.acidity ~ quality, data = baja)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  fixed.acidity by quality
## Fligner-Killeen:med chi-squared = 0.17837, df = 1, p-value =
## 0.6728
```

```
fligner.test(fixed.acidity ~ quality, data = alta)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  fixed.acidity by quality
## Fligner-Killeen:med chi-squared = 5.1235, df = 2, p-value =
## 0.07717
```

En ambos casos el p-valor es superior a 0.05 y se acepta la hipótesis nula, por lo que los tests paramétricos no serán aplicables en este caso.

residual.sugar vs quality

```
fligner.test(residual.sugar ~ quality, data = baja)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  residual.sugar by quality
## Fligner-Killeen:med chi-squared = 0.37191, df = 1, p-value = 0.542
```

```r
fligner.test(residual.sugar ~ quality, data = alta)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  residual.sugar by quality
## Fligner-Killeen:med chi-squared = 5.926, df = 2, p-value = 0.05166
```

En ambos casos el p-valor es superior a 0.05 y se acepta la hipótesis nula, por lo que los tests paramétricos no serán aplicables en este caso.

total.sulfur.dioxide vs. quality

```r
fligner.test(total.sulfur.dioxide ~ quality, data = baja)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  total.sulfur.dioxide by quality
## Fligner-Killeen:med chi-squared = 8.7007, df = 1, p-value =
## 0.003181
```

```r
fligner.test(total.sulfur.dioxide ~ quality, data = alta)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  total.sulfur.dioxide by quality
## Fligner-Killeen:med chi-squared = 5.456, df = 2, p-value = 0.06535
```

En este caso la hipótesis nula se rechaza para los vinos de baja calidad y se acepta para los de alta.

chlorides vs. quality

```r
fligner.test(chlorides ~ quality, data = baja)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chlorides by quality
## Fligner-Killeen:med chi-squared = 1.7167, df = 1, p-value = 0.1901
```

```r
fligner.test(chlorides ~ quality, data = alta)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chlorides by quality
## Fligner-Killeen:med chi-squared = 5.2679, df = 2, p-value = 0.0718
```

En este caso la hipótesis nula se rechaza para los vinos de baja calidad y se acepta para los de alta.

sulphates vs. pH

```
fligner.test(pH ~ quality, data = baja)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  pH by quality
## Fligner-Killeen:med chi-squared = 0.04192, df = 1, p-value =
## 0.8378

fligner.test(pH ~ quality, data = alta)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  pH by quality
## Fligner-Killeen:med chi-squared = 0.47769, df = 2, p-value =
## 0.7875
```

sulphates vs. quality

```
fligner.test(sulphates ~ quality, data = baja)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  sulphates by quality
## Fligner-Killeen:med chi-squared = 0.96113, df = 1, p-value =
## 0.3269

fligner.test(sulphates ~ quality, data = alta)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  sulphates by quality
## Fligner-Killeen:med chi-squared = 0.88391, df = 2, p-value =
## 0.6428
```

En este caso la hipótesis nula se rechaza para los vinos de baja calidad y se acepta para los de alta.

En resumen: se podrán aplicar tests paramétricos para todas las variables anteriores cuyo p-valor es >0.05. NO se deberían aplicar tests paramétricos para: alcohol (baja), total.sulphur.dioxide (baja), chlorides (baja) y sulphates (baja).

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

```
library(Kendall)

# Calcular el coeficiente de correlación
summary(Kendall(practica02$fixed.acidity,practica02$quality))

## Score =  89235 , Var(Score) = 389487232
## denominator =  1014425
## tau = 0.088, 2-sided pvalue =6.1989e-06

summary(Kendall(practica02$volatile.acidity,practica02$quality))

## Score =  -305885 , Var(Score) = 389576224
## denominator =  1016977
## tau = -0.301, 2-sided pvalue =< 2.22e-16

summary(Kendall(practica02$citric.acid,practica02$quality))

## Score =  169549 , Var(Score) = 389343808
## denominator =  1013334
## tau = 0.167, 2-sided pvalue =< 2.22e-16

summary(Kendall(practica02$residual.sugar,practica02$quality))

## Score =  25658 , Var(Score) = 388179680
## denominator =  996659.5
## tau = 0.0257, 2-sided pvalue =0.19284

summary(Kendall(practica02$chlorides,practica02$quality))

## Score =  -151214 , Var(Score) = 389513824
## denominator =  1015409
## tau = -0.149, 2-sided pvalue =1.8342e-14

summary(Kendall(practica02$free.sulfur.dioxide,practica02$quality))

## Score =  -45902 , Var(Score) = 388972672
## denominator =  1005610
## tau = -0.0456, 2-sided pvalue =0.019946

summary(Kendall(practica02$total.sulfur.dioxide,practica02$quality))

## Score =  -159477 , Var(Score) = 389611488
## denominator =  1018293
## tau = -0.157, 2-sided pvalue =6.5083e-16

summary(Kendall(practica02$density,practica02$quality))
```

```
## Score =  -139527 , Var(Score) = 389658176
## denominator =  1021343
## tau = -0.137, 2-sided pvalue =1.5688e-12

summary(Kendall(practica02$sulphates,practica02$quality))

## Score =  303045 , Var(Score) = 389412640
## denominator =  1012615
## tau = 0.299, 2-sided pvalue =< 2.22e-16

summary(Kendall(practica02$alcohol,practica02$quality))

## Score =  382953 , Var(Score) = 389028960
## denominator =  1006798
## tau = 0.38, 2-sided pvalue =< 2.22e-16
```

Para el conjunto del dataset, las variables con una correlación más alta con respecto a la calidad son: volatile.acidity sulphates alcohol

```
summary(Kendall(baja$fixed.acidity,baja$quality))

## Score =  5737 , Var(Score) = 8836316
## denominator =  97408.7
## tau = 0.0589, 2-sided pvalue =0.053653

summary(Kendall(baja$volatile.acidity,baja$quality))

## Score =  -11272 , Var(Score) = 8839338
## denominator =  97772.98
## tau = -0.115, 2-sided pvalue =0.00015005

summary(Kendall(baja$citric.acid,baja$quality))

## Score =  9616 , Var(Score) = 8832024
## denominator =  97297.66
## tau = 0.0988, 2-sided pvalue =0.001215

summary(Kendall(baja$residual.sugar,baja$quality))

## Score =  407 , Var(Score) = 8807753
## denominator =  95869.01
## tau = 0.00425, 2-sided pvalue =0.89119

summary(Kendall(baja$chlorides,baja$quality))

## Score =  5597 , Var(Score) = 8837605
## denominator =  97599.88
## tau = 0.0573, 2-sided pvalue =0.059783

summary(Kendall(baja$free.sulfur.dioxide,baja$quality))

## Score =  10392 , Var(Score) = 8829076
## denominator =  96855.96
## tau = 0.107, 2-sided pvalue =0.00047052
```

```
summary(Kendall(baja$total.sulfur.dioxide,baja$quality))

## Score =  12603 , Var(Score) = 8841332
## denominator =  98043.75
## tau = 0.129, 2-sided pvalue =2.2531e-05

summary(Kendall(baja$density,baja$quality))

## Score =  7638 , Var(Score) = 8841240
## denominator =  98114.22
## tau = 0.0778, 2-sided pvalue =0.010216

summary(Kendall(baja$sulphates,baja$quality))

## Score =  7148 , Var(Score) = 8832820
## denominator =  97135.04
## tau = 0.0736, 2-sided pvalue =0.016183

summary(Kendall(baja$alcohol,baja$quality))

## Score =  -8274 , Var(Score) = 8795621
## denominator =  95540.52
## tau = -0.0866, 2-sided pvalue =0.0052786
```

Para los vinos de calidad baja del dataset, las variables con una correlación más alta con respecto a la calidad son: volatile.acidity free.sulfur total.sulfur

```
summary(Kendall(alta$fixed.acidity,alta$quality))

## Score =  23930 , Var(Score) = 39745632
## denominator =  225694.3
## tau = 0.106, 2-sided pvalue =0.00014734

summary(Kendall(alta$volatile.acidity,alta$quality))

## Score =  -48591 , Var(Score) = 39747708
## denominator =  225803.5
## tau = -0.215, 2-sided pvalue =1.2871e-14

summary(Kendall(alta$citric.acid,alta$quality))

## Score =  41219 , Var(Score) = 39734812
## denominator =  225416.3
## tau = 0.183, 2-sided pvalue =< 2.22e-16

summary(Kendall(alta$residual.sugar,alta$quality))

## Score =  14832 , Var(Score) = 39601972
## denominator =  221381.1
## tau = 0.067, 2-sided pvalue =0.018436

summary(Kendall(alta$chlorides,alta$quality))
```

```
## Score =  -26732 , Var(Score) = 39745864
## denominator =  225711
## tau = -0.118, 2-sided pvalue =2.2349e-05

summary(Kendall(alta$free.sulfur.dioxide,alta$quality))

## Score =  -19744 , Var(Score) = 39666236
## denominator =  223144.2
## tau = -0.0885, 2-sided pvalue =0.00172

summary(Kendall(alta$total.sulfur.dioxide,alta$quality))

## Score =  -29972 , Var(Score) = 39749996
## denominator =  225978.6
## tau = -0.133, 2-sided pvalue =1.9972e-06

summary(Kendall(alta$density,alta$quality))

## Score =  -25157 , Var(Score) = 39760768
## denominator =  227130.2
## tau = -0.111, 2-sided pvalue =6.6224e-05

summary(Kendall(alta$sulphates,alta$quality))

## Score =  48700 , Var(Score) = 39735456
## denominator =  225111.1
## tau = 0.216, 2-sided pvalue =< 2.22e-16

summary(Kendall(alta$alcohol,alta$quality))

## Score =  65756 , Var(Score) = 39730284
## denominator =  224762.6
## tau = 0.293, 2-sided pvalue =< 2.22e-16
```

Para los vinos de calidad alta del dataset, las variables con una correlación más alta con respecto a la calidad son: volatile.acidity sulphates alcohol

z.test para vinos de calidad baja. Intervalos de confianza al 95% para las 3 variables con mayor correlación (Calidad estimada del vino)

```
library(BSDA)

z <- z.test(x = baja$volatile.acidity, y = baja$quality, # Two samples with
normal distribution
            alt = "two.sided",          # Dos colas
            mu = 0,                     # H_0: mu_1 - mu_2 = 0
            sigma.x = sd(baja$volatile.acidity),     # desviación estándar m
            sigma.y = sd(baja$quality),     # desviación estándar n
            conf.level = 0.95)          # IC: error alpha_a/2 = 0.01/2

z
```

```
##
##  Two-sample z-Test
##
## data:  baja$volatile.acidity and baja$quality
## z = -378.45, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.364798 -4.319820
## sample estimates:
## mean of x mean of y
## 0.5854837 4.9277929
```

```
z <- z.test(x = baja$free.sulfur.dioxide, y = baja$quality, # Two samples
with normal distribution
            alt = "two.sided",          # Dos colas
            mu = 0,                     # H_0: mu_1 - mu_2 = 0
            sigma.x = sd(baja$free.sulfur.dioxide),     # desviación estándar
m
            sigma.y = sd(baja$quality),     # desviación estándar n
            conf.level = 0.95)          # IC: error alpha_a/2 = 0.01/2
```

```
z
```

```
##
##  Two-sample z-Test
##
## data:  baja$free.sulfur.dioxide and baja$quality
## z = 29.134, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.92713 12.50339
## sample estimates:
## mean of x mean of y
## 16.643052  4.927793
```

```
z <- z.test(x = baja$total.sulfur.dioxide, y = practica02$quality, # Two
samples with normal distribution
            alt = "two.sided",          # Dos colas
            mu = 0,                     # H_0: mu_1 - mu_2 = 0
            sigma.x = sd(baja$total.sulfur.dioxide),     # desviación
estándar m
            sigma.y = sd(baja$quality),     # desviación estándar n
            conf.level = 0.95)          # IC: error alpha_a/2 = 0.01/2
```

```
z
```

```
##
##  Two-sample z-Test
##
## data:  baja$total.sulfur.dioxide and practica02$quality
## z = 36.421, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  46.75522 52.07355
## sample estimates:
## mean of x mean of y
## 55.050409  5.636023
```

z.test para vinos de calidad alta. Intervalos de confianza al 95% para las variables con mayor correlación (Calidad estimada del vino)

```
z <- z.test(x = alta$volatile.acidity, y = alta$quality, # Two samples with normal distribution
            alt = "two.sided",          # Dos colas
            mu = 0,                     # H_0: mu_1 - mu_2 = 0
            sigma.x = sd(alta$volatile.acidity),     # desviación estándar m
            sigma.y = sd(alta$quality),     # desviación estándar n
            conf.level = 0.95)          # IC: error alpha_a/2 = 0.01/2

z

##
##  Two-sample z-Test
##
## data:  alta$volatile.acidity and alta$quality
## z = -327.68, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.835404 -5.766011
## sample estimates:
## mean of x mean of y
## 0.4741462 6.2748538

z <- z.test(x = alta$sulphates, y = alta$quality, # Two samples with normal distribution
            alt = "two.sided",          # Dos colas
            mu = 0,                     # H_0: mu_1 - mu_2 = 0
            sigma.x = sd(alta$sulphates),     # desviación estándar m
            sigma.y = sd(alta$quality),     # desviación estándar n
            conf.level = 0.95)          # IC: error alpha_a/2 = 0.01/2

z

##
##  Two-sample z-Test
##
## data:  alta$sulphates and alta$quality
## z = -316.55, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.616798 -5.547670
## sample estimates:
```

```
## mean of x mean of y
## 0.6926199 6.2748538

z <- z.test(x = alta$alcohol, y = practica02$quality, # Two samples with
normal distribution
            alt = "two.sided",          # Dos colas
            mu = 0,                     # H_0: mu_1 - mu_2 = 0
            sigma.x = sd(alta$alcohol),     # desviación estándarr m
            sigma.y = sd(alta$quality),     # desviación estándar n
            conf.level = 0.95)          # IC: error alpha_a/2 = 0.01/2

z

##
##  Two-sample z-Test
##
## data:  alta$alcohol and practica02$quality
## z = 131.21, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.141047 5.296966
## sample estimates:
## mean of x mean of y
## 10.855029  5.636023
```

Modelos de regresión

Para obtener un modelo de regresión lineal considerablemente eficiente, lo que haremos será obtener varios modelos de regresión utilizando las variables que están más correlacionadas, escogeremos el mejor utilizando como criterio aquel que presente un mayor coeficiente de determinación (R2):

Calidad baja

volatile.acidity free.sulfur total.sulfur

```
modelo1baja <- lm(quality~volatile.acidity + total.sulfur.dioxide +
free.sulfur.dioxide , data = baja)
modelo2baja <- lm(quality~residual.sugar + alcohol + density , data = baja)
modelo3baja <- lm(quality~pH+ total.sulfur.dioxide+density+ alcohol, data =
baja)
```

Calidad alta

volatile.acidity sulphates alcohol

```
modelo1alta <- lm(quality~volatile.acidity + sulphates + alcohol , data =
alta)
modelo2alta <- lm(quality~residual.sugar + total.sulfur.dioxide + density ,
data = alta)
```

```
modelo3alta <- lm(quality~residual.sugar+ total.sulfur.dioxide+density+
alcohol, data = alta)
```

Para los anteriores modelos de regresión lineal múltiple obtenidos, podemos utilizar el coeficiente de determinación para medir la bondad de los ajustes y quedarnos con aquel modelo que mejor coeficiente presente.

```
# Tabla con los coeficientes de determinación de cada modelo - Calidad baja
tabla.coeficientes <- matrix(c(1, summary(modelo1baja)$r.squared,
2, summary(modelo2baja)$r.squared,
3, summary(modelo3baja)$r.squared),
ncol = 2, byrow = TRUE)
colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes
```

```
##      Modelo        R^2
## [1,]      1 0.05021959
## [2,]      2 0.02125658
## [3,]      3 0.03790278
```

El mejor modelo para elegir es el 1, lo cual coincide con las variables con mayor correlación que calculamos anteriormente.

```
# Tabla con los coeficientes de determinación de cada modelo - Calidad alta
tabla.coeficientes <- matrix(c(1, summary(modelo1alta)$r.squared,
2, summary(modelo2alta)$r.squared,
3, summary(modelo3alta)$r.squared),
ncol = 2, byrow = TRUE)
colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes
```

```
##      Modelo        R^2
## [1,]      1 0.18161284
## [2,]      2 0.05004662
## [3,]      3 0.14113542
```

El modelo más ajustado sería el 2, que en este caso incluiría el azúcar, los sulfitos y la densidad, lo cual no se corresponde con las variables que presentaban una correlación más alta.

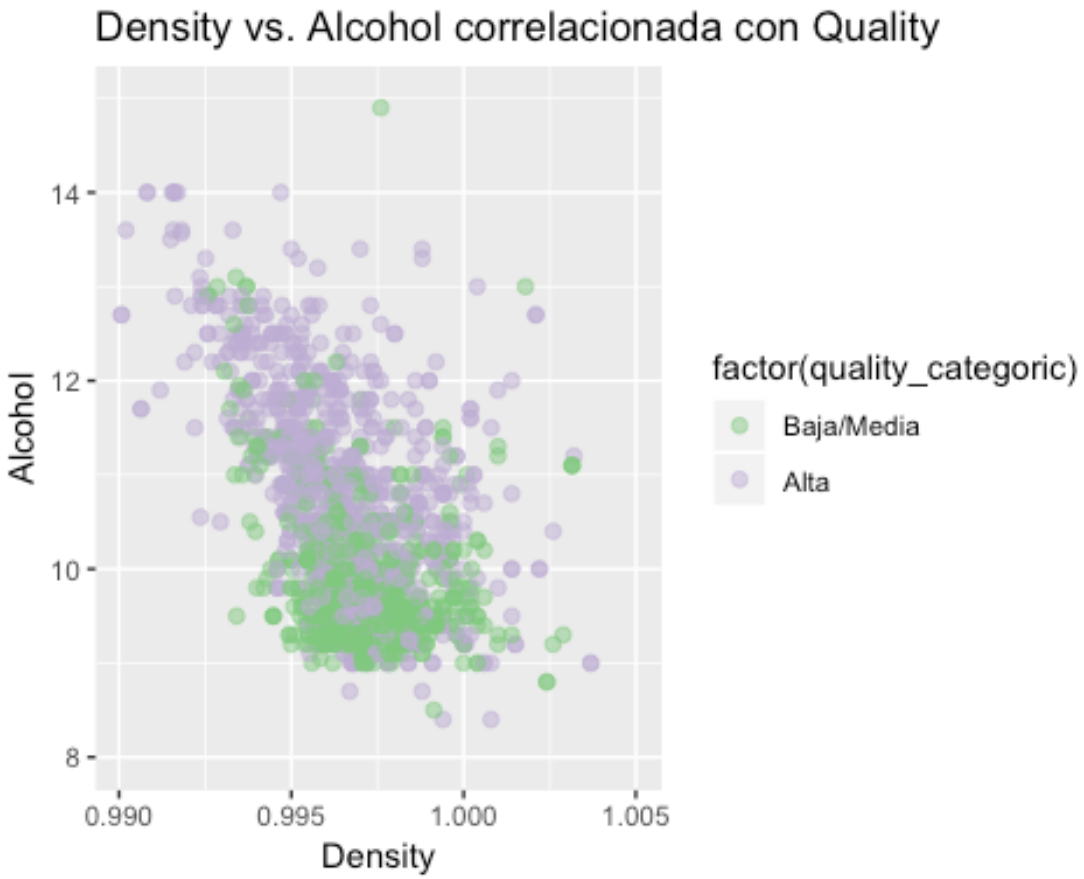
```
newdata <- data.frame(
residual.sugar=18,
total.sulfur.dioxide=34,
density=0.998,
alcohol=9.2
)
# Predecir el precio
predict(modelo3alta, newdata)
```

```
##        1
## 6.314543
```

Por ejemplo, para un vino con un azúcar de 18, sulfitos de 34, densidad 0.998 y alcohol 9.2, la calidad predicha sería de 6.3, por lo que estaría clasificado como de calidad alta.

5. Representación de los resultados a partir de tablas y gráficas.

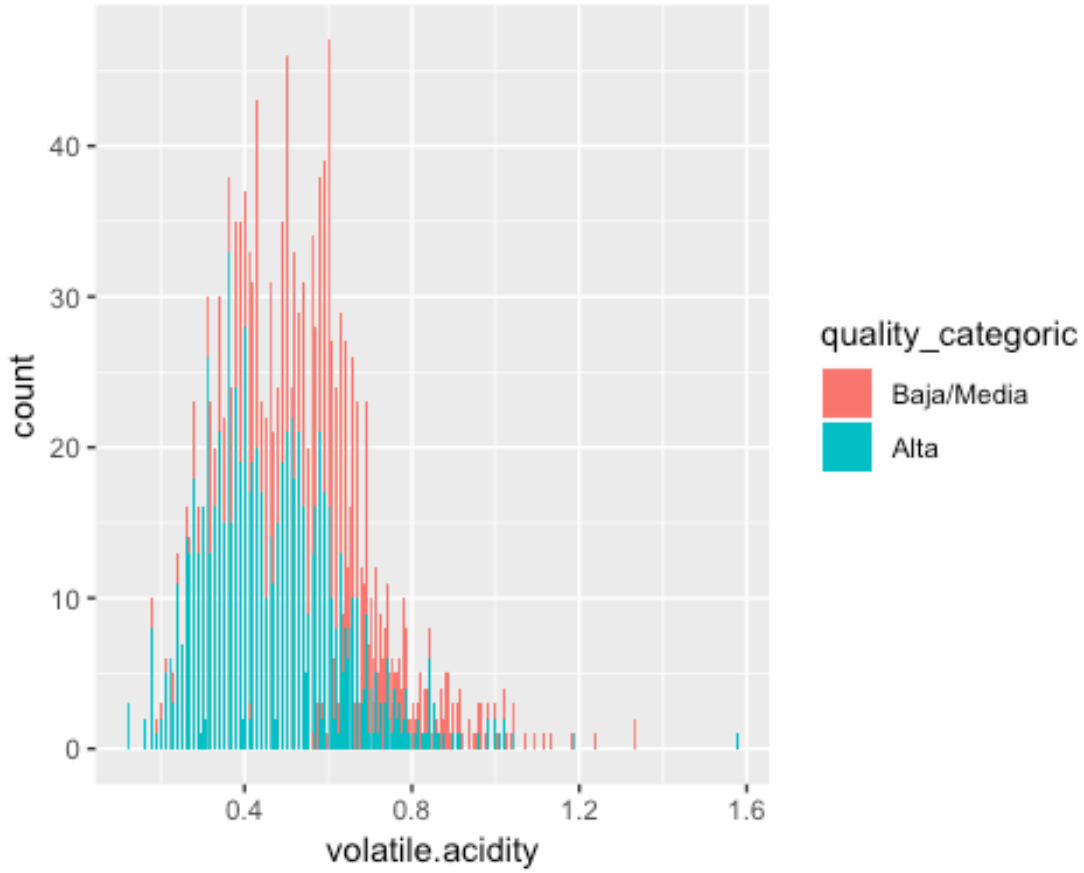
```
library(ggplot2)
ggplot(data = practica02,
       aes(x = density, y = alcohol, color = factor(quality_categoric))) +
   geom_point(alpha = 1/2, position = position_jitter(h = 0), size = 2) +
   coord_cartesian(xlim=c(min(practica02$density),1.005), ylim=c(8,15)) +
   scale_color_brewer(type='qual') +
   xlab('Density') +
   ylab('Alcohol') +
   ggtitle('Density vs. Alcohol correlacionada con Quality')
```



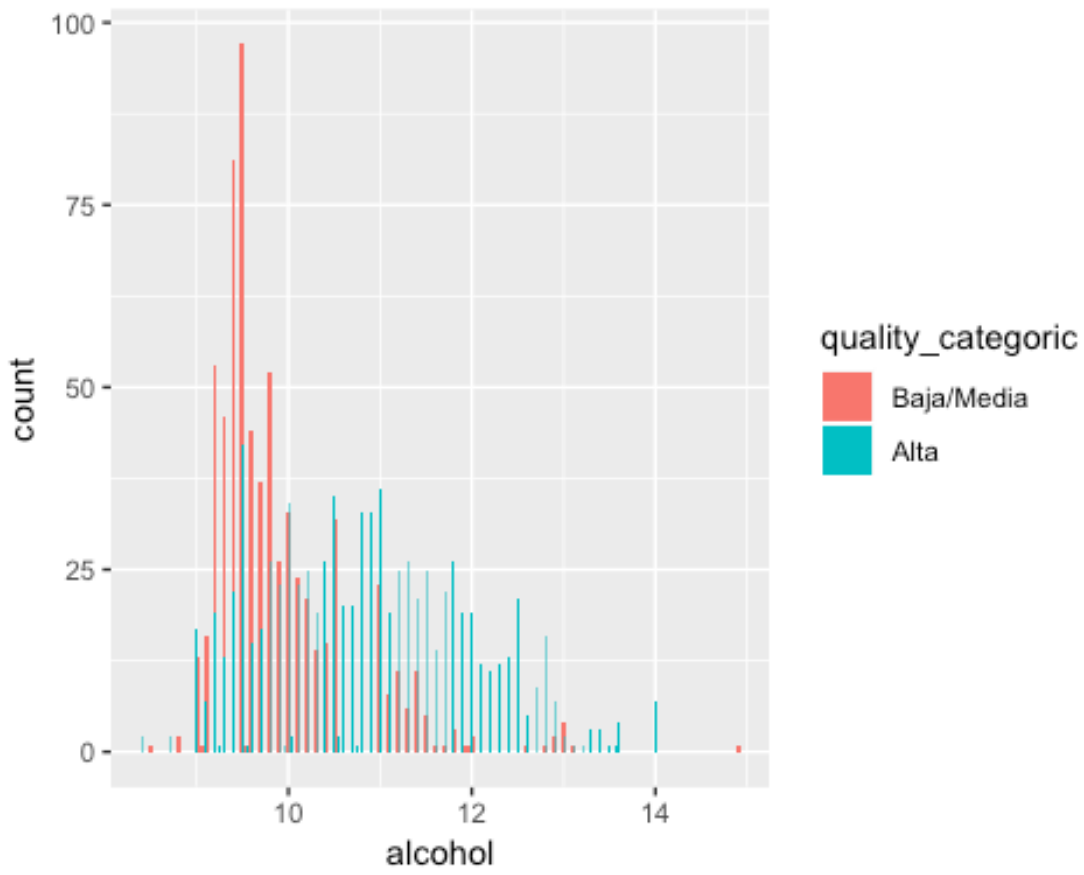
De la gráfica se ve que los vinos con una relación alcohol / densidad más alta son aquellos asociados a calidad alta, mientras que si la relación es al contrario serán más probablemente de calidad baja.

```
filas=dim(practica02)[1]

# Visualizamos la relación entre las variables "volatile acidity" y "quality"
para los vinos de ambas calidades:
```

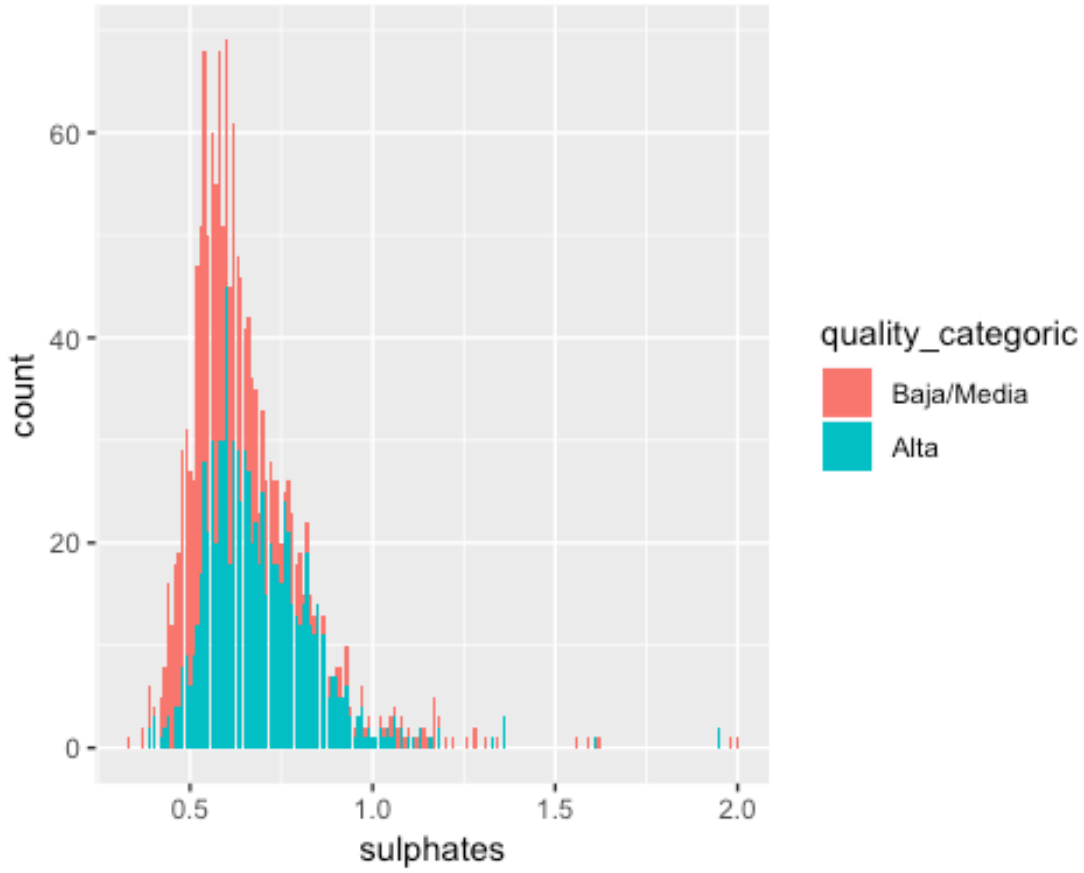
```
ggplot(data=practica02[1:filas,],aes(x=volatile.acidity,fill=quality_categori
c))+geom_bar()
```




```
# Calidad en función del alcohol para ambos grupos:
ggplot(data=practica02[1:filas,],aes(x=alcohol,fill=quality_categoric))+geom_
bar()

## Warning: position_stack requires non-overlapping x intervals
```



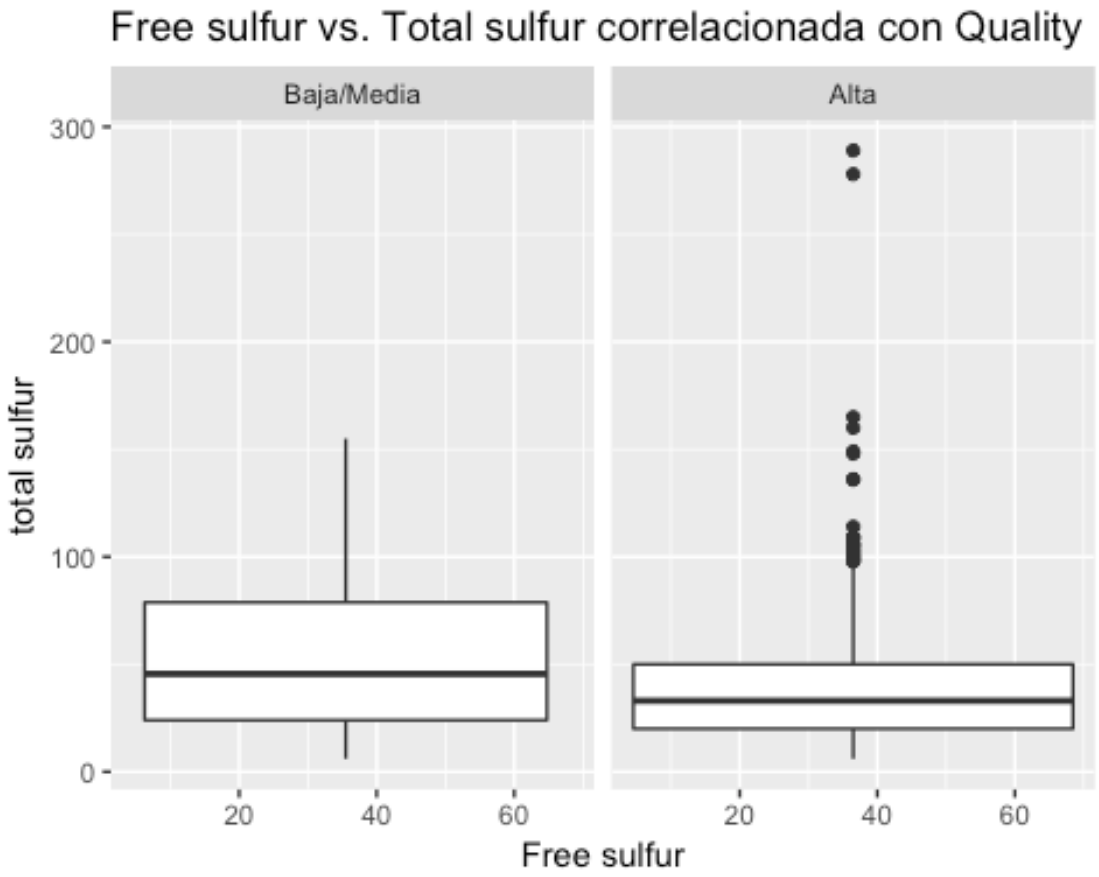

```
# Calidad en función de los sulfitos para ambos grupos:
ggplot(data=practica02[1:filas,],aes(x=sulphates,fill=quality_categoric))+geo
m_bar()
```



Los vinos de calidad alta tienen una acidez volátil más baja que los de menor calidad.Su graduación alcohólica es más alta que los de menor calidad y la cantidad de sulfitos presentes en los mismos es menor.

```
ggplot(data = practica02,
       aes(x = free.sulfur.dioxide, y = total.sulfur.dioxide) )+
   facet_wrap( ~ quality_categoric) +
   geom_boxplot() +
   xlab('Free sulfur') +
   ylab('total sulfur') +
   ggtitle('Free sulfur vs. Total sulfur correlacionada con Quality')
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



La cantidad de sulfitos libres con respecto a los totales es mayor para los vinos de calidad alta.

6. Resolución del problema. A partir de los resultados obtenidos, cuáles son las conclusiones? Los resultados permiten responder al problema?

A partir del análisis de correlaciones entre las distintas características del vino y su calidad, se ha visto que para los vinos etiquetados como de calidad media baja las variables más correlacionadas son volatile.acidity, free.sulfur, total.sulfur. Para los de calidad alta, las variables más correlacionadas son: volatile.acidity, sulphates y alcohol.

Al aplicar el análisis de regresión lineal, se confirma mediante el coeficiente R^2 que efectivamente las tres variables con alta correlación generan un buen modelo de regresión para los vinos de calidad baja. Sin embargo, para los de calidad alta el modelo de regresión más ajustado se obtiene teniendo en cuenta las variables residual.sugar, total.sulfur.dioxide y density.

Se puede decir que los vinos de calidad alta tienen una acidez volátil más baja que los de menor calidad.Su graduación alcohólica es más alta que los de menor calidad y la cantidad

de sulfitos presentes en los mismos es menor.La cantidad de sulfitos libres con respecto a los totales es mayor también para los vinos de calidad alta.

10. Finalmente, crear el archivo de datos corregido.

```
write.csv(practica02, file = "red_wine_quality_clean.csv",row.names=FALSE)
```