Tipología y ciclo de vida de los datos: Práctica2

Antonio Guzmán Martín

Diciembre 2017

Contents


```
library(readr)
wine_red <- read.csv(file="~/Desktop/Tipología\ y\ ciclo\ de\ vida\ de\ los\ datos/practica2/wine/wineq

#numero de filas por dataset
nrow(wine_red)
```

```
## [1] 1599
```

```
ncol<-ncol(wine_red)
#sacamos 5 primeras filas
head(wine_red[,1:ncol])
```

```
##   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
```

```
sapply(wine_red,class)
```

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

1. Descripción del dataset

Variables del dataset

- **fixed acidity:** Conjunto de los ácidos naturales procedentes de la uva (tartárico, málico, cítrico y succínico) o formados en la fermentación maloláctica (láctico). En general, los ácidos (acidez fija) son preservante naturales del vino y ayuda a mantener el color y cualidades aromáticas.
- **volatile acidity:** Conjunto de ácidos formados durante la fermentación o como consecuencia de alteraciones microbianas. Estos ácidos son, principalmente: ácido Acético, ácido Propionico, ácido Butírico y ácido Sulfúrico. Si la acidez volátil, presente en todos los vinos, es muy elevada el vino se picará y avigranará con el paso del tiempo. Es conveniente que la acidez volatil de un vino sea lo más baja posible.

El contenido en acidez volátil no puede ser superior a: a) 18 miliequivalentes por litro para los mostos de uva parcialmente fermentados, b) 18 miliequivalentes por litro para los vinos blancos y rosados, c) 20 miliequivalentes por litro para los vinos tintos.

- **citric acid:** En pequeñas cantidades este hacido puede añadir frescor y sabor a los vinos (dentro de acido fijo)
- **residual sugar:** Azúcar que queda en el vino después de la fermentación. Es raro encontrar vinos con menos de 1 g/l y vinos con más de 45 g/l son considerados dulces
- **chlorides:** cantidad de sal en el vino
- **free sulfur dioxide :** Previene del crecimiento microbial y de la oxidación del vino.La oxidación enturbia sus colores característicos (tornándolos en amarillos intensos e, incluso, marrones).Por lo que respecta al gusto, al beberlo notaremos sabores más secos y ásperos, incluso amargos en algunos casos.
- **total sulfur dioxide:** suma de concentraciones libres y amarradas de S02; concentraciones de dioxodo de sulfuro libres superiores a 50 ppm se vuelven evidentes en el sabor y olor
- **density:** densidad del vino, suele ser similar al del agua dependiendo de la concentración de azucar y alcohol
- **pH:** Describe como de ácido o básico es el vino 0 (very acidic) to 14 (very basic); mayoria vinos en escala 3-4.(principalmente 3,55 a 4)
- **sulphates:** Actua como un antimicrobial and antioxidante. Los sulfatos de sodio y calcio aparecen en el agua y por lo tanto la uva y el vino pueden contenerlos. Un agua con una cantidad de sulfatos inferior a 250mg/l se considera en este aspecto un agua de calidad y con valores superiores a 400mg/l insalubre.

- **alcohol:** cantidad de alcohol del vino. No es muy útil para hallar la calidad
- **quality:** calidad del vino entre 0 y 10

1.1 ¿Por qué es importante y qué pregunta/problema pretende responder?

Pregunta: ¿Qué componentes fisico-químicos influyen en que un vino sea bueno?. Obtener un modelo cuya combinación de variables permita determinar si es un buen vino.

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

La mayoría de los atributos corresponden con características necesarias para determinar la calidad del vino.

Sin embargo, a priori, podemos prescindir de la variable `total sulfur dioxide` (indica elsuma de concentraciones libres y amarradas, solo nos interesan las libres) y `density` (indica proporción de alcohol y esta no es interesante para determinar la calidad) (Según el estudio de: https://www.vinopack.es/criterios-que-determinan-la-calidad-en-el-vino).

No obstante, en los siguientes apartados comprobaremos si esto es cierto, o por el contrario si que afecta en la calidad.

3 Limpieza de datos

3.1 Elementos vacíos

```
# Números de valores desconocidos por campo
sapply(wine_red, function(x) sum(is.na(x)))
```

```
##        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
```

```
#No se han encontrado valores vacíos o NAs.

# Resumen de las variables
summary(wine_red)
```

```
##  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
```

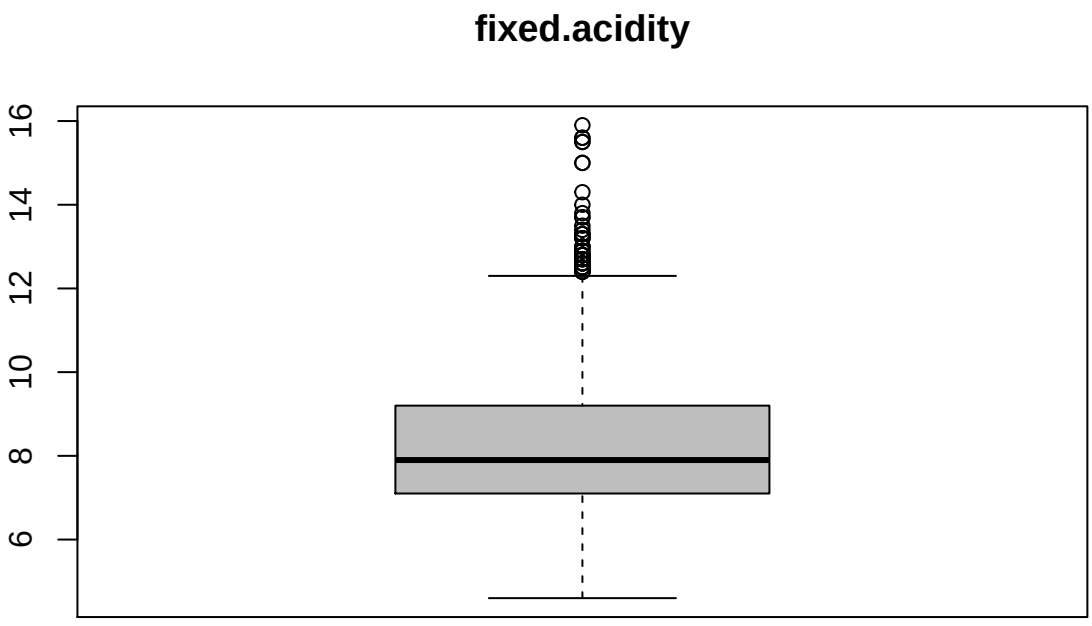
```
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  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
```

```
#ph: correcto (entre 2 y 4)
```

3.2 Valores extremos

Para cada una de las variables observemos si existen valores atípicos:

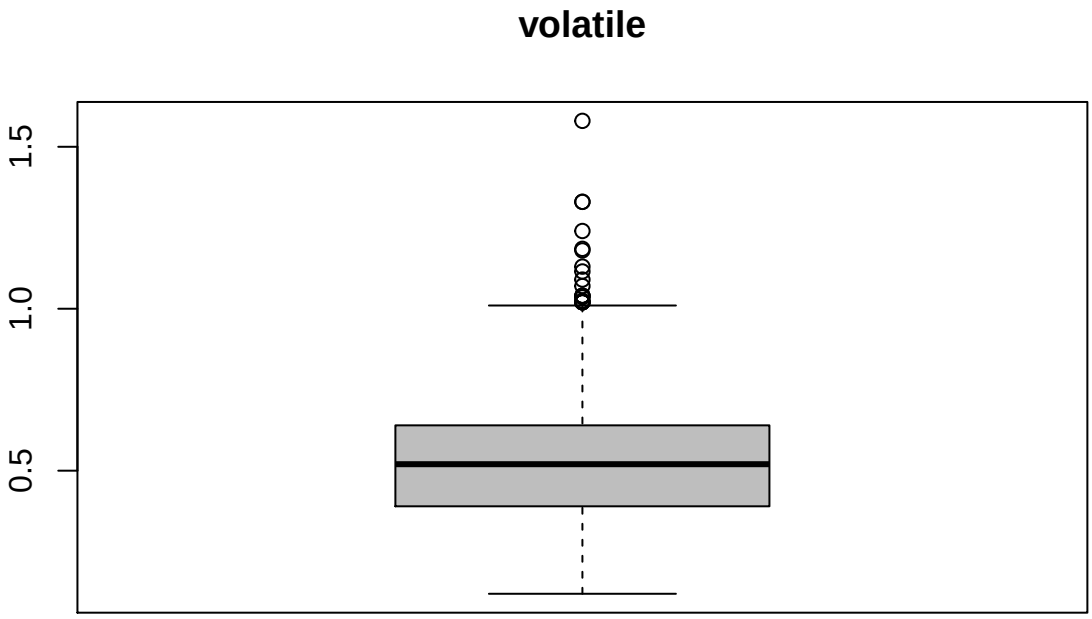
```
boxplot(wine_red$fixed.acidity,main = "fixed.acidity",col="gray")
```




```
boxplot.stats(wine_red$fixed.acidity)$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
```

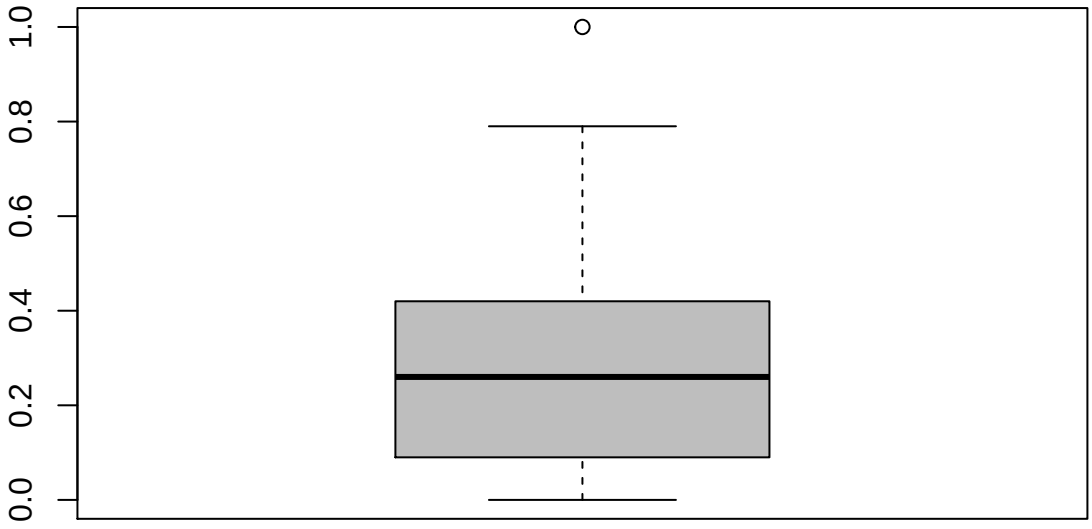
```r
boxplot(wine_red$volatile.acidity,main = "volatile",col="gray")
```




```r
boxplot.stats(wine_red$volatile.acidity)$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
```

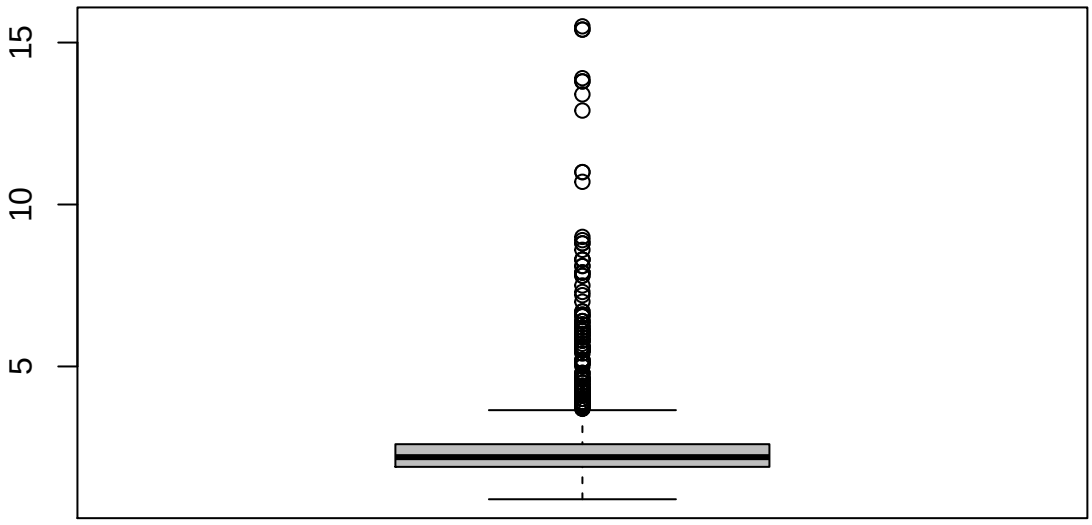
```r
boxplot(wine_red$citric.acid,main = "citric.acid",col="gray")
```




```r
boxplot.stats(wine_red$citric.acid)$out
```

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

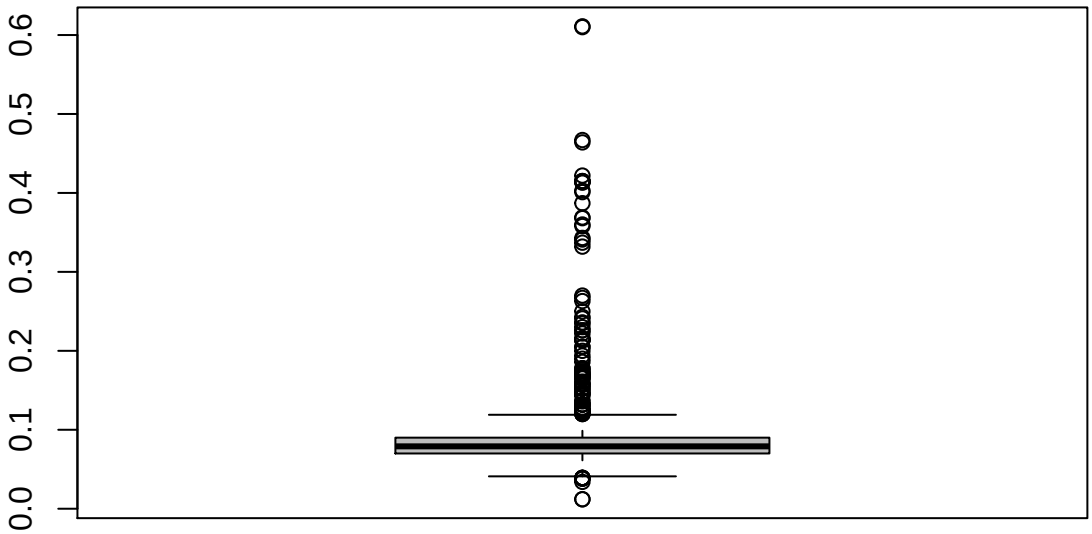
```r
boxplot(wine_red$residual.sugar,main = "residual.sugar",col="gray")
```




```
boxplot.stats(wine_red$residual.sugar)$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
```

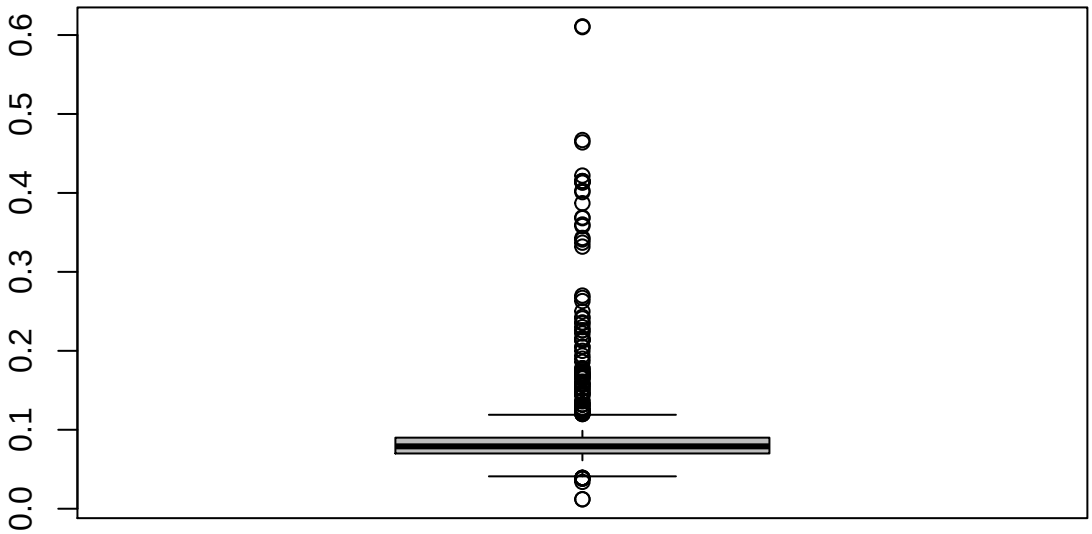
```
boxplot(wine_red$chlorides,main = "chlorides",col="gray")
```




```
boxplot.stats(wine_red$residual.sugar)$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
```

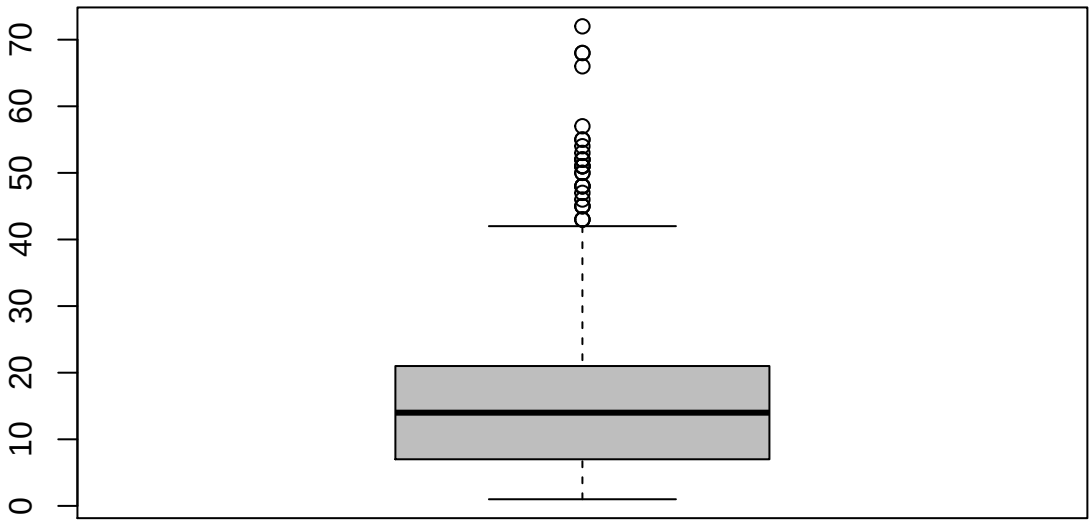
```
boxplot(wine_red$chlorides,main = "residual.sugar",col="gray")
```




```
boxplot.stats(wine_red$residual.sugar)$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
```

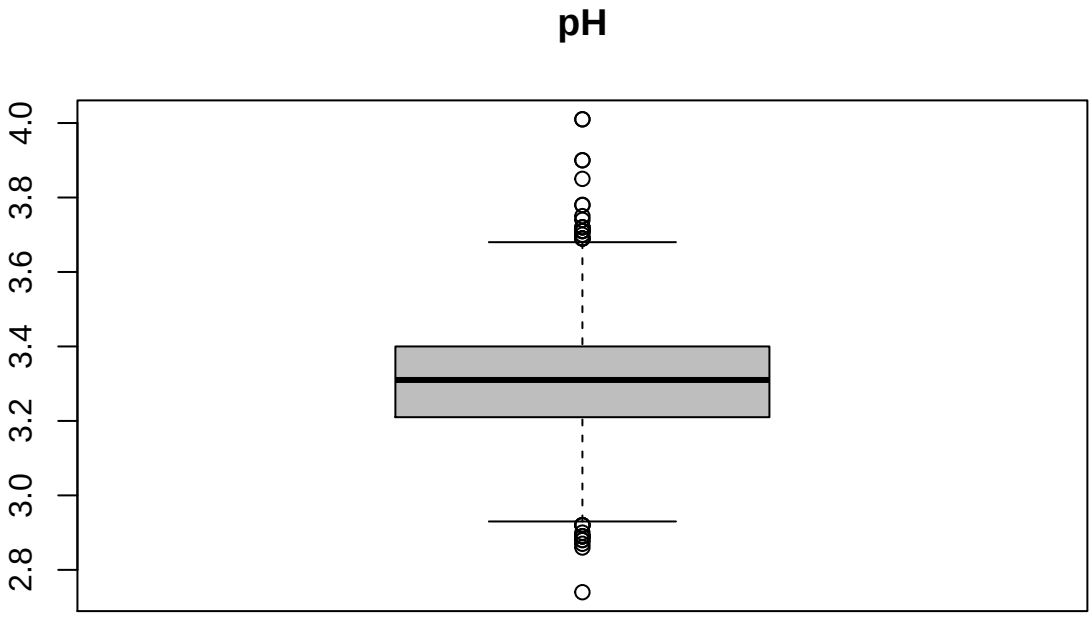
```
boxplot(wine_red$free.sulfur.dioxide,main = "free.sulfur.dioxide",col="gray")
```




```
boxplot.stats(wine_red$free.sulfur.dioxide)$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
```

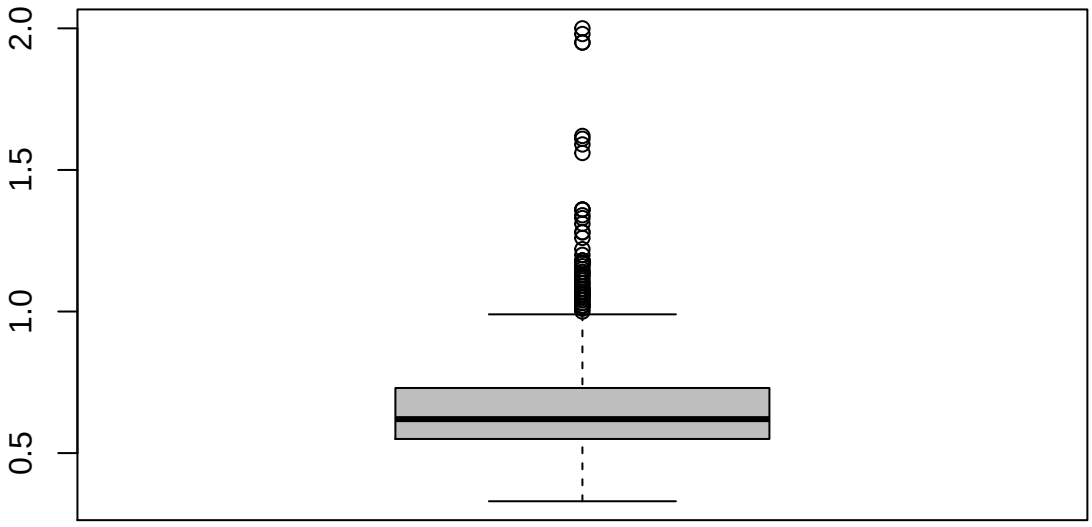
```
boxplot(wine_red$pH,main = "pH",col="gray")
```




```
boxplot.stats(wine_red$pH)$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
```

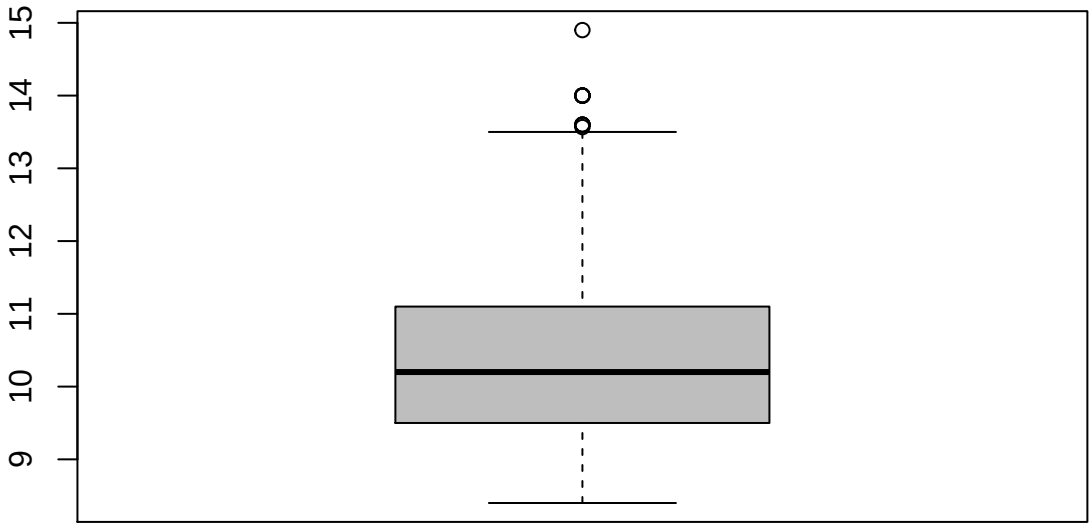
```
#datos correctos porque los valores de pH estan entre 2 y 7

boxplot(wine_red$sulphates,main = "sulphates",col="gray")
```




```
boxplot.stats(wine_red$sulphates)$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
```

```
boxplot(wine_red$ alcohol,main = " alcohol",col="gray")
```




```
boxplot.stats(wine_red$ 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
```

```
#alcohol está dentro de unos rangos adecuados
```

Todas las gráficas presentan outliers pero dado a que, la creación de un vino es puramente una reacción química, se puede decir que los valores que se presentan son posibles y simplemente podemos decir que dichos vinos tienen características muy diferentes al resto.

4. Análisis de datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Vamos a analizar si son significativas cada variable del dataset para ello haremos un test t con un nivel de significación=0.05.

Diferentes hipótesis se pueden establecer cambiando el valor de parámetro alternative en la función t.test

1. **Alternative='two.sided':** Hipótesis nula H0: $\mu 1 = \mu 2$ Hipótesis alternativa H1: $\mu 1 != \mu 2$
2. **Alternative='greater':** Hipótesis nula H0: $\mu 1 = \mu 2$ Hipótesis alternativa H1: $\mu 1 > \mu 2$
3. **Alternative='less':** Hipótesis nula H0: $\mu 1 = \mu 2$ Hipótesis alternativa H1: $\mu 1 < \mu 2$

```
 # A más alcohol, hay más calidad, significativo

high_alcohol<-quantile(wine_red$alcohol, probs =0.75)
wine_red.altoAlcohol<-wine_red[wine_red$alcohol>=high_alcohol,]$quality
wine_red.bajoAlcohol<-wine_red[wine_red$alcohol<high_alcohol,]$quality

t.test(wine_red.altoAlcohol, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoAlcohol and wine_red$quality
## t = 12.555, df = 628.55, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4891266       Inf
## sample estimates:
## mean of x mean of y
##  6.199017  5.636023
```

```
t.test(wine_red.bajoAlcohol, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoAlcohol and wine_red$quality
## t = -6.6585, df = 2711.5, p-value = 1.668e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1447275
## sample estimates:
## mean of x mean of y
##  5.443792  5.636023
```

```r
# A más azucar no hay más calidad y a menos azucar tampoco , no significativo porque no se rechaza la h

high_sugar<-quantile(wine_red$residual.sugar, probs =0.75)
wine_red.altoGradoAzucar<-wine_red[wine_red$residual.sugar>=high_sugar,]$quality
wine_red.bajoGradoAzucar<-wine_red[wine_red$residual.sugar<high_sugar,]$quality
t.test(wine_red.altoGradoAzucar, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoGradoAzucar and wine_red$quality
## t = 0.049196, df = 664.01, p-value = 0.4804
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.07231395         Inf
## sample estimates:
## mean of x mean of y
##  5.638249  5.636023
```

```r
t.test(wine_red.bajoGradoAzucar, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoGradoAzucar and wine_red$quality
## t = -0.026921, df = 2532, p-value = 0.4893
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.04986301
## sample estimates:
## mean of x mean of y
##  5.635193  5.636023
```

```r
#  Volatile.acidity, significativo
#Para un alto valor de acido volátil está claro que el vino no es mejor, es mejor el vino cuanto menos

high_volatile.acidity<-quantile(wine_red$volatile.acidity, probs =0.75)
wine_red.altoVolatile.acidity<-wine_red[wine_red$volatile.acidity>=high_volatile.acidity,]$quality
wine_red.bajoVolatile.acidity<-wine_red[wine_red$volatile.acidity<high_volatile.acidity,]$quality

# Es significativo el nivel de volatile acidity
t.test(wine_red.altoVolatile.acidity, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoVolatile.acidity and wine_red$quality
## t = -8.6351, df = 669.24, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2918063
## sample estimates:
## mean of x mean of y
##  5.275434  5.636023
```

```r
t.test(wine_red.bajoVolatile.acidity, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoVolatile.acidity and wine_red$quality
## t = 3.9702, df = 2595, p-value = 3.689e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.07114568        Inf
## sample estimates:
## mean of x mean of y
##  5.757525  5.636023
```

```r
#sulphates, significativo

high_sulphates<-quantile(wine_red$volatile.acidity, probs =0.75)
wine_red.altoSulphates<-wine_red[wine_red$sulphates>=high_sulphates,]$quality
wine_red.bajoSulphates<-wine_red[wine_red$sulphates<high_sulphates,]$quality

t.test(wine_red.altoSulphates, wine_red$quality, alternative = "greater") # significativo sulfato alto
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoSulphates and wine_red$quality
## t = 8.5186, df = 1382.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2500704       Inf
## sample estimates:
## mean of x mean of y
##  5.945983  5.636023
```

```r
t.test(wine_red.bajoSulphates, wine_red$quality, alternative = "less") #
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoSulphates and wine_red$quality
## t = -8.1554, df = 2011.5, p-value = 3.027e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2036886
## sample estimates:
## mean of x mean of y
##  5.380844  5.636023
```

```r
#PH, no significativo

high_pH<-quantile(wine_red$pH, probs =0.75)
wine_red.altopH<-wine_red[wine_red$pH>=high_pH,]$quality
wine_red.bajopH<-wine_red[wine_red$pH<high_pH,]$quality

t.test(wine_red.altopH, wine_red$quality, alternative = "greater") # NO significativo  PH
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altopH and wine_red$quality
## t = -1.2978, df = 654.48, p-value = 0.9026
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1320493        Inf
## sample estimates:
## mean of x mean of y
##  5.577830  5.636023
```

```
t.test(wine_red.bajopH, wine_red$quality, alternative = "less") # NO significativo  PH
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajopH and wine_red$quality
## t = 0.68001, df = 2541.6, p-value = 0.7517
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.07181025
## sample estimates:
## mean of x mean of y
##  5.657021  5.636023
```

```
# Citric acid es significativo
high_citric_acid<-quantile(wine_red$citric.acid, probs =0.75)
wine_red.altoCitricAcid<-wine_red[wine_red$citric.acid>=high_citric_acid,]$quality
wine_red.bajoCitricAcid<-wine_red[wine_red$citric.acid<high_citric_acid,]$quality

t.test(wine_red.altoCitricAcid, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoCitricAcid and wine_red$quality
## t = 5.371, df = 644.36, p-value = 5.475e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1711653       Inf
## sample estimates:
## mean of x mean of y
##  5.882904  5.636023
```

```
t.test(wine_red.bajoCitricAcid, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoCitricAcid and wine_red$quality
## t = -2.973, df = 2585.7, p-value = 0.001488
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.04016442
## sample estimates:
```

```
## mean of x mean of y
##  5.546075  5.636023
```

```
# fixed acidity es significativo
high_fixed_acidity<-quantile(wine_red$fixed.acidity, probs =0.75)
wine_red.altoFixedAcidity<-wine_red[wine_red$fixed.acidity>=high_fixed_acidity,]$quality
wine_red.bajoFixedAcidity<-wine_red[wine_red$fixed.acidity<high_fixed_acidity,]$quality

t.test(wine_red.altoFixedAcidity, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoFixedAcidity and wine_red$quality
## t = 3.8769, df = 618.02, p-value = 5.857e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.1032353       Inf
## sample estimates:
## mean of x mean of y
##  5.815534  5.636023
```

```
t.test(wine_red.bajoFixedAcidity, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoFixedAcidity and wine_red$quality
## t = -2.0469, df = 2594, p-value = 0.02039
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01221934
## sample estimates:
## mean of x mean of y
##  5.573715  5.636023
```

```
#Chlorides, significativo, a más chlorides peor calidad

high_chlorides<-quantile(wine_red$chlorides, probs =0.75)
wine_red.altoChlorides<-wine_red[wine_red$chlorides>=high_chlorides,]$quality
wine_red.bajoChlorides<-wine_red[wine_red$chlorides<high_chlorides,]$quality

t.test(wine_red.altoChlorides, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoChlorides and wine_red$quality
## t = -2.9858, df = 674.92, p-value = 0.001466
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.05722153
## sample estimates:
## mean of x mean of y
##  5.508393  5.636023
```

```r
t.test(wine_red.bajoChlorides, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoChlorides and wine_red$quality
## t = 1.4446, df = 2529.7, p-value = 0.07434
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.006259493          Inf
## sample estimates:
## mean of x mean of y
##  5.681049  5.636023
```

```r
#free sulfur dioxide, no significativo

high_free_sulfur<-quantile(wine_red$free.sulfur.dioxide, probs =0.75)
wine_red.altoFreeSulfur<-wine_red[wine_red$free.sulfur.dioxide>=high_free_sulfur,]$quality
wine_red.bajoFreeSulfur<-wine_red[wine_red$free.sulfur.dioxide<high_free_sulfur,]$quality

t.test(wine_red.altoFreeSulfur, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoFreeSulfur and wine_red$quality
## t = -1.5957, df = 758.2, p-value = 0.9445
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.130598       Inf
## sample estimates:
## mean of x mean of y
##  5.571754  5.636023
```

```r
t.test(wine_red.bajoFreeSulfur, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoFreeSulfur and wine_red$quality
## t = 0.76627, df = 2449.8, p-value = 0.7782
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.07655152
## sample estimates:
## mean of x mean of y
##  5.660345  5.636023
```

```r
# total sulfur dioxide,  significativo, a más cantidad de  total sulfur peor calidad

high_total_sulfur<-quantile(wine_red$total.sulfur.dioxide, probs =0.75)
wine_red.altoTotalSulfur<-wine_red[wine_red$total.sulfur.dioxide>=high_total_sulfur,]$quality
wine_red.bajoTotalSulfur<-wine_red[wine_red$total.sulfur.dioxide<high_total_sulfur,]$quality

t.test(wine_red.altoTotalSulfur, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoTotalSulfur and wine_red$quality
## t = -6.677, df = 752.79, p-value = 2.368e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1922452
## sample estimates:
## mean of x mean of y
##  5.380835  5.636023
```

```
t.test(wine_red.bajoTotalSulfur, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoTotalSulfur and wine_red$quality
## t = 2.7632, df = 2516.4, p-value = 0.002883
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.03524473        Inf
## sample estimates:
## mean of x mean of y
##  5.723154  5.636023
```

```
# Density, significativo a mayor densidad peor calidad

high_density<-quantile(wine_red$density, probs =0.75)
wine_red.altoDensity<-wine_red[wine_red$density>=high_density,]$quality
wine_red.bajoDensity<-wine_red[wine_red$density<high_density,]$quality

t.test(wine_red.altoDensity, wine_red$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.altoDensity and wine_red$quality
## t = -2.0039, df = 645.06, p-value = 0.02275
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01531291
## sample estimates:
## mean of x mean of y
##  5.550000  5.636023
```

```
t.test(wine_red.bajoDensity, wine_red$quality, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_red.bajoDensity and wine_red$quality
## t = 0.92086, df = 2556.1, p-value = 0.1786
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.0225815        Inf
```

```
## sample estimates:
## mean of x mean of y
##  5.664721  5.636023
```

Las variables significativas son : alcohol , volatile.acidity , sulphates , citric.acid , fixed.acidity , chlorides , total.sulfur.dioxide , density

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Comprobar que variables siguen distribución normal

Aplicando la siguiente función podemos comprobar si siguen una distribución normal o si por el contrario no lo hacen.

```
#install.packages("nortest")
library(nortest)

alpha = 0.05
col.names = colnames(wine_red)
for (i in 1:ncol(wine_red)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(wine_red[,i]) | is.numeric(wine_red[,i])) {
      p_val = ad.test(wine_red[,i])$p.value
        if (p_val < alpha) {
            cat(col.names[i])
                  # Format output
            if (i < ncol(wine_red) - 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, alcoholquality
```

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.

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre la calidad del vino:

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa # con respecto al campo "preci
for (i in 1:(ncol(wine_red) - 1)) {
if (is.integer(wine_red[,i]) | is.numeric(wine_red[,i]))
    {
    spearman_test = cor.test(wine_red[,i],wine_red[,length(wine_red)],method = "spearman")
      corr_coef = spearman_test$estimate
      p_val = spearman_test$p.value
```

```
      # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(wine_red)[i]
  }
}
```

```
## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties

## Warning in cor.test.default(wine_red[, i], wine_red[, length(wine_red)], :
## Cannot compute exact p-value with ties
```

```
a <- corr_matrix[,'p-value']
corr_matrix[order(a),]
```

```
##                         estimate      p-value
## alcohol               0.47853169 2.726838e-92
## volatile.acidity     -0.38064651 2.734944e-56
## sulphates             0.37706020 3.477695e-55
## citric.acid           0.21348091 6.158952e-18
## total.sulfur.dioxide -0.19673508 2.046488e-15
## chlorides            -0.18992234 1.882858e-14
## density              -0.17707407 9.918139e-13
## fixed.acidity         0.11408367 4.801220e-06
## free.sulfur.dioxide  -0.05690065 2.288322e-02
```

```
## pH                   -0.04367193 8.084594e-02
## residual.sugar        0.03204817 2.002454e-01
```

Así, identificamos cuáles son las variables más correlacionadas con la calidad según su proximidad con los valores -1 y +1.

Teniendo esto en cuenta, queda patente cómo la variable más relevante para **la calidad** es la variable `alcohol`. Pero en términos generales podemos decir que los valores que obtenemos son bastante modestos y no sería adecuado obtener ninguna conclusión por ahora. Lo único que podemos hacer es utilizar estos valores como tendencias.

Nota. Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

Modelo de regresión lineal logística

Tal y como se planteó en los objetivos de la actividad, resultará de mucho interés poder realizar predicciones sobre la calidad del vino dadas sus características. Así, se calculará un modelo de regresión lógistica utilizando regresores tanto cuantitativos como cualitativos con el que poder realizar las predicciones sobre la calidad.

Se usa regresión logística ya que nuestro modelo predicirá si el vino es de calidad o no (variable dicotómica).

Para obtener un modelo de regresión logística considerablemente eficiente, lo que haremos será obtener varios modelos de regresión utilizando las variables que estén más correladas con respecto a la calidad.

```
# Regresores cuantitativos con mayor coeficiente  de correlación con respecto al precio
alcohol = wine_red$alcohol
volatile.acidity = wine_red$volatile.acidity
sulphates=wine_red$sulphates
citric.acid=wine_red$citric.acid
total.sulfur.dioxide=wine_red$total.sulfur.dioxide
chlorides=wine_red$chlorides
density=wine_red$density
fixed.acidity=wine_red$fixed.acidity
free.sulfur.dioxide=wine_red$free.sulfur.dioxide
pH=wine_red$pH

good_wine <-ifelse(test=wine_red$quality>=7,yes=1,no=0)
wine_red$good_wine=good_wine
quality<- wine_red$quality

library(InformationValue)
# El modelo 1 usa las variables que hemos detectado como significativas según el test t para un valor s
#El modelo 2 usa todas las variables
GLM.1 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + citric.acid + fixed.acidity
AIC_GLM1<-summary(GLM.1)$aic

GLM.2 <- glm( wine_red$good_wine ~ . -quality , family=binomial(logit),data=wine_red)
AIC_GLM2<-summary(GLM.2)$aic

AIC_GLM1
```

```
## [1] 898.8508
```

```
AIC_GLM2
```

```
## [1] 894.8644
```

```
p_model1 <- predict(GLM.1, type = 'response')
p_model2 <- predict(GLM.2, type = 'response')
wine_red$prob<-p_model1
wine_red$prob2<-p_model2

confusionMatrix(wine_red$good_wine,p_model1,0.7)
```

```
##      0   1
## 0 1372 202
## 1   10  15
```

```
confusionMatrix(wine_red$good_wine,p_model2,0.7)
```

```
##      0   1
## 0 1372 196
## 1   10  21
```

El mejor modelo, el número 2 que usa todas las variables, aunque no mejora demasiado respecto al modelo 1: AIC_GLM1=898.8508 AIC_GLM2=894.8644

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

Calculo curva ROC

```
library(pROC)
```

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

```
##
## Attaching package: 'pROC'
```

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

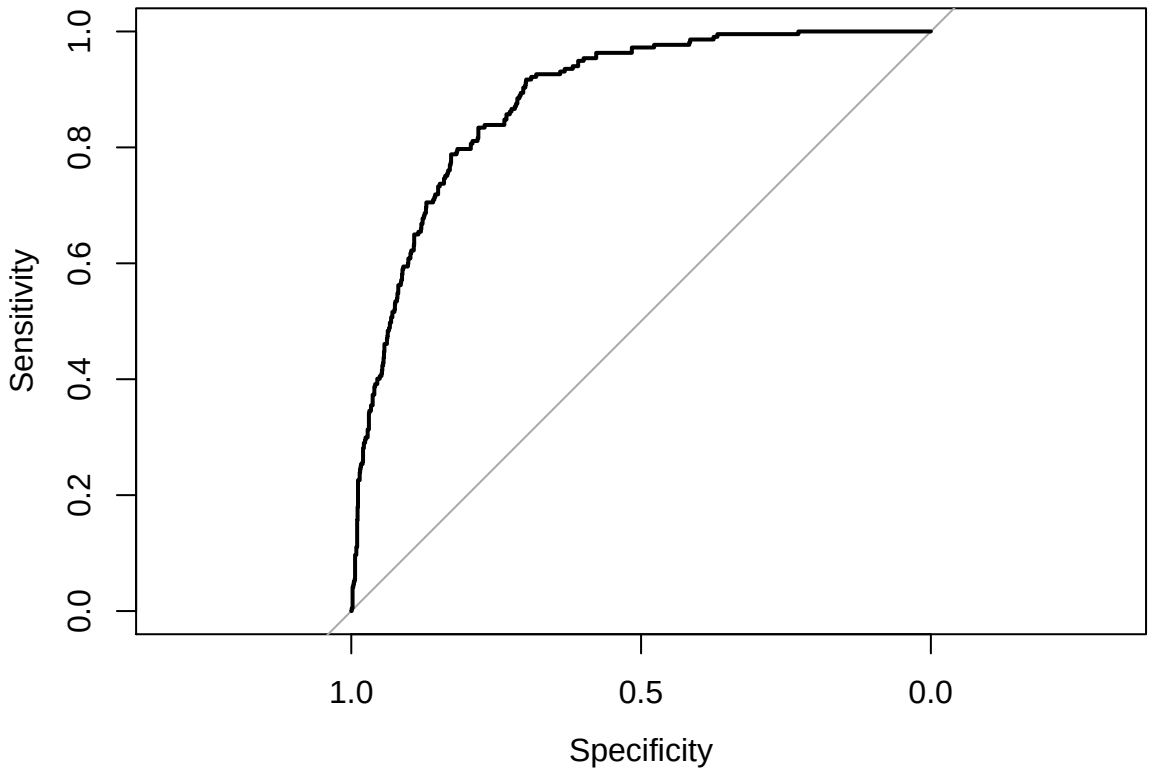
```
#Cáculo curva
g<-roc(wine_red$good_wine,wine_red$prob2)
g
```

```
##
## Call:
## roc.default(response = wine_red$good_wine, predictor = wine_red$prob2)
##
## Data: wine_red$prob2 in 1382 controls (wine_red$good_wine 0) < 217 cases (wine_red$good_wine 1).
## Area under the curve: 0.8822
```

```
#Área bajo la curva
auc(g)
```

```
## Area under the curve: 0.8822
```

```
plot(g)
```


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity