Tipología y ciclo de vida de los datos: Práctica2

Antonio Guzmán Martín

Diciembre 2017

Contents


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

#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 podemos prescindir del 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). **DUDA:¿Siempre hay que hacer caso al estudio de correlación (cercania -1 y +1)? Resulta curioso como en un primer momento las variables que creía no importantes resultan ser las que más correlación guardan con la calidad para esta muestra. Inicialmente creía que el alcohol o la densidad no eran importantes, basándome en la búsqueda que realicé en internet: https://www.vinopack.es/criterios-que-determinan-la-calidad-en-el-vino.**

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
```

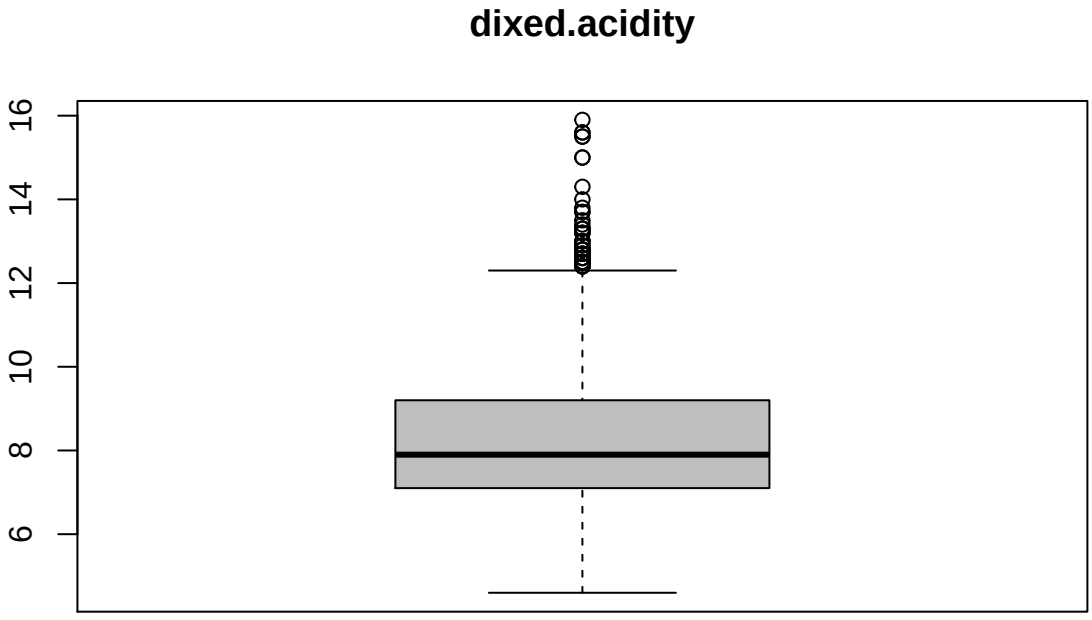
```
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd 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
```

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

3.1 Valores extremos

DUDA: ¿Es necesario justificar para cada variable, la razón de porque los valores que resultan ser atípicos son en realidad legítimos? Por ejemplo: No se que rango es el adecuado para fixed.acidity

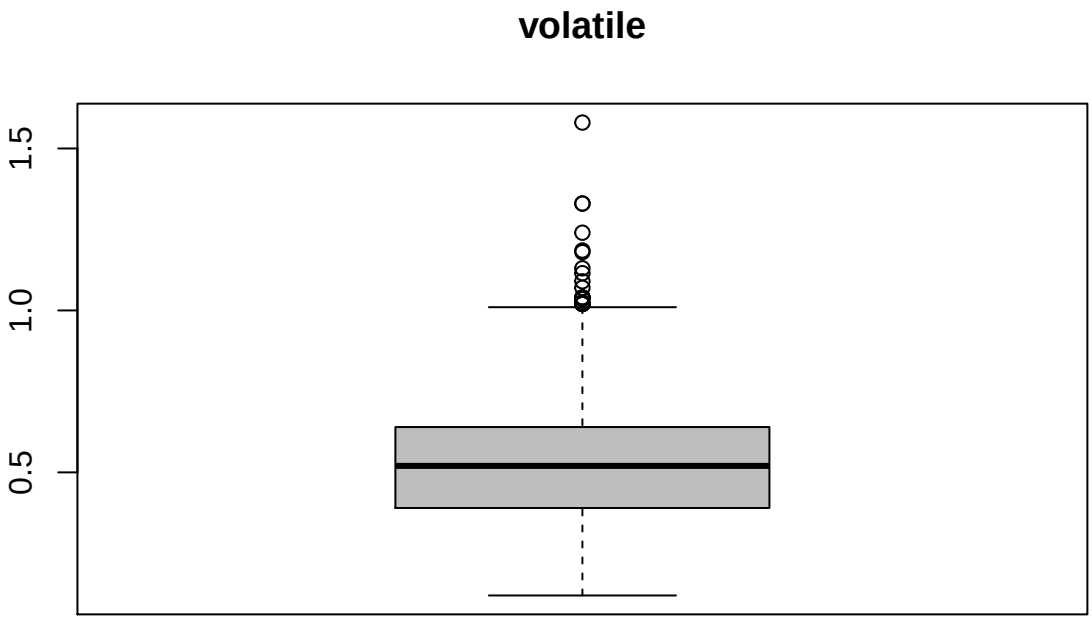
```r
# Para cada una de las variables observemos si existen valores atípicos

boxplot(wine_red$fixed.acidity,main = "dixed.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
```

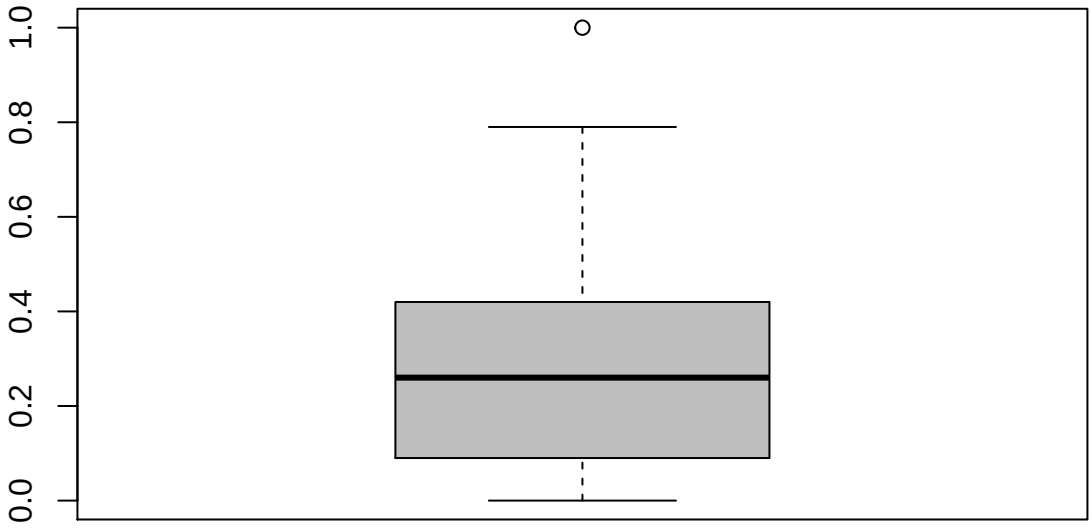
```
boxplot(wine_red$volatile.acidity,main = "volatile",col="gray")
```




```
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
```

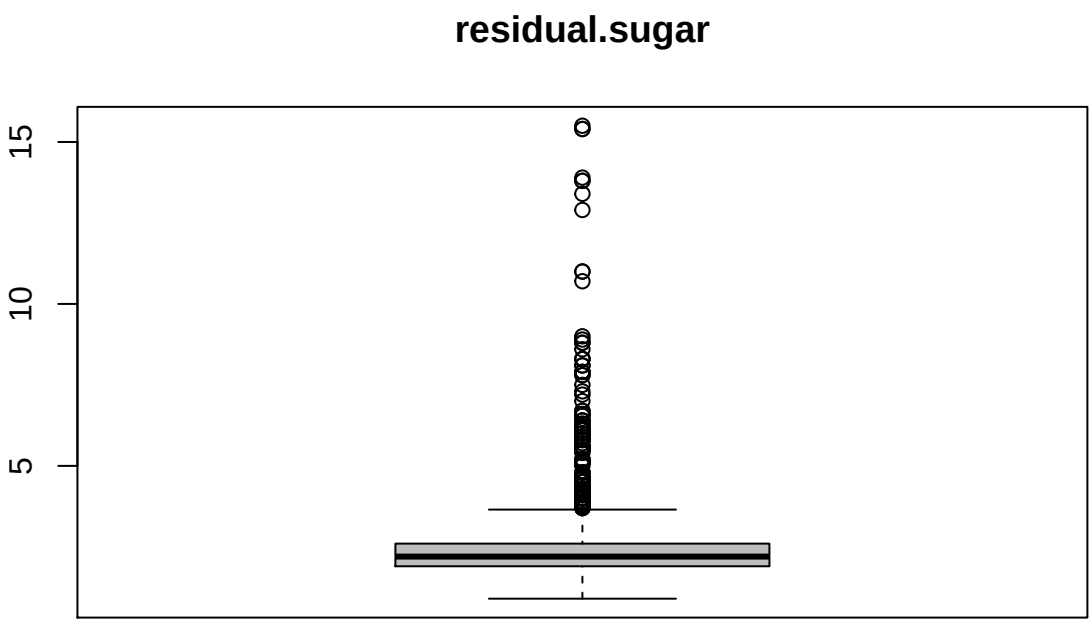
```
boxplot(wine_red$citric.acid,main = "citric.acid",col="gray")
```



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

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

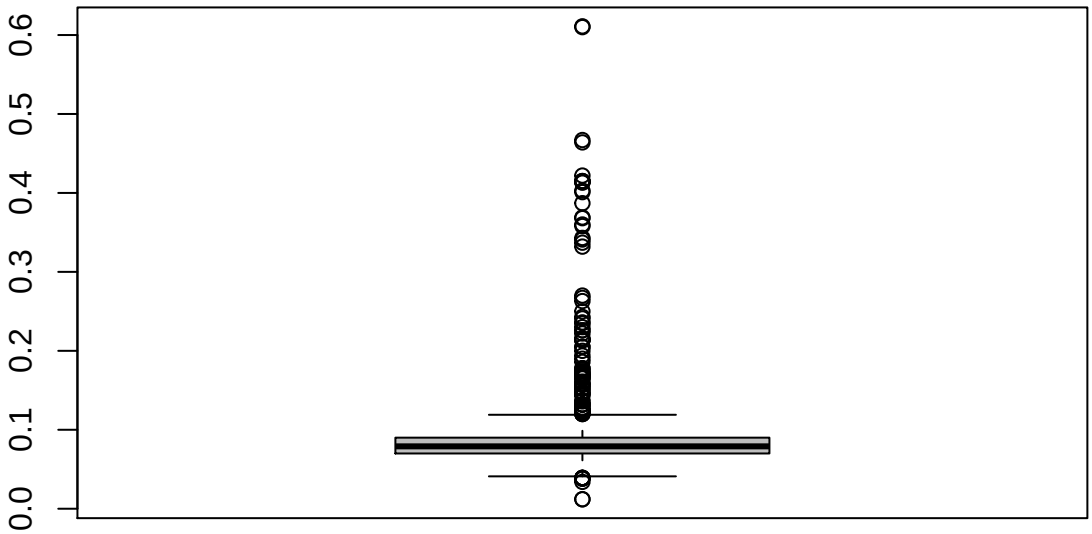
```r
boxplot(wine_red$residual.sugar,main = "residual.sugar",col="gray")
```




```r
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
```

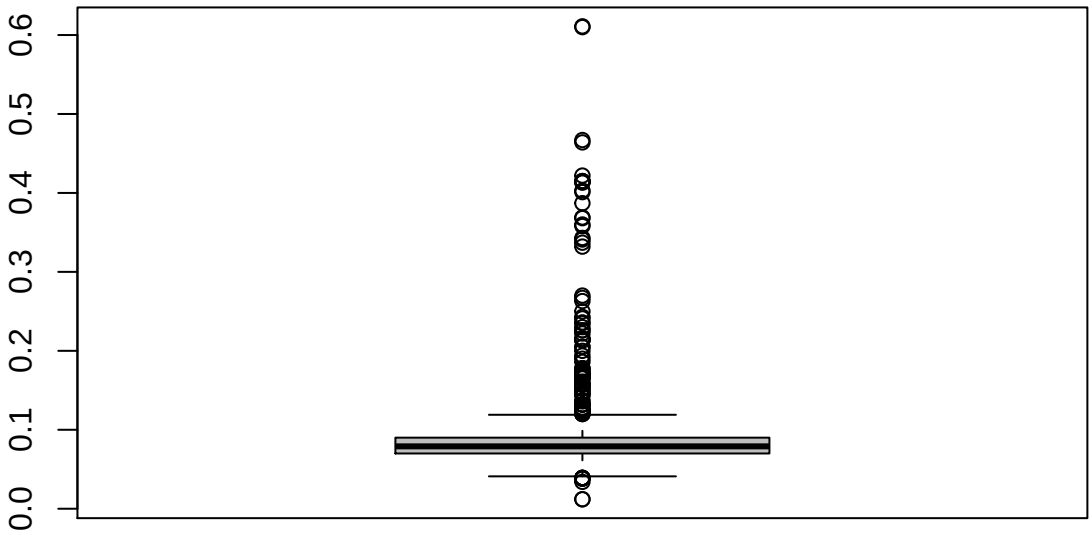
```r
boxplot(wine_red$chlorides,main = "chlorides",col="gray")
```




```
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
```

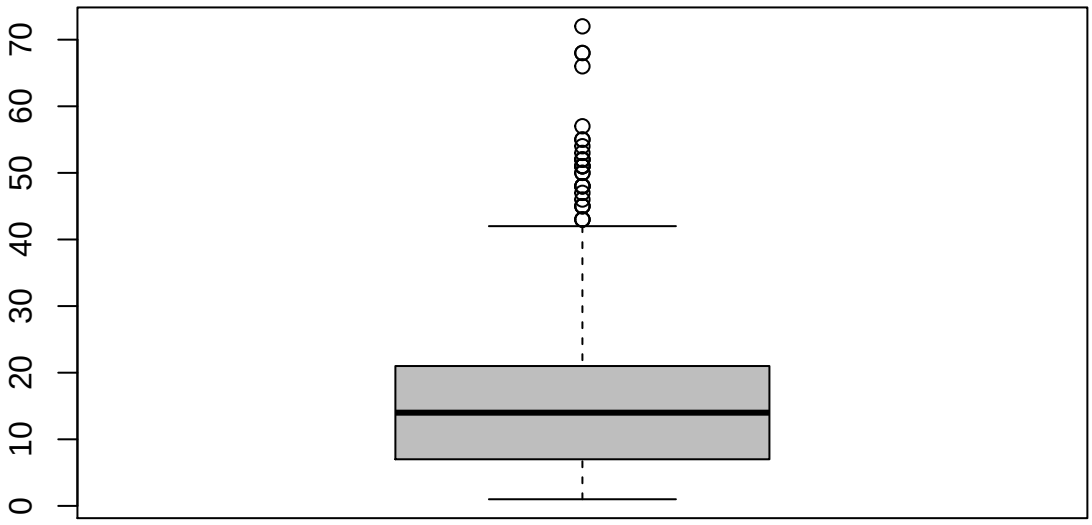
```
boxplot(wine_red$chlorides,main = "residual.sugar",col="gray")
```




```
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
```

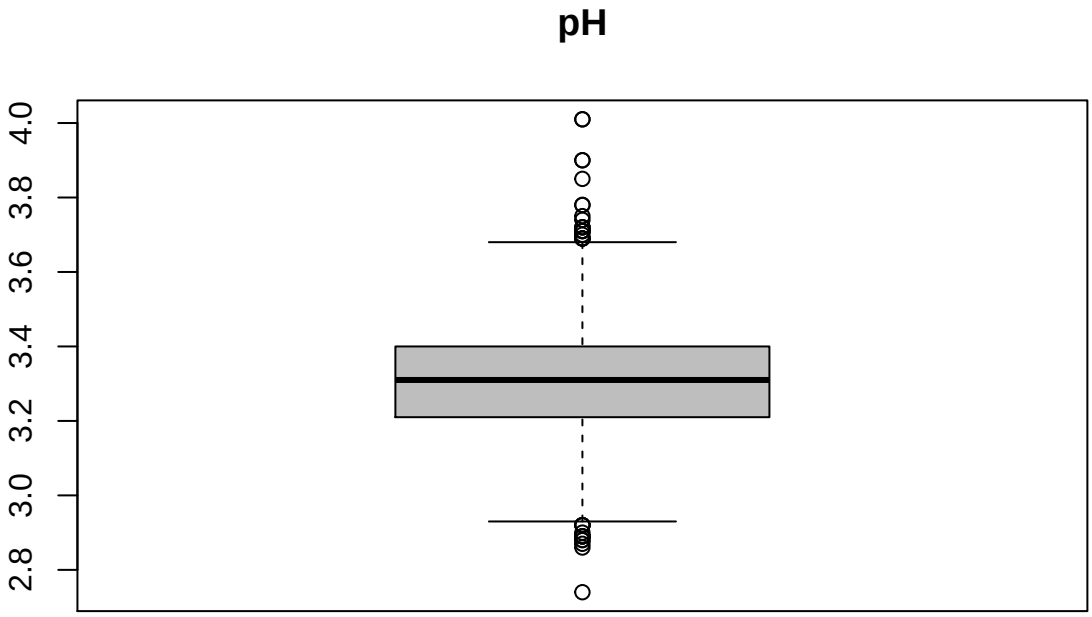
```
boxplot(wine_red$free.sulfur.dioxide,main = "free.sulfur.dioxide",col="gray")
```




```
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
```

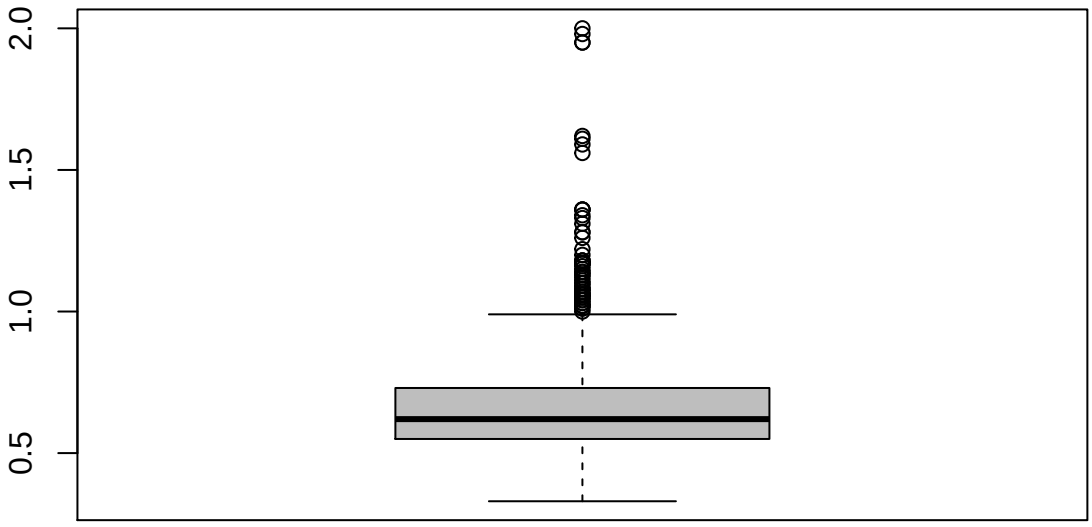
```
boxplot(wine_red$pH,main = "pH",col="gray")
```




```
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
```

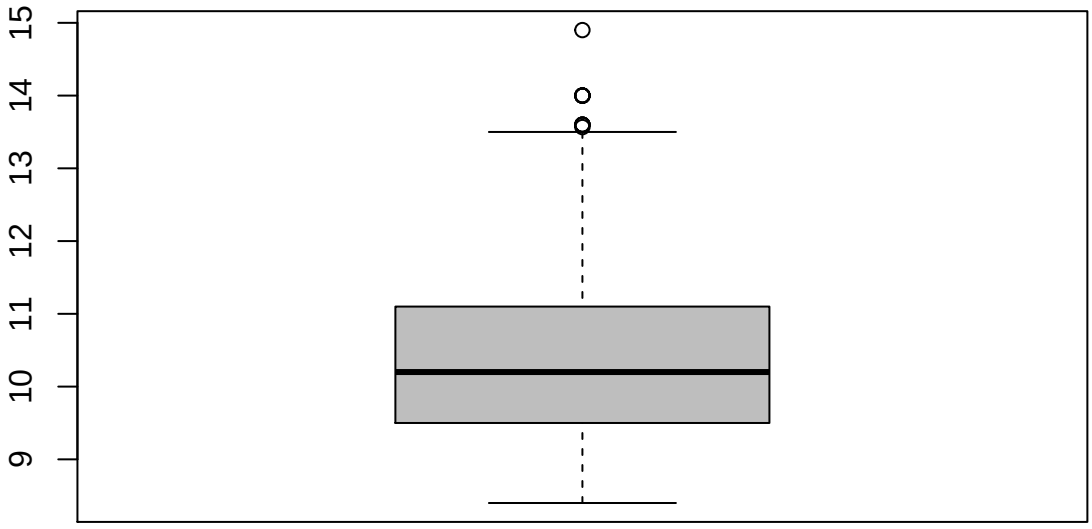
```
#datos correctos porque los valores de pH estan entre 2 y 7

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
```

4. Análisis de datos.

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

Comprobar que variables siguen distribución normal

DUDA: Todas las variables NO siguen distribución normal. ¿Debo normalizarlas?¿Siguen otro tipo de distribución y por tanto tengo que seguir otros pasos? ¿Como identifico que distribución sigue?

```
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
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza

DUDA: ¿Cual es la variable que se escoge para aplicar el test de Fligner-Killen?¿ Es cualquiera que sigue una distribución normal?

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
```

```r
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
```

```r
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.

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.

DUDA: ¿Es buen criterio plantear hipótesis sobre la variable "aparentemente" más correlada con la variable dependiente?

¿ A más alcohol, mejor es la calidad del vino?

```
high_alcohol<-quantile(wine_red$alcohol, probs =0.75)

quality_alcohol<-wine_red[wine_red$alcohol>=high_alcohol,]$quality

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

```
##
##  Welch Two Sample t-test
##
## data:  quality_alcohol 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
```

```
 #a mas alcohol por lo general más calidad
```

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.

DUDA: ¿Cómo saber cuantas variables coger: las 3, 4 o 5 primeras?¿Añado otra más que no esté en ese TOP? ¿Cómo averiguar si las variables están relacionadas y por tanto no debo añadir ambas? ¿Solo tengo como herramienta la información que tengo de las variables?

```r
# 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

GLM.1 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + fixed.acidity, family=binom
#summary(GLM.1)
AIC_GLM1<-summary(GLM.1)$aic
#exp(coef(GLM.1))

GLM.2 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + citric.acid, family=binomial
#summary(GLM.2)
AIC_GLM2<-summary(GLM.2)$aic
#exp(coef(GLM.2))


GLM.3 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + density, family=binomial(lo
AIC_GLM3<-summary(GLM.3)$aic


GLM.4 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide, family
AIC_GLM4<-summary(GLM.4)$aic


GLM.5 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + free.sulfur.dioxide  , famil
AIC_GLM5<-summary(GLM.5)$aic

GLM.6 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + chlorides  , family=binomial
AIC_GLM6<-summary(GLM.6)$aic

GLM.7 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + pH  , family=binomial(logit)
AIC_GLM7<-summary(GLM.7)$aic

# DUDA: ¿¿A MÁS VARIABLES ES MEJOR??
GLM.8 <- glm( wine_red$good_wine ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + citri
AIC_GLM8<-summary(GLM.8)$aic

AIC_GLM1
```

```
## [1] 915.7066
```

```r
AIC_GLM2
```

```
## [1] 924.1143
```

```
AIC_GLM3
```

```
## [1] 924.824
```

```
AIC_GLM4
```

```
## [1] 909.5529
```

```
AIC_GLM5
```

```
## [1] 920.4162
```

```
AIC_GLM6
```

```
## [1] 921.2041
```

```
AIC_GLM7
```

```
## [1] 921.6982
```

```
AIC_GLM8
```

```
## [1] 911.1082
```

```
p_model8 <- predict(GLM.8, type = 'response')
p_model4 <- predict(GLM.4, type = 'response')
summary(p_model4)
```

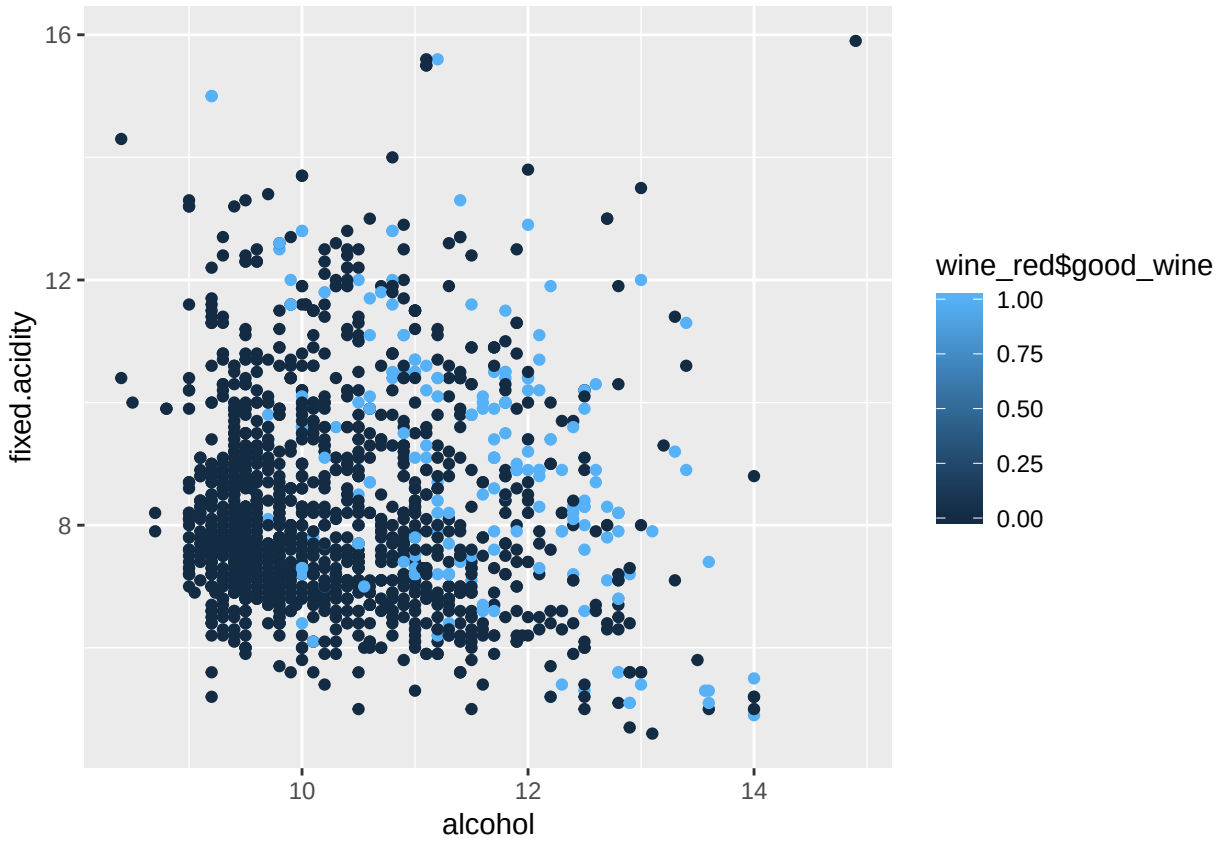
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000257 0.017740 0.052318 0.135710 0.183642 0.930569
```

```
wine_red$prob<-p_model4
wine_red$prob2<-p_model8
pred <-ifelse(test=wine_red$prob>0.7,yes=1,no=0)
pred2 <-ifelse(test=wine_red$prob2>0.7,yes=1,no=0)
wine_red$pred=pred
wine_red$pred2=pred2
```

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

Se observa una clara tendencia a que a más alcohol, mejor es la calidad

```
library(ggplot2)
ggplot(wine_red, aes(x = alcohol, y = fixed.acidity, color = wine_red$good_wine)) + geom_point()
```




```
aux=summary(wine_red$alcohol)
```

Calculo curva ROC

El mejor modelo, el número 4 tan solo me da un AUC=0.5451, es bastante malo. **¿DUDA: A partir de que AUC considero un modelo aceptable? ¿mayor que 88%?**

```
library(pROC)
```

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

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

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

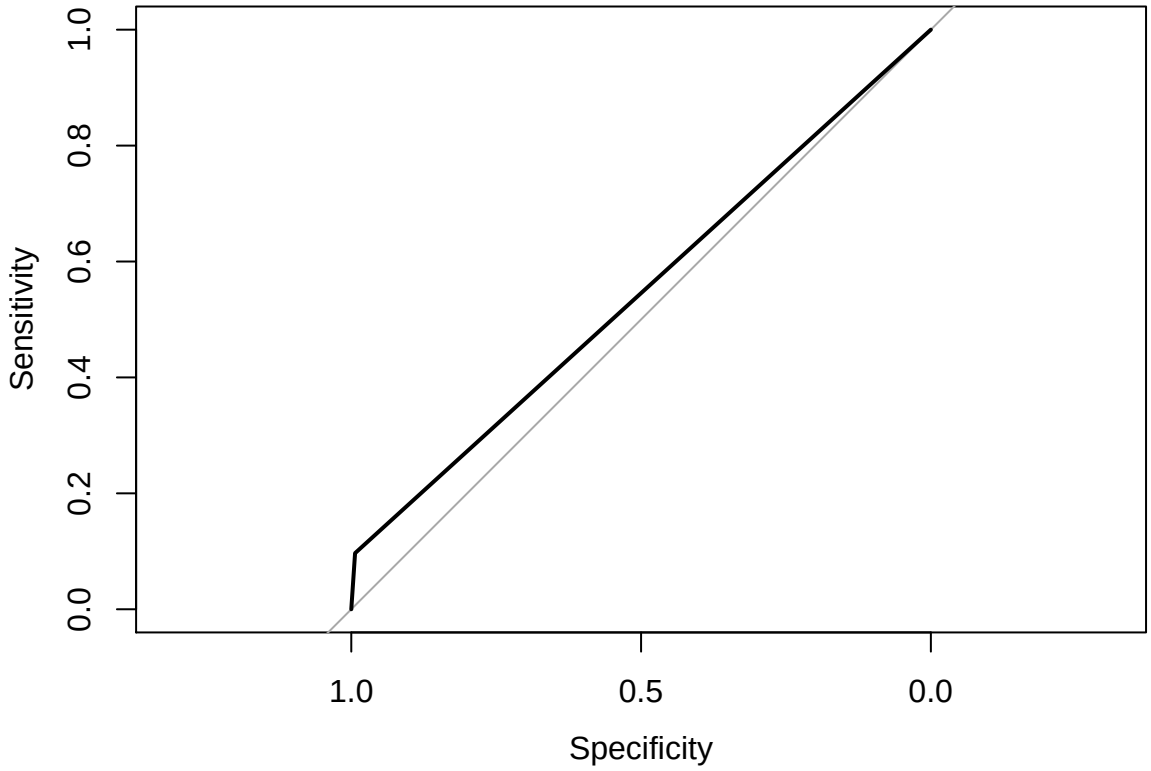
```
#Cáculo curva
g<-roc(wine_red$good_wine,wine_red$pred)
g
```

```
##
## Call:
## roc.default(response = wine_red$good_wine, predictor = wine_red$pred)
##
## Data: wine_red$pred in 1382 controls (wine_red$good_wine 0) < 217 cases (wine_red$good_wine 1).
## Area under the curve: 0.5451
```

```r
#Área bajo la curva
auc(g)
```

```
## Area under the curve: 0.5451
```

```r
plot(g)
```



6. Conclusiones.