

PRACTICA2TIPOLOGIA

Jorge Arias Martín

9/6/2018

PRACTICA 2 TIPOLOGIA DE DATOS

Jorge Arias Martín

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

El dataset proporcionado en la practica es el Red Wine Quality <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009> (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>)

El Dataset contienen los valores fisicoquímicos y sensoriales de las variantes rojas del vino portugués “Vinho Verde”. Para más detalles, consulte la referencia [Cortez et al., 2009]. No hay datos sobre los tipos de uva, la marca del vino, el precio de venta del vino, etc.).

Los conjuntos de datos se pueden tomar como tareas de Regresión Lineal Las clases están ordenadas y no equilibradas (por ejemplo, hay muchos más vinos normales que excelentes o malos).

Pretendo determinar qué propiedades fisicoquímicas hacen que un vino sea clasificado como “bueno”

Integración y selección de los datos de interés a analizar.

Los datos proporcionados están disponibles en CSV; el fichero proporciona en la primera linea, el nombre de los campos para facilitar la tarea de clasificación y dispone de un total de 1599 lineas de datos Las columnas disponibles son las siguientes:

fixed acidity (fisicoquímico) volatile acidity (fisicoquímico) citric acid (fisicoquímico) residual sugar (fisicoquímico) chlorides (fisicoquímico) free sulfur dioxide (fisicoquímico) total sulfur dioxide (fisicoquímico) density (fisicoquímico) pH (fisicoquímico) sulphates (fisicoquímico) alcohol (fisicoquímico) quality (subjetivo)

Es importante destacar que las primeras 11 columnas son datos objetivos obtenidos a través de métodos científicos mientras que la última columna, se trata de un dato subjetivo y es obtenido por un metodo desconocido, clasificando entre 3 y 8 la calidad del vino.

Procedo a la carga del CSV con los siguientes parámetros generando un dataframe denominado winequality

```
library(readr)
winequality <- read_csv("~/Downloads/winequality-red.csv")
```

```
## Parsed with column specification:
## cols(
##   `fixed acidity` = col_double(),
##   `volatile acidity` = col_double(),
##   `citric acid` = col_double(),
##   `residual sugar` = col_double(),
##   chlorides = col_double(),
##   `free sulfur dioxide` = col_double(),
##   `total sulfur dioxide` = col_double(),
##   density = col_double(),
##   pH = col_double(),
##   sulphates = col_double(),
##   alcohol = col_double(),
##   quality = col_integer()
## )
```

```
#View(winequality)
```

Para comprobar la calidad de los datos, realizo un Summary para ver el aspecto que tiene la estadística de los datos

```
summary(winequality)
```

```
## 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
```

Podemos observar información interesante en los datos proporcionados por el summary. En primer lugar observamos que los máximos y mínimos de las 12 columnas del dataframe contienen datos numéricos

```
str(winequality)
```

```
## Classes 'tbl_df', 'tbl' and '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 0.07
3 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 ...
## - attr(*, "spec")=List of 2
## ..$ cols      :List of 12
## .. ..$ fixed acidity      : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ volatile acidity   : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ citric acid        : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ residual sugar     : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ chlorides          : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ free sulfur dioxide : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ total sulfur dioxide: list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ density            : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ pH                 : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ sulphates          : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ alcohol            : list()
## .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
## .. ..$ quality            : list()
## .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
## ..$ default: list()
## .. ..- attr(*, "class")= chr  "collector_guess" "collector"
## ..- attr(*, "class")= chr "col_spec"
```

Observamos que hay 1599 observaciones y 12 variables en el dataframe Las primeras columnas son numéricas y la última contiene valores enteros En concreto la columna quality es la variable objetivo del estudio; para ver detalles de la misma utilizamos el comando table

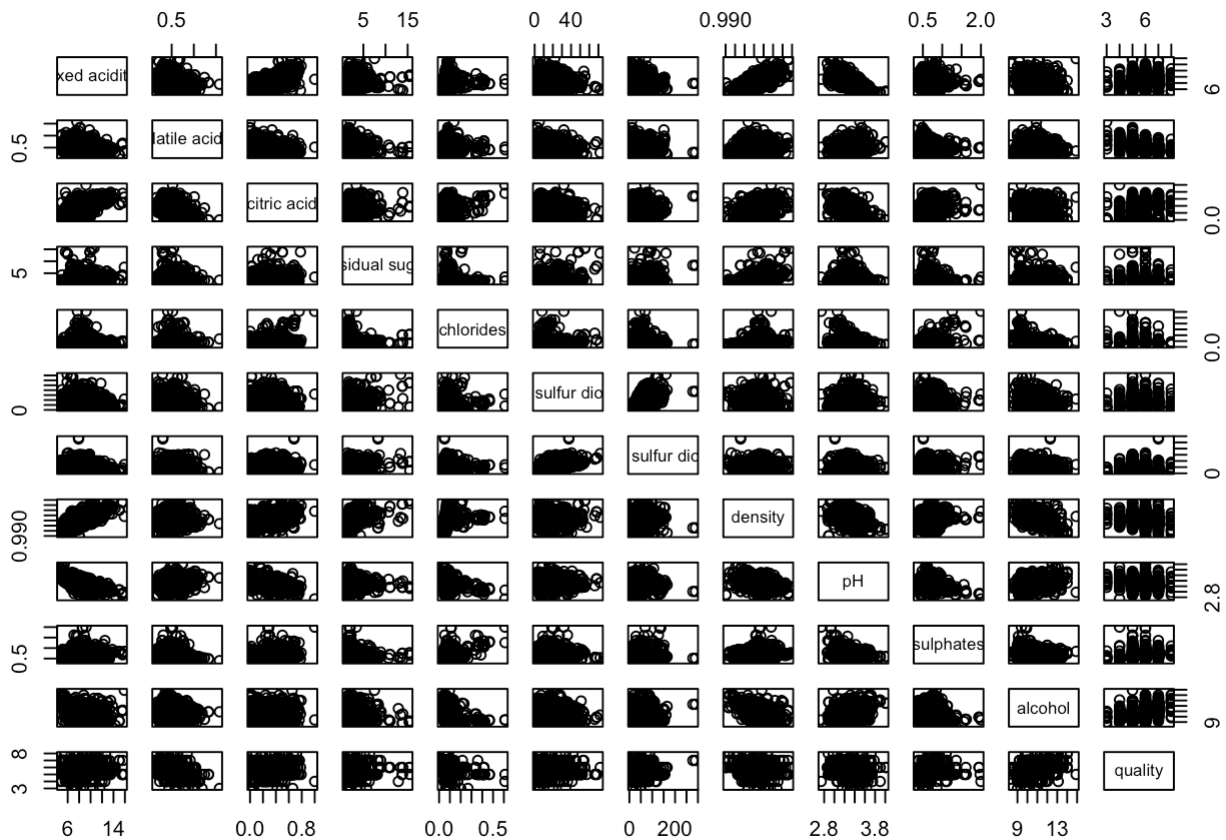
```
table(winequality$quality)
```

```
##
##    3    4    5    6    7    8
##  10   53 681 638 199   18
```

Observamos una concentración en los valores 5 y 6 de calidad; con valores entre 3 y 8, en el summary, vemos como valor medio 5.636

Procedemos


```
plot(winequality)
```











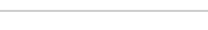


como resumen del análisis utilizamos la función skimr

```
skim(winequality)
```

```
## Skim summary statistics
## n obs: 1599
## n variables: 12
##
## — Variable type:integer —
```

```
## variable missing complete      n mean   sd p0 p25 p50 p75 p100      hist
## quality          0        1599 1599 5.64 0.81  3   5   6   6   8  
##
## — Variable type:numeric —
```

```
##          variable missing complete      n   mean      sd    p0    p25
##          alcohol          0      1599 1599 10.42    1.07    8.4    9.5
##          chlorides         0      1599 1599  0.087    0.047   0.012   0.07
##          citric acid        0      1599 1599  0.27     0.19     0      0.09
##          density           0      1599 1599  1         0.0019  0.99    1
##          fixed acidity      0      1599 1599  8.32     1.74     4.6    7.1
##          free sulfur dioxide 0      1599 1599 15.87    10.46     1      7
##          pH                0      1599 1599  3.31     0.15     2.74   3.21
##          residual sugar     0      1599 1599  2.54     1.41     0.9    1.9
##          sulphates          0      1599 1599  0.66     0.17     0.33   0.55
##          total sulfur dioxide 0      1599 1599 46.47    32.9      6      22
##          volatile acidity   0      1599 1599  0.53     0.18     0.12   0.39
##          p50    p75    p100      hist
## 10.2    11.1    14.9  
##  0.079  0.09    0.61  
##  0.26   0.42    1      
##  1       1       1      
##  7.9     9.2    15.9  
## 14      21      72    
##  3.31    3.4     4.01  
##  2.2     2.6    15.5  
##  0.62    0.73    2      
## 38      62     289    
##  0.52    0.64    1.58  
```

Limpieza de los datos.

El dataframe no tiene campos vacios ni nulos, es un buen dataset desde el punto de vista de limpieza.

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

El dataset no tiene elementos vacios, lo podemos comprobar con el siguiente comando

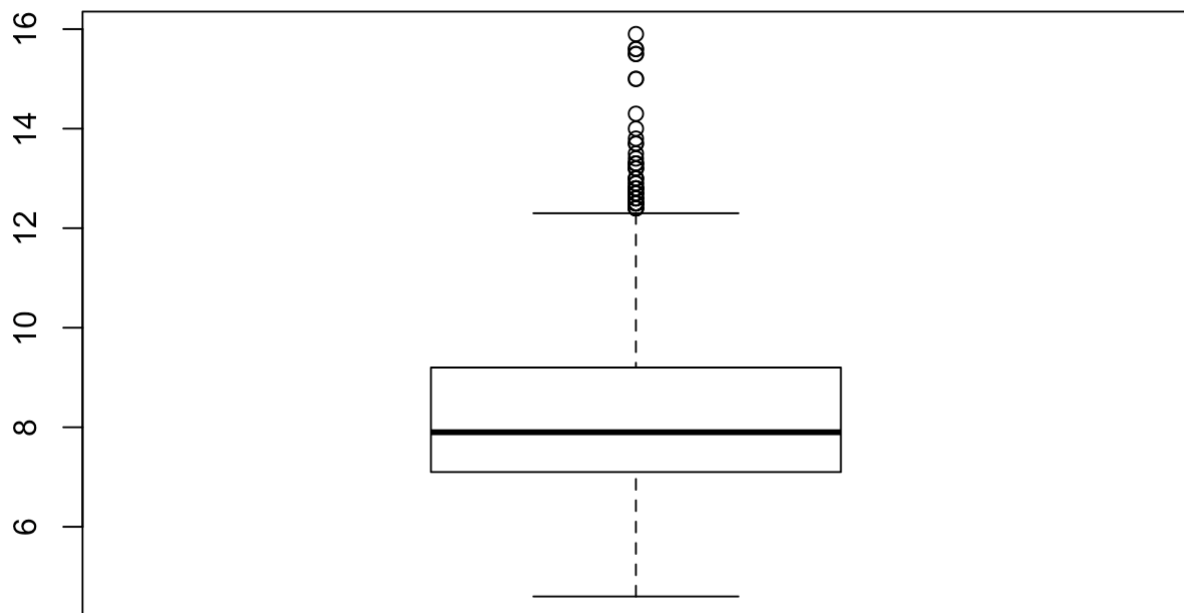
```
sapply(winequality, function(x) sum(length(which(is.na(x)))))
```

##	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

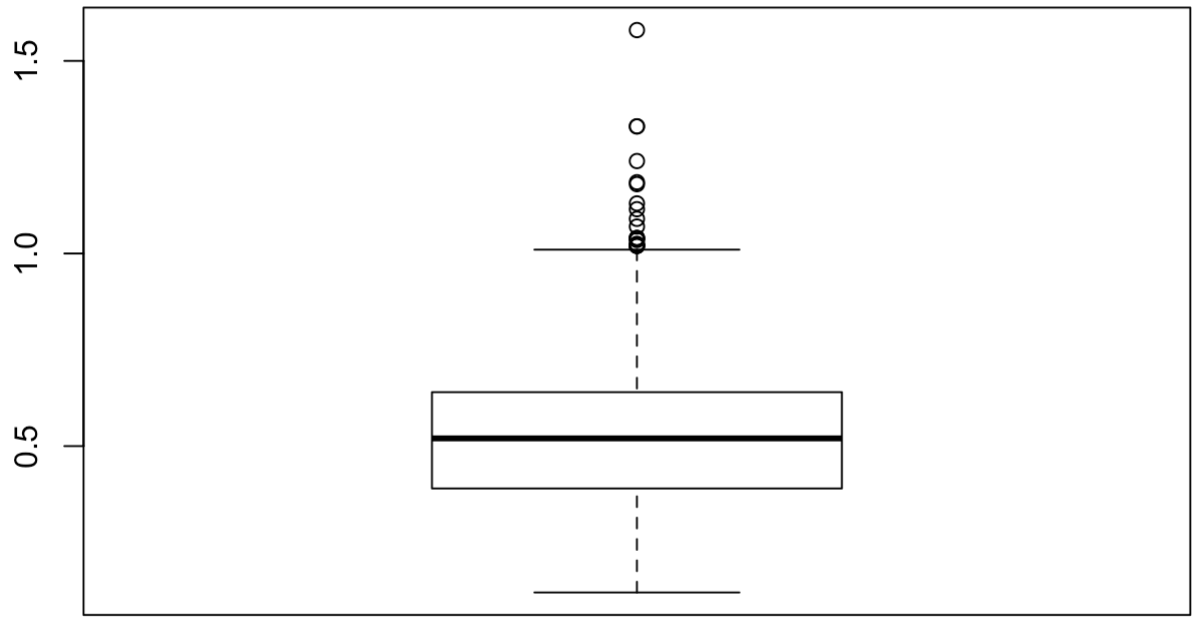
Identificación y tratamiento de valores extremos.

vamos a comprobar los Outlier que tiene las variables con el comando

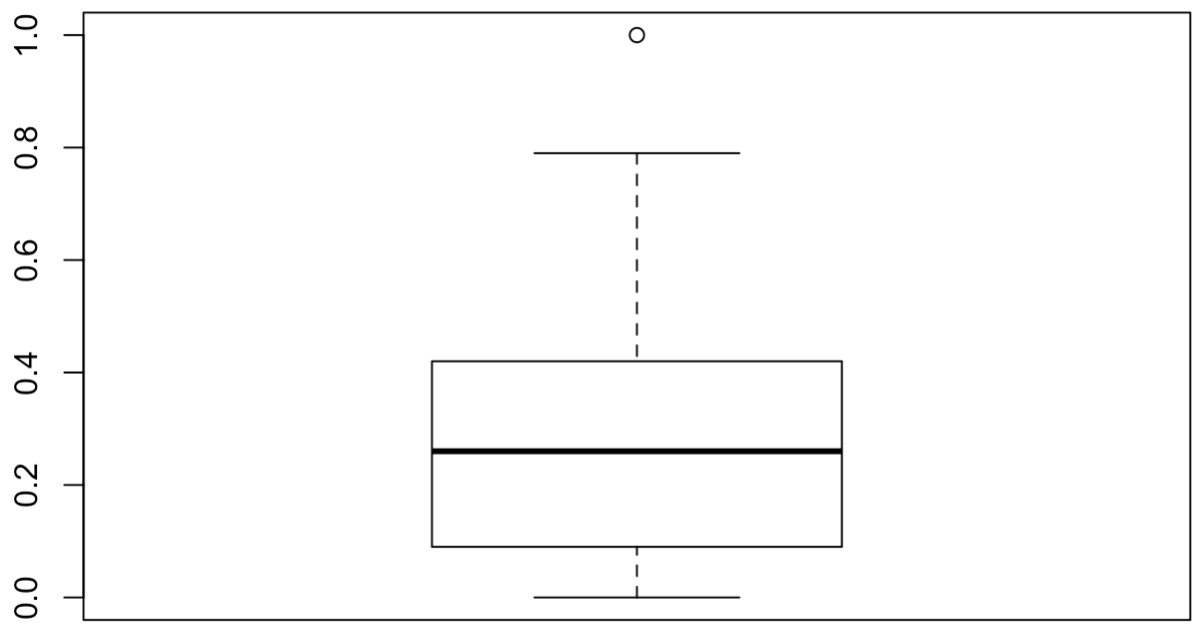
```
boxplot(winequality$`fixed acidity`)
```



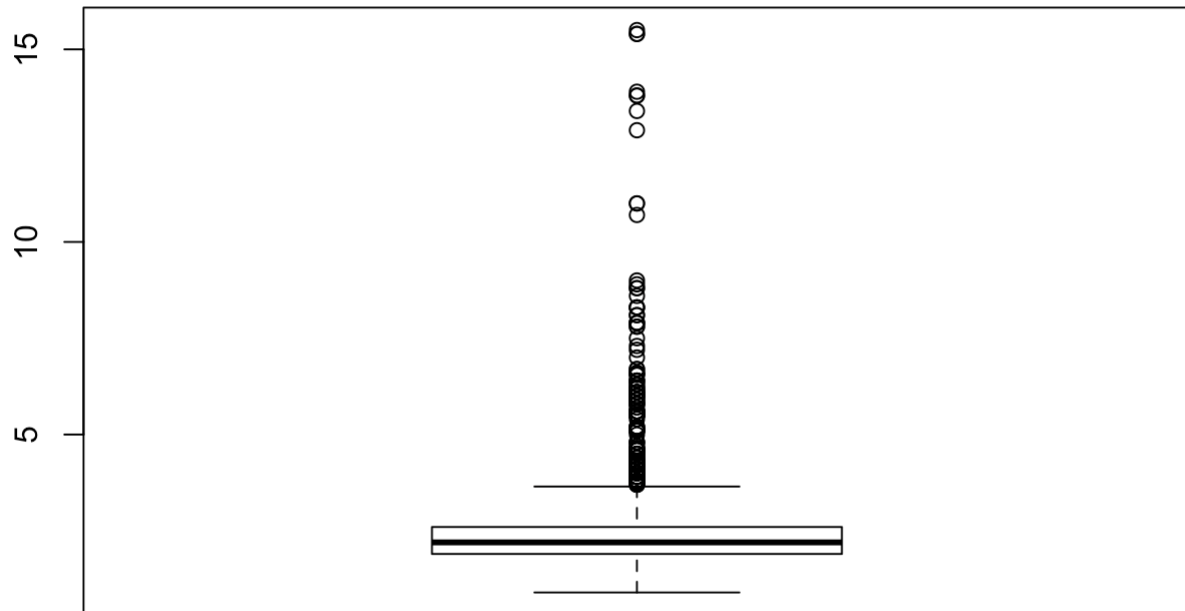
```
boxplot(winequality$`volatile acidity`)
```



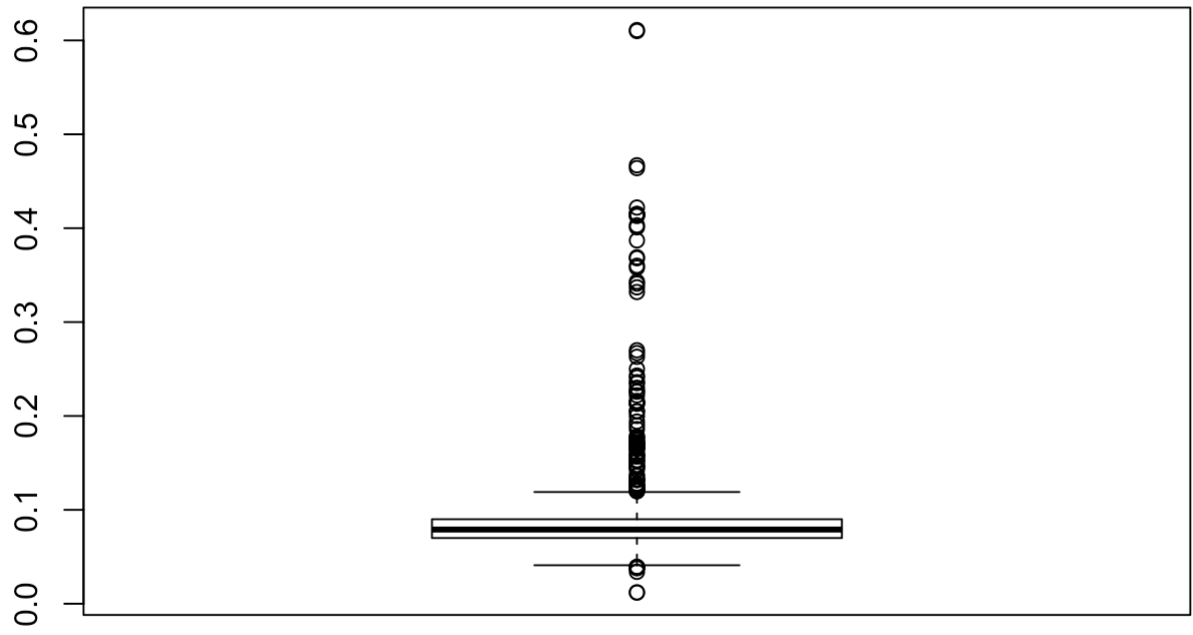
```
boxplot(winequality$`citric acid`)
```



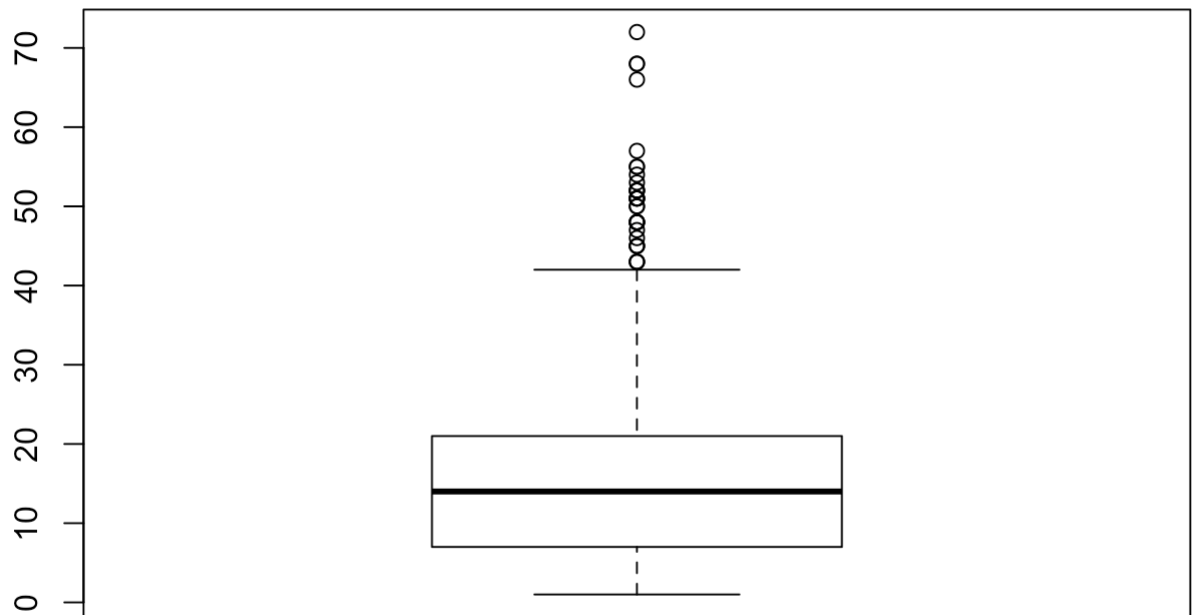
```
boxplot(winequality$`residual sugar`)
```



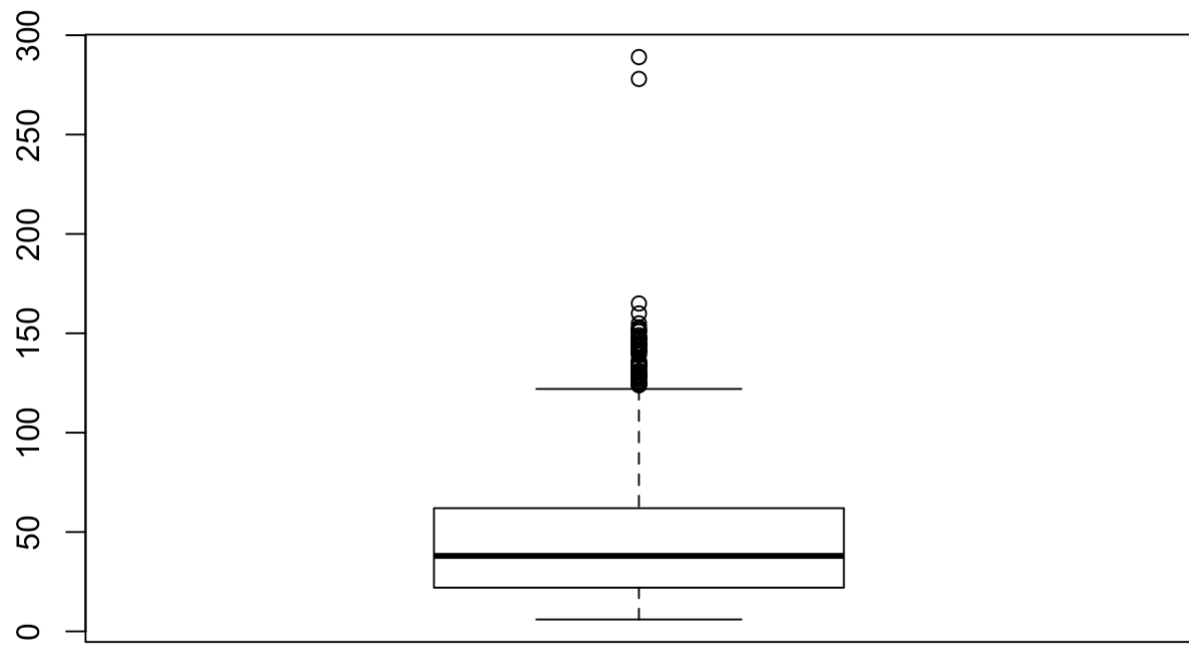
```
boxplot(winequality$chlorides)
```

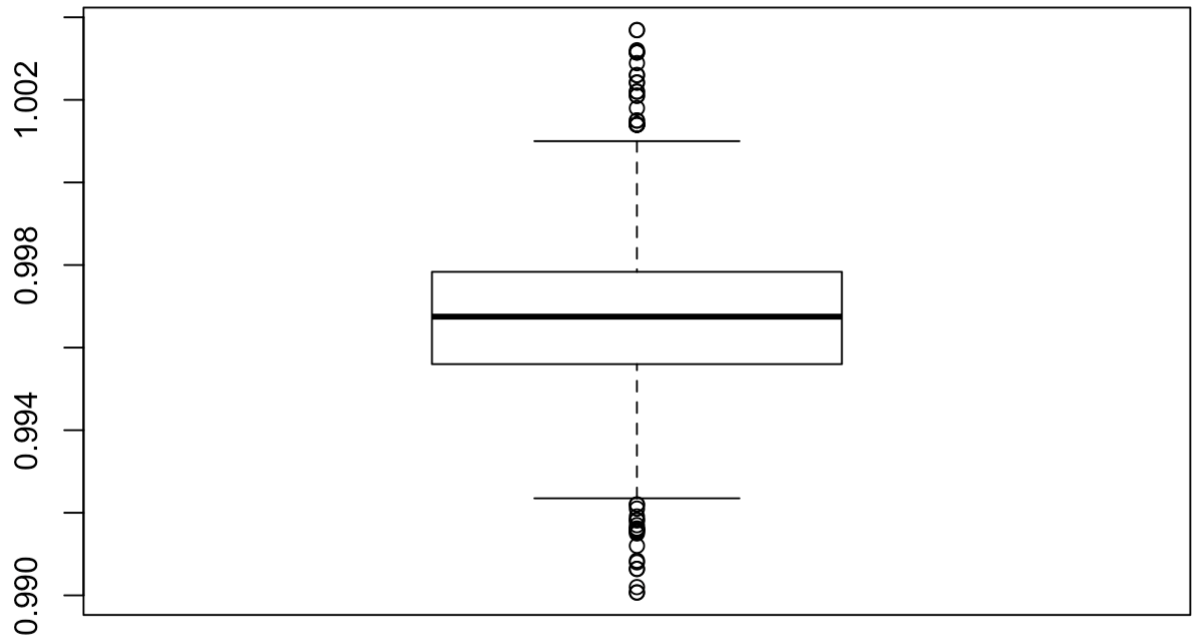
```
boxplot(winequality$`free sulfur dioxide`)
```



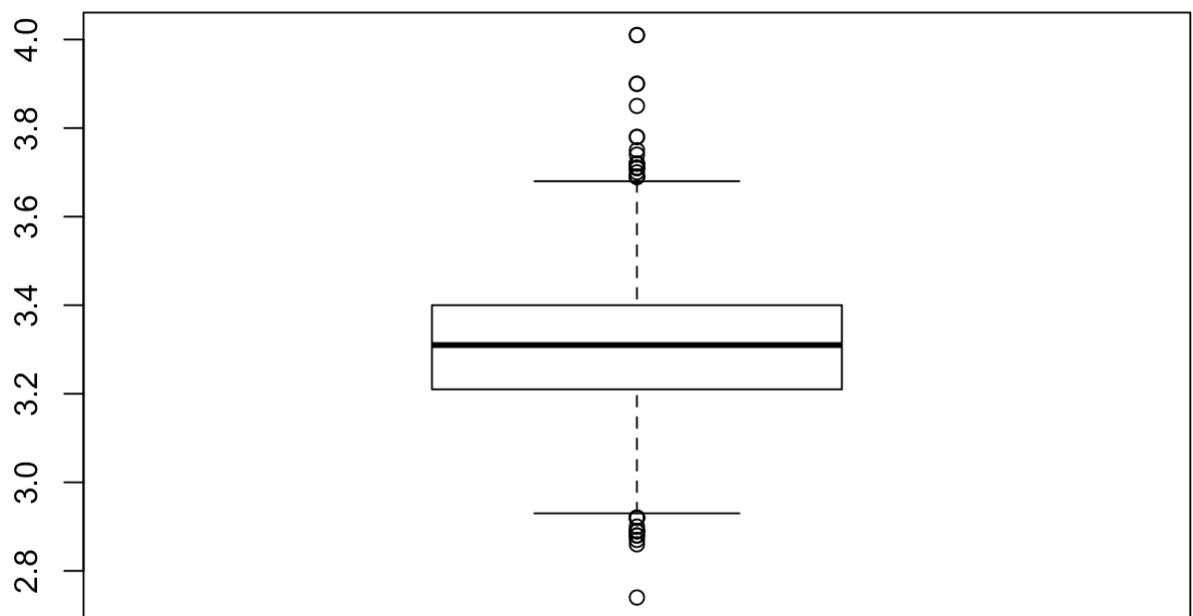
```
boxplot(winequality$`total sulfur dioxide`)
```



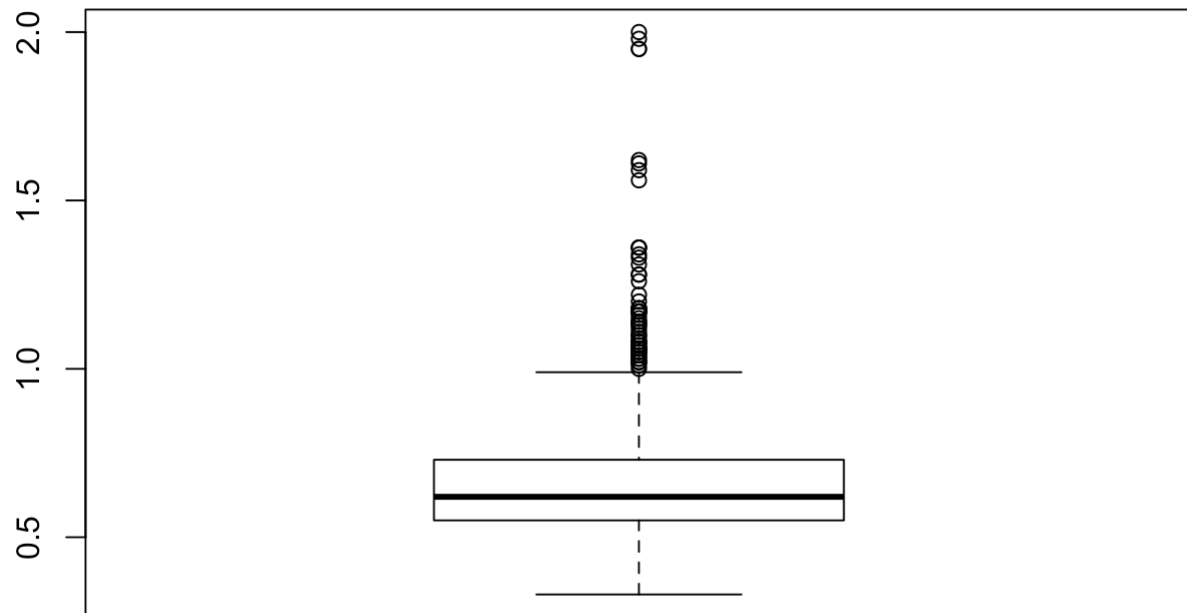
```
boxplot(winequality$density)
```



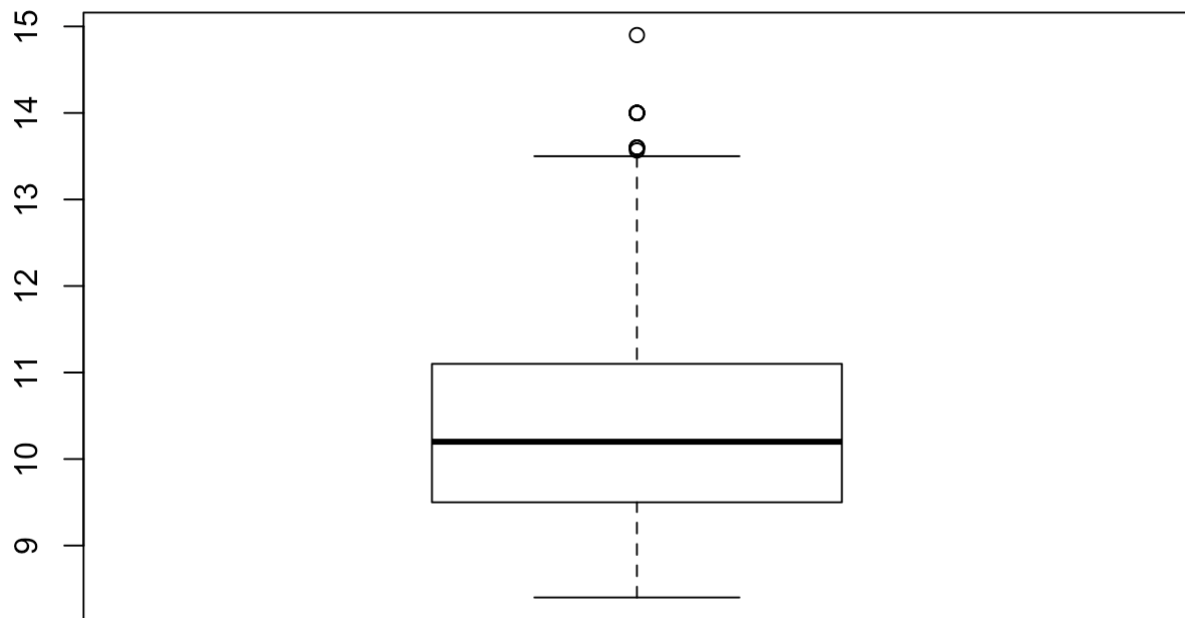
```
boxplot(winequality$pH)
```



```
boxplot(winequality$sulphates)
```



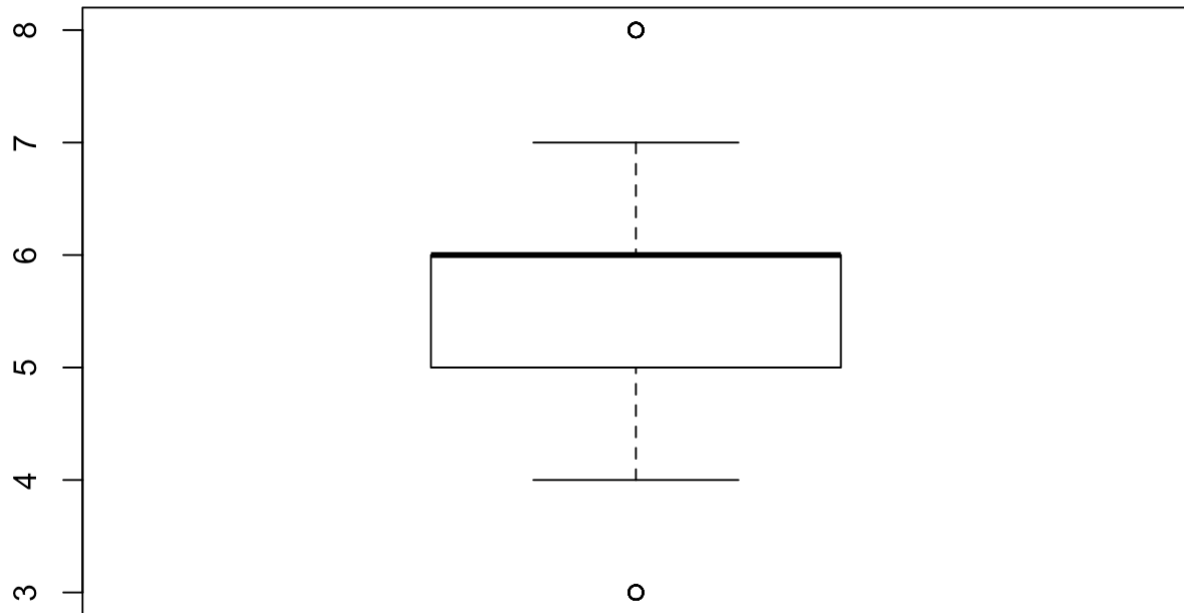
```
boxplot(winequality$alcohol)
```



En esta variable encontramos una serie de valores extremos a la media los cuales son necesarios estudiar. En este caso en particular, considero necesario dejar los valores atípicos de los datos objetivos como son los de los valores fisicoquímicos obtenidos Otro caso es el Quality, puesto que es subjetivo, tampoco conocemos el margen de error que puede tener la decisión de haberlo marcado con el valor 8.

Para la variable a buscar obtenemos los siguientes outliers

```
boxplot(winequality$quality)
```



```
fivenum(winequality$quality)
```

```
## [1] 3 5 6 6 8
```

Lo cual quiere decir que hay dos valores extremos en 3 y en 8 y la media coincide con los cinco números de Tukey

- minimum = 3
- lower-hinge = 5
- median = 6
- upper-hinge = 6
- maximum = 8

La conclusión es que un gran número de valores se concentran alrededor del 5 y 6 de calidad (el valor de la media y mediana están entre estos dos valores)

Análisis de los datos.

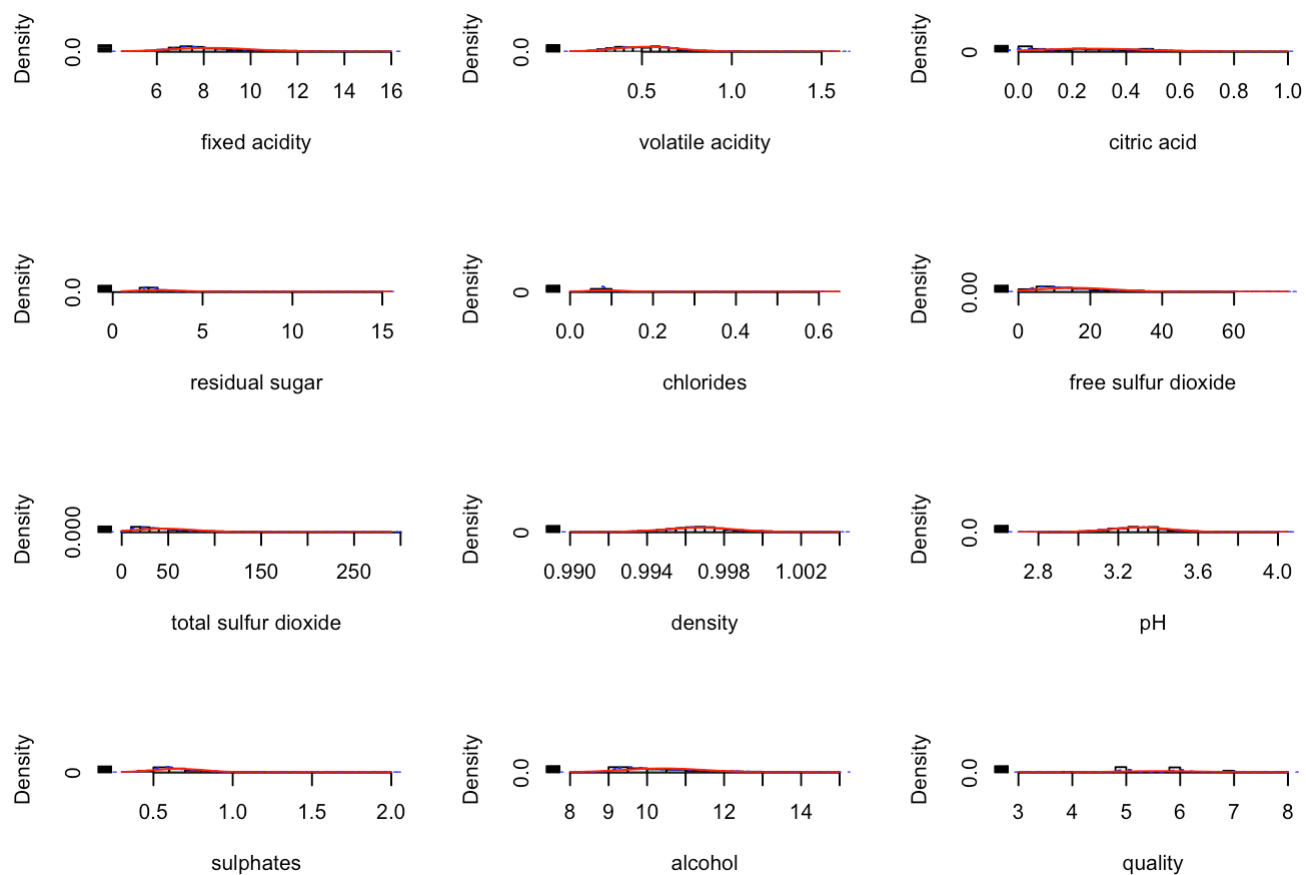
Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En este caso en particular, es necesario conocer el grado de influencia que tienen las variables respecto a la calidad del vino, así que se analizará como una regresión lineal

En primer lugar queremos comprobar que relación existe entre las variables para buscar que no existe colinealidad (variables que se influyen entre ellas). Esta información es crítica para identificar las mejores variables predictoras

Para realizar dicho análisis es necesario calcular el coeficiente de correlación de cada par de variables

```
multi.hist(x=winequality,dcol=c("blue", "red"),dltty = c("dotted","solid"),main="")
```



```
round(cor(x=winequality,method = "pearson"),3)
```

##	fixed acidity	volatile acidity	citric acid			
##	fixed acidity	1.000	-0.256	0.672		
##	volatile acidity	-0.256	1.000	-0.552		
##	citric acid	0.672	-0.552	1.000		
##	residual sugar	0.115	0.002	0.144		
##	chlorides	0.094	0.061	0.204		
##	free sulfur dioxide	-0.154	-0.011	-0.061		
##	total sulfur dioxide	-0.113	0.076	0.036		
##	density	0.668	0.022	0.365		
##	pH	-0.683	0.235	-0.542		
##	sulphates	0.183	-0.261	0.313		
##	alcohol	-0.062	-0.202	0.110		
##	quality	0.124	-0.391	0.226		
##	residual sugar	chlorides	free sulfur dioxide			
##	fixed acidity	0.115	0.094	-0.154		
##	volatile acidity	0.002	0.061	-0.011		
##	citric acid	0.144	0.204	-0.061		
##	residual sugar	1.000	0.056	0.187		
##	chlorides	0.056	1.000	0.006		
##	free sulfur dioxide	0.187	0.006	1.000		
##	total sulfur dioxide	0.203	0.047	0.668		
##	density	0.355	0.201	-0.022		
##	pH	-0.086	-0.265	0.070		
##	sulphates	0.006	0.371	0.052		
##	alcohol	0.042	-0.221	-0.069		
##	quality	0.014	-0.129	-0.051		
##	total sulfur dioxide	density	pH	sulphates	alcohol	
##	fixed acidity	-0.113	0.668	-0.683	0.183	-0.062
##	volatile acidity	0.076	0.022	0.235	-0.261	-0.202
##	citric acid	0.036	0.365	-0.542	0.313	0.110
##	residual sugar	0.203	0.355	-0.086	0.006	0.042
##	chlorides	0.047	0.201	-0.265	0.371	-0.221
##	free sulfur dioxide	0.668	-0.022	0.070	0.052	-0.069
##	total sulfur dioxide	1.000	0.071	-0.066	0.043	-0.206
##	density	0.071	1.000	-0.342	0.149	-0.496
##	pH	-0.066	-0.342	1.000	-0.197	0.206
##	sulphates	0.043	0.149	-0.197	1.000	0.094
##	alcohol	-0.206	-0.496	0.206	0.094	1.000
##	quality	-0.185	-0.175	-0.058	0.251	0.476
##	quality					
##	fixed acidity	0.124				
##	volatile acidity	-0.391				
##	citric acid	0.226				
##	residual sugar	0.014				
##	chlorides	-0.129				
##	free sulfur dioxide	-0.051				
##	total sulfur dioxide	-0.185				
##	density	-0.175				
##	pH	-0.058				
##	sulphates	0.251				
##	alcohol	0.476				
##	quality	1.000				

Conclusiones: Las variables que tienen una mayor relación lineal con la variable quality son:

- Alcohol 0.476
- Sulfatos 0.251

Comprobamos si están relacionadas entre ellas

```
cor(winequality$alcohol,winequality$sulphates)
```

```
## [1] 0.09359475
```

Comprobamos que no hay colinialidad entre ambas variables (0.09359475).

Ahora comprobamos con un modelos de regresión lineal multiple la influencia de las variables predictoras (todas menos quality) sobre la variable dependiente (quality)

```
calidadlm <- lm(quality ~ .,data=winequality)
summary(calidadlm)
```

```
##
## Call:
## lm(formula = quality ~ ., data = winequality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.197e+01  2.119e+01   1.036   0.3002
## `fixed acidity`    2.499e-02  2.595e-02   0.963   0.3357
## `volatile acidity` -1.084e+00  1.211e-01  -8.948 < 2e-16 ***
## `citric acid`     -1.826e-01  1.472e-01  -1.240   0.2150
## `residual sugar`   1.633e-02  1.500e-02   1.089   0.2765
## chlorides         -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## `free sulfur dioxide` 4.361e-03  2.171e-03   2.009   0.0447 *
## `total sulfur dioxide` -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density           -1.788e+01  2.163e+01  -0.827   0.4086
## pH                -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates          9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol            2.762e-01  2.648e-02  10.429 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF, p-value: < 2.2e-16
```

Este modelo explica el 35% de los casos (variabilidad) con todos los datos del dataset

Observamos que las variables independientes más influyentes en la variable dependiente quality (calidad del vino) son las siguientes:

- alcohol
- sulphates

Así creamos el siguiente modelo con solo las dos variables más predictoras (Sulfatos y Alcohol)

```
modelo<-lm(quality ~ sulphates+alcohol,data=winequality)
summary(modelo)
```

```
##
## Call:
## lm(formula = quality ~ sulphates + alcohol, data = winequality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.6685 -0.3781 -0.1005  0.4992  2.4187
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37497     0.17745   7.748 1.64e-14 ***
## sulphates    0.99409     0.10235   9.713 < 2e-16 ***
## alcohol      0.34604     0.01628  21.256 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6905 on 1596 degrees of freedom
## Multiple R-squared:  0.2699, Adjusted R-squared:  0.269
## F-statistic: 295 on 2 and 1596 DF,  p-value: < 2.2e-16
```

Ahora, con un modelo más simplificado, podemos explicar el 27% de la variabilidad de la calidad.

Comprobación de la normalidad y homogeneidad de la varianza.

Realizamos un análisis de inflación de varianza de las variables predictoras anteriores

```
vif(modelo)
```

```
## sulphates    alcohol
##  1.008837  1.008837
```

```
sapply(winequality, ad.test)
```

```
##          fixed acidity
## statistic 28.14296
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          volatile acidity
## statistic 5.683075
## p.value   5.318894e-14
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          citric acid
## statistic 17.54209
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          residual sugar
## statistic 188.0644
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          chlorides
## statistic 210.4492
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          free sulfur dioxide
## statistic 38.60991
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          total sulfur dioxide
## statistic 52.48865
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          density
## statistic 3.867595
## p.value   1.227494e-09
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          pH
## statistic 1.864112
## p.value   9.245208e-05
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          sulphates
## statistic 46.9322
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          alcohol
## statistic 34.91706
## p.value   3.7e-24
## method    "Anderson-Darling normality test"
## data.name "X[[i]]"
##          quality
## statistic 110.6328
```

```
## p.value      3.7e-24
## method      "Anderson-Darling normality test"
## data.name    "X[[i]]"
```

Como el valor de inflación de la varianza es inferior a 5 , no se considera colinialidad: confirmamos que son buenos predictores

El valor del P-Value es inferior a 0.05 de Alcohol y Sulfatos por lo tanto no sigue una distribución normal, es más ninguna de las columnas lo siguen

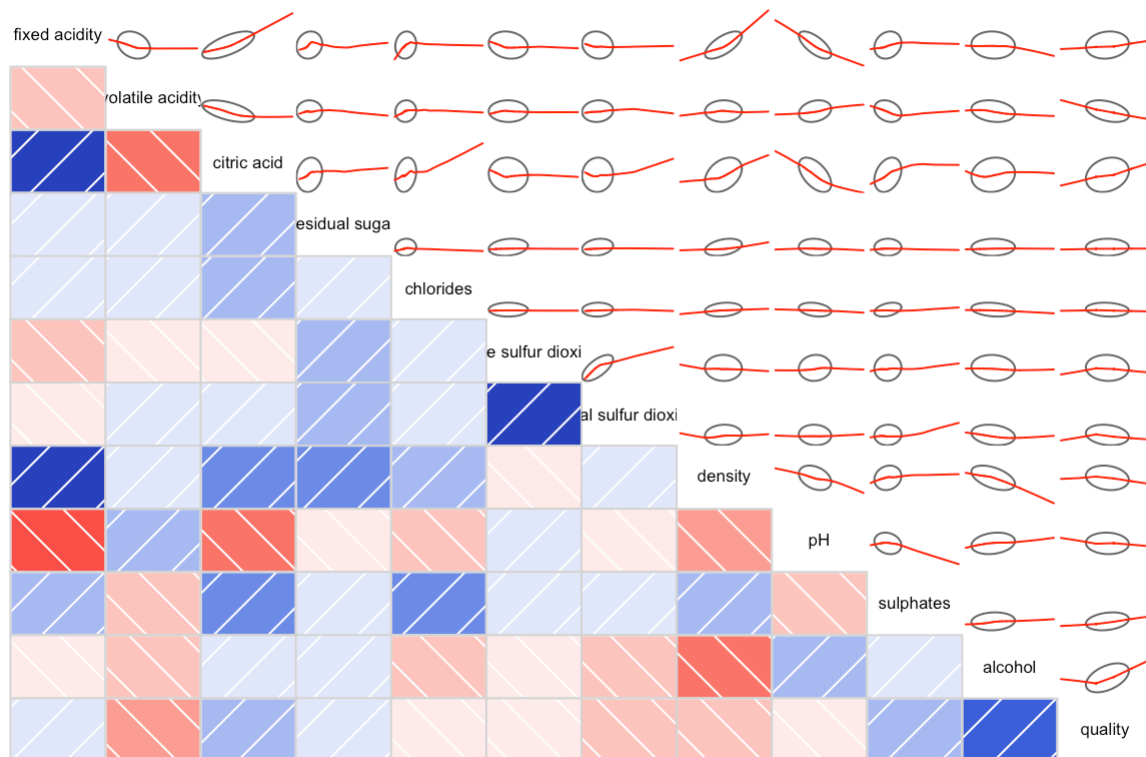
De lo que podemos asegurar que no hay correlación lineal muy alta entre los predictores, por lo tanto son las variables que más influyen en la calidad del vino

Aplicación de pruebas estadísticas para comparar los grupos de datos.

Se ha realizado la regresión lineal en el apartado anterior

Representación de los resultados a partir de tablas y gráficas.

```
corrgram(winequality, lower.panel=panel.shade, upper.panel=panel.ellipse)
```



Comprobamos en la última final (quality) que los colores más azules son variables sin correlación que más influyen en la calidad

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

Los resultados obtenidos nos demuestran que cuanto máyor sea el valor de la variable alcohol o sulfatos, mayor será la calidad del vino