Estudio de la influencia de las características estructurales en la calidad del vino blanco

Autores: José Miguel Castellano Sierra y Pedro Jesús de Barrios Roncero

Junio 2020

Contents

1 Descripción del conjunto de datos

```
if(!require(knitr)){
    install.packages('knitr', repos='http://cran.us.r-project.org')
    library(knitr)
}

if(!require(arules)){
    install.packages('arules', repos='http://cran.us.r-project.org')
    library(arules)
}

if(!require(arulesViz)){
    install.packages('arulesViz', repos='http://cran.us.r-project.org')
    library(arulesViz)
}

 if(!require(scales)){
    install.packages('scales', repos='http://cran.us.r-project.org')
    library(scales)
}

if(!require(ggplot2)){
    install.packages('ggplot2', dependencies = TRUE, repos='http://cran.us.r-project.org')
    library(ggplot2)
}

if(!require(dplyr)){
    install.packages('dplyr', repos='http://cran.us.r-project.org')
    library(dplyr)
}

if(!require(scales)){
    install.packages('scales', repos='http://cran.us.r-project.org')
    library(scales)
}

if(!require(corrplot)){
    install.packages('corrplot', repos='http://cran.us.r-project.org')
    library(corrplot)
}

if(!require(ngram)){
    install.packages('ngram', repos='http://cran.us.r-project.org')
    library(ngram)
}

if(!require(car)){
    install.packages('car', repos='http://cran.us.r-project.org')
    library(car)
}
```

El juego de datos elegido trata sobre la calidad del vino blanco en función de sus características y está

disponible en el siguiente enlace: http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/ Este juego de datos está compuesto por 12 variables, siendo 11 de las cuales características del vino (tales como la densidad o el PH) y una variable que califica la calidad del vino (asignando una nota entre 0 y 10 a cada muestra).

El dataset cuenta con 11 atributos numéricos, así como una variable que califica la calidad de las muestras.

Los atributos que tenemos son los siguientes:

1. fixed acidity (acidez fija)
2. volatile acidity (acidez volátil)
3. citric acid (ácido cítrico)
4. residual sugar (azúcar residual)
5. chlorides (cloruros)
6. free sulfur dioxide (dióxido de azufre libre)
7. total sulfur dioxide (dióxido de azufre total)
8. density (densidad)
9. pH
10. sulphates (sulfatos)
11. alcohol
12. quality (calidad)

La variable de salida es la calidad. Tiene un rango de valores comprendido entre 0 y 10 y ha sido estimada en función de datos de percepción sensorial.

Este estudio pretende analizar la influencia de los distintos atributos estructurales en la calidad final del vino, de manera que se detecten cuáles son más preponderantes, para así poder modificar los procesos de elaboración del vino. El fin que se persigue es el de mejorar la calidad de la producción para obtener vinos de gama alta y así aumentar los márgenes de beneficio de la bodega.

```
wine_read <- read.csv(
  'http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv',
  sep=';', header = TRUE)
colnames(wine_read)<-c("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar",
                       "chlorides", "free_sulfur_dioxide", "total_sulfur_dioxide", "density",
                       "pH", "sulphates", "alcohol", "quality")
```

2 Integración, selección y limpieza de los datos de interés

En primer lugar, realizaremos un estudio de cómo influye cada uno de los atributos en la calidad del vino, sin descartar ninguno. Posteriormente, estudiaremos la correlacción que puedan tener los distintos atributos y aplicaremos el método de reducción de la dimensionalidad ACP para obtener un conjunto reducido de variables no correlacionadas que permitan predecir con fiabilidad la calidad de las muestras de vino.

En primer lugar, analizaremos si el conjunto de datos contienen elementos vacíos o con valores NA:

```
registrosNA<-matrix(nrow=1,ncol=ncol(wine_read));
registrosNulos<-matrix(nrow=1,ncol=ncol(wine_read));

for (m in 1:ncol(wine_read)) {
    registrosNA[1,m]<-length(wine_read[(is.na(wine_read[,m])==TRUE),m])
    print(concatenate("El número de muestras que contienen NA en el atributo",
                      names(wine_read)[m]," es:", registrosNA[1,m]))
```

```
    registrosNulos[1,m]<-length(wine_read[(wine_read[,m]==""),m])
    print(concatenate("El número de muestras que no traen informado el atributo",
                      names(wine_read)[m]," es:", registrosNulos[1,m]))

    print("")
    m<-m+1
}
```

```
## [1] "El número de muestras que contienen NA en el atributo fixed_acidity  es: 0"
## [1] "El número de muestras que no traen informado el atributo fixed_acidity  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo volatile_acidity  es: 0"
## [1] "El número de muestras que no traen informado el atributo volatile_acidity  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo citric_acid  es: 0"
## [1] "El número de muestras que no traen informado el atributo citric_acid  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo residual_sugar  es: 0"
## [1] "El número de muestras que no traen informado el atributo residual_sugar  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo chlorides  es: 0"
## [1] "El número de muestras que no traen informado el atributo chlorides  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo free_sulfur_dioxide  es: 0"
## [1] "El número de muestras que no traen informado el atributo free_sulfur_dioxide  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo total_sulfur_dioxide  es: 0"
## [1] "El número de muestras que no traen informado el atributo total_sulfur_dioxide  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo density  es: 0"
## [1] "El número de muestras que no traen informado el atributo density  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo pH  es: 0"
## [1] "El número de muestras que no traen informado el atributo pH  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo sulphates  es: 0"
## [1] "El número de muestras que no traen informado el atributo sulphates  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo alcohol  es: 0"
## [1] "El número de muestras que no traen informado el atributo alcohol  es: 0"
## [1] ""
## [1] "El número de muestras que contienen NA en el atributo quality  es: 0"
## [1] "El número de muestras que no traen informado el atributo quality  es: 0"
## [1] ""
```

Las muestras tienen informados todos sus atributos, por lo que no es necesario acometer ninguna acción.

Haremos un análisis previo, viendo las primeras filas de nuestro dataset, así como el tipo de las variables que tenemos y un estudio para ver la distribución de los valores máximos y mínimos teniendo en cuenta la media, la mediana y los cuartiles.

```r
dim(wine_read)
```

```
## [1] 4898   12
```

```r
head(wine_read, 5)
```

```
##   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
##   free_sulfur_dioxide total_sulfur_dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
```

Vemos que todos los atributos contienen valores numéricos, pudiendo tomar cualquier valor perteneciente al conjunto de los número reales.

```r
str(wine_read)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed_acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile_acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric_acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual_sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free_sulfur_dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total_sulfur_dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```r
summary(wine_read)
```

```
##  fixed_acidity    volatile_acidity  citric_acid     residual_sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
```

```
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free_sulfur_dioxide total_sulfur_dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

Ahora mostraremos el número de valores distintos que puede tomar cada uno de las variables en estudio.

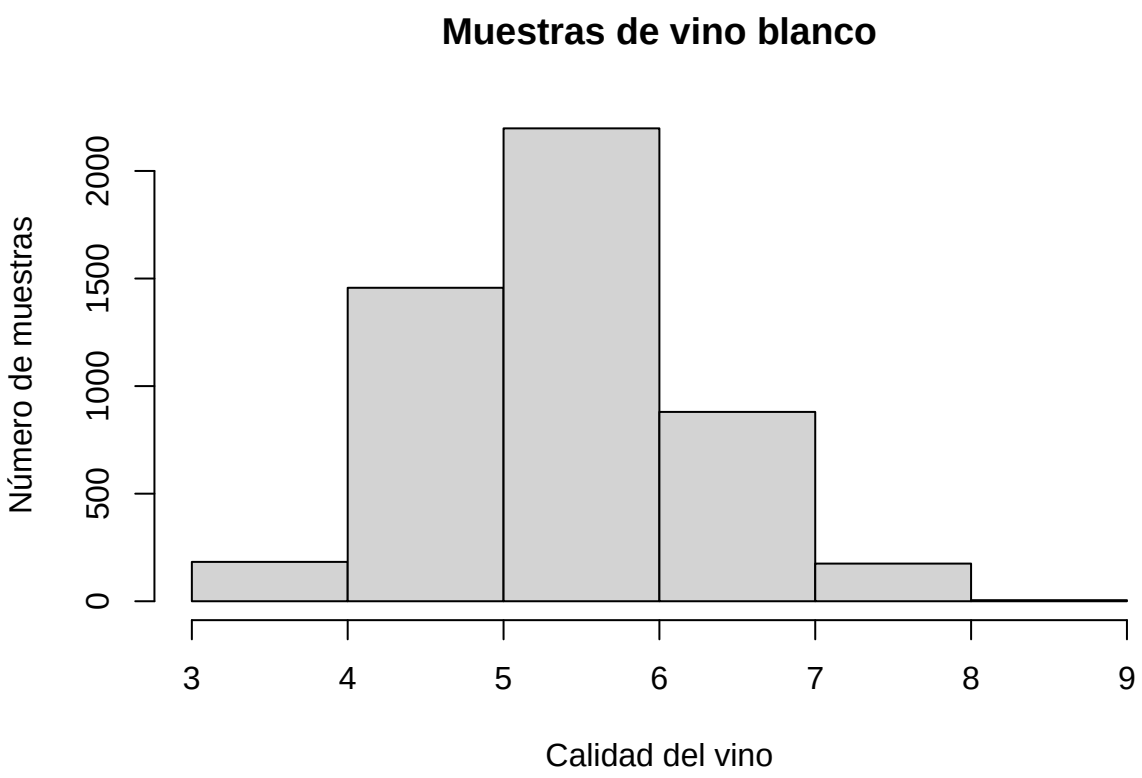
```
apply(wine_read, 2, function(x) length(unique(x)))
```

```
##        fixed_acidity     volatile_acidity          citric_acid
##                   68                  125                   87
##       residual_sugar            chlorides  free_sulfur_dioxide
##                  310                  160                  132
## total_sulfur_dioxide              density                   pH
##                  251                  890                  103
##            sulphates              alcohol              quality
##                   79                  103                    7
```

Como se puede apreciar, el campo quality tiene únicamente 7 valores distintos, mientras que el resto de variables numéricas toman un gran número de valores distintos.

A continuación se muestra en un histograma la frecuencia de las muestras en el juego de datos en función de la calidad:

```
hist(wine_read[["quality"]], xlab="Calidad del vino", ylab="Número de muestras",
     main="Muestras de vino blanco" , breaks=length(unique(wine_read[["quality"]])),
     cex.names=1, names=T)
```



Como se puede observar, los datos del número de muestras en función de la calidad tienen una distribución que se asemeja a la distribución normal, aunque posteriormente realizaremos el test de normalidad de Shapiro-Wilk para comprobarlo.

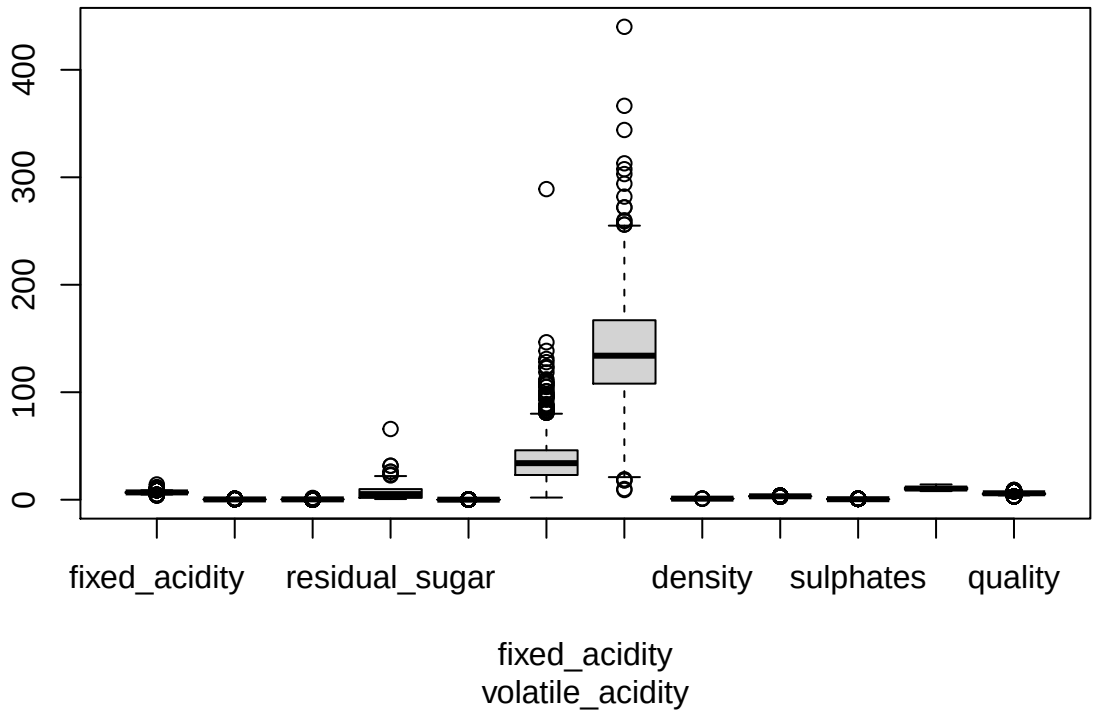
Para detectar visualmente los outliers que tiene el conjunto de datos, mostraremos las variables en un diagrama de cajas a través de la función boxplot.

```
summary(wine_read)
```

```
##  fixed_acidity    volatile_acidity  citric_acid     residual_sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free_sulfur_dioxide total_sulfur_dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
```

```
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

```
boxplot(wine_read, xlab=colnames(wine_read))
```



Al analizar la gráfica, se observa que el atributo "free_sulfur_dioxide" presenta varios valores extremos muy marcados En menor medida, también se aprecia la existencia de un destacado número de outliers en el atributo "total_sulfur_dioxide".

Viendo que la variable free_sulfur_dioxide presenta outliers muy marcados, procederemos a eliminar tales muestras.

```
outliersFreeSulfur <- boxplot(wine_read$free_sulfur_dioxide, plot=FALSE)$out
wine_read[which(wine_read$free_sulfur_dioxide %in% outliersFreeSulfur),]
```

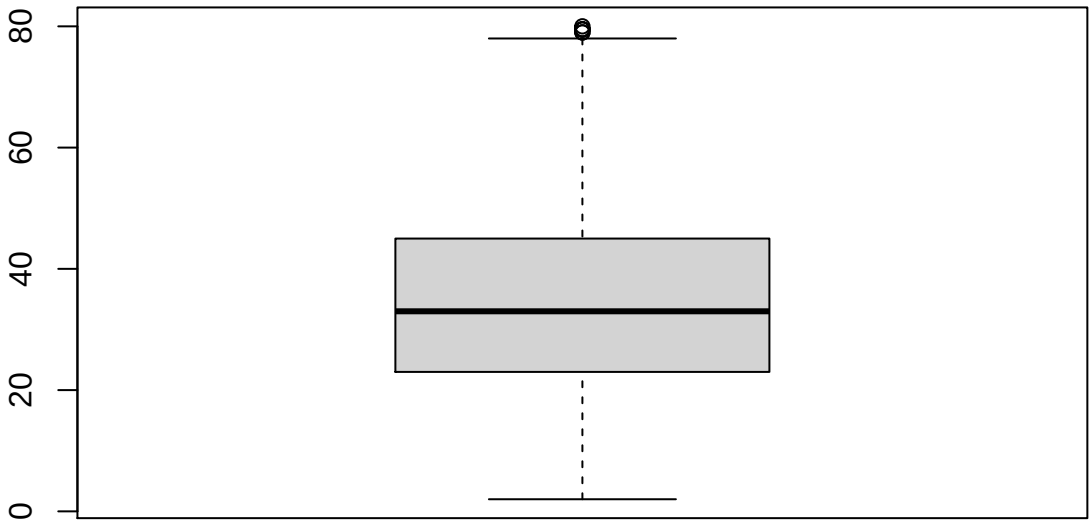
```
##      fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
## 68             6.7            0.250        0.13           1.20     0.041
## 298            7.2            0.190        0.46           3.80     0.041
## 326            7.5            0.270        0.31           5.80     0.057
## 388            6.3            0.390        0.35           5.90     0.040
## 396            6.8            0.270        0.12           1.30     0.040
## 406            6.8            0.270        0.12           1.30     0.040
## 460            6.1            0.430        0.35           9.10     0.059
## 660            6.8            0.290        0.16           1.40     0.038
```

```
## 753            6.7           0.200        0.42          14.00     0.038
## 767            6.7           0.500        0.63          13.40     0.078
## 1258           6.4           0.170        0.27           6.70     0.036
## 1675           6.8           0.220        0.29           8.90     0.046
## 1689           6.7           0.250        0.26           1.55     0.041
## 1760           6.6           0.620        0.20           8.70     0.046
## 1843           8.2           0.180        0.31          11.80     0.039
## 1856           8.0           0.220        0.28          14.00     0.053
## 1860           8.0           0.220        0.28          14.00     0.053
## 1932           7.1           0.490        0.22           2.00     0.047
## 2335           7.5           0.230        0.35          17.80     0.058
## 2337           6.8           0.260        0.22           4.80     0.041
## 2576           6.7           0.170        0.42          10.40     0.038
## 2626           4.5           0.190        0.21           0.95     0.033
## 2729           6.5           0.350        0.28          12.40     0.051
## 2736           6.5           0.350        0.28          12.40     0.051
## 2749           5.8           0.170        0.34           1.80     0.045
## 2751           5.8           0.170        0.34           1.80     0.045
## 2873           4.2           0.170        0.36           1.80     0.029
## 2894           6.2           0.160        0.34           1.70     0.038
## 2931           5.6           0.210        0.40           1.30     0.041
## 3051           6.2           0.255        0.24           1.70     0.039
## 3073           7.5           0.250        0.47           4.10     0.041
## 3308           9.4           0.240        0.29           8.50     0.037
## 3380           7.6           0.360        0.49          11.30     0.046
## 3388           7.6           0.360        0.49          11.30     0.046
## 3462           6.7           0.240        0.30           3.85     0.042
## 3471           6.7           0.240        0.30           3.85     0.042
## 3521           6.0           0.230        0.15           9.70     0.048
## 3524           6.0           0.230        0.15           9.70     0.048
## 3621           6.3           0.200        0.26           4.70     0.040
## 3862           6.2           0.220        0.30          12.40     0.054
## 3863           6.5           0.270        0.19           6.60     0.045
## 3864           6.5           0.270        0.19           6.60     0.045
## 3869           6.0           0.220        0.25          11.10     0.056
## 3870           6.2           0.220        0.30          12.40     0.054
## 3872           6.5           0.270        0.19           6.60     0.045
## 3982           6.3           0.220        0.27           4.50     0.036
## 4180           6.9           0.240        0.40          15.40     0.052
## 4186           6.9           0.240        0.40          15.40     0.052
## 4746           6.1           0.260        0.25           2.90     0.047
## 4842           5.7           0.220        0.25           1.10     0.050
##      free_sulfur_dioxide total_sulfur_dioxide density   pH sulphates alcohol
## 68                  81.0                174.0 0.99200 3.14      0.42     9.8
## 298                 82.0                187.0 0.99320 3.19      0.60    11.2
## 326                131.0                313.0 0.99460 3.18      0.59    10.5
## 388                 82.5                260.0 0.99410 3.12      0.66    10.1
## 396                 87.0                168.0 0.99200 3.18      0.41    10.0
## 406                 87.0                168.0 0.99200 3.18      0.41    10.0
## 460                 83.0                249.0 0.99710 3.37      0.50     8.5
## 660                122.5                234.5 0.99220 3.15      0.47    10.0
## 753                 83.0                160.0 0.99870 3.16      0.50     9.4
## 767                 81.0                238.0 0.99880 3.08      0.44     9.2
## 1258                88.0                223.0 0.99480 3.28      0.35    10.2
```

```
## 1675                82.0                188.0 0.99550 3.30      0.44    10.3
## 1689               118.5                216.0 0.99490 3.55      0.63     9.4
## 1760                81.0                224.0 0.99605 3.17      0.44     9.3
## 1843                96.0                249.0 0.99760 3.07      0.52     9.5
## 1856                83.0                197.0 0.99810 3.14      0.45     9.8
## 1860                83.0                197.0 0.99810 3.14      0.45     9.8
## 1932               146.5                307.5 0.99240 3.24      0.37    11.0
## 2335               128.0                212.0 1.00241 3.44      0.43     8.9
## 2337               110.0                198.0 0.99437 3.29      0.67    10.6
## 2576                85.0                182.0 0.99628 3.04      0.44     8.9
## 2626                89.0                159.0 0.99332 3.34      0.42     8.0
## 2729                86.0                213.0 0.99620 3.16      0.51     9.9
## 2736                86.0                213.0 0.99620 3.16      0.51     9.9
## 2749                96.0                170.0 0.99035 3.38      0.90    11.8
## 2751                96.0                170.0 0.99035 3.38      0.90    11.8
## 2873                93.0                161.0 0.98999 3.65      0.89    12.0
## 2894                85.0                153.0 0.99090 3.33      0.86    12.0
## 2931                81.0                147.0 0.99010 3.22      0.95    11.6
## 3051               138.5                272.0 0.99452 3.53      0.53     9.6
## 3073                95.0                163.0 0.99184 2.92      0.59    11.3
## 3308               124.0                208.0 0.99395 2.90      0.38    11.0
## 3380                87.0                221.0 0.99840 3.01      0.43     9.2
## 3388                87.0                221.0 0.99840 3.01      0.43     9.2
## 3462               105.0                179.0 0.99189 3.04      0.59    11.3
## 3471               105.0                179.0 0.99189 3.04      0.59    11.3
## 3521               101.0                207.0 0.99571 3.05      0.30     9.1
## 3524               101.0                207.0 0.99571 3.05      0.30     9.1
## 3621               108.0                168.0 0.99278 3.07      0.75    10.7
## 3862               108.0                152.0 0.99728 3.10      0.47     9.5
## 3863                98.0                175.0 0.99364 3.16      0.34    10.1
## 3864                98.0                175.0 0.99364 3.16      0.34    10.1
## 3869               112.0                177.0 0.99610 3.08      0.36     9.4
## 3870               108.0                152.0 0.99728 3.10      0.47     9.5
## 3872                98.0                175.0 0.99364 3.16      0.34    10.1
## 3982                81.0                157.0 0.99280 3.05      0.76    10.7
## 4180                81.0                198.0 0.99860 3.20      0.69     9.4
## 4186                81.0                198.0 0.99860 3.20      0.69     9.4
## 4746               289.0                440.0 0.99314 3.44      0.64    10.5
## 4842                97.0                175.0 0.99099 3.44      0.62    11.1
##      quality
## 68         5
## 298        7
## 326        5
## 388        5
## 396        5
## 406        5
## 460        5
## 660        4
## 753        6
## 767        5
## 1258       6
## 1675       6
## 1689       3
## 1760       5
```

```
## 1843       6
## 1856       6
## 1860       6
## 1932       3
## 2335       5
## 2337       5
## 2576       6
## 2626       5
## 2729       6
## 2736       6
## 2749       8
## 2751       8
## 2873       7
## 2894       7
## 2931       8
## 3051       4
## 3073       6
## 3308       3
## 3380       5
## 3388       5
## 3462       8
## 3471       8
## 3521       5
## 3524       5
## 3621       7
## 3862       6
## 3863       6
## 3864       6
## 3869       6
## 3870       6
## 3872       6
## 3982       7
## 4180       5
## 4186       5
## 4746       3
## 4842       6
```

```
wine_read <- wine_read[-which(wine_read$free_sulfur_dioxide %in% outliersFreeSulfur),]
boxplot(wine_read$free_sulfur_dioxide)
```

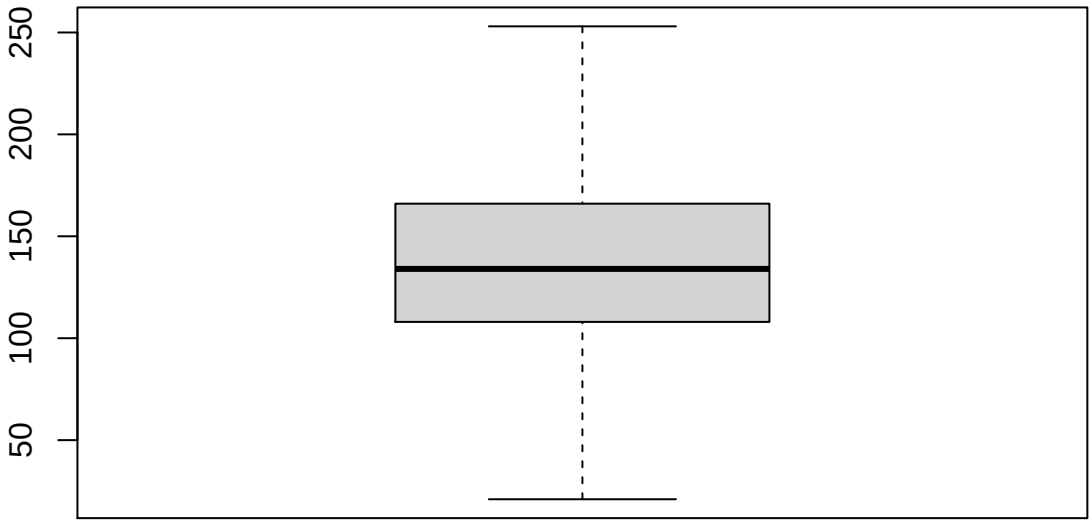



```
outliersTotalSulfur <- boxplot(wine_read$total_sulfur_dioxide, plot=FALSE)$out
wine_read[which(wine_read$total_sulfur_dioxide %in% outliersTotalSulfur),]
```

```
##      fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
## 112            7.2             0.27        0.46          18.75     0.052
## 228            7.1             0.25        0.32          10.30     0.041
## 741            6.9             0.39        0.40           4.60     0.022
## 1418           8.6             0.55        0.35          15.55     0.057
## 1941           8.3             0.36        0.57          15.00     0.052
## 1943           8.3             0.36        0.57          15.00     0.052
## 2128           9.1             0.33        0.38           1.70     0.062
## 2379           9.4             0.23        0.56          16.45     0.063
## 2655           6.9             0.40        0.22           5.95     0.081
## 3095           9.7             0.24        0.49           4.90     0.032
## 3096           9.7             0.24        0.49           4.90     0.032
## 3153           7.6             0.25        1.23           4.60     0.035
## 3711           4.7             0.67        0.09           1.00     0.020
## 3902           4.8             0.65        0.12           1.10     0.013
## 4515           6.1             0.40        0.18           9.00     0.051
##      free_sulfur_dioxide total_sulfur_dioxide density   pH sulphates alcohol
## 112                 45.0                255.0 1.00000 3.04      0.52     8.9
## 228                 66.0                272.0 0.99690 3.17      0.52     9.1
## 741                  5.0                 19.0 0.99150 3.31      0.37    12.6
## 1418                35.5                366.5 1.00010 3.04      0.63    11.0
## 1941                35.0                256.0 1.00010 2.93      0.64     8.6
```

```
## 1943                35.0                256.0 1.00010 2.93      0.64     8.6
## 2128                50.5                344.0 0.99580 3.10      0.70     9.5
## 2379                52.5                282.0 1.00098 3.10      0.51     9.3
## 2655                76.0                303.0 0.99705 3.40      0.57     9.4
## 3095                 3.0                 18.0 0.99368 2.85      0.54    10.0
## 3096                 3.0                 18.0 0.99368 2.85      0.54    10.0
## 3153                51.0                294.0 0.99018 3.03      0.43    13.1
## 3711                 5.0                  9.0 0.98722 3.30      0.34    13.6
## 3902                 4.0                 10.0 0.99246 3.32      0.36    13.5
## 4515                28.5                259.0 0.99640 3.19      0.50     8.8
##      quality
## 112        5
## 228        6
## 741        3
## 1418       3
## 1941       5
## 1943       5
## 2128       5
## 2379       5
## 2655       5
## 3095       6
## 3096       6
## 3153       6
## 3711       5
## 3902       4
## 4515       5
```

```r
wine_read <- wine_read[-which(wine_read$total_sulfur_dioxide %in% outliersTotalSulfur),]
boxplot(wine_read$total_sulfur_dioxide)
```



3 Análisis de los datos

Ahora normalizamos las variables para reducir el sesgo causado por la combinación de valores medidos a diferentes escalas para favorecer el agrupamiento que vamos a realizar.

```
vinoBlancoMatriz <- as.matrix(wine_read,nrow=nrow(wine_read), ncol=ncol(wine_read))
vinoBlancoMatrizRescalado<-matrix(ncol=ncol(wine_read), nrow=nrow(wine_read))
colnames(vinoBlancoMatrizRescalado) <- colnames(wine_read)

for (m in 1:ncol(vinoBlancoMatriz)) {
  vinoBlancoMatrizRescalado[,m] <- rescale(vinoBlancoMatriz[,m], to = c(0, 1),
                                           from = range(vinoBlancoMatriz[,m],
                                                        na.rm = TRUE, finite = TRUE))
}
```

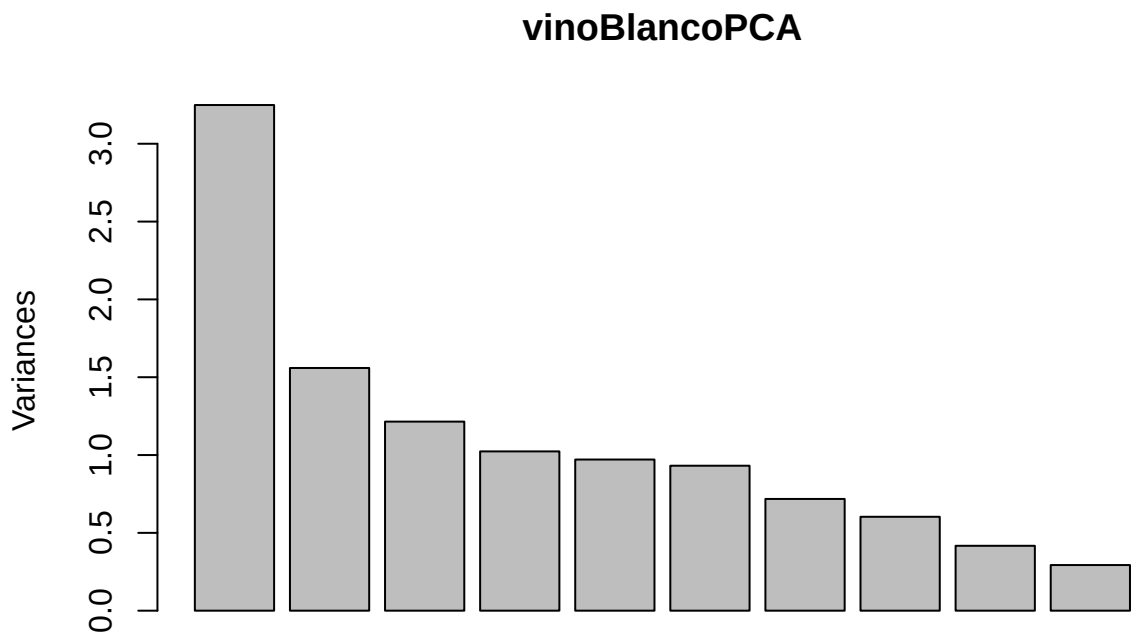
Para determinar qué atributos son clave en la calidad del vino, aplicaremos el método del análisis de los componentes principales (ACP), que permite reducir las dimensiones del problema. El ACP es un procedimiento estadístico que utiliza una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas (cada una de las cuales toma valores numéricos) en un conjunto de valores no correlacionados linealmente.

Desde una perspectiva simplificada, el procedimiento ACP transforma los datos linealmente en nuevas propiedades que no tienen correlación entre sí.

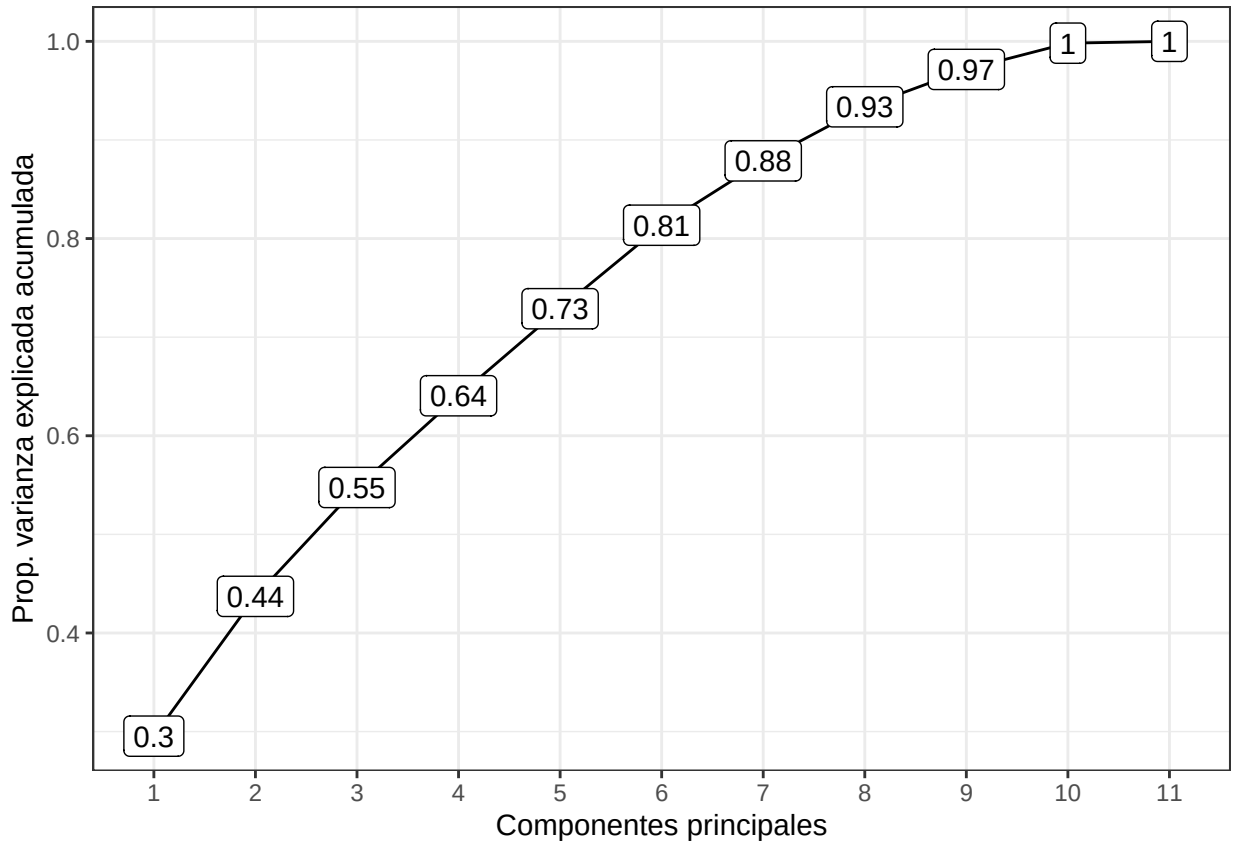
```r
vinoBlancoPCA <- prcomp(vinoBlancoMatrizRescalado[,1:11], center = TRUE, scale = TRUE)
summary(vinoBlancoPCA)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8025 1.2487 1.1022 1.01159 0.98543 0.96509 0.84724
## Proportion of Variance 0.2954 0.1418 0.1104 0.09303 0.08828 0.08467 0.06526
## Cumulative Proportion  0.2954 0.4371 0.5475 0.64057 0.72885 0.81352 0.87878
##                           PC8     PC9    PC10    PC11
## Standard deviation     0.77666 0.64561 0.54144 0.14228
## Proportion of Variance 0.05484 0.03789 0.02665 0.00184
## Cumulative Proportion  0.93362 0.97151 0.99816 1.00000
```

```r
plot(vinoBlancoPCA)
```




```r
prop_varianza <- vinoBlancoPCA$sdev^2/sum(vinoBlancoPCA$sdev^2)
prop_varianza_acum <- cumsum(prop_varianza)

ggplot(data = data.frame(prop_varianza_acum, pc = factor(1:11)),
       aes(x = pc, y = prop_varianza_acum, group = 1)) +
 geom_point() +
 geom_line() +
 geom_label(aes(label = round(prop_varianza_acum,2))) +
 theme_bw() +
 labs(x = "Componentes principales", y = "Prop. varianza explicada acumulada")
```



Aplicando el análisis de componentes principales obtenemos que somos capaces de predecir aproximadamente el 90% del valor de la calidad de las muestras analizando únicamente 7 componentes principales que no tienen correlación entre sí.

Las componentes principales son las siguientes:

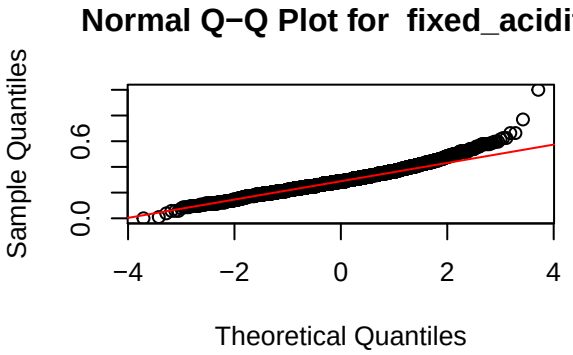
```
vinoBlancoPCA$rotation
```

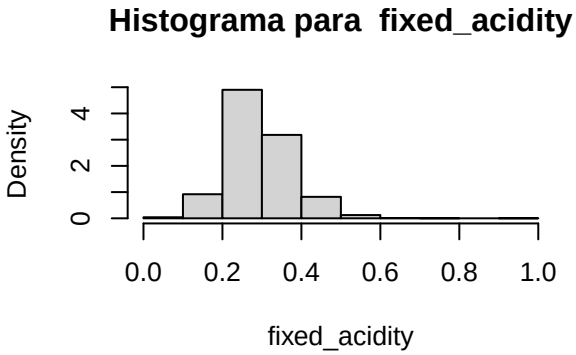
```
##                              PC1          PC2         PC3         PC4
## fixed_acidity        -0.151505534 -0.598410758  0.09304876 -0.05116650
## volatile_acidity     -0.004601277  0.061326958 -0.60794118 -0.30727755
## citric_acid          -0.145064046 -0.350511585  0.50538178 -0.14107525
## residual_sugar       -0.426637027  0.021433739 -0.20648806  0.28186111
## chlorides            -0.209592150 -0.008375099 -0.10199625 -0.68209148
## free_sulfur_dioxide  -0.313251195  0.257251825  0.27352274  0.29107923
## total_sulfur_dioxide -0.407499707  0.230041903  0.10911981  0.03320130
## density              -0.509418108  0.014324268 -0.12109007  0.02674108
## pH                    0.125162028  0.588969085  0.15632069 -0.07997910
## sulphates            -0.042943776  0.217041481  0.42197810 -0.47868974
## alcohol               0.434550881 -0.038224314  0.09701470  0.12759604
##                              PC5         PC6         PC7         PC8
## fixed_acidity         0.26878087 -0.06472727  0.16922865  0.58357684
## volatile_acidity      0.53830588  0.31422082 -0.26823733  0.02074247
## citric_acid           0.01675699  0.19774098 -0.68700728 -0.16528670
## residual_sugar        0.11205410 -0.24747698 -0.21845419 -0.39726652
## chlorides            -0.48424774  0.26118796  0.08255278 -0.10067996
```

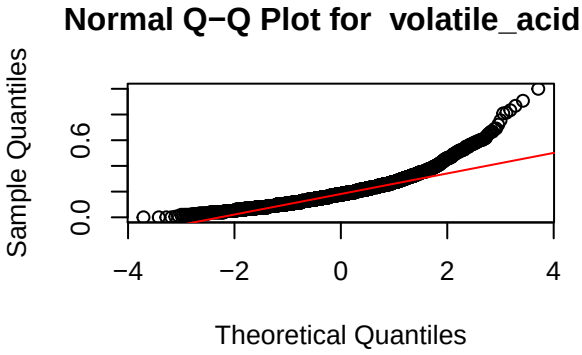
```
## free_sulfur_dioxide   0.04686088  0.52646138  0.21483231 -0.07049206
## total_sulfur_dioxide  0.20618801  0.34550860  0.11352340  0.27480036
## density               0.01564586 -0.33640646 -0.14135481  0.05684108
## pH                   -0.05330177 -0.18021821 -0.44779960  0.51661289
## sulphates             0.50729596 -0.34716409  0.29053571 -0.27572590
## alcohol               0.29241466  0.25651816 -0.09453600 -0.19261896
##                              PC9        PC10        PC11
## fixed_acidity         0.339448738  0.13278152 -0.17083659
## volatile_acidity     -0.138541793  0.22780850 -0.01502882
## citric_acid          -0.204053766  0.01880930 -0.01060051
## residual_sugar        0.401498212 -0.11794187 -0.49090065
## chlorides             0.394603773 -0.06556196 -0.02592329
## free_sulfur_dioxide   0.155274276  0.56608848  0.03327958
## total_sulfur_dioxide -0.173553302 -0.69730025 -0.03707260
## density               0.090320388  0.05922604  0.75935396
## pH                    0.269197328  0.10791836 -0.14060742
## sulphates            -0.008032817  0.06080346 -0.04207132
## alcohol               0.610007775 -0.29389615  0.35802288
```

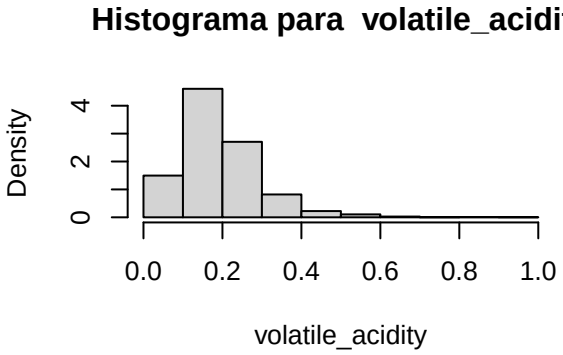
Para revisar si las variables pueden ser candidatas a la normalización, analizaremos las gráficas de quantile-quantile plot y el histograma.

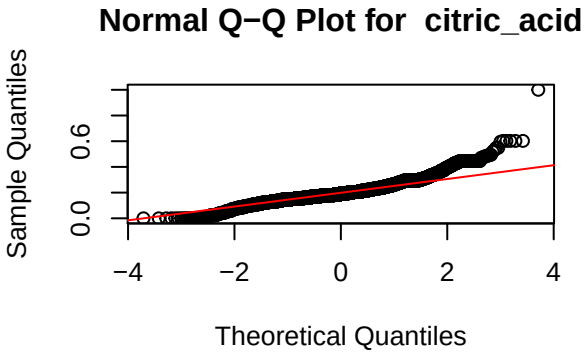
```
par(mfrow=c(2,2))

for(i in 1:ncol(vinoBlancoMatrizRescalado)) {
  qqnorm(vinoBlancoMatrizRescalado[,i],
         main = paste("Normal Q-Q Plot for ",colnames(vinoBlancoMatrizRescalado)[i]))
  qqline(vinoBlancoMatrizRescalado[,i],col="red")
  hist(vinoBlancoMatrizRescalado[,i],
  main=paste("Histograma para ", colnames(vinoBlancoMatrizRescalado)[i]),
  xlab=colnames(vinoBlancoMatrizRescalado)[i], freq = FALSE)
}
```

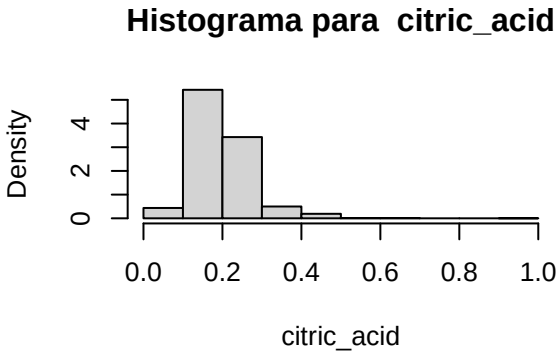

Normal Q−Q Plot for fixed_acidity
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

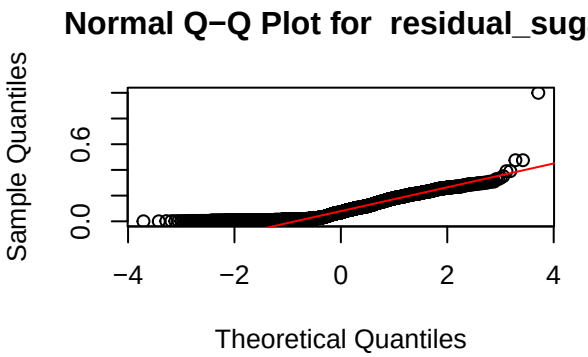

Histograma para fixed_acidity
Density
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
fixed_acidity

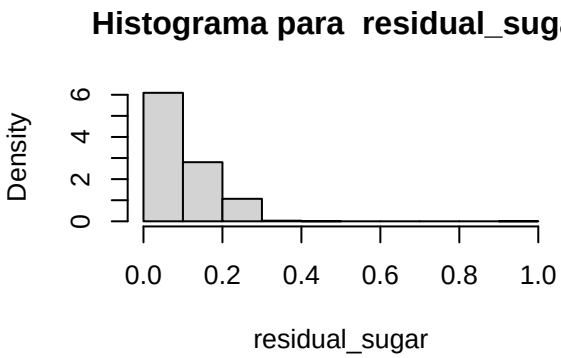

Normal Q−Q Plot for volatile_acidity
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

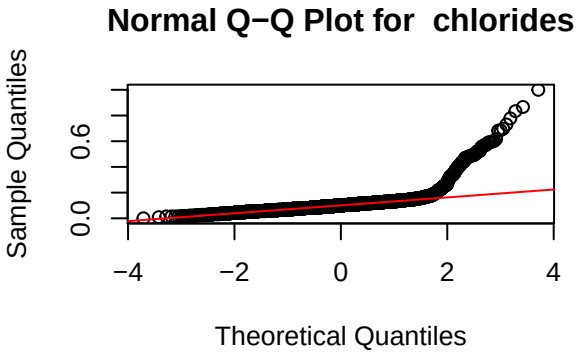

Histograma para volatile_acidity
Density
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
volatile_acidity

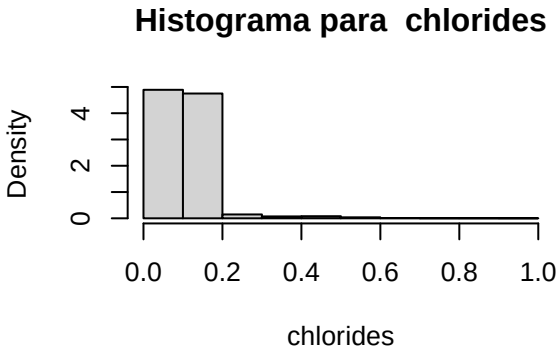

Normal Q−Q Plot for citric_acid
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

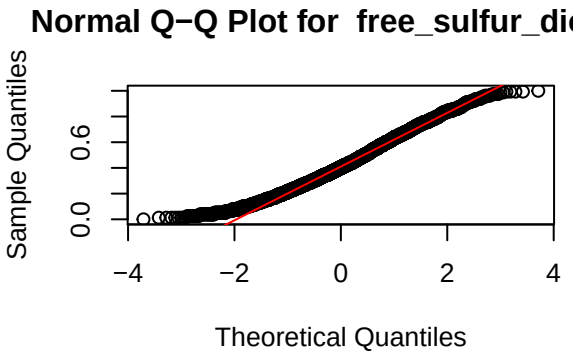

Histograma para citric_acid
Density
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
citric_acid

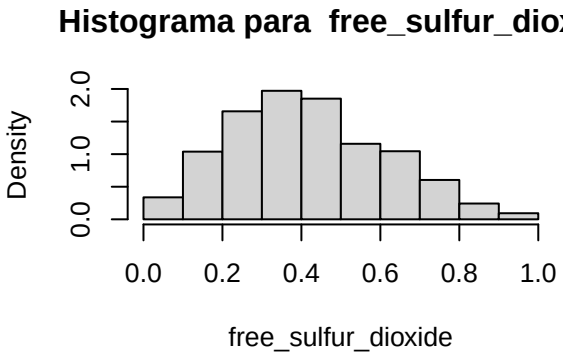

Normal Q−Q Plot for residual_sugar
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

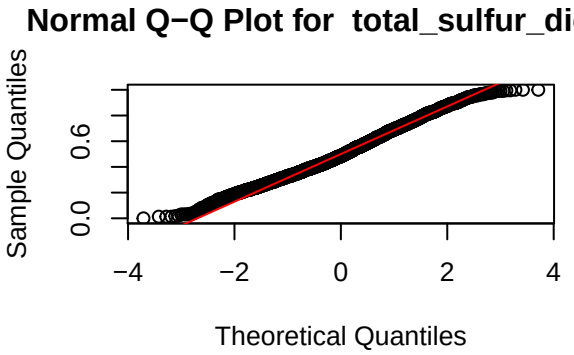

Histograma para residual_sugar
Density
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
residual_sugar

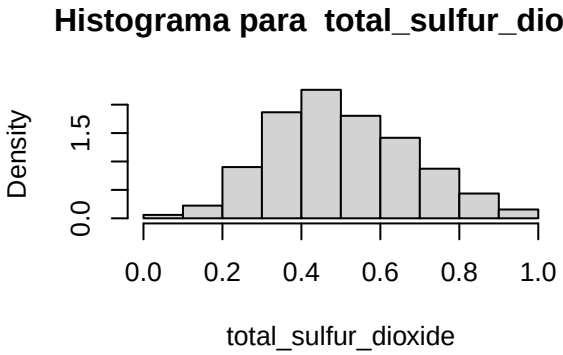

Normal Q−Q Plot for chlorides
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

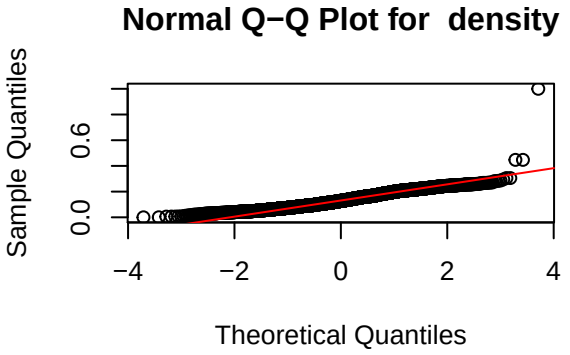

Histograma para chlorides
Density
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
chlorides

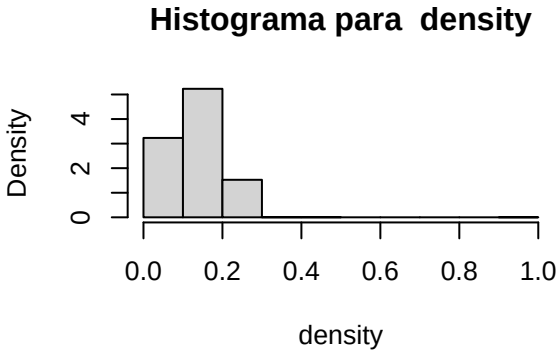

Normal Q−Q Plot for free_sulfur_dioxi
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

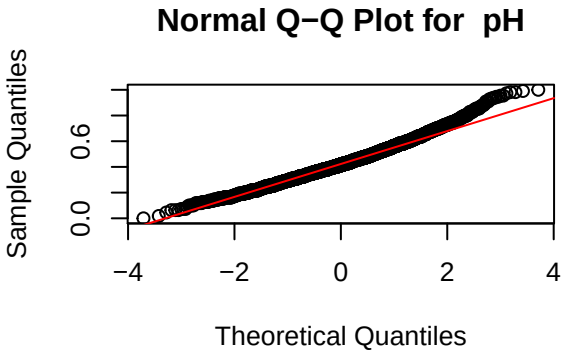

Histograma para free_sulfur_dioxide
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
free_sulfur_dioxide

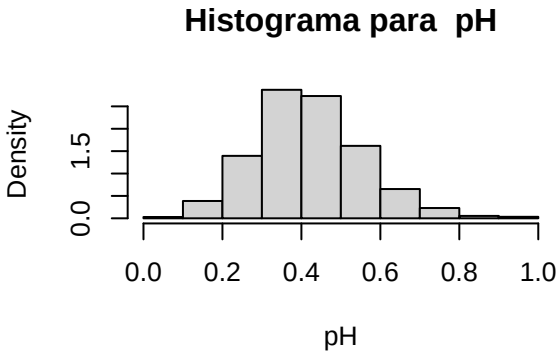

Normal Q−Q Plot for total_sulfur_dioxi
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

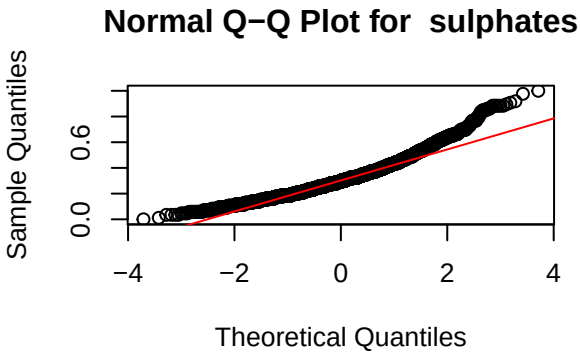

Histograma para total_sulfur_dioxide
Density
0.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
total_sulfur_dioxide

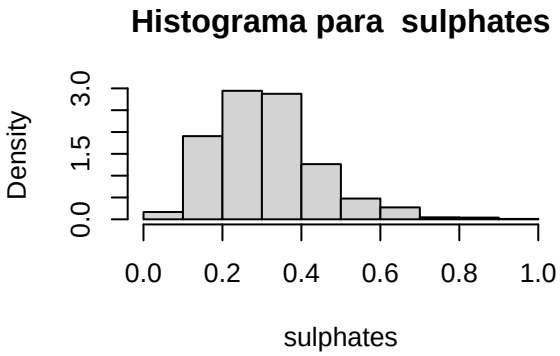

Normal Q−Q Plot for density
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

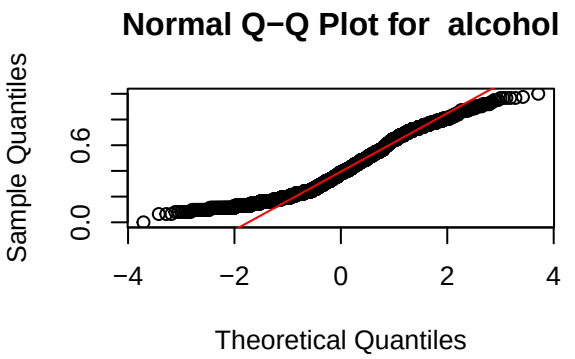

Histograma para density
Density
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
density

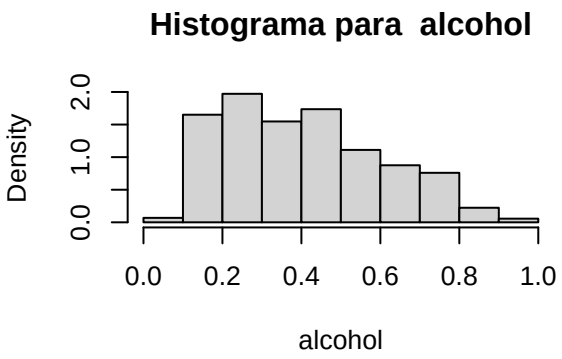

Normal Q−Q Plot for pH
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles

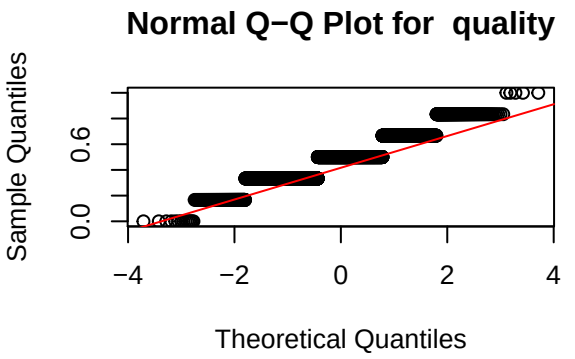

Histograma para pH
Density
0.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
pH

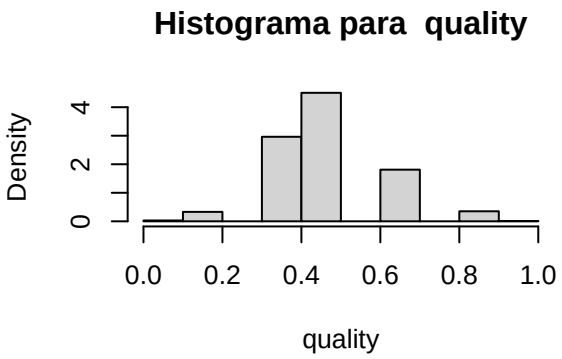

Normal Q−Q Plot for sulphates
Sample Quantiles
0.0
0.6
−4
−2
0
2
4
Theoretical Quantiles


Histograma para sulphates
Density
0.0
1.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
sulphates









Los resultados del quantile-quantile plot nos indica que si las variables siguen una distribución normalizada o no.

Para revisar si las variables están normalizadas, aplicaremos el test de Shapiro Wilk en cada atributo numérico. Éste se basa en asumir como hipótesis nula que la población está distribuida normalmente y, si el p-valor es menor al nivel de significancia (se suele tomar el valor 0,05), entonces la hipótesis nula es rechazada y se concluye que los datos no tienen una distribución normal. Si, por el contrario, el p-valor es mayor a 0,05, se concluye que no se puede rechazar dicha hipótesis y se asume que los datos siguen una distribución normal.

```
for(i in 1:ncol(vinoBlancoMatrizRescalado)) {
  print(concatenate("Para el atributo ",names(vinoBlancoMatrizRescalado)[i],":"))
  print(shapiro.test(vinoBlancoMatrizRescalado[,i]))
}
```

```
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.9768, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
```

```
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.9055, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.92427, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.88385, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.58954, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.98635, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.99222, p-value = 1.288e-15
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.95404, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.98811, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
```

```
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.95258, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.95516, p-value < 2.2e-16
##
## [1] "Para el atributo   :"
##
##  Shapiro-Wilk normality test
##
## data:  vinoBlancoMatrizRescalado[, i]
## W = 0.88776, p-value < 2.2e-16
```

El test de Shapiro-Wilk nos indica que ninguna variable está normalizada, ya que el p-value es menor que 0,05, por lo que podemos rechazar la hipótesis nula y entender que el conjunto de datos no sigue una distribución normal.

Ahora estudiaremos si existe correlación entre los atributos mediante el modelo matemático de la regresión lineal.

Para ello, nos llevaremos a arrays la información de cada campo para que quede más claro el código al invocar a la función linear models. La variable quality no la usaremos aqu al ser la variable sobre la que estableceremos el grado de relación.

```
vinoBlancoMatrizRescalado <- as.data.frame(vinoBlancoMatrizRescalado)
colnames(vinoBlancoMatrizRescalado)<-c("fixed_acidity", "volatile_acidity", "citric_acid",
                                       "residual_sugar", "chlorides", "free_sulfur_dioxide",
                                       "total_sulfur_dioxide", "density", "pH", "sulphates",
                                       "alcohol", "quality")

fixed_acidity = vinoBlancoMatrizRescalado$fixed_acidity
volatile_acidity = vinoBlancoMatrizRescalado$volatile_acidity
citric_acid = vinoBlancoMatrizRescalado$citric_acid
residual_sugar = vinoBlancoMatrizRescalado$residual_sugar
chlorides = vinoBlancoMatrizRescalado$chlorides
free_sulfur_dioxide = vinoBlancoMatrizRescalado$free_sulfur_dioxide
total_sulfur_dioxide = vinoBlancoMatrizRescalado$total_sulfur_dioxide
density = vinoBlancoMatrizRescalado$density
pH = vinoBlancoMatrizRescalado$pH
sulphates = vinoBlancoMatrizRescalado$sulphates
alcohol = vinoBlancoMatrizRescalado$alcohol
```

Aquí utilizaremos las variables modeloX para invocar al modelo linear model. La variable quality será la variable respuesta y el resto de variables las utilizaremos como variables regresoras.

```
modelo1 <- lm(quality ~ fixed_acidity + volatile_acidity +
                citric_acid, data = vinoBlancoMatrizRescalado)
```

```
modelo2 <- lm(quality ~ residual_sugar + chlorides +
                free_sulfur_dioxide, data = vinoBlancoMatrizRescalado)
```

```
modelo3 <- lm(quality ~ total_sulfur_dioxide +
                density + pH, data = vinoBlancoMatrizRescalado)
```

```
modelo4 <- lm(quality ~ sulphates +
                alcohol, data = vinoBlancoMatrizRescalado)
```

En los objetos modeloX hemos guardado una lista con información relevante sobre el análisis de nuestras variables del data frame respecto a la variable quality. A través del comando summary se muestra la información más relevante.

Nos vamos a centrar en el valor del coeficiente de correlación al cuadrado Multiple R-squared, ya que este valor es el que nos indica cómo de buena es la relación entre la variable respuesta y la variable regresora.

Puesto que obtenemos unos valores de Multiple R-squared bajos para todos los modelos, vamos a realizar un único modelo con todas las variables.

```
tabla_coeficientes <- matrix(
  c(1, summary(modelo1)$r.squared,
    2, summary(modelo2)$r.squared,
    3, summary(modelo3)$r.squared,
    4, summary(modelo4)$r.squared),
ncol = 2, byrow = TRUE)
colnames(tabla_coeficientes) <- c("Modelo", "R^2")
tabla_coeficientes
```

```
##      Modelo        R^2
## [1,]      1 0.04884424
## [2,]      2 0.05924516
## [3,]      3 0.09978506
## [4,]      4 0.19639304
```

En este caso, el valor de Adjusted R-squared mejora, pero sigue siendo muy bajo. Ésto se debe a que los datos no están distribuidos normalmente, tal y como hemos comprobado anteriormente, por lo que la regresión no nos da garantías para realizar predicciones del valor de la variable en cuestión.

```
model_tot <- lm(quality ~ fixed_acidity + volatile_acidity + citric_acid +
                  chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
                  density + pH + sulphates + alcohol,  data = vinoBlancoMatrizRescalado)
summary(model_tot)
```

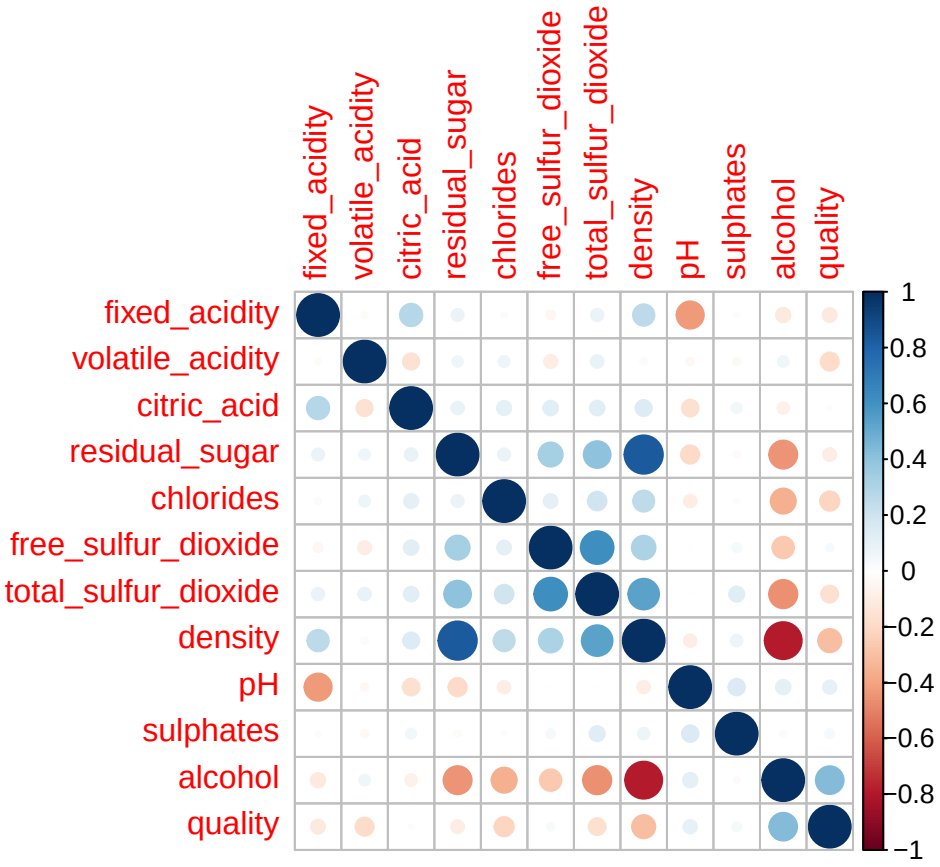
```
##
## Call:
## lm(formula = quality ~ fixed_acidity + volatile_acidity + citric_acid +
##     chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
##     density + pH + sulphates + alcohol, data = vinoBlancoMatrizRescalado)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52992 -0.08359 -0.00550  0.07782  0.53954
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.337444   0.016860  20.014  < 2e-16 ***
## fixed_acidity        -0.127784   0.026969  -4.738 2.22e-06 ***
## volatile_acidity     -0.312468   0.019444 -16.070  < 2e-16 ***
## citric_acid          -0.022607   0.026786  -0.844  0.39871
## chlorides            -0.085038   0.030117  -2.824  0.00477 **
## free_sulfur_dioxide   0.105486   0.011986   8.801  < 2e-16 ***
## total_sulfur_dioxide -0.034614   0.014916  -2.321  0.02035 *
## density               0.350611   0.056170   6.242 4.69e-10 ***
## pH                    0.002379   0.014998   0.159  0.87398
## sulphates             0.043367   0.014030   3.091  0.00201 **
## alcohol               0.411394   0.015732  26.150  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1251 on 4822 degrees of freedom
## Multiple R-squared:  0.2721, Adjusted R-squared:  0.2706
## F-statistic: 180.3 on 10 and 4822 DF,  p-value: < 2.2e-16
```

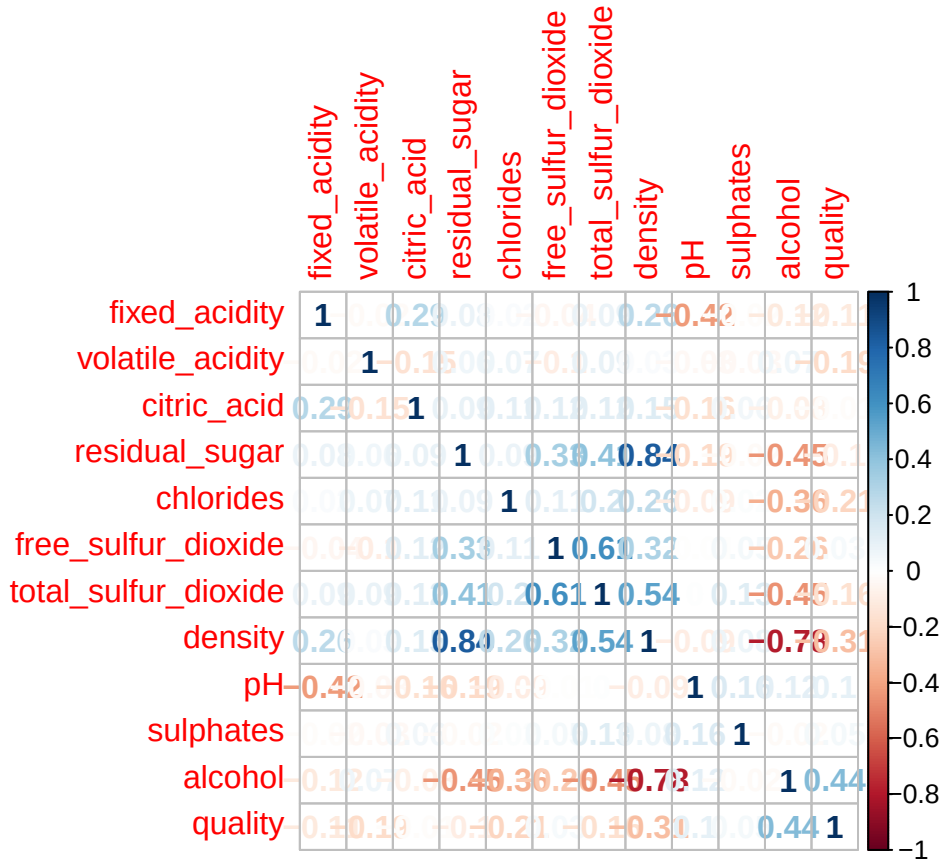
Ahora mostraremos gráficamente la correlación que presentan nuestras variables, es decir, la relación de dependencia que guarda cada variable con el resto de variables. Nos centraremos en la variable quality, para ver qué variable ejerce mayor influencia sobre la variable quality.

Con las 3 presentaciones de los resultados, vemos que la variable que tiene mayor relación es la variable alcohol, cuyo valor es de 0,44. Este valor indica que existe poca relación de las variables respecto a la variable quality.

```
correlacion <- cor(vinoBlancoMatrizRescalado)
corrplot(correlacion)
```




```
corrplot(correlacion, method = "number")
```




```
print(correlacion)
```

```
##                       fixed_acidity volatile_acidity citric_acid residual_sugar
## fixed_acidity            1.00000000      -0.02199479  0.28570302     0.08359885
## volatile_acidity        -0.02199479       1.00000000 -0.15033166     0.06461660
## citric_acid              0.28570302      -0.15033166  1.00000000     0.09078646
## residual_sugar           0.08359885       0.06461660  0.09078646     1.00000000
## chlorides                0.02098254       0.07112885  0.11484435     0.08592385
## free_sulfur_dioxide     -0.04452320      -0.09908667  0.12164124     0.33117568
## total_sulfur_dioxide     0.08845628       0.09101142  0.12304592     0.40929991
## density                  0.26195304       0.02616325  0.14887519     0.83797017
## pH                      -0.42215123      -0.03205815 -0.16039788    -0.19120232
## sulphates               -0.01568552      -0.03098503  0.05933434    -0.02443065
## alcohol                 -0.11874281       0.06684190 -0.07852277    -0.44877069
## quality                 -0.11143033      -0.18813532 -0.01319817    -0.09633614
##                        chlorides free_sulfur_dioxide total_sulfur_dioxide
## fixed_acidity         0.02098254       -0.044523198          0.088456280
## volatile_acidity      0.07112885       -0.099086674          0.091011418
## citric_acid           0.11484435        0.121641241          0.123045925
## residual_sugar        0.08592385        0.331175678          0.409299915
## chlorides             1.00000000        0.110569991          0.199789242
## free_sulfur_dioxide   0.11056999        1.000000000          0.610789327
## total_sulfur_dioxide  0.19978924        0.610789327          1.000000000
## density               0.25532714        0.319644631          0.537987157
## pH                   -0.09087247       -0.009576975          0.002623095
```

```
## sulphates             0.01673504         0.047317031          0.130389742
## alcohol              -0.35969884        -0.263689418         -0.454675917
## quality              -0.21307358         0.033695167         -0.163712503
##                          density           pH   sulphates     alcohol
## fixed_acidity         0.26195304 -0.422151226 -0.01568552 -0.11874281
## volatile_acidity      0.02616325 -0.032058146 -0.03098503  0.06684190
## citric_acid           0.14887519 -0.160397877  0.05933434 -0.07852277
## residual_sugar        0.83797017 -0.191202321 -0.02443065 -0.44877069
## chlorides             0.25532714 -0.090872472  0.01673504 -0.35969884
## free_sulfur_dioxide   0.31964463 -0.009576975  0.04731703 -0.26368942
## total_sulfur_dioxide  0.53798716  0.002623095  0.13038974 -0.45467592
## density               1.00000000 -0.090519888  0.07840775 -0.78040372
## pH                   -0.09051989  1.000000000  0.15502567  0.11987421
## sulphates             0.07840775  0.155025672  1.00000000 -0.02049640
## alcohol              -0.78040372  0.119874213 -0.02049640  1.00000000
## quality              -0.30674549  0.102842224  0.04633325  0.43969153
##                          quality
## fixed_acidity        -0.11143033
## volatile_acidity     -0.18813532
## citric_acid          -0.01319817
## residual_sugar       -0.09633614
## chlorides            -0.21307358
## free_sulfur_dioxide   0.03369517
## total_sulfur_dioxide -0.16371250
## density              -0.30674549
## pH                    0.10284222
## sulphates             0.04633325
## alcohol               0.43969153
## quality               1.00000000
```

4 Conclusiones finales

De los resultados obtenidos con el test de Shapiro-Wilk se deduce que el juego de datos no sigue una distribución normal.

Del estudio de la correlación entre las variables se deduce que la densidad y el azúcar residual guardan una fuerte correlación entre sí (del 84%). Asimismo, se puede afirmar que el alcohol y la densidad tienen un alto grado de correlación (de más del 70%).

Aplicando el análisis de componentes principales obtenemos que somos capaces de predecir aproximadamente el 90% del valor de la calidad de las muestras analizando únicamente 7 componentes principales que no tienen correlación entre sí.

El éxito de este estudio ha sido parcial, ya que no se ha conseguido establecer ninguna correlación fuerte entre la calidad y los atributos del juego de datos de forma directa. Sin embargo, se ha conseguido obtener una combinación de 7 componentes principales (cada una de las cuales podría representar a un tipo de vino que tuviese las características correspondientes a cada componente principal) para así determinar la composición de cada tipo de vino en función de la calidad de la muestra que se desee analizar.