

Práctica 2: Limpieza y análisis de datos

Daria Gracheva, Zechao Jin

24 de December, 2020

Contents

Descripción del dataset.	1
Integración y selección de los datos de interés a analizar.	2
Limpieza de los datos	3
Valores nulos	5
Unidades de medida	5
Outliers	6
Discretización	10
Balanceo de clase color	14
Análisis	14
Analisis inferencial	17
Estudio de correlación entre las variables	23
Regresión multiple	35
Modelo no supervisado (clustering)	38
Conclusión / Resolución del problema.	47

Descripción del dataset.

Hemos elegido el dataset “Wine quality dataset”, un dataset de kaggle cuya fuente es el repositorio UCI Machine Learning (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>; <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>). Se trata de un dataset de 6497 observaciones, 11 atributos y 1 marca de clase.

Resumen de datos: más precisamente son 2 datasets separados, para tipos tinto y blanco del vino verde, el de vinos tintos contiene 1599 observaciones, y el de vinos blancos - 4898 observaciones, con que tenemos opción de trabajar con datasets separadamente o bien unir todas las observaciones y obtener la variebilidad natural tinto/blanco.

Las clases de dataset corresponden a la calidad del vino basandose en sus características físicas y químicas (sensory data), en una escala de 0 a 10. El paradigma de calidad y las características son facilmente interpretables y el ámbito del tema es generalmente conocido pues tiene un alto potencial de dar un resultado divulgativo.

En cuanto a la finalidad del estudio, sería interesante jugar con los datos para estimar que características influyen a la calidad y/o representan el color. ¿Las características de calidad varian segun el color del vino? Como podemos distribuir la calidad para entenderlo empiricamente -qué agrupaciones podemos obtener y cual es el número de “categorías de calidad” más óptimo según los algoritmos?, o bien si podemos estudiar los vinos partiendo de otras variables.

Librerías

```
# Cargamos las librerías
library(ggplot2) # visualization
library(patchwork)
```

```

library(tidyverse)
library(corrplot)
library(factoextra) # clustering
library(Hmisc) # impute
library(arules) # discretize
library(DMwR) # smote
library(DescTools) # box-cox

```

Integración y selección de los datos de interés a analizar.

Cargamos los dos datasets y concatenamos los datos de vinos tintos y blancos:

```

# Carga del dataset
red <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv')

white <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv')

# Nombres de los atributos
names(red) <- c("fixed acidity", "volatile acidity", "citric acid", "residual sugar", "chlorides", "free sulfur dioxide", "density", "pH", "sulphates", "alcohol", "quality")
names(white) <- c("fixed acidity", "volatile acidity", "citric acid", "residual sugar", "chlorides", "free sulfur dioxide", "density", "pH", "sulphates", "alcohol", "quality")

# Creamos las columnas de color
red['color'] <- 'red'
white['color'] <- 'white'

# Fusionamos los dos datasets
vinos <- rbind(white,red)

# Factorizamos la variable color
vinos$color <- as.factor(vinos$color)

# Verificamos la estructura del conjunto de datos
str(vinos)

## 'data.frame': 6497 obs. of 13 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 ...
## $ color : Factor w/ 2 levels "red","white": 2 2 2 2 2 2 2 2 2 2 ...

```

Se observa que tenemos todas las variables numéricas menos el atributo de color.

Descripción de variables: son las características físicas o químicas del vino más su calidad y su color/tipo (tinto o blanco)

fixed acidity : acidez fija (g/l), v. continua

volatile acidity : acidez volatil (g/l), v. continua
 citric acid : acido citrico (g/l), v. continua
 residual sugar : azucar residual (g/l), v. continua
 chlorides : cloruros (g/l), v. continua
 free sulfur dioxide : dioxido de azufre libre (mg/l), v. continua
 total sulfur dioxide: dioxido de azufre total (mg/l), v. continua
 density : densidad (g/l), v. continua
 pH : pH, v. continua
 sulphates : sulfatos (g/l), v. continua
 alcohol : concentración de alcohol (%), v. continua
 quality : calidad, v. discreta
 color : color, v. cualitativa dicotómica

Veamos como son algunas de las observaciones:

```
rbind(head(vinos,5), tail(vinos,5))
```

```

##      fixed acidity volatile acidity citric acid residual sugar chlorides
## 1          7.0          0.270       0.36        20.7     0.045
## 2          6.3          0.300       0.34        1.6      0.049
## 3          8.1          0.280       0.40        6.9      0.050
## 4          7.2          0.230       0.32        8.5      0.058
## 5          7.2          0.230       0.32        8.5      0.058
## 6493       6.2          0.600       0.08        2.0      0.090
## 6494       5.9          0.550       0.10        2.2      0.062
## 6495       6.3          0.510       0.13        2.3      0.076
## 6496       5.9          0.645       0.12        2.0      0.075
## 6497       6.0          0.310       0.47        3.6      0.067
##      free sulfur dioxide total sulfur dioxide density   pH sulphates
## 1              45          170 1.00100 3.00      0.45
## 2              14          132 0.99400 3.30      0.49
## 3              30             97 0.99510 3.26      0.44
## 4              47            186 0.99560 3.19      0.40
## 5              47            186 0.99560 3.19      0.40
## 6493         32             44 0.99490 3.45      0.58
## 6494         39             51 0.99512 3.52      0.76
## 6495         29             40 0.99574 3.42      0.75
## 6496         32             44 0.99547 3.57      0.71
## 6497         18             42 0.99549 3.39      0.66
##      alcohol quality color
## 1          8.8      6 white
## 2          9.5      6 white
## 3         10.1      6 white
## 4          9.9      6 white
## 5          9.9      6 white
## 6493       10.5      5 red
## 6494       11.2      6 red
## 6495       11.0      6 red
## 6496       10.2      5 red
## 6497       11.0      6 red

```

Limpieza de los datos

Veamos las estadísticas básicas:

```
summary(vinos)
```

```
## fixed acidity      volatile acidity      citric acid      residual sugar
## Min.   : 3.800    Min.   :0.0800    Min.   :0.0000    Min.   : 0.600
## 1st Qu.: 6.400    1st Qu.:0.2300    1st Qu.:0.2500    1st Qu.: 1.800
## Median : 7.000    Median :0.2900    Median :0.3100    Median : 3.000
## Mean   : 7.215    Mean   :0.3397    Mean   :0.3186    Mean   : 5.443
## 3rd Qu.: 7.700    3rd Qu.:0.4000    3rd Qu.:0.3900    3rd Qu.: 8.100
## Max.   :15.900    Max.   :1.5800    Max.   :1.6600    Max.   :65.800
## chlorides        free sulfur dioxide total sulfur dioxide
## Min.   :0.00900   Min.   : 1.00     Min.   : 6.0
## 1st Qu.:0.03800   1st Qu.: 17.00    1st Qu.: 77.0
## Median :0.04700   Median : 29.00    Median :118.0
## Mean   :0.05603   Mean   : 30.53    Mean   :115.7
## 3rd Qu.:0.06500   3rd Qu.: 41.00    3rd Qu.:156.0
## Max.   :0.61100   Max.   :289.00    Max.   :440.0
## density          pH            sulphates      alcohol
## Min.   :0.9871    Min.   :2.720    Min.   :0.2200    Min.   : 8.00
## 1st Qu.:0.9923    1st Qu.:3.110    1st Qu.:0.4300    1st Qu.: 9.50
## Median :0.9949    Median :3.210    Median :0.5100    Median :10.30
## Mean   :0.9947    Mean   :3.219    Mean   :0.5313    Mean   :10.49
## 3rd Qu.:0.9970    3rd Qu.:3.320    3rd Qu.:0.6000    3rd Qu.:11.30
## Max.   :1.0390    Max.   :4.010    Max.   :2.0000    Max.   :14.90
## quality          color
## Min.   :3.000    red   :1599
## 1st Qu.:5.000    white:4898
## Median :6.000
## Mean   :5.818
## 3rd Qu.:6.000
## Max.   :9.000
```

Todas las columnas parecen ser bastante limpias, no obstante aquí se observa la heterogeneidad de las variables (por ejemplo, cloruros que no superan 0.611 g/l y dioxido de sulfuro total que “alcanza” 440 mg/l: dado que tienen las unidades distintas se produce esta brecha).

En cuanto a la calidad, el valor mínimo es 3 y el máximo es 9. Por lo que 1-10 es una escala “común”, y la escala real sería 3-9. Comprobamos la distribución de calidad:

```
table(vinos$quality)
```

```
##
##   3    4    5    6    7    8    9
##  30  216 2138 2836 1079  193   5
```

Con el atributo “quality” podemos tener 7 marcas de clase diferentes, aunque puede ser conveniente agruparlo en una variable discreta como vemos más adelante.

Distribución de valores únicos de atributos:

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

```
##      fixed acidity      volatile acidity      citric acid
##             106                  187                  89
##      residual sugar      chlorides      free sulfur dioxide
##             316                  214                  135
##      total sulfur dioxide      density          pH
##              276                 998                 108
```

```

##          sulphates           alcohol          quality
##          111                  111                  7
##          color
##          2

```

Algunos atributos tienen la distribución bastante dispersa, que, si fuera necesario, podríamos agrupar en intervalos.

Valores nulos

Comprobamos si hay valores nulos en en dataset

```
any(is.na(vinos))
```

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

```
any(vinos=="")
```

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

A partir de las estadísticas se ve que todas las variables tienen un mínimo distinto del cero menos la variable “citric acid”, y podría tomar un 0 como un valor desconocido.

Veamos la distribución de “citric acid”:

```
table(vinos$`citric acid`)
```

```

##
##      0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 0.11 0.12 0.13 0.14
##    151   40   56   32   41   25   30   34   37   42   49   16   46   35   48
##  0.15 0.16 0.17 0.18 0.19 0.2 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
##   42   42   43   71   69   95   99  131  108  232  163  257  236  301  244
##  0.3 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.4 0.41 0.42 0.43 0.44
##  337  230  289  208  249  150  197  153  136  129  146  98  124  52  86
##  0.45 0.46 0.47 0.48 0.49 0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
##   68   70   56   62  283   55   38   40   30   32   23   30   22   30   14
##  0.6 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.7 0.71 0.72 0.73 0.74
##   15   11   15   14   15   15   21    9   18    9    5   10    6    8   45
##  0.75 0.76 0.78 0.79 0.8 0.81 0.82 0.86 0.88 0.91 0.99    1  1.23  1.66
##    1    3    3    3    2    2    2    1    1    2    1    6    1    1

```

Observación “0” aparece 151 veces que supone apr. 2% de la distribución y es comparable con algunas otras frecuencias, por lo que puede ser valor real (=“no siempre se añade el acido citrico”) y no perdido o nulo.

Unidades de medida

Como hemos visto, hay variables que tienen distintas unidades de medidas (hablando de variables de misma naturaleza), podemos reducirlas a las mismas medidas (quedamos con g/l).

Hay dos variables que usan otras unidades: “free/total sulfur dioxide”, que en g/l tendrían la distribución parecida a la de “chlorides”.

```

# Cambio de unidades de mg/l a g/l
vinos$`free sulfur dioxide` <- vinos$`free sulfur dioxide` * 0.001
vinos$`total sulfur dioxide` <- vinos$`total sulfur dioxide` * 0.001

```

Visualizamos de nuevo las estadísticas:

```
summary(vinos)
```

```

##   fixed acidity   volatile acidity   citric acid   residual sugar
##   Min. : 3.800   Min. : 0.0800   Min. : 0.0000   Min. : 0.600
##   1st Qu.: 7.048   1st Qu.: 0.3580   1st Qu.: 0.0450   1st Qu.: 1.000
##   Median : 10.000  Median : 0.5140  Median : 0.0580  Median : 1.680
##   Mean   : 11.287  Mean   : 0.5287  Mean   : 0.0523  Mean   : 1.805
##   3rd Qu.: 14.370  3rd Qu.: 0.5680  3rd Qu.: 0.0698  3rd Qu.: 2.450
##   Max.  : 16.000  Max.  : 1.0000  Max.  : 0.0960  Max.  : 3.995

```

```

## 1st Qu.: 6.400 1st Qu.:0.2300 1st Qu.:0.2500 1st Qu.: 1.800
## Median : 7.000 Median :0.2900 Median :0.3100 Median : 3.000
## Mean   : 7.215 Mean   :0.3397 Mean   :0.3186 Mean   : 5.443
## 3rd Qu.: 7.700 3rd Qu.:0.4000 3rd Qu.:0.3900 3rd Qu.: 8.100
## Max.   :15.900 Max.   :1.5800 Max.   :1.6600 Max.   :65.800
## chlorides      free sulfur dioxide total sulfur dioxide
## Min.   :0.00900 Min.   :0.00100 Min.   :0.0060
## 1st Qu.:0.03800 1st Qu.:0.01700 1st Qu.:0.0770
## Median :0.04700 Median :0.02900 Median :0.1180
## Mean   :0.05603 Mean   :0.03053 Mean   :0.1157
## 3rd Qu.:0.06500 3rd Qu.:0.04100 3rd Qu.:0.1560
## Max.   :0.61100 Max.   :0.28900 Max.   :0.4400
## density          pH           sulphates      alcohol
## Min.   :0.9871  Min.   :2.720  Min.   :0.2200  Min.   : 8.00
## 1st Qu.:0.9923  1st Qu.:3.110  1st Qu.:0.4300  1st Qu.: 9.50
## Median :0.9949  Median :3.210  Median :0.5100  Median :10.30
## Mean   :0.9947  Mean   :3.219  Mean   :0.5313  Mean   :10.49
## 3rd Qu.:0.9970  3rd Qu.:3.320  3rd Qu.:0.6000  3rd Qu.:11.30
## Max.   :1.0390  Max.   :4.010  Max.   :2.0000  Max.   :14.90
## quality          color
## Min.   :3.000  red   :1599
## 1st Qu.:5.000  white:4898
## Median :6.000
## Mean   :5.818
## 3rd Qu.:6.000
## Max.   :9.000

```

Outliers

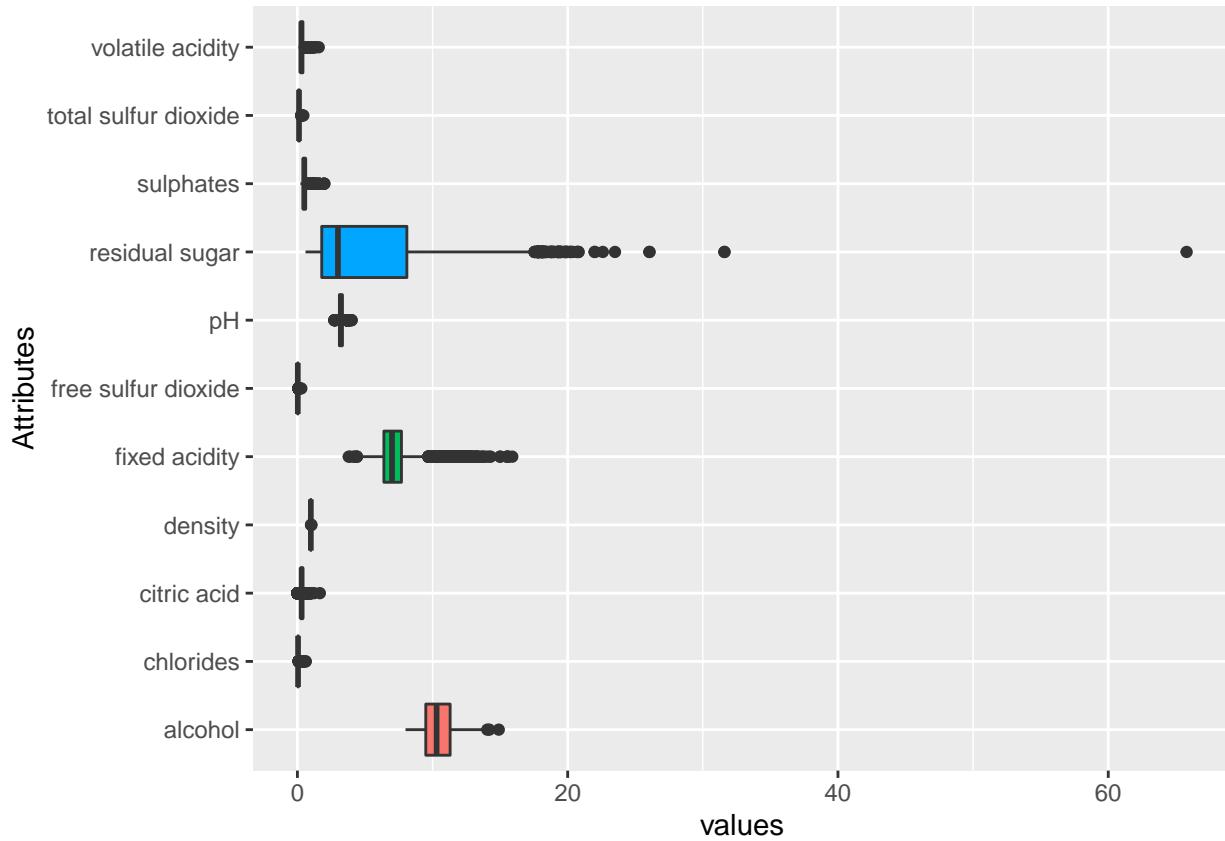
El dataset parece tener outliers ya que en muchas variables la diferencia entre tercer cuantil y el máximo es considerable.

Visualizamos los boxplots de todas las columnas:

```

bp1 <- vinos %>%
  gather(Attributes, values, c(1:11)) %>%
  ggplot(aes(x=Attributes, y=values, fill=Attributes)) + geom_boxplot(show.legend=FALSE) + coord_
bp1

```



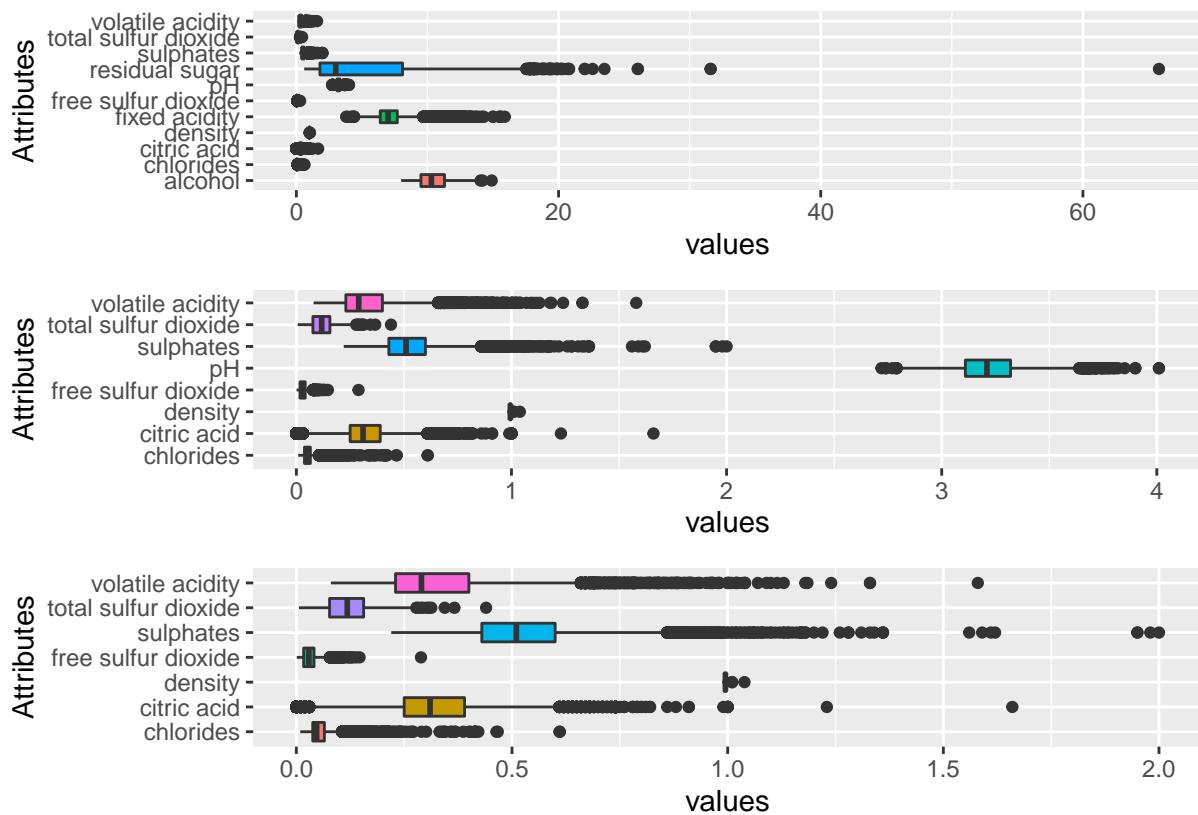
Se observa que la variable “residual sugar” tiene unos outliers muy distantes. Otros atributos tienen una escala diferente por lo que procedemos a visualizarlos sin “residual sugar” con distinta ampliación:

```
p1 <- vinos %>%
  gather(Attributes, values, c(1:3,5:11)) %>%
  ggplot(aes(x=Attributes, y=values, fill=Attributes)) + geom_boxplot(show.legend=FALSE) + coord_flip()

p2 <- vinos %>%
  gather(Attributes, values, c(2:3,5:10)) %>%
  ggplot(aes(x=Attributes, y=values, fill=Attributes)) + geom_boxplot(show.legend=FALSE) + coord_flip()

p3 <- vinos %>%
  gather(Attributes, values, c(2:3,5:8,10)) %>%
  ggplot(aes(x=Attributes, y=values, fill=Attributes)) + geom_boxplot(show.legend=FALSE) + coord_flip()

bp1 + p2 + p3 + plot_layout(nrow = 3)
```



Tenemos valores bastante anómalos, sobre todo en “citric acid”, “free sulfur dioxide”, “sulphates”, “volatile acidity”, “sulphates”, “chlorides”.

No obstante, son relativamente pocas observaciones por lo que se puede realizar una imputación por la mediana.

Por ello, remplazamos los valores extremos segun el estadístico de boxplot por NA:

```
for (x in c("residual sugar", "citric acid", "free sulfur dioxide", "sulphates", "volatile acidity", "chlorides")) {
  vinos[,x][vinos[,x] %in% boxplot.stats(vinos[,x])$out] <- NA
}
```

Cantidad de outliers detectados:

```
sapply(vinos, function(vinos) sum(is.na(vinos)))
```

```
##      fixed acidity      volatile acidity      citric acid
##            357                  377                  509
##      residual sugar      chlorides  free sulfur dioxide
##            118                  286                  62
##      total sulfur dioxide      density          pH
##              10                  3                  73
##      sulphates      alcohol      quality
##             191                  3                  0
##      color
##            0
```

Un ejemplo de observaciones con outliers:

```
vinos[is.na(vinos$alcohol),]
```

```
##      fixed acidity      volatile acidity      citric acid      residual sugar      chlorides
##
```

```

## 3919      6.4      0.35      0.28      1.6      0.037
## 4504      5.8      0.61      NA        8.4      0.041
## 5551      NA       0.36      NA        7.5      0.096
##   free sulfur dioxide total sulfur dioxide density    pH sulphates
## 3919          0.031           0.113 0.98779 3.12      0.40
## 4504          0.031           0.104 0.99090 3.26      0.72
## 5551          0.022           0.071 0.99760 2.98      0.84
##   alcohol quality color
## 3919      NA      7 white
## 4504      NA      7 white
## 5551      NA      5 red

```

Podemos ver que algunos valores atípicos se encuentran en las mismas observaciones.

Imputación por la mediana:

```
vinos[,c(1:11)] <- apply(vinos[,c(1:11)], 2, impute)
```

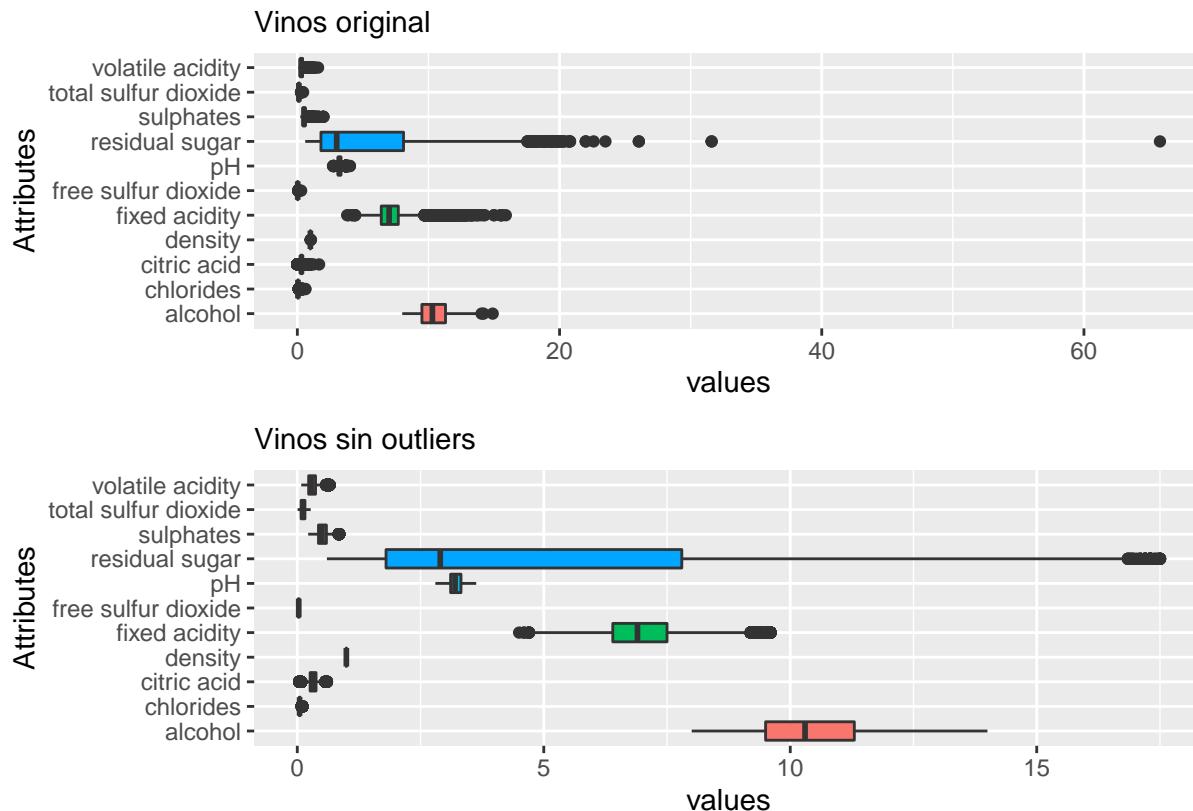
Visualizamos los atributos de nuevo comparando con los boxplots anteriores:

```

bp2 <- vinos %>%
  gather(Attributes, values, c(1:11)) %>%
  ggplot(aes(x=Attributes, y=values, fill=Attributes)) + geom_boxplot(show.legend=FALSE) + coord_flip()

bp1 + labs(subtitle = "Vinos original") + bp2 + labs(subtitle = "Vinos sin outliers") + plot_layout(nrow=2)

```



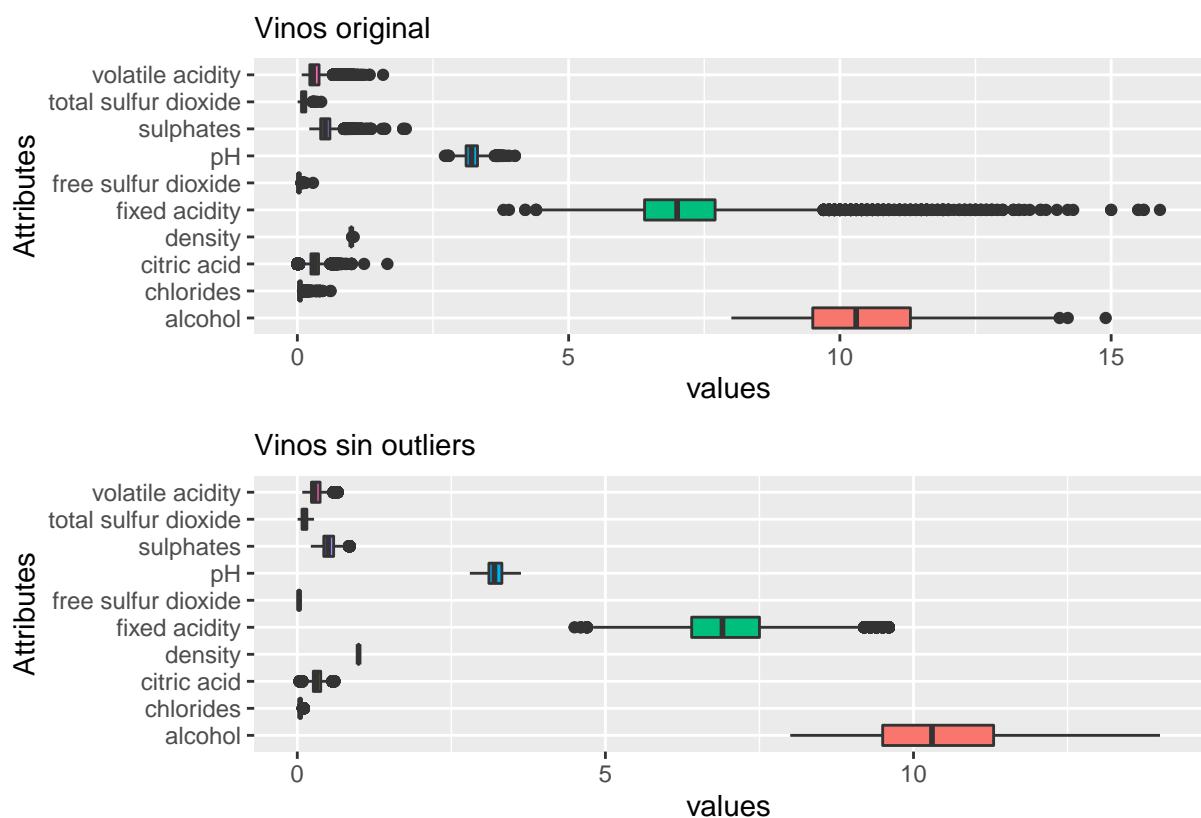
O bien omitiendo residual sugar:

```

p2 <- vinos %>%
  gather(Attributes, values, c(1:3,5:11)) %>%
  ggplot(aes(x=Attributes, y=values, fill=Attributes)) + geom_boxplot(show.legend=FALSE) + coord_flip()

```

```
p1 + labs(subtitle = "Vinos original") + p2 + labs(subtitle = "Vinos sin outliers") + plot_layout(nrow =
```



Aunque seguimos teniendo outliers segun estos boxplots en su definición más teórica, están más agrupados y, considerando que representan la variebilidad real del dataset, eliminamos los valores muy anómalos y demasiado dispersos que podían ser errores o inconsistencias. Por ello, hemos obtenido la distribución mucho menos sesgada.

Discretización

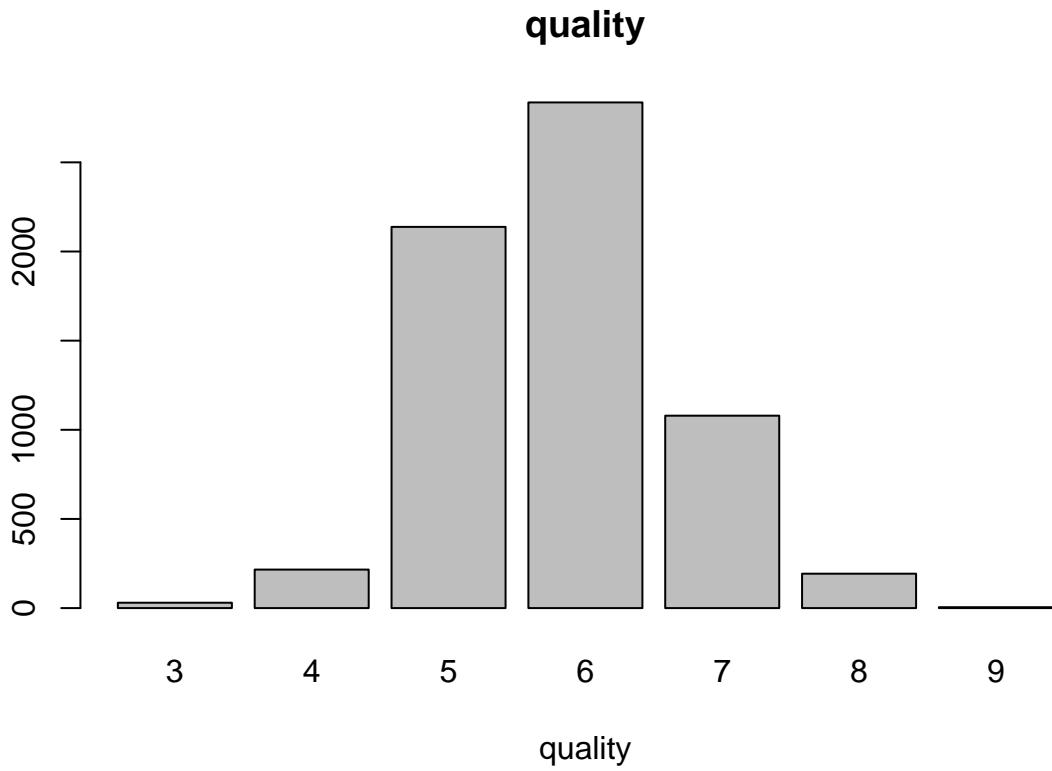
Como se observa, la variable quality no está balanceada, y las clases que tienen pocas observaciones pueden presentar problemas en análisis así que es conveniente crear particiones con más observaciones. Por ello con el fin de equilibrar la marca de calidad y agruparlo de manera natural, se puede realizar la discretización de la variable.

Veamos de nuevo sus estadísticas:

```
summary(vinos$quality)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 3.000   5.000   6.000   5.818   6.000   9.000

barplot(table(vinos$quality), xlab="quality", main = "quality")
```



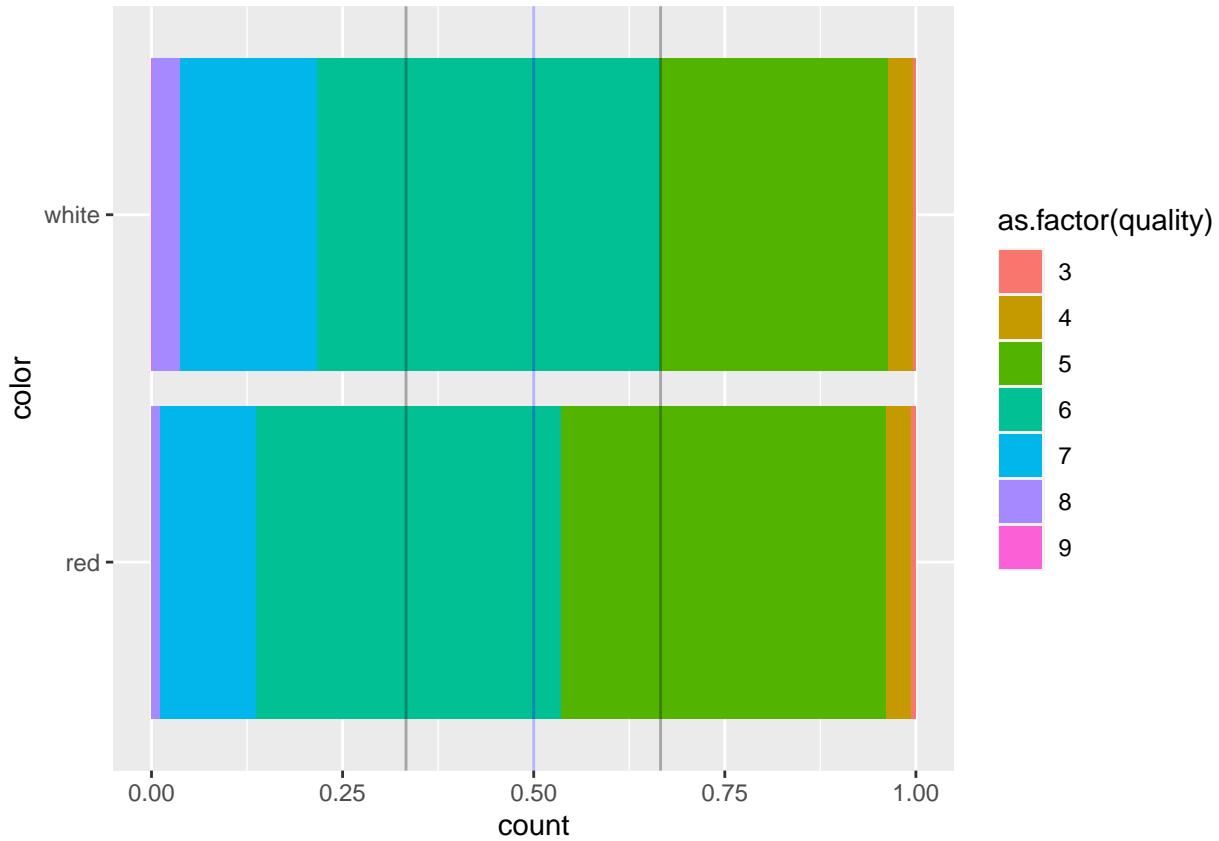
```
# Distribución de valores únicos
table(vinos$quality)
```

```
##
##      3     4     5     6     7     8     9
##    30   216  2138  2836 1079   193     5
```

Tenemos 7 marcas de calidad ordinales de 6471 observaciones. En principio, no sabemos en cuantas particiones podemos agrupar las observaciones.

Logicamente, por frecuencias no se puede crear 4 o más particiones equilibradas ($6471 \div 4 = 1617$, los números de observaciones de clases 5,6 son mayores). Por lo que podríamos crear 2 o 3 clases. Tenemos que seguir la lógica real para agrupar la calidad, por lo que puede ser: alta/media/baja calidad o bien alta/no alta calidad. Visualizamos la distribución de clases segun el color:

```
ggplot(data = vinos,aes(x=color,fill=as.factor(quality)))+geom_bar(position="fill") + geom_hline(yintercept=
```



Para poder trabajar posteriormente con una variable dicotómica de calidad, fijaremos el número de bins de 2, que representaría calidad alta/no alta.

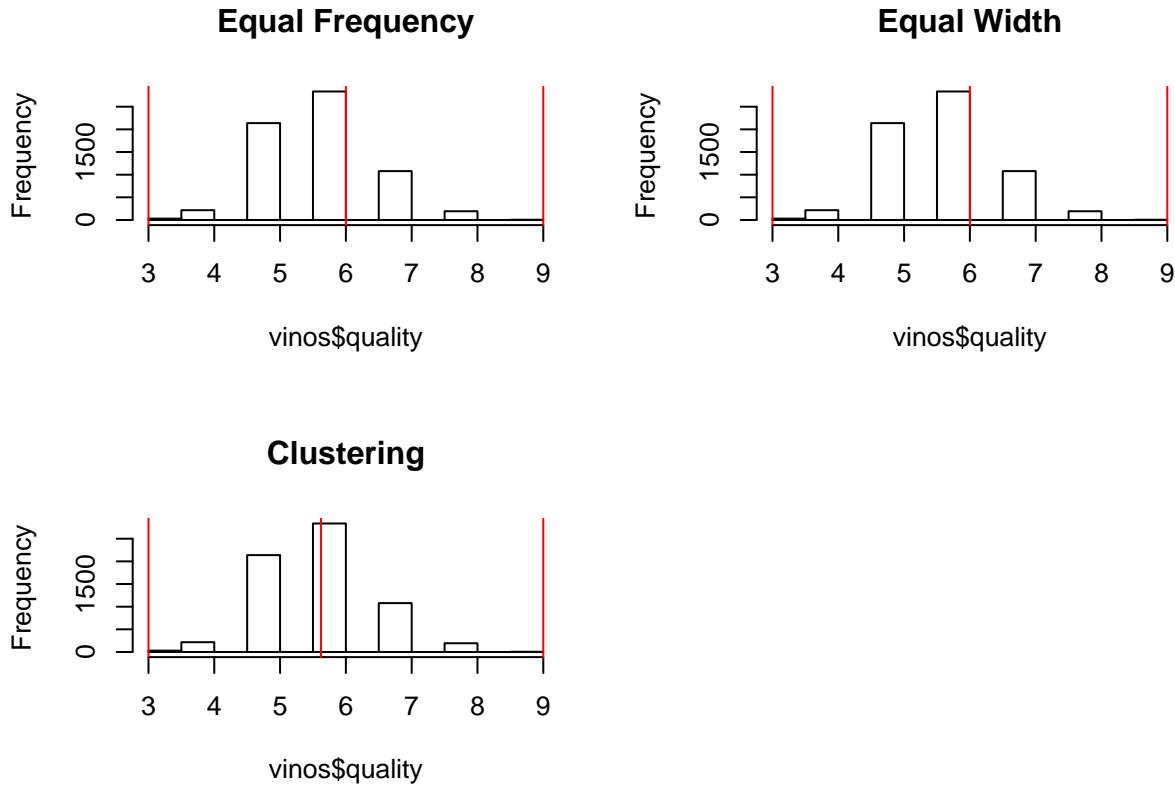
Para poder determinar qué observaciones agrupamos en qué clase de calidad, visualización de clases discretas segun si la particion se hace por igual frecuencia / igual amplitud o clustering:

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

hist(vinos$quality, breaks = 20, main = "Equal Frequency")
abline(v = discretize(vinos$quality, breaks = 2, onlycuts = TRUE), col = "red")

hist(vinos$quality, breaks = 20, main = "Equal Width")
abline(v = discretize(vinos$quality, method = "interval", breaks = 2, onlycuts = TRUE), col = "red")

hist(vinos$quality, breaks = 20, main = "Clustering")
abline(v = discretize(vinos$quality, method = "cluster", breaks = 2, onlycuts = TRUE), col = "red")
```



Creamos las particiones como otro atributo para poder comparar las futuras agrupaciones con valores de quality. Aunque pueda ser redundante ahora, nos podría servir para ajustar las particiones de calidad en futuro.

```

vinos$`quality class`[vinos$quality<=5]="low"
vinos$`quality class`[vinos$quality>5]="high"

vinos$`quality class` <- as.factor(vinos$`quality class`)

# Creamos un factor ordinal
vinos$`quality class` <- factor(vinos$`quality class`, levels = c("low","high"))

# Estructura de la variable
table(vinos$`quality class`)

## 
##   low  high
## 2384 4113

```

Finalmente, convertimos la calidad en factor también:

```

vinos$quality <- as.factor(vinos$quality)

# Variable factorizada
str(vinos$quality)

##  Factor w/ 7 levels "3","4","5","6",...: 4 4 4 4 4 4 4 4 4 4 ...

```

Ahora tenemos un dataset de 6471 observaciones y 14 columnas, 3 de cuales se puede considerar clases / grupos de vinos reales: quality (7 niveles), quality class (2 niveles), color (2 niveles).

Balanceo de clase color

Si podemos considerar que los datos de clase de calidad están ahora balanceados, las observaciones de color no lo estan, por ello usamos el algortimo SMOTE para poder equilibrar las marcas de clase para tareas posteriores de análisis.

```
vinos <- SMOTE(color ~ ., vinos)

dim(vinos)

## [1] 11193    14

table(vinos$color)

## 
##   red white
## 4797 6396

table(vinos$quality)

##
##      3     4     5     6     7     8     9
##     46   368  4061  4870 1568   276     4
```

Con ello tenemos un dataset de 11193 observaciones con vinos tintos/blancos más balanceados gracias a las nuevas observaciones “sinteticas” y como se observa, la calidad tiene la distribución parecida a la anterior.

Análisis

Volvemos a ver las estadísticas básicas de las variables:

```
summary(vinos)

##   fixed acidity  volatile acidity  citric acid  residual sugar
##   Min. :4.500  Min. :0.0800  Min. :0.0400  Min. : 0.600
##   1st Qu.:6.600 1st Qu.:0.2500 1st Qu.:0.2500 1st Qu.: 1.900
##   Median :7.000 Median :0.3000  Median :0.3100  Median : 2.500
##   Mean   :7.124 Mean   :0.3382  Mean   :0.3125  Mean   : 4.465
##   3rd Qu.:7.600 3rd Qu.:0.4200 3rd Qu.:0.3800 3rd Qu.: 6.100
##   Max.   :9.600  Max.   :0.6550  Max.   :0.6000  Max.   :17.500
##
##   chlorides      free sulfur dioxide total sulfur dioxide
##   Min. :0.00900  Min. :0.00100  Min. :0.00600
##   1st Qu.:0.04100 1st Qu.:0.01300 1st Qu.:0.04285
##   Median :0.05088 Median :0.02400 Median :0.09900
##   Mean   :0.05624 Mean   :0.02629 Mean   :0.09819
##   3rd Qu.:0.07400 3rd Qu.:0.03600 3rd Qu.:0.14300
##   Max.   :0.10500  Max.   :0.07700  Max.   :0.27200
##
##   density          pH        sulphates      alcohol
##   Min. :0.9871  Min. :2.800  Min. :0.2200  Min. : 8.00
##   1st Qu.:0.9931 1st Qu.:3.130 1st Qu.:0.4500 1st Qu.: 9.50
##   Median :0.9956 Median :3.230 Median :0.5299 Median :10.26
##   Mean   :0.9952 Mean   :3.236 Mean   :0.5359 Mean   :10.45
##   3rd Qu.:0.9972 3rd Qu.:3.340 3rd Qu.:0.6100 3rd Qu.:11.20
##   Max.   :1.0037  Max.   :3.630  Max.   :0.8500  Max.   :14.00
##
##   quality      color      quality class
##   
```

```

## 3: 46    red :4797    low :4469
## 4: 368    white:6396   high:6724
## 5:4061
## 6:4870
## 7:1568
## 8: 276
## 9:   4

```

Y la desviación típica de las variables continuas:

```
apply(vinos[,c(1:11)], 2, sd)
```

```

##          fixed acidity      volatile acidity      citric acid
##            0.891495850        0.130682910        0.112688389
##          residual sugar      chlorides  free sulfur dioxide
##            3.950839565        0.019803656        0.015978696
##          total sulfur dioxide      density           pH
##            0.058905459        0.002813668        0.150782603
##          sulphates      alcohol
##            0.117652642        1.117547112

```

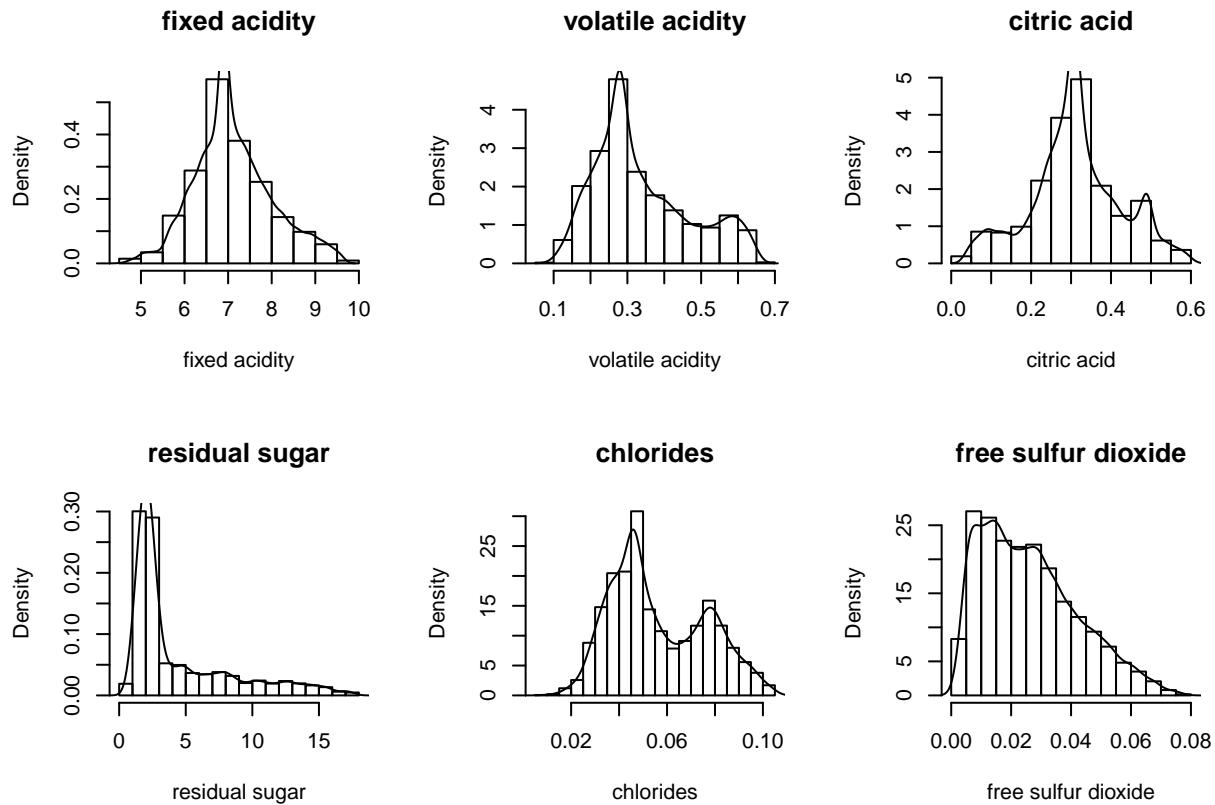
Visualizamos las variables (en histograma / gráfico de barras):

```

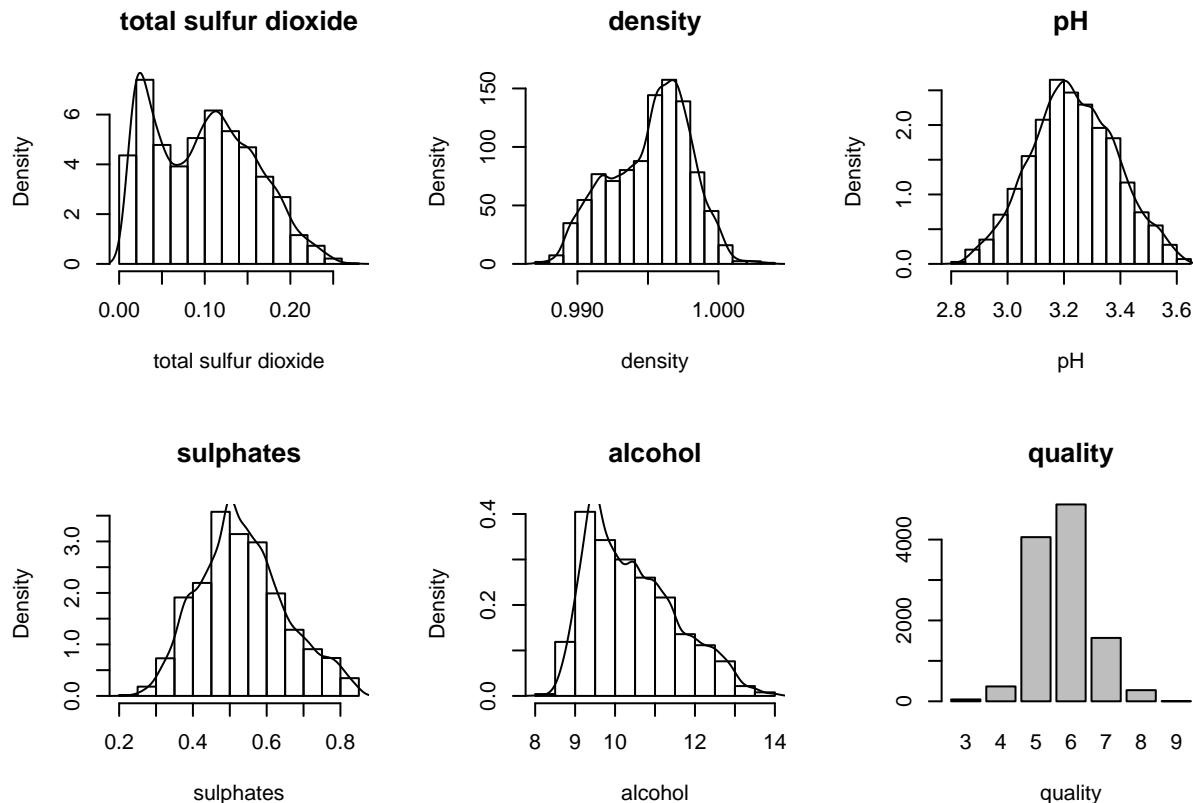
col <- c("fixed acidity", "volatile acidity", "citric acid", "residual sugar", "chlorides", "free sulfur dioxide")
par(mfrow=c(2,3))

for (name in col) {
  hist(vinos[,name], prob=TRUE, xlab=name, main = name)
  lines(density(na.omit(vinos[,name])))
}

```



```
barplot(table(vinos$quality), xlab="quality", main = "quality")
```



Se observa que la distribución de la mayoría de las variables es bastante sesgada y parece visualmente a la F-distribution. Suponemos que eso es debido a la naturaleza de indicadores físico-químicos. Además, la variable residual sugar tiene una cola por la derecha siquiera después de tratamiento de los outliers, y su desviación estándar también es relativamente alta.

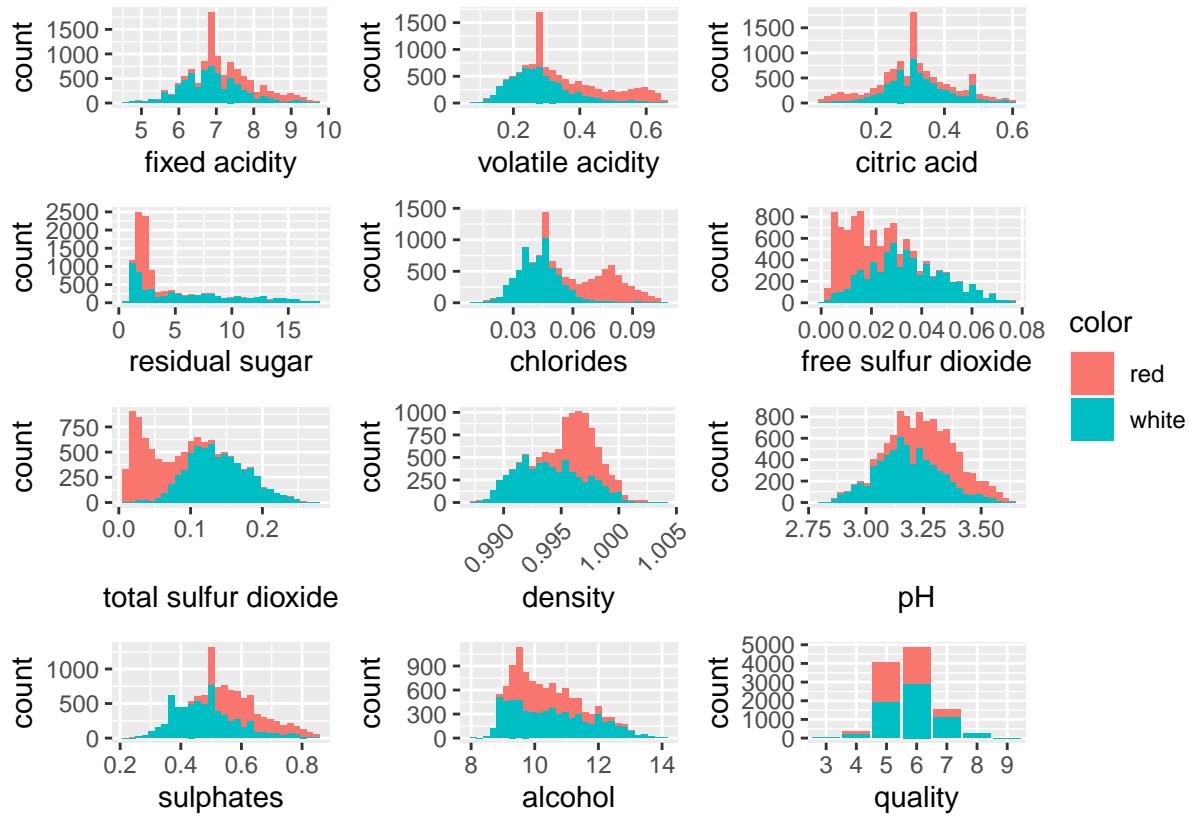
Hemos visto las distribuciones univariantes del dataset, ahora nos interesa ver las relaciones entre variables, correlaciones, cuáles de las características son más representativas en cuanto a la calidad / color del vino, qué explica la variebilidad natural del dataset y/u otras inferencias que podemos obtener de visualizaciones.

Veamos primero la distribución y las frecuencias relativas de variables por color:

```
p1 <- ggplot(data=vinos,aes(x=`fixed acidity`,fill=color))+geom_histogram()
p2 <- ggplot(data=vinos,aes(x=`volatile acidity`,fill=color))+geom_histogram()
p3 <- ggplot(data=vinos,aes(x=`citric acid`,fill=color))+geom_histogram()
p4 <- ggplot(data=vinos,aes(x=`residual sugar`,fill=color))+geom_histogram()
p5 <- ggplot(data=vinos,aes(x=`chlorides`,fill=color))+geom_histogram()
p6 <- ggplot(data=vinos,aes(x=`free sulfur dioxide`,fill=color))+geom_histogram()
p7 <- ggplot(data=vinos,aes(x=`total sulfur dioxide`,fill=color))+geom_histogram()
p8 <- ggplot(data=vinos,aes(x=`density`,fill=color))+geom_histogram() +theme(axis.text.x = element_text())
p9 <- ggplot(data=vinos,aes(x=`pH`,fill=color))+geom_histogram()
p10 <- ggplot(data=vinos,aes(x=`sulphates`,fill=color))+geom_histogram()
p11 <- ggplot(data=vinos,aes(x=`alcohol`,fill=color))+geom_histogram()

p12 <- ggplot(data=vinos,aes(x=`quality`,fill=color))+geom_bar()

p1 + p2 + p3 + p4 + p5 + p6 + p7 + p8 + p9 + p10 + p11 + p12 + plot_layout(ncol = 3) + plot_layout(guides = "none")
```



Se ve que en la mayoría de las variables la distribución de vinos tintos y vinos blancos es relativamente diferente. En 'free sulfur dioxide' y 'total sulfur dioxide' la distribución de tintos tiende bastante más a la izquierda (valores menores) y en otras características hay sesgo a la derecha (valores más altos) - densidad, sulfatos, acidez volátil/fija, cloruros. Y algunas no tienen un sesgo muy pronunciado (alcohol, residual sugar). Muchas características bien deben de depender del color del vino.

Analisis inferencial

Como bien tenemos dos conjuntos distintos por el color de vino, podemos comparar las cualidades de los dos, por ejemplo si existe una relación entre el color de vino y su calidad en el dataset (o bien si son independientes) o por otro lado si hay diferencia en alcohol para vinos tintos/blancos y vinos de calidad baja/alta -con lo que podemos obtener conclusiones tangibles para explicar la naturaleza de vinos.

Selección de grupos

Hacemos los subconjuntos de datos por el color y por la calidad:

```
red <- vinos[vinos$color=='red',]
white <- vinos[vinos$color=='white',]

high <- vinos[vinos$`quality class`=='high',]
low <- vinos[vinos$`quality class`=='low',]
```

Test sobre independencia de color y calidad alta/baja del vino

Como calidad es la variable discreta, realizamos el test chi-squared para poder hacer inferencia sobre la calidad de los vinos, con la hipótesis nula que la calidad y color son estadísticamente independientes y la hipótesis alternativa que existe una relación entre el color y la calidad.

```

# Tabla de contingencia
tc<-table( vinos$`quality class`, vinos$color )
tc

##
##      red white
##  low  2321  2148
##  high 2476  4248

# Test
chisq.test(tc, correct=FALSE)

##
##  Pearson's Chi-squared test
##
##  data: tc
##  X-squared = 250.36, df = 1, p-value < 2.2e-16

```

Segun el p-value podemos concluir que sí hay relacion entre el color de vino y la calidad en los datos observados, pudiendo concluir que los vinos blancos generalmente tienen mejor calidad.

Test sobre la proporción de vinos de alta calidad según el color

Para comprobar los resultados anteriores, planteamos el contraste para determinar si la proporción de vinos de alta calidad es igual para vinos tintos y blancos. Puesto que tenemos una muestra lo suficientemente grande, podemos también realizar un test parametrico sobre la proporción.

```

pR <- dim(vinos[which(vinos$color=='red' & vinos$`quality class`=='high'),])[1]/dim(vinos[vinos$color==
pW <- dim(vinos[which(vinos$color=='white' & vinos$`quality class`=='high'),])[1]/dim(vinos[vinos$color==
nR <- dim(vinos[vinos$color=='red',])[1]
nW <- dim(vinos[vinos$color=='white',])[1]

success<-c(pR*nR, pW*nW) # vector de casos de "exito"
nn<-c(nR,nW) # vector de tamaño de muestras
prop.test(success, nn, alternative="two.sided", conf.level=0.95, correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
##  data: success out of nn
##  X-squared = 250.36, df = 1, p-value < 2.2e-16
##  alternative hypothesis: two.sided
##  95 percent confidence interval:
##  -0.1662837 -0.1297347
##  sample estimates:
##      prop 1      prop 2
##  0.5161559  0.6641651

```

Por el p-value no podemos aceptar la hipotesis nula de igualdad de proporciones, por ello concluimos que las proporciones de vinos de alta calidad son distintos para vinos de distinto color, siendo igualmente el color blanco el que tiende a tener calidad más alta.

Test sobre alcohol en vinos blancos/tintos y de calidad alta/baja

Test de hipotesis de contraste sobre la igualdad estadística de la media de alcohol en dos muestras respectivas es un test parametrico si podemos comprobar la suposición de normalidad de variable y segun la igualdad de

varianzas elegiremos el test a realizar.

Vinos blancos / tintos

Test de normalidad

Test de normalidad Shapiro-Wilk (que es mas robusto que el test de Kolmogorov-Smirnov)

```
shapiro.test(red$alcohol)

##
##  Shapiro-Wilk normality test
##
## data: red$alcohol
## W = 0.9333, p-value < 2.2e-16
shapiro.test(sample(white$alcohol, 5000))
```

```
##
##  Shapiro-Wilk normality test
##
## data: sample(white$alcohol, 5000)
## W = 0.95524, p-value < 2.2e-16
```

No podemos aceptar la normalidad, por ello hacemos la transformación de Box-Cox:

```
red$alcohol <- BoxCox(red$alcohol, lambda = BoxCoxLambda(red$alcohol))
white$alcohol <- BoxCox(white$alcohol, lambda = BoxCoxLambda(white$alcohol))
```

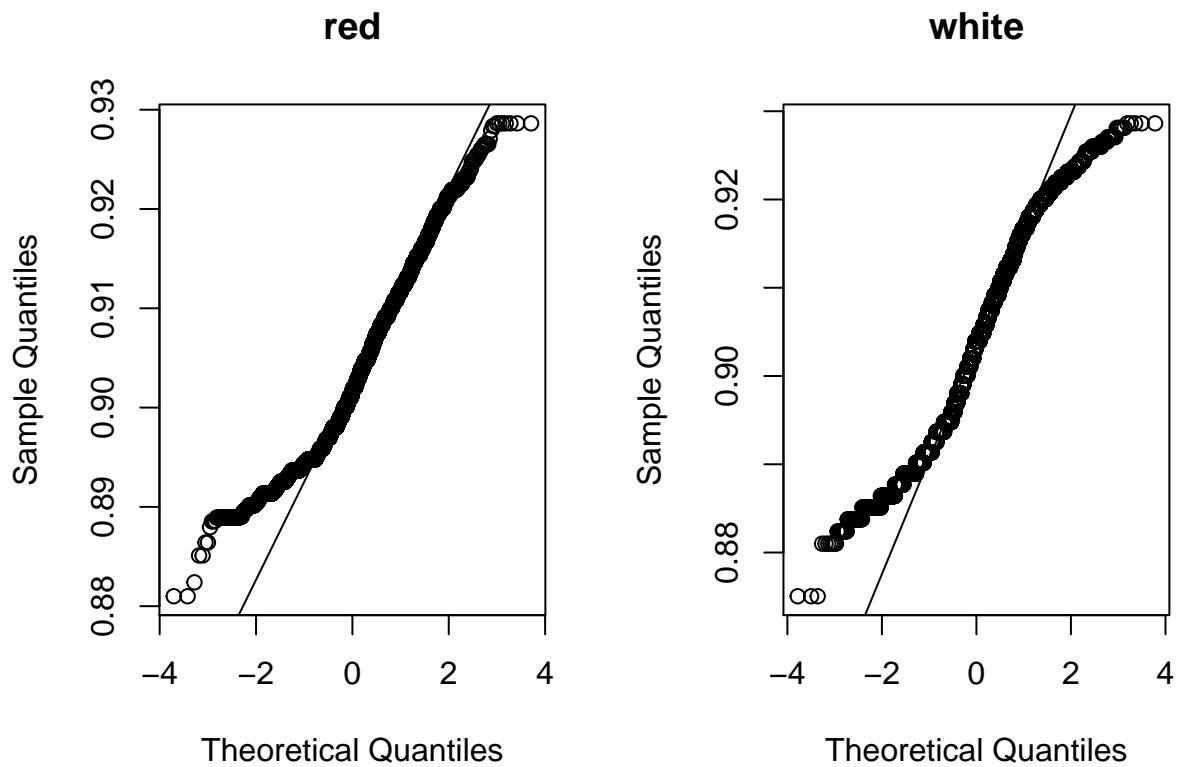
```
shapiro.test(red$alcohol)

##
##  Shapiro-Wilk normality test
##
## data: red$alcohol
## W = 0.96334, p-value < 2.2e-16
shapiro.test(sample(white$alcohol, 5000))

##
##  Shapiro-Wilk normality test
##
## data: sample(white$alcohol, 5000)
## W = 0.96799, p-value < 2.2e-16
```

A pesar de realizacion de transformacion de Box-Cox, rechazamos la hipotesis nula de normalidad de distribucion. Veamos las gráficas Q-Q:

```
par(mfrow=c(1, 2))
qqnorm(red$alcohol, main='red')
qqline(red$alcohol)
qqnorm(white$alcohol, main='white')
qqline(white$alcohol)
```



Aunque rechazamos la hipótesis nula de normalidad de datos, con el tamaño de la muestra por el teorema del límite central podemos asumir la normalidad de distribución de alcohol. Con ello realizamos el test de igualdad de varianzas de las muestras:

Test de homocedasticidad

```
var.test(red$alcohol, white$alcohol, alternative = "two.sided")
##
## F test to compare two variances
##
## data: red$alcohol and white$alcohol
## F = 0.58787, num df = 4796, denom df = 6395, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5576098 0.6199034
## sample estimates:
## ratio of variances
## 0.5878725
```

De la misma manera, rechazamos la hipótesis nula de igualdad de varianzas por lo que el test estadístico que se realiza es de dos muestras con varianza poblacional desconocida pero distinta.

Contraste sobre la media de alcohol en vinos blancos/tintos

```
t.test(red$alcohol, white$alcohol, alternative='two.sided', var.equal=FALSE)
```

```
##
## Welch Two Sample t-test
##
## data: red$alcohol and white$alcohol
```

```

## t = -5.0598, df = 11186, p-value = 4.264e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.001274738 -0.000562852
## sample estimates:
## mean of x mean of y
## 0.9027990 0.9037178

```

Por los resultados del test, con la confianza de 95% aceptamos la hipótesis de que el nivel promedio del alcohol es estadísticamente distinto en vinos blancos y tintos, con la diferencia real entre las dos medias en intervalo [-0.0015334921,-0.0008158837].

Vinos calidad alta / baja

Seguimos los mismos pasos para los subconjuntos según la calidad:

Test de normalidad

Test de normalidad Shapiro-Wilk:

```
shapiro.test(sample(high$alcohol, 5000))
```

```

##
##  Shapiro-Wilk normality test
##
## data: sample(high$alcohol, 5000)
## W = 0.97162, p-value < 2.2e-16
shapiro.test(low$alcohol)

##
##  Shapiro-Wilk normality test
##
## data: low$alcohol
## W = 0.93553, p-value < 2.2e-16

```

Transformación de Box-Cox:

```
high$alcohol <- BoxCox(high$alcohol, lambda = BoxCoxLambda(high$alcohol))
```

```
low$alcohol <- BoxCox(low$alcohol, lambda = BoxCoxLambda(low$alcohol))
```

```
shapiro.test(sample(high$alcohol, 5000))
```

```

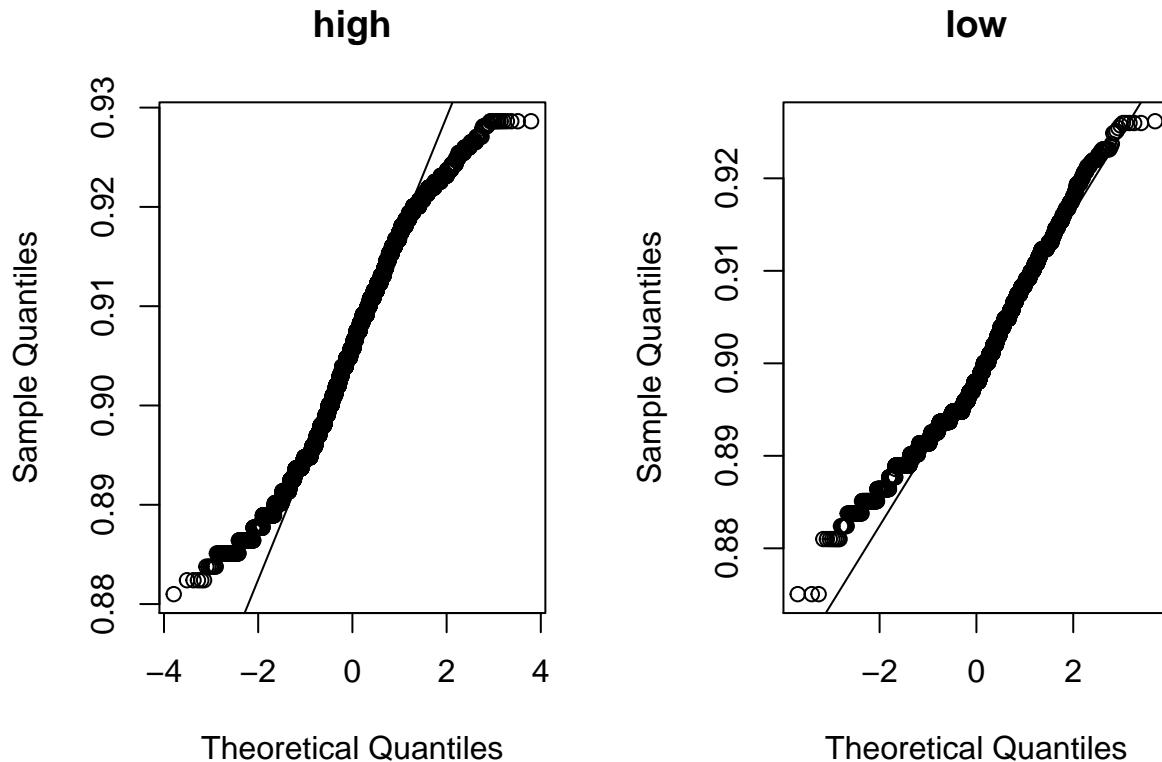
##
##  Shapiro-Wilk normality test
##
## data: sample(high$alcohol, 5000)
## W = 0.98031, p-value < 2.2e-16
shapiro.test(low$alcohol)

##
##  Shapiro-Wilk normality test
##
## data: low$alcohol
## W = 0.97072, p-value < 2.2e-16

```

A pesar de realizacion de transformación de Box-Cox, rechazamos la hipótesis nula de normalidad de distribución. Veamos las gráficas Q-Q:

```
par(mfrow=c(1,2))
qqnorm(high$alcohol, main='high')
qqline(high$alcohol)
qqnorm(low$alcohol, main='low')
qqline(low$alcohol)
```



Resultado es parecido que en los subconjuntos anteriores, por ello de la misma manera aplicamos el teorema del límite central para asumir la normalidad.

Test de homocedasticidad

```
var.test(high$alcohol, low$alcohol, alternative = "two.sided")
```

```
##
## F test to compare two variances
##
## data: high$alcohol and low$alcohol
## F = 1.4774, num df = 6723, denom df = 4468, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.400268 1.558427
## sample estimates:
## ratio of variances
## 1.477448
```

De la misma manera, rechazamos la hipótesis nula de igualdad de varianzas por lo que el test estadístico que se realiza es de dos muestras con varianza poblacional desconocida pero distinta.

Contraste sobre la media de alcohol en vinos blancos/tintos

```
t.test(high$alcohol, low$alcohol, alternative='two.sided', var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data: high$alcohol and low$alcohol
## t = 35.963, df = 10696, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.005925997 0.006609237
## sample estimates:
## mean of x mean of y
## 0.9058265 0.8995589
```

Por los resultados del test, igualmente, con la confianza de 95% aceptamos la hipótesis de que el valor promedio del alcohol es estadísticamente distinto en vinos de calidad alta y baja, con los vinos de alta calidad teniendo el valor promedio de alcohol mas alto.

Conclusiones

Por ello, podemos concluir que la variable alcohol puede ser importante tanto para la calidad como para el color del vino, y que hay además una relación estadística entre la calidad y el color.

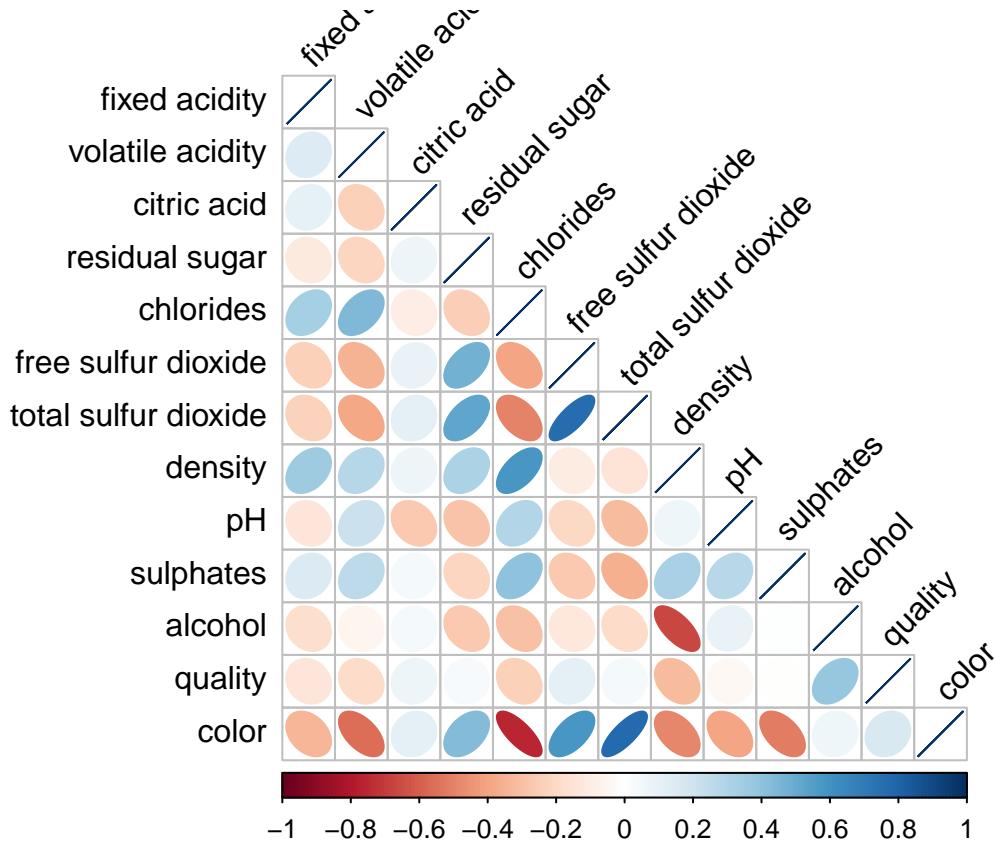
Estudio de correlación entre las variables

Podemos ver el correlograma visualizando las correlaciones entre las variables:

```
d <- vinos[1:11]
d$quality = as.numeric(vinos$quality)
d$color = as.numeric(vinos$color)

M<-cor(d)

corrplot(M, method="ellipse", type='lower', tl.col="black", tl.srt=45)
```



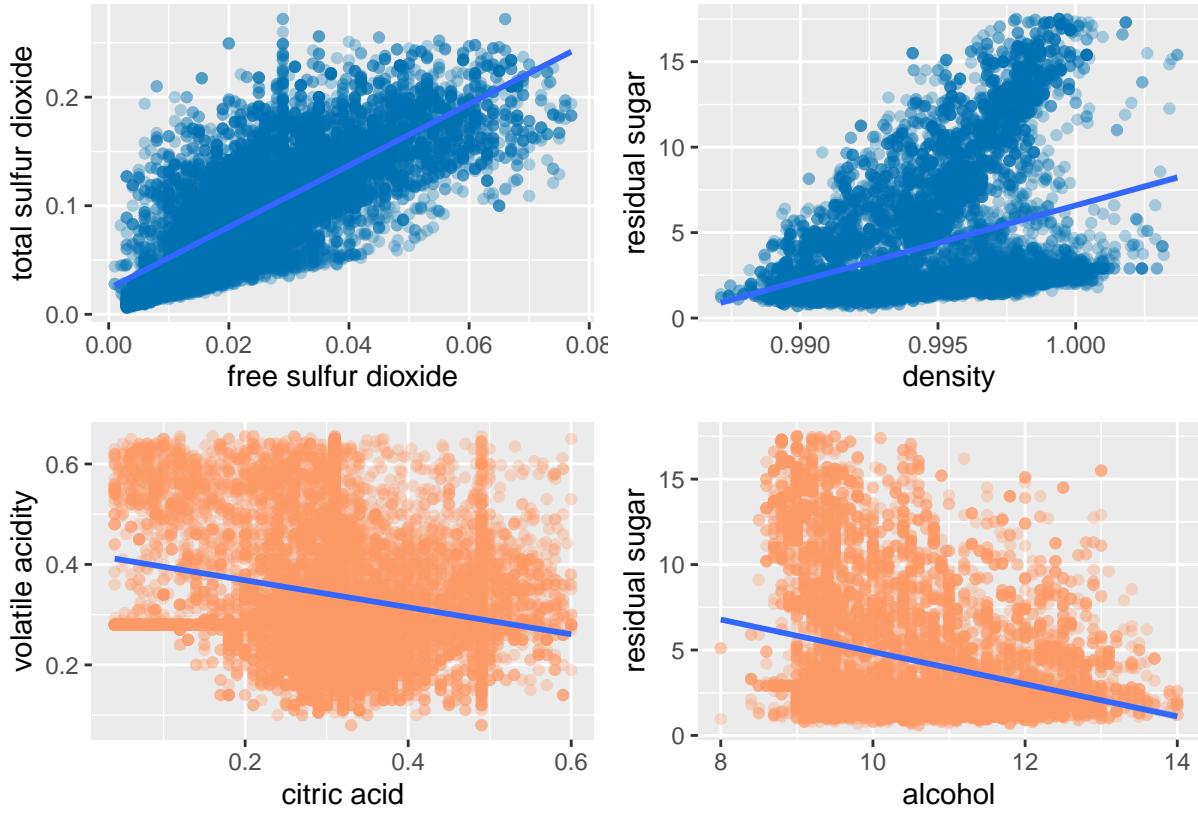
Se observa que, muchos atributos tienen correlación entre sí. Por ejemplo, fuerte correlación positiva presentan total sulfur dioxide y free sulfur dioxide, esperadamente. En general, density tiene correlación positiva con casi todas las características (sin contar color y calidad), bastante alta con “residual sugar”, y solo con “alcohol” está fuerte y negativamente correlacionada. Las correlaciones entre atributos químicos pueden ser explicados por adición de las sustancias con el fin de nivelar/acentuar las otras características .

Por otro lado, calidad no parece tener correlaciones muy importantes positivas o negativas con muchas de las características, no obstante las que más le puedan influir son alcohol (positivamente) y densidad (negativamente).

No obstante, el color está más correlacionado con los atributos físico-químicos, como con “total sulfur dioxide” (positivamente), “volatile acidity” (negativamente), etc, donde los signos de correlación representaran: negativo -> vino tinto, positivo -> vino blanco. Hay que tener en cuenta que la dimensionalidad de la parte de vinos tintos en el dataset es distinta a la de vinos blancos, aunque debe de ser suficiente para obtener las correlaciones correctas.

Algunas de correlaciones más significativas:

```
p1 <- ggplot(data = vinos,aes(x=`free sulfur dioxide`,y=`total sulfur dioxide`))+geom_jitter(color="#0072B2", alpha=0.3)
p2 <- ggplot(data = vinos,aes(x=`density`,y=`residual sugar`))+geom_jitter(color="#0072B2", alpha=0.3)
p3 <- ggplot(data = vinos,aes(x=`citric acid`,y=`volatile acidity`)) + geom_jitter(color="#FF9966", alpha=0.3)
p4 <- ggplot(data = vinos,aes(x=`alcohol`,y=`residual sugar`)) + geom_jitter(color="#FF9966", alpha=0.3)
p1 + p2 + p3 + p4
```

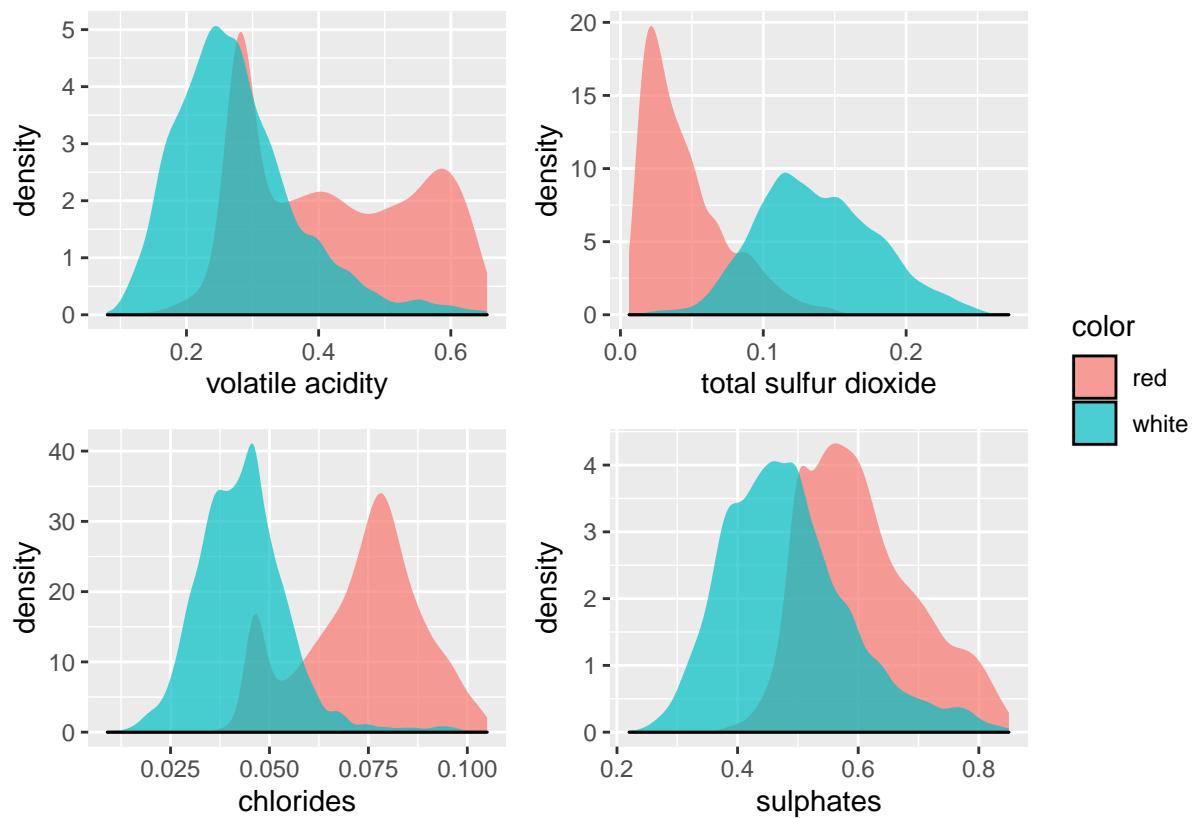


Color

Es interesante ver qué color tiene las correlaciones más fuertes con atributos físicos y químicos de vinos que calidad.

Veamos algunas de las características:

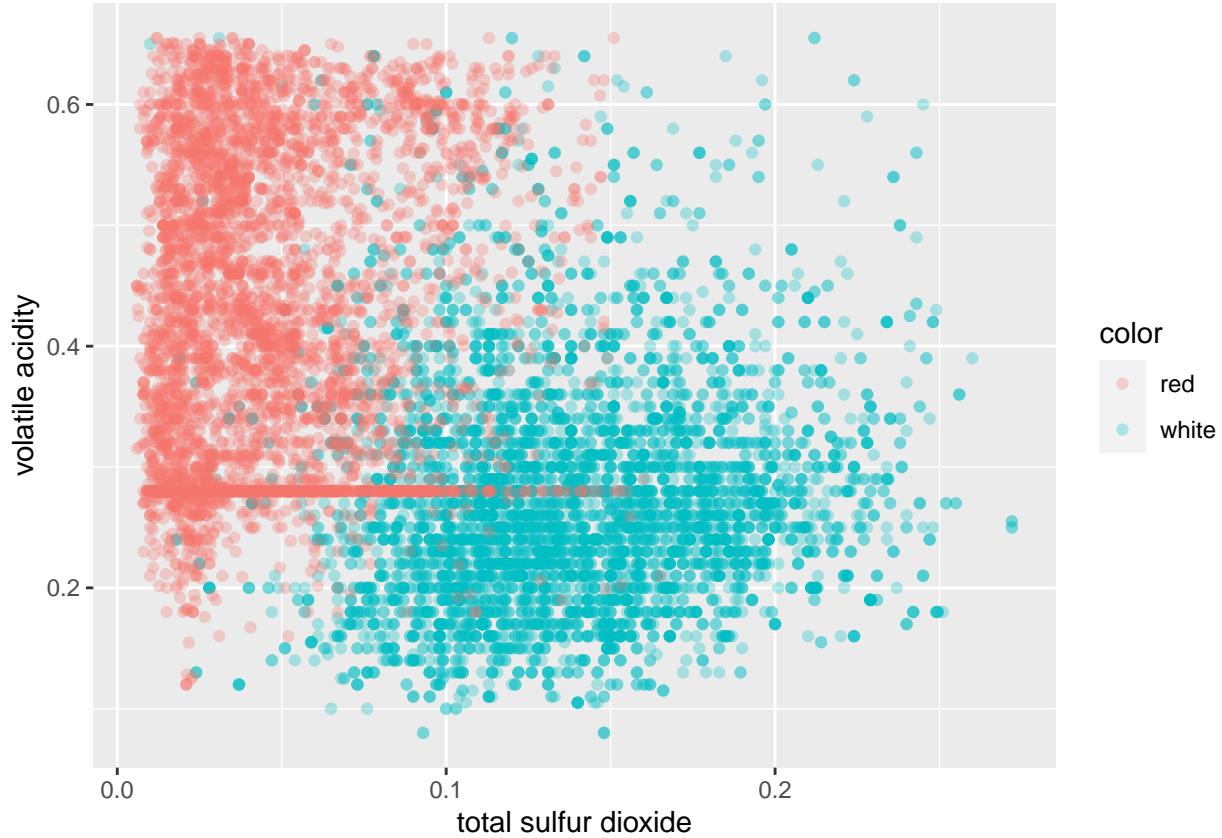
```
p1 <- ggplot(data = vinos,aes(x='volatile acidity',fill=color))+geom_density(alpha=0.7, outline.type = 'upper')
p2 <- ggplot(data = vinos,aes(x='total sulfur dioxide',fill=color))+geom_density(alpha=0.7, outline.type = 'upper')
p3 <- ggplot(data = vinos,aes(x='chlorides',fill=color))+geom_density(alpha=0.7, outline.type = 'lower')
p4 <- ggplot(data = vinos,aes(x='sulphates',fill=color))+geom_density(alpha=0.7, outline.type = 'lower')
p1 + p2 + p3 + p4 + plot_layout(guides = 'collect')
```



Las dos primeras características (“total sulphur dioxide” y “volatile acidity”) resultan en distribuciones de densidad muy distintas entre tipos de vinos que confirma la correlación.

Además, podemos ver las agrupaciones por color visualizando dos variables opuestamente correlacionadas:

```
ggplot(data = vinos, aes(x = `total sulfur dioxide`, y = `volatile acidity`, color = `color`)) + geom_p
```



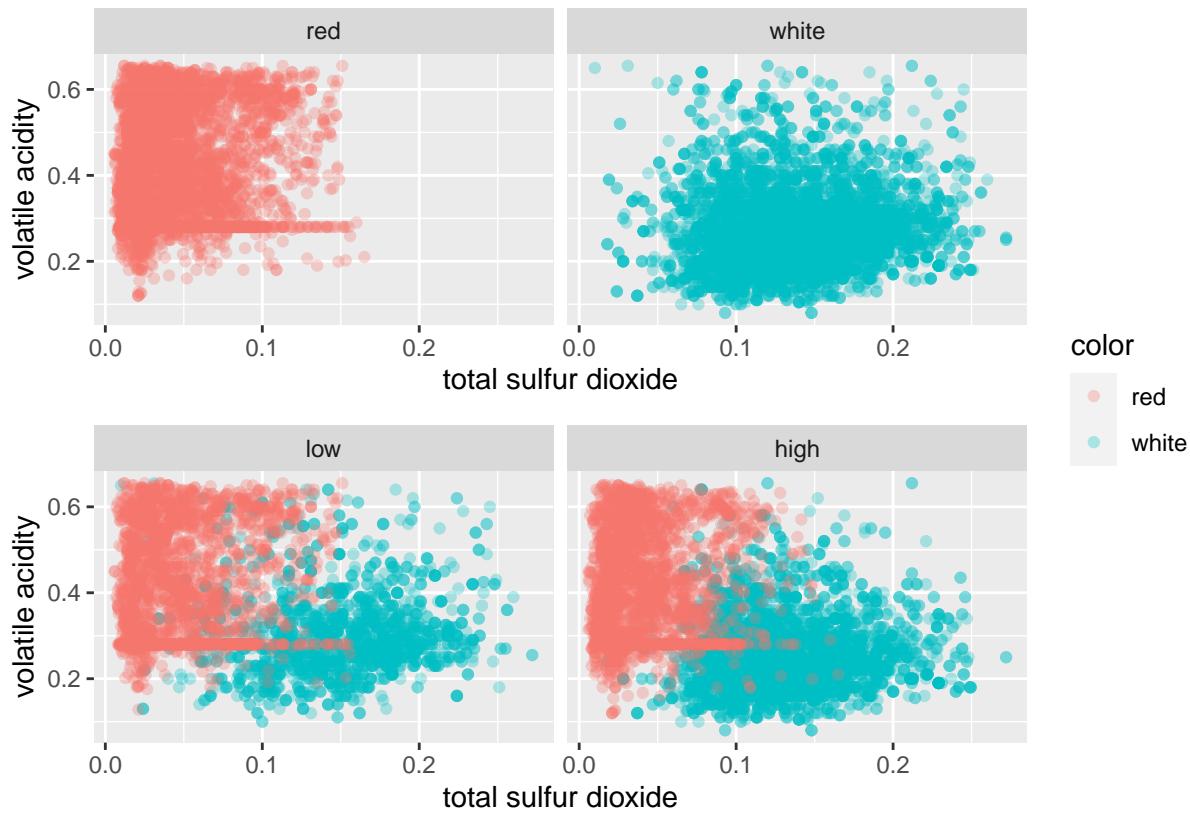
Hay cierta tendencia bastante definida visualmente de agrupaciones de vinos tintos/blancos aunque las observaciones están todavía entremezcladas.

Fijémonos en estas dos características:

```
p1 <- ggplot(data = vinos,
    aes(x = `total sulfur dioxide`, y = `volatile acidity`, color = `color`)) +
  geom_point(alpha = 0.3)+facet_wrap(~color)

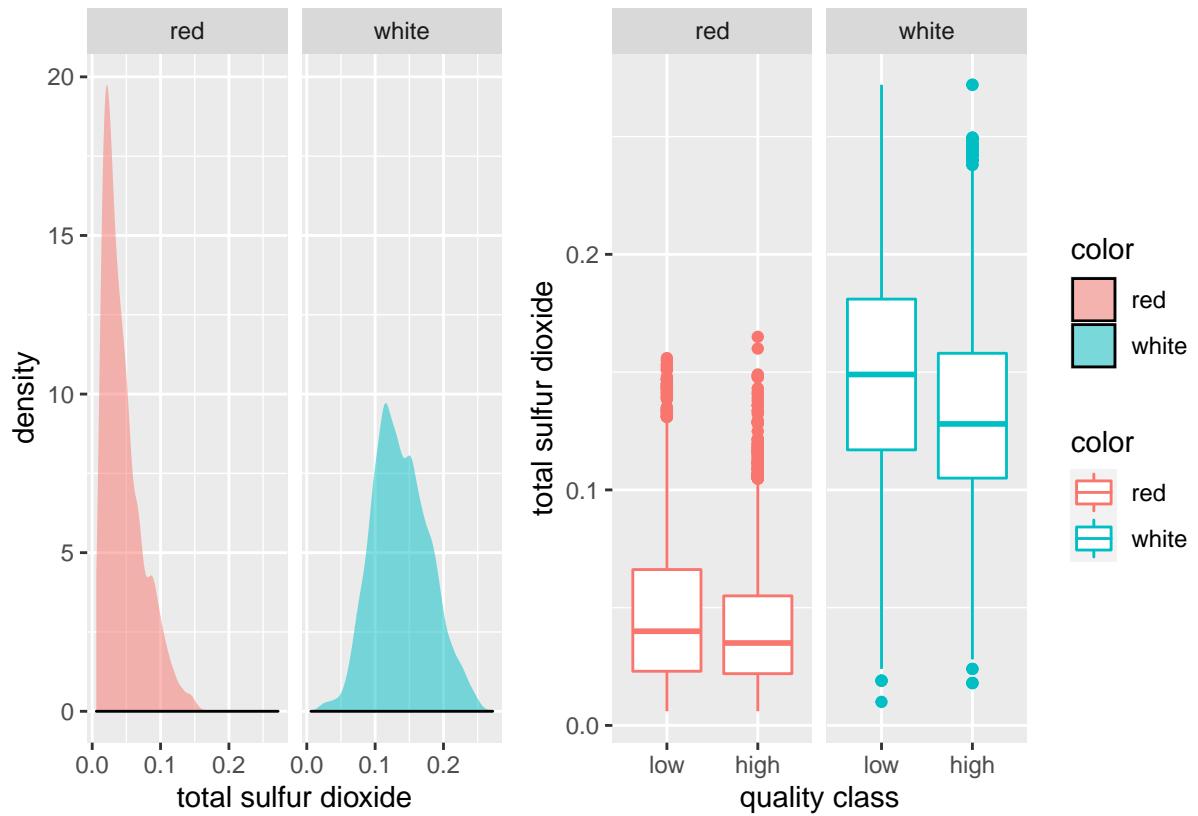
p2 <- ggplot(data = vinos,
    aes(x = `total sulfur dioxide`, y = `volatile acidity`, color = `color`)) +
  geom_point(alpha = 0.3)+facet_wrap(~quality class)

p1 + p2 + plot_layout(guides = 'collect') + plot_layout(ncol = 1)
```

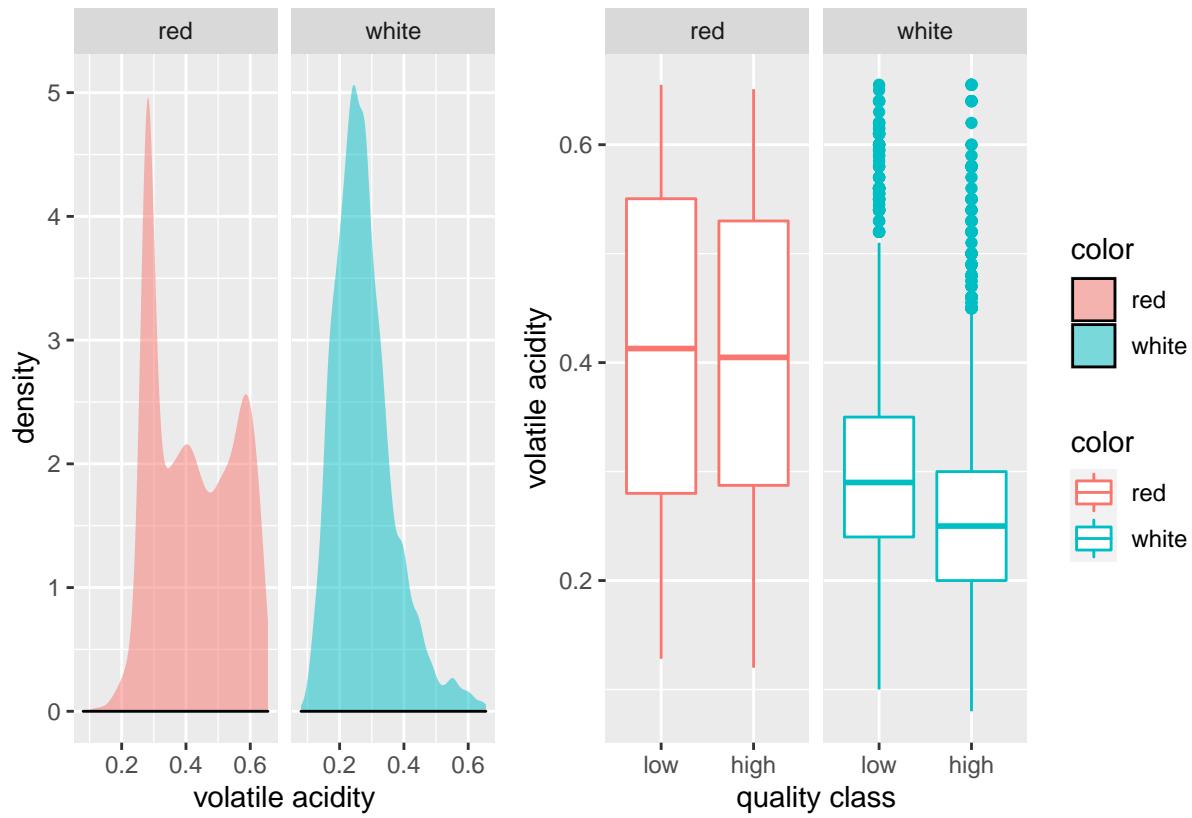


Las agrupaciones se mantienen en distintos grupos de calidad de vino.

```
# Total sulfur dioxide
p1 <- ggplot(data = vinos,aes(x=`total sulfur dioxide`,fill=`color`))+geom_density(alpha=0.5, outline.type="stroke")
p2 <- ggplot(data = vinos,aes(x=`quality class`, y=`total sulfur dioxide`, color = `color`))+geom_boxplot()
p1 + p2 + plot_layout(guides = 'collect')
```



```
# Volatile acidity
p1 <- ggplot(data = vinos,aes(x=`volatile acidity`,fill=`color`))+geom_density(alpha=0.5, outline.type = "line")
p2 <- ggplot(data = vinos,aes(x=`quality class` , y=`volatile acidity` , color = `color`))+geom_boxplot()
p1 + p2 + plot_layout(guides = 'collect')
```



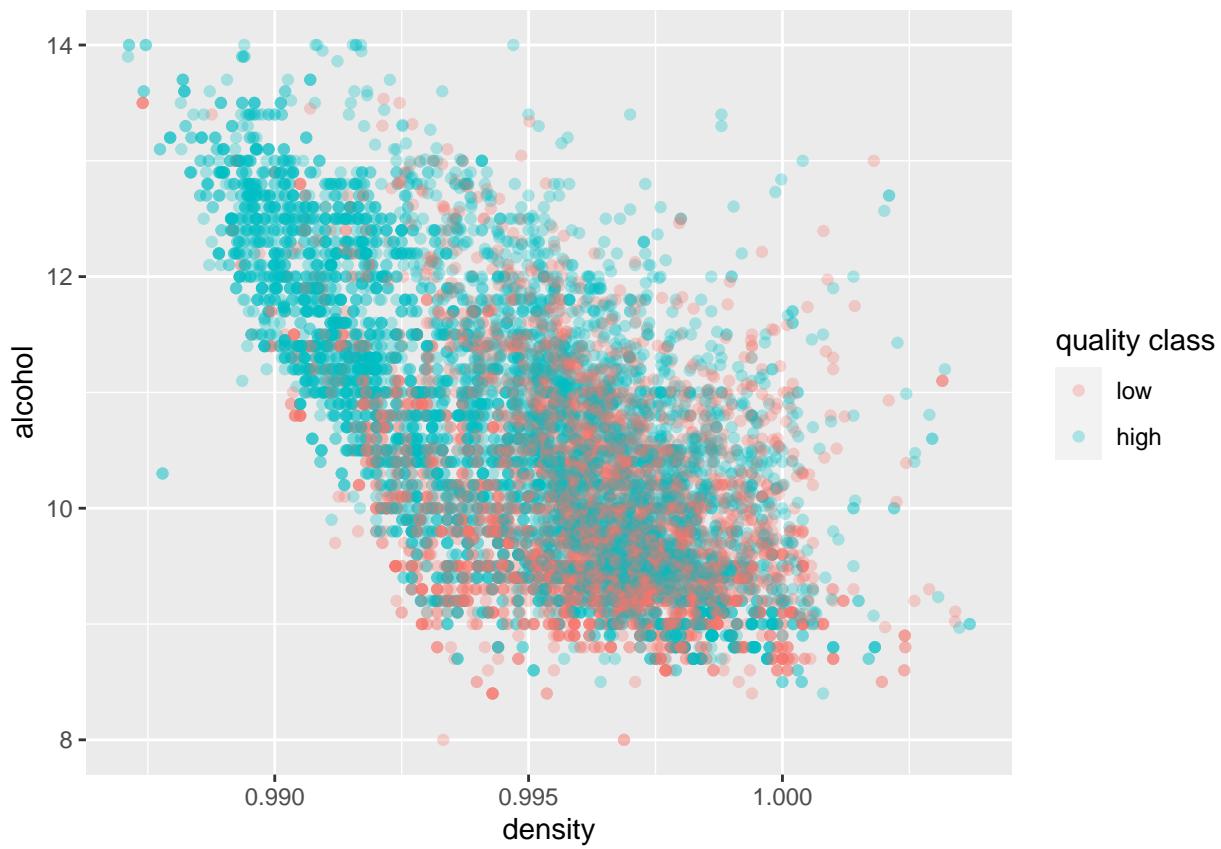
Se ven claramente las diferencias en distribuciones de ambas características por color de vino y correlaciones positiva y negativa. Ya que “volatile acidity” también presntaba correlación con “quality”, se observan diferencias por cada clase de calidad, pero principalmente en vinos tintos.

Calidad

La densidad y el alcohol pueden ser las características del vino que son las más responsables por la calidad del vino del dataset. También, hay correlaciones más bajas con volatile acidity, sin embargo, como hemos visto, principalmente se manifiestan en un tipo de vinos (vinos tintos).

Veamos las dos variables con más detalle para ver la relación entre ellas:

```
ggplot(data = vinos, aes(x = density, y = alcohol, color = `quality class`)) + geom_point(alpha = 0.3)
```



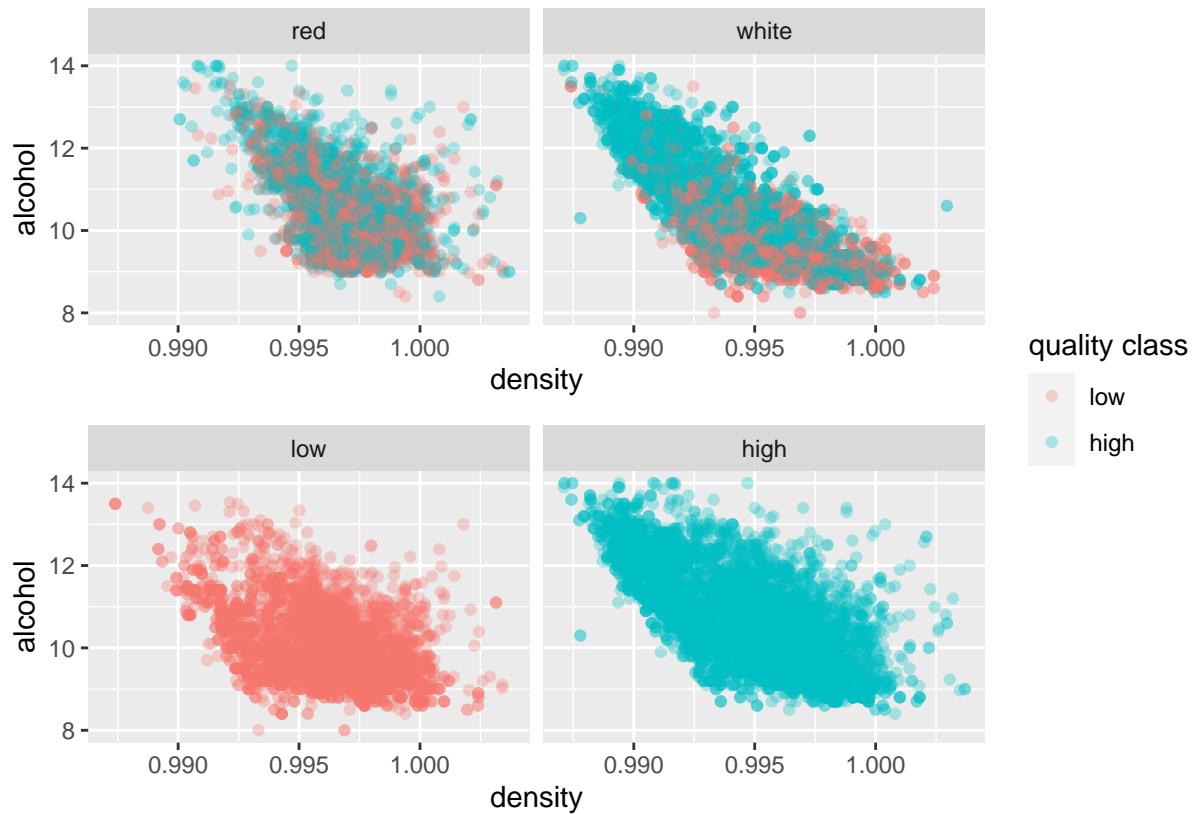
```

p1 <- ggplot(data = vinos,
  aes(x = density, y = alcohol, color = `quality class`)) +
  geom_point(alpha = 0.3)+facet_wrap(~color)

p2 <- ggplot(data = vinos,
  aes(x = density, y = alcohol, color = `quality class`)) +
  geom_point(alpha = 0.3)+facet_wrap(~`quality class`)

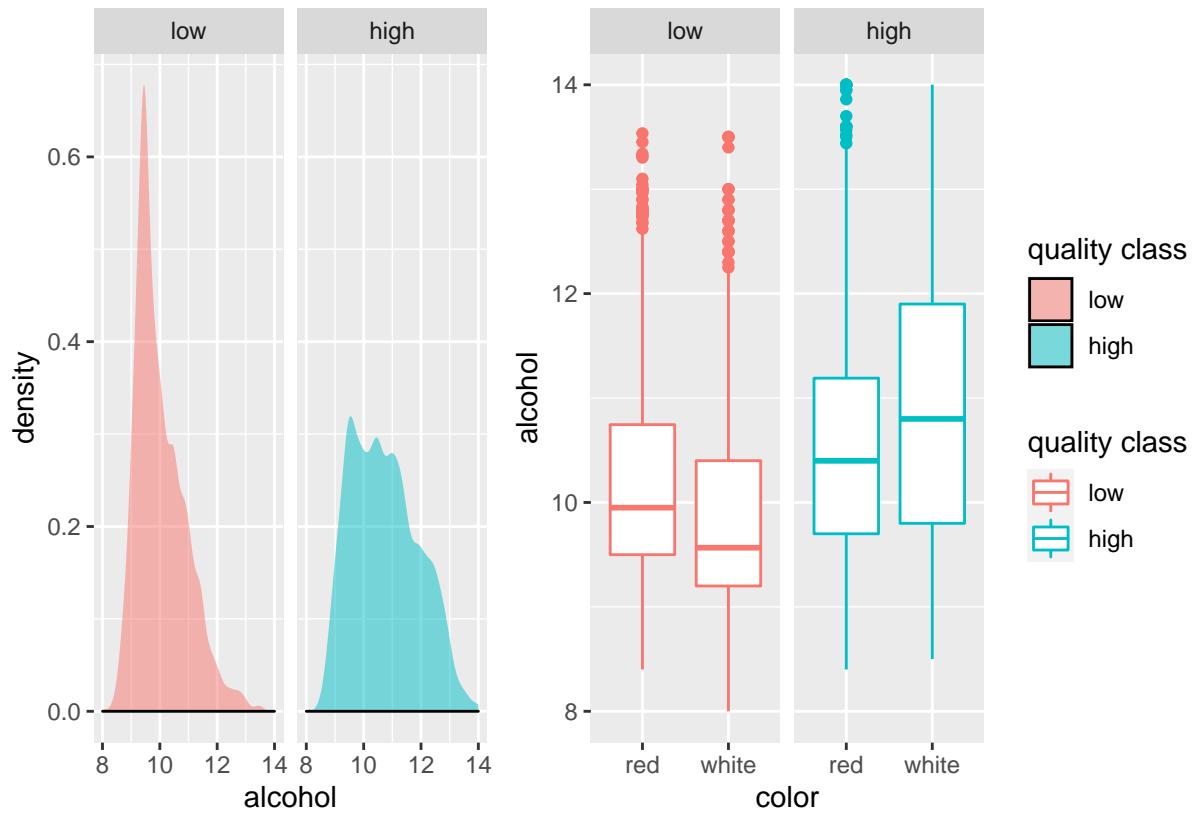
p1 + p2 + plot_layout(guides = 'collect') + plot_layout(ncol = 1)

```



Aunque las clases están entremezclados, también se ve la tendencia de agrupaciones.

```
# Alcohol
p1 <- ggplot(data = vinos,aes(x=alcohol,fill= `quality class`))+geom_density(alpha=0.5, outline.type = 'line')
p2 <- ggplot(data = vinos,aes(x=color, y=alcohol, color = `quality class`))+geom_boxplot() + facet_wrap(~color)
p1 + p2 + plot_layout(guides = 'collect')
```

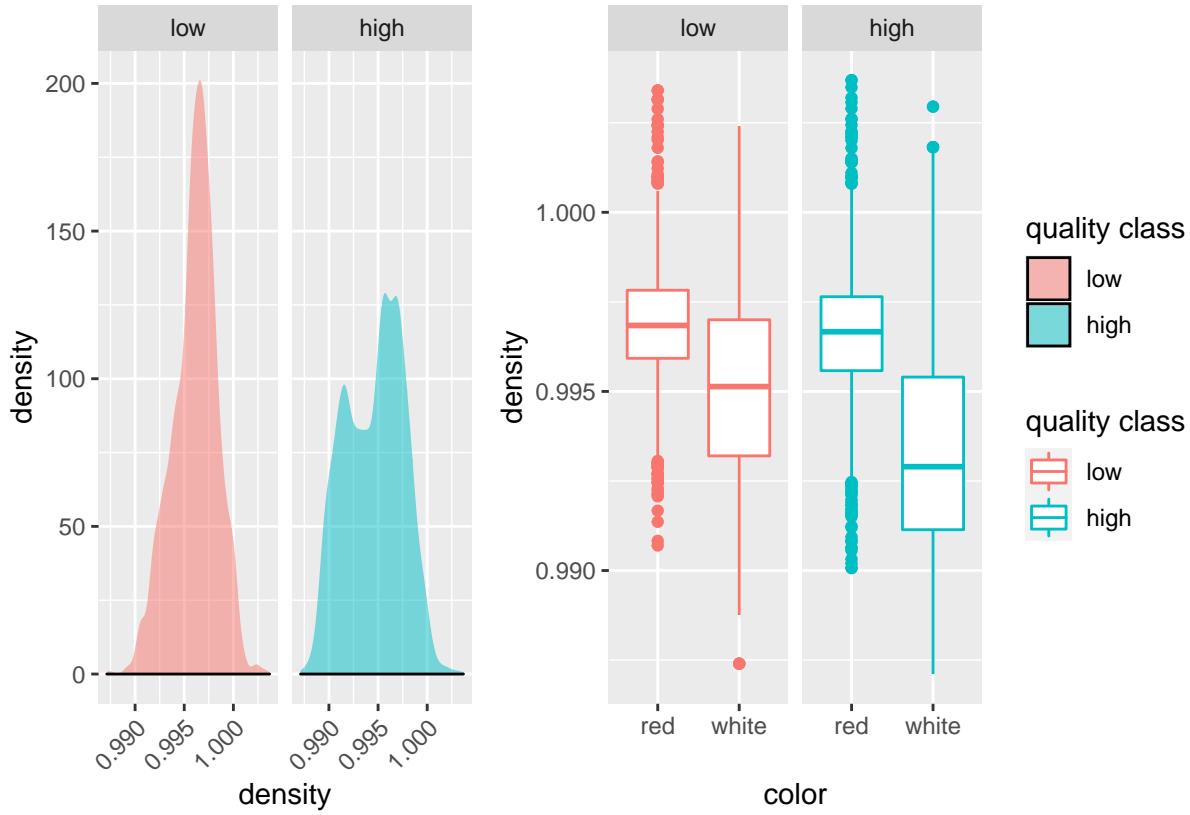


Density

```
p1 <- ggplot(data = vinos,aes(x=density,fill=`quality class`))+geom_density(alpha=0.5, outline.type = 'line')

p2 <- ggplot(data = vinos,aes(x=color, y=density, color = `quality class`))+geom_boxplot() + facet_wrap(~quality class)

p1 + p2 + plot_layout(guides = 'collect')
```

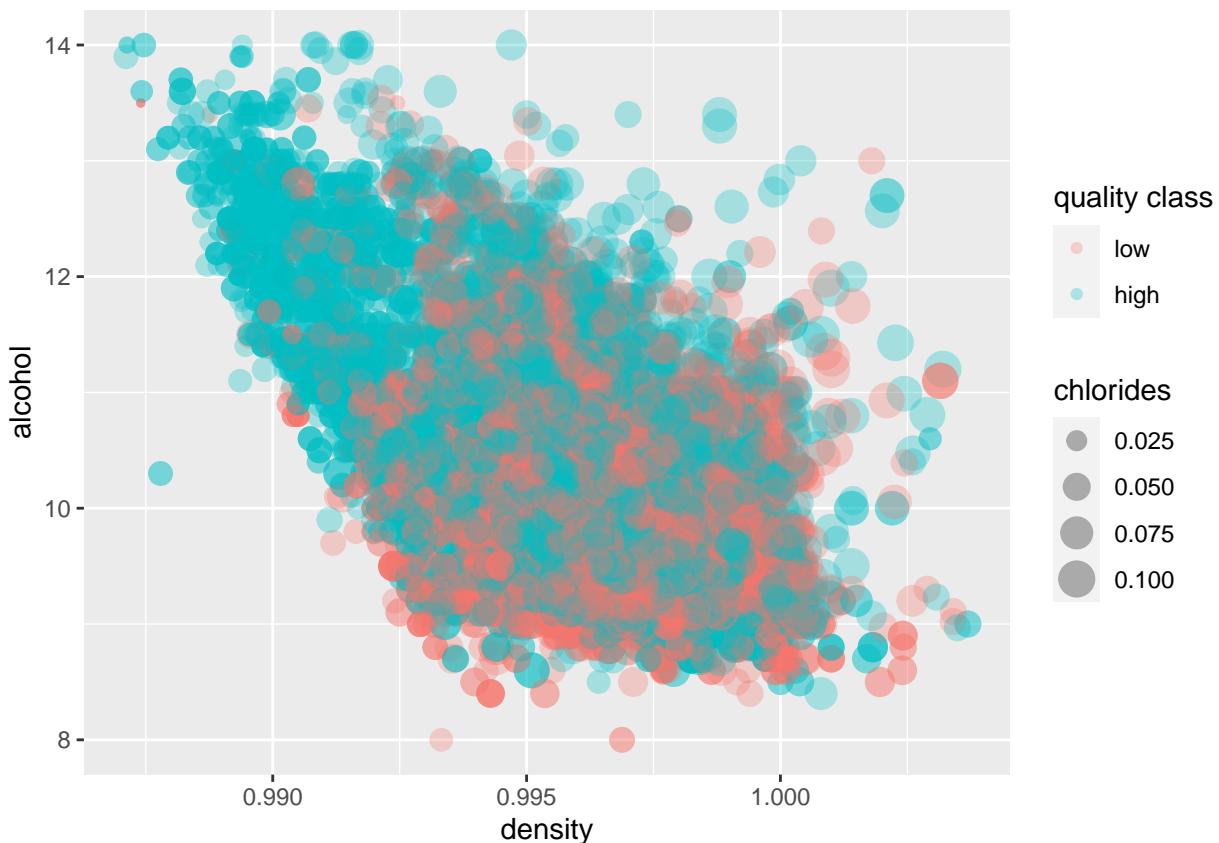


Observamos las correlaciones (positiva con alcohol y negativa con densidad) en ambos tipos/colores de vinos, aunque la densidad es la característica con una influencia más fuerte en vinos blancos. Se visualiza bien cómo el nivel de alcohol y densidad cambian en distintas particiones de la calidad del vino del dataset.

Hay tendencias de agrupamiento por valores de nivel de alcohol y densidad, pero al final la calidad debe de depender de un conjunto más grande de características físico-químicas ya que dos variables no explican totalmente la calidad, ni calidad explica la variebilidad del dataset.

Finalmente, para aproximarnos al conjunto de características, podemos visualizar la relación entre 4 variables: density, alcohol, chlorides y quality class. Chlorides es la siguiente característica que tenía correlación fuerte con quality después de volatile acidity que lo presentaba mayormente en vinos tintos.

```
ggplot(data = vinos, aes(x = density, y = alcohol, color = `quality class`)) + geom_point(aes(size = `chlorides`))
```



No hay relación lineal, pero se explican algunas observaciones “alejadas” (pertenecen a una partición de calidad pero se visualizan donde prevalece otra partición, pues tienen chlorides, por ejemplo, bajos, frente a la mayoría de su partición. Al final y a cabo muchos de los atributos influyen a la densidad).

Regresión multiple

Como hemos observado, no hay una relación lineal simple entre las variables, por ello se puede intentar explicar la calidad y el color linealmente con varios regresores a través de algoritmo de regresión logística, dado que tanto el color como la clase de calidad son variables dicotómicas.

Calidad

```
model1 <- glm(`quality class` ~ `fixed acidity` + `volatile acidity` + `citric acid` + `residual sugar` + `chlorides` + `free sulfur dioxide` + `total sulfur dioxide` + `density` + `pH` + `sulphates` + `alcohol`, family = binomial(link = logit), data = vinos)
summary(model1)

##
## Call:
## glm(formula = `quality class` ~ `fixed acidity` + `volatile acidity` +
##       `citric acid` + `residual sugar` + `chlorides` + `free sulfur dioxide` +
##       `total sulfur dioxide` + `density` + `pH` + `sulphates` + `alcohol`,
##       family = binomial(link = logit), data = vinos)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.4001   -1.0981    0.5659    0.9558    1.8747
##
## Coefficients:
```

```

##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 92.581975  16.404028   5.644 1.66e-08 ***
## `fixed acidity`            -0.015724   0.027284  -0.576  0.5644
## `volatile acidity`         -2.218213   0.190165 -11.665 < 2e-16 ***
## `citric acid`              -0.036810   0.202077  -0.182  0.8555
## `residual sugar`           0.075021   0.008299   9.040 < 2e-16 ***
## chlorides                  -3.917256   1.645374  -2.381  0.0173 *
## `free sulfur dioxide`      16.184696   2.121194   7.630 2.35e-14 ***
## `total sulfur dioxide`     -5.186445   0.653160  -7.941 2.01e-15 ***
## density                     -98.060623  16.488839 -5.947 2.73e-09 ***
## pH                          -0.032880   0.165127  -0.199  0.8422
## sulphates                  1.602855   0.224250   7.148 8.83e-13 ***
## alcohol                     0.535267   0.031166  17.175 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 15059 on 11192 degrees of freedom
## Residual deviance: 13290 on 11181 degrees of freedom
## AIC: 13314
##
## Number of Fisher Scoring iterations: 4

```

Según el modelo, las variables fixed acidity, citric acid, pH no tienen significancia estadística para el modelo ($p.value > 0.05$ por lo que aceptamos la hipótesis que significancia es igual a cero).

Para ver los odds ratio para cada unidad de las características:

```
exp(coefficients(model1))
```

```

##             (Intercept)          `fixed acidity`        `volatile acidity`
## 1.613767e+40          9.843986e-01          1.088034e-01
## `citric acid`          `residual sugar`       chlorides
## 9.638595e-01          1.077907e+00          1.989562e-02
## `free sulfur dioxide` `total sulfur dioxide`    density
## 1.068868e+07          5.591853e-03          2.587096e-43
## pH                      sulphates            alcohol
## 9.676543e-01          4.967194e+00          1.707904e+00

```

Puesto que las características con valores superiores a uno son alcohol y residual sugar, podemos considerar alcohol el más influyente a la probabilidad en el modelo obtenido.

La bondad de ajuste se obtiene a través del índice de Akaike AIC, que en este caso asciende a 13336, un valor muy alto. Podremos compararlo con el modelo posterior de regresión sobre el color.

Color

```

model2 <- glm(color ~ `fixed acidity` + `volatile acidity` + `citric acid` + `residual sugar` + `chlorides` + `free s
summary(model2)

##
## Call:
## glm(formula = color ~ `fixed acidity` + `volatile acidity` +
##     `citric acid` + `residual sugar` + `chlorides` + `free sulfur dioxide` +
##     `total sulfur dioxide` + density + pH + sulphates + alcohol,

```

```

##      family = binomial(link = logit), data = vinos)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.1295 -0.0127  0.0061  0.0386  3.2453
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.248e+03  6.808e+01 18.332 < 2e-16 ***
## `fixed acidity`     -4.476e-01  1.181e-01 -3.789 0.000151 ***
## `volatile acidity`  -6.827e+00  8.099e-01 -8.429 < 2e-16 ***
## `citric acid`        4.327e+00  1.038e+00  4.170 3.05e-05 ***
## `residual sugar`    4.875e-01  4.195e-02 11.620 < 2e-16 ***
## chlorides            -9.283e+01  6.931e+00 -13.393 < 2e-16 ***
## `free sulfur dioxide` -8.078e+01  1.128e+01 -7.161 8.03e-13 ***
## `total sulfur dioxide` 7.978e+01  4.314e+00 18.495 < 2e-16 ***
## density              -1.228e+03  6.738e+01 -18.232 < 2e-16 ***
## pH                   -9.664e-01  7.132e-01 -1.355 0.175394
## sulphates            -8.482e+00  1.013e+00 -8.374 < 2e-16 ***
## alcohol              -1.491e+00  1.459e-01 -10.218 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 15287.58 on 11192 degrees of freedom
## Residual deviance: 826.55 on 11181 degrees of freedom
## AIC: 850.55
##
## Number of Fisher Scoring iterations: 9

```

Se observan en este caso todas las variables significativas para el modelo, con los odds ratio respectivos de:

```
exp(coefficients(model2))
```

##	(Intercept)	`fixed acidity`	`volatile acidity`
##	Inf	6.391473e-01	1.084384e-03
##	`citric acid`	`residual sugar`	chlorides
##	7.569745e+01	1.628220e+00	4.836992e-41
##	`free sulfur dioxide`	`total sulfur dioxide`	density
##	8.238911e-36	4.459334e+34	0.000000e+00
##	pH	sulphates	alcohol
##	3.804458e-01	2.071528e-04	2.251544e-01

según los que las variables explicativas parecen ser total sulfur dioxide, alcohol.

El modelo es mas explicativo que el anterior, puesto que la bondad de ajuste es mejor con el AIC de 823.

Conclusiones

Podemos decir que no es facil encontrar un modelo lineal multiple que explique la varianza del dataset tanto para la probabilidad del color como de la calidad, entonces que el dataset no es linealmente separable. Otra manera poder explicar la varianza es realizar un estudio no supervisado y analizar las agrupaciones resultantes.

Modelo no supervisado (clustering)

Normalizamos el dataset (solo los features), puesto que el algoritmo de clustering se basa en las distancias entre las observaciones.

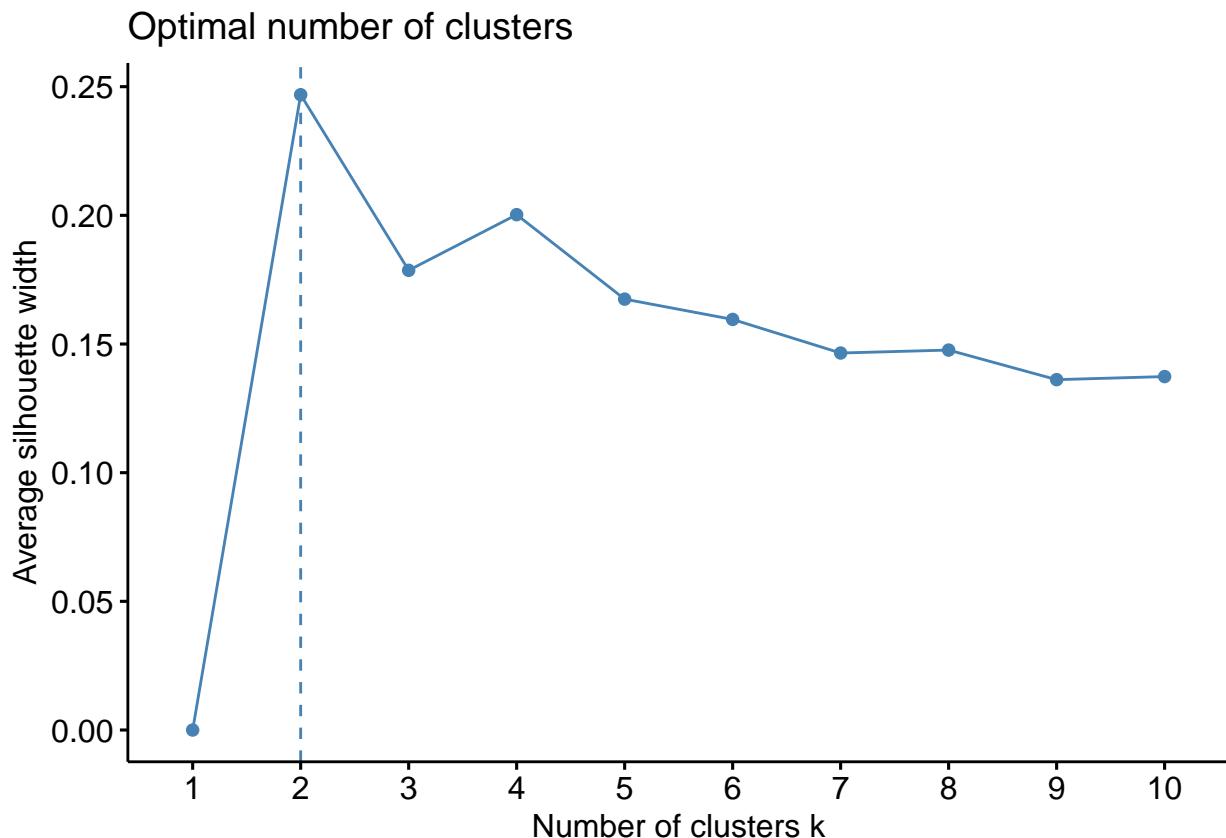
```
vinos_norm <- scale(vinos[,1:11])
```

Tal y como hemos dicho anteriormente, aunque hemos estimado el numero de intervalos de calidad de 2 (alta/baja), no tiene por que ser un número óptimo de clusters segun las características físico-químicas, por lo que estudiaremos las agrupaciones sin referencia a la marca de calidad y las aproximaremos al dataset con atributos caracrtísticos del vino.

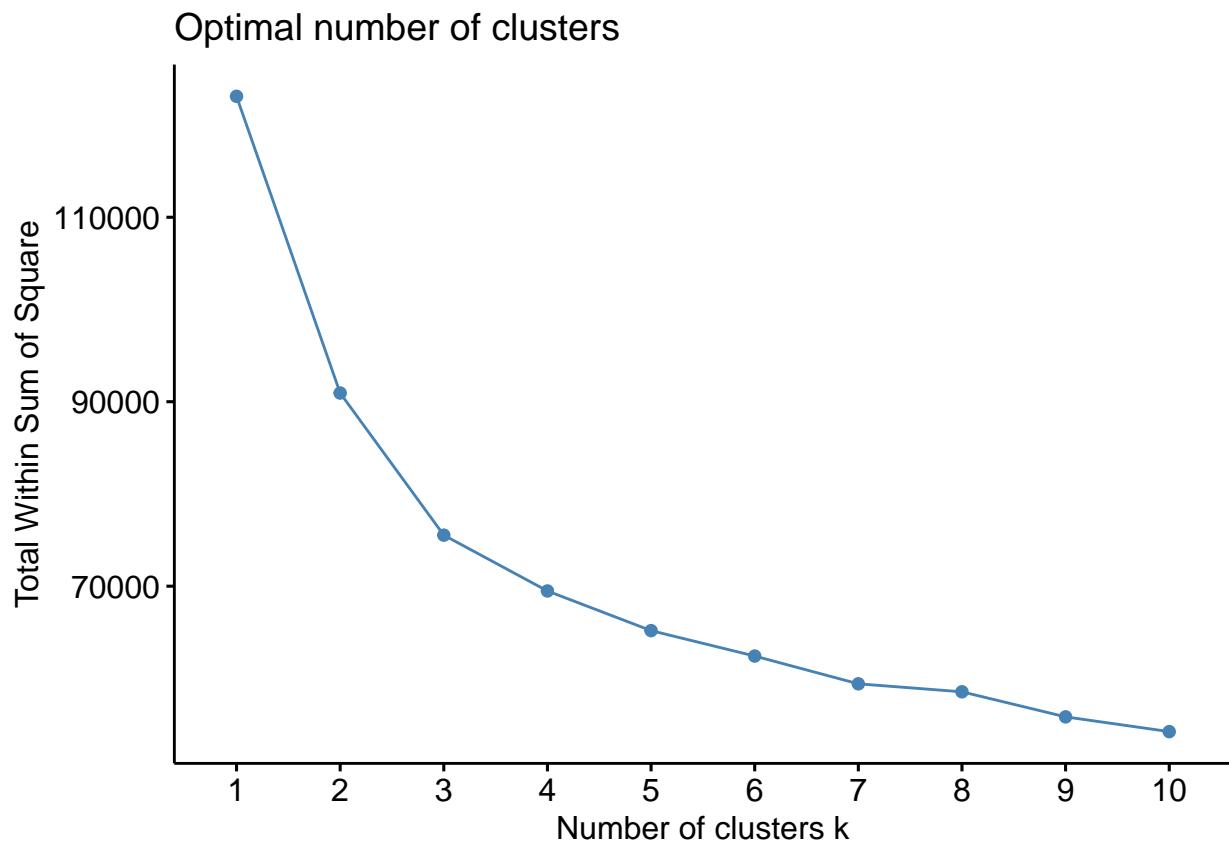
Número óptimo de clusters

Con la librería factoextra y la función fviz_nbclust podemos visualizar el número óptimo calculado con metodos por siluetas medias y wss (suma de cuadrados):

```
# Metodo por siluetas medias  
fviz_nbclust(vinos_norm, kmeans, method = "silhouette")
```



```
# Metodo wss  
fviz_nbclust(vinos_norm, kmeans, method = "wss")
```



El número óptimo de clusters parece ser 2, lo que está más claro por el metodo de siluetas pero no por el metodo de “codo”, como se ha preliminarmente definido para el número de particiones de calidad.

Modelo clusters

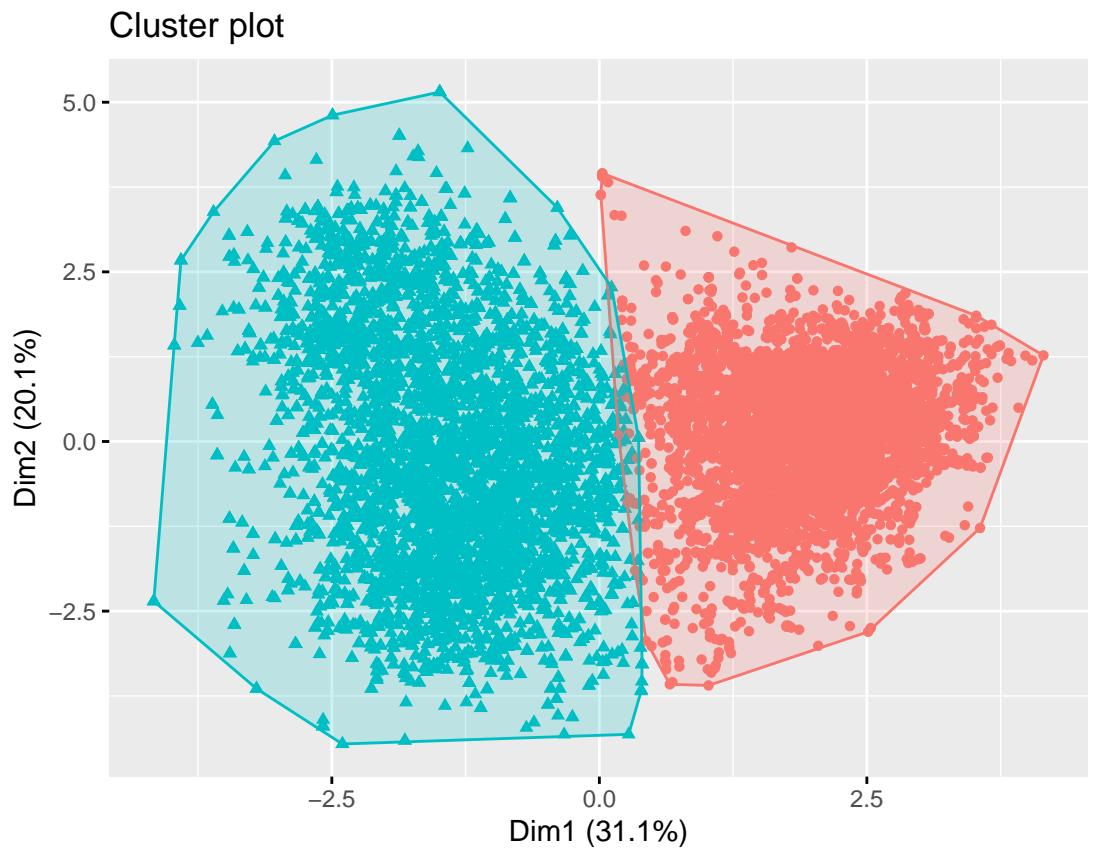
```
set.seed(1234)
fit <- kmeans(vinos_norm, 2)

# Tamaño de los clusters obtenidos
fit$size

## [1] 4790 6403
```

Visualizamos los clusters:

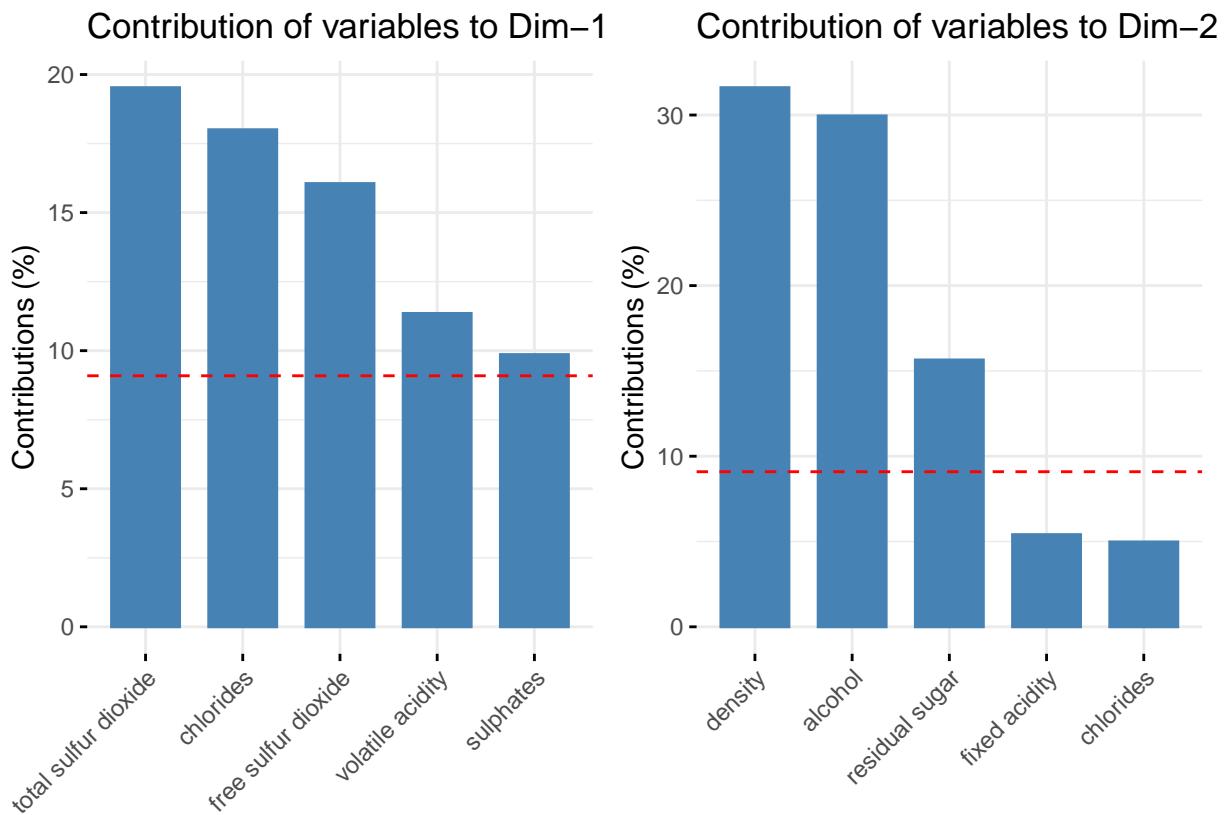
```
fviz_cluster(fit, geom = "point", data = vinos_norm)
```



Se observan que están bastante poco entremezclados, siendo el segundo cluster más poblado.

Se puede ver la feature importance de las dimensiones de la gráfica que serán las mismas que los 2 primeros componentes de PCA:

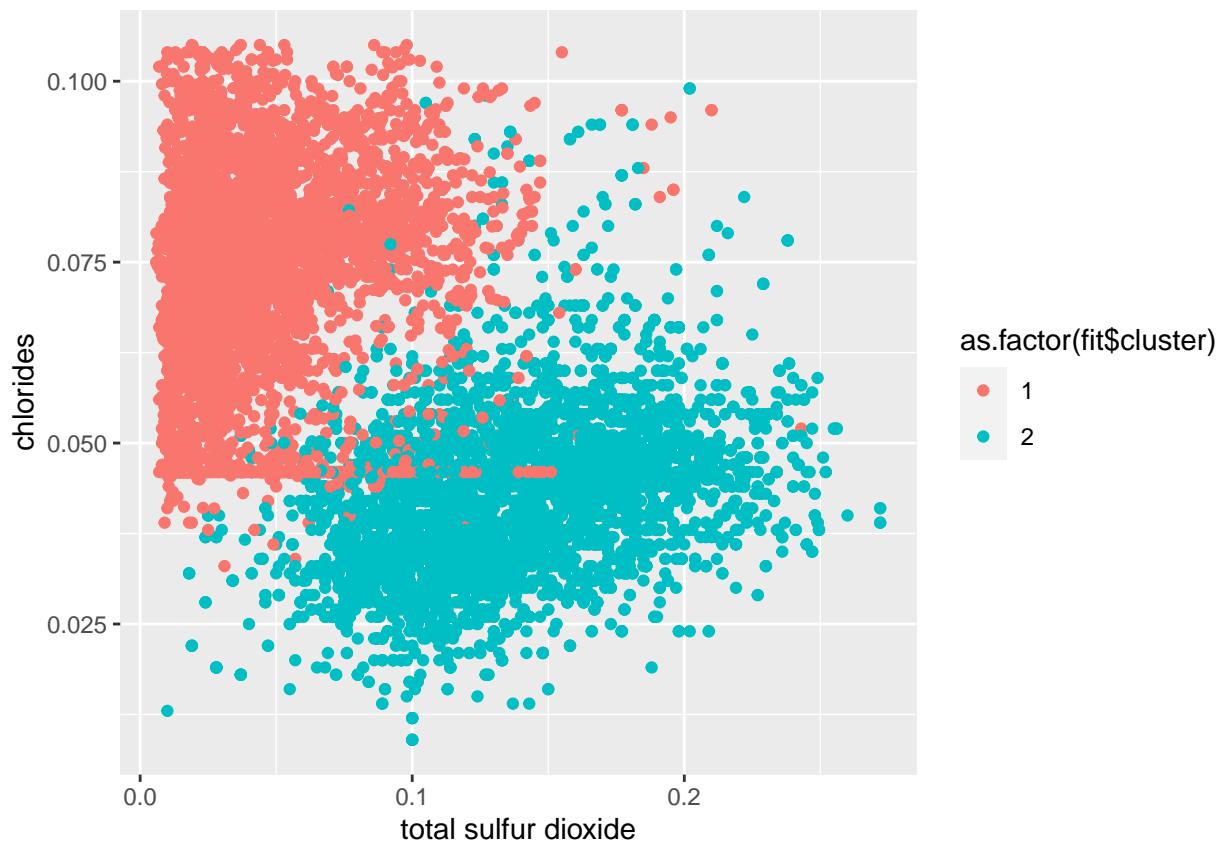
```
var <- get_pca_var(prcomp(vinos_norm))
p1 <- fviz_contrib(prcomp(vinos_norm), choice = "var", axes = 1, top = 5)
p2 <- fviz_contrib(prcomp(vinos_norm), choice = "var", axes = 2, top = 5)
p1 + p2
```

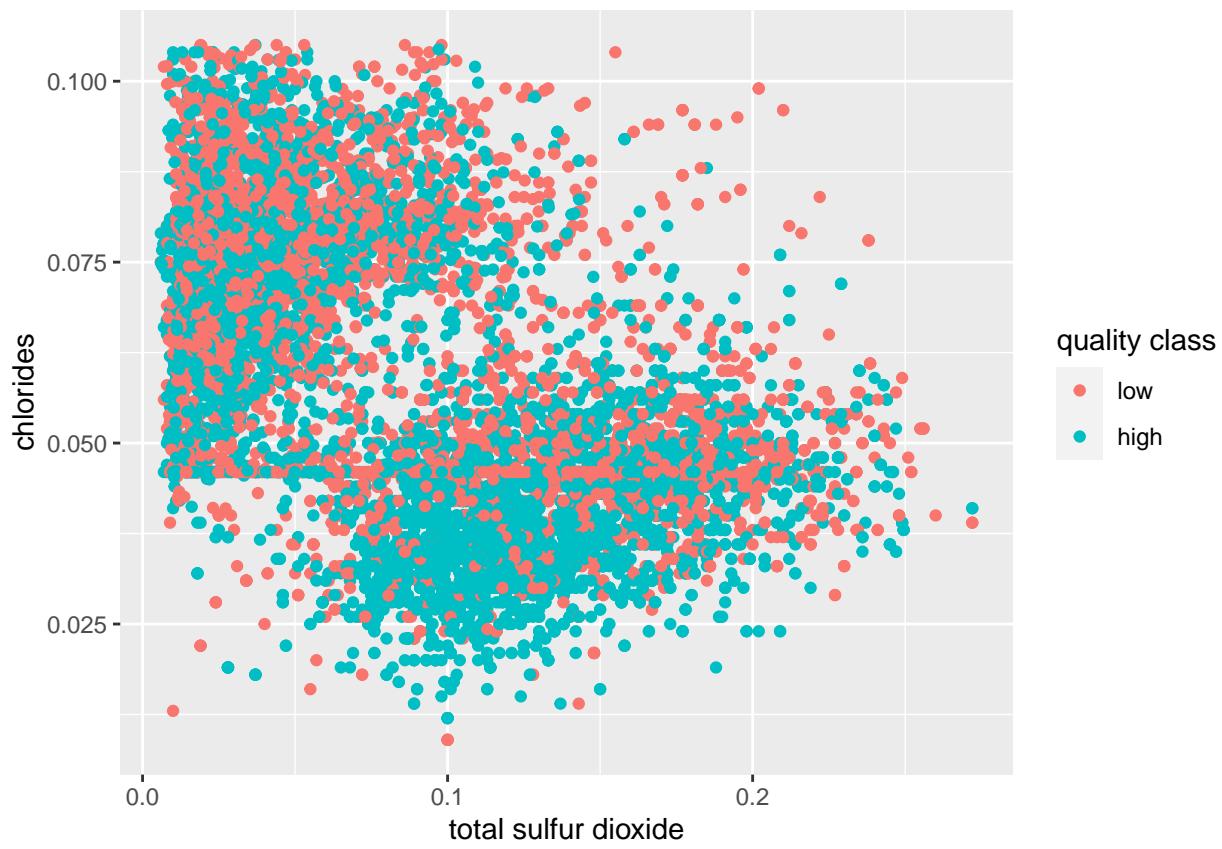


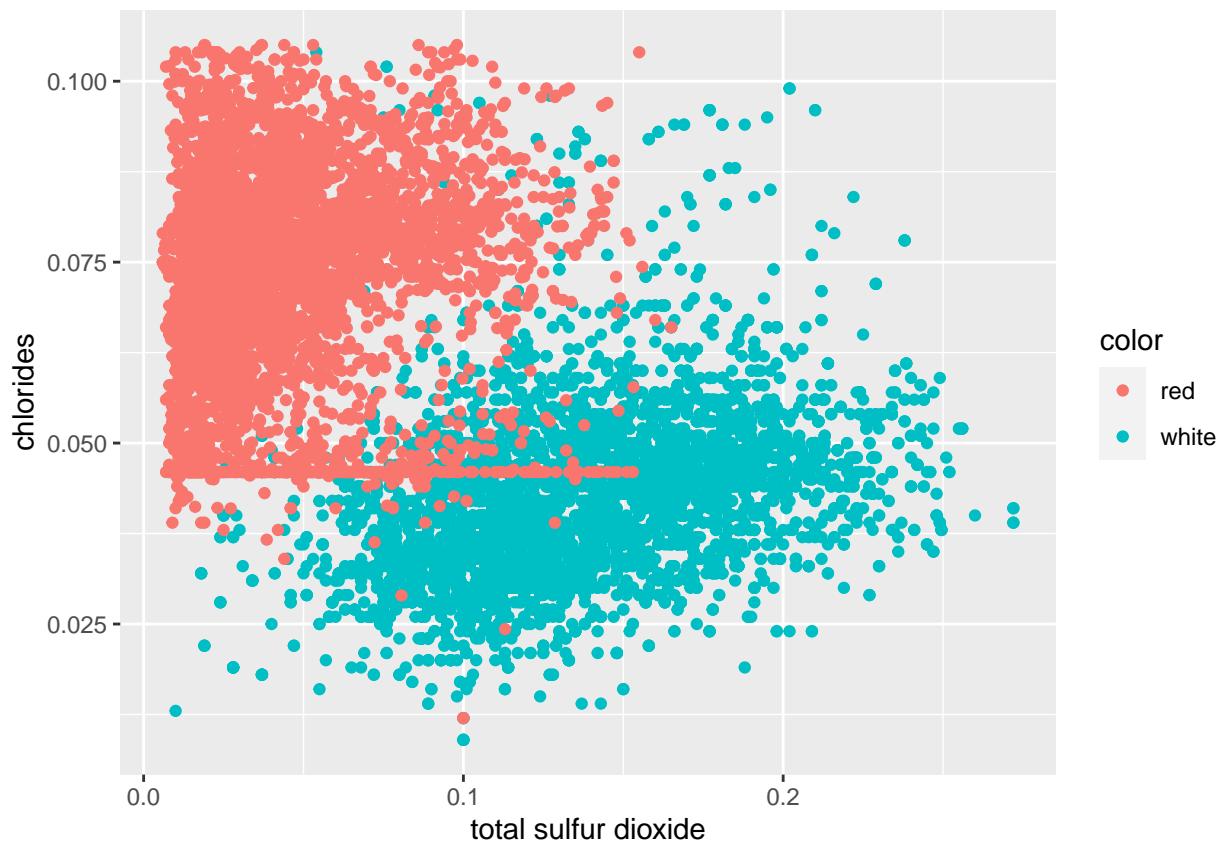
Para poder aproximar las agrupaciones al dataset, visualizamos un gráfico con los atributos explicativos del dataset (no componentes).

Para visualizar los gráficos con los atributos total sulfur dioxide y chlorides (dimensión 1):

```
ggplot(vinos,aes(`total sulfur dioxide`, `chlorides`, color=as.factor(fit$cluster)))+geom_point()
```

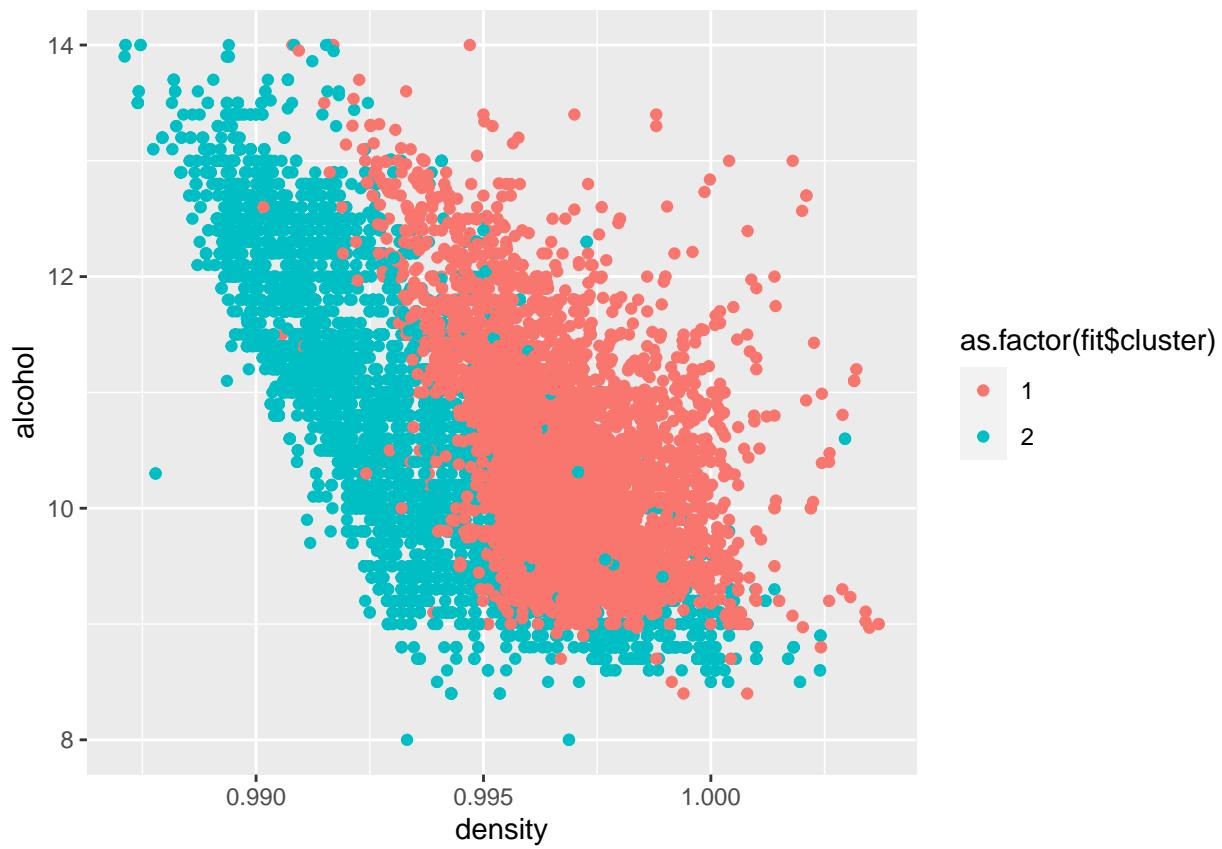


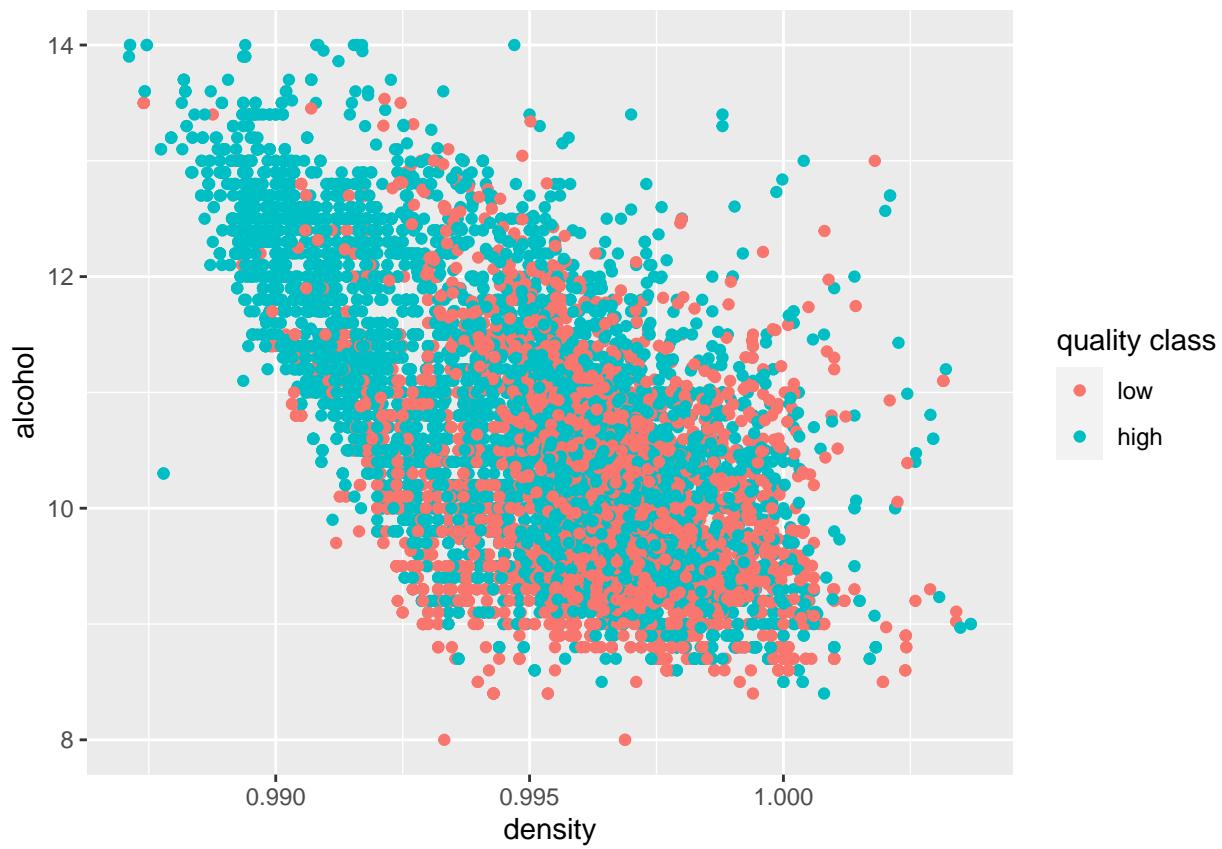


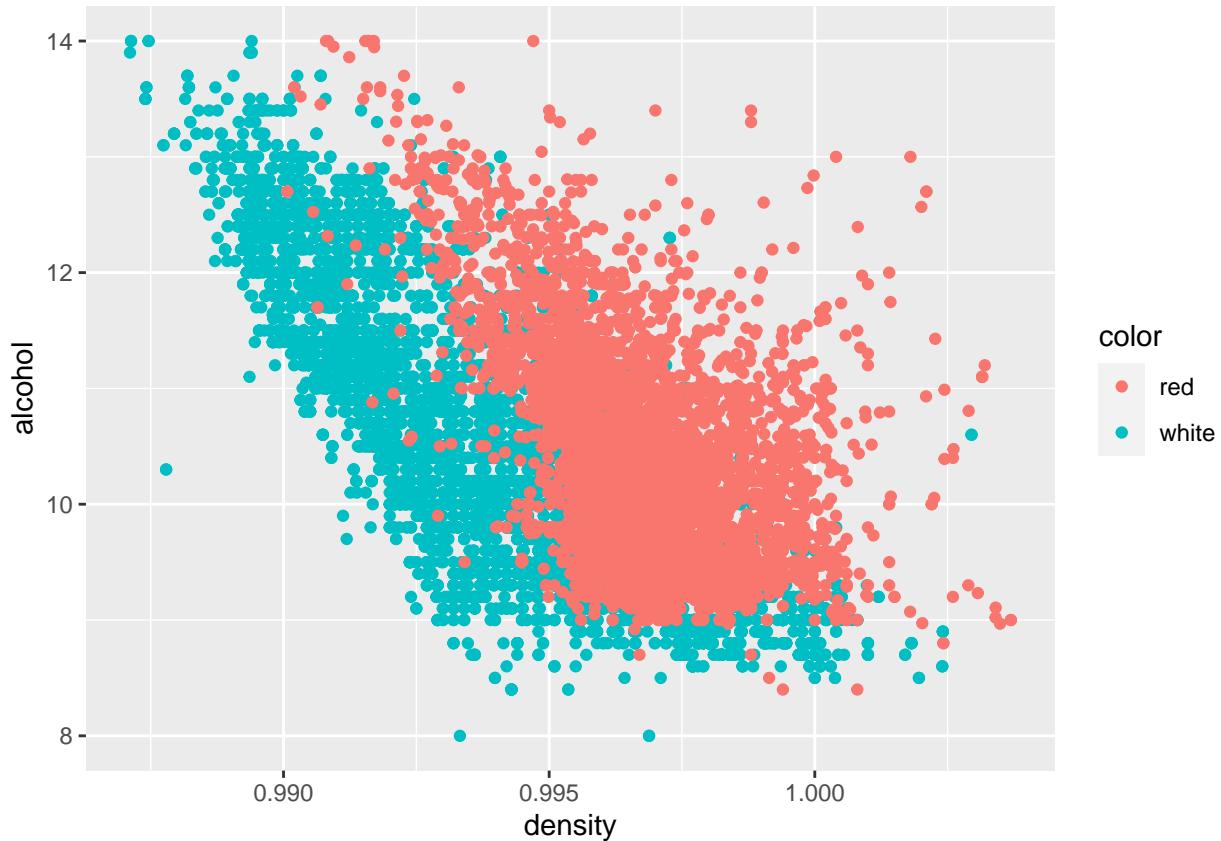


Para visualizar ahora los gráficos con los atributos density y alcohol (dimensión 2):

```
ggplot(vinos,aes(`density`, `alcohol`, color=as.factor(fit$cluster)))+geom_point()
```







En este caso la aproximación de clusters puede captar la naturaleza de las variables segun su color pero no segun quality class, lo que significa que la discretización realizada no es muy significativa, pero que las características tienen que ser distintas segun el color.

Conclusión / Resolución del problema.

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

En principio no podemos concluir que sea posible determinar la calidad de vino según su sensory data con alta precisión. Tenemos evidencia que el color sí que está correlacionado con los atributos.

(no acabado)