Práctica 2: Clasificación vino Blanco

Autores: Jorge Miranda Álamo y José Manuel García Rodes

31 de diciembre 2020

Contents

Descripción de las variables contenidas en el fichero

A partir del conjunto de datos* disponible en el siguiente enlace https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv, vamos a realizar un modelo de agregación no supervisado y un modelo de regresión, para clasificar el vino blanco según su calidad, basandonos en los valores aportados por una serie de mediciones fisioquímicas (entradas) y sensoriales (salidas). También vamos a realizar diferentes contrastes de hiótesis para ver si existen diferencias estadísticas entres las medias para los diferentes atributos para los vinos de calidad Alta y de calidad Baja.

* Para más información sobre los datos se puede visitar el siguiente enlace: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names

*P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

El fichero en estudio está formado por las siguientes 12 variables:

Variables de entrada (basadas en pruebas fisicoquímicas):

- **fixed acidity**. *acidez fija:* Continua.
- **volatile acidity**. *acidez volátil:* Continua.
- **citric acid**. *ácido cítrico:* Continua.
- **residual sugar**. *azúcar residual:* Continua.
- **chlorides**. *cloruros:* continuous.
- **free sulfur dioxide**. *dióxido de azufre libre*: Continua.
- **total sulfur dioxide**. *dióxido de azufre total:* Continua.
- **density**. *densidad:* Continua.
- **pH**. *pH:* Continua.
- **sulphates**. *sulfatos:* Continua.
- **alcohol**. *alcohol:* Continua.

Variable de salida (basada en datos sensoriales):

- **quality** . *calidad:* Discreta. Esta variable puede tomar valores entre 0 y 10.

Carga del conjunto de datos

```
# Cargamos los paquetes R que vamos a usar
library(ggplot2)
library(dplyr)
library(tidyr)
library(cluster)

# Establecemos el directorio de trabajo
setwd("./")

url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv"
winequality <- read.csv(url,stringsAsFactors = FALSE, header =TRUE, sep =";" )
```

Procesos de limpieza del conjunto de datos

Primer contacto con el conjunto de datos, visualizamos su estructura.

```
str(winequality)
```

```
## '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 ...
```

Estadísticas básicas.

Pasamos a estudiar las estadisticas básicas del conjunto de datos

```
summary(winequality)
```

```
##  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
```

Tras un primer análisis básico del conjunto de datos, cabe destacar los distintos rangos que toman las variables, siendo todos positivos pero muy dispares, por lo que sugiere la posibilidad de realizar una normalización de los datos.

Trabajamos los atributos con valores vacíos.

Pasamos a comprobar si existen valores vacios o nulos

```
colSums(is.na(winequality))
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
```

```
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

```
colSums(winequality=="")
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

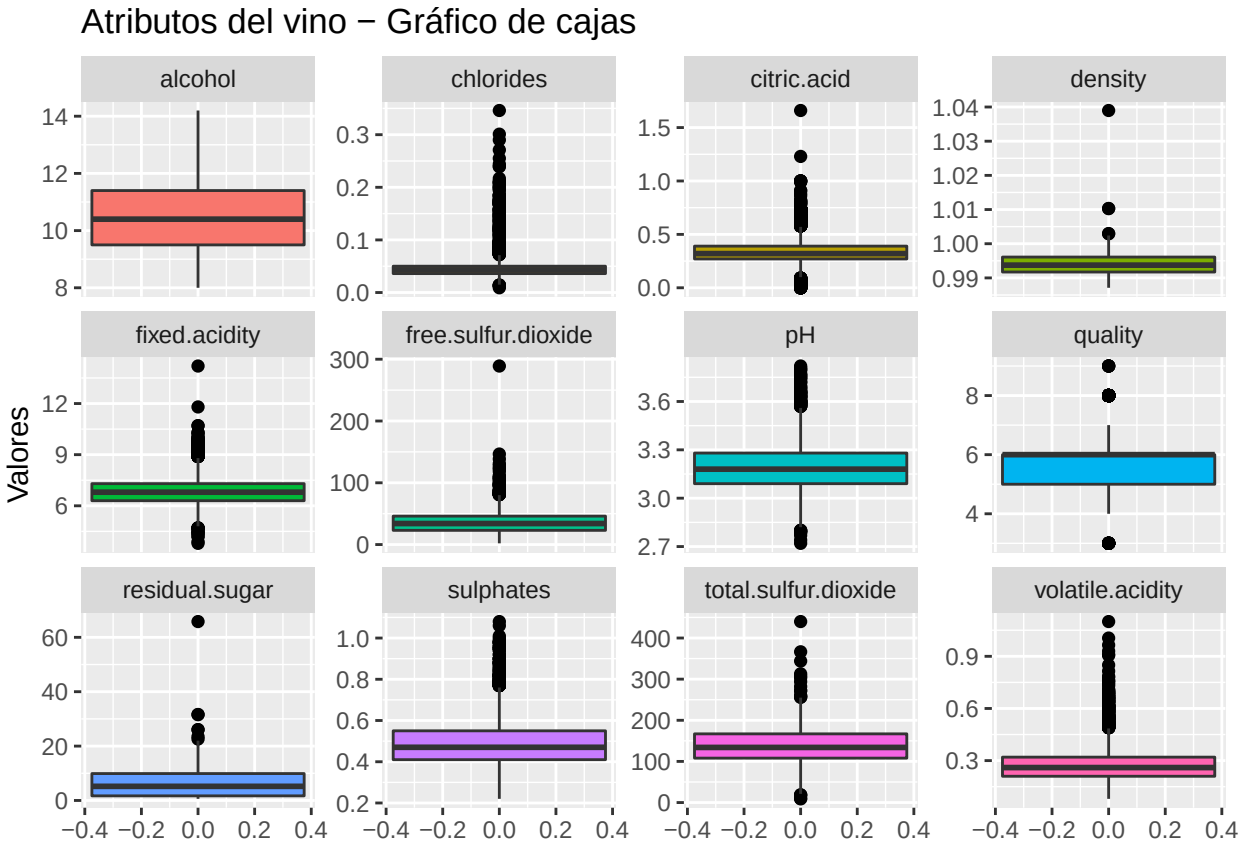
Como podemos apreciar, no existen valores perdidos.

Valores extremos

Los valores extremos (extreme scores o outliers) son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población.

Para detectarlos utilizaremos la ténica de la representación de los datos mediante gráficos de cajas (boxplots), con el objetivo de detectar dichos outliers viendo los valores que distan mucho de la media. Esto sólo es válido para atributos numéricos.

```
winequality %>%
  gather(atributos, valores) %>%
  ggplot() +
  aes(y=valores, fill=atributos) +
  geom_boxplot(outlier.colour="black", outlier.shape=16, outlier.size=2, notch=FALSE) +
  theme(legend.position="none") + # Remove legend
  facet_wrap(~atributos, scales="free_y") +
  labs(y="Valores", title="Atributos del vino - Gráfico de cajas")
```



Vamos a investigar que puntos están muy alejados de del tercer cuartil y del primer cuartil. Para las variables *citric.acid*, *density*, *fixed.acidity*, *free.sulfur.dioxide*, *quality*,*residual.sugar* y *total.sulfur.dioxide* tenemos valores aislados que se alejan del tercer cuartil por la parte superior, y para *quality* y *total.sulfur.dioxide* hay valores que se alejan por la parte inferior.

Mostramos los valores que se alejan más de dos desviaciones típicas de la media.

```
boxplot.stats(winequality$citric.acid)$out
```

```
##   [1] 0.62 0.04 0.59 0.07 0.03 0.61 0.62 0.63 0.61 0.62 0.63 0.66 0.66 0.00 0.04
##  [16] 0.67 0.67 0.04 0.04 0.07 0.88 0.08 0.59 0.07 0.07 0.07 0.07 0.58 0.70 0.00
##  [31] 0.00 0.60 0.07 0.09 0.04 0.62 0.58 0.62 0.70 0.62 0.62 0.58 0.02 0.65 0.65
##  [46] 0.71 0.66 0.66 0.07 0.06 0.07 0.06 0.68 0.68 0.68 0.68 0.06 0.72 0.69 0.58
##  [61] 0.70 1.66 0.04 0.63 0.60 0.00 0.08 0.58 0.58 0.05 0.58 0.00 0.00 0.65 0.58
##  [76] 0.00 0.05 0.05 0.62 0.62 0.58 0.58 1.00 0.09 0.01 0.71 0.71 0.60 0.06 0.74
##  [91] 0.81 0.69 0.58 0.69 0.00 0.07 0.64 0.72 0.73 0.65 0.68 0.65 0.74 0.71 0.59
## [106] 0.68 0.08 0.72 0.64 0.02 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74
## [121] 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.99 0.74 0.74 0.74
## [136] 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.01 0.74 0.01 0.74
## [151] 0.74 1.00 0.04 0.58 0.07 1.00 0.00 0.58 0.61 0.61 0.61 0.02 0.67 0.67 0.67
## [166] 0.58 0.65 0.58 0.09 0.08 0.71 0.04 0.03 0.05 0.64 0.64 0.58 0.58 0.81 0.58
## [181] 0.61 0.62 0.59 0.00 0.04 0.63 0.73 0.68 0.09 0.78 0.79 0.09 0.64 0.65 0.65
## [196] 0.00 0.73 0.73 0.64 0.60 0.71 0.72 0.82 0.07 0.58 0.58 1.00 0.66 0.80 0.80
## [211] 1.23 0.59 0.02 0.00 1.00 0.62 0.00 0.71 0.71 0.71 0.61 0.61 0.00 0.60 0.58
## [226] 0.09 0.09 0.72 0.62 0.62 0.79 0.82 0.67 0.01 0.01 0.86 0.61 0.02 0.05 0.00
## [241] 0.69 0.69 0.59 0.01 0.66 0.66 0.78 0.00 0.04 0.91 0.91 0.06 0.06 0.04 0.04
## [256] 0.74 0.09 0.09 0.60 0.62 0.73 0.00 0.09 0.00 0.09 0.67 0.01 0.09 0.00 0.02
```

```
boxplot.stats(winequality$density)$out
```

```
## [1] 1.01030 1.01030 1.03898 1.00295 1.00295
```

```
boxplot.stats(winequality$fixed.acidity)$out
```

```
##   [1]  9.8  9.8 10.2  9.1 10.0  9.2  9.2  9.0  9.1  9.2 10.3  9.4  9.2  9.8  9.6
##  [16]  9.2  9.0  9.3  9.2  9.1  8.9  9.8  8.9  9.2  9.7  9.4 10.3  9.6  9.0  9.7
##  [31]  9.2  9.4  9.6  9.2  9.0  9.2 10.7 10.7  9.0  9.2  9.8  9.2 14.2  8.9  8.9
##  [46]  9.1  9.1  9.8  9.0  9.3  8.9  9.0  9.0  8.9  9.0  9.3  9.2  9.6  9.4  9.4
##  [61] 10.0  8.9  8.9 10.0  9.2  9.2  9.2  9.9  9.5  9.0  9.0  8.9  9.5 11.8  9.4
##  [76]  9.1  9.8  9.9  9.2  8.9  9.2  9.4  9.4  9.4  4.6  8.9  9.4  9.2  9.2  9.8
##  [91]  9.0  9.0  9.0  8.9  8.9  4.5  9.2  9.6  4.2  9.7  9.7  9.0  4.2  9.4  8.9
## [106]  8.9  8.9  4.7  4.7  3.8  4.4  4.7  9.0  9.0  4.7  4.4  3.9  4.7  4.4
```

```
boxplot.stats(winequality$free.sulfur.dioxide)$out
```

```
##  [1]  81.0  82.0 131.0  82.5  87.0  87.0  83.0 122.5  83.0  81.0  88.0  82.0
## [13] 118.5  81.0  96.0  83.0  83.0 146.5 128.0 110.0  85.0  89.0  86.0  86.0
## [25]  96.0  96.0  93.0  85.0  81.0 138.5  95.0 124.0  87.0  87.0 105.0 105.0
## [37] 101.0 101.0 108.0 108.0  98.0  98.0 112.0 108.0  98.0  81.0  81.0  81.0
## [49] 289.0  97.0
```

```
boxplot.stats(winequality$quality)$out
```

```
##   [1] 8 8 8 8 8 8 8 8 3 3 8 8 8 3 8 8 8 8 8 3 8 8 8 8 8 3 9 8 8 8 9 9 8 8 8 8 8
##  [38] 8 8 8 8 3 9 8 8 8 8 8 3 8 8 8 8 8 8 8 8 8 8 8 3 8 8 8 8 8 8 8 8 8 8 8 8 8
##  [75] 3 8 3 8 8 8 9 8 8 8 3 8 8 8 8 8 3 8 8 8 8 8 8 3 8 8 8 3 8 8 8 8 8 8 8 8 8
## [112] 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 3 8 8 8 8 8 8 8 3 8 8
## [149] 8 3 8 8 8 3 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 3 8 8 8 8 8 8
## [186] 8 8 8 8 8 8 8 8 8 3 8 8 8 8 8
```

```
boxplot.stats(winequality$residual.sugar)$out
```

```
## [1] 23.50 31.60 31.60 65.80 26.05 26.05 22.60
```

```
boxplot.stats(winequality$total.sulfur.dioxide)$out
```

```
##  [1] 272.0 313.0 260.0  19.0 366.5 307.5 256.0 256.0 344.0 282.0 303.0 272.0
## [13]  18.0  18.0 294.0   9.0  10.0 259.0 440.0
```

```
boxplot.stats(winequality$total.sulfur.dioxide)$out
```

```
##  [1] 272.0 313.0 260.0  19.0 366.5 307.5 256.0 256.0 344.0 282.0 303.0 272.0
## [13]  18.0  18.0 294.0   9.0  10.0 259.0 440.0
```

Estudiamos con más detenimiento los valores más alejados del rango intercuartílico, ordeno de forma descendente el conjunto de datos para cada una de las variables con posibles outliers superiores y ascendente para las variables con posibles outliers inferiores, mostrando las 5 primeras filas. De esta forma podemos comporbar que son valores aislados y no varios valores juntos que el el gráfico apararecen representados por un único punto.

Ordenación ascendente

```
winequality <- arrange(winequality, -citric.acid)
head(winequality$citric.acid, n=5)
```

```
## [1] 1.66 1.23 1.00 1.00 1.00
```

```
winequality <- arrange(winequality, -density)
head(winequality$density, n=5)
```

```
## [1] 1.03898 1.01030 1.01030 1.00295 1.00295
```

```
winequality <- arrange(winequality, -fixed.acidity)
head(winequality$fixed.acidity, n=5)
```

```
## [1] 14.2 11.8 10.7 10.7 10.3
```

```
winequality <- arrange(winequality, -free.sulfur.dioxide)
head(winequality$free.sulfur.dioxide, n=5)
```

```
## [1] 289.0 146.5 138.5 131.0 128.0
```

```
winequality <- arrange(winequality, -quality)
head(winequality$quality, n=5)
```

```
## [1] 9 9 9 9 9
```

```
winequality <- arrange(winequality, -residual.sugar)
head(winequality$residual.sugar, n=5)
```

```
## [1] 65.80 31.60 31.60 26.05 26.05
```

```
winequality <- arrange(winequality, -total.sulfur.dioxide)
head(winequality$total.sulfur.dioxide, n=5)
```

```
## [1] 440.0 366.5 344.0 313.0 307.5
```

Ordenación descendene

```
winequality <- arrange(winequality, quality)
head(winequality$quality, n=5)
```

```
## [1] 3 3 3 3 3
```

```
winequality <- arrange(winequality, total.sulfur.dioxide)
head(winequality$total.sulfur.dioxide, n=5)
```

```
## [1]  9 10 18 18 19
```

A la vista de estos resultados, podemos concluir que son todos valores únicos en los extremos superiores excepto para la variable *quality* que hay varios valores representados por el mismo punto al igual que por la parte inferior.

Para *total.sulfur.dioxide*, como son varios los valores que se escapan por la parte inferior, estos no los vamos a eliminar. Vamos a eliminar las observaciones correspondientes a los outlaiers superiores.

Elimino los valores únicos de los extremos superiores.

```
winequality <- winequality[-which.max(winequality$citric.acid),]
winequality <- winequality[-which.max(winequality$density),]
winequality <- winequality[-which.max(winequality$fixed.acidity),]
winequality <- winequality[-which.max(winequality$free.sulfur.dioxide),]
winequality <- winequality[-which.max(winequality$residual.sugar),]
winequality <- winequality[-which.max(winequality$total.sulfur.dioxide),]
```

Despues de eliminar estos valores aislados y alejados del resto del grupo, habría que ver que hacemos con los siguientes valores alejados del grupo. LLegados a este punto, sería importante tener la opinión de un experto que nos dijera si son observaciones anómalas que podrían deberse a un error de medida o efectivamente son valores que están dentro de un rango posible.

Discretizaión de variables

Vamos a discretizar la variable *quality*. Para ello añadimos una variable nueva a los datos que se llamará *quality_disc*.

Vemos como se distribuyen os valores:

```
# Vemos cómo se distribuyen los valores
summary(winequality[,"quality"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.879   6.000   9.000
```

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

```
##
##    3    4    5    6    7    8    9
##   18  163 1457 2194  880  175    5
```

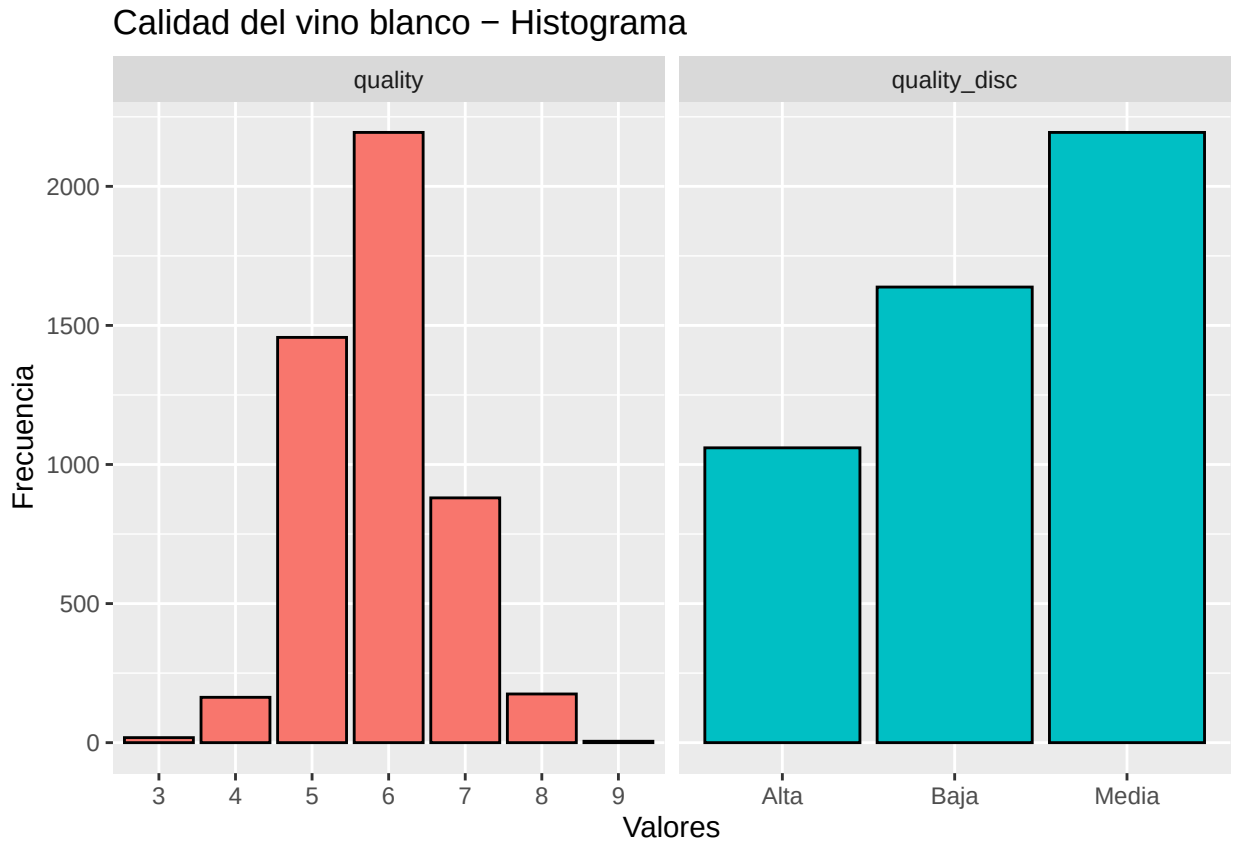
Tras estudiar sus distribuciones de frecuencias, discretizamos agrupando en tres intervalos, estando formado el intervalo central solo por valor “6”, ya que éste es el valor de la calidad que agrupa cerca del 45 % de las observaciones. Los otros dos intervalos estarán formados por los valores de la calidad por encima y por debajo de éste respectivamente.

```
winequality["quality_disc"] <- cut(winequality$"quality", breaks = c(0,5,6,9), labels = c("Baja", "Media
```

Comparamos la distribución de la variable ***quality*** con ***quality_disc***, mediante su representación gráfica.

```
df_quality <- winequality[,12:13]

df_quality %>%
  gather(atributos, valores) %>%
  ggplot() +
  aes(x=valores, fill=atributos) +
  geom_bar(colour="black", show.legend=FALSE) +
  facet_wrap(~atributos, scales="free_x") +
  labs(x="Valores", y="Frecuencia",
       title="Calidad del vino blanco - Histograma")
```



El gráfico ***quality_disc*** nos indica un ligero desbalanzeamiento, habría que aplicar técnicas para corregirlo, pero no vamos a entrar en ello. El gráfico ***quality***, indica que apenas existen vinos blancos excelentes ni malos.

Exportación de los datos preprocesados

Una vez integrados, validados y guardados los datos creamos un nuevo fichero llamado "winequality_clean.csv".

```
write.csv(winequality, "winequality_clean.csv")
```

Análisis de los datos

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

Estudio de la normalidad

En función de los resultados del test de normalidad y homegeneidad de varianza se podrán aplicar tests paramétricos o no paramétricos a nuestro conjunto de datos.

```
library(nortest)
alpha = 0.05
col.names = colnames(winequality)
for (i in 1:ncol(winequality)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(winequality[,i]) | is.numeric(winequality[,i])) {
    p_val = ad.test(winequality[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(winequality) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## fixed.acidity, volatile.acidity, citric.acid,
## residual.sugar, chlorides, free.sulfur.dioxide,
## total.sulfur.dioxide, density, pH,
## sulphates, alcohol, quality
```

Podemos comprobar como ninguna de las variables cuantitativas se distribuye de forma normal, por lo que sería conveniente usar métodos no paramétricos para la realizacion de los tests, pero como tenemos un dataset de 4892 observaciones también se podrán usar métodos paramétricos.

Estudio de la Homogeneidad de varianzas

En este apartado vamos a realizar el test estadisticos de *Fligner-Killeen* para contras tar la hipótesis nula de que las varianzas son iguales o por el contratio son son distintas con un nivel de confianza del 5%.

```
# fixed.acidity
fligner.test(fixed.acidity ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  fixed.acidity by quality_disc
## Fligner-Killeen:med chi-squared = 9.4526, df = 2, p-value = 0.008859
```

```r
# volatile.acidity
fligner.test(volatile.acidity ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  volatile.acidity by quality_disc
## Fligner-Killeen:med chi-squared = 35.732, df = 2, p-value = 1.742e-08
```

```r
# citric.acid
fligner.test(citric.acid ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  citric.acid by quality_disc
## Fligner-Killeen:med chi-squared = 300.32, df = 2, p-value < 2.2e-16
```

```r
# residual.sugar
fligner.test(residual.sugar ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  residual.sugar by quality_disc
## Fligner-Killeen:med chi-squared = 140.42, df = 2, p-value < 2.2e-16
```

```r
# chlorides
fligner.test(chlorides ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chlorides by quality_disc
## Fligner-Killeen:med chi-squared = 26.394, df = 2, p-value = 1.857e-06
```

```r
# free.sulfur.dioxide
fligner.test(free.sulfur.dioxide ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  free.sulfur.dioxide by quality_disc
## Fligner-Killeen:med chi-squared = 194.43, df = 2, p-value < 2.2e-16
```

```r
# total.sulfur.dioxide
fligner.test(total.sulfur.dioxide ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  total.sulfur.dioxide by quality_disc
## Fligner-Killeen:med chi-squared = 135.2, df = 2, p-value < 2.2e-16
```

```
# density
fligner.test(density ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  density by quality_disc
## Fligner-Killeen:med chi-squared = 32.157, df = 2, p-value = 1.04e-07
```

```
# pH
fligner.test(pH ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  pH by quality_disc
## Fligner-Killeen:med chi-squared = 33.512, df = 2, p-value = 5.285e-08
```

```
# sulphates
fligner.test(sulphates ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  sulphates by quality_disc
## Fligner-Killeen:med chi-squared = 83.694, df = 2, p-value < 2.2e-16
```

```
# alcohol
fligner.test(alcohol ~ quality_disc, data = winequality)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  alcohol by quality_disc
## Fligner-Killeen:med chi-squared = 229.55, df = 2, p-value < 2.2e-16
```

Tras realizar el test de homgeneidad de varianzas a todas las variales para cada una de las categorías de *quality_disc*, podemos comprobar como han salido todos los p-value $<$ de 0,05, por lo que rechazamos la hipótesis nula de homogeneidad de varianzas.

Reducción de la dimensionalidad

En este apartado vamos a tratar de identificar variables que tienen el mismo comportamiento, para incluir a una sola en nuestro conjunto de datos y de esta forma no tener información redundante y hacer más facil la interpretación de los datos.

Estudio de la correlación Una forma de reducir la dimensionalidad es sacar del modelo las variables que estén correlacionadas, para ello podemos calcular la matriz de correlaciones de las variables sacando del conjunto de datos la variable *quality_disc* ya que esta es una variable categórica.

Cargo los datos sin la variable *quality_disc*

```
df <- winequality[,1:12]
head(df)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           4.7             0.67        0.09            1.0     0.020
## 2           4.8             0.65        0.12            1.1     0.013
## 3           9.7             0.24        0.49            4.9     0.032
## 4           9.7             0.24        0.49            4.9     0.032
## 5           6.9             0.39        0.40            4.6     0.022
## 6           5.9             0.19        0.37            0.8     0.027
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                   5                    9 0.98722 3.30      0.34    13.6
## 2                   4                   10 0.99246 3.32      0.36    13.5
## 3                   3                   18 0.99368 2.85      0.54    10.0
## 4                   3                   18 0.99368 2.85      0.54    10.0
## 5                   5                   19 0.99150 3.31      0.37    12.6
## 6                   3                   21 0.98970 3.09      0.31    10.8
##   quality
## 1       5
## 2       4
## 3       6
## 4       6
## 5       3
## 6       5
```

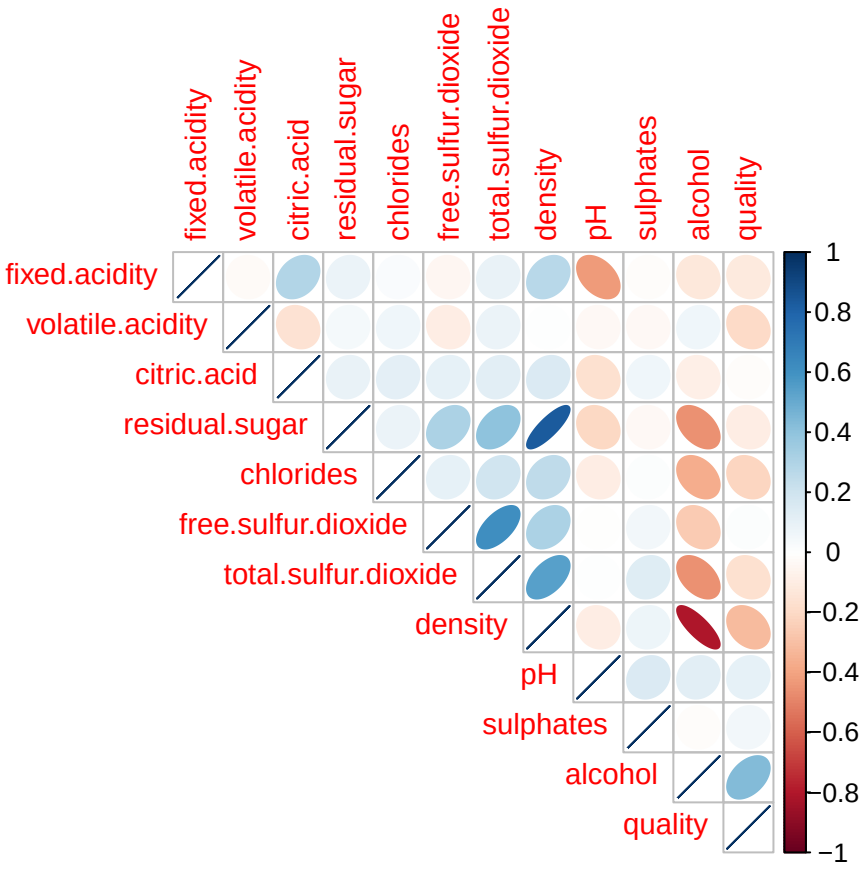
Calculamos la matriz de correlaciones y su representación gráfica

```
library(corrplot)
cor(df)
```

```
##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.025678165  0.29115325     0.08842823
## volatile.acidity       -0.02567817      1.000000000 -0.15368330     0.04736838
## citric.acid             0.29115325     -0.153683299  1.00000000     0.09426223
## residual.sugar          0.08842823      0.047368376  0.09426223     1.00000000
## chlorides               0.02356784      0.068626332  0.11793134     0.08632128
## free.sulfur.dioxide    -0.04756683     -0.097037275  0.10074500     0.31774285
## total.sulfur.dioxide    0.09061777      0.086874592  0.12573916     0.40869437
## density                 0.27047647      0.004503821  0.15111040     0.83287758
## pH                     -0.42894784     -0.033281060 -0.16739309    -0.20001685
## sulphates              -0.01860695     -0.038853321  0.06116074    -0.03051843
## alcohol                -0.12307072      0.066939266 -0.08058657    -0.45949490
## quality                -0.11444305     -0.194742498 -0.01012870    -0.09916974
##                       chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.02356784        -0.047566827         0.0906177728
## volatile.acidity      0.06862633        -0.097037275         0.0868745916
## citric.acid           0.11793134         0.100745005         0.1257391578
```

```
## residual.sugar        0.08632128         0.317742854         0.4086943657
## chlorides             1.00000000         0.104072072         0.1996954049
## free.sulfur.dioxide   0.10407207         1.000000000         0.6132594111
## total.sulfur.dioxide  0.19969540         0.613259411         1.0000000000
## density               0.25953116         0.315683332         0.5445822021
## pH                   -0.09069207        -0.005385662         0.0009748636
## sulphates             0.01639283         0.057170035         0.1325269871
## alcohol              -0.36036025        -0.255805665        -0.4530191865
## quality              -0.21008758         0.018590674        -0.1682514193
##                          density            pH   sulphates     alcohol
## fixed.acidity         0.270476471 -0.4289478416 -0.01860695 -0.12307072
## volatile.acidity      0.004503821 -0.0332810602 -0.03885332  0.06693927
## citric.acid           0.151110396 -0.1673930907  0.06116074 -0.08058657
## residual.sugar        0.832877579 -0.2000168477 -0.03051843 -0.45949490
## chlorides             0.259531163 -0.0906920681  0.01639283 -0.36036025
## free.sulfur.dioxide   0.315683332 -0.0053856621  0.05717003 -0.25580567
## total.sulfur.dioxide  0.544582202  0.0009748636  0.13252699 -0.45301919
## density               1.000000000 -0.0996097395  0.07188928 -0.80353108
## pH                   -0.099609740  1.0000000000  0.15540861  0.12120811
## sulphates             0.071889284  0.1554086054  1.00000000 -0.01836207
## alcohol              -0.803531081  0.1212081075 -0.01836207  1.00000000
## quality              -0.315692476  0.1001227238  0.05548129  0.43696486
##                         quality
## fixed.acidity        -0.11444305
## volatile.acidity     -0.19474250
## citric.acid          -0.01012870
## residual.sugar       -0.09916974
## chlorides            -0.21008758
## free.sulfur.dioxide   0.01859067
## total.sulfur.dioxide -0.16825142
## density              -0.31569248
## pH                    0.10012272
## sulphates             0.05548129
## alcohol               0.43696486
## quality               1.00000000
```

```
corrplot(cor(df), type="upper", method="ellipse", tl.cex=0.9)
```



Variables a sacar del conjunto de datos:

- El **alcohol** esta fuertemente correlacionado con la densidad, con un coeficiente de correlación del -0,80, lo que significa que a mayor cantidad de alcohol menor **densidad**. También está correlacionado inversamente proporcional, en menor medida, con el **azúcar residual** y con el **total de dióxido de sulfuro** con coeficientes de correlación -0,46 y -0,45 respectivamente.
- El coeficiente de correlación entre el **pH** y la **acided fija** es del -0.43, aunque no es muy fuerte, lo vamos a tener en cuenta para reducir el numero de variables.
- La correlación entre **densidad** y **azúcar residual** es fuerte con un coeficiente de 0,83, por lo que a mayor densidad mayor azucar residual. La **densidad** también está directamente correlacionada de forma más debil con **total de dióxido de sulfuro** (0,55), a mayor densidad mayor dioxdido de sulfuro total.
- El **total de dioxido de sulfuro** esta directamente correlacionado con el **dióxido de sulfuro libre** (0,61).

Por lo tanto sacamos del modelo la **densidad**, el **azúcar residual**, el **total de dióxido de sulfuro** y la **acided fija**.

```
df <- select(df,-1,-4,-7,-8)
head(df)
```

```
##   volatile.acidity citric.acid chlorides free.sulfur.dioxide   pH sulphates
## 1             0.67        0.09     0.020                   5 3.30      0.34
```

```
## 2            0.65        0.12     0.013                    4 3.32      0.36
## 3            0.24        0.49     0.032                    3 2.85      0.54
## 4            0.24        0.49     0.032                    3 2.85      0.54
## 5            0.39        0.40     0.022                    5 3.31      0.37
## 6            0.19        0.37     0.027                    3 3.09      0.31
##   alcohol quality
## 1    13.6       5
## 2    13.5       4
## 3    10.0       6
## 4    10.0       6
## 5    12.6       3
## 6    10.8       5
```

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes

Contrastes de hipótesis

A pesar de que los datos no se distribuyen normales y dado que tenemos más de 30 observaciones (4892), vamos a utilizar métodos paramétricos para la realización de los contrastes de hipótesis.

En primer lugar vamos a contrastar la hipótesis de que los vinos con calidad **Alta** tienen más graduación alcohólica que los vinos de calidad **Baja**.

```
winequality_alta_alcohol <- winequality[winequality$quality_disc == "Alta",]$alcohol
winequality_baja_alcohol <- winequality[winequality$quality_disc == "Baja",]$alcohol

t.test(winequality_baja_alcohol, winequality_alta_alcohol, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  winequality_baja_alcohol and winequality_alta_alcohol
## t = -35.453, df = 1721.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.494823
## sample estimates:
## mean of x mean of y
##  9.848431 11.416022
```

Del resultado del test anterior se desprende que la graduación media de alcohol de los vinos con calidad **Alta** es de 11.42 grados y la de los vinos de calidad **Baja** es de 9.85 grados, siendo la diferecia de estas medias lo suficientemente grande (desde el punto de vista estadítico) como para poder afirmar con una probabilidad del 0.95 que; **la graduación media de los vinos de calidad alta es mayor a la de los vinos de calidad baja**, dado que p-value = 2.2e-16 < 0.05.

Repetimos el contraste de hipótesis para el **pH**.

```r
winequality_alta_pH <- winequality[winequality$quality_disc == "Alta",]$pH
winequality_baja_pH <- winequality[winequality$quality_disc == "Baja",]$pH

t.test(winequality_baja_pH, winequality_alta_pH, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  winequality_baja_pH and winequality_alta_pH
## t = -7.4601, df = 2119.6, p-value = 6.269e-14
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.03488648
## sample estimates:
## mean of x mean of y
##  3.170372  3.215132
```

A la vista de los resultados del test, rechazamos la hipótesis nula de que las medias del **pH** sean las mismas para los vinos de calidad alta que para los de calidad baja (p-value = 6.269e-14 < 0.05), por lo que podemos afirmar que el pH de los vinos de calidad **Alta** es mayor que los de calidad **Baja**.

Repetimos el contraste de hipótesis para el **sulphates**.

```r
winequality_alta_sulphates <- winequality[winequality$quality_disc == "Alta",]$sulphates
winequality_baja_sulphates <- winequality[winequality$quality_disc == "Baja",]$sulphates

t.test(winequality_baja_sulphates, winequality_alta_sulphates, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  winequality_baja_sulphates and winequality_alta_sulphates
## t = -3.9366, df = 1824.3, p-value = 4.288e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01095392
## sample estimates:
## mean of x mean of y
## 0.4813187 0.5001415
```

Obtenemos un p-value = 4.288e-05, por lo que podemos afirmar que la media de **sulfitos** es mayor en los vinos de calidad **Alta** que en los de calidad **Baja**.

Repetimos el contraste de hipótesis para el **free.sulfur.dioxide**.

```r
winequality_alta_free.sulfur.dioxide <- winequality[winequality$quality_disc == "Alta",]$free.sulfur.dio
winequality_baja_free.sulfur.dioxide <- winequality[winequality$quality_disc == "Baja",]$free.sulfur.dio

t.test(winequality_alta_free.sulfur.dioxide, winequality_baja_free.sulfur.dioxide, alternative = "less")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  winequality_alta_free.sulfur.dioxide and winequality_baja_free.sulfur.dioxide
## t = -0.99459, df = 2667.9, p-value = 0.16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.4144068
## sample estimates:
## mean of x mean of y
##  34.55047  35.18376
```

El p-value para *free.sulfur.dioxide* es de 0.16 > que 0.05 por lo que no podemos rechazar la hipótesis nula de que las medias son iguales.

Repetimos el contraste de hipótesis para el **chlorides**.

```
winequality_alta_chlorides <- winequality[winequality$quality_disc == "Alta",]$chlorides
winequality_baja_chlorides <- winequality[winequality$quality_disc == "Baja",]$chlorides

t.test(winequality_alta_chlorides, winequality_baja_chlorides, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  winequality_alta_chlorides and winequality_baja_chlorides
## t = -17.83, df = 2369.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01204978
## sample estimates:
##  mean of x  mean of y
## 0.03816038 0.05143529
```

Rechazamos la hipótesis nula (p-value = 2.2e-16 < 0.05), los vinos de calidad **Alta** tienen menor cantidad de **chlorides** que los vinos de calidad **Baja**.

Repetimos el contraste de hipótesis para el **citric.acid**.

```
winequality_alta_citric.acid <- winequality[winequality$quality_disc == "Alta",]$citric.acid
winequality_baja_citric.acid <- winequality[winequality$quality_disc == "Baja",]$citric.acid

t.test(winequality_alta_citric.acid, winequality_baja_citric.acid, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  winequality_alta_citric.acid and winequality_baja_citric.acid
## t = -1.9249, df = 2648.5, p-value = 0.02717
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.001204547
## sample estimates:
## mean of x mean of y
## 0.3260566 0.3343529
```

Rechazamos la hipótesis nula (p-value = 0.02717 < 0.05), los vinos de calidad **Alta** tienen menor cantidad de **ácido cítrico** que los vinos de calidad **Baja**.

Repetimos el contraste de hipótesis para el **volatile.acidity**.

```
winequality_alta_volatile.acidity <- winequality[winequality$quality_disc == "Alta",]$volatile.acidity
winequality_baja_volatile.acidity <- winequality[winequality$quality_disc == "Baja",]$volatile.acidity

t.test(winequality_alta_volatile.acidity, winequality_baja_volatile.acidity, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  winequality_alta_volatile.acidity and winequality_baja_volatile.acidity
## t = -11.175, df = 2525.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.03820363
## sample estimates:
## mean of x mean of y
## 0.2653491 0.3101496
```

Rechazamos la hipótesis nula (p-value = 2.2e-16 < 0.05), los vinos de calidad **Alta** tienen menor cantidad de **acided volátil** que los vinos de calidad **Baja**.

Conclusión Los vinos de calidad **Alta** se caracterizán mayoritariamente por tener una **mayor graduación alcohólica**, un **mayor pH** y una **mayor cantidad de sulfitos** que los vinos de calidad **Baja** .Por otra parte también se caracterizan por tener una **menor** cantidad de **chlorides**, **ácido cítrico** y **acided volátil** que los de calidad baja.

Modelo de regresión lineal

Vamos a tratar de predecir la calidad del vino en fución a sus características conocidas.

```
#Regresores cuantitativos
acidez = winequality$volatile.acidity
acido_citrico = winequality$citric.acid
calorias = winequality$chlorides
dioxido = winequality$free.sulfur.dioxide
ph = winequality$pH
sulfitos = winequality$sulphates
alcohol = winequality$alcohol

# Variable a predecir
calidad = winequality$quality

# Generación de varios modelos
modelo1 <- lm(calidad ~ acidez + acido_citrico + calorias + dioxido +
              ph + sulfitos + alcohol, data = winequality)
```

Mostramos los resultados del modelo.

```r
# Resultado del modelo
summary(modelo1)
```

```
##
## Call:
## lm(formula = calidad ~ acidez + acido_citrico + calorias + dioxido +
##     ph + sulfitos + alcohol, data = winequality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6125 -0.4925 -0.0448  0.4820  3.1926
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.1159248  0.2645996   7.997 1.58e-15 ***
## acidez        -1.8584887  0.1112908 -16.699  < 2e-16 ***
## acido_citrico -0.0838949  0.0948474  -0.885 0.376456
## calorias      -1.5239591  0.5404997  -2.820 0.004829 **
## dioxido        0.0063645  0.0006828   9.321  < 2e-16 ***
## ph             0.1455511  0.0748910   1.944 0.052012 .
## sulfitos       0.3598785  0.0971417   3.705 0.000214 ***
## alcohol        0.3341186  0.0098383  33.961  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7614 on 4884 degrees of freedom
## Multiple R-squared:  0.2596, Adjusted R-squared:  0.2585
## F-statistic: 244.6 on 7 and 4884 DF,  p-value: < 2.2e-16
```

Conclusión El coeficiente de determinación (R2) nos mide la bondad del ajuste. Ha tomado un valor de **0.2585** lejos de **1**, lo que nos indica que este modelo no va a ser bueno para predecir la calidad del vino.

Modelo de agrupamiento K-Means

En este apartado vamos a tratar de aplicar el modelo de agrupamiento K-Means a nuestro dataframe resultante de los procesos de limpieza y reducción de la dimensionalidad, para tratar de agrupar de manera no supervisada a las distintas mediciones de las características de los vinos blancos.

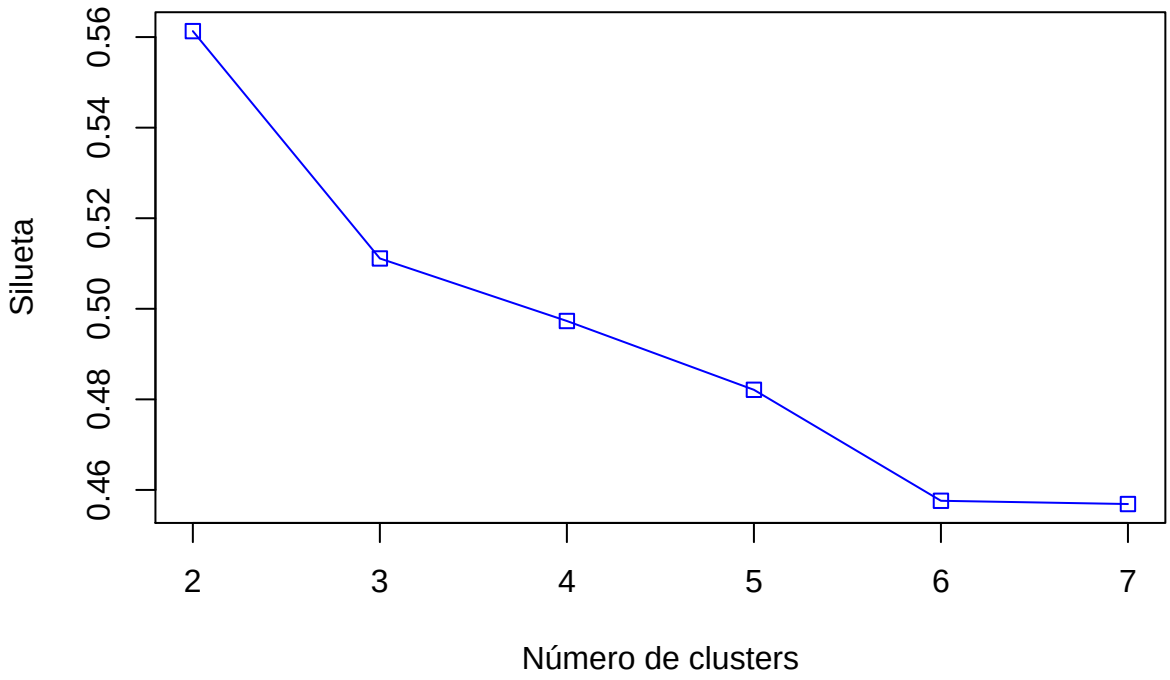
Utilazamos el modelo de agrupamiento K-Means para $\mathbf{k = 2,3,4,5,6,7}$, elegimos **7** por ser el número de categorías de la variable ***quality*** sin discretizar. Como medida de la calidad del proceso de agregación utilizamos la función silhouette.

Visualizamos los clusters.

```r
d <- daisy(df)
resultados <- rep(0, 10)
for (i in c(2,3,4,5,6,7))
{
  fit           <- kmeans(df, i)
  y_cluster     <- fit$cluster
  sk            <- silhouette(y_cluster, d)
  resultados[i] <- mean(sk[,3])
}
```

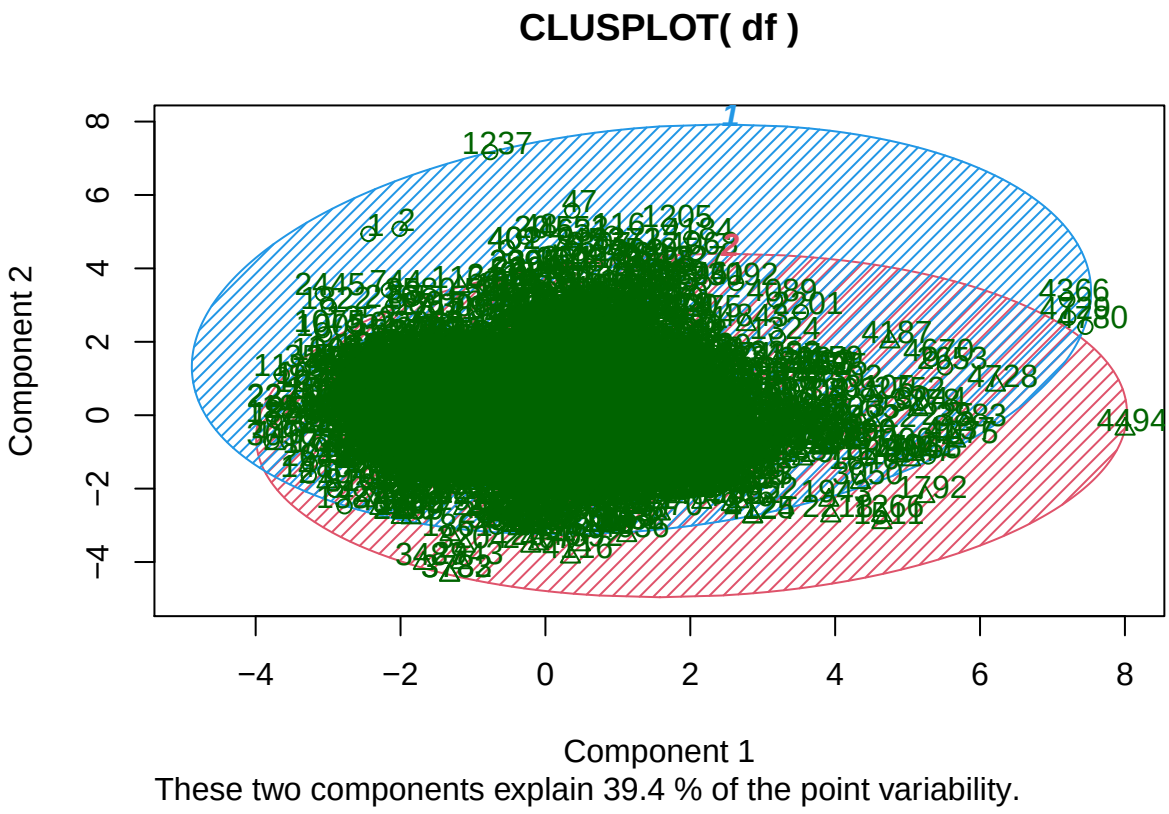
Mostramos en un gráfica los valores de las siluetas medias de cada prueba para comprobar que número de clústers es el mejor.

```
plot(2:7,resultados[2:7],type="o",col="blue",pch=0,xlab="Número de clusters",ylab="Silueta")
```



A la vista de la gráfica anterior, el número de clusters óptimo sería de 2, con un valor medio de la silueta de 0,57, lo que nos indica que la agrupación no va a ser muy buena.

```
fit           <- kmeans(df, 2)
 y_cluster     <- fit$cluster
 sk            <- silhouette(y_cluster, d)
 resultados[2] <- mean(sk[,3])
 clusplot(df, fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```



Fijandonos en el gráfico anterior, no parece que los datos se vayan a poder agrupar en clústers separados.

Conclusión El método ***k-means*** no nos aporta una agrupación clara para las distintas variables.

Representación de los resultados a partir de tablas y gráficas

Estas representaciones y gráficas se han ido realizando a lo largo del desarrollo de esta práctica.

Conclusión final

Hemos agrupado la varialble **quality** en tres categorías y aplicando técnicas básicas estadísticas como son los contrastes de hipótesis hemos llegado a una caracterización de los vinos de calidad alta y de calidad baja en función de sus propiedades químicas, llgando a la siguiente conclusión:

Los vinos de calidad **Alta** se caracterizán mayoritariamente por tener una **mayor graduación alcohólica**, un **mayor pH** y una **mayor cantidad de sulfitos** que los vinos de calidad **Baja** .Por otra parte también se caracterizan por tener una **menor** cantidad de **chlorides**, **ácido cítrico** y **acided volátil** que los de calidad baja.

Por otra parte, este conjunto de datos no nos ha permitido una agrupación de los vinos blancos en función de sus propiedades químicas utilizando los métodos **K-means** y **regresión**, con estos métodos no podemos clasificar o hacer predicciones para la calidad medida por medios sensoriales a través de las medidas químicas.

Contribuciones al trabajo

Contribuciones	Contribuciones
Investigación previa	Jorge Miranda Álamo, José Manuel García Rodes
Redacción de las respuestas	Jorge Miranda Álamo, José Manuel García Rodes
Desarrollo código	Jorge Miranda Álamo, José Manuel García Rodes