Tipología y ciclo de vida de los datos: PRA2

Autor: Iván López-Baltasar Benito | David Quiles Gómez

Junio 2019

1 Introducción

1.1 Presentación

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, dataset), orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2 Objetivos

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.3 Competencias

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis

1.4 Descripcion del dataset

En ésta práctica vamos a trabajar con el juego de datos de https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/ (https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/) el cual contiene dos datasets, uno de vinos blancos y otro de vinos tintos.

Ambos datasets contienen 11 atributos de entrada, correspondientes a pruebas fisioquímicas, y uno de salida: “quality”.

El objetivo del análisis será por un lado construir un modelo que nos pueda predecir la calidad de un vino, y por otro, construir un modelo que nos permita clasificar un vino en un determinado tipo (blanco/tinto).

---

2 Carga y limpieza del dataset

Cargamos los paquetes R que vamos a usar

```
library(ggplot2)
library(dplyr)
```

```
blanco<-read.csv("vinos/winequality-white.csv", header=T, sep=";")
tinto<-read.csv("vinos/winequality-red.csv", header=T, sep=";")
```

Vamos a añadirle la clase a cada juego de datos para después unir ambos datasets.

```
blanco$tipo<-'B'
tinto$tipo<-'T'

nomCols <- c("acidez_fija", "acidez_volatil", "acido_citrico", "azucar_residual", "clo
ruros","diox_azufre_libre","diox_azufre_total","densidad","pH","sulfatos", "alcohol",
"calidad", "tipo")

colnames(blanco) <- nomCols
colnames(tinto) <- nomCols

#str(blanco)
summary(blanco)
```

```
##   acidez_fija     acidez_volatil   acido_citrico    azucar_residual
##  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
##     cloruros       diox_azufre_libre diox_azufre_total    densidad
##  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           sulfatos         alcohol         calidad
##  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
##      tipo
##  Length:4898
##  Class :character
##  Mode  :character
##
##
##
```

```
#str(tinto)
summary(tinto)
```

```
##   acidez_fija    acidez_volatil   acido_citrico   azucar_residual
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##     cloruros       diox_azufre_libre diox_azufre_total    densidad
##  Min.   :0.01200   Min.   : 1.00     Min.   :  6.00    Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00     1st Qu.: 22.00    1st Qu.:0.9956
##  Median :0.07900   Median :14.00     Median : 38.00    Median :0.9968
##  Mean   :0.08747   Mean   :15.87     Mean   : 46.47    Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00     3rd Qu.: 62.00    3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00     Max.   :289.00    Max.   :1.0037
##        pH           sulfatos         alcohol         calidad
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
##      tipo
##  Length:1599
##  Class :character
##  Mode  :character
##
##
##
```

Ahora unimos ambos datasets

```
# Unimos los dos juetos de datos en uno solo
totalData <- bind_rows(blanco,tinto)
filas=dim(totalData)[1]

# Factorizamos la variable tipo
totalData$tipo <- as.factor(totalData$tipo)

str(totalData)
```

```
## 'data.frame':    6497 obs. of  13 variables:
##  $ acidez_fija      : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ acidez_volatil   : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ acido_citrico    : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ azucar_residual  : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ cloruros         : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.04
4 ...
##  $ diox_azufre_libre: num  45 14 30 47 47 30 30 45 14 28 ...
##  $ diox_azufre_total: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ densidad         : 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 ...
##  $ sulfatos         : 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 ...
##  $ calidad          : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ tipo             : Factor w/ 2 levels "B","T": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(totalData)
```

```
##   acidez_fija     acidez_volatil   acido_citrico    azucar_residual
##  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
##     cloruros       diox_azufre_libre diox_azufre_total    densidad
##  Min.   :0.00900   Min.   :  1.00    Min.   :  6.0     Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00    1st Qu.: 77.0     1st Qu.:0.9923
##  Median :0.04700   Median : 29.00    Median :118.0     Median :0.9949
##  Mean   :0.05603   Mean   : 30.53    Mean   :115.7     Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00    3rd Qu.:156.0     3rd Qu.:0.9970
##  Max.   :0.61100   Max.   :289.00    Max.   :440.0     Max.   :1.0390
##        pH           sulfatos         alcohol         calidad      tipo
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000   B:4898
##  1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   T:1599
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

2.1 Nulos y/o elementos vacíos

Comprobamos que no haya valores vacíos o nulos.

```
# Estadísticas de valores vacios
colSums(is.na(totalData))
```

```
##       acidez_fija    acidez_volatil     acido_citrico   azucar_residual
##                 0                 0                 0                 0
##          cloruros diox_azufre_libre diox_azufre_total          densidad
##                 0                 0                 0                 0
##                pH          sulfatos           alcohol           calidad
##                 0                 0                 0                 0
##             tipo
##                 0
```

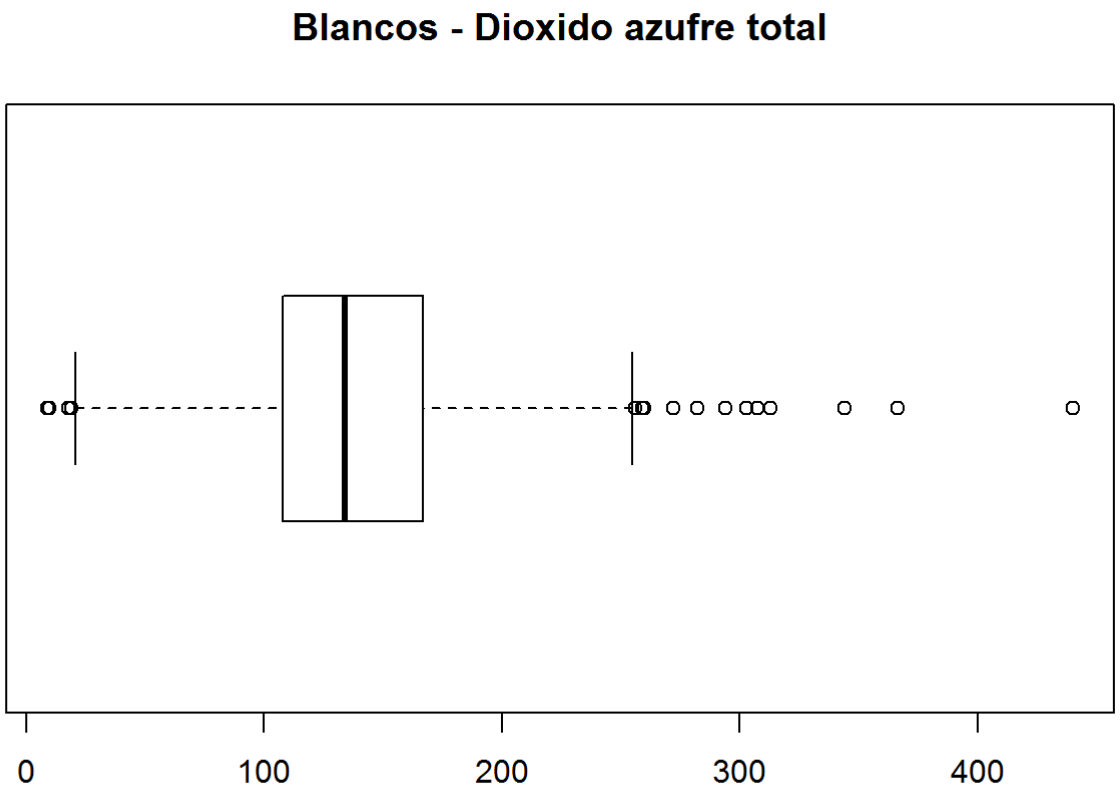
```
colSums(totalData=="")
```

```
##       acidez_fija    acidez_volatil     acido_citrico   azucar_residual
##                 0                 0                 0                 0
##          cloruros diox_azufre_libre diox_azufre_total          densidad
##                 0                 0                 0                 0
##                pH          sulfatos           alcohol           calidad
##                 0                 0                 0                 0
##             tipo
##                 0
```

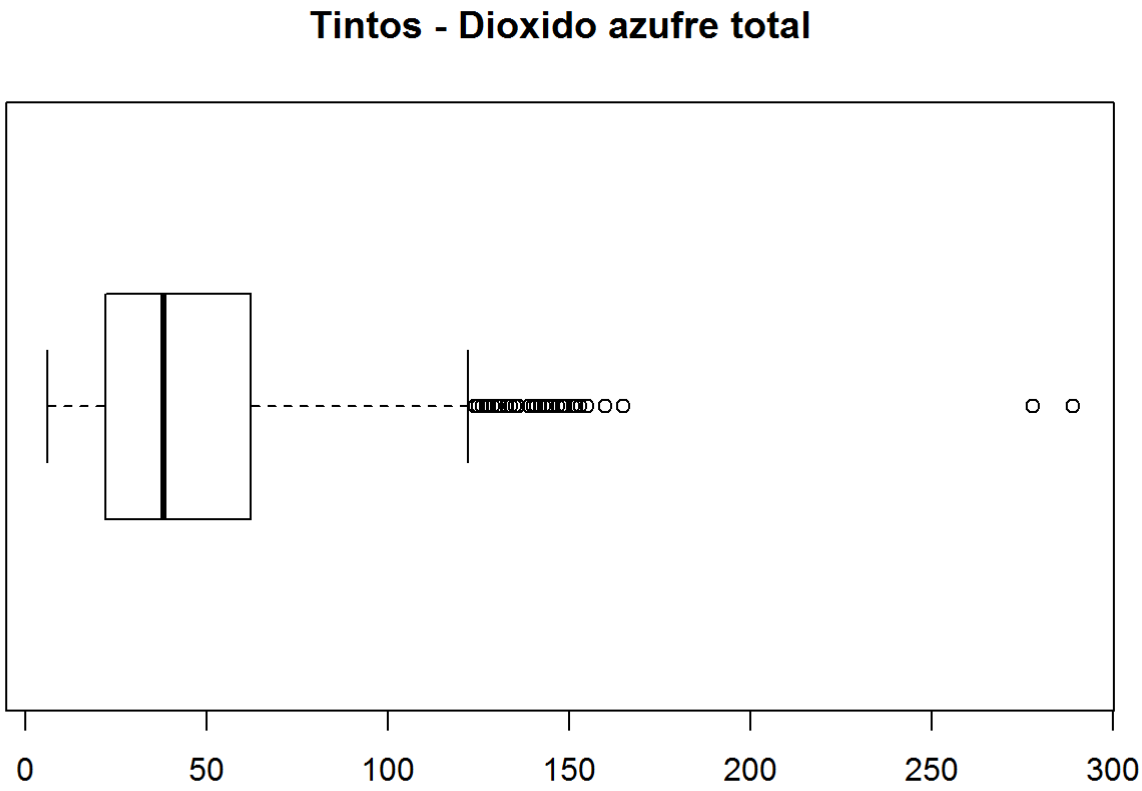
2.2 Valores extremos

En el resumen descriptivo pudimos observar tanto en el grupo vinos tintos como en el de blancos, los valores máximos de dioxido de azufre total parecen muy distantes de sus medidas de tencencia central. Vamos a identificarlos de manera gráfica con un diagrama box plot.

```
# comprobamos outliers en dioxido de azufre total en los blancos
#ggplot(totalData, aes(x=tipo, y=diox_azufre_total)) +  geom_point(size=2, shape=23)
datos.bp <-boxplot(blanco$diox_azufre_total, main="Blancos - Dioxido azufre total", ho
rizontal = T)
```




```
# comprobamos outliers en dioxido de azufre total en los tintos
datos.bp <-boxplot(tinto$diox_azufre_total, main="Tintos - Dioxido azufre total", hori
zontal = T)
```




```
boxplot.stats(blanco$diox_azufre_total)$out
```

```
##  [1] 272.0 313.0 260.0  19.0 366.5 307.5 256.0 256.0 344.0 282.0 303.0
## [12] 272.0  18.0  18.0 294.0   9.0  10.0 259.0 440.0
```

```
boxplot.stats(tinto$diox_azufre_total)$out
```

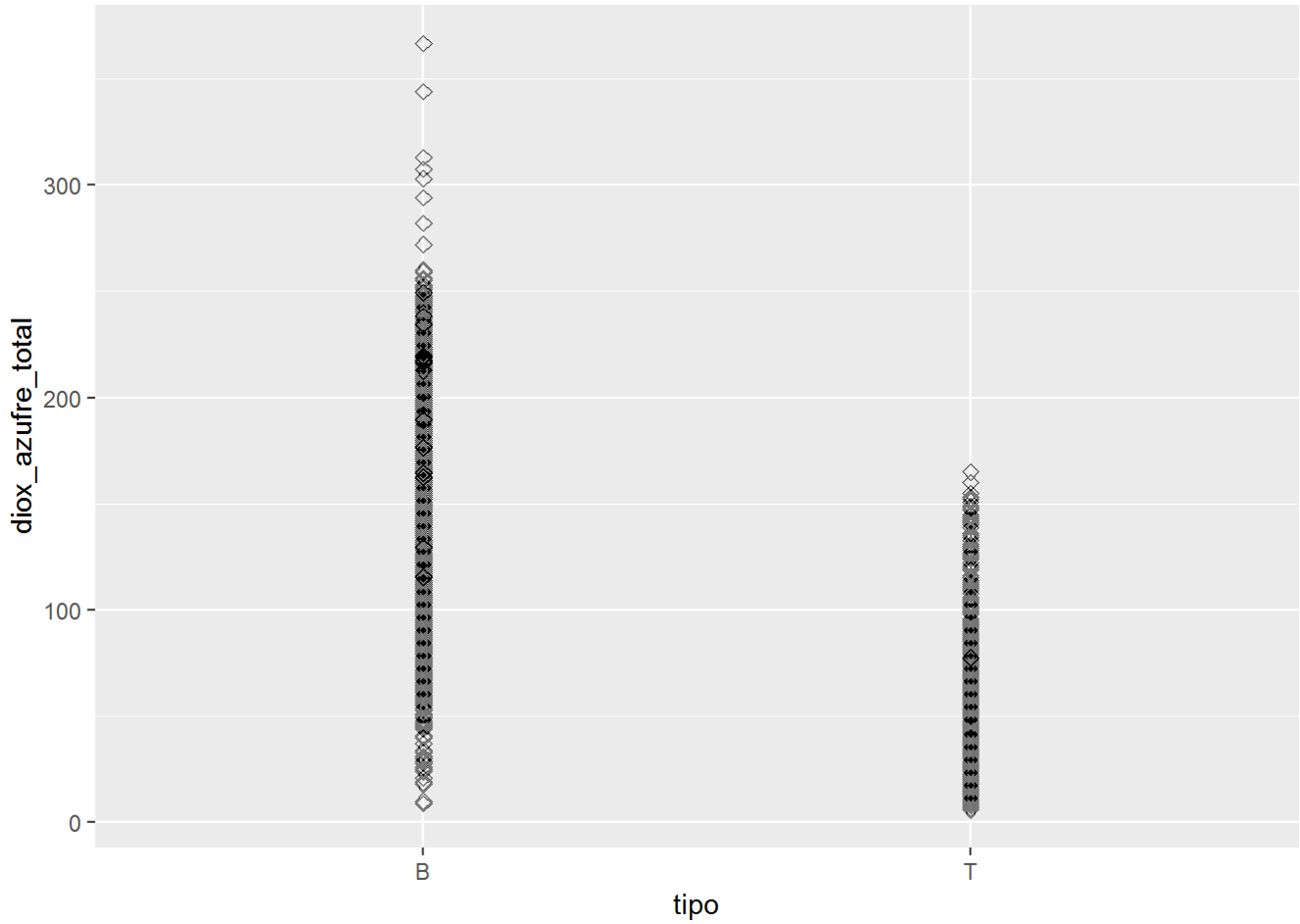
```
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127
## [18] 126 145 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148
## [35] 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147
## [52] 147 131 131 131
```

Vemos que el sistema detecta 20 valores atípicos en los vinos blancos y 56 en los tintos. No tenemos un conocimiento suficiente para valorar si se han producido por errores o por diferentes metodologías de medición o si por el contrario, son valores correctos por lo que vamos solamente vamos a sacar de la muestra los que están más alejados del rango intercuartílico.

De la muestra total sacamos el que tiene un valor > 400 y es blanco y de la de tintos los dos que tienen un valor superior a 250.

```
blanco <-subset(blanco, diox_azufre_total<400)
tinto <- subset(tinto, diox_azufre_total<250)
totalData <- bind_rows(blanco,tinto)
filas=dim(totalData)[1]
# Factorizamos la variable tipo
totalData$tipo <- as.factor(totalData$tipo)


#©totalData <- subset(totalData, (tipo=="B"& diox_azufre_total < 400) | (tipo == "T" & diox_azufre_total < 250))
ggplot(totalData, aes(x=tipo, y=diox_azufre_total)) +  geom_point(size=2, shape=23)
```



3 Análisis de los datos

A continuación vamos a realizar un análisis descriptivo de la variable calidad.

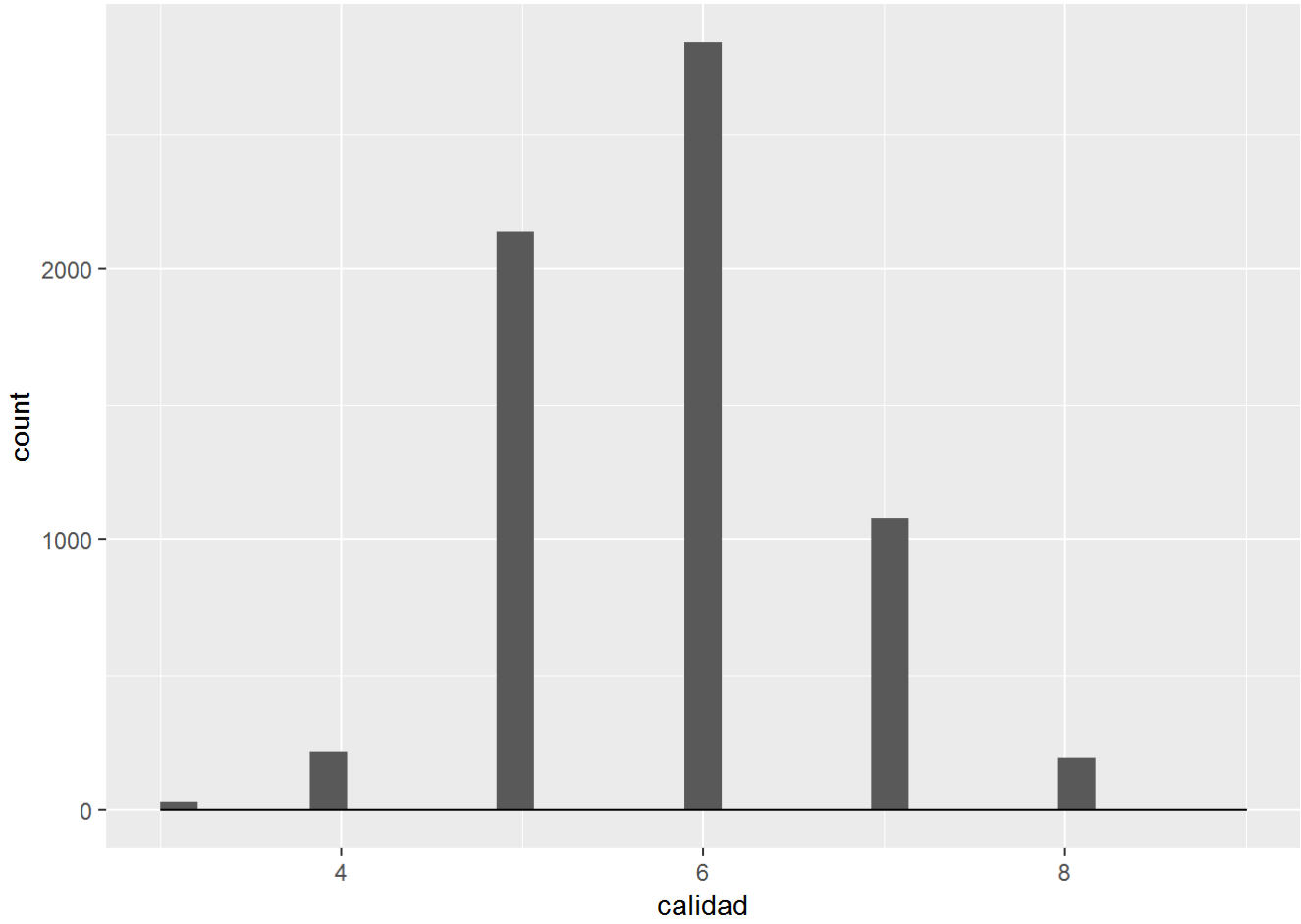
```
summary(totalData$calidad)
```

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

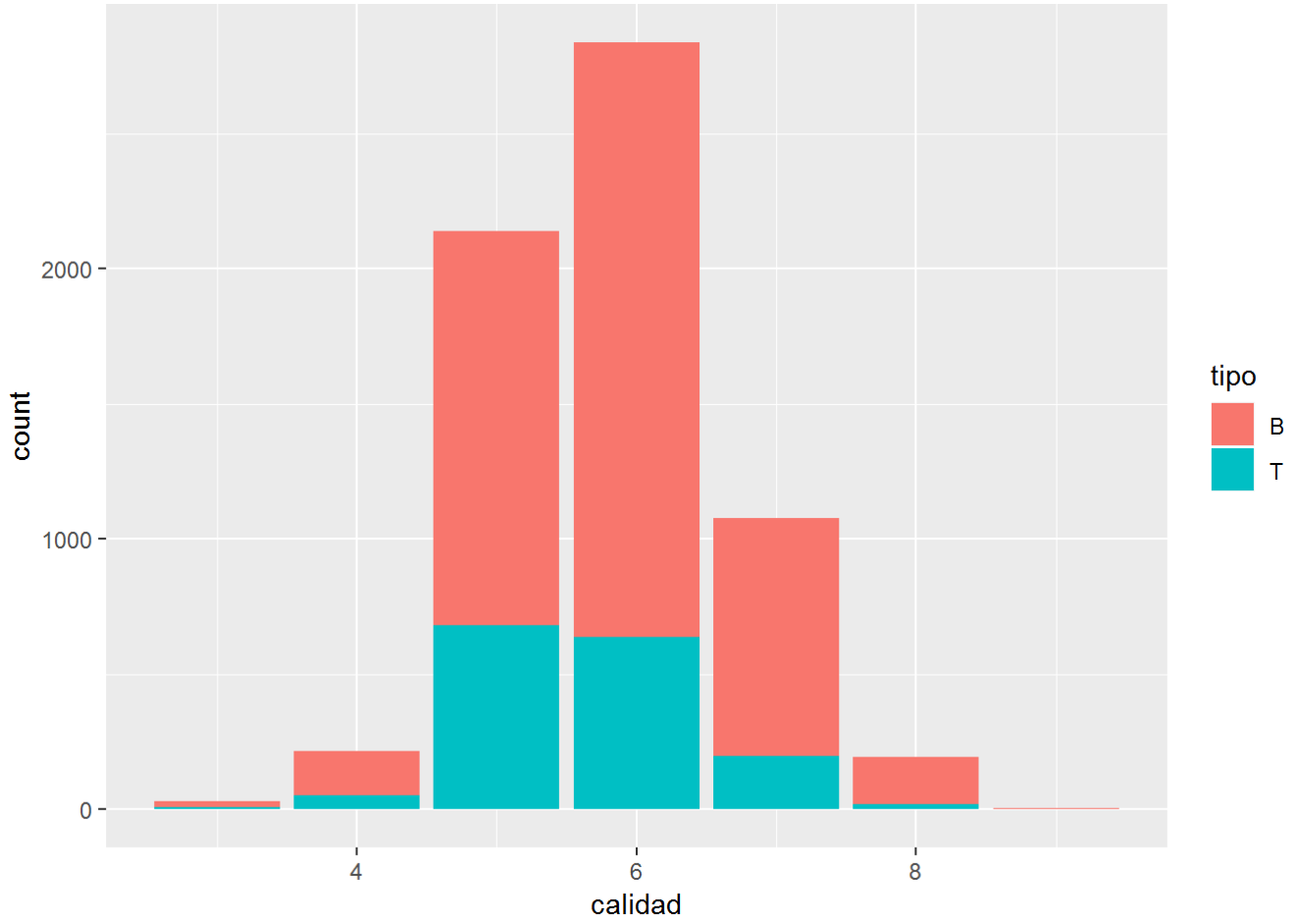
```
#desviacion estandara
sd(totalData$calidad)
```

```
## [1] 0.87251
```

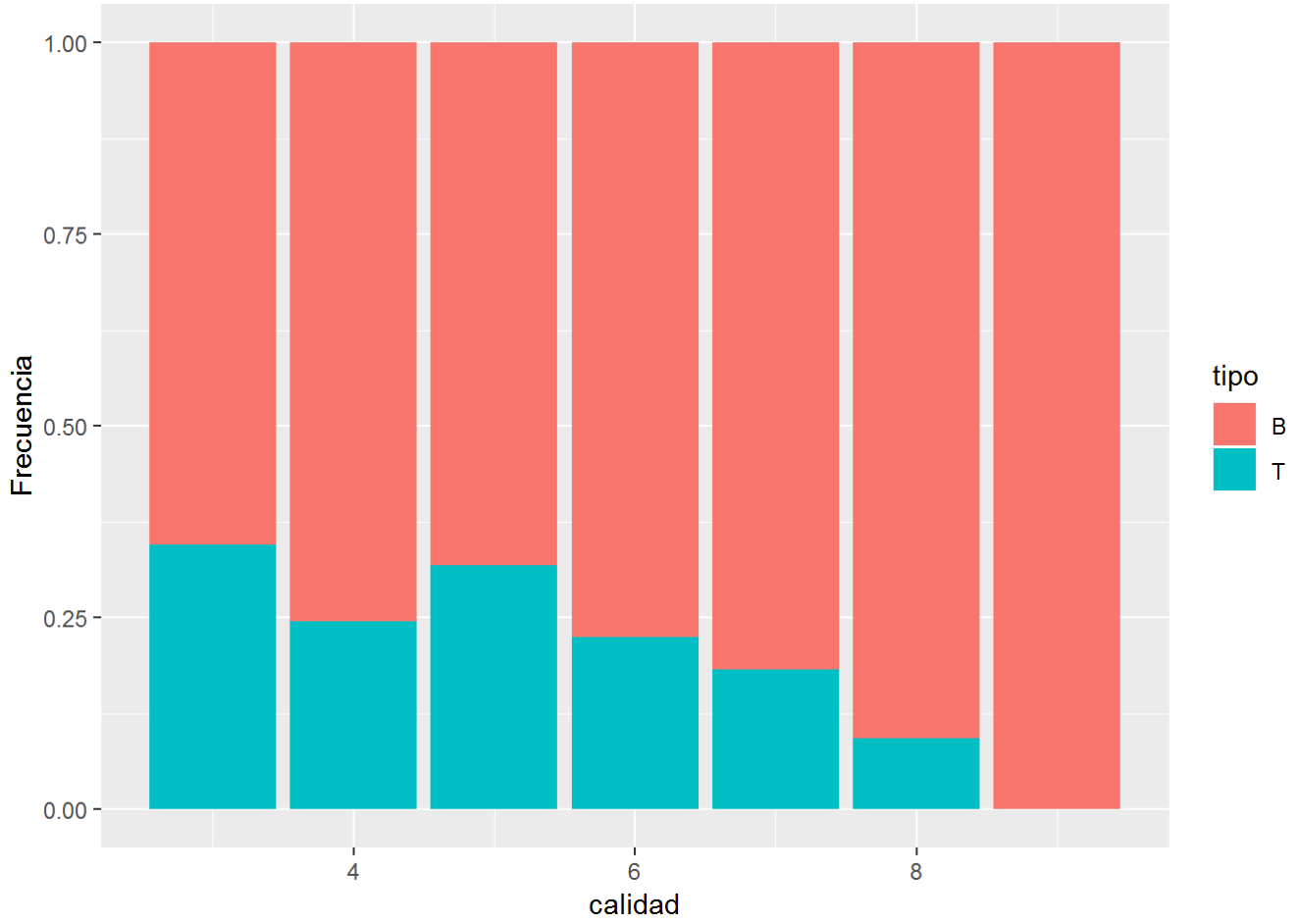
```
# mostramos un histograma de la calidad
ggplot(data = totalData[1:filas,],aes(x=calidad))+geom_histogram()+ geom_density(alpha
=.2, fill="#FF6666")
```




```
# Relacion entre calidad y tipo de vino
ggplot(data=totalData[1:filas,],aes(x=calidad,fill=tipo))+geom_bar()
```




```
# Grafico de frecuencias
ggplot(data = totalData[1:filas,],aes(x=calidad,fill=tipo))+geom_bar(position="fill")+
ylab("Frecuencia")
```



Se puede deducir de los gráficos que los vinos blancos de la muestra tienen más calidad que los tintos.

3.1 Análisis de la normalidad y homogeneidad de la varianza

Vamos a comprobar la normalidad de la calidad en ambos grupos de vinos. Utilizaremos los tests de **Kolmogorov-Smirnov** y **Shapiro-Wilk**

```
##
ks.test(tinto$calidad, pnorm, mean(tinto$calidad), sd(tinto$calidad))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  tinto$calidad
## D = 0.25005, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
shapiro.test(tinto$calidad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tinto$calidad
## W = 0.85728, p-value < 2.2e-16
```

```
ks.test(blanco$calidad, pnorm, mean(blanco$calidad), sd(blanco$calidad))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  blanco$calidad
## D = 0.22893, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
shapiro.test(blanco$calidad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  blanco$calidad
## W = 0.88887, p-value < 2.2e-16
```

En ambos test se rechaza la hipótesis nula, por tanto consideramos que la calidad no se distribuye mediante una distribución normal en ninguno de los dos grupos. No obstante, por el **teorema central del límite** se podría considerar que los datos siguen una distribución normal.

Analizaremos la homocedasticidad de la varianza mediante el **test de Flinger-Killen** en cuanto a los grupos conformados por los vinos tintos y los blancos.

```
##
b <- blanco$calidad
t <- tinto$calidad
fligner.test(calidad ~ tipo, data= totalData)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  calidad by tipo
## Fligner-Killeen:med chi-squared = 0.6346, df = 1, p-value = 0.4257
```

Dado que el p-valor es > 0.05 podemos aceptar la hipóstesis nula de que las varianzas de ambas muestras son homogéneas.

4 Pruebas estadísticas

4.1 ¿Que tipo de vino tiene más calidad?

En los histogramas y gráficos de frecuencias pudimos observar que la calidad de los vinos blancos de la muestra era más alta que la de los tintos, vamos a realizar un contraste de hipótesis para comprobar si tenemos diferencias estadísticamente significativas en la media de la calidad de ambos grupos de vinos.

Considerando el análisis de la normalidad y homogeneidad de la varianza del punto anterior, aplicaremos la prueba **t de Student** formulando las siguientes hipótesis:

```
H0: nuB - nuT = 0
H1: nuB - nuT > 0
```

donde nuB es la media muestral de la calidad de los vinos blancos y nuT es la media muestral de la calidad de los vinos tintos.

```
## Realizamos el test por tipo de vino
t.test(calidad ~ tipo, data = totalData, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  calidad by tipo
## t = 10.252, df = 2946.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2049939       Inf
## sample estimates:
## mean in group B mean in group T
##        5.878497        5.634314
```

Dado que el p-valor es inferior al nivel de significancia (0.05), debemos rechazar la hipótesis nula, por tanto podemos concluir que efectivamente, la calidad de los vinos blancos es superior que la de los vinos tintos de la muestra.

4.2 ¿Qué prueba fisioquímica es más determinante para la calidad de un vino?

Vamos a calcular la matriz de correlaciones de las variables cuantitativas de cada grupo de vinos.

```
round(cor(blanco[,-13]),2)
```

```
##                   acidez_fija acidez_volatil acido_citrico azucar_residual
## acidez_fija              1.00          -0.02          0.29            0.09
## acidez_volatil          -0.02           1.00         -0.15            0.06
## acido_citrico            0.29          -0.15          1.00            0.09
## azucar_residual          0.09           0.06          0.09            1.00
## cloruros                 0.02           0.07          0.11            0.09
## diox_azufre_libre       -0.05          -0.10          0.10            0.31
## diox_azufre_total        0.09           0.09          0.12            0.40
## densidad                 0.27           0.03          0.15            0.84
## pH                      -0.43          -0.03         -0.16           -0.19
## sulfatos                -0.02          -0.04          0.06           -0.03
## alcohol                 -0.12           0.07         -0.08           -0.45
## calidad                 -0.11          -0.20         -0.01           -0.10
##                   cloruros diox_azufre_libre diox_azufre_total densidad
## acidez_fija           0.02             -0.05              0.09     0.27
## acidez_volatil        0.07             -0.10              0.09     0.03
## acido_citrico         0.11              0.10              0.12     0.15
## azucar_residual       0.09              0.31              0.40     0.84
## cloruros              1.00              0.10              0.20     0.26
## diox_azufre_libre     0.10              1.00              0.61     0.30
## diox_azufre_total     0.20              0.61              1.00     0.53
## densidad              0.26              0.30              0.53     1.00
## pH                   -0.09             -0.01              0.00    -0.09
## sulfatos              0.02              0.06              0.13     0.07
## alcohol              -0.36             -0.26             -0.45    -0.78
## calidad              -0.21              0.02             -0.17    -0.31
##                      pH sulfatos alcohol calidad
## acidez_fija       -0.43    -0.02   -0.12   -0.11
## acidez_volatil    -0.03    -0.04    0.07   -0.20
## acido_citrico     -0.16     0.06   -0.08   -0.01
## azucar_residual   -0.19    -0.03   -0.45   -0.10
## cloruros          -0.09     0.02   -0.36   -0.21
## diox_azufre_libre -0.01     0.06   -0.26    0.02
## diox_azufre_total  0.00     0.13   -0.45   -0.17
## densidad          -0.09     0.07   -0.78   -0.31
## pH                 1.00     0.16    0.12    0.10
## sulfatos           0.16     1.00   -0.02    0.05
## alcohol            0.12    -0.02    1.00    0.44
## calidad            0.10     0.05    0.44    1.00
```

Observando la matriz de correlaciones vemos que las variables dioxido de azufre libre y ácido cítrico, no tienen prácticamente ninguna correlación con la calidad, podríamos sacarlas del modelo. Por el contrario, el alcohol y la densidad son las variables que más correlación tienen con la calidad, positiva y negativa respectivamente, aunque la correlación es más bien baja.

Probamos la significancia de la correlación entre la calidad y el alcohol:

```
Hipótesis nula H0: no hay relación
Hipótesis alternativa H1: hay relación.
```

```
cor.test(blanco$alcohol, blanco$calidad, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  blanco$alcohol and blanco$calidad
## t = 33.899, df = 4895, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4130731 0.4584477
## sample estimates:
##       cor
## 0.4360375
```

Comprobamos que el test nos arroja un p-value inferior a 0.05 por lo que rechazamos la hipótesis nula.

Comprobamos que para la densidad también rechazamos la hipótesis nula.

```
cor.test(blanco$densidad, blanco$calidad, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  blanco$densidad and blanco$calidad
## t = -22.622, df = 4895, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3327965 -0.2820763
## sample estimates:
##        cor
## -0.3076549
```

Obtenemos la matriz de correlaciones para el grupo de vinos tinto

```
round(cor(tinto[,-13]),2)
```

```
##                    acidez_fija acidez_volatil acido_citrico azucar_residual
## acidez_fija               1.00          -0.26          0.67            0.12
## acidez_volatil           -0.26           1.00         -0.55            0.01
## acido_citrico             0.67          -0.55          1.00            0.13
## azucar_residual           0.12           0.01          0.13            1.00
## cloruros                  0.09           0.06          0.21            0.06
## diox_azufre_libre        -0.15          -0.01         -0.07            0.18
## diox_azufre_total        -0.11           0.09          0.02            0.17
## densidad                  0.67           0.02          0.37            0.37
## pH                       -0.69           0.23         -0.54           -0.08
## sulfatos                  0.18          -0.26          0.32            0.01
## alcohol                  -0.06          -0.20          0.11            0.03
## calidad                   0.12          -0.39          0.22            0.01
##                    cloruros diox_azufre_libre diox_azufre_total densidad
## acidez_fija            0.09             -0.15             -0.11     0.67
## acidez_volatil         0.06             -0.01              0.09     0.02
## acido_citrico          0.21             -0.07              0.02     0.37
## azucar_residual        0.06              0.18              0.17     0.37
## cloruros               1.00              0.01              0.06     0.20
## diox_azufre_libre      0.01              1.00              0.67    -0.02
## diox_azufre_total      0.06              0.67              1.00     0.09
## densidad               0.20             -0.02              0.09     1.00
## pH                    -0.27              0.08             -0.05    -0.35
## sulfatos               0.37              0.05              0.05     0.15
## alcohol               -0.22             -0.07             -0.23    -0.49
## calidad               -0.13             -0.06             -0.21    -0.17
##                          pH sulfatos alcohol calidad
## acidez_fija           -0.69     0.18   -0.06    0.12
## acidez_volatil         0.23    -0.26   -0.20   -0.39
## acido_citrico         -0.54     0.32    0.11    0.22
## azucar_residual       -0.08     0.01    0.03    0.01
## cloruros              -0.27     0.37   -0.22   -0.13
## diox_azufre_libre      0.08     0.05   -0.07   -0.06
## diox_azufre_total     -0.05     0.05   -0.23   -0.21
## densidad              -0.35     0.15   -0.49   -0.17
## pH                     1.00    -0.20    0.21   -0.05
## sulfatos              -0.20     1.00    0.10    0.25
## alcohol                0.21     0.10    1.00    0.47
## calidad               -0.05     0.25    0.47    1.00
```

Tampoco obtenemos unas correlaciones altas de la calidad con el resto de variables, por lo que consideramos que un modelo de regresión lineal no va a ser de mucha utilidad para predecir la calidad de los vinos.

4.3 Regresion lineal

Partiendo del dataset de vinos blancos, vamos a hacer un análisis de regresión para estimar la calidad del vino. Nos quedamos solamente con el dataset de vinos blancos y le quitamos la variable de tipo.

```
blancoQ <- blanco[,1:12]
str(blancoQ)
```

```
## 'data.frame':    4897 obs. of  12 variables:
##  $ acidez_fija      : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ acidez_volatil   : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ acido_citrico    : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ azucar_residual  : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ cloruros         : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.04
4 ...
##  $ diox_azufre_libre: num  45 14 30 47 47 30 30 45 14 28 ...
##  $ diox_azufre_total: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ densidad         : 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 ...
##  $ sulfatos         : 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 ...
##  $ calidad          : int  6 6 6 6 6 6 6 6 6 6 ...
```

Vamos a dividir las observaciones en dos grupos, uno de entrenamiento para ajustar el modelo (2/3 de los datos) y uno de test (1/3 de los datos)

```
library(rminer)
set.seed(123)
h <- holdout(blancoQ$calidad,ratio=2/3,mode="stratified")
training <- blancoQ[h$tr,]
test <- blancoQ[h$ts,]
str(training)
```

```
## 'data.frame':    3264 obs. of  12 variables:
##  $ acidez_fija      : num  7.3 7.3 6.7 6.3 6.9 6.2 6.8 7.1 5.7 6.9 ...
##  $ acidez_volatil   : num  0.25 0.25 0.31 0.28 0.23 0.25 0.18 0.31 0.23 0.25 ...
##  $ acido_citrico    : num  0.36 0.26 0.08 0.22 0.35 0.25 0.3 0.25 0.28 0.35 ...
##  $ azucar_residual  : num  13.1 7.2 1.3 11.5 6.9 1.4 12.8 11.2 9.65 9.2 ...
##  $ cloruros         : num  0.05 0.048 0.038 0.036 0.03 0.03 0.062 0.048 0.025 0.034
...
##  $ diox_azufre_libre: num  35 52 58 27 45 35 19 32 26 42 ...
##  $ diox_azufre_total: num  200 207 147 150 116 105 171 136 121 150 ...
##  $ densidad         : num  0.999 0.996 0.992 0.994 0.992 ...
##  $ pH               : num  3.04 3.12 3.18 3 2.8 3.3 3 3.14 3.28 3.21 ...
##  $ sulfatos         : num  0.46 0.37 0.46 0.33 0.54 0.44 0.52 0.4 0.38 0.36 ...
##  $ alcohol          : num  8.9 9.2 10 10.6 11 11.1 9 9.5 11.3 11.5 ...
##  $ calidad          : int  7 5 5 6 6 7 7 5 6 6 ...
```

```
str(test)
```

```
## 'data.frame':    1633 obs. of  12 variables:
##  $ acidez_fija      : num  7 6.3 8.1 6.2 6.3 8.6 6.6 6.4 6.8 6.9 ...
##  $ acidez_volatil   : num  0.27 0.3 0.28 0.32 0.3 0.23 0.16 0.31 0.26 0.24 ...
##  $ acido_citrico    : num  0.36 0.34 0.4 0.16 0.34 0.4 0.4 0.38 0.42 0.35 ...
##  $ azucar_residual  : num  20.7 1.6 6.9 7 1.6 4.2 1.5 2.9 1.7 1 ...
##  $ cloruros         : num  0.045 0.049 0.05 0.045 0.049 0.035 0.044 0.038 0.049 0.0
52 ...
##  $ diox_azufre_libre: num  45 14 30 30 14 17 48 19 41 35 ...
##  $ diox_azufre_total: num  170 132 97 136 132 109 143 102 122 146 ...
##  $ densidad         : num  1.001 0.994 0.995 0.995 0.994 ...
##  $ pH               : num  3 3.3 3.26 3.18 3.3 3.14 3.54 3.17 3.47 3.45 ...
##  $ sulfatos         : num  0.45 0.49 0.44 0.47 0.49 0.53 0.52 0.35 0.48 0.44 ...
##  $ alcohol          : num  8.8 9.5 10.1 9.6 9.5 9.7 12.4 11 10.5 10 ...
##  $ calidad          : int  6 6 6 6 6 5 7 7 8 6 ...
```

```
#training <- sample_frac(blancoQ, .7)
#test <- setdiff(blancoQ, training)

modelo <- lm(calidad ~ ., data = training)
summary(modelo)
```

```
##
## Call:
## lm(formula = calidad ~ ., data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3366 -0.5014 -0.0421  0.4546  3.0711
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.311e+02  2.117e+01   6.193 6.65e-10 ***
## acidez_fija         7.542e-02  2.449e-02   3.079  0.00209 **
## acidez_volatil     -1.778e+00  1.392e-01 -12.768  < 2e-16 ***
## acido_citrico       3.377e-02  1.173e-01   0.288  0.77337
## azucar_residual     7.300e-02  8.709e-03   8.382  < 2e-16 ***
## cloruros           -9.228e-01  6.514e-01  -1.417  0.15670
## diox_azufre_libre   5.383e-03  1.050e-03   5.127 3.11e-07 ***
## diox_azufre_total -1.134e-04  4.583e-04  -0.248  0.80453
## densidad           -1.313e+02  2.147e+01  -6.115 1.08e-09 ***
## pH                  6.802e-01  1.251e-01   5.438 5.79e-08 ***
## sulfatos            5.499e-01  1.210e-01   4.544 5.73e-06 ***
## alcohol             2.169e-01  2.766e-02   7.841 6.00e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7486 on 3252 degrees of freedom
## Multiple R-squared:  0.2821, Adjusted R-squared:  0.2797
## F-statistic: 116.2 on 11 and 3252 DF,  p-value: < 2.2e-16
```

Vemos que el valor R2 ajustado es bajo, 0.2797 por lo que el modelo no es capaz de predecir con precisión la calidad.

Vamos a verificarlo calculando la MSE

```
# funcion que calcula la media de los cuadrados de las desviaciones
dm <- function(actual, predicted){
  mean((actual - predicted)^2)
}

# MSE empleando las observaciones de entrenamiento
training_mse <- dm(modelo$fitted.values, training$calidad)
training_mse
```

```
## [1] 0.5582977
```

```
# MSE empleando nuevas observaciones
predicciones <- predict(modelo, newdata = test)
test_mse <- dm(predicciones, test$calidad)
test_mse
```

```
## [1] 0.5666391
```

4.4 Modelo supervisado

4.4.1 Clasificación. Random forest

A continuación vamos a aplicar un método de clasificación random forest mediante una validación cruzada con 4 folds para clasificar los vinos en tintos o blancos.

```
library(caret)


h <- holdout(totalData$tipo,ratio=2/3,mode="stratified")
vino_entrenamiento <- totalData[h$tr,]
vino_prueba <- totalData[h$ts,]

train_control <-trainControl(method = "cv", number = 4)
mod<-train(tipo~., data=vino_entrenamiento, method="rf",trControl=train_control)
pred <- predict(mod, newdata=vino_prueba)
```

Obtenemos la matriz de confusión para comprobar la bondad del modelo.

```
confusionMatrix(pred,vino_prueba$tipo, positive="T")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    B    T
##          B 1627    7
##          T    5  525
##
##                Accuracy : 0.9945
##                  95% CI : (0.9903, 0.9971)
##     No Information Rate : 0.7542
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.985
##
##  Mcnemar's Test P-Value : 0.7728
##
##             Sensitivity : 0.9868
##             Specificity : 0.9969
##          Pos Pred Value : 0.9906
##          Neg Pred Value : 0.9957
##              Prevalence : 0.2458
##          Detection Rate : 0.2426
##    Detection Prevalence : 0.2449
##       Balanced Accuracy : 0.9919
##
##        'Positive' Class : T
##
```

Vemos que el resultado es excelente, el modelo nos clasifica los vinos con una precisión del 99.45% con un índice **kappa=0.985** que nos indica que nuestra clasificación es un 98.5% mejor que una clasificación aleatoria.

5 Conclusiones