

TIPOLOGÍA Y CICLO DEL DATO.

Limpieza y Análisis de los datos

PRÁCTICA 2

Autor: Juan Ramón Tonda Barberá

1.- Descripción del dataset.

El conjunto de datos a utilizar en la práctica se denomina “winequality-red.csv” y en él se recoge un total de 1.599 muestras de vino tinto, cada una de ellas con un total de 11 variables que se corresponden con la composición química del vino, así como una variable que determina la calidad del vino, siendo 3 el vino con peor calidad y 8 el vino con mayor calidad.

Las variables son las siguientes:

Componente	Descripción del componente
fixed.acidity	Es el resto de ácidos (excepto el acético)
volatile.acidity	Es fundamentalmente el ácido acético. El olor a vinagre de un vino se debe al ácido acético.
citric.acid	Encontrado en pequeñas cantidades, puede agregar frescura y sabor a los vinos.
residual.sugar	Es la cantidad de azúcar restante una vez se detiene la fermentación
chlorides	La cantidad de sal en el vino
free.sulfur.dioxide	Evita la oxidación y suprime la actividad de las enzimas que causan oscurecimiento y otros problemas en el vino.
total.sulfur.dioxide	
density	Es la densidad del agua dependiendo del porcentaje de contenido de alcohol y azúcar
pH	Describe cuán de ácido o básico es un vino 0, muy ácido y 14 muy básico.
sulphates	Presente de forma natural o artificial e incide en la conservación del vino y su calidad. Tienen función antioxidante, antioxidásico y antimicrobiano.
alcohol	El grado de alcohol es uno de los factores más importantes para conservar las propiedades de un vino.
quality	Determina la calidad del vino (3 peor calidad, 8 mejor calidad)

OBJETIVO DEL ANÁLISIS: A partir de estos datos, se va a realizar un análisis para establecer qué componentes químicos son los que determinan la calidad de un vino. Asimismo, otro de los objetivos será el de proponer un modelo que muestre la calidad de un determinado vino dado en base a sus componentes químicos.

El Dataset se carga en RStudio para el inicio del análisis:

```
#Carga de datos
wineData<-read.csv("C:/Users/Juan/OneDrive/MASTER EN CIENCIA DE DATOS/TIPOLOGÍA Y CICLO DEL DATO/PRACTICA2/winequality-red.csv",sep=";")
```

2.- Selección de los datos de interés a analizar.

La primera aproximación a los datos se realiza a partir de una visión de conjunto, en el que por un lado podemos observar el nombre de la variable y el tipo de dato y por otro unos estadísticos básicos para cada una de dichas variables.

```
1 glimpse(wineData)
2
```

```

Observations: 1,599
Variables: 12
 $ fixed.acidity      <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5...
 $ volatile.acidity  <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, ...
 $ citric.acid       <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0...
 $ residual.sugar    <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1,...
 $ chlorides         <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069, ...
 $ free.sulfur.dioxide <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, 16...
 $ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 102,...
 $ density           <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0...
 $ pH               <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3...
 $ sulphates         <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0...
 $ alcohol           <dbl> 9.4, 9.8, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10...
 $ quality           <int> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 7...

```

```
summary(wineData)
```

```

fixed.acidity    volatile.acidity    citric.acid
Min.   : 4.60      Min.   :0.1200      Min.   :0.000
1st Qu.: 7.10      1st Qu.:0.3900      1st Qu.:0.090
Median : 7.90      Median :0.5200      Median :0.260
Mean   : 8.32      Mean   :0.5278      Mean   :0.271
3rd Qu.: 9.20      3rd Qu.:0.6400      3rd Qu.:0.420
Max.   :15.90      Max.   :1.5800      Max.   :1.000
residual.sugar   chlorides          free.sulfur.dioxide
Min.   : 0.900     Min.   :0.01200     Min.   : 1.00
1st Qu.: 1.900     1st Qu.:0.07000     1st Qu.: 7.00
Median : 2.200     Median :0.07900     Median :14.00
Mean   : 2.539     Mean   :0.08747     Mean   :15.87
3rd Qu.: 2.600     3rd Qu.:0.09000     3rd Qu.:21.00
Max.   :15.500     Max.   :0.61100     Max.   :72.00
total.sulfur.dioxide density          pH
Min.   : 6.00      Min.   :0.9901      Min.   :2.740
1st Qu.: 22.00     1st Qu.:0.9956      1st Qu.:3.210
Median : 38.00     Median :0.9968      Median :3.310
Mean   : 46.47     Mean   :0.9967      Mean   :3.311
3rd Qu.: 62.00     3rd Qu.:0.9978      3rd Qu.:3.400
Max.   :289.00     Max.   :1.0037      Max.   :4.010
sulphates        alcohol          quality
Min.   :0.3300     Min.   : 8.40      Min.   :3.000
1st Qu.:0.5500     1st Qu.: 9.50      1st Qu.:5.000
Median :0.6200     Median :10.20      Median :6.000
Mean   :0.6581     Mean   :10.42      Mean   :5.636
3rd Qu.:0.7300     3rd Qu.:11.10      3rd Qu.:6.000
Max.   :2.0000     Max.   :14.90      Max.   :8.000

```

Realizada la visualización preliminar determinamos que el “dataset” contiene 1.599 registros y 12 atributos, de los cuales once pertenecen a los componentes químicos del vino y uno “quality” determina la calidad del vino. A excepción de la variable quality, que es de tipo entero, el resto de variables es de tipo “double”.

No se observan inicialmente valores nulos, y sí parece que existen valores atípicos, pero esto se analizará más adelante.

Por último, se quiere conocer cuántos tipos (registros) de vinos están asociados a una clase u otra y para ello se establece una tabla de contingencia para la variable “quality”.

Los valores son los siguientes:

```
table(wineData$quality)
```

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

3.- Limpieza de los datos.

3.1 Análisis de elementos vacíos y elementos con valor cero.

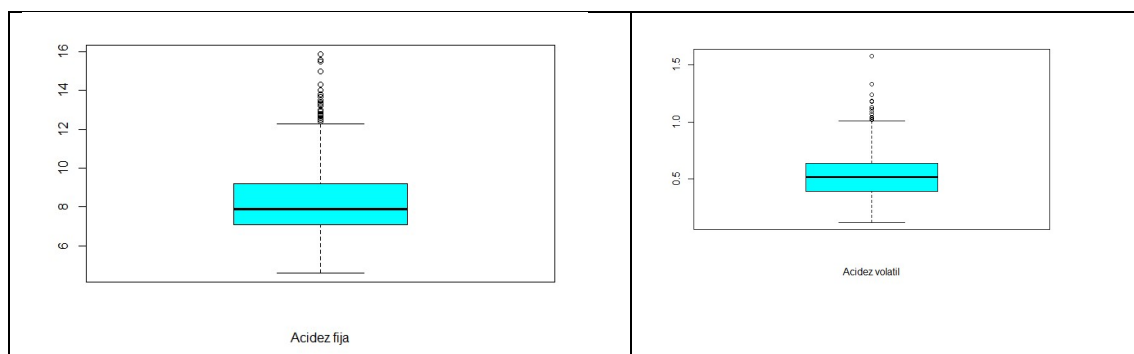
De análisis realizado en el apartado anterior, se observa que no existen valores nulos en todo el conjunto de datos. Por otra parte, aquellos valores cero que se localizan en la variable “citric.acid” se consideran como valores aceptables debido a que el valor cero podría ser ausencia de dato en lugar de un error.

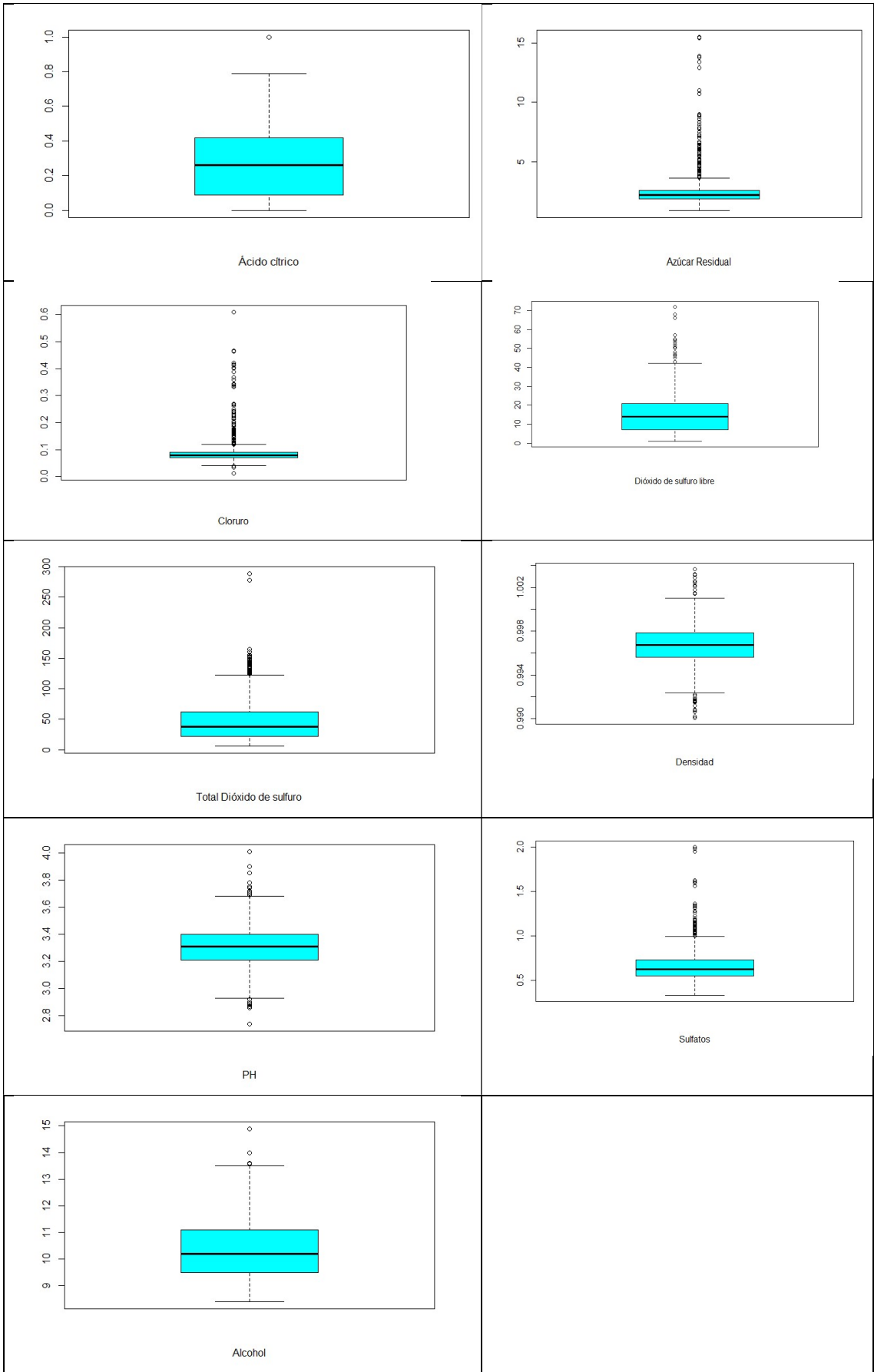
Por lo tanto no se requiere ningún tratamiento en este sentido.

3.2 Identificación y tratamiento de valores extremos.

Mediante la técnica de BoxPlot se va a analizar la existencia de valores extremos, y a continuación se decidirá qué se hace con ellos.

En un primer momento se han analizado todos los atributos del vino para localizar dichos valores atípicos.





Al no ser un experto en el análisis de vinos no puedo determinar si los valores atípicos son un error en la recogida de datos o valores excepcionales de cada uno de esos vinos. Podría inclinarme por realizar la sustitución de los valores atípicos por la media o la mediana del resto de valores, pero no tengo evidencia científica de que tenga que eliminarlos porque éstos superen un número determinado de anchos intercuartílicos, por lo que opto por mantener los valores originales.

De todos modos, incluyo el script que permitiría sustituir dichos valores atípicos por la media del resto de valores.

Eliminar los valores atípicos, pasaría por:

- Localizarlos:
 - `atipicos<-boxplot.stats(wineData$residual.sugar)&out`
- Encontrar la posición que ocupan en el repositorio de datos:
 - `Indice<-which(wineData$residual.sugar %in% valorAtipico)`
- Realizar la operación que se desee:
 - Eliminarlos:
 - `wineData<-wineData[-Indice,]`
 - Sustituirlos por ejemplo por la mediana.
 - `winData$residual.sugar<-
ifelse(wineData$residual.sugar>min(atipicos),median(wineData$
residual.sugar,wineData$residual.sugar)`

Existe otra manera de poder eliminar los atípicos y sustituirlos por la mediana o la media a través de una librería denominada “outlier”, del siguiente modo:

```
rm.outlier(wineData$residual.sugar, fill=FALSE, median=FALSE, opposite=FALSE)
```

En este caso, si fill es FALSE el outlier se eliminará, en el caso de TRUE, se sustituirá por la media o la mediana en función del booleano del parámetro median, si éste es FALSE se sustituirá por la media y si es true por la mediana.

De todos modos, estas dos técnicas para eliminar los outliers no se van a emplear en la práctica ya que he decidido conservar los outliers tal y como se ha comentado anteriormente

3.3 Conclusión

No existen valores nulos en el conjunto de datos, los valores cero del atributo “citric.acid” se consideran valores válidos así como todos los valores outliers de cada una de las variables.

4. Análisis de los datos

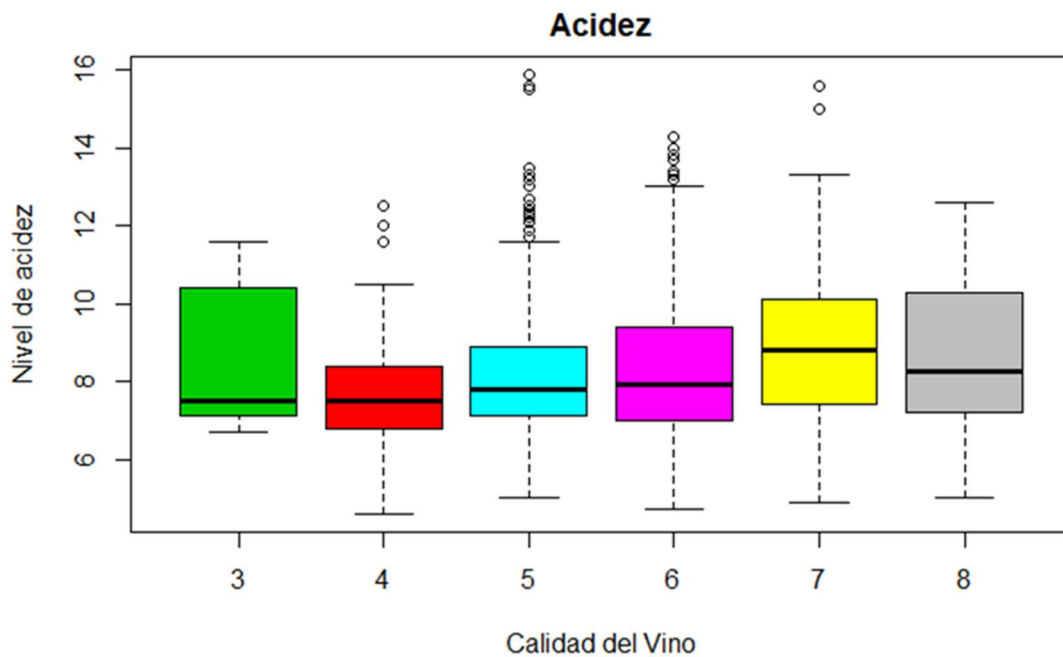
4.1 Selección de los grupos de datos que se quieren analizar/comparar

El análisis que se va a realizar a continuación del conjunto de datos del vino tiene como objetivo determinar qué grupos de datos se van a utilizar en el análisis. Las variables seleccionadas serán aquellas que determinan la calidad del vino.

Y para ello se va a volver a utilizar la técnica del boxplot para cada variable, pero esta vez se va a poner en relación con la calidad del vino y de esa manera poder determinar visualmente qué variable es la más significativa.

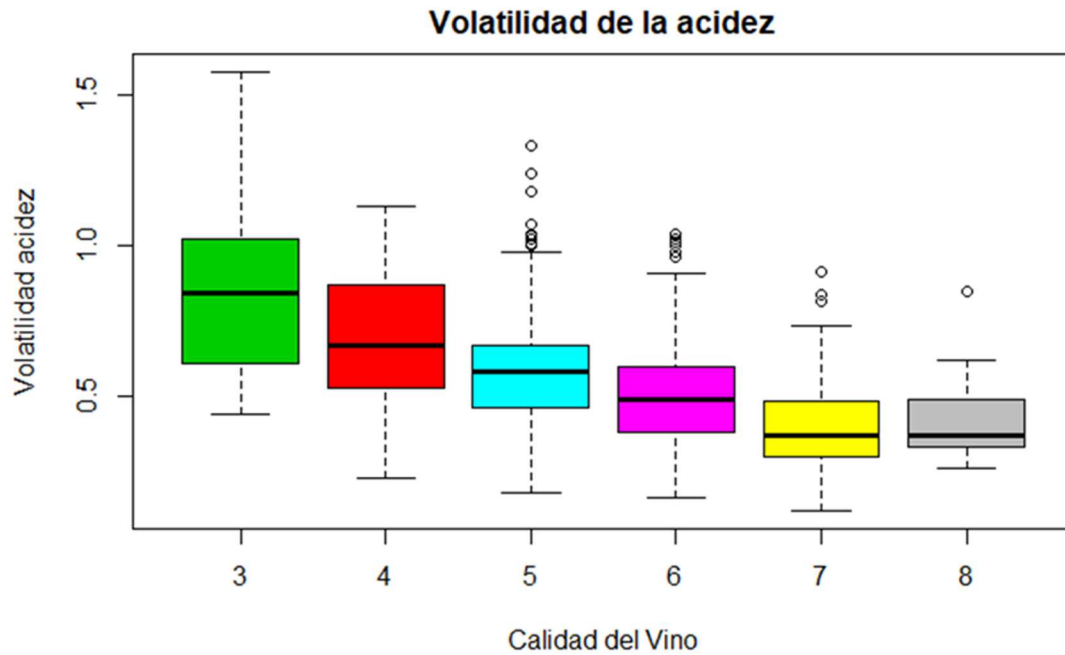
A continuación, se realiza dicho análisis:

a) Variable Fixed acidity



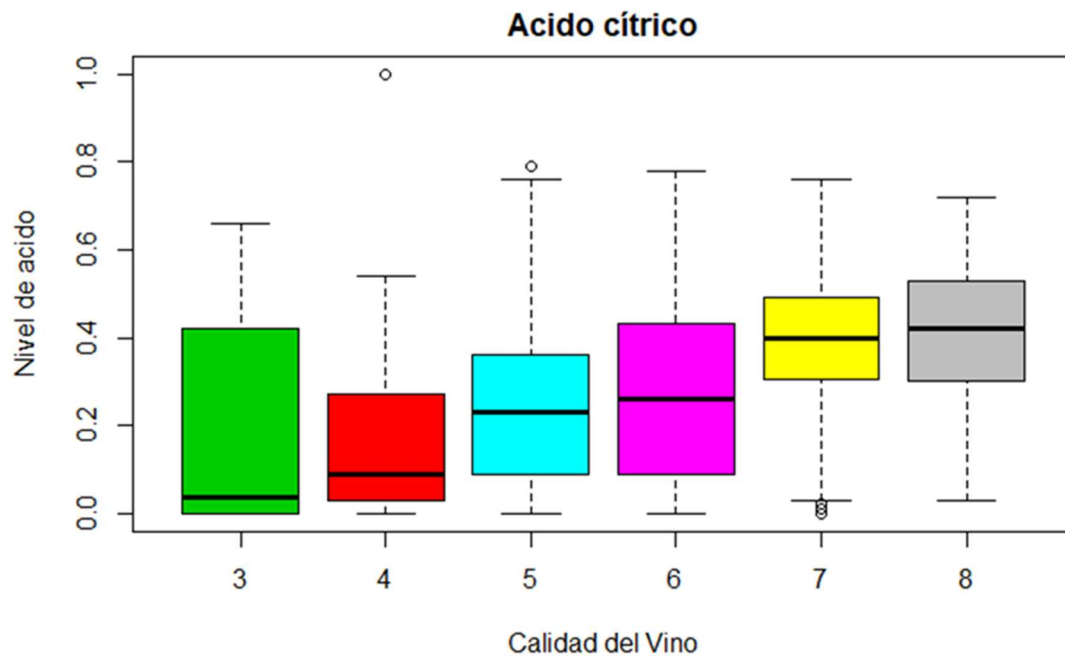
Conclusión: El nivel de acidez se sitúa en la mayor parte de la variable "quality" entre 7.5 y 10 por lo que no se puede concluir que este atributo determine el nivel de calidad del vino.

b) Variable Volatilidad de la acidez



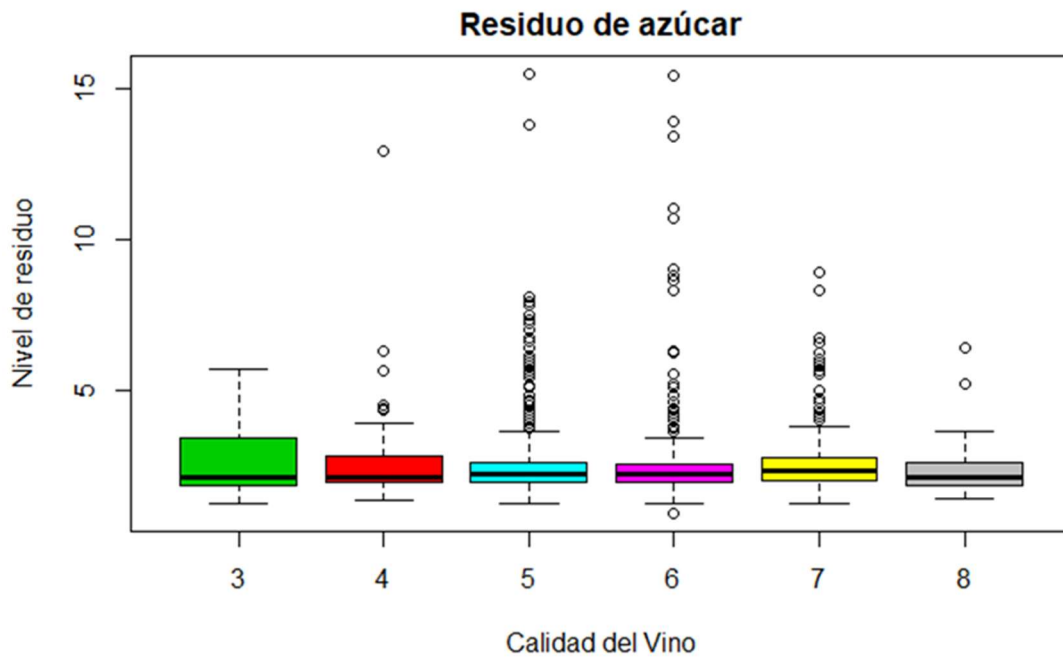
Conclusión: En este caso sí que se puede apreciar cómo la volatilidad disminuye a medida que aumenta la calidad del vino, por lo tanto, en un primer momento se selecciona dicha variable como candidata para el análisis.

c) Variable Ácido Cítrico



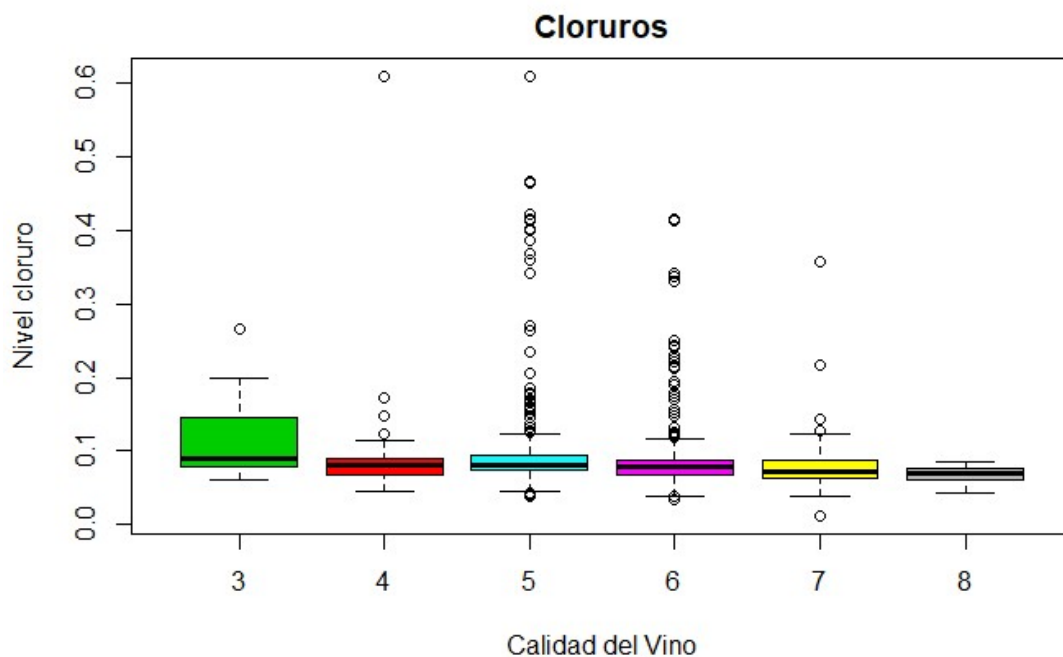
Conclusión: También se observa como el ácido cítrico aumenta a medida que la calidad del vino es mejor, por lo tanto, en un primer momento se tendrá en cuenta para el análisis.

d) Variable Residuo de azúcar.



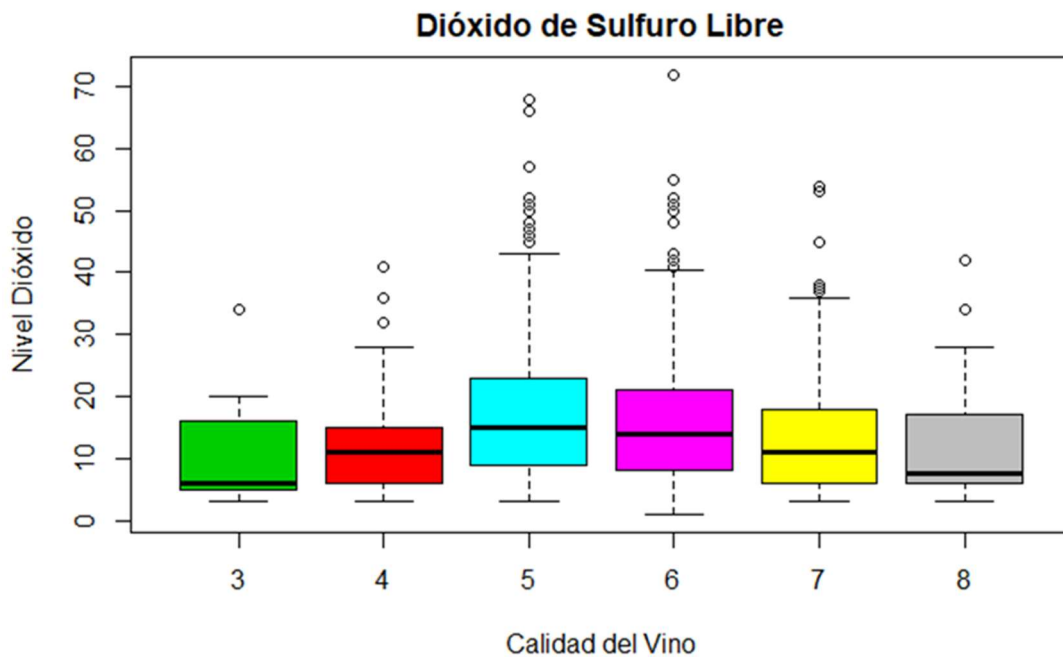
Conclusión: Presenta muchísimos atípicos, sin embargo, como se comentó en el apartado anterior éstos no se van a tener en cuenta. El nivel de residuo es muy similar en todas las calidades, por lo que no se tendrá en cuenta esta variable en los análisis posteriores.

e) Variable Cloruros



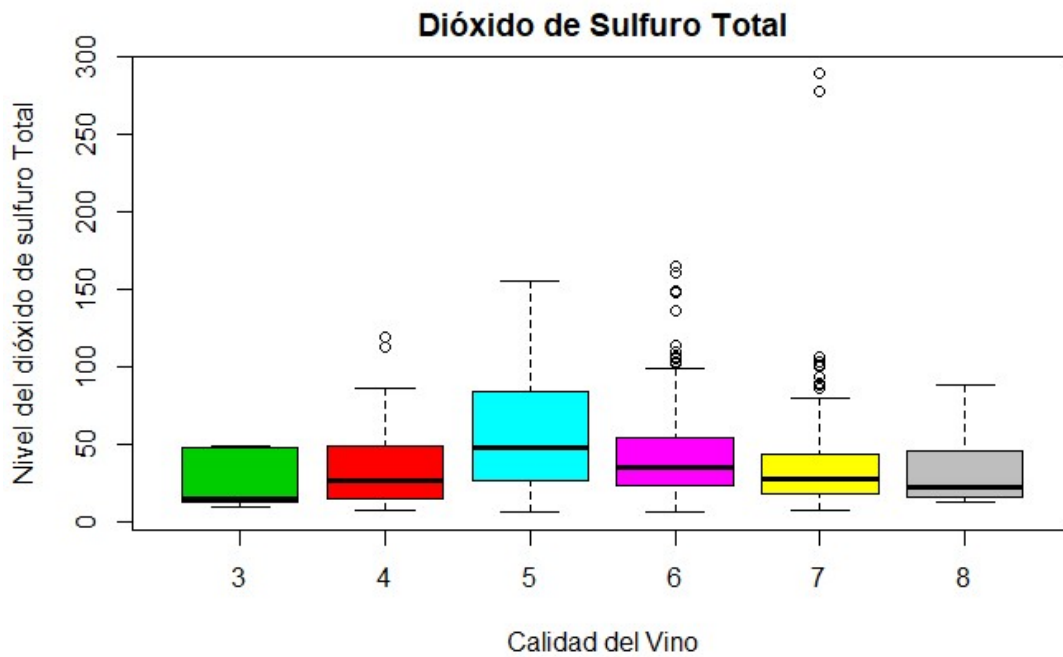
Conclusión: Presenta muchos atípicos, sin embargo, el nivel de cloruro no es determinante en la calidad del vino, por lo tanto dicha variable no se tendrá en cuenta en los análisis posteriores.

f) Variable Dióxido de Sulfuro Libre



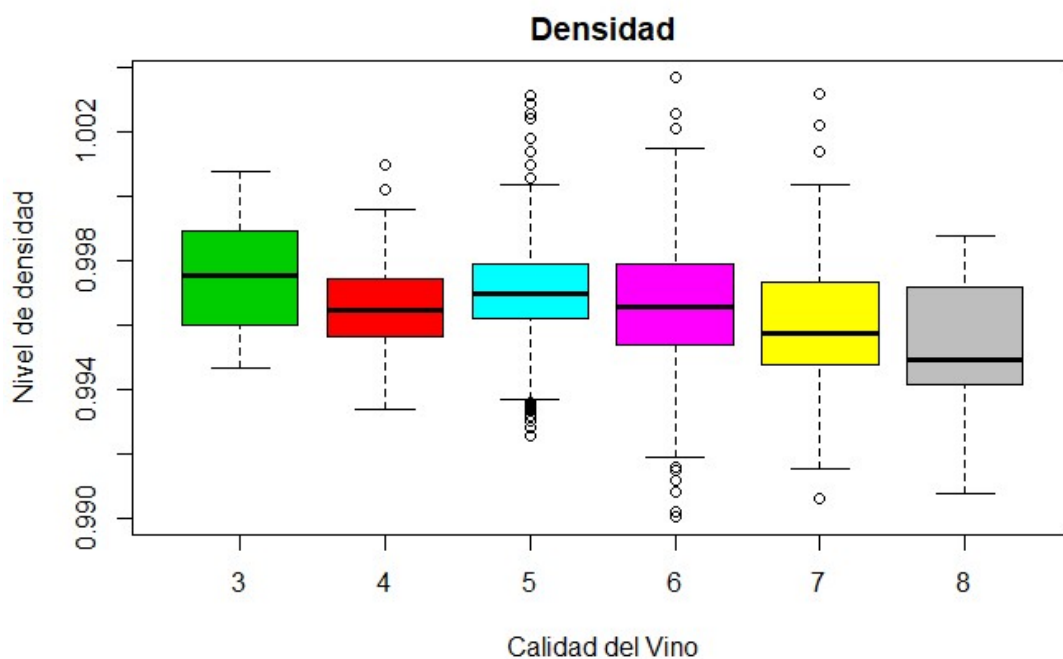
Conclusión: Los niveles son superiores en los vinos intermedios, sin embargo, se observa cómo el nivel de Dióxido de sulfuro libre es similar tanto en un vino de baja calidad como en otro de superior calidad. Por lo tanto, no se va a tener en cuenta en los análisis posteriores.

g) Variable Dióxido de sulfuro Total



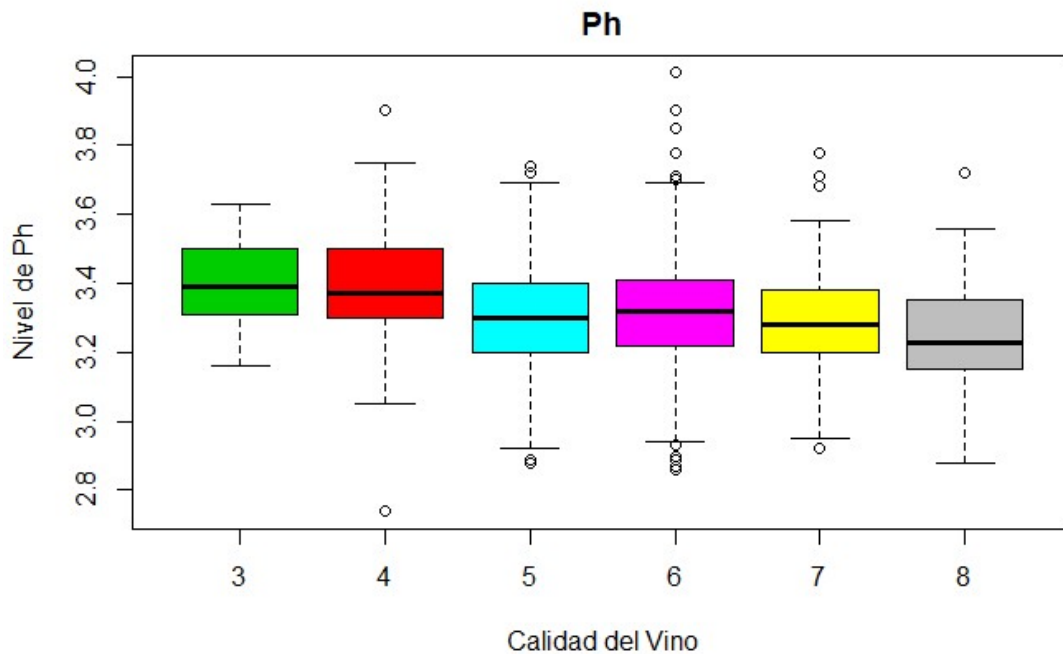
Conclusión: Es un análisis similar al Dióxido de sulfuro libre. Niveles más altos en calidad del vino medio, sin embargo, los vinos de más baja calidad y los de mayor nivel de calidad tienen valores similares. Por lo tanto, esta variable no se tendrá en cuenta en los análisis posteriores.

h) Variable Densidad



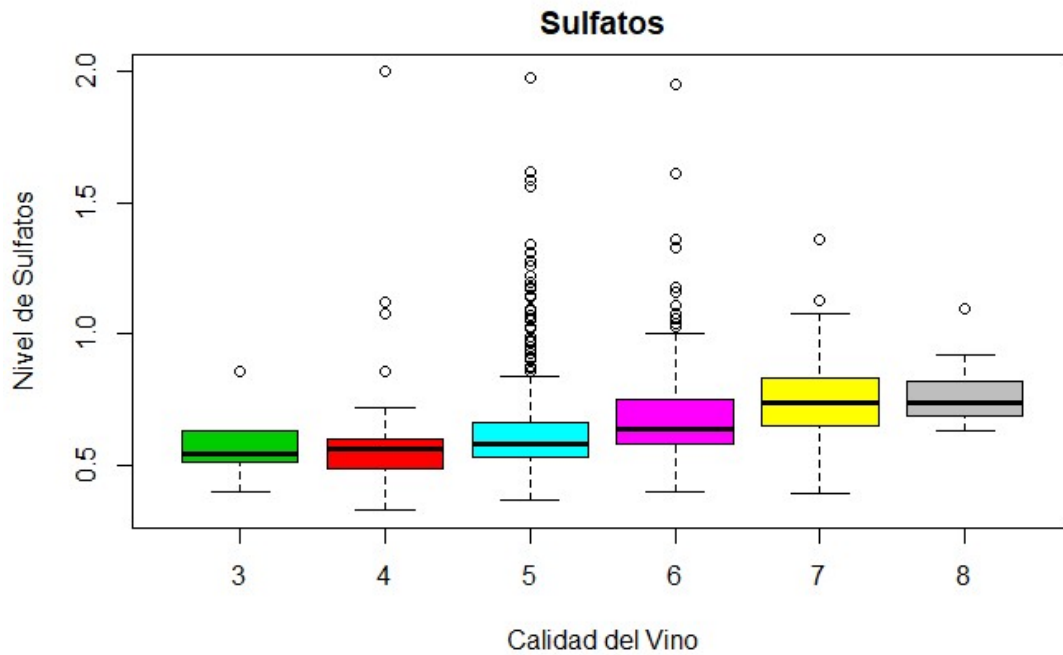
Conclusión: En esta variable sí que se aprecia que la densidad del vino disminuye a medida que aumenta la calidad del vino. Por lo tanto, sí que es una variable que se tendrá en cuenta para el análisis posterior.

i) Variable Ph



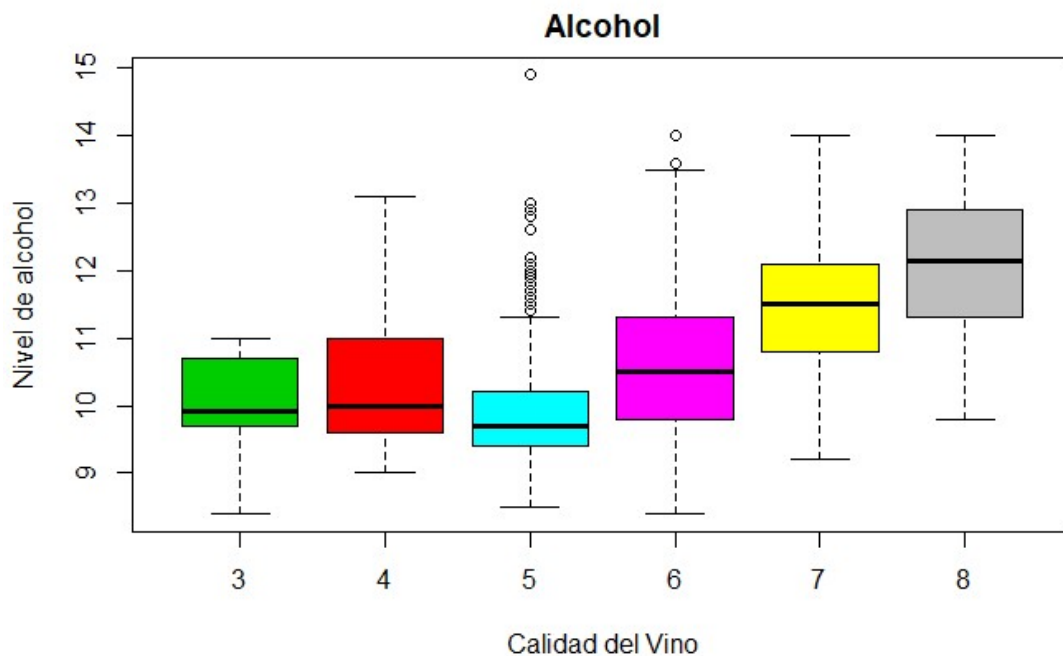
Conclusión: En la gráfica se observa cómo el Ph desciende a medida que aumenta la calidad del vino, salvo en los vinos intermedios, por lo que se utilizará esta variable para análisis posteriores.

j) Variable Sulfatos



Conclusión: Se observa una gran cantidad de atípicos en los valores medios del vino, sin embargo, podemos concluir que el nivel de sulfatos aumenta a medida que se incrementa la calidad del vino, por lo tanto, será una de las variables a tener en cuenta en sucesivos análisis.

k) Variable Alcohol



Conclusión: Salvo en los vinos de peor calidad, se aprecia un incremento del alcohol a medida que aumenta la calidad del vino, por lo que será una variable a tener en cuenta en posteriores análisis.

Por lo tanto, la conclusión final de este análisis nos indica que las variables que determinan la calidad del vino son las siguientes:

- Volatilidad de la acidez
- Ácido cítrico
- Densidad
- Ph
- Sulfatos
- Alcohol

Es decir, de los once componentes analizados, utilizaremos 6 para determinar la calidad del vino.

Sin embargo, vamos a buscar si existe alguna correlación entre las variables seleccionadas que nos permitan afinar un poco más la selección de los grupos. Es decir, si encontramos que existe una fuerte correlación proporcional entre algunas de estas variables se podrá tomar la decisión de utilizar una de ellas y descartar la otra.

Para ello, en primer lugar, se realizará una correlación entre los atributos del conjunto de datos para analizar qué variables de las analizadas están más estrechamente relacionadas. Las relaciones que presentan un color azul son las correlaciones positivas, mientras que las marcadas en rojo las correlaciones inversas.

```
corrgram(x=cor(wineData))
```



Y a continuación se realiza el análisis numérico de las correlaciones entre las variables.

```
R = cor(wineData[, 1:12])
round(R, 3)
```

```

fixed.acidity volatile.acidity citric.acid
fixed.acidity      1.000      -0.256      0.672
volatile.acidity   -0.256      1.000     -0.552
citric.acid         0.672     -0.552      1.000
residual.sugar      0.115      0.002      0.144
chlorides           0.094      0.061      0.204
free.sulfur.dioxide -0.154     -0.011     -0.061
total.sulfur.dioxide -0.113      0.076      0.036
density             0.668      0.022      0.365
pH                  -0.683      0.235     -0.542
sulphates           0.183     -0.261      0.313
alcohol             -0.062     -0.202      0.110
quality             0.124     -0.391      0.226

residual.sugar chlorides free.sulfur.dioxide
fixed.acidity      0.115      0.094     -0.154
volatile.acidity    0.002      0.061     -0.011
citric.acid         0.144      0.204     -0.061
residual.sugar      1.000      0.056      0.187
chlorides           0.056      1.000      0.006
free.sulfur.dioxide 0.187      0.006      1.000
total.sulfur.dioxide 0.203      0.047      0.668
density             0.355      0.201     -0.022
pH                  -0.086     -0.265      0.070
sulphates           0.006      0.371      0.052
alcohol             0.042     -0.221     -0.069
quality             0.014     -0.129     -0.051

total.sulfur.dioxide density pH sulphates
fixed.acidity      -0.113      0.668 -0.683      0.183
volatile.acidity     0.076      0.022  0.235     -0.261
citric.acid         0.036      0.365 -0.542      0.313
residual.sugar      0.203      0.355 -0.086      0.006
chlorides           0.047      0.201 -0.265      0.371
free.sulfur.dioxide 0.668     -0.022  0.070      0.052
total.sulfur.dioxide 1.000      0.071 -0.066      0.043
density             0.071      1.000 -0.342      0.149
pH                  -0.066     -0.342  1.000     -0.197
sulphates           0.043      0.149 -0.197      1.000
alcohol             -0.206     -0.496  0.206      0.094
quality             -0.185     -0.175 -0.058      0.251

alcohol quality
fixed.acidity   -0.062      0.124
volatile.acidity -0.202     -0.391
citric.acid      0.110      0.226
residual.sugar   0.042      0.014
chlorides        -0.221     -0.129
free.sulfur.dioxide -0.069     -0.051
total.sulfur.dioxide -0.206     -0.185
density          -0.496     -0.175
pH               0.206     -0.058
sulphates        0.094      0.251
alcohol          1.000      0.476
quality          0.476      1.000
```

(Top Level) ↕

En el resultado numérico las mejores correlaciones se encuentran en las siguientes parejas de variables:

Citric acid vs fixed-acid	0.67
Density vs fixed_acid	0.668

Ph vs fixed_acidity	-0.683
Citric acid vs Ph	-0.542
Total suf dióxido vs free sulfuro dióxido	0.668

En el análisis visual sí que se apreciaban ciertas correlaciones entre dichas variables, sin embargo, en el análisis numérico se aprecia que la correlación entre las variables es débil ya que considero que una correlación fuerte se encontraría desde mi punto de vista por encima de 0.75. Además, solo hay dos variables de las seleccionadas que presentan algo de correlación y que podría afectar al análisis, como son el ácido cítrico y el PH no presentan indicios de colinialidad entre ellas.

Por lo tanto, el grupo de variables a estudiar será el determinado anteriormente.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Se realiza un estudio para ver si las variables que se van a analizar presentan una distribución normal. Para determinarlo se va a aplicar el test de kolmogorov-Smirno que es el indicado para conjuntos de datos superior a 50 registros. Sin embargo, éste asume que se conoce la media y la variación población, lo que en la mayoría de los casos no es posible y para resolverlo existe una modificación denominada test de Lilliefors que asume que la media y la varianza son desconocidas. En el caso de que tuviésemos menos de 50 registros utilizaríamos el de Shapiro-Wilk.

Así pues, para el alcohol el test dice que el p-valor está por debajo de 0.05 por lo que no tiene normalidad.

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: wineData$alcohol
D = 0.12145, p-value < 2.2e-16
```

Para el Ph

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: wineData$pH
D = 0.040368, p-value = 2.244e-06
```

Para la densidad

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: wineData$pH
D = 0.040368, p-value = 2.244e-06
```

Para el sulfato

Lilliefors (Kolmogorov-Smirnov) normality test

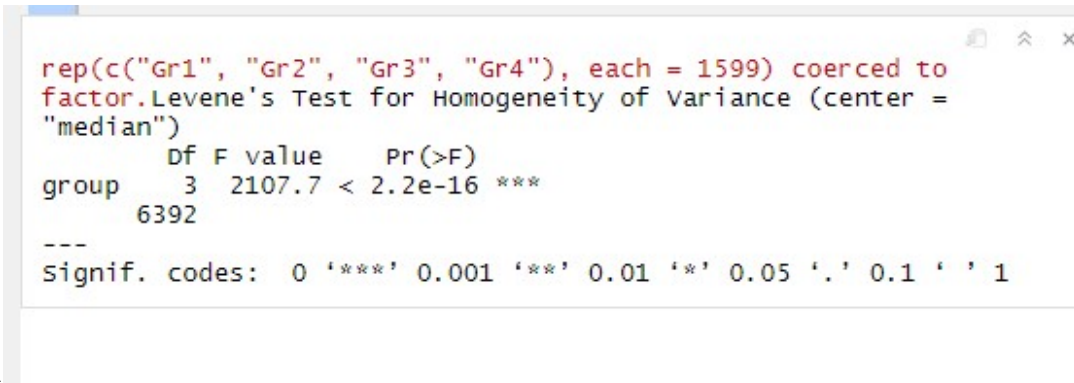
```
data: wineData$sulphates
D = 0.12479, p-value < 2.2e-16
```


Conclusión. - Como en los cuatro casos el p-valor está por debajo de 0.05 se puede asumir que no hay normalidad, por lo que se descarta la hipótesis nula.

A continuación, se va a estudiar la homogeneidad de las varianzas para los grupos analizados anteriormente: alcohol, pH, densidad y sulfatos.

Según el análisis de normalidad realizado en el apartado anterior, los cuatro grupos de datos no presenta normalidad en su distribución por lo tanto se va a estudiar la homogeneidad de la varianza a partir del test de Levene.

```
leveneTest(c(wineData$alcohol,wineData$pH,wineData$density,wineData$sulphates),
rep(c("Gr1","Gr2","Gr3","Gr4"), each=1599), center="median")
```



```
rep(c("Gr1", "Gr2", "Gr3", "Gr4"), each = 1599) coerced to
factor. Levene's Test for Homogeneity of Variance (center =
"median")
      Df F value    Pr(>F)
group   3 2107.7 < 2.2e-16 ***
      6392
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al ser el valor p-valor inferior a 0.05 se rechaza la hipótesis nula y por tanto se determina de que no hay homogeneidad de varianzas entre los valores seleccionados.

De todos modos, el teorema del límite central determina que si la muestra supera los 30-50 elementos se puede tener en cuenta que sigue una distribución normal de media 0 y desviación estándar 1, por lo que al tener el conjunto de datos 1599 registros suponemos que sigue distribución normal para continuar con el análisis.

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

En primer lugar, y de acuerdo con el objetivo del ejercicio, se va a realizar un modelo de regresión múltiple para poder clasificar un nuevo vino del que tengamos las características de su composición química.

Tal y como se ha analizado anteriormente, la variable dependiente o de respuesta será la calidad del vino, mientras el conjunto de variables independientes o predictores serán las seis variables seleccionadas anteriormente.

En este caso, el modelo de regresión múltiple va a constar de dos análisis, en primer lugar, se utilizará como predictor, mientras que por otra parte se utilizará para evaluar la influencia que tienen los predictores sobre dicha variable dependiente.

Predicción de la calidad del vino.

Para empezar a trabajar tendremos en cuenta que los predictores que se van a utilizar en el modelo de regresión múltiple no presentan indicios de colinialidad tal y como se puede comprobar el apartado 4.1, por lo tanto, se van a utilizar de momento los 6 predictores seleccionados.:

- Volatilidad de la acidez
- Ácido cítrico
- Densidad
- Ph
- Sulfatos
- Alcohol

Como comentado en el apartado 4.2 se asume que los predictores siguen una distribución normal de acuerdo con el teorema del límite central.

Se aplica un modelo de regresión múltiple teniendo en cuenta todas las variables descritas anteriormente del siguiente modo:

```
Call:
lm(formula = quality ~ volatile.acidity + citric.acid + density +
    pH + sulphates + alcohol, data = wineData)

Residuals:
    Min       1Q   Median       3Q      Max
-2.64141 -0.38701 -0.06721  0.45480  2.11572

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -11.77058    11.92162  -0.987  0.323631
volatile.acidity -1.32197     0.11597 -11.399 < 2e-16 ***
citric.acid    -0.37834     0.13479  -2.807  0.005064 **
density        15.84518    11.88503   1.333  0.182655
pH             -0.47787     0.13381  -3.571  0.000366 ***
sulphates       0.65627     0.10367   6.330  3.17e-10 ***
alcohol         0.34190     0.01985  17.222 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6563 on 1592 degrees of freedom
Multiple R-squared:  0.3421,    Adjusted R-squared:  0.3396
F-statistic: 138 on 6 and 1592 DF,  p-value: < 2.2e-16
```

Al analizar los resultados se observa que la variable densidad tiene un p-valor de 0.18 (superior a 0.05) por lo que existe un 18% de probabilidad de que este predictor no

sea significativo para la regresión, sin embargo, el resto de predictores sí que son una buena elección para el modelo.

Por otra parte, el R-squared ajustado y no ajustado no son muy elevados. 0.3396 y 0.3421 respectivamente, de todos modos, un valor alto no quiere decir que sea necesariamente bueno ni al revés, un valor bajo sea necesariamente malo.

Los valores residuales entre el 1Q y el 3Q se encuentra son aproximadamente cero, sin embargo, se aprecia desvíos importantes en sentido al mínimo y al máximo.

Del análisis se concluye de que no se va a utilizar en el modelo la variable densidad del, siendo el resultado el siguiente:

```
call:
lm(formula = quality ~ volatile.acidity + citric.acid + pH +
    sulphates + alcohol, data = wineData)

Residuals:
    min       1q   median       3q      max
-2.63642 -0.37783 -0.07067  0.46233  2.10817

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.11161    0.45915   8.955  < 2e-16 ***
volatile.acidity -1.28363    0.11238  -11.422  < 2e-16 ***
citric.acid     -0.29730    0.12034   -2.471   0.0136 *
pH              -0.47476    0.13382   -3.548   0.0004 ***
sulphates        0.67306    0.10293    6.539 8.31e-11 ***
alcohol          0.32731    0.01657   19.756  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6564 on 1593 degrees of freedom
Multiple R-squared:  0.3413,    Adjusted R-squared:  0.3393
F-statistic: 165.1 on 5 and 1593 DF,  p-value: < 2.2e-16
```

Por lo tanto, la función para obtener la calidad de un vino determinado será la siguiente:

Quality=4.11 – 1.28363*volatile.acidity -0.2973*citric.acid -0.47476*pH+0.67306*sulphates+0.32731*alcohol

A continuación, se va a determinar qué importancia relativa tiene cada uno de los componentes, así que se va a realizar una estandarización de los valores.

```
242 ~~~{r}
243 modelo_standard<-lm(quality~volatile.acidity+citric.acid+pH+sulphates+alcohol,
244   data=wineDataStd)
```

Utilizamos la librería relaimpo: library(relaimpo)

Y a continuación calculamos:

```

####{r}
calc.relimp(modelo_standard, type=c("lmg"),rela=TRUE)

```

Response variable: quality
 Total response variance: 1
 Analysis based on 1599 observations

5 Regressors:
 volatile.acidity citric.acid pH sulphates alcohol
 Proportion of variance explained by model: 34.13%
 Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

	lmg
volatile.acidity	0.27389241
citric.acid	0.05381378
pH	0.01574112
sulphates	0.09976349
alcohol	0.55678920

El resultado determina qué la importancia relativa es la siguiente:

1. El alcohol con un 55.67%
2. La acidez volátil con un 27.39%
3. Los sulfatos con un 9.98%
4. El ácido cítrico con un 5.3%
5. El pH con un 1.57%

5.- Representación de los resultados a partir de tablas y gráficas

Con el fin de verificar la bondad del modelo se ha decidido utilizar los mismos datos del fichero de vinos y ya clasificados para comprobar como clasifica el modelo. Y para ello se va a comprobar la clasificación de la calidad del vino original con la clasificación de la calidad del vino estimada por el modelo.

En primer lugar, se crea dicho conjunto de datos:

```

####{r}
datosTestX<-select(wineData,volatile.acidity,citric.acid,pH,sulphates,alcohol)

```

Y se hace la predicción sobre el modelo:

```

####{r}
y_predict<-predict(modelo,datosTestX)

```

Se crea una nueva columna en el conjunto de datos original con la predicción.

```

####{r}
wineData$prediccion<-as.integer(y_predict+0.5)

```

total.sulfur.dioxide <dbl>	density <dbl>	pH <dbl>	sulphates <dbl>	alcohol <dbl>	quality <int>	prediccion <int>
24	0.99695	3.22	0.82	10.3	7	6
145	0.99750	3.04	1.03	9.3	5	6
49	0.99545	3.36	0.79	9.5	5	5
39	0.99610	3.34	0.55	9.2	5	5
37	0.99615	3.34	0.56	9.2	6	5
28	0.99940	3.20	0.77	10.8	7	6
28	0.99940	3.20	0.77	10.8	7	6
120	0.99625	3.29	0.53	9.3	5	5
95	0.99660	3.22	0.67	9.4	5	5
19	0.99800	3.31	0.88	10.5	7	6

49 rows | 7-13 of 13 columns

Previous 1 ... 19 20 21 22 23 ... 100 Next

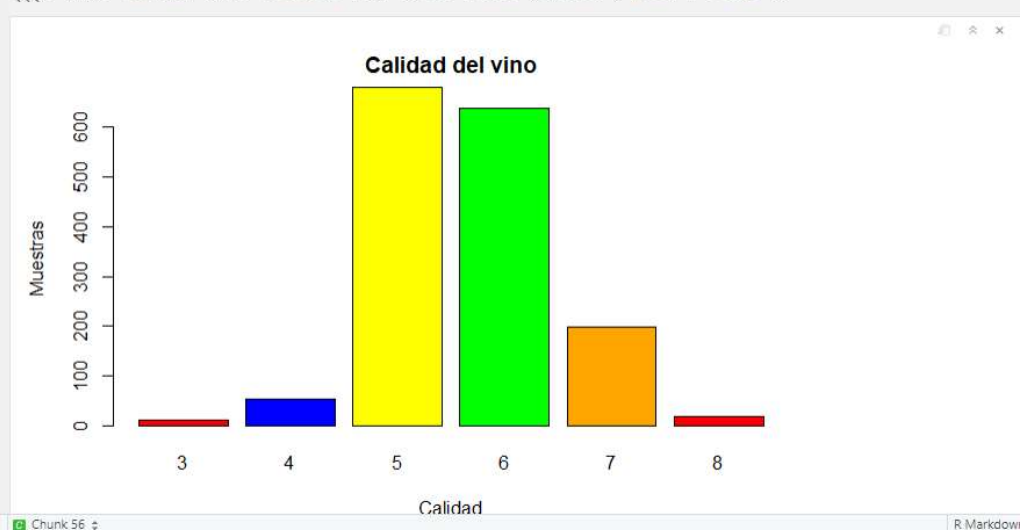
La tabla y el gráfico original son los siguientes:

```
{r}
quality1<-table(wineData$quality)
quality1
```

```

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

```
{r}
barplot(quality1,main="Calidad del
vino",xlab="Calidad",ylab="Muestras",col=c("red","blue","yellow","green","orange"))
```

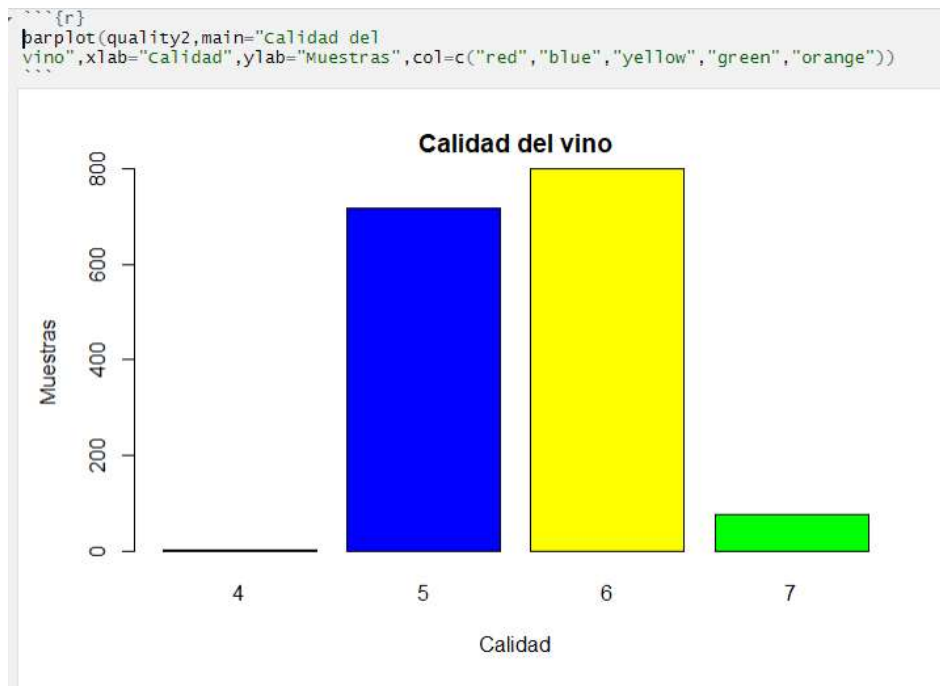


Y la tabla y el gráfico con la estimación es el siguiente:

```
{r}
quality2<-table(wineData$prediccion)
quality2
```

```

 4   5   6   7
3 719 801  76
```

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

Se verifica que el modelo no detecta la calidad de vino más alta, así como la más baja del conjunto original, por lo que es previsible que éstas se hayan agregado a la clase 7 y a la clase 4 respectivamente, siendo éstas por lo tanto en el conjunto estimado la calidad más alta y la calidad más baja. En el conjunto original la clase 5 y 6 comprende el 82.5% de todas las muestras de vino, mientras que en el conjunto estimado comprende el 95% de las muestras por lo que se deduce desde este punto de vista un exceso de concentración de las muestras en estas dos clases. Si generalizamos un poco más, es decir, si tenemos en cuenta que las muestras de vino con calidad 3, 4 y 5 serían muestras de baja calidad y las muestras de vino con calidad 6, 7 y 8 de alta calidad, tendríamos la siguiente distribución. El 46.2% de los vinos son de baja calidad, mientras que el 53.8% serían de buena calidad en el conjunto de datos original. Si atendemos a la predicción, es decir, 4 y 5 baja calidad y 6 y 7 alta calidad, el 45.15% del vino sería de baja calidad y el 54.85% de alta calidad. Esta distribución se parece bastante a la distribución original.

Como conclusión, el modelo ha funcionado para determinar si un vino puede ser de buena o de mala calidad, es decir, para estimar una variable dicotómica podría funcionar, sin embargo, para segregar dentro de cada grupo por niveles de calidad no sería suficiente como de hecho nos informa el R^2 del modelo (un valor bajo).

Por lo tanto, habría que reconsiderar el modelo para una variable dicotómica, bueno o malo, y, además, habría que probar con otro conjunto de muestras para determinar definitivamente su precisión.

