Práctica 2: Limpieza y análisis de datos

Elisabeth Anna López Simpson y Jesús Antonio Blay Tamarit

7 de June 2021

Contents

0. Carga del dataset y de los paquetes de R necesarios

```
# Cargamos los paquetes de R que vamos a usar
library(ggplot2)
library(dplyr)
library(gridExtra)
library(corrplot)
library(relaimpo)
library(psych)
library(lmtest)
library(nortest)

# Cargamos el fichero de datos
wine <- read.csv('winequality-red.csv', header=TRUE)
```

1. Descripción del dataset.

¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset contiene distintas observaciones de vino tinto y para cada observación se recogen 12 atributos. Los 11 primeros atributos son mediciones físicoquímicas (de tipo numérico), mientras que el último atributo

(la variable de interés), es una clasificación sensorial de la calidad del vino, en una escala de 0 a 10 (esta variable sólo puede tomar valores enteros).

Nuestro objetivo es estudiar como influyen las 11 variables de entrada en la calidad del vino y cuáles de ellas son más decisivas en una mayor calidad.

Variables de entrada:

1 - *fixed acidity*: mide la acidez fija debida a los ácidos orgánicos presentes en la uva, como los ácidos tartárico, málico y cítrico. 2 - *volatile acidity*: mide el ácido acético presente en el vino por la fermentación. Puede dar lugar a sensación de olor a vinagre como consecuencia de fermentos dañinos. 3 - *citric acid*: acidificante para corregir la acidez en mostos y vinos. 4 - *residual sugar*: mide la cantidad de azúcar presente tras la fermentación. 5 - *chlorides*: mide la concentración de iones de cloruro, generalmente indicativa de la presencia de cloruro de sodio, que aumenta la salinidad de un vino, lo que puede contribuir o restar valor al sabor y la calidad general del vino. 6 - *free sulfur dioxide*: mide la cantidad de dioxido de azufre o SO2 que está libre en el vino. 7 - *total sulfur dioxide*: mide el dióxido de azufre total (TSO2), es decir, la porción de SO2 que está libre en el vino más la porción que está unida a otras sustancias químicas en el vino, como aldehídos, pigmentos o azúcares. 8 - *density*: mide la densidad del vino, en g/cm3 9 - *pH*: mide la acidez del vino. 10 - *sulphates*: aditivo para el vino que actúa como antimicrobiano y antioxidante. Puede contribuir a los niveles de dióxido de azufre. 11 - *alcohol*: porcentaje de alcohol en el vino.

Variable de interés:

12 - *quality*: clasificación sensorial de la calidad del vino, en una escala de 0 a 10.

Comenzamos con un primer vistazo a las distintas variables.

```
# Verificamos la dimensión y la estructura del conjunto de datos
dim(wine)
```

```
## [1] 1599   12
```

```
summary(wine)
```

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

Tenemos 1599 observaciones. Vemos también que efectivamente todas las variables son de tipo numérico. Además, todas las variables son continuas, a excepción de nuestra variable de interés, ***quality***, que es una variable discreta.

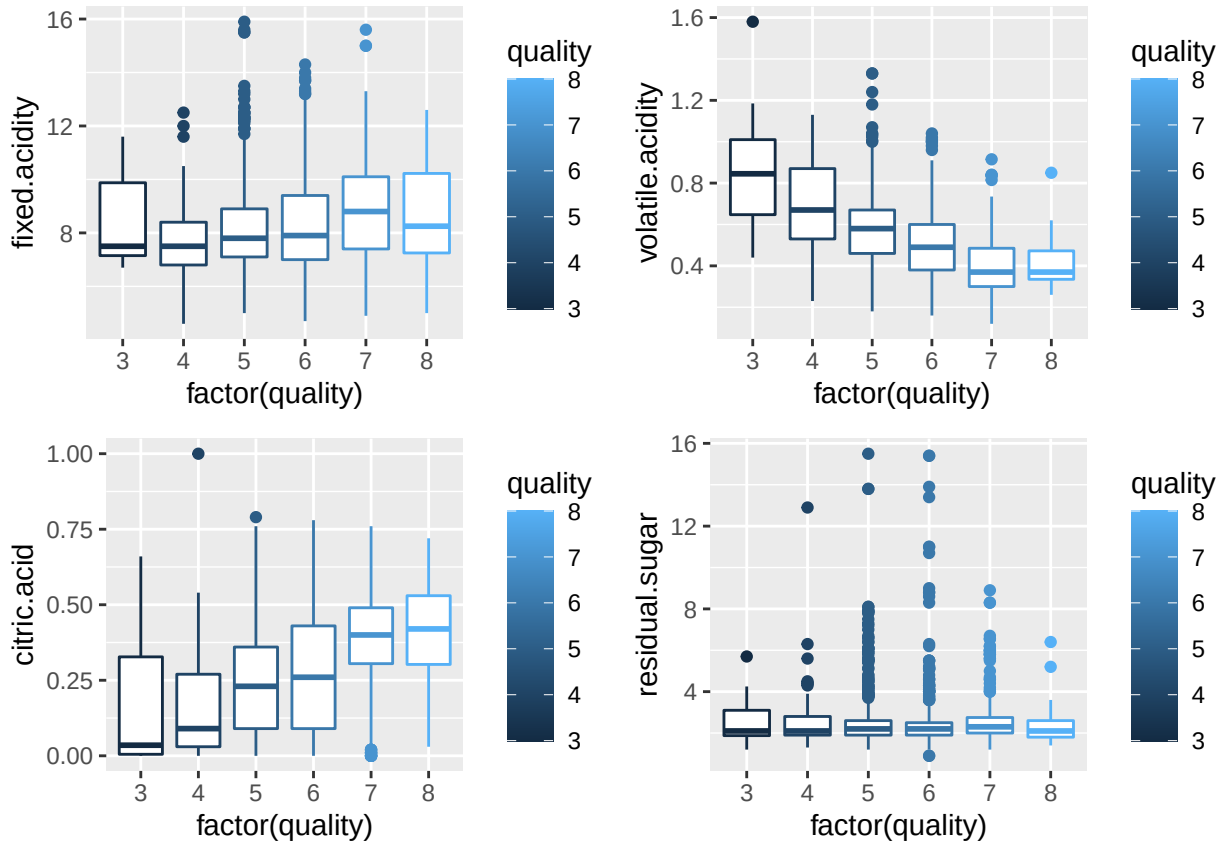
2. Integración y selección de los datos de interés a analizar.

De las 11 variables de entrada, a priori desconocemos como afecta cada una y su peso a la calidad final del vino, pero por la descripción de las variables, ya podemos identificar variables que estarán altamente correlacionadas.

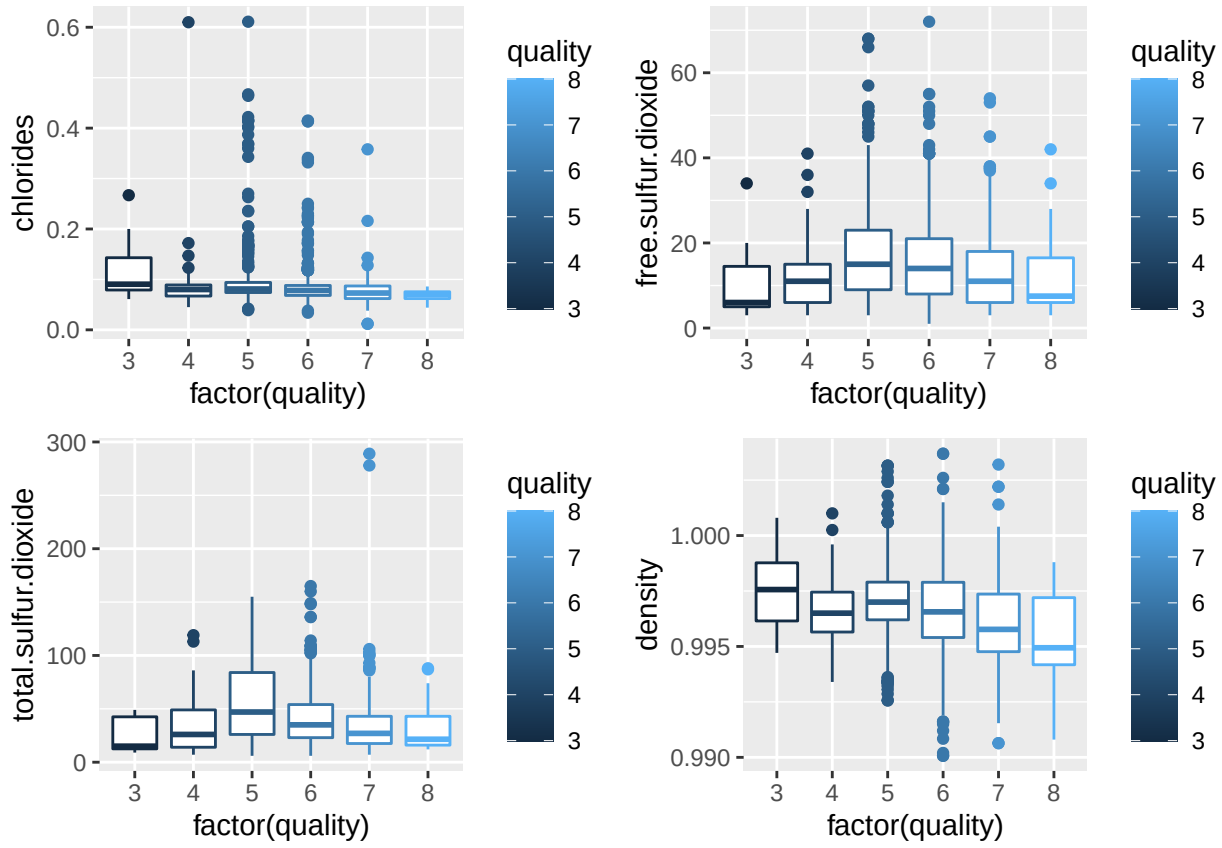
- ***ph*** - ***citric.acid*** - ***fixed.acidity*** - ya que son variables relacionas a la acidez.
- ***free.sulfur.dioxide*** y ***total.sulfur.dioxide***, ya que la segunda contiene la primera.

Para empezar vamos a estudiar la distribución de cada una de estas variables en función de la calidad del vino (cuyos valores, en nuestra muestra, observamos que van del 3 al 8).

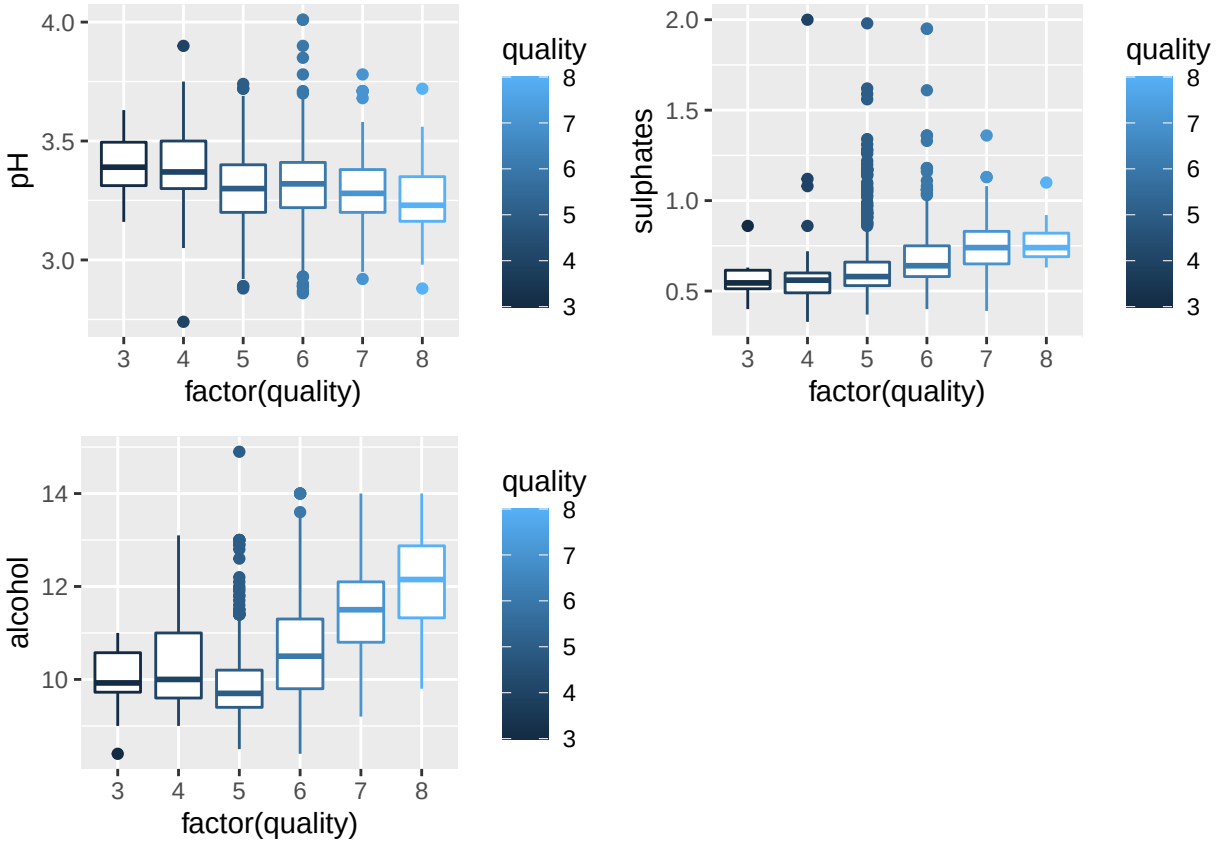
```
# Distribución de fixed.acidity, volatile.acidity, citric.acid y residual.sugar en función de la calida
v1 <- ggplot(data=wine, aes(x=factor(quality), y=fixed.acidity, color=quality)) +
  geom_boxplot()
v2 <- ggplot(data=wine, aes(x=factor(quality), y=volatile.acidity, color=quality)) +
  geom_boxplot()
v3 <- ggplot(data=wine, aes(x=factor(quality), y=citric.acid, color=quality)) +
  geom_boxplot()
v4 <- ggplot(data=wine, aes(x=factor(quality), y=residual.sugar, color=quality)) +
  geom_boxplot()

grid.arrange(v1, v2, v3, v4, ncol = 2, nrow = 2)
```




```
# Distribución de chlorides, free.sulfur.dioxide, total.sulfur.dioxide y density en función de la calida
v5 <- ggplot(data=wine, aes(x=factor(quality), y=chlorides, color=quality)) +
  geom_boxplot()
v6 <- ggplot(data=wine, aes(x=factor(quality), y=free.sulfur.dioxide, color=quality)) +
  geom_boxplot()
v7 <- ggplot(data=wine, aes(x=factor(quality), y=total.sulfur.dioxide, color=quality)) +
  geom_boxplot()
v8 <- ggplot(data=wine, aes(x=factor(quality), y=density, color=quality)) +
  geom_boxplot()

grid.arrange(v5, v6, v7, v8, ncol = 2, nrow = 2)
```




```
# Distribución de pH, sulphates, y alcohol en función de la calidad del vino
v9 <- ggplot(data=wine, aes(x=factor(quality), y=pH, color=quality)) +
  geom_boxplot()
v10 <- ggplot(data=wine, aes(x=factor(quality), y=sulphates, color=quality)) +
  geom_boxplot()
v11 <- ggplot(data=wine, aes(x=factor(quality), y=alcohol, color=quality)) +
  geom_boxplot()

grid.arrange(v9, v10, v11, ncol = 2, nrow = 2)
```



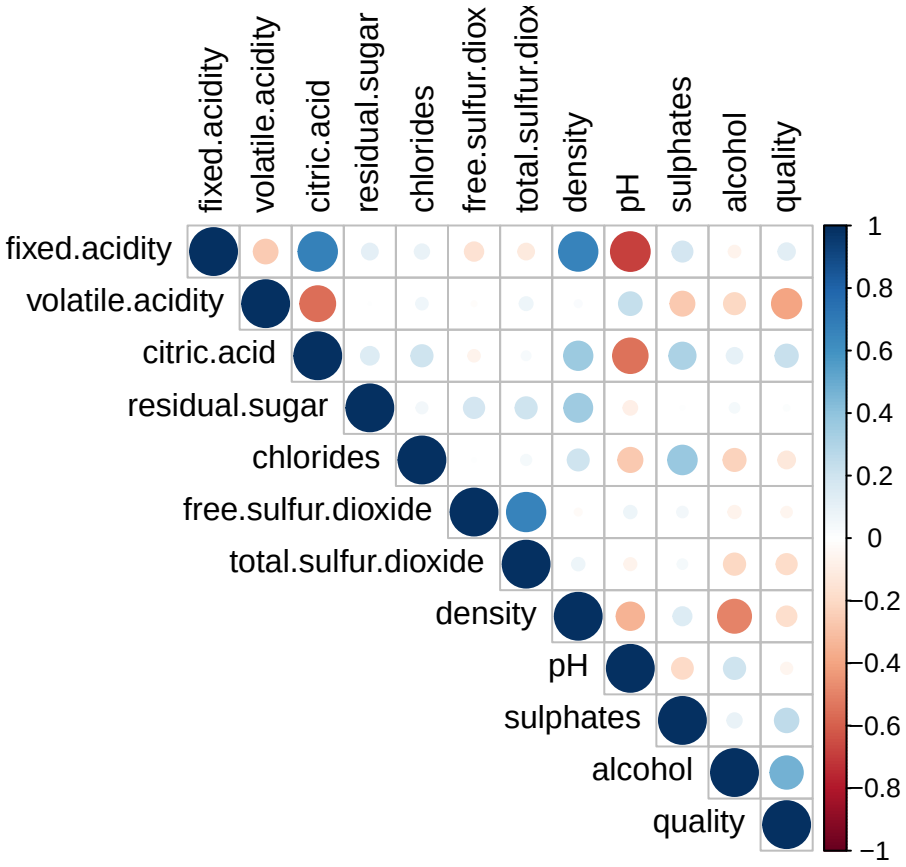
A primera vista es difícil sacar conclusiones a través de los boxplot, pero estos datos ya nos ofrecen un primer punto de partida a partir del cuál comenzar nuestro análisis:

Según muestran los *boxplots* anteriores, parece existir una correlación negativa entre la variable ***quality*** y las variables ***volatile.acidity***, ***chlorides*** y ***pH***, que por otra parte sería completamente coherente con la descripción de estas variables, detallada en el punto 1 (descripción del dataset). También observamos una aparente relación inversa entre ***quality*** y ***density***, que a priori no parece tan obvia y que estudiaremos posteriormente.

Igualmente, también parece existir una relación positiva entre la variable ***quality*** y las variables ***citric.acid*** , ***sulphates*** y ***alcohol***.

Si bien analizaremos la correlación más adelante, podemos apoyarnos en un gráfico de correlación para corroborar las observaciones hechas anteriormente con los boxplots.

```
# Estudio de la correlción
correlation <- cor(wine)
corrplot(correlation, type ="upper", tl.col="black")
```



Como podemos ver, comprobamos que nuestra variable ***quality*** se relaciona positivamente con ***alcohol***, ***sulphates***, y ***acid.citric*** y de forma negativa con ***volatile.acidity***, y ***chlorides***. Según este gráfico, no queda tan clara la correlación entre ***quality*** y **pH**.

Además, parece claro que la variable ***residual.sugar*** no va a tener efecto en la variable ***quality***, pero no la descartaremos por ahora.

Por último, como ya preveíamos parece que hay una correlación alta entre las variables ***pH*** y ***citric.acid***, y entre ***free.sulfur.dioxide*** y ***total.sulfur.dioxide***.

Analizaremos estas posibles correlaciones en profundidad más adelante. En todo caso, por ahora no podemos descartar ninguna de las variables ya que todas ellas corresponden a distintas mediciones de factores que podrían ser determinantes en la calidad del vino.

Sin embargo, nos interesaremos especialmente por analizar como afectan a la calidad del vino las variables ***density***, ***citric.acid*** , ***sulphates*** y ***alcohol***, así como también el ***pH*** (acidez total del vino) y el ***total sulfur dioxide*** (dióxido de azufre total).

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Antes de comenzar con el análisis, comprobamos la existencia de vacíos o nulos.

```
# Estadísticas de valores vacíos
colSums(is.na(wine))
```

```
##        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(wine=='')
```

```
##        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
```

```
# Estadísticas de valores cero
colSums(wine==0)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                  132
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

No existen valores vacíos (en caso de que los hubiera, no deberíamos prescindir de la observación, pero hubiese sido recomendable saberlo en su caso, para ignorar estos valores en las distintas funciones que usaremos en nuestro análisis, o para imputar un valor medio, si no queremos perder los registros.).

Por otra parte, sí existen valores nulos, concretamente 132 en la variable *citric.acid*. El ácido cítrico es un ácido que está presente de forma natural en la uva, con lo que probablemente todos estos valores "0", son valores que realmente deberían ser nulos.

En este caso tendríamos dos opciones. La primera, eliminar los registros de nuestra muestra, pero no es demasiado conveniente, porque corresponden a un 8% del total de los datos. La segunda opción es imputar su valor medio.

```
# Identificar valores 0
row_sub <- apply(wine, 1, function(row) all(row !=0 ))

# Sustituir por NA

wine[wine == 0] <- NA

# Sustituir por valor medio
wine$citric.acid[is.na(wine$citric.acid)] <- mean(wine$citric.acid, na.rm = TRUE)
```

Lo primero que hemos hecho es cambiar los ceros por valores nulos, y luego los cambiaremos por la media de la variable.

De esta manera ya contamos con un dataset que tiene valores válidos.

3.2. Identificación y tratamiento de valores extremos.

En lo boxplots del punto 2 observamos que efectivamente la mayoría de variables presentan un buen número de *outliers* si bien es razonable que se produzca la aparicicón de valores extremos en la medida en que estos valores se obtienen como resultado de un reacción química. Entendemos, dada su distribución, que se trata de valores normales derivados de los distintos procesos químicos y composiciones distintas de cada vino, por lo que en principio es razonable considerar que todos estos son valores legítimos y deberíamos considerarlos en nuestro análisis.

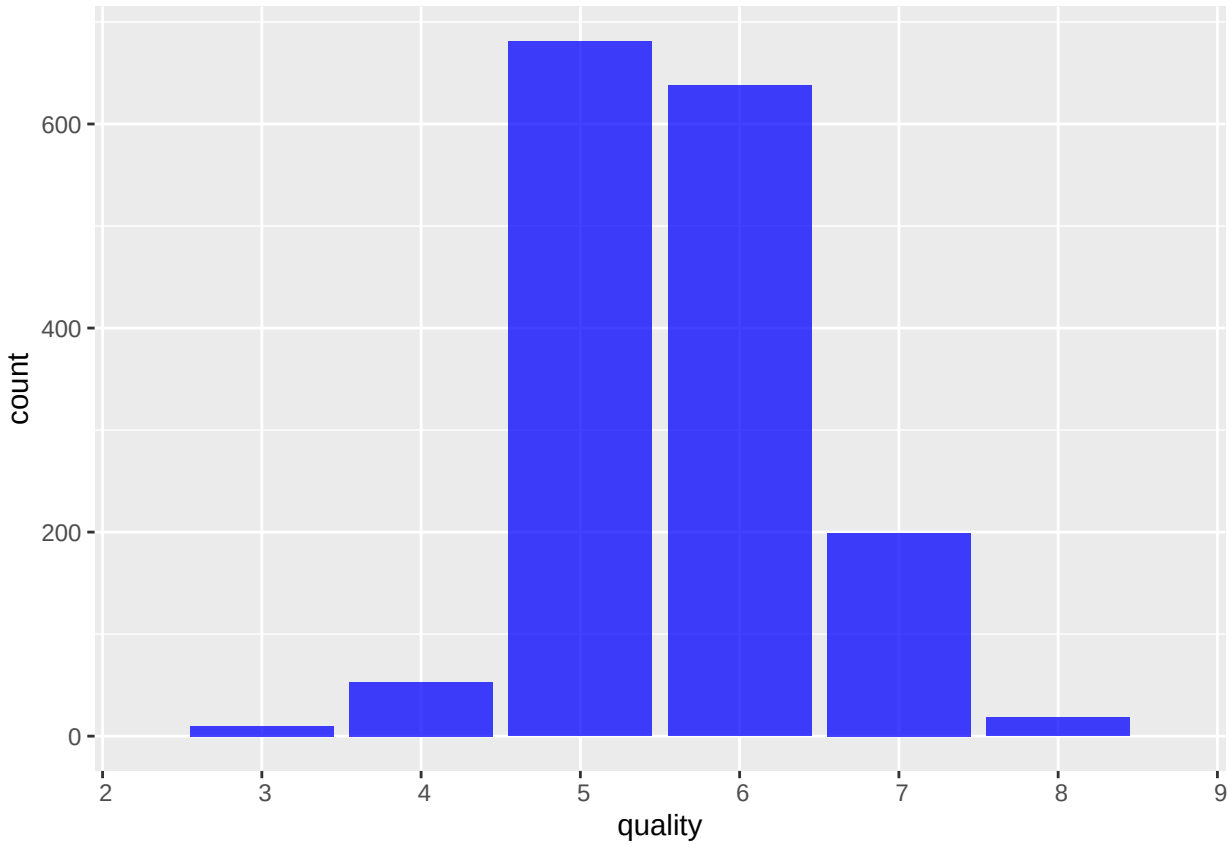
Respecto de la variable ***quality***, que en principio debe estar limitada a la escala de 0 a 10, observamos que efectivamente nuestros valores se encuentran dentro de dicha escala e incluso se concentran dentro de un rango más pequeño, entre 3-8, por lo que los no existen valores anómalos.

4. Análisis de los datos.

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

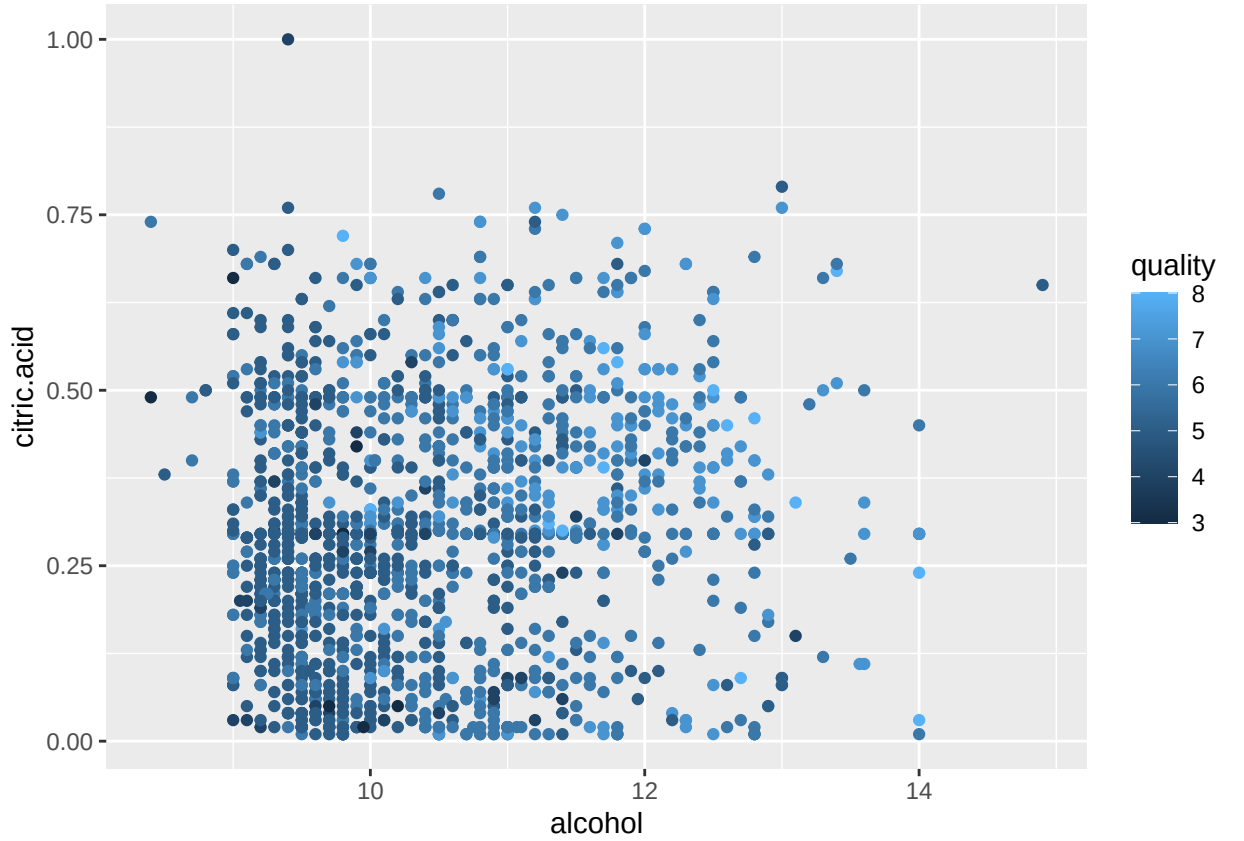
Sabemos que la calidad del vino se mide en una escala de 0 a 10. Sin embargo, hemos visto que de nuestras 1.599 observaciones, la mejor nota obtenida es 8, situándose la gran mayoría entre 5 y 6 (rango intercuartílico). Observamos con más detalle la distribución de todas las puntuaciones.

```
# Verificamos la dimensión y la estructura del conjunto de datos
ggplot(data=wine, aes(x=quality)) + geom_bar(fill='blue', alpha=0.75) + scale_x_discrete(limits=c(0:10)
```



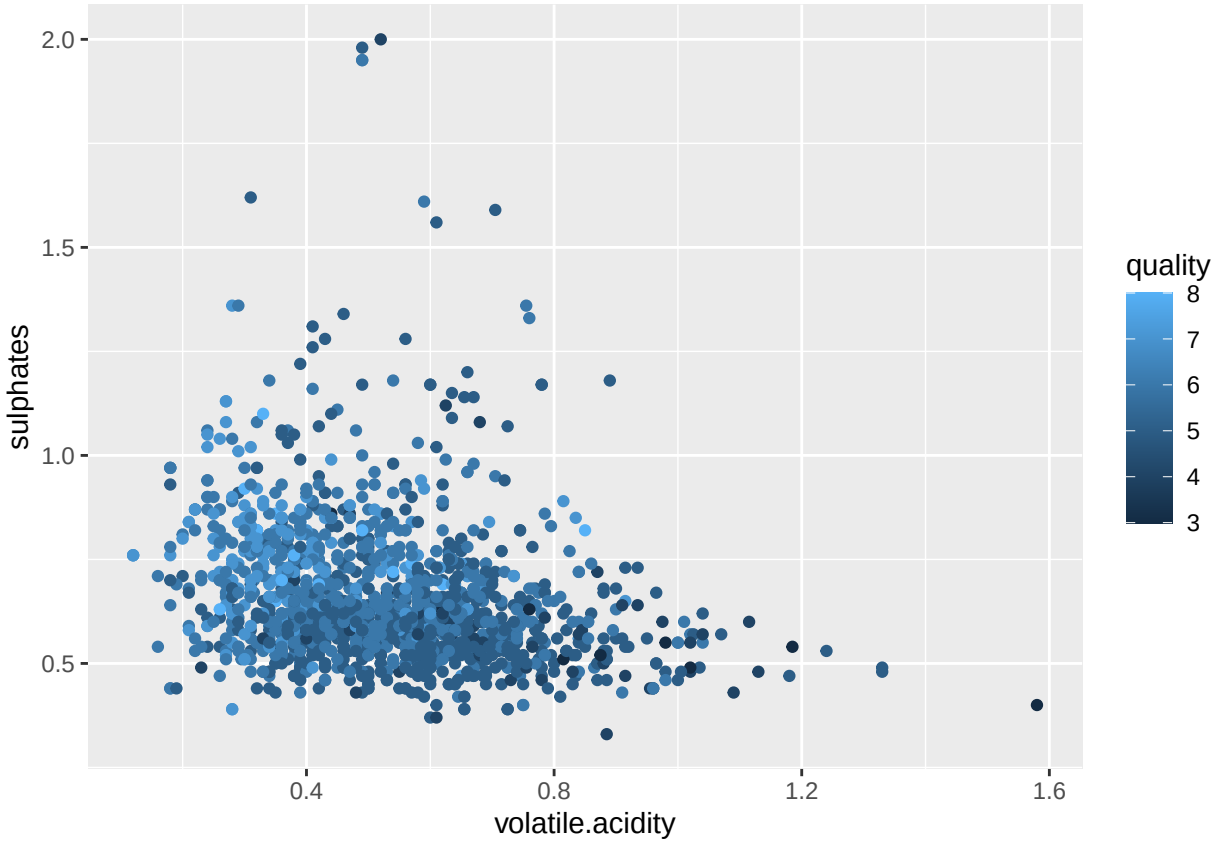
Vamos a explorar también como varían de manera conjunta algunas de las parejas de variables en las que ya hemos visto que parecen estar relacionadas con la variable ***quality***, distinguiendo la calidad del vino para cada uno de los puntos en el diagrama de dispersión.

```
# Distribución bivariante de las variables alcohol y citric.acid
ggplot(data=wine, aes(x=alcohol, y=citric.acid, color=quality)) + geom_point()
```



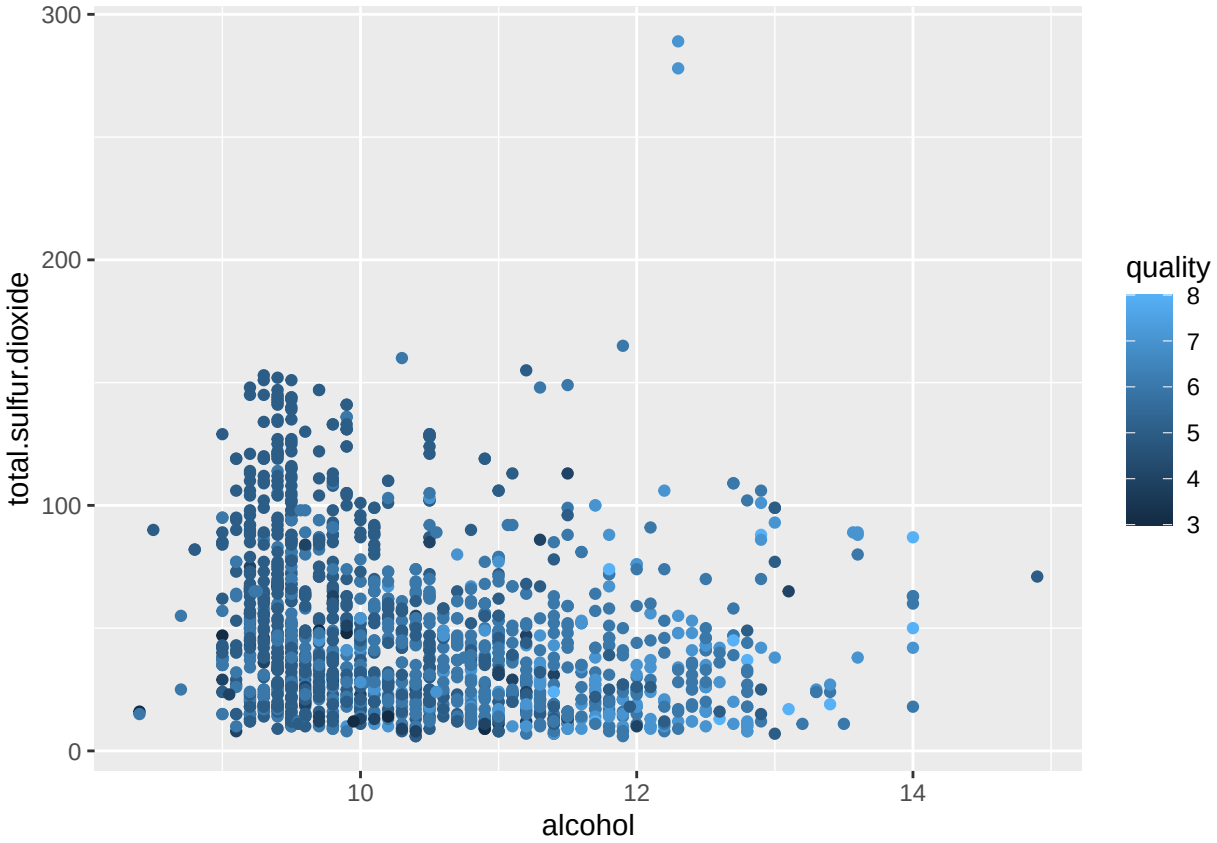
Parece apreciarse que los vinos con mayor calidad tienden a presentar una gran cantidad relativa de alcohol y/o de ácido cítrico.

```
# Distribución bivariante de las variables volatile.acidity y sulphates
ggplot(data=wine, aes(x=volatile.acidity, y=sulphates, color=quality)) +
  geom_point()
```



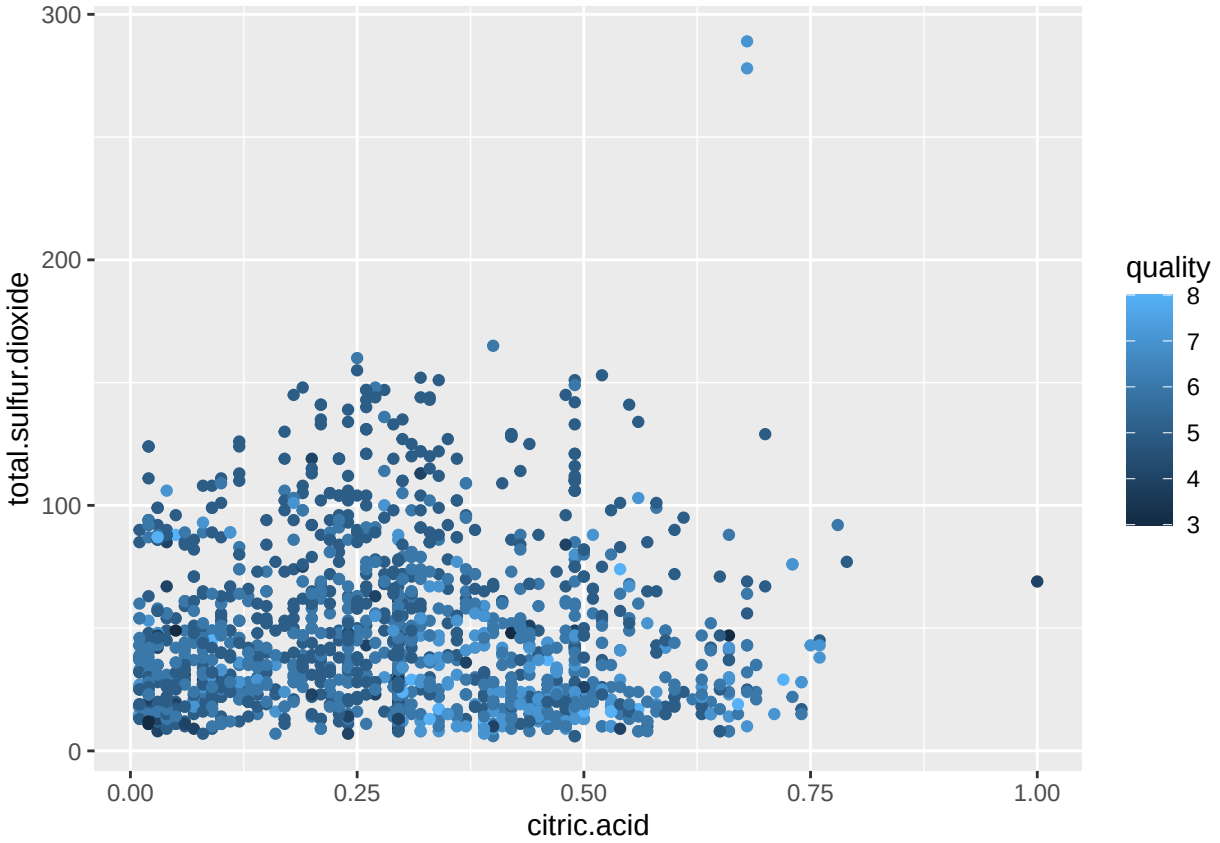
En general este gráfico de dispersión no parece sugerir una clara relación entre las dos variables, si bien se aprecia claramente que los vinos de mayor calidad tienden a presentar una menor cantidad de ácido acético (medido por la ***volatile.acidity***).

```
# Distribución bivariante de las variables alcohol y total.sulfur.dioxide
ggplot(data=wine, aes(x=alcohol, y=total.sulfur.dioxide, color=quality)) +
  geom_point()
```



Este gráfico de dispersión muestra que los vinos de mayor calidad suelen presentar una mezcla de alcohol en cantidad relativa grande y una cantidad relativa pequeña de dioxido de azufre.

```
# Distribución bivariante de las variables citric.acid y total.sulfur.dioxide
ggplot(data=wine, aes(x=citric.acid, y=total.sulfur.dioxide, color=quality)) +
  geom_point()
```



Finalmente, en este último caso observamos claramente como los vinos de calidad tienen una cantidad relativa pequeña de dioxido de azufre (con dos claras excepciones en la parte superior del gráfico), independientemente de la cantidad de ácido cítrico.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para comprobar la normalidad vamos a trabajar con el **test de Lilliefors** (una variante del test de **Kolmogorov-Smirnov** que asume que la media y la varianza son desconocidas, siendo especialmente desarrollado para testear la normalidad), en nuestro caso, del paquete **nortest**.

Partimos de las hipótesis H0 y H1, donde: H0: La muestra proviene de una distribución normal. H1: La muestra no proviene de una distribución normal.

El nivel de significancia con el que trabajaremos es de 0.05 (Alfa=0.05). Así, el criterio de decisión es el siguiente: Si P < Alfa Se rechaza H0 Si p >= Alfa No se rechaza H0

Analizamos la normalidad para todas las variables mediante un bucle con la función *lapply*:

```
# Automatizamos el test de Lilliefors para todas las variables del dataset
w_ntest <- lapply(wine, lillie.test)
w_ntest
```

```
## $fixed.acidity
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
```

```
## D = 0.1105, p-value < 2.2e-16
##
##
## $volatile.acidity
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.054662, p-value = 4.489e-12
##
##
## $citric.acid
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.06615, p-value < 2.2e-16
##
##
## $residual.sugar
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.26068, p-value < 2.2e-16
##
##
## $chlorides
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.25964, p-value < 2.2e-16
##
##
## $free.sulfur.dioxide
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.11124, p-value < 2.2e-16
##
##
## $total.sulfur.dioxide
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.12098, p-value < 2.2e-16
##
##
## $density
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  X[[i]]
## D = 0.044787, p-value = 6.252e-08
##
##
## $pH
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.040368, p-value = 2.244e-06
##
##
## $sulphates
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.12479, p-value < 2.2e-16
##
##
## $alcohol
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.12145, p-value < 2.2e-16
##
##
## $quality
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[[i]]
## D = 0.24982, p-value < 2.2e-16
```

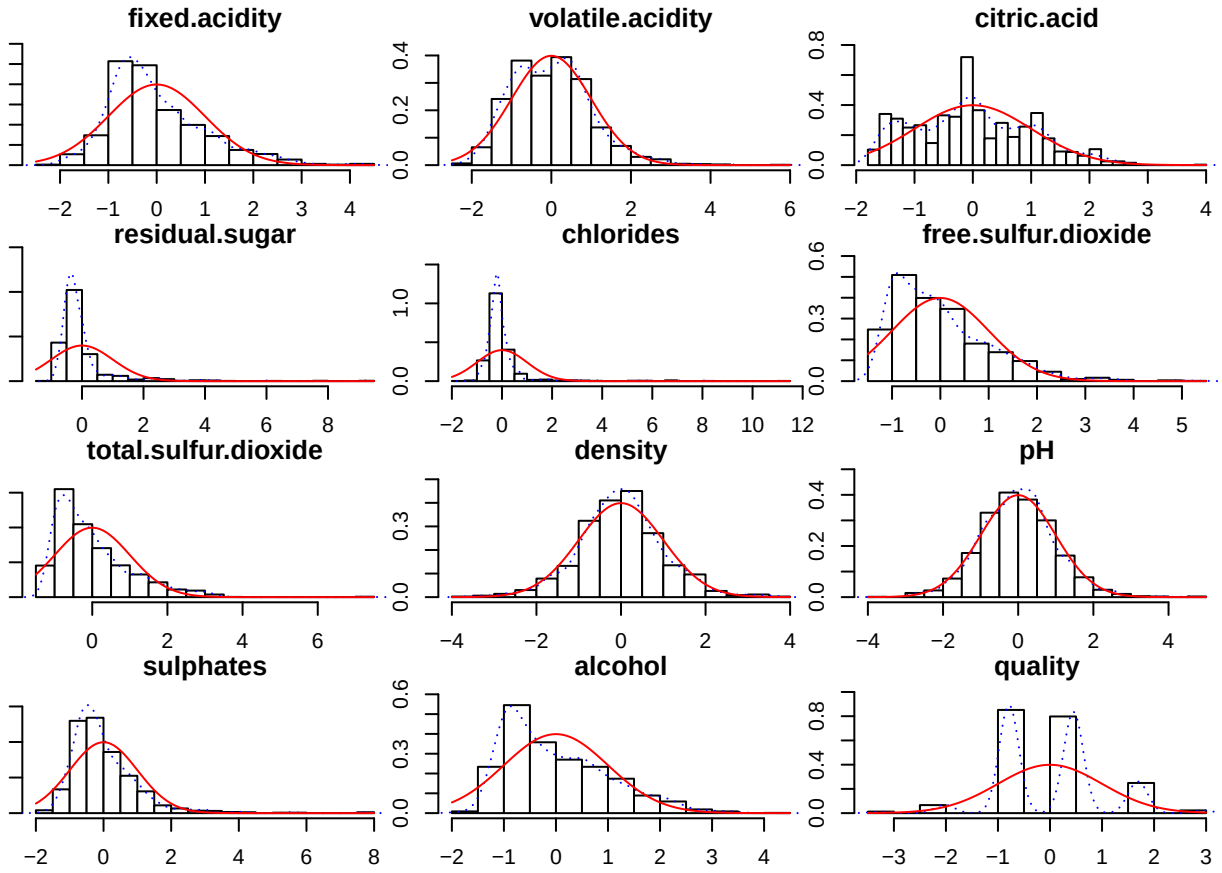
Para todas las variables, el valor de p se aproxima a cero (2.2e-16, que es uno de los resultados más comunes por ejemplo, sería la notación científica de 0.00000000000000022).

Dado que el valor de p es menor que 0.05 en todos los casos, rechazamos pata todas las variables la hipótesis H0 ('La muestra proviene de una distribución normal'); y concluimos, por tanto, que las variables no siguen una distribución normal.

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 contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Vamos a revisar la distribución de las variables mediante un histograma. Para ello vamos primero a escalar nuestros datos

```
wine_scaled <- data.frame(scale(wine))
multi.hist(x = wine_scaled, dcol = c("blue", "red"), dlty = c("dotted", "solid"))
```



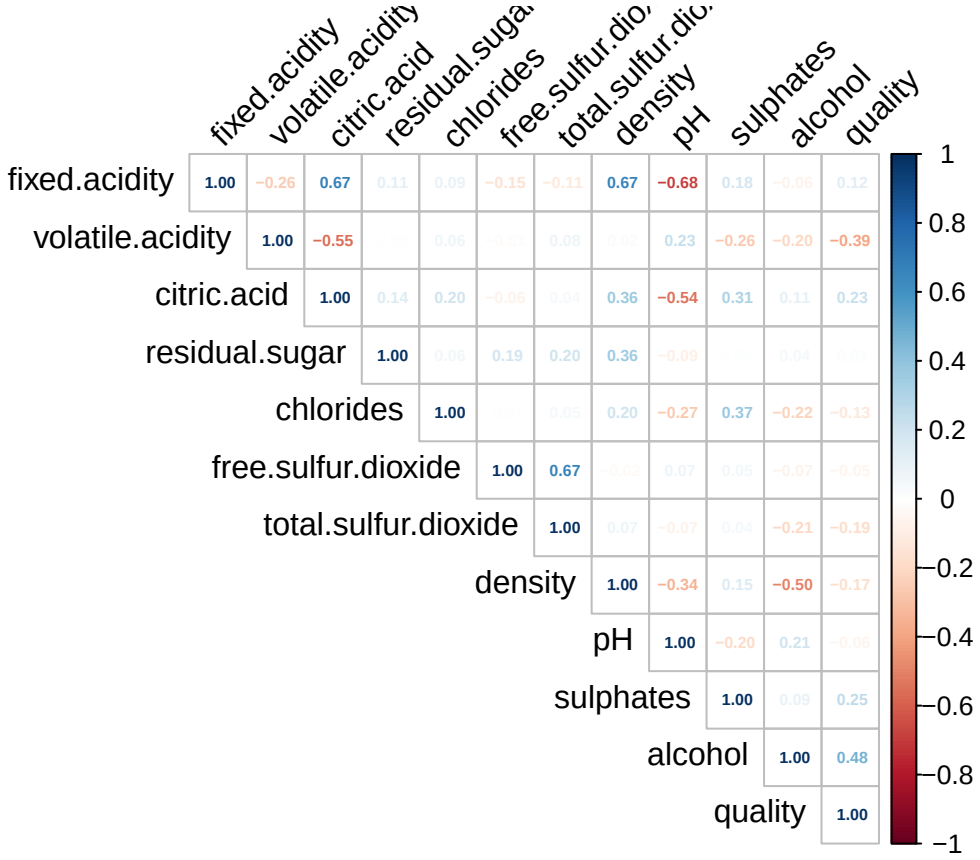
Como lo que nos interesa saber es la contribución de las variables independientes en la explicación de la vairable dependiente (***quality***), realizaremos un análisis de regresión lineal múltiple. Este análisis nos devolverá un coeficiente para cada variable que explica la variación de la variable dependiente como la suma de los efectos de las variables independients.

Además, como lo que nos interesa es realizar un modelo inferencial, y no solo descriptivo, debemos realizar una serie de pruebas que verifiquen que nuestras variables son adecuadas para el modelo que queremos conseguir.

1- La variable dependiente y los errores deben seguir una distribución normal 2- Nos encontramos ante un supuesto de homoscedasticidad; es decir, las varianzas de la variable dependiente son iguales para los diferentes valores de la variable independiente (homogenieddad de varianza) 3- No hay correlación entre las variables independientes 4- No hay colinealidad, es decir, que no hay ninguna variables que sea combinación lineal de otras.

Vamos a ir confirmando los puntos anteriores a lo largo de nuestro análisis.

```
# Vamos a analizar la correlación de nuestras variables
corrplot(correlation, type ="upper", tl.srt=45, number.cex=0.5, tl.col="black",
         method = 'number')
```



Si estudiamos la correlación entre las variables, podemos observar que:

Correlación positiva entre variable dependiente e independiente Existe una correlación positiva y relativamente alta (de 0.48), entre la ***quality*** del vino y el ***alcohol***. La siguiente correlación más alta y positica, sería con ***Sulphates***, con una correlación de 0.25. Y la tercera con ***citric.acid*** con una correlación de 0.23

Correlación negativa entre variable dependiente e independiente Si estudiamos la correlación negativa, podemos ver que hay una correlació negativa alta entre ***quality*** y ***volatile.acidity*** de -0.39.

Correlación entre vairables independientes Como ya hemos comentado, es importante que nuestras variables independientes tengan una correlación baja entre ellas, porque una correlación alta, puede influir negativamente en nuestro modelo.

Vemos que ***fixed.acidity*** y ***pH*** tienen una correlación negativa altísima, de hecho la más alta del moldeo con -0.68,***fixed.acidity*** y ***citric.acid*** en este caso positiva con un 0.67, ***volatile.acidity*** y ***citric.acid*** de -0.55 y ***free.sulfur.dioxide*** con ***total.sulfur.dioxide*** de 0.67, cosa que ya habíamos previsto en puntos anteriores.

Como ***ph*** tiene una correlación tan baja con nuestra variable dependiente,la excluiremos de nuestro modelo. Haremos lo mismo con ***free.sulfur.dioxide***, puesto que está incluida en ***total.sulfur.dioxide***, y con ***citric.acid***, puesto que está incluida en ***fixed.acidity*** y el hecho de incluirlas podría afectar negativamente al modelo.

Vamos a realizar nuestro modelo:

```
# Realizamos nuestro modelo excluyendo PH - FREE.SULFUR.DIOXIDE - CITRIC.ACID
mlm <- lm(quality ~ alcohol + sulphates + volatile.acidity + citric.acid +
            fixed.acidity + total.sulfur.dioxide + density + chlorides +
```

```
            residual.sugar,  data = wine )

summary(mlm)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     citric.acid + fixed.acidity + total.sulfur.dioxide + density +
##     chlorides + residual.sugar, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70523 -0.37004 -0.05798  0.43043  1.98341
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           46.706752  17.722735   2.635 0.008485 **
## alcohol                0.251043   0.022715  11.052  < 2e-16 ***
## sulphates              0.956911   0.113470   8.433  < 2e-16 ***
## volatile.acidity      -1.116825   0.111378 -10.027  < 2e-16 ***
## citric.acid           -0.264641   0.140605  -1.882 0.059997 .
## fixed.acidity          0.066273   0.017231   3.846 0.000125 ***
## total.sulfur.dioxide  -0.002214   0.000531  -4.169 3.23e-05 ***
## density              -44.174071  17.764168  -2.487 0.012996 *
## chlorides             -1.611678   0.406583  -3.964 7.70e-05 ***
## residual.sugar         0.028885   0.014119   2.046 0.040945 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6489 on 1589 degrees of freedom
## Multiple R-squared:  0.358,  Adjusted R-squared:  0.3544
## F-statistic: 98.46 on 9 and 1589 DF,  p-value: < 2.2e-16
```

De este primer modelo podemos extraer las siguientes conclusiones:

1) Dada la hipótesis nula de No relación entre las Variables Independientes y la Dependiente y viendo el p.valor de nuestro modelo, podemos ver que todas las variables que hemos incluido son relevantes para el modelo.
2) El R-cuadrado (R-Squared) de nuestro modelo es de 0.3534, lo que significa que nuestro modelo sólo explica un 35% de la varianza.

Utilizaremos el método ***Step Wise*** para que nos ayude a identificar si eliminando alguna variable, nuestro modelo mejora:

```
#Realizamos Step Wise para ver si vale la pena quitar alguna variable:

Step_wise <- step(mlm, direction = "both", trace = 1)
```

```
## Start:  AIC=-1373.17
## quality ~ alcohol + sulphates + volatile.acidity + citric.acid +
##     fixed.acidity + total.sulfur.dioxide + density + chlorides +
##     residual.sugar
```

```
##
##                         Df Sum of Sq    RSS     AIC
## <none>                                669.05 -1373.2
## - citric.acid            1     1.492 670.54 -1371.6
## - residual.sugar         1     1.762 670.81 -1371.0
## - density                1     2.604 671.65 -1369.0
## - fixed.acidity          1     6.228 675.28 -1360.3
## - chlorides              1     6.616 675.67 -1359.4
## - total.sulfur.dioxide   1     7.318 676.37 -1357.8
## - sulphates              1    29.945 698.99 -1305.2
## - volatile.acidity       1    42.336 711.39 -1277.1
## - alcohol                1    51.428 720.48 -1256.8
```

como vemos, no se nos recomienda eliminar ninguna variable del modelo.

Una vez hecho esto, nos interesa saber cuáles de nuestras variables son las que tienen más importancia en el modelo. Para ello, podemos hacer dos cosas:

1) Realizar el análisis de regresión con los datos escalados, para comparar los coeficientes más altos en valor absoluto.
2) Uitlizar la función calc.relimp para saber la importancia relativa de cada una de nuestras muestras.

```
# Modelo de regresión lineal con datos escalados:

mlm_scaled <- lm(quality ~ alcohol + sulphates + volatile.acidity + fixed.acidity +
                   total.sulfur.dioxide + density + chlorides + residual.sugar,
                 data = wine_scaled)
summary(mlm_scaled)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     fixed.acidity + total.sulfur.dioxide + density + chlorides +
##     residual.sugar, data = wine_scaled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3336 -0.4486 -0.0753  0.5627  2.4056
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.039e-15  2.011e-02   0.000 1.000000
## alcohol               3.215e-01  2.954e-02  10.882  < 2e-16 ***
## sulphates             1.996e-01  2.383e-02   8.378  < 2e-16 ***
## volatile.acidity     -2.288e-01  2.261e-02 -10.123  < 2e-16 ***
## fixed.acidity         1.118e-01  3.331e-02   3.357 0.000806 ***
## total.sulfur.dioxide -9.488e-02  2.150e-02  -4.413 1.09e-05 ***
## density              -1.041e-01  4.155e-02  -2.505 0.012332 *
## chlorides            -1.057e-01  2.287e-02  -4.622 4.11e-06 ***
## residual.sugar        4.883e-02  2.466e-02   1.980 0.047831 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.8041 on 1590 degrees of freedom
## Multiple R-squared:  0.3566, Adjusted R-squared:  0.3534
## F-statistic: 110.1 on 8 and 1590 DF,  p-value: < 2.2e-16
```

```r
# Importancia de las variablas independientes

calc.relimp(mlm, type = c("lmg"), rela = TRUE, rank = TRUE)
```

```
## Response variable: quality
## Total response variance: 0.6521684
## Analysis based on 1599 observations
##
## 9 Regressors:
## alcohol sulphates volatile.acidity citric.acid fixed.acidity total.sulfur.dioxide density chlorides
## Proportion of variance explained by model: 35.8%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##                              lmg
## alcohol              0.375034935
## sulphates            0.126127903
## volatile.acidity     0.230784600
## citric.acid          0.045670156
## fixed.acidity        0.042011002
## total.sulfur.dioxide 0.050327258
## density              0.079849543
## chlorides            0.041480241
## residual.sugar       0.008714362
##
## Average coefficients for different model sizes:
##
##                                1X           2Xs           3Xs           4Xs
## alcohol                0.360841765   0.353664145   0.342467705   0.329269822
## sulphates              1.197712323   1.163895043   1.124526536   1.088074326
## volatile.acidity      -1.761437780  -1.666588674  -1.550990777  -1.434631039
## citric.acid            0.959758374   0.866492841   0.724272841   0.550413179
## fixed.acidity          0.057538644   0.060622947   0.063734360   0.066479960
## total.sulfur.dioxide  -0.004544151  -0.004131247  -0.003717622  -0.003344923
## density              -74.846013601 -83.993268547 -86.678970319 -84.260826875
## chlorides             -2.211841716  -2.272887162  -2.252753056  -2.179587811
## residual.sugar         0.007865118   0.012010901   0.016972471   0.021590250
##                               5Xs           6Xs           7Xs           8Xs
## alcohol                0.315262942   0.300900960   0.285979552   0.269739528
## sulphates              1.058245648   1.034079236   1.011360213   0.985597469
## volatile.acidity      -1.331615975  -1.250021655  -1.191442598  -1.150663977
## citric.acid            0.361626412   0.173112861  -0.001910245  -0.151793772
## fixed.acidity          0.068498558   0.069539485   0.069494493   0.068391205
## total.sulfur.dioxide  -0.003026863  -0.002759705  -0.002534866  -0.002349146
## density              -78.172396948 -69.824651730 -60.548350368 -51.603624276
## chlorides             -2.076500819  -1.959127423  -1.835861932  -1.713595400
## residual.sugar         0.025237019   0.027659111   0.028866953   0.029103716
##                               9Xs
## alcohol                0.251042467
```

```
## sulphates              0.956911381
## volatile.acidity      -1.116824618
## citric.acid           -0.264641306
## fixed.acidity          0.066273008
## total.sulfur.dioxide  -0.002213525
## density              -44.174070615
## chlorides             -1.611678289
## residual.sugar         0.028884653
```

En ambos casos podemos comprobar que las variables que más influyen a nuestro modelo son ***alcohol***, ***sulphates*** y ***volatile.acidity***, mientras que la que menos aporta al modelo es ***residual.sugar*** como ya imaginabamos.

Ahora que sabemos qué variables influyen mas en la calidad del vino tinto, vamos a intentar averiguar por qué nuestro modelo explica un porcentaje tan bajo de la varianza.

Ya hemos analizado la correlación de las cariables independientes y hemos descartado variables que son combinación lineal de otras, pero aún nos queda revisar, la colinealidad, la distribución de las variables y los errores, y el supuesto de homoscedasticidad.

Para la revisión de la colinealidad, podemos utilizar el factor de inflación de la varianza (VIF). Valores bajos (entre 1 y 5) nos indican que podría verse afectada por la colinealidad, pero que no es motivo de preocupación.

```
# Colinealidad

require(car)
vif(mlm)
```

```
##              alcohol            sulphates     volatile.acidity
##             2.223883             1.404030             1.509514
##          citric.acid       fixed.acidity total.sulfur.dioxide
##             2.351222             3.415983             1.157819
##              density            chlorides       residual.sugar
##             4.266081             1.389767             1.504055
```

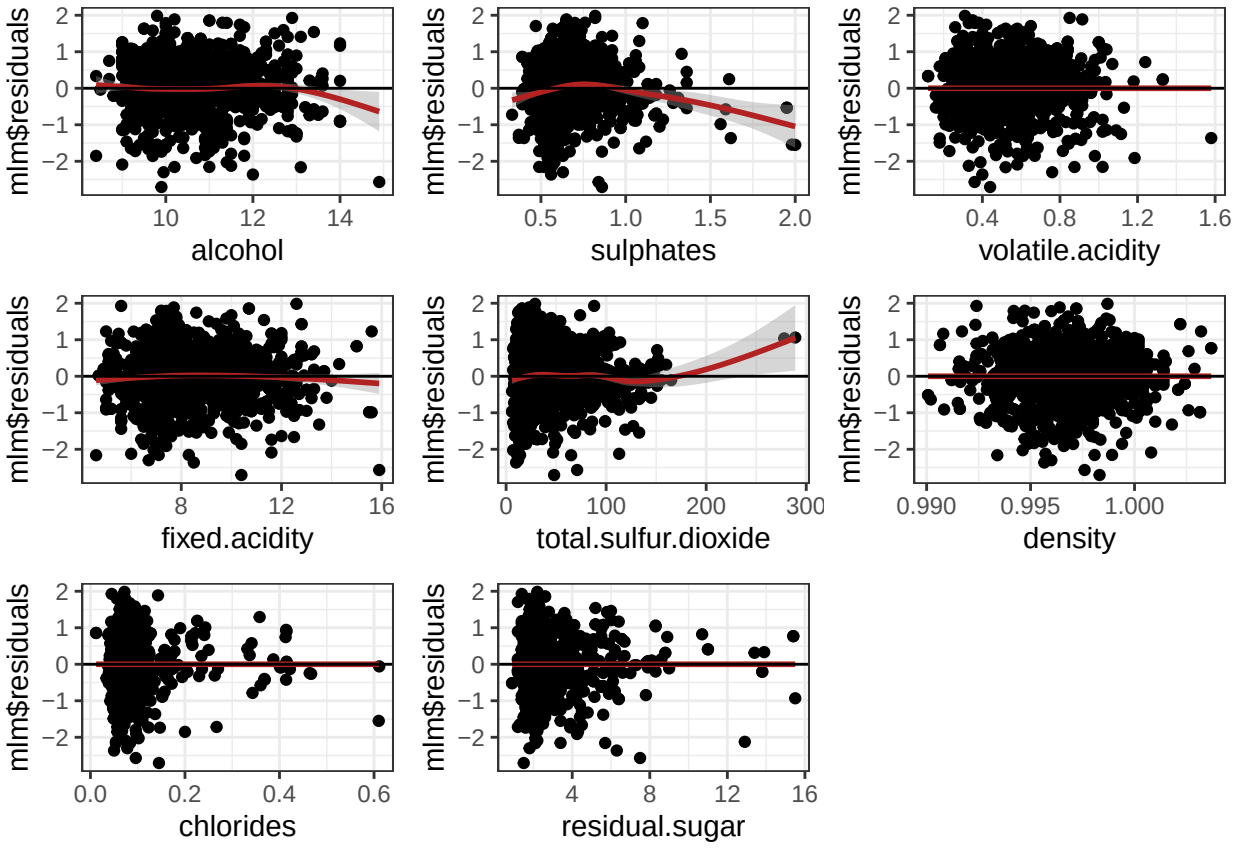
Como vemos, todos nuestros factores están entre 1 y 5, por lo que aceptaremos el supuesto de que no hay colinealidad.

Ahora pasemos a mirar la normalidad. Podemos observar la normalidad de los residuos de diferentes maneras.

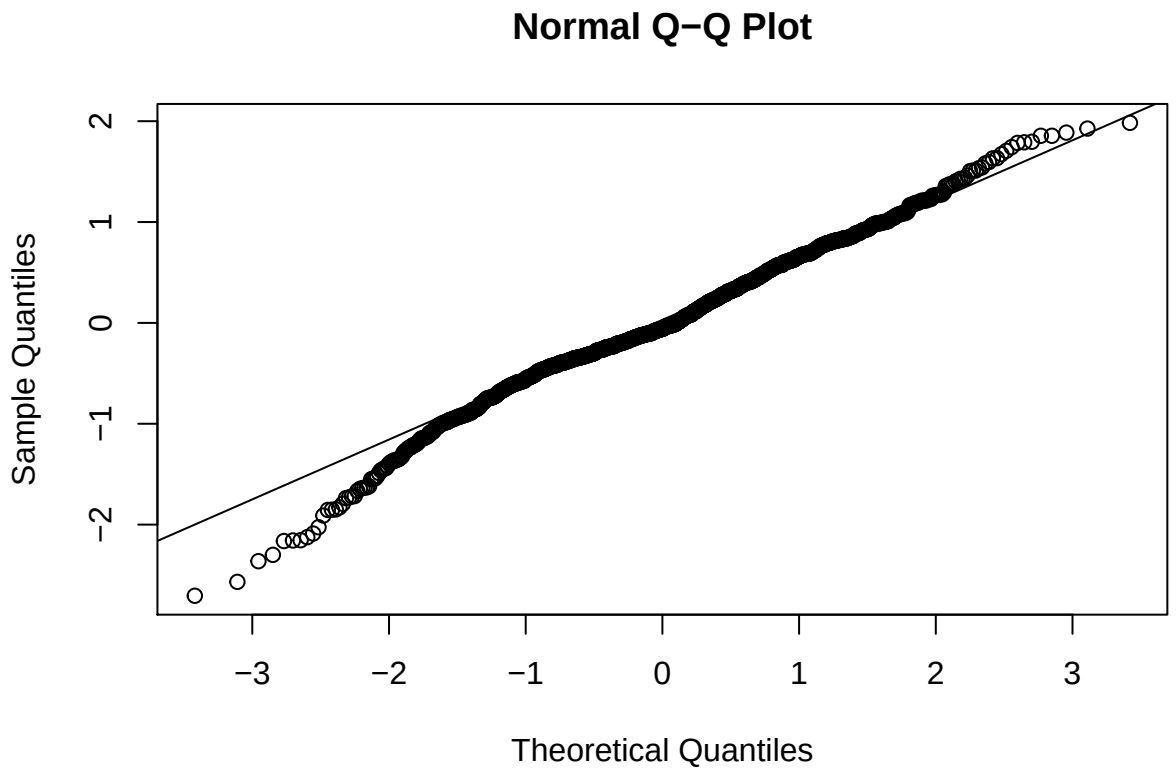
Una de las maneras es realizando un plot con los residuos de cada una de las variables:

```
# Análisis de la normalidad de los residuos:

plot1 <- ggplot(data = wine, aes(alcohol, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot2 <- ggplot(data = wine, aes(sulphates, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()

plot3 <- ggplot(data = wine, aes(volatile.acidity, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
```

```
plot4 <- ggplot(data = wine, aes(fixed.acidity, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()

plot5 <- ggplot(data = wine, aes(total.sulfur.dioxide, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()

plot6 <- ggplot(data = wine, aes(density, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()

plot7 <- ggplot(data = wine, aes(chlorides, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()

plot8 <- ggplot(data = wine, aes(residual.sugar, mlm$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()


grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6,plot7, plot8)
```



```r
qqnorm(mlm$residuals)
qqline(mlm$residuals)
```




```r
shapiro.test(mlm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mlm$residuals
## W = 0.98986, p-value = 4.23e-09
```

Como podemos observar tanto por las gráficas, como por el test de Shapiro-Wilk de normalidad de los residuos, es que no siguen una distribución normal. El test de Shapiro-Wilk, nos da un p-valor de 2.043e-09, lo que nos hace rechazar la hopotesis nula de normalidad.

Esto puede deberse a muchos factores, como por ejemplo, que las variables independientes que tenemos no son las únicas involucradas en la calidad del vino. Tampoco sabemos cómo se recogió la información de calidad, podría haber sido algo muy subjetivo y no relacionado a las variables que tenemos.

Uno de los problemas del incumplimiento de la hipótesis de normalidad en los residuos, suele ser la existencia de heteroscedasticidad, que implica que la precisión del modelo no es constante.

```r
# Test de Breusch-Pagan para la heteroscedasticidad

bptest(mlm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mlm
## BP = 78.637, df = 9, p-value = 3.013e-13
```

Vemos que el test de Breusch-Pagan nos da un valor de 1.735e-13, volviendo a rechazar la hipótesis nula de homoscedasticidad, y confirmando que nuestras variables independientes, no son las adecuadas para nuestro modelo.

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

Podemos ver todas las tablas y las gráficas introducidas a medida que explicamos el análisis.

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

Para el análisis de nuestro dataset, hemos optado por realizar una regresión lineal multiple, para ver si nuestras variables independientes tienen influencia en la explicación de nuestra variable dependiente (***quality***).

Tras el análisis inicial de nuestras variables independientes, y tras cambiar los valores 0 de ***citric.acid*** a la media de la variable, ya podíamos ver que varias de nuestras variables eran combinación lineal de otras, y por tanto las hemos excluido del análisis.

Estas son, el PH, por ser un medido de acidez y tener otras variables específicas de acidez; el ácido cítrico, por ser un acidificante, y ya teníamos otras variables como fixed.acidity que contemplaba el acído cítrico, y el free sulfur dioxide, porque ya teníamos el total.sulfur.dioxide que tiene una correlación mayor con la variable dependiente e incluye la primera.

Una vez sabemos qué variantes queremos introducir en nuestro modelo, generamos nuestro modelo de regresión lineal, y vemos que todas las variables introducidas son significativas. Las variables que más influyen a nuestro modelo son ***alcohol***, ***sulphates*** y ***volatile.acidity***, mientras que la que menos aporta al modelo es ***residual.sugar*** .A pesar de ello nuestro modelo sólo se explica un 35% de la varianza, que no es demasiado alta.

Podemos comprobar que nuestras variables independientes no son adecuadas, mediante los tests que hemos realizado. Hemos descartado un posible porblema de colinealidad, pero hemos identificado que los residuos no siguen una distribución normal y que nos encontramos ante un supuesto de heteroscedasticidad, es decir que no hay homogeneidad en la varianza.

Como conclusión, podemos decir que las variables independientes escogidas no determinan con claridadla calidad del vino, y esto puede deberse a que las variables escogidas, o no son las adecuadas, o son insuficientes. Además, como no sabemos cómo se ha obtenido el índice de calidad, podría deberse a que se han tenido otros factores en cuenta que no estamos analizando.

Por tanto, aunque nuestros resultados no son óptimos para indicar qué factores son los que determinan la calidad del vino, sí que podemos decir que de las variables que tenemos, el alcohol, los sulfatos y la acidez volátil, son los que más inflyen.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Todo el código de R empleado en este análisis se encuentra resumido en el archivo cleaning_script.R en el fichero https://github.com/jesusblay/red-wine-quality

Además exportamos el fichero con los datos finales que hemos usado:

```
wine_final <- wine[ ,c("alcohol","sulphates","volatile.acidity","fixed.acidity",
                       "total.sulfur.dioxide","density",
                      "chlorides","residual.sugar")]
write.csv(wine_final, "wine_final.csv")
```

Contribuciones

Hemos estado en contacto contínuo para cada una de las preguntas.

Investigación previa: EALS & JABT

Redacción de las respuestas: EALS & JABT

Desarrollo código: EALS & JABT