PRACTICA 2. ANÁLISIS DE LA CALIDAD DE VINOS ROJOS

Rubén Gómez Marquez y Andrés Felipe Bello Zapata

30/5/2020

# 1. DESCRIPCIÓN DEL DATASET

En esta primera parte visualizaremos el dataset y describiremos las variables que lo componen, tratando de entender que preguntas o problemas pueden ayudarnos a responder.

#CARGA DEL ARCHIVO  
redwine<-read.csv("redwine.csv",sep=",",na.string="NA",header = TRUE)  
#PRIMERA VISUALIZACIÓN  
head(redwine, max(10))

## fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1 7.4 0.70 0.00 1.9 0.076  
## 2 7.8 0.88 0.00 2.6 0.098  
## 3 7.8 0.76 0.04 2.3 0.092  
## 4 11.2 0.28 0.56 1.9 0.075  
## 5 7.4 0.70 0.00 1.9 0.076  
## 6 7.4 0.66 0.00 1.8 0.075  
## 7 7.9 0.60 0.06 1.6 0.069  
## 8 7.3 0.65 0.00 1.2 0.065  
## 9 7.8 0.58 0.02 2.0 0.073  
## 10 7.5 0.50 0.36 6.1 0.071  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 1 11 34 0.9978 3.51 0.56 9.4  
## 2 25 67 0.9968 3.20 0.68 9.8  
## 3 15 54 0.9970 3.26 0.65 9.8  
## 4 17 60 0.9980 3.16 0.58 9.8  
## 5 11 34 0.9978 3.51 0.56 9.4  
## 6 13 40 0.9978 3.51 0.56 9.4  
## 7 15 59 0.9964 3.30 0.46 9.4  
## 8 15 21 0.9946 3.39 0.47 10.0  
## 9 9 18 0.9968 3.36 0.57 9.5  
## 10 17 102 0.9978 3.35 0.80 10.5  
## quality  
## 1 5  
## 2 5  
## 3 5  
## 4 6  
## 5 5  
## 6 5  
## 7 5  
## 8 7  
## 9 7  
## 10 5

#ESTRUCTURA DEL DATASET  
str(redwine)

## 'data.frame': 1599 obs. of 12 variables:  
## $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...  
## $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...  
## $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...  
## $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...  
## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...  
## $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...  
## $ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...  
## $ density : num 0.998 0.997 0.997 0.998 0.998 ...  
## $ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...  
## $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...  
## $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...  
## $ quality : int 5 5 5 6 5 5 5 7 7 5 ...

## 1.1 Descripción de los atributos

En este data set se describen las propiedades químicas de un conjunto de vinos rojos procedentes del norte de Portugal. Estos vinos están clasificados por puntuaciones basadas en la opinión personal.

Todos los atributos del dataset son variables numéricas y continuas, excepto quality que es una variable tipo int, es decir, numérica discreta (enteros):

fixed.acidity: acidez fija. (ácido tartárico - g / dm^3) volatile.acidity: acidez volatil. (ácido acético - g / dm^3) citric.acid: acidez citrica.(g / dm^3) residual.sugar:azucar residual.(g / dm^3) chlorides:cloruros.(cloruro sódico - g / dm^3 ) free.sulfur.dioxide:dióxido de azufre libre. (mg / dm^3) total.sulfur.dioxide:dióxido de azufre total.(mg / dm^3) density:densidad.(g / cm^3) pH:ph. sulphates:sulfatos.(sulfato potásico - g / dm3) alcohol:alcohol.(% por volumen) quality: calidad del vino según una opinión personal 0-10.

## 1.2 Motivaciones

Este dataset puede ayudarnos a comprender, desde un punto de vista químico, que propiedades debe tener un vino para que guste a los consumidores. Preguntas que siguen la linea de: ¿qué grado de alcohol prefieren los consumidores?. Y de está forma con cualquier propiedad química del vino.

# 2. INTEGRACIÓN Y SELECCIÓN DE LOS DATOS DE INTERÉS A ANALIZAR

En esta etapa no haremos ningún tipo de integración de datos ya que solo tenemos este dataset con el cual trabajar.

## 2.1 Comprobación de la normalidad y homogeneidad de la varianza.

Previo a realizar la reducción de dimensionalidad es necesario verificar si cada una de las columnas presenta una distribución normal. Lo anterior con el fin de determinar que modelo estadístico implementar. Para ello se emplea el test Shapiro que, determina si una columna tiene una distribución normal basado en el valor de la variable p, la cual cuando asume valores superiores a 0.05 permite confirmar la hipotesis nula, concluyendo que los datos presentan una distribución normal.

Test Shapiro para la columna fixed acidity

shapiro.test(redwine$fixed.acidity)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$fixed.acidity  
## W = 0.94203, p-value < 2.2e-16

Test Shapiro para la columna volatile acidity

shapiro.test(redwine$volatile.acidity)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$volatile.acidity  
## W = 0.97434, p-value = 2.693e-16

Test Shapiro para la columna citric acid

shapiro.test(redwine$citric.acid)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$citric.acid  
## W = 0.95529, p-value < 2.2e-16

Test Shapiro para la columna residual sugar

shapiro.test(redwine$residual.sugar)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$residual.sugar  
## W = 0.56608, p-value < 2.2e-16

Test Shapiro para la columna chlorides

shapiro.test(redwine$chlorides)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$chlorides  
## W = 0.48425, p-value < 2.2e-16

Test Shapiro para la columna total sulfur dioxide

shapiro.test(redwine$total.sulfur.dioxide)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$total.sulfur.dioxide  
## W = 0.87322, p-value < 2.2e-16

Test Shapiro para la columna density

shapiro.test(redwine$density)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$density  
## W = 0.99087, p-value = 1.936e-08

Test Shapiro para la columna ph

shapiro.test(redwine$pH)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$pH  
## W = 0.99349, p-value = 1.712e-06

Test Shapiro para la columna sulphates

shapiro.test(redwine$sulphates)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$sulphates  
## W = 0.83304, p-value < 2.2e-16

Test Shapiro para la columna alcohol

shapiro.test(redwine$alcohol)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$alcohol  
## W = 0.92884, p-value < 2.2e-16

Test Shapiro para la columna quality

shapiro.test(redwine$quality)

##   
## Shapiro-Wilk normality test  
##   
## data: redwine$quality  
## W = 0.85759, p-value < 2.2e-16

Se evidencia que todas las columnas tienen valores inferiores a 0.05 por lo tanto se rechaza la hipotesis nula y se confirma que nignuna las columnas relacionadas presentan una distribución normal.

Se aplica un modelo de PCA (Principal Component Analysis) con el fin de determinar si es viable realizar un proceso de reducción de datos, de tal forma que se obtenga un conjunto de menos dimensiones. Lo anterior considerando que este modelo no supervisado puede ser aplicado a datos que no presentan una distribución normal, como se identificó en el análisis de Shapiro realizada a cada una de las columnas. Inicialmente se aplica el metodo y se realiza un proceso de escala para todas las variables del dataset, de tal forma que los valores de las columnas queden en el rango de 0 a 1.

El metodo rotation permite identificar que efectivamente se genera un componente principal por cada una de las variables, relacionando en cada una de ellas el porcentaje de participación que tendría cada una de las variables.

pca <- prcomp(redwine, scale = TRUE)  
pca$rotation

## PC1 PC2 PC3 PC4  
## fixed.acidity 0.487883358 -0.004173212 0.16482854 0.231098077  
## volatile.acidity -0.265128984 0.338967858 0.22708884 -0.041858245  
## citric.acid 0.473335467 -0.137358104 -0.10022856 0.056735802  
## residual.sugar 0.139154423 0.167736336 -0.24362014 0.383037581  
## chlorides 0.197426792 0.189788185 0.02660785 -0.654777820  
## free.sulfur.dioxide -0.045880713 0.259483136 -0.61611132 0.033711483  
## total.sulfur.dioxide 0.004066746 0.363971374 -0.54073214 0.028459726  
## density 0.370301191 0.330780789 0.16872267 0.200693412  
## pH -0.432720849 -0.065440145 -0.06977056 0.005466181  
## sulphates 0.254535354 -0.109333620 -0.21291324 -0.560502367  
## alcohol -0.073176777 -0.502708647 -0.22497138 0.091701428  
## quality 0.112488776 -0.473166214 -0.22336929 0.036669226  
## PC5 PC6 PC7 PC8 PC9  
## fixed.acidity -0.07877938 0.05553130 -0.3072150 0.20052866 -0.17457815  
## volatile.acidity 0.29937933 0.29728700 -0.6262337 0.14612614 -0.06022334  
## citric.acid -0.12014871 0.13663328 0.2441486 0.29633271 -0.22097505  
## residual.sugar 0.70936319 0.10931059 0.2838543 -0.17062614 0.27818728  
## chlorides 0.26623723 0.33733656 0.2305470 -0.18692254 -0.41993639  
## free.sulfur.dioxide -0.15941286 -0.04264807 -0.1382604 -0.01935607 -0.31800012  
## total.sulfur.dioxide -0.21845284 0.11595360 -0.1102087 0.08989655 0.12182276  
## density 0.20879298 -0.42566742 -0.1225465 0.07950023 -0.24907449  
## pH 0.25764682 -0.48035396 0.1856917 0.31469303 -0.46191598  
## sulphates 0.21483493 -0.40374303 -0.2334021 0.27549158 0.45268884  
## alcohol 0.25972635 0.39217625 -0.1217188 0.47118865 -0.09652795  
## quality 0.13758414 -0.14183046 -0.4123879 -0.61224719 -0.24024309  
## PC10 PC11 PC12  
## fixed.acidity 0.182956014 -0.256437921 0.638579761  
## volatile.acidity -0.155105626 0.377161229 0.004661681  
## citric.acid -0.346085556 0.624327833 -0.070036908  
## residual.sugar 0.052236558 0.088077871 0.183646374  
## chlorides 0.003862734 -0.208616670 0.053931176  
## free.sulfur.dioxide 0.585388580 0.237933171 -0.051921666  
## total.sulfur.dioxide -0.589188239 -0.355046842 0.069792953  
## density -0.043538098 -0.231453058 -0.566644992  
## pH -0.207609889 -0.005599072 0.341230056  
## sulphates 0.071918570 0.097637028 0.067792979  
## alcohol 0.110605247 -0.319948696 -0.317640325  
## quality -0.260239788 0.052465707 0.008470476

Una vez se visualizan los componentes principales se procede a verificar la desviación estandar con el fin de identificar que efectivamente cada una de las columnas del PCA presenta una desviación estandar menor. Lo anterior indica que el modelo es aplicado de manera correcta abarcando en cada componente prinicipal una mayor cantidad de datos.

print(pca$sdev)

## [1] 1.7666827 1.4972916 1.2972739 1.1022799 0.9865412 0.8139977 0.7863319  
## [8] 0.7112472 0.6413326 0.5726425 0.4245216 0.2439629

Finalmente se analiza la proporción acumulada con el fin de determinar el nivel de precisión que nos daria el empleo de cada uno de los componentes principales y determinar si es viable reducir la dimensionalidad de los datos.

summary(pca)

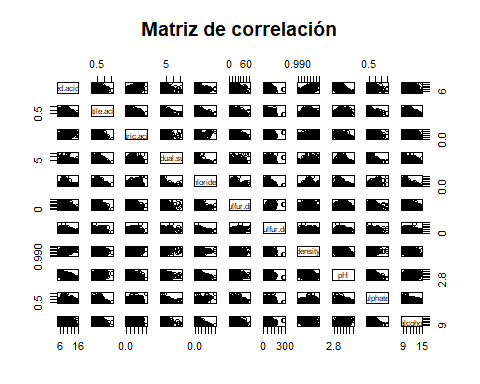
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.7667 1.4973 1.2973 1.1023 0.98654 0.81400 0.78633  
## Proportion of Variance 0.2601 0.1868 0.1402 0.1013 0.08111 0.05522 0.05153  
## Cumulative Proportion 0.2601 0.4469 0.5872 0.6884 0.76952 0.82474 0.87626  
## PC8 PC9 PC10 PC11 PC12  
## Standard deviation 0.71125 0.64133 0.57264 0.42452 0.24396  
## Proportion of Variance 0.04216 0.03428 0.02733 0.01502 0.00496  
## Cumulative Proportion 0.91842 0.95270 0.98002 0.99504 1.00000

Del anterior análisis estadístico se logra identificar que mediante el uso de Componente Principales, es posible eliminar una o máximo 2 columnas de la base de datos, manteniendo un nivel presición del 98%, siendo esta una reducción no tan cosiderable, por lo tanto se toma la determinación de analizar el impacto que tiene cada una de las variables independientes en la variable dependiente de QUALITY ques es finalmente la que se desea analizar. De esta forma es mas viable descartar las variables que no influyen de manera considerable en la calidad del vino.

## 2.2 Selección de datos

Por tanto, para realizar la selección de los datos analizaremos la correlación entre las variables con el fin de ver cuales están relacionadas entre sí.De esta forma podremos observar que variables mantienen una mayor relación con la variable calidad, que es la variable que nos interesa analizar y podremos hacer una reducción de variables, quedandonos unicamente con las variables que más contribuyan.

#Mostramos la matriz de dispersión para visualizar la linealidad entre las variables  
pairs(fixed.acidity ~ volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, data = redwine, main="Matriz de correlación", size= 20)



#Mostramos la matriz de correlaciones que nos indica las relaciones entre las diferentes variables.  
round(cor(redwine),2)

## fixed.acidity volatile.acidity citric.acid residual.sugar  
## fixed.acidity 1.00 -0.26 0.67 0.11  
## volatile.acidity -0.26 1.00 -0.55 0.00  
## citric.acid 0.67 -0.55 1.00 0.14  
## residual.sugar 0.11 0.00 0.14 1.00  
## chlorides 0.09 0.06 0.20 0.06  
## free.sulfur.dioxide -0.15 -0.01 -0.06 0.19  
## total.sulfur.dioxide -0.11 0.08 0.04 0.20  
## density 0.67 0.02 0.36 0.36  
## pH -0.68 0.23 -0.54 -0.09  
## sulphates 0.18 -0.26 0.31 0.01  
## alcohol -0.06 -0.20 0.11 0.04  
## quality 0.12 -0.39 0.23 0.01  
## chlorides free.sulfur.dioxide total.sulfur.dioxide density  
## fixed.acidity 0.09 -0.15 -0.11 0.67  
## volatile.acidity 0.06 -0.01 0.08 0.02  
## citric.acid 0.20 -0.06 0.04 0.36  
## residual.sugar 0.06 0.19 0.20 0.36  
## chlorides 1.00 0.01 0.05 0.20  
## free.sulfur.dioxide 0.01 1.00 0.67 -0.02  
## total.sulfur.dioxide 0.05 0.67 1.00 0.07  
## density 0.20 -0.02 0.07 1.00  
## pH -0.27 0.07 -0.07 -0.34  
## sulphates 0.37 0.05 0.04 0.15  
## alcohol -0.22 -0.07 -0.21 -0.50  
## quality -0.13 -0.05 -0.19 -0.17  
## pH sulphates alcohol quality  
## fixed.acidity -0.68 0.18 -0.06 0.12  
## volatile.acidity 0.23 -0.26 -0.20 -0.39  
## citric.acid -0.54 0.31 0.11 0.23  
## residual.sugar -0.09 0.01 0.04 0.01  
## chlorides -0.27 0.37 -0.22 -0.13  
## free.sulfur.dioxide 0.07 0.05 -0.07 -0.05  
## total.sulfur.dioxide -0.07 0.04 -0.21 -0.19  
## density -0.34 0.15 -0.50 -0.17  
## pH 1.00 -0.20 0.21 -0.06  
## sulphates -0.20 1.00 0.09 0.25  
## alcohol 0.21 0.09 1.00 0.48  
## quality -0.06 0.25 0.48 1.00

Nos fijamos en la variable que realmente nos interesa para saber que vinos son los mejor valorados, que es la calidad. Esta variable presenta una mayor correlación con la acidez volatil y cítrica, los sulfatos y en mayor medida con el alcohol. Por tanto, estas serán las variables que analizaremos.

#Dividimos el dataset en 2. Por una parte las variables que queremos analizar y por otra parte la variable predictora o analítica (quality) que nos servirá para analizar los vinos.  
variables <- redwine[c(2,3,10,11)]  
calidad<-redwine[12]

# 3. LIMPIEZA DE DATOS

## 3.1 Ceros o elementos vacíos.

Analizaremos si el dataset contiene ceros o elementos vacios.

summary(variables)

## volatile.acidity citric.acid sulphates alcohol   
## Min. :0.1200 Min. :0.000 Min. :0.3300 Min. : 8.40   
## 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.:0.5500 1st Qu.: 9.50   
## Median :0.5200 Median :0.260 Median :0.6200 Median :10.20   
## Mean :0.5278 Mean :0.271 Mean :0.6581 Mean :10.42   
## 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.:0.7300 3rd Qu.:11.10   
## Max. :1.5800 Max. :1.000 Max. :2.0000 Max. :14.90

summary(calidad)

## quality   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.636   
## 3rd Qu.:6.000   
## Max. :8.000

Podemos ver con la función summary un resumen de las variables que queremos analizar. Observamos que no hay elementos vacios (NA’s) y a su vez que tenemos ceros en la variable citric.acid.

En caso de encontrarnos con elementos vacíos, se puede proceder de varias formas:

* Eliminando las filas que contienen los valores vacíos. Esto puede ser útil y ahorrarnos bastante trabajo pero solo si tenemos muchas filas y pocos elementos vacíos, de forma que no alteremos la estadística.
* Sustituir los valores vacíos por nulos (NA).
* Aproximarlos por métodos de estadísticos, por ejemplo, sustituirlos por la media. Hay otros métodos como el método de los k vecinos más próximos (k-neighbors), que tratan de rellenar el valor perdido en función de los valores que se encuentran más cerca en el dataset.

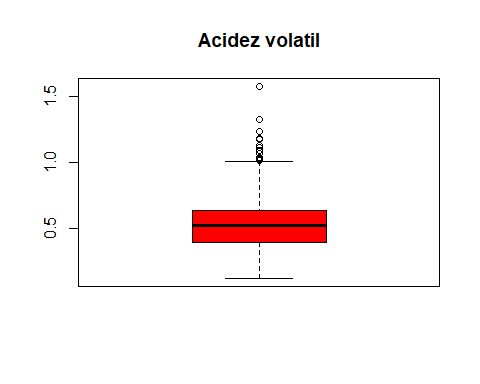
Si nos encontrásemos con ceros, habría que interpretar para cada variable si tiene sentido o no. En nuestro caso la variable que indica el ácido cítrico es numérica y continua por lo que en este aspecto si que podríamos tener ceros. Por otra parte en <https://www.lavinoteca.info/que-es-la-acidez> nos indican que el ácido cítrico es poco abundante en la uva y que despues de la fermentación este desaparece, por tanto si tiene sentido encontrarnos ceros en esta variable.

## 3.2 Identificación y tratamiento de valores extremos.

Analizaremos las variables cuantitativas para detectar valores extremos, mediante la generación de diagramas de caja y la función boxplot.stats.

### ÁCIDO VOLATIL (<http://urbinavinos.blogspot.com/2011/08/acidez-volatil-en-vinos.html>)

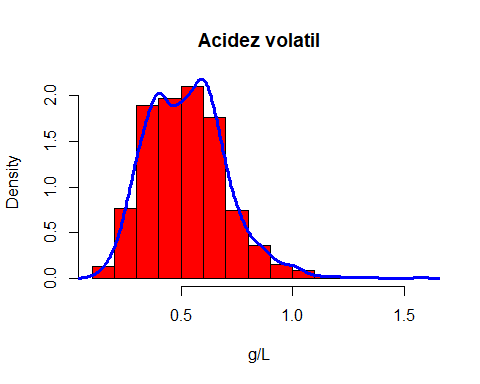
#Boxplot para obtener los valores extremos  
boxplot(variables$volatile.acidity,main="Acidez volatil",col="red")



boxplot.stats(variables$volatile.acidity)$out

## [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035  
## [13] 1.025 1.115 1.020 1.020 1.580 1.180 1.040

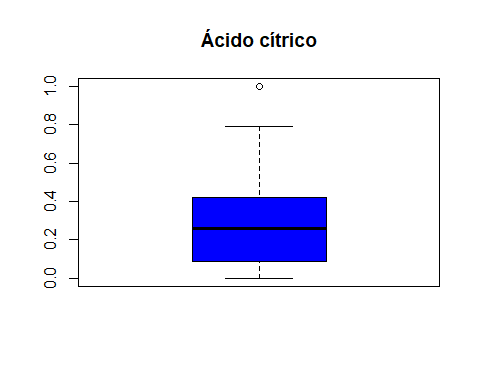
#Diagrama de densidad para observar la distribución  
den.vol<-density(variables$volatile.acidity)  
hist(variables$volatile.acidity, main="Acidez volatil", col="red", freq = FALSE, xlab="g/L")  
lines(den.vol,col="blue",lwd=3)



Según el blog relacionado en el link (parte inicial del enunciado), que nos habla de la acidez volatil en los vinos, la concentración de ácido acético debe encontrarse entre los 0.2 - 0.6 g/L, ademas cuando esta concentración supera el 1.4 g/L se considera que el vino se ha estropeado.Por tanto, dentro de los valores extremos que tenemos solo podemos considerar que uno de ellos se ha estropeado. De todas formas, observando la distribución vemos que la media esta dentro de lo esperado y sigue una distribución normal, por tanto no es necesario excluir estos datos porque no afectan a la estadística.

### ÁCIDO CÍTRICO (<https://www.vinosycavasonline.es/los-aacidos-del-vino.-quae-son-y-caomo-se-forman>.)

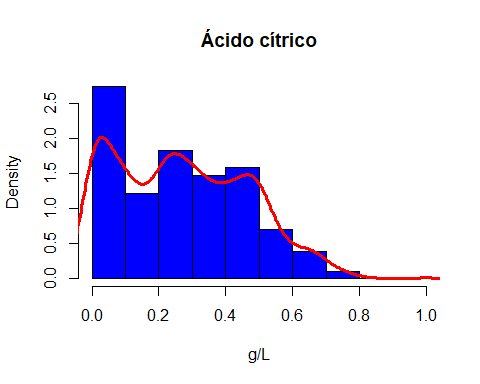
#Boxplot para obtener los valores extremos  
boxplot(variables$citric.acid,main="Ácido cítrico",col="blue")



boxplot.stats(variables$citric.acid)$out

## [1] 1

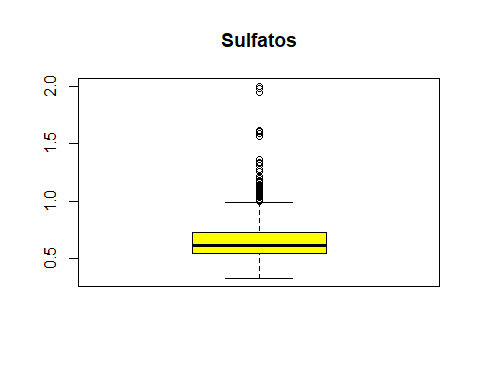
#Diagrama de densidad para observar la distribución  
den.cit<-density(variables$citric.acid)  
hist(variables$citric.acid, main="Ácido cítrico", col="blue", freq = FALSE, xlab="g/L")  
lines(den.cit,col="red",lwd=3)



Solo encontramos un valor (g/L) de ácido cítrico que se podría considerar como extremo. Según la página relacionada en el link (parte inicial del enunciado), el ácido cítrico se puede encontrar hasta concentraciones de 1 g/L, por lo cual no es un valor erroneo que podemos incluir en la estadística, ya que al ser solo uno no afectará. En cuanto a la distribución, no corresponde a una distribución normal, pero sus valores se encuentran todos dentro de lo esperado (0-1 g/L).

### SULFATOS

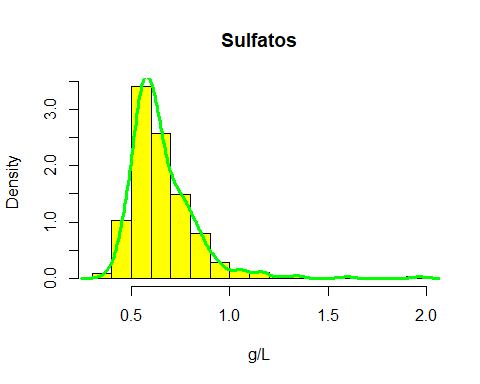
#Boxplot para obtener los valores extremos  
boxplot(variables$sulphates,main="Sulfatos",col="yellow")



boxplot.stats(variables$sulphates)$out

## [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59  
## [16] 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06  
## [31] 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18  
## [46] 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01

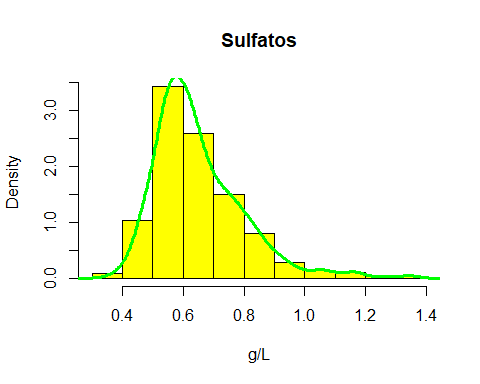
#Diagrama de densidad para observar la distribución  
den.sulf<-density(variables$sulphates)  
hist(variables$sulphates, main="Sulfatos", col="yellow", freq = FALSE, xlab="g/L")  
lines(den.sulf,col="green",lwd=3)



#Observamos los cuantiles  
x<-quantile(variables$sulphates)  
x

## 0% 25% 50% 75% 100%   
## 0.33 0.55 0.62 0.73 2.00

#Eliminamos los outliers superiores a la mitad del ultimo cuantil y creamos un nuevo dataset con estos datos  
redwine$quality[redwine$sulphates>(x[5]+x[4])/2]<-NA  
redwine$sulphates[redwine$sulphates>(x[5]+x[4])/2]<-NA  
data.sulfatos<-redwine[c(10,12)]  
data.sulfatos<-data.sulfatos[!is.na(data.sulfatos$quality),]  
  
#Diagrama sin outliers  
den.sulf<-density(data.sulfatos$sulphates)  
hist(data.sulfatos$sulphates, main="Sulfatos", col="yellow", freq = FALSE, xlab="g/L")  
lines(den.sulf,col="green",lwd=3)



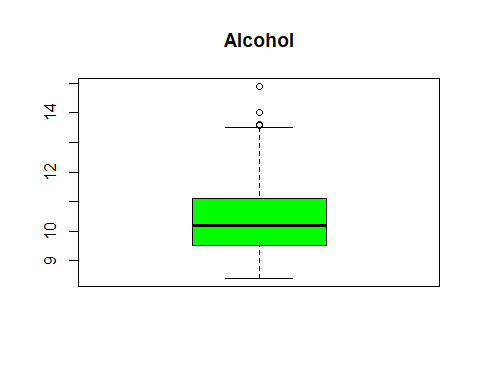
#Número de outliers eliminados  
sum(is.na(redwine$sulphates))

## [1] 8

En este caso nos encontrabamos con una gran presencia de valores extremos que si podría afectar a la estadística. Por tanto, se eliminan los que tienen valores superiores a la mitad del último percentil. De esta forma reducimos la influencia de los valores extremos, pero a su vez también los tenemos en cuenta, ya que las puntuaciones son opiniones personales y vinos con una mayor cantidad de concentración de sulfato podrían entrar dentro de los gustos de un tipo de público. Se eliminan 8 outliers correspondientes a los casos más extremos.

### ALCOHOL

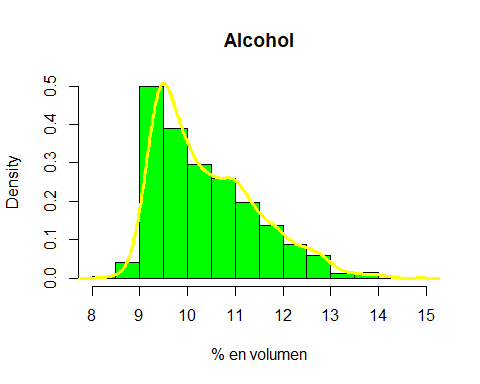
#Boxplot para obtener los valores extremos  
boxplot(variables$alcohol,main="Alcohol",col="green")



boxplot.stats(variables$alcohol)$out

## [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000  
## [9] 13.60000 14.00000 14.00000 13.56667 13.60000

#Diagrama de densidad para observar la distribución  
den.al<-density(variables$alcohol)  
hist(variables$alcohol, main="Alcohol", col="green", freq = FALSE, xlab="% en volumen")  
lines(den.al,col="yellow",lwd=3)



En el caso del alcohol tenemos muy pocos valores extremos, por lo que podemos incluirlos, ya que tampoco son valores incoherentes.

# 4. ANÁLISIS DE DATOS

## 4.1 Selección de los datos y planificación de análisis

Dividiremos el análisis en tres partes sobre las variables que anteriormente hemos seleccionado. El objetivo y la respuesta a contestar es el tratar de aclarar que tipo de vinos son aquellos con la mayor calidad. Para ello analizaremos esta variable, la calidad, frente a la composición química de los vinos, concretamente, la acidez volátil, el ácido cítrico, los sulfatos y el alcohol. Los análisis a realizar son:

1. Regresión lineal simple. Trataremos de ajustar a la ecuación de una recta los datos de la calidad en función de los componentes químicos del vino. La pendiente nos dirá si debemos aumentar o reducir dicho componente para mejorar su calidad. Veremos la realineación que hay entre las diferentes variables que tenemos y la puntuación que se le hace a los vinos.Para ello haremos una regresión lineal con el fin de observar los gustos de la gente.
2. Intervalos de confianza. Calcularemos la media y los intervalos de confianza de nuestras variables para poder realizar resoluciones de contrastes de hipótesis bilaterales al 95% de confianza. Preguntas del estilo de ¿Puede ser la media de la calidad de una muestra 8?.
3. Arbol de decisión: Se implementa un modelo supervisado de clasificación de tipo C50, con el fin de identificar las propiedades de las variables seleccionadas que son características de los vinos de alta, media y baja calidad, de acuerdo a la calificación de las diferentes muestras de vino, las cuales se encuentran agrupadas en la variable QUALITY.

### 4.2.1 Regresión lineal simple (variable dependiente calidad)

Haremos las rectas de regresión de las diferentes variables frente a la variable dependiente calidad. De esta forma obtenemos un primer modelo que nos permite obtener la calidad a partir del resto de variables.

Para saber si nuestro modelo es viable primero haremos un gráfico para analizar los valores ajustados frente a los residuos. En este gráfico se representan en el eje de ordenadas el valor de los residuos y en el de abcisas los valores ajustados, de forma que si el modelo es bueno no se debería apreciar ninguna estructura y los valores serían cercanos al eje de abcisas.

Por último con un gráfico QQ podemos ver el efectos de los valores atípicos.

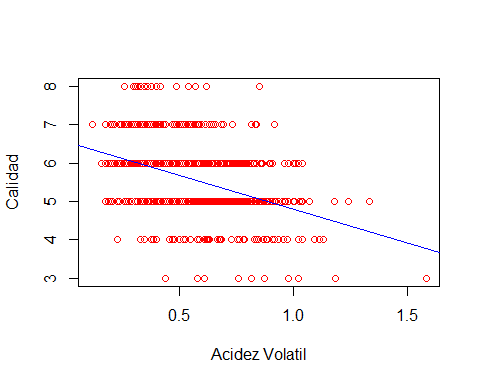
#LIBRERIAS PARA LA REPRESENTACIÓN  
library(ggplot2)

#### ACIDEZ VOLATIL

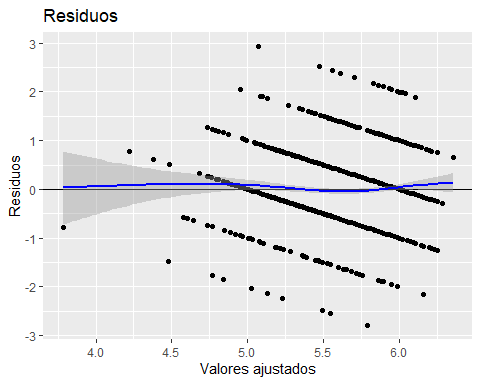
#AJUSTE DE LOS DATOS A LA ECUACIÓN y=ax+b  
lineal<-lm(calidad$quality~volatile.acidity,data=variables)  
summary(lineal)

##   
## Call:  
## lm(formula = calidad$quality ~ volatile.acidity, data = variables)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.79071 -0.54411 -0.00687 0.47350 2.93148   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.56575 0.05791 113.39 <2e-16 \*\*\*  
## volatile.acidity -1.76144 0.10389 -16.95 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7437 on 1597 degrees of freedom  
## Multiple R-squared: 0.1525, Adjusted R-squared: 0.152   
## F-statistic: 287.4 on 1 and 1597 DF, p-value: < 2.2e-16

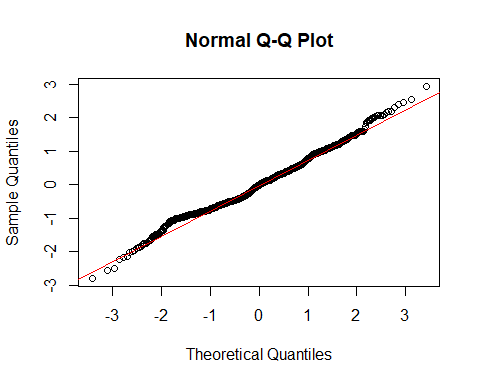
#REPRESENTACIÓN GRÁFICA  
plot(variables$volatile.acidity,calidad$quality, xlab = "Acidez Volatil", ylab = "Calidad", col="red")  
#RECTA DE REGRESIÓN  
abline(lineal, col="blue")



#ANÁLISIS POR RESIDUOS  
ggplot(data=variables, aes(lineal$fitted.values,lineal$residuals)) + geom\_point() + geom\_hline(yintercept = 0) + geom\_smooth(color = "blue") + labs(title = "Residuos", x="Valores ajustados", y="Residuos")



#FUNCION QQNORM PARA REPRESENTAR EL GRÁFICO QUANTIL QUANTIL  
qqnorm(lineal$residuals)  
#LINEA DE TENDENCIA  
qqline(lineal$residuals, col="red")

 La recta se ajusta a los siguientes valores: y = -1.76144x + 6.56575

En el análisis por residuos vemos que los datos presentan una estructura definida, lo cual nos indica que el modelo lineal no se ajusta bien a los datos. Tambien podemos observar este detalle con el valor de R cuadrado, que es de 0.1525, muy bajo.

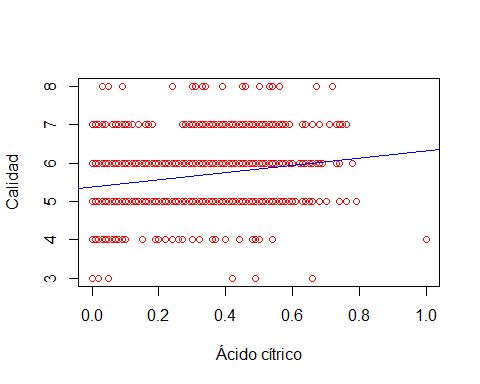
Vemos que el el gráfico QQ que tenemos valores atípicos y que estos, cuanto más se alejan de la normal, más desajustan los datos. En los primeros cuantiles si se observa normalidad.

#### ÁCIDO CÍTRICO

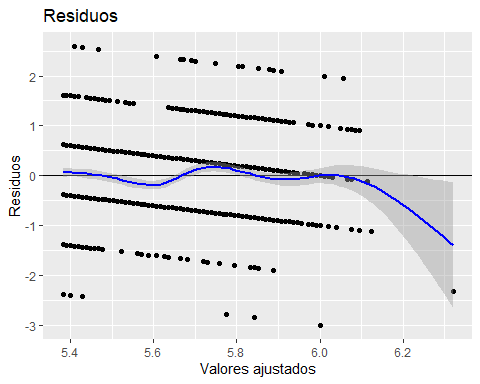
#AJUSTE DE LOS DATOS A LA ECUACIÓN y=ax+b  
lineal<-lm(calidad$quality~citric.acid,data=variables)  
summary(lineal)

##   
## Call:  
## lm(formula = calidad$quality ~ citric.acid, data = variables)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0011 -0.5976 0.1021 0.5057 2.5901   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.38172 0.03372 159.610 <2e-16 \*\*\*  
## citric.acid 0.93845 0.10104 9.288 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7869 on 1597 degrees of freedom  
## Multiple R-squared: 0.05124, Adjusted R-squared: 0.05065   
## F-statistic: 86.26 on 1 and 1597 DF, p-value: < 2.2e-16

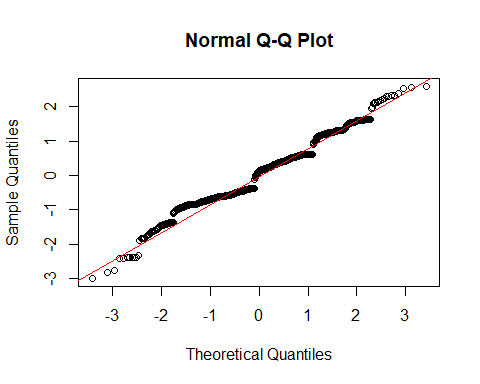
#REPRESENTACIÓN GRÁFICA  
plot(variables$citric.acid,calidad$quality, xlab = "Ácido cítrico", ylab = "Calidad", col="red")  
#RECTA DE REGRESIÓN  
abline(lineal, col="blue")



#ANÁLISIS POR RESIDUOS  
ggplot(data=variables, aes(lineal$fitted.values,lineal$residuals)) + geom\_point() + geom\_hline(yintercept = 0) + geom\_smooth(color = "blue") + labs(title = "Residuos", x="Valores ajustados", y="Residuos")



#FUNCION QQNORM PARA REPRESENTAR EL GRÁFICO QUANTIL QUANTIL  
qqnorm(lineal$residuals)  
#LINEA DE TENDENCIA  
qqline(lineal$residuals, col="red")

 La recta se ajusta a los siguientes valores: y = 0.93845x + 5.38172

En el análisis por residuos vemos que los datos presentan una estructura definida, lo cual nos indica que el modelo lineal no se ajusta bien a los datos de la misma forma que podemos ver el efecto de algunos valores atípicos. Tambien podemos observar este detalle con el valor de R cuadrado, que es de 0.05065, practicamente cero.

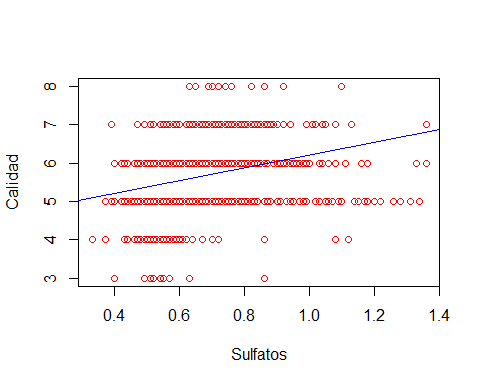
Vemos que el el gráfico QQ no se ajusta bien en ningún cuantil, por lo que indica que la variable no es normal.

#### SULFATOS

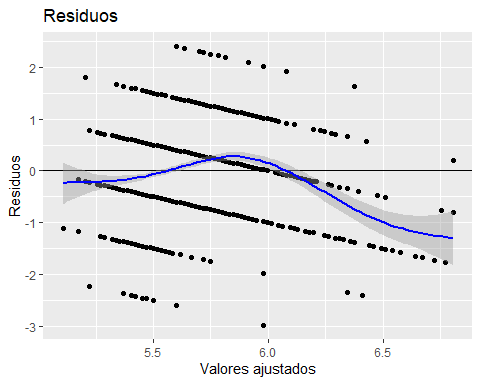
#Nuevo dataset para la representacion  
  
#AJUSTE DE LOS DATOS A LA ECUACIÓN y=ax+b  
lineal<-lm(quality~sulphates,data=data.sulfatos)  
summary(lineal)

##   
## Call:  
## lm(formula = quality ~ sulphates, data = data.sulfatos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.98046 -0.51831 0.01954 0.46519 2.39917   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.56099 0.08649 52.74 <2e-16 \*\*\*  
## sulphates 1.65055 0.12921 12.77 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7695 on 1589 degrees of freedom  
## Multiple R-squared: 0.09313, Adjusted R-squared: 0.09256   
## F-statistic: 163.2 on 1 and 1589 DF, p-value: < 2.2e-16

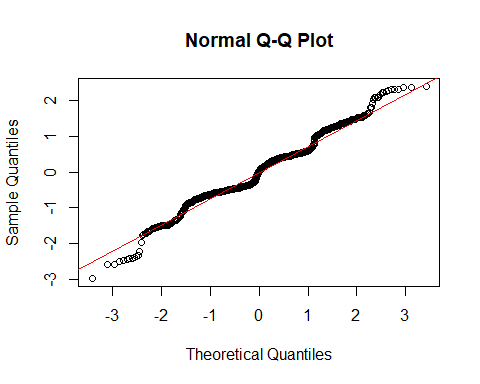
#REPRESENTACIÓN GRÁFICA  
plot(data.sulfatos$sulphates,data.sulfatos$quality, xlab = "Sulfatos", ylab = "Calidad", col="red")  
#RECTA DE REGRESIÓN  
abline(lineal, col="blue")



#ANÁLISIS POR RESIDUOS  
ggplot(data=data.sulfatos, aes(lineal$fitted.values,lineal$residuals)) + geom\_point() + geom\_hline(yintercept = 0) + geom\_smooth(color = "blue") + labs(title = "Residuos", x="Valores ajustados", y="Residuos")



#FUNCION QQNORM PARA REPRESENTAR EL GRÁFICO QUANTIL QUANTIL  
qqnorm(lineal$residuals)  
#LINEA DE TENDENCIA  
qqline(lineal$residuals, col="red")

 La recta se ajusta a los siguientes valores: y = 1.65055x + 4.56099

En el análisis por residuos vemos que los datos presentan una estructura definida, lo cual nos indica que el modelo lineal no se ajusta bien a los datos de la misma forma que podemos ver el efecto de algunos valores atípicos. Tambien podemos observar este detalle con el valor de R cuadrado, que es de 0.09256, practicamente cero.

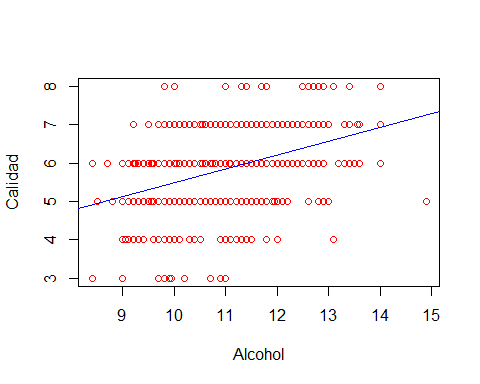
Vemos que el el gráfico QQ no se ajusta bien en ningún cuantil, por lo que indica que la variable no es normal.

#### ALCOHOL

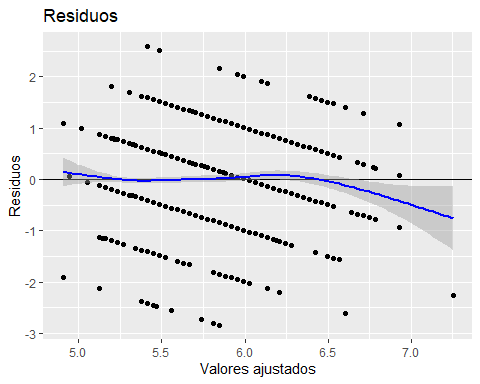
#AJUSTE DE LOS DATOS A LA ECUACIÓN y=ax+b  
lineal<-lm(calidad$quality~alcohol,data=variables)  
summary(lineal)

##   
## Call:  
## lm(formula = calidad$quality ~ alcohol, data = variables)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8442 -0.4112 -0.1690 0.5166 2.5888   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.87497 0.17471 10.73 <2e-16 \*\*\*  
## alcohol 0.36084 0.01668 21.64 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7104 on 1597 degrees of freedom  
## Multiple R-squared: 0.2267, Adjusted R-squared: 0.2263   
## F-statistic: 468.3 on 1 and 1597 DF, p-value: < 2.2e-16

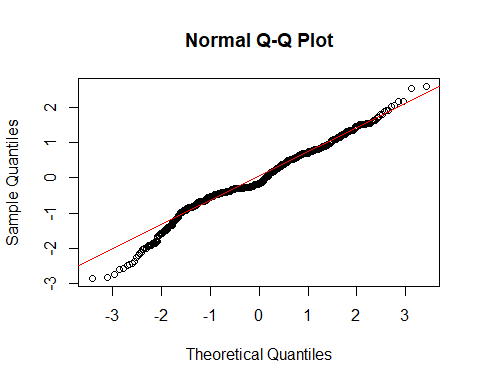
#REPRESENTACIÓN GRÁFICA  
plot(variables$alcohol,calidad$quality, xlab = "Alcohol", ylab = "Calidad", col="red")  
#RECTA DE REGRESIÓN  
abline(lineal, col="blue")



#ANÁLISIS POR RESIDUOS  
ggplot(data=variables, aes(lineal$fitted.values,lineal$residuals)) + geom\_point() + geom\_hline(yintercept = 0) + geom\_smooth(color = "blue") + labs(title = "Residuos", x="Valores ajustados", y="Residuos")



#FUNCION QQNORM PARA REPRESENTAR EL GRÁFICO QUANTIL QUANTIL  
qqnorm(lineal$residuals)  
#LINEA DE TENDENCIA  
qqline(lineal$residuals, col="red")

 La recta se ajusta a los siguientes valores: y = 1.87497x + 0.36084

En el análisis por residuos vemos que los datos presentan una estructura definida, lo cual nos indica que el modelo lineal no se ajusta bien a los datos de la misma forma que podemos ver el efecto de algunos valores atípicos. Tambien podemos observar este detalle con el valor de R cuadrado, que es de 0.2263, muy baja, a pesar de ser la que mejor resultados da..

Vemos que el el gráfico QQ se ajusta bien en el primer y segundo cuantil pero no en el resto, donde volvemos a ver presencia de valores atípicos que desajustan mucho el modelo.

### 4.2.2 Intervalos de confianza 95%.

Calculando los intervalos de confianza de nuestras variables podremos realizar contrastes de hipótesis bilaterales para conocer en que rangos puede estar la media de una muestra distinta a la nuestra, con una confianza del 95%.

#Para calcular el intervalo de confianza seguiremos la siguiente expresión:  
  
#(M - 1,96\*error, M + 1,96\*error)  
  
#Donde M es la media muestral, es decir, la media de los valores de nuestra muestra, que es diferente a la media de un bebe al nacer, ya que para ello deberiamos de tener los datos de todos los niños.  
#1,96 es el valor de z\*, que se obtiene de las tablas de distribución normal (http://halweb.uc3m.es/esp/Personal/personas/amalonso/esp/Tablas.pdf) para un intervalo de confianza del 95%, ya que equivale a un 2,5% o 0,025 en las tablas de uno de los extremos del area bajo la curva de la distribución.  
#El error no es mas que la desviación típica de la muestra, dividido por la raiz del número de casos, en nuestro caso son 1599 filas.  
  
  
#Para hacerlo implementaremos la función IQ, la cual calculará la media, la desviación tipica, el error y los dos intervalos de confianza, los cuales sacará por pantalla.  
IC<-function(x)  
{desv<-sd(x)  
media<-mean(x)  
error<-desv/(sqrt(length(x)))  
ICmenos<-media-1.96\*error  
ICmas<-media+1.96\*error  
paste("El intervalo de confianza es","(",ICmenos,",",ICmas,")")  
}

Ahora calculamos los intervalos de confianza y las medias de nuestras variables.

#CALIDAD  
  
mean(calidad$quality)

## [1] 5.636023

IC(calidad$quality)

## [1] "El intervalo de confianza es ( 5.59643923981771 , 5.67560578832488 )"

#ACIDEZ VOLATIL  
  
mean(variables$volatile.acidity)

## [1] 0.5278205

IC(variables$volatile.acidity)

## [1] "El intervalo de confianza es ( 0.519043844179357 , 0.536597181461669 )"

#ACIDO CITRICO  
  
mean(variables$citric.acid)

## [1] 0.2709756

IC(variables$citric.acid)

## [1] "El intervalo de confianza es ( 0.261427369731861 , 0.280523849780334 )"

#ALCOHOL  
  
mean(variables$alcohol)

## [1] 10.42298

IC(variables$alcohol)

## [1] "El intervalo de confianza es ( 10.3707490772481 , 10.475217151645 )"

#SULFATOS  
  
mean(data.sulfatos$sulphates)

## [1] 0.6524953

IC(data.sulfatos$sulphates)

## [1] "El intervalo de confianza es ( 0.645156490042222 , 0.659834081925094 )"

### 4.2.3 Arbol de decisión

Como parte de la ejecución de modelos estadísticos, se toma la determinación de implementar un modelo supervisado de Decision Tree, con el fin de identificar cuales son las cualidades mas comunes de los vinos que presentan una calificación mayor. Inicialmente filtramos los datos que se van a emplear para dicho analisis. Para ello se emplean las columnas de Acidez volatil, Acido citrico, Alcohol, Sulfatos; las cuales segun el criterio de reducción de dimensión son las que inciden de manera mas considerablen en la variable QUALITY.

De igual forma como parte del proceso de preparación de los datos para implementar el modelo de decision tree, se realiza una discretización de la variablel QUALITY, de tal forma que se sintetice las variables númericas existentes, pasando de tener 10 categorías, siendo cero las mas baja y 10 la mas alta, a tener tan solo 3 categorías: pobre, normal y bueno.

#Previamente para facilitarnos el ver que vinos son mejores o peores los agruparemos por puntuaciones. Asi los catalogaremos de 0-3 como "pobres", de 4-6 como "normales" y de 7-10 como "buenos"  
redwine$quality<-as.factor(redwine$quality)  
levels(redwine$quality)

## [1] "3" "4" "5" "6" "7" "8"

levels(redwine$quality)<-c("pobre","pobre","normal","normal","bueno","bueno")  
#Aqui podemos observar  
table(redwine$quality)

##   
## pobre normal bueno   
## 62 1312 217

variables\_tree <- redwine[c(2,3,10,11,12)]

Una vez seleccionadas se inicia la generación de el arbol de decisión. Para este caso no se realiza como tal partición de los datos para entrenamiento y/o evaluación ya que no se tiene el objetivo de generar un modelo de predicción. Ya que el objetivo del arbol es únicamente identificar si existe algun tipo de cualidad en los vinos que este mas relacionada con una buena calificiació, por lo tanto se emplean el 100% de los datos para entrenamiento

y <- variables\_tree[,5]  
X <- variables\_tree[,1:4]

Una vez divididas las variables independientes de la variable dependiente se define la variable dependiente quality como un factor.

y <- as.factor(y)

Una vez se tienen los datos listos se implementa inicialmente el modelo de arbol de decision C50.

tree\_model <- C50::C5.0(X,y,rules=TRUE )

Se ejecura el comando print() con el fin de identificar la cantidad de reglas o ramificaciones generadas a partir de la creación del arbol y la cantidad de datos analizados. Previamente se habia generado el arbol empleando los valores categóricos que venian por default en la columna QUALITY, generandose mas de 70 nodo, lo cual indica que los datos no presentan un comportamiento uniforme, siendo la calificación de los vinos altamente subjetividad.

Una vez implementado el metodo de discretización, se logra identificar que se obtiene un mejor comportamiento de los datos, generandose únicamente 9 nodos, lo cual demuestra un comportamiento mas uniforme en la información.

print(tree\_model)

##   
## Call:  
## C5.0.default(x = X, y = y, rules = TRUE)  
##   
## Rule-Based Model  
## Number of samples: 1599   
## Number of predictors: 4   
##   
## Number of Rules: 9   
##   
## Non-standard options: attempt to group attributes

Se imprime un resumen del arbol de decision generado logrando identificar cada una de las ramificaciones, en donde se especifica la clase a la que hace referencias y los patrones definidas (mayor, menos, igual) de cada una de las 4 variables. Se logra identificar de igual forma que el arbol presenta un margen de error que aunque no es despreciable si es menor al que se tenía con los datos originales, pasando de tener un error del 22.5% a tan solo un 13%.

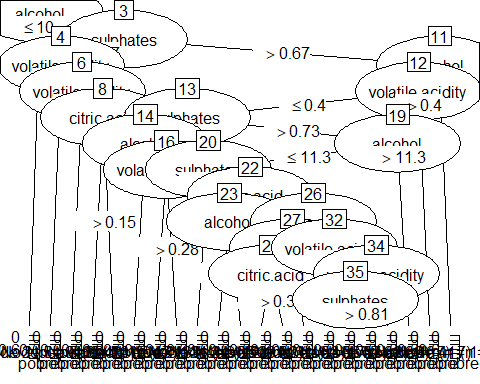
Otra factor importante que se logra concluir del presente resumen es el hecho de que de las 4 variables el alcohol es la que tiene mayor determinacion en el momento de definir la calidad de un vino, seguido por los sulfatos, la acidez volatil, y el acido cítrico. Tambien se logra extraer información relevante respecto a las calificaciones mas altas de calidad, las cuales aunque no presentan un comportamiento uniforme, tienden a presentar niveles bajos de acides volatil, niveles altos de acido cítrico y niveles de sulfato alto. En lo referente al alcohol, es comun que presenten niveles altos, superiores a 10.5 %.

summary(tree\_model)

##   
## Call:  
## C5.0.default(x = X, y = y, rules = TRUE)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Tue Jun 09 17:49:02 2020  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
## \*\*\* ignoring cases with bad or unknown class  
##   
## Read 1591 cases (5 attributes) from undefined.data  
##   
## Rules:  
##   
## Rule 1: (1042/120, lift 1.1)  
## sulphates <= 0.67  
## -> class normal [0.884]  
##   
## Rule 2: (1356/172, lift 1.1)  
## alcohol <= 11.6  
## -> class normal [0.873]  
##   
## Rule 3: (17/1, lift 6.6)  
## volatile.acidity <= 0.4  
## citric.acid > 0.4  
## citric.acid <= 0.47  
## sulphates > 0.77  
## alcohol > 10.4  
## alcohol <= 11.3  
## -> class bueno [0.895]  
##   
## Rule 4: (6, lift 6.4)  
## volatile.acidity <= 0.4  
## citric.acid <= 0.34  
## sulphates > 0.67  
## alcohol > 11.2  
## alcohol <= 11.6  
## -> class bueno [0.875]  
##   
## Rule 5: (13/1, lift 6.4)  
## volatile.acidity <= 0.4  
## citric.acid > 0.34  
## citric.acid <= 0.39  
## sulphates > 0.73  
## alcohol > 10.4  
## -> class bueno [0.867]  
##   
## Rule 6: (12/1, lift 6.3)  
## volatile.acidity <= 0.4  
## sulphates > 0.73  
## sulphates <= 0.77  
## alcohol > 10.4  
## alcohol <= 11.3  
## -> class bueno [0.857]  
##   
## Rule 7: (8/1, lift 5.9)  
## volatile.acidity > 0.29  
## volatile.acidity <= 0.33  
## citric.acid > 0.34  
## sulphates > 0.81  
## alcohol > 10.4  
## alcohol <= 11.3  
## -> class bueno [0.800]  
##   
## Rule 8: (110/38, lift 4.8)  
## sulphates > 0.67  
## alcohol > 11.6  
## -> class bueno [0.652]  
##   
## Rule 9: (115/45, lift 4.4)  
## volatile.acidity <= 0.4  
## sulphates > 0.73  
## alcohol > 10.4  
## -> class bueno [0.607]  
##   
## Default class: normal  
##   
##   
## Evaluation on training data (1591 cases):  
##   
## Rules   
## ----------------  
## No Errors  
##   
## 9 207(13.0%) <<  
##   
##   
## (a) (b) (c) <-classified as  
## ---- ---- ----  
## 62 (a): class pobre  
## 1271 41 (b): class normal  
## 104 113 (c): class bueno  
##   
##   
## Attribute usage:  
##   
## 92.14% alcohol  
## 77.18% sulphates  
## 7.42% volatile.acidity  
## 2.64% citric.acid  
##   
##   
## Time: 0.0 secs

Se genera una gráfica del arbol de decisión la cual no permite evidenciar con claridad los nodos generados, considerando la cantidad de datos generados. Se logra identificar sin embargo que el ROOT o raiz del arbol esta definida por la variable que mas tiene incidencia (Alcohol)

models\_graph <- C50::C5.0(X, y, trials=10)  
plot(models\_graph)



# 5. BASE DE DATOS LIMPIA

Una vez se realiza el proceso de limpieza de la base de datos y se realiza el análisis se guarda la información en el siguiente archivo .csv (comma separated values)

clean<-redwine[c(2,3,10,11,12)]  
write.csv(clean,'redwine\_clean.csv')

# 6. CONCLUSIONES

1. Las diferentes columnas que presenta la base de datos no presentan una distribución normal. Lo anterior se logra concluir al rechazar la hipotesis nula, basado en el valor de p el cual es muy inferior al rango de 0.05. Esto se refleja en la alta subjetividad que presenta la base de datos, al estar basada en la percepción de calidad de diferentes personas.
2. El tipo de base de datos que presenta el documento analizado para la presente práctica presenta una mejor reducción de dimensionalidad a través del análisis de correlación existente entre las variables independientes y la dependiente (QUALITY), ya que el modelo PCA únicamente permite reducir dos dimensiones con un 98% de precisión. Del análisis de correlación se logra determinar que las variables que presentan una incidencia mayor en el QUALITY del vino son: alcohol, sulfatos, la acidez volatil,y el acido cítrico.
3. Las variables seleccionadas presentan valores extremos. Sin embargo, son valores que se encuentran dentro de los rangos normales de un vino, por tanto no se pueden considerar como valores extremos. La variable sulfatos es la cual presenta una mayor cantidad de valores extremos y se realiza una suave limpieza de los mismos para eliminar los valores más extremos.
4. Se observa que Los modelos de regresión lineal no se ajustan a la variable “quality”. A pesar de esto, si podemos deducir que la variable alcohol (presenta el valor de R más alto) es la variable más influyente para determinar la calidad del vino. De los gráficos cuantil- cuantil volvemos a concluir que la distribución de las variables no son normales y presentan una dispersión muy elevada debido a la subjetividad de la variable “quality”.
5. De las cuatro variables escogidas a través del modelo de reducción de dimensionalidad, se logra implementar el modelo de arbol de decisión el cual permite confirmar que la variable nivel de alcohol es la que genera un mayor impacto al momento de determinar el nivel de calidad del vino, recurriendo en el 92% de los casos a su uso, seguido por el sulfato con un 77.18%. Las otras dos variables (volatile.acidity y citric.acid) no presentan una incidencia considerable en la calidad del vino
6. Aunque la base de datos tiene un comportamiento bastante subjetivo, se logra identificar que en general los vinos que presentan una calificación mayor (QUALITY), tienden a presentar niveles bajos de acidez volatil, niveles altos de ácido cítrico y niveles de sulfato alto. En lo referente al alcohol, es común que presenten niveles altos, superiores a 10.4 %.
7. Aplicando técnicas de discretización es posible aumentar de manera considerable la precisión del arbol de decisión, considerando que se reduce la volatilildad que presenta la base de datos debido a la subjetividad de los datos. De esta forma se logra reducir la cantidad de valores categóricos de la columna QUALITY de 8 a tan solo 3, permitiendo extraer información mas objetiva de los datos.