Practica 2 - Tipología y Ciclo de vida de datos

Maria JOsé MOrte Ruiz

Enero de 2020

# 1. Descripcion del dataset

\*\* Por qué es importante y qué pregunta/problema pretende responder?

Por medio del análisis y exploración del dataset vamos a averiguar si se puede mplantar alguna relación entre las propiedades fisicoquímicas de los vinos analizados (datos objetivos) y su calidad (subjetiva) y definir la importancia de cada una de ellas. Las conclusiones, es determinar qué características hay que potenciar y qué procesos de producción se pueden mejorar la elaboración de nuevos vinos, ya sean de buena calidad o no.

El objetivo es comparar las características volatile acidity, Residual sugar, Chlorides, Density y pH, para determinar la calidad del vino.

# INTRODUCCION

En ésta práctica vamos a trabajar con el juego de datos de es el Red Wine Quality <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>. o <https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/> el cual contiene dos datasets, uno de vinos blancos y otro de vinos tintos. Ambos datasets contienen 11 atributos de entrada, correspondientes a pruebas fisioquímicas, y uno de salida: “quality”. El objetivo del análisis será por un lado construir un modelo que nos pueda predecir la calidad de un vino, y por otro, construir un modelo que nos permita clasificar un vino en un determinado tipo (tinto), segun el enunciado

# EL DATASET

El Dataset contienen los valores fisicoquímicos y sensoriales de las variantes rojas del vino portugués “Vinho Verde”. Para más detalles, consulte la referencia [Cortez et al., 2009].No hay datos sobre los tipos de uva, la marca del vino, el precio de venta del vino, etc.). Los conjuntos de datos se pueden tomar como tareas de Regresión Lineal Las clases están ordenadas y no equilibradas (por ejemplo, hay muchos más vinos normales que excelentes o malos).

Pretendo determinar qué propiedades fisicoquímicas hacen que un vino sea clasificado como “bueno”

#Instalar paquetes

#install.packages(c("psych","ggplot2","car","Hmisc","corrplot"))  
#install.packages("tidyverse")  
#install.packages("ggplot2")  
#install.packages("devtools")  
#install.packages("DescTools")  
#install.packages("usethis")  
#install.packages("lawstat")

#Librarias necesarias

#Librarias necesarias  
library(ggplot2)  
library(dplyr)  
library(skimr)  
library(stringr)  
#library(ggthemes)  
library(GGally)  
library(corrplot)  
library(corrgram)  
library(nortest)  
library(psych)  
library(reshape2)  
library(lawstat)  
  
library(randomForest)  
  
  
  
library(gridExtra)  
  
library(pROC)  
  
library(caret)  
library(car)  
library(stats)

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

Este analisis nos permite comprobar que efectivamente nuestro archivo tiene 1599 registros y 12 variables. Asi mismo, podemos ver que tipo de variable es cada una. Los datos proporcionados están disponibles en CSV; el fichero proporciona en la primera linea, el nombre de los campos para facilitar la tarea de clasificación y dispone de un total de 1599 lineas de datos

Los datos son como se ven de un solo dataset. Después de leer los atributos e parece que hay algunos relacionados (la acidez y el dióxido de azufre). En caso de existir, esta relación deberá ver reflejada por ejemplo, por la existencia de un coeficiente de correlació con un cierto grado de significación, ya sea positivo o negativo. A pesar de esta observación, a priori, no podemos descartar ninguno de los atributos del conjunto inicial de datos pues todavía no disponemos de ningún indicio que permita fundamentar la eliminación de ningún

# Cargo el dataset

#Lectura fichero (winequality-red.csv)  
redwine<-read.csv("winequality-red.csv")  
#Se muestran los datos para asegurarnos una correcta lectura de ellos  
head(redwine)

## 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  
## 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  
## quality  
## 1 5  
## 2 5  
## 3 5  
## 4 6  
## 5 5  
## 6 5

#nombre de las variable  
names(redwine)

## [1] "fixed.acidity" "volatile.acidity" "citric.acid"   
## [4] "residual.sugar" "chlorides" "free.sulfur.dioxide"   
## [7] "total.sulfur.dioxide" "density" "pH"   
## [10] "sulphates" "alcohol" "quality"

## DESCRIPCION

Las columnas disponibles son las siguientes:

**acidez fija:** La mayoría de los ácidos involucrados con el vino o fijos o no volátiles (no se evaporan fácilmente)

**acidez volátil:** la cantidad de ácido acético en el vino, que a niveles demasiado altos puede dar lugar a un sabor desagradable y vinagre

**ácido cítrico:** encontrado en pequeñas cantidades, el ácido cítrico puede agregar ‘frescura’ y sabor a los vinos

**azúcar residual:** La cantidad de azúcar restante después de que se detiene la fermentación, es raro encontrar vinos con menos de 1 gramo / litro y los vinos con más de 45 gramos / litro se consideran dulces

**cloruros:** la cantidad de sal en el vino

**dióxido de azufre:** libre la forma libre de SO2 existe en equilibrio entre SO2 molecular (como gas disuelto) y el ion bisulfito; Previene el crecimiento microbiano y la oxidación del vino.

**dióxido de azufre:** total cantidad de formas libres y ligadas de SO2; en bajas concentraciones, el SO2 es principalmente indetectable en el vino, pero a concentraciones de SO2 libres superiores a 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino

**densidad:** la densidad del agua es cercana a la del agua dependiendo del porcentaje de contenido de alcohol y azúca

**pH:** describe cuán ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3-4 en la escala de pH

**sulfatos:** un aditivo para el vino que puede contribuir a los niveles de dióxido de azufre (S02), que actúa como antimicrobiano y antioxidante

**alcohol:** el porcentaje de contenido de alcohol del vino

**calidad:** variable de salida (basada en datos sensoriales, puntaje entre 0 y 10)

Es importante destacar que las primeras 11 columnas son datos objetivos obtenidos a través de métodos cientificos mientras que la última columna, se trata de un dato subjetivo y es obtenido por un metodo desconocido, clasificando entre 3 y 8 la calidad del vino.

Procedo a la carga del CSV con los siguientes parámetros generando un dataframe denominado redwine

#Crea una variable que indique si un vino es bueno o malo

#AHora tiene 13 variables con la nueva  
redwine$good.wine<-ifelse(redwine$quality>6,1,0)  
#Se muestran los datos para asegurarnos una correcta lectura de ellos  
tail(redwine)

## fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1594 6.8 0.620 0.08 1.9 0.068  
## 1595 6.2 0.600 0.08 2.0 0.090  
## 1596 5.9 0.550 0.10 2.2 0.062  
## 1597 6.3 0.510 0.13 2.3 0.076  
## 1598 5.9 0.645 0.12 2.0 0.075  
## 1599 6.0 0.310 0.47 3.6 0.067  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 1594 28 38 0.99651 3.42 0.82 9.5  
## 1595 32 44 0.99490 3.45 0.58 10.5  
## 1596 39 51 0.99512 3.52 0.76 11.2  
## 1597 29 40 0.99574 3.42 0.75 11.0  
## 1598 32 44 0.99547 3.57 0.71 10.2  
## 1599 18 42 0.99549 3.39 0.66 11.0  
## quality good.wine  
## 1594 6 0  
## 1595 5 0  
## 1596 6 0  
## 1597 6 0  
## 1598 5 0  
## 1599 6 0

# Printem el nom i el tipus de variable  
nombretipo <- sapply(redwine,class)  
data.frame(Variables=names(nombretipo),Classe=as.vector(nombretipo))

## Variables Classe  
## 1 fixed.acidity numeric  
## 2 volatile.acidity numeric  
## 3 citric.acid numeric  
## 4 residual.sugar numeric  
## 5 chlorides numeric  
## 6 free.sulfur.dioxide numeric  
## 7 total.sulfur.dioxide numeric  
## 8 density numeric  
## 9 pH numeric  
## 10 sulphates numeric  
## 11 alcohol numeric  
## 12 quality integer  
## 13 good.wine numeric

#Estadsiticas Para comprobar la calidad de los datos, realizo un str nos describe el tipo de variablesy con Summary para ver el aspecto que tiene la estadistica de los datos y

#Estructura de los datos  
str(redwine)

## 'data.frame': 1599 obs. of 13 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 ...  
## $ good.wine : num 0 0 0 0 0 0 0 1 1 0 ...

# Es una función genérica utilizada para resúmenes de resultados, de varias funciones.  
# Muestra la media, mediana, cuartiles, valor mínimo y valor máximo, para variables cuantitativas y la frecuencia absoluta para variables cualitativas.  
  
summary(redwine)

## 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   
## good.wine   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.1357   
## 3rd Qu.:0.0000   
## Max. :1.0000

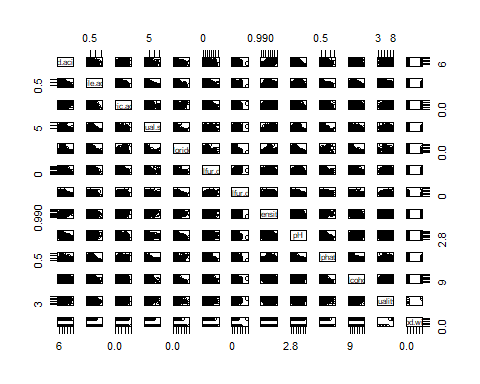
Podemos observar información interesante en los datos proporcionados por el summary. En primer lugar observamos que los máximos y mínimos de las 12 columnas del dataframe contienen datos numéricos. Observamos que hay 1599 observaciones y 12 variables en el dataframe Las primeras columnas son numéricas y la última contiene valores enteros En concreto la columna quality es la variable objetivo del estudio; para ver detalles de la misma utilizamos el comando table

# La distribución de frecuencia de los niveles del factor  
table(redwine$quality)

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

Observamos una concentración en los valores 5 y 6 de calidad; con valores entre 3 y 8, en el summary, vemos como valor medio 5.636 como resumen del análisis utilizamos la función skimr Realizmao estudio

# Graficas del dataset  
plot(redwine)



# Resumen de estadisticas  
skim(redwine)

Data summary

|  |  |
| --- | --- |
| Name | redwine |
| Number of rows | 1599 |
| Number of columns | 13 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 13 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| fixed.acidity | 0 | 1 | 8.32 | 1.74 | 4.60 | 7.10 | 7.90 | 9.20 | 15.90 | ▂▇▂▁▁ |
| volatile.acidity | 0 | 1 | 0.53 | 0.18 | 0.12 | 0.39 | 0.52 | 0.64 | 1.58 | ▅▇▂▁▁ |
| citric.acid | 0 | 1 | 0.27 | 0.19 | 0.00 | 0.09 | 0.26 | 0.42 | 1.00 | ▇▆▅▁▁ |
| residual.sugar | 0 | 1 | 2.54 | 1.41 | 0.90 | 1.90 | 2.20 | 2.60 | 15.50 | ▇▁▁▁▁ |
| chlorides | 0 | 1 | 0.09 | 0.05 | 0.01 | 0.07 | 0.08 | 0.09 | 0.61 | ▇▁▁▁▁ |
| free.sulfur.dioxide | 0 | 1 | 15.87 | 10.46 | 1.00 | 7.00 | 14.00 | 21.00 | 72.00 | ▇▅▁▁▁ |
| total.sulfur.dioxide | 0 | 1 | 46.47 | 32.90 | 6.00 | 22.00 | 38.00 | 62.00 | 289.00 | ▇▂▁▁▁ |
| density | 0 | 1 | 1.00 | 0.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | ▁▃▇▂▁ |
| pH | 0 | 1 | 3.31 | 0.15 | 2.74 | 3.21 | 3.31 | 3.40 | 4.01 | ▁▅▇▂▁ |
| sulphates | 0 | 1 | 0.66 | 0.17 | 0.33 | 0.55 | 0.62 | 0.73 | 2.00 | ▇▅▁▁▁ |
| alcohol | 0 | 1 | 10.42 | 1.07 | 8.40 | 9.50 | 10.20 | 11.10 | 14.90 | ▇▇▃▁▁ |
| quality | 0 | 1 | 5.64 | 0.81 | 3.00 | 5.00 | 6.00 | 6.00 | 8.00 | ▁▇▇▂▁ |
| good.wine | 0 | 1 | 0.14 | 0.34 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| # Como se ve en good.wi | ne la Mean e | s Mean :0.1357 |  |  |  |  |  |  |  |  |
| # trabajaremos con un c | onjunto de d | atos desequilibr | ado, |  |  |  |  |  |  |  |
| # en el que solo alrede | dor del 13.5 | 7% de 1599 vinos | se cons | idera un | buen v | ino. |  |  |  |  |

NOTA: Ms adelante de este documento se adjuntan gráficas relativas a la distribución de los valores de los diferentes atributos en formato boxplot y histog

# 3. LIMPIEZA DE DATOS

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

El dataframe no tiene campos vacios ni nulos, es un buen dataset desde el punto de vista de limpieza.

### VALORES NULOS

Uno de los primeros pasos cuando se trabaja con datos, es observar si las variables tienen valores perdidos o nulos. Esto significa, que por alguna razon no se pude obtener o registrar el valor.

Si bien, lo ideal es que nunca falten ningun dato, lo cierto es que en la realidad eso pocas veces pasa, por lo que se deben de aplicar tecnicas para poder estimar esos valores que no estan disponibles.

Lo primero que se hara, es verificar que no existan valores perdidos en el conjunto de datos a analizar.

El dataset no tiene elementos vacios, lo podemos comprobar con el siguiente comando. HAy que tenr en cuenta que comúnmente, se utilizan los ceros como centinela para indicar la ausencia de ciertos valores. Vamos a proceder a conocer a continuación qué campos contienen elementos con valores ceros o elementos vacíos

sapply(redwine, function(x) sum(length(which(is.na(x)))))

## 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   
## good.wine   
## 0

#colSums(is.na(redwine))

## Verificar en que columnas hay valores cero  
lapply(redwine, function(x) sum(x==0))

## $fixed.acidity  
## [1] 0  
##   
## $volatile.acidity  
## [1] 0  
##   
## $citric.acid  
## [1] 132  
##   
## $residual.sugar  
## [1] 0  
##   
## $chlorides  
## [1] 0  
##   
## $free.sulfur.dioxide  
## [1] 0  
##   
## $total.sulfur.dioxide  
## [1] 0  
##   
## $density  
## [1] 0  
##   
## $pH  
## [1] 0  
##   
## $sulphates  
## [1] 0  
##   
## $alcohol  
## [1] 0  
##   
## $quality  
## [1] 0  
##   
## $good.wine  
## [1] 1382

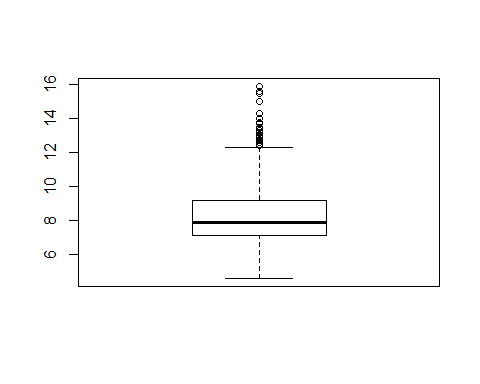
#sum(is.na(redwine))

Podemos ver que no hay problema en un caso **good.wine** nuestra varible de uso, es de usode clasificacion por asi decirlo, y en caso de citric.acid con 132 es un valor que pude etar permitido para eata variable, por lo tanto son valores queno decartamos sino buenos Solo el campo **citric.acid** tiene valores cero. Podemos ver en el siguiente enlace (<http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity>) la concentración de ácido cítrico puede ser entre 0 y 500 mg/L, por lo tanto, podemos dar esos valores como buenos.

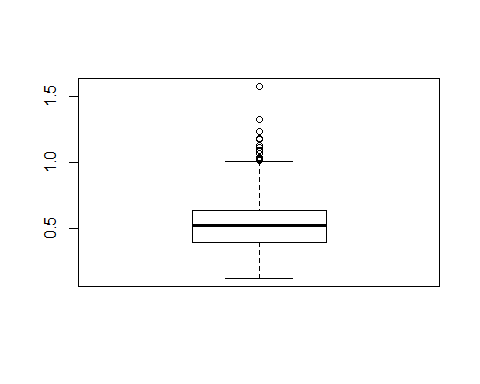
### Identificación y tratamiento de valores extremos.

Vamos a comprobar los Outlier que tiene las variables con el comando

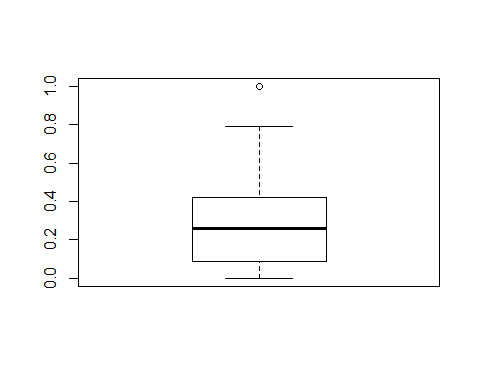
boxplot(redwine$fixed.acidity )



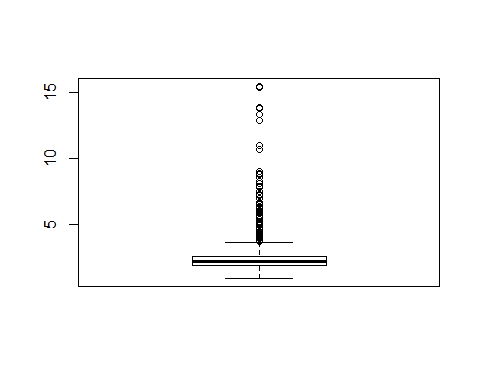
boxplot(redwine$volatile.acidity)



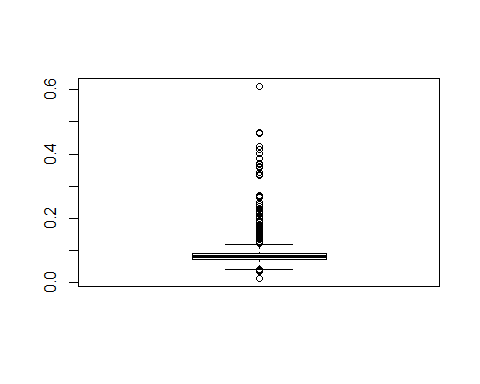
boxplot(redwine$citric.acid)



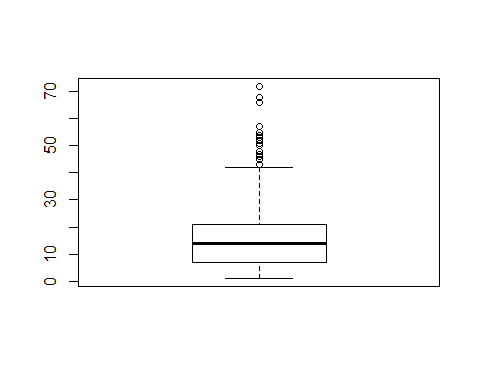
boxplot(redwine$residual.sugar)



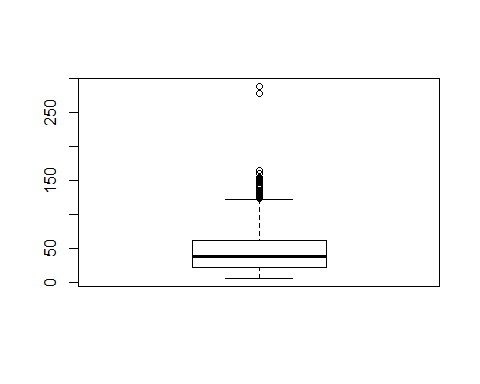
boxplot(redwine$chlorides)



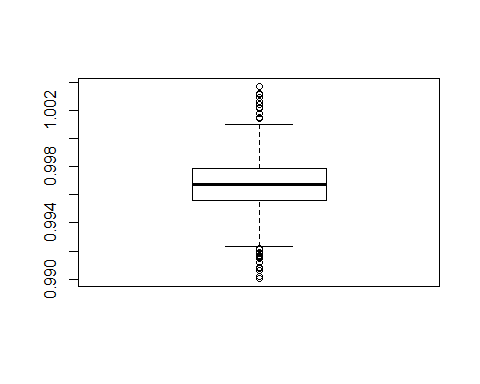
boxplot(redwine$free.sulfur.dioxide)



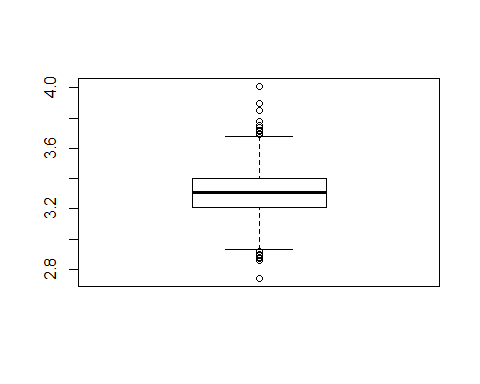
boxplot(redwine$total.sulfur.dioxide)



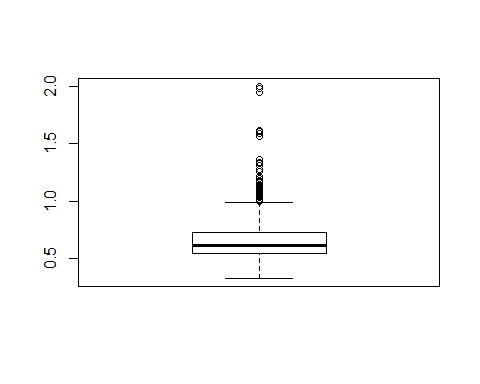
boxplot(redwine$density)



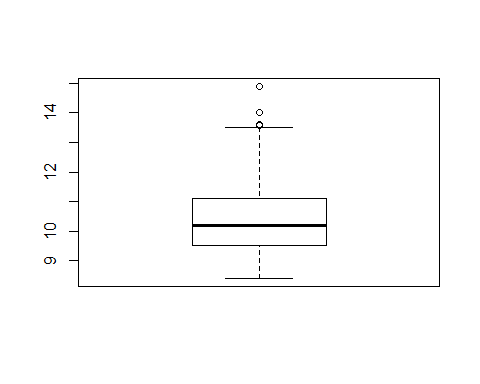
boxplot(redwine$pH)



boxplot(redwine$sulphates)

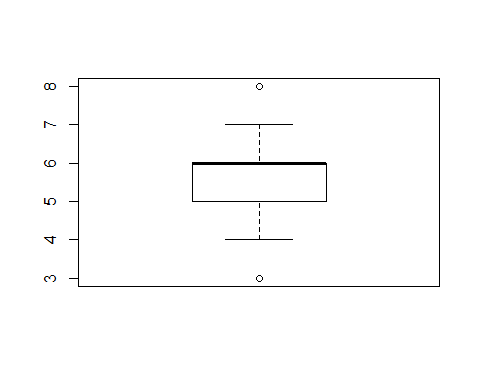


boxplot(redwine$alcohol)

 En esta variable encontramos una serie de valores extremos a la media los cuales son necesarios estudiar. En este caso en particular, considero necesario dejar los valores atípicos de los datos objetivos como son los de los valores fisicoquímicos obtenidos Otro caso es el Quality, puesto que es subjetivo, tampoco conocemos el margen de error que puede tener la decisión de haberlo marcado con el valor 8.

Para la variable a buscar obtenemos los siguients outlers

boxplot(redwine$quality)



fivenum(redwine$quality)

## [1] 3 5 6 6 8

Lo cual quiere decir que hay dos valores extremos en 3 y en 8 y la media coincide con los cinco números de Tukey

* minimum = 3
* lower-hinge = 5
* median = 6
* upper-hinge = 6
* maximum = 8

La conclusión es que un gran número de valores se concentran alrededor del 5 y 6 de calidad (el valor de la media y mediana están ente estos dos valores)

Se dice que un valor extremo se puede describir como una observación que se desvía tanto de otros observaciones como para despertar sospechas de que fue generado por un mecanismo diferente.

Como punto de partida, podemos considerar que aquellas observaciones alejadas 3 o más desviaciones estándar de la media de la muestra son susceptibles de ser valores extremos, salvo que la muestra sea particularmente pequeña. Esta valores extremos son un caso particular de los valores que, convencionalmente y de manera generalizada, se conocen como outliers y que por definición son aquellos valores que están situados en: • una distancia superior a 1,5 veces el rango interquartíl.lic por sobre el 3er cuartil • una distancia inferior a 1,5 veces el rango interquartíl.lic por debajo del 1er cuartil Usaremos esta última definición (configuración por defecto del parámetro coef de la función boxplot.stats) para el cálculo de los valores aípics

c <- names(redwine)  
total\_vatipicos<- 0  
for (i in 1:ncol(redwine))  
{  
cat(" \n")  
a <- boxplot.stats(redwine[, i],)$out  
cat("\*\*Atributos '",c[i],"'\*\* \n", sep = '')  
cat("\*Num. valores atipicos : ", length(a), "\* \n", sep = '')  
cat(sort(a), " \n")  
total\_vatipicos <- total\_vatipicos + length(a)  
}

##   
## \*\*Atributos 'fixed.acidity'\*\*   
## \*Num. valores atipicos : 49\*   
## 12.4 12.4 12.4 12.4 12.5 12.5 12.5 12.5 12.5 12.5 12.5 12.6 12.6 12.6 12.6 12.7 12.7 12.7 12.7 12.8 12.8 12.8 12.8 12.8 12.9 12.9 13 13 13 13.2 13.2 13.2 13.3 13.3 13.3 13.4 13.5 13.7 13.7 13.8 14 14.3 15 15 15.5 15.5 15.6 15.6 15.9   
##   
## \*\*Atributos 'volatile.acidity'\*\*   
## \*Num. valores atipicos : 19\*   
## 1.02 1.02 1.02 1.02 1.025 1.035 1.04 1.04 1.04 1.07 1.09 1.115 1.13 1.18 1.185 1.24 1.33 1.33 1.58   
##   
## \*\*Atributos 'citric.acid'\*\*   
## \*Num. valores atipicos : 1\*   
## 1   
##   
## \*\*Atributos 'residual.sugar'\*\*   
## \*Num. valores atipicos : 155\*   
## 3.7 3.7 3.7 3.7 3.75 3.8 3.8 3.8 3.8 3.8 3.8 3.8 3.8 3.9 3.9 3.9 3.9 3.9 3.9 4 4 4 4 4 4 4 4 4 4 4 4.1 4.1 4.1 4.1 4.1 4.1 4.2 4.2 4.2 4.2 4.2 4.25 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.4 4.4 4.4 4.4 4.5 4.5 4.5 4.5 4.6 4.6 4.6 4.6 4.6 4.6 4.65 4.65 4.7 4.8 4.8 4.8 5 5.1 5.1 5.1 5.1 5.1 5.15 5.2 5.2 5.2 5.4 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.6 5.6 5.6 5.6 5.6 5.6 5.7 5.8 5.8 5.8 5.8 5.9 5.9 5.9 6 6 6 6 6.1 6.1 6.1 6.1 6.2 6.2 6.2 6.3 6.3 6.4 6.4 6.4 6.55 6.55 6.6 6.6 6.7 6.7 7 7.2 7.3 7.5 7.8 7.8 7.9 7.9 7.9 8.1 8.1 8.3 8.3 8.3 8.6 8.8 8.8 8.9 9 10.7 11 11 12.9 13.4 13.8 13.8 13.9 15.4 15.4 15.5   
##   
## \*\*Atributos 'chlorides'\*\*   
## \*Num. valores atipicos : 112\*   
## 0.012 0.012 0.034 0.038 0.038 0.039 0.039 0.039 0.039 0.12 0.12 0.12 0.121 0.121 0.122 0.122 0.122 0.122 0.122 0.122 0.122 0.123 0.123 0.123 0.123 0.123 0.123 0.124 0.124 0.124 0.125 0.126 0.127 0.128 0.132 0.132 0.132 0.132 0.136 0.137 0.143 0.145 0.146 0.147 0.148 0.152 0.152 0.153 0.157 0.157 0.157 0.159 0.161 0.165 0.166 0.166 0.166 0.168 0.169 0.17 0.171 0.171 0.172 0.174 0.176 0.178 0.178 0.186 0.19 0.194 0.2 0.205 0.205 0.213 0.214 0.214 0.214 0.216 0.222 0.226 0.226 0.23 0.235 0.236 0.241 0.243 0.25 0.263 0.267 0.27 0.332 0.337 0.341 0.343 0.358 0.36 0.368 0.369 0.387 0.401 0.403 0.413 0.414 0.414 0.415 0.415 0.415 0.422 0.464 0.467 0.61 0.611   
##   
## \*\*Atributos 'free.sulfur.dioxide'\*\*   
## \*Num. valores atipicos : 30\*   
## 43 43 43 45 45 45 46 47 48 48 48 48 50 50 51 51 51 51 52 52 52 53 54 55 55 57 66 68 68 72   
##   
## \*\*Atributos 'total.sulfur.dioxide'\*\*   
## \*Num. valores atipicos : 55\*   
## 124 124 124 125 125 126 127 127 128 128 129 129 129 130 131 131 131 133 133 133 134 134 135 135 136 136 139 140 141 141 141 142 143 143 144 144 144 145 145 145 147 147 147 148 148 149 151 151 152 153 155 160 165 278 289   
##   
## \*\*Atributos 'density'\*\*   
## \*Num. valores atipicos : 45\*   
## 0.99007 0.99007 0.9902 0.99064 0.99064 0.9908 0.99084 0.9912 0.9915 0.99154 0.99157 0.9916 0.9916 0.99162 0.9917 0.99182 0.99182 0.99191 0.9921 0.9922 0.9922 1.0014 1.0014 1.0014 1.0014 1.0014 1.0014 1.0015 1.0015 1.0018 1.0021 1.0021 1.0022 1.0022 1.00242 1.00242 1.0026 1.0026 1.00289 1.00315 1.00315 1.00315 1.0032 1.00369 1.00369   
##   
## \*\*Atributos 'pH'\*\*   
## \*Num. valores atipicos : 35\*   
## 2.74 2.86 2.87 2.88 2.88 2.89 2.89 2.89 2.89 2.9 2.92 2.92 2.92 2.92 3.69 3.69 3.69 3.69 3.7 3.71 3.71 3.71 3.71 3.72 3.72 3.72 3.74 3.75 3.78 3.78 3.85 3.9 3.9 4.01 4.01   
##   
## \*\*Atributos 'sulphates'\*\*   
## \*Num. valores atipicos : 59\*   
## 1 1.01 1.02 1.02 1.02 1.03 1.03 1.04 1.04 1.05 1.05 1.05 1.06 1.06 1.06 1.06 1.07 1.07 1.08 1.08 1.08 1.09 1.1 1.1 1.11 1.12 1.13 1.13 1.14 1.14 1.15 1.16 1.17 1.17 1.17 1.17 1.17 1.18 1.18 1.18 1.2 1.22 1.26 1.28 1.28 1.31 1.33 1.34 1.36 1.36 1.36 1.56 1.59 1.61 1.62 1.95 1.95 1.98 2   
##   
## \*\*Atributos 'alcohol'\*\*   
## \*Num. valores atipicos : 13\*   
## 13.56667 13.6 13.6 13.6 13.6 14 14 14 14 14 14 14 14.9   
##   
## \*\*Atributos 'quality'\*\*   
## \*Num. valores atipicos : 28\*   
## 3 3 3 3 3 3 3 3 3 3 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8   
##   
## \*\*Atributos 'good.wine'\*\*   
## \*Num. valores atipicos : 217\*   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Total outliers: 601, pues aunque salen 808 no tenemoe encuenta ‘good.wine’ A priori aparecen 601 valores susceptibles de ser considerados ‘atípicos’ y distribuidos de forma heterogénea entre los diferentes atributos pero, dado que no tenemos suficiente conocimiento sobre el mundo de los vinos para saber si se trata de errores de medición o son valores posibles, optaremos para dejarlos, sin hacer ningún tratamiento, aunque ya podemos anticipar que su dispersión puede impactar en el análisis de los datos Otra técnica que a menudo se usa para encontrar valores atípicos es utilizando la distancia de Mahalanobis que por medio del concepto de similitud, permite identificar aquellas observaciones que más se alejan (o son menos similares) respecto al resto de valores. A continuación, a modo de ejemplo práctico, realizaremos las siguientes operaciones • calcular la distancia de Mahalanobis para cada una de las filas del fecha siete • calcular los estadísticos básicos de la distancia de Mahalanobis y los outliers • Visualizaremos gráficamente qué elementos del conjuntos de datos son menos similares (aquellos que estén dispersos y más alejados)

## Grafico

#Distancia de Mahalanobis  
#Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales  
options(repr.plot.width=6, repr.plot.height=4)  
#Calsulamso la distancia de Mahalanobis  
mahaladistancia <- mahalanobis(redwine[, c(1:11)], colMeans(redwine[, c(1:11)]), cov(redwine[, c(1:11)]))  
# Estadistica basica de la distancia de Mahalanobis- mahaladistncia  
summary(mahaladistancia)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.092 5.120 7.738 10.993 12.172 155.546

# Rango intercuartílico  
iqr <- IQR(mahaladistancia)  
iqr

## [1] 7.052017

## [1] 7.052017

# Umbral  
Umbral <- IQR(mahaladistancia)\*(1.5) + summary(mahaladistancia)[[5]] # 3er quartil  
Umbral

## [1] 22.74971

#[1] 22.74971

# Representación gráfica  
plot(mahaladistancia, pch=1, cex=.2, main="Distancia de Mahalanibus",  
xlab = "Índice de elementos", ylab="Distancia")  
#liena de corte  
abline(h = Umbral, col="blue")

# Nombre valores atipicos  
vp <- length(boxplot.stats(mahaladistancia)$out)  
vp

## [1] 110

#[1] 110

redwine$mahaladistncia <- NULL

#Se guardan los datos en un nuevo archivo csv  
write.csv(redwine, "redwine.csv")

# 4 Análisis DE DATOS

### Selccionde lso datos

Vamos a usar tods lso atributos, etudairemso la homogeneidad de la varianza, discretizar la variable quality en una nueva variable clase, cualitativa, que contendrá 3 valores posibles (Malo, Normal o Bueno) relativas a la valoración del vino Normalidad Planteamos el contraste de hipotesis par saber si los atribitos siguen una distribucion normal con uan significacion α del 0.05.

• H0: La muestra de tamaño n sigue una distribución Normal • H1: La muestra de tamaño n NO sigue una distribución Normal Este contraste lo llevaremos a cabo con el test de normalidad de Shapiro-Wilk para cada uno de los atributos. Si el p-valor obtenido para cada atributo es menor al nivel de significación α (<0,05), rechazaremos la hipótesis nul.la (H0) y afirmaremos que la muestra NO sigue una distribución normal

# Normalitat  
for (i in 1:ncol(redwine))  
{  
 cat("Atribut '",c[i],"', ", sep = '')  
 pvalor <- shapiro.test(redwine[, i])[["p.value"]]  
 cat("p-valor '", pvalor,"'\n", sep = '')  
}

## Atribut 'fixed.acidity', p-valor '1.525012e-24'  
## Atribut 'volatile.acidity', p-valor '2.692935e-16'  
## Atribut 'citric.acid', p-valor '1.021932e-21'  
## Atribut 'residual.sugar', p-valor '1.020162e-52'  
## Atribut 'chlorides', p-valor '1.179056e-55'  
## Atribut 'free.sulfur.dioxide', p-valor '7.694597e-31'  
## Atribut 'total.sulfur.dioxide', p-valor '3.573451e-34'  
## Atribut 'density', p-valor '1.936053e-08'  
## Atribut 'pH', p-valor '1.712237e-06'  
## Atribut 'sulphates', p-valor '5.82314e-38'  
## Atribut 'alcohol', p-valor '6.644057e-27'  
## Atribut 'quality', p-valor '9.515085e-36'  
## Atribut 'good.wine', p-valor '3.726209e-58'

Observamso que no seiguen una distribucion normañ Aplicaremso la homogeneidad de la varianza aplicando también un contraste de hipótesis con un nivel de significación α del 0.05. Para ello, tal y como se he dicho hemos discretizado la variable quality en una nueva variable cualitativa clase, cualitativa, que contendrá 3 valores posibles (Malo, Normal o Bueno) relativas a la valoración del vino.

Aplicaremos la homogeneidad comparando las varianzas de las muestras de vinos agrupados por este nuevo atributo clase • H0: Las varianzas poblacionales son iguales (Homoscedasticidad): • H1: Las varianzas poblacionales son diferentes (heteroscedasticidad) Este contraste lo llevaremos a cabo mediante deltest de Levene Si el p-valor obtenido es menor que el nivel de significación α (<0,05) rechazaremos la hipótesis nul.la (H0) y afirmaremos que las varianzas poblacionales son diferentes (heteroscedasticidad). De lo contrario, no podremos rechazar la hipótesis nul.la H0

# Homogeneitat de les variances  
redwine2 <- redwine  
redwine2$classe <- 0  
idx <- which(redwine2$quality < 5)  
redwine2$classe[idx] <- "Dolent"  
idx <- which(redwine2$quality >= 5 & redwine2$quality <= 6)  
redwine2$classe[idx] <- "Normal"  
idx <- which(redwine2$quality >= 7)  
redwine2$classe[idx] <- "Bo"  
redwine2$quality <- NULL  
redwine2$classe <- as.factor(redwine2$classe)  
table(redwine2$classe)

##   
## Bo Dolent Normal   
## 217 63 1319

# El test de Levene  
leveneTest(redwine2$alcohol,redwine2$classe)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.7375 0.4785  
## 1596

Conclusion: Como el p-valor (0.4785) es mayor que el nivel de signifació α (0.05), no podemos rechazar la hipótesis nula H0 que las varianzas poblacionales son iguales y por lo tanto, podemos afirmar que hay homoscedasticidad

## Comparar dos grupos de daros:

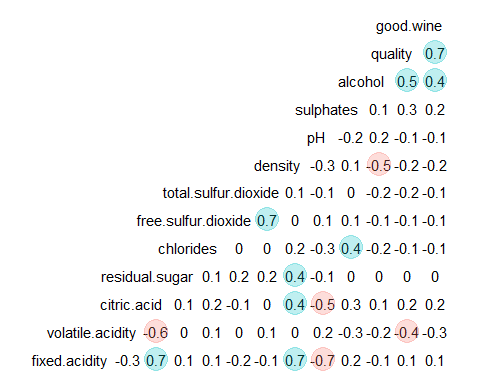
Queremos ver si hay relacion entre las propiedades fisicoquímicas de los vinos analizados (datos objetivos) y su calidad (Valoración subjetiva) y determinar la importancia de cada una de ellas.

Para llevarlo a cabo, realizaremos 3 pruebas estadísticas La matriz de correlación entre los diferentes atributos, hallarla Reducir la cardinalidad aplicando un algoritmo de PCA Modelos de regresión lineal para ver si estos cuentan o no la calidad del vino en función del resto de atributos y si el modelo fomenta dudas

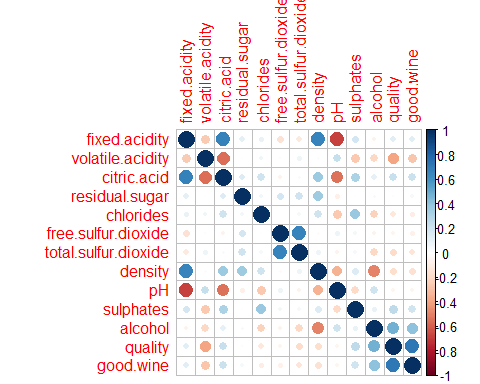
### Matriz de correlacion

Dos graficas que representa la ralcion directa o inveras de lso atributos del dataset

# Grafico  
options(repr.plot.width=4, repr.plot.height=3)  
ggcorr(redwine, geom = "blank", label = TRUE,  
hjust = 0.9, layout.exp = 2) +  
geom\_point(size = 8, aes(color = coefficient > 0,  
alpha = abs(coefficient) > 0.35)) +  
scale\_alpha\_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +  
guides(color = FALSE, alpha = FALSE)



# Visualizacion grafica de una matriz a color  
   
cor\_redwine<-cor(redwine)  
#cor\_redwine  
library(corrplot)  
corrplot(cor\_redwine,method = "circle")



corrplot(cor\_redwine,method = "number")

 Observaciones • La calidad (**quality**)del vino parece estar principalmente relacionada con los niveles de alcohol (**alcohol**), de manera directa y con la concentración de ácido acético (**volatile.acidity**), de manera **inversa**. • El% de alcohol de un vino (**alcohol**) está inversamente relacionado con su densidad (**density**) • El valor de pH (**pH**) está inversa y significativamente relacionado con atributos que tienen que ver con la acidez (**citric.acid** y **fixed.acidity**) • La densidad (**density**) es la propiedad con un *mayor* número de correlaciones con otros atributos (directa o inversamente): **alcohol**, **residual.sugar**, **citric.acid** y **fixed.acidity** • Los atributos que miden sustancias similares (**total.sulfur.dioxide** y **free.sulfur.dioxide**, por un lado y **citric.acid**, **volatile.acidity** y **fixed.acidity**, por el otro) están fuertemente correlacionades (Directa o inversamente), como era de esperar y ya habíamos enunciado. •Para terminar, el atributo **fixed.acidity** es la que está más fuertemente correlacionada (directa o inversamente) con otros atributos: ""citric.acid**,** density\*\* y **pH** Hay atributos que parecen mejores candidatos a ser explicados con métodos de regresión lineal que a la calidad (**quality**)

### PCA

DESpes de la correlación que hay entre los diferentes atributos, aunque su cardinalidad NO es muy alta, parece que NO son necesarios todos atributos para determinar la calidad del vino. Vamos a ver que el atributo **quality** sólo está muy relacionado con los atributos **alcohol** y **volatile.acidity** APlicamso PCA para reducir la cardinalidad del conjunto, para encontrar qué atributos son significativos

# Elegimos 'quality' y aplicamos el PCA  
redwine.pca <- prcomp(redwine[,1:11], scale. = TRUE)

summary(redwine.pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.7604 1.3878 1.2452 1.1015 0.97943 0.81216 0.76406  
## Proportion of Variance 0.2817 0.1751 0.1410 0.1103 0.08721 0.05996 0.05307  
## Cumulative Proportion 0.2817 0.4568 0.5978 0.7081 0.79528 0.85525 0.90832  
## PC8 PC9 PC10 PC11  
## Standard deviation 0.65035 0.58706 0.42583 0.24405  
## Proportion of Variance 0.03845 0.03133 0.01648 0.00541  
## Cumulative Proportion 0.94677 0.97810 0.99459 1.00000

Se ve que que hace falat 7 atributos apr explciar la varianza al 90%

#   
redwine.pca$rotation

## PC1 PC2 PC3 PC4  
## fixed.acidity 0.48931422 -0.110502738 0.12330157 -0.229617370  
## volatile.acidity -0.23858436 0.274930480 0.44996253 0.078959783  
## citric.acid 0.46363166 -0.151791356 -0.23824707 -0.079418256  
## residual.sugar 0.14610715 0.272080238 -0.10128338 -0.372792562  
## chlorides 0.21224658 0.148051555 0.09261383 0.666194756  
## free.sulfur.dioxide -0.03615752 0.513566812 -0.42879287 -0.043537818  
## total.sulfur.dioxide 0.02357485 0.569486959 -0.32241450 -0.034577115  
## density 0.39535301 0.233575490 0.33887135 -0.174499758  
## pH -0.43851962 0.006710793 -0.05769735 -0.003787746  
## sulphates 0.24292133 -0.037553916 -0.27978615 0.550872362  
## alcohol -0.11323206 -0.386180959 -0.47167322 -0.122181088  
## PC5 PC6 PC7 PC8  
## fixed.acidity 0.08261366 -0.10147858 0.35022736 -0.17759545  
## volatile.acidity -0.21873452 -0.41144893 0.53373510 -0.07877531  
## citric.acid 0.05857268 -0.06959338 -0.10549701 -0.37751558  
## residual.sugar -0.73214429 -0.04915555 -0.29066341 0.29984469  
## chlorides -0.24650090 -0.30433857 -0.37041337 -0.35700936  
## free.sulfur.dioxide 0.15915198 0.01400021 0.11659611 -0.20478050  
## total.sulfur.dioxide 0.22246456 -0.13630755 0.09366237 0.01903597  
## density -0.15707671 0.39115230 0.17048116 -0.23922267  
## pH -0.26752977 0.52211645 0.02513762 -0.56139075  
## sulphates -0.22596222 0.38126343 0.44746911 0.37460432  
## alcohol -0.35068141 -0.36164504 0.32765090 -0.21762556  
## PC9 PC10 PC11  
## fixed.acidity 0.194020908 0.24952314 0.639691452  
## volatile.acidity -0.129110301 -0.36592473 0.002388597  
## citric.acid -0.381449669 -0.62167708 -0.070910304  
## residual.sugar 0.007522949 -0.09287208 0.184029964  
## chlorides 0.111338666 0.21767112 0.053065322  
## free.sulfur.dioxide 0.635405218 -0.24848326 -0.051420865  
## total.sulfur.dioxide -0.592115893 0.37075027 0.068701598  
## density 0.020718675 0.23999012 -0.567331898  
## pH -0.167745886 0.01096960 0.340710903  
## sulphates -0.058367062 -0.11232046 0.069555381  
## alcohol 0.037603106 0.30301450 -0.314525906

Usamos lo que acabamso de obtener de PCA 7 componentes y intentamos establecer una regresion lineal

fitPCA <- lm(redwine$quality ~ redwine.pca$x[,1:7])  
summary(fitPCA)

##   
## Call:  
## lm(formula = redwine$quality ~ redwine.pca$x[, 1:7])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.70114 -0.37034 -0.06334 0.49300 1.94724   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.636023 0.016415 343.343 < 2e-16 \*\*\*  
## redwine.pca$x[, 1:7]PC1 0.050621 0.009327 5.427 6.61e-08 \*\*\*  
## redwine.pca$x[, 1:7]PC2 -0.225087 0.011832 -19.023 < 2e-16 \*\*\*  
## redwine.pca$x[, 1:7]PC3 -0.258946 0.013187 -19.637 < 2e-16 \*\*\*  
## redwine.pca$x[, 1:7]PC4 -0.032376 0.014908 -2.172 0.030 \*   
## redwine.pca$x[, 1:7]PC5 -0.083721 0.016765 -4.994 6.57e-07 \*\*\*  
## redwine.pca$x[, 1:7]PC6 0.025114 0.020218 1.242 0.214   
## redwine.pca$x[, 1:7]PC7 0.095257 0.021491 4.432 9.95e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6564 on 1591 degrees of freedom  
## Multiple R-squared: 0.3422, Adjusted R-squared: 0.3393   
## F-statistic: 118.3 on 7 and 1591 DF, p-value: < 2.2e-16

Se ve qeu el parámetro R-squared es muy bajo, por lo tanto la precisión del modelo no es correcta

### Regresion lineal

lm0 <- lm(quality ~ ., data=redwine)  
summary(lm0)

##   
## Call:  
## lm(formula = quality ~ ., data = redwine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.51156 -0.30857 -0.05853 0.41108 1.18315   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -24.730865 16.887574 -1.464 0.14327   
## fixed.acidity -0.019823 0.020643 -0.960 0.33707   
## volatile.acidity -0.848632 0.096403 -8.803 < 2e-16 \*\*\*  
## citric.acid -0.297001 0.116847 -2.542 0.01112 \*   
## residual.sugar -0.016876 0.011954 -1.412 0.15822   
## chlorides -1.010177 0.333909 -3.025 0.00252 \*\*   
## free.sulfur.dioxide 0.005090 0.001723 2.954 0.00319 \*\*   
## total.sulfur.dioxide -0.002387 0.000579 -4.123 3.93e-05 \*\*\*  
## density 30.456146 17.238997 1.767 0.07747 .   
## pH -0.436387 0.152038 -2.870 0.00416 \*\*   
## sulphates 0.453033 0.091986 4.925 9.32e-07 \*\*\*  
## alcohol 0.175803 0.021270 8.265 2.91e-16 \*\*\*  
## good.wine 1.317907 0.043115 30.567 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5142 on 1586 degrees of freedom  
## Multiple R-squared: 0.5976, Adjusted R-squared: 0.5946   
## F-statistic: 196.3 on 12 and 1586 DF, p-value: < 2.2e-16

lm1 <- lm(quality ~ volatile.acidity+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+pH+alcohol+sulphates, data=redwine)  
summary(lm1)

##   
## Call:  
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +   
## total.sulfur.dioxide + pH + alcohol + sulphates, data = redwine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.68918 -0.36757 -0.04653 0.46081 2.02954   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.4300987 0.4029168 10.995 < 2e-16 \*\*\*  
## volatile.acidity -1.0127527 0.1008429 -10.043 < 2e-16 \*\*\*  
## chlorides -2.0178138 0.3975417 -5.076 4.31e-07 \*\*\*  
## free.sulfur.dioxide 0.0050774 0.0021255 2.389 0.017 \*   
## total.sulfur.dioxide -0.0034822 0.0006868 -5.070 4.43e-07 \*\*\*  
## pH -0.4826614 0.1175581 -4.106 4.23e-05 \*\*\*  
## alcohol 0.2893028 0.0167958 17.225 < 2e-16 \*\*\*  
## sulphates 0.8826651 0.1099084 8.031 1.86e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6477 on 1591 degrees of freedom  
## Multiple R-squared: 0.3595, Adjusted R-squared: 0.3567   
## F-statistic: 127.6 on 7 and 1591 DF, p-value: < 2.2e-16

Casi precision total

# Comparemos   
summary(lm0)$r.squared

## [1] 0.5976089

summary(lm1)$r.squared

## [1] 0.3594709

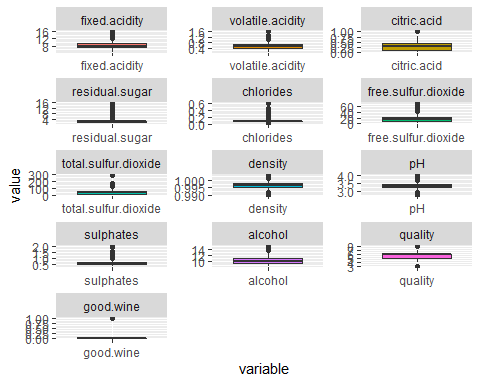
# 5. Representacion a partir de tablas y otros graficso y datos.

## Boxplot por cada varaible o atributo

redwine3<- melt(redwine)

## No id variables; using all as measure variables

ggplot(redwine3, aes(x=variable, y=value, fill=variable)) +  
geom\_boxplot() + facet\_wrap(~variable, scales="free", ncol=3) +   
theme(legend.position="none")



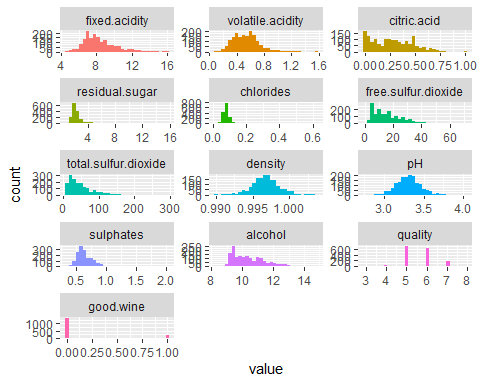
## Histogramas

redwine4 <- melt(redwine)

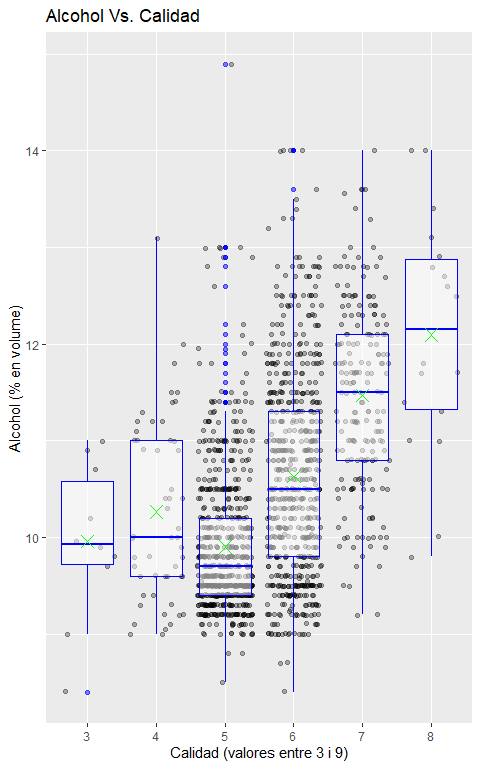
## No id variables; using all as measure variables

ggplot(redwine4, aes(value, fill=variable)) + facet\_wrap(~variable, scales = 'free', ncol=3) + geom\_histogram() + theme(legend.position="none")

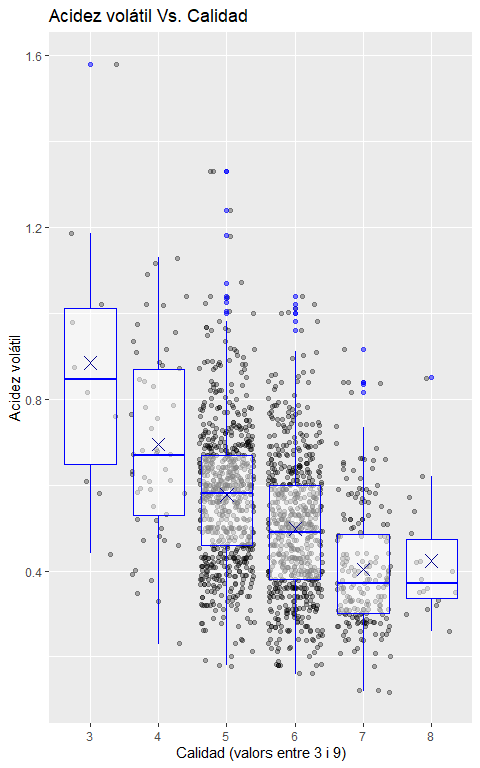
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



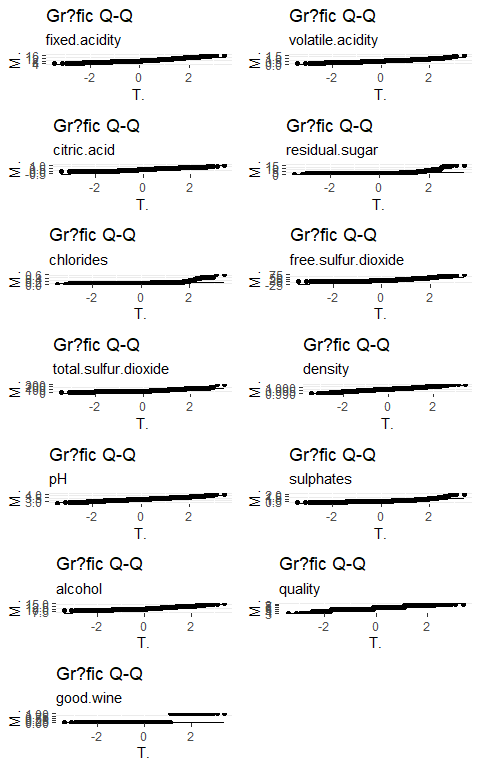
## Alcohol vs Qualitat



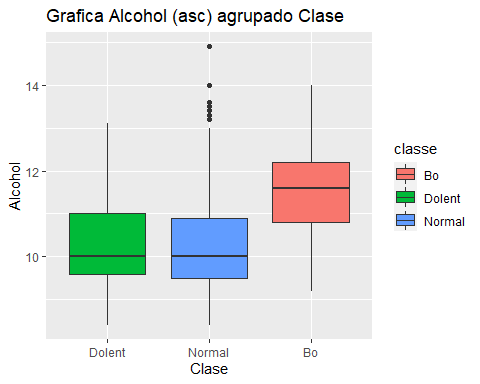
## Volatile acidity vs Calidiad



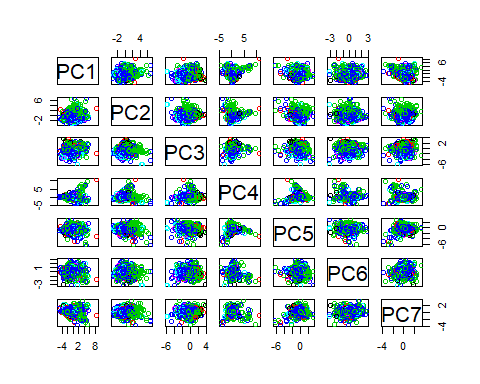
## Normalidad : Q-Q



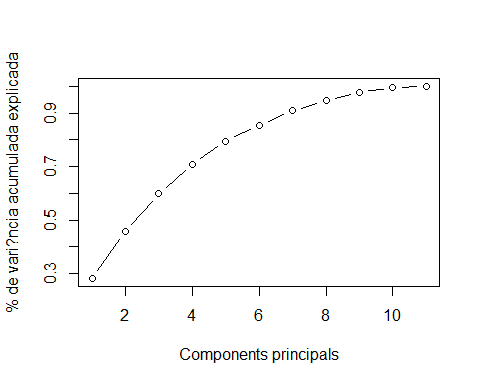
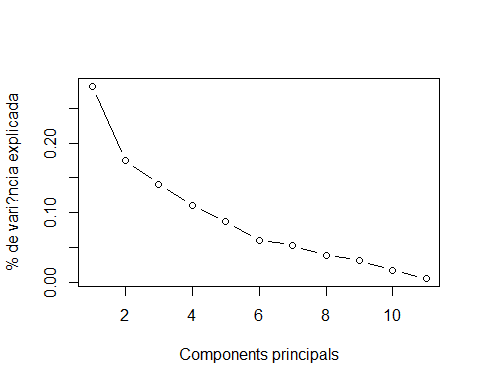
## Variancia : varainza segun los tipos de vino



## PCA : Grafico que muestran la relación



## PCA : graficos acumulado y parcial qu emuestra el % de la varianza



# 6. Resolucion del problema.

## A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Con los resultados podemos decir

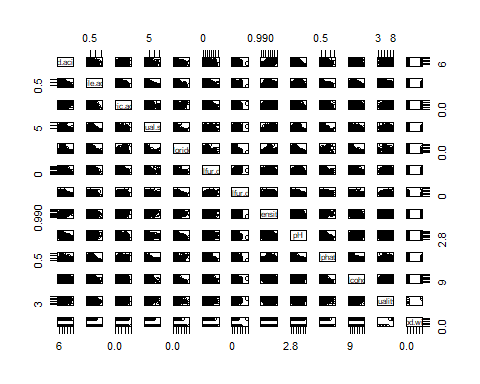
* la *regresion lineal* NO es un buen modelo. NO explica la calidad de un vino en funcionde los valore o atributos qu ehemso estudiado. Como hemos usado un valor subjetivo (**quality**) a partir de *valores objetivos* (alguna propiedd que se pueda medir). Seria mejor hacer una *discretitzacion* la calidad del vino en funcion de dos valore sbinarios (bueno=1, malo=0) y aplicar *regresion logistica*
* Per altra banda, el model de regressi? lineal probablement serviria per explicar, amb un cert grau de bondat, altres atributs, sense tenir en compte la qualitat.

# Bibliografia y referencias usadas

* Materials de l’assignatura ‘Tipolog?a i cicle de vida de les dades’, UOC
* Dataset de mostra : <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>
* <https://stackoverflow.com/questions/49044753/scale-kable-table-to-fit-page-width>
* <https://www.uv.es/conesa/CursoR/material/>
* <https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/>
* <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>
* <https://owi.usgs.gov/blog/boxplots/>
* <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Box%20Plot>
* <https://www.cyclismo.org/tutorial/R/pValues.html>
* <http://www.sthda.com>
* <http://r-statistics.co/Outlier-Treatment-With-R.html>
* <https://www.kaggle.com/tsilveira/wine-r/comments>
* <https://owi.usgs.gov/blog/boxplots/>
* <https://stackoverflow.com/questions/7196450/create-a-data-frame-of-unequal-lengths>
* <https://stackoverflow.com/questions/34004008/transposing-in-dplyr>
* <https://briatte.github.io/ggcorr/#controlling-the-main-geometry>
* <https://rpubs.com/Joaquin_AR/287787>
* <https://towardsdatascience.com>
* <https://tereom.github.io/est_computacional/01-Intro-R/R_analisis_datos_visualizacion>

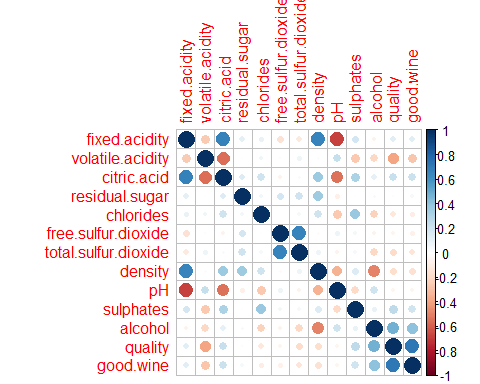
\*\*\* Otros trbajos preparado pero no terminados ##Analisis de las variables #Correlacion de Variables #Matriz de dispersion

plot(redwine)



# Visualizacion grafica de una matriz a color

#!!!  
#corrplot (corr (redwine))   
cor\_redwine<-cor(redwine)  
#cor\_redwine  
library(corrplot)  
corrplot(cor\_redwine,method = "circle")



corrplot(cor\_redwine,method = "number")



summary(redwine$quality)

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

### Mostrar el coeficiente de correlación :

### corrplot(redwine, method=“color”) !!!!!!!

### Como intento predecir la calidad del vino, lo que me intera las 2 ultimas filas/columnas

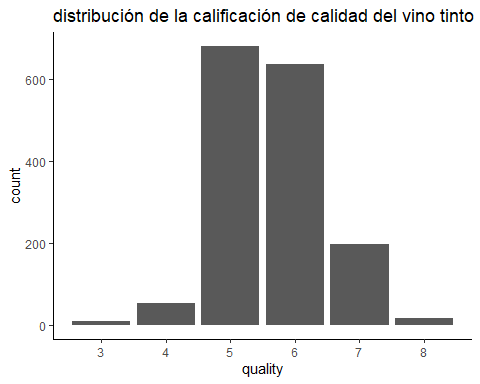
### quality y vino\_bueno

### Como sugiere el mapa de color, el alcohol tiene la correlación más fuerte con la calidad del vino.

### Calidad del vino

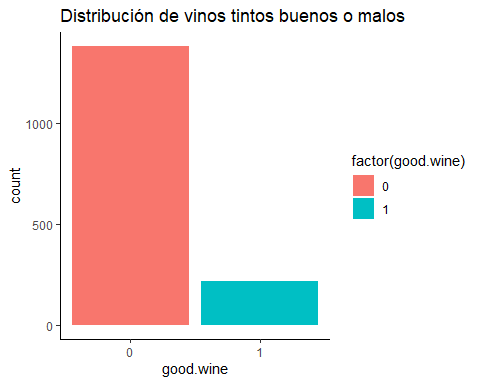
### Distribución de las calificaciones de calidad del vino tinto

library(ggplot2)  
  
ggplot(redwine,aes(x=quality))+geom\_bar(stat = "count",position = "dodge")+  
 scale\_x\_continuous(breaks = seq(3,8,1))+  
 ggtitle("distribución de la calificación de calidad del vino tinto")+  
 theme\_classic()



### Distribución de vinos tintos buenos o malos

ggplot(redwine,aes(x=good.wine,fill=factor(good.wine)))+geom\_bar(stat = "count",position = "dodge")+  
 scale\_x\_continuous(breaks = seq(0,1,1))+  
 ggtitle("Distribución de vinos tintos buenos o malos")+  
 theme\_classic()



### El grafico anterior, nso dice que hay mas buenos vinos que malos

### La mayoria son mediocres pues su calidad esta entre 5 o 6, hay algunsopobres entre 3 o 4.

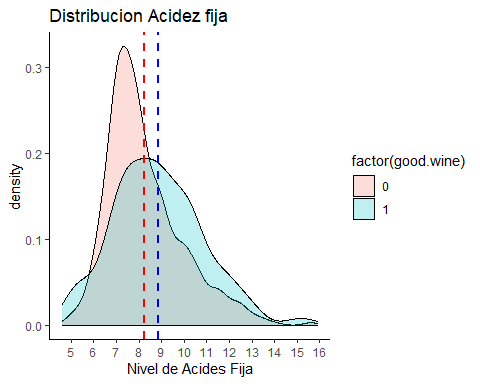
### La matoria de lso que podemso decir que son buenos estan 6.

### Propiedades fisicoquímicas y calidad del vino.

### la relación entre las propiedades fisicoquímicas y si un vino es bueno o no.

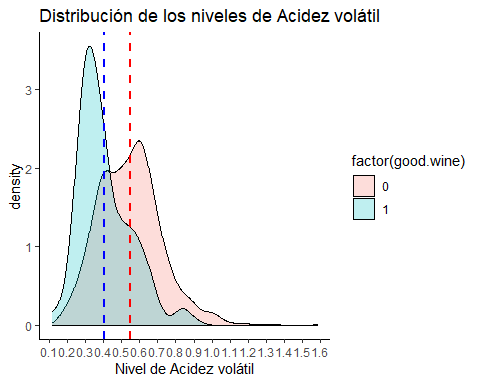
## Acidez fija y calidad del vino

ggplot(redwine,aes(x=fixed.acidity,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(fixed.acidity[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(fixed.acidity[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(4,16,1))+  
 xlab(label = "Nivel de Acides Fija")+  
 ggtitle("Distribucion Acidez fija")+  
 theme\_classic()



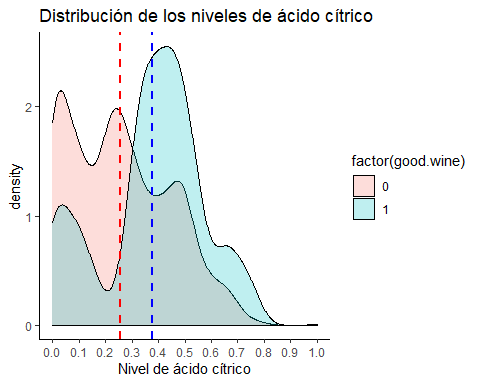
##Acidez volátil y calidad del vino

ggplot(redwine,aes(x=volatile.acidity,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(volatile.acidity[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(volatile.acidity[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0,1.6,0.1))+  
 xlab(label = "Nivel de Acidez volátil")+  
 ggtitle("Distribución de los niveles de Acidez volátil")+  
 theme\_classic()



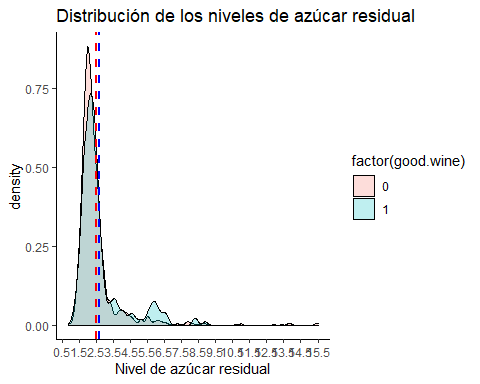
## Ácido cítrico y calidad del vino

ggplot(redwine,aes(x=citric.acid,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(citric.acid[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(citric.acid[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0,1,0.1))+  
 xlab(label = "Nivel de ácido cítrico")+  
 ggtitle("Distribución de los niveles de ácido cítrico")+  
 theme\_classic()



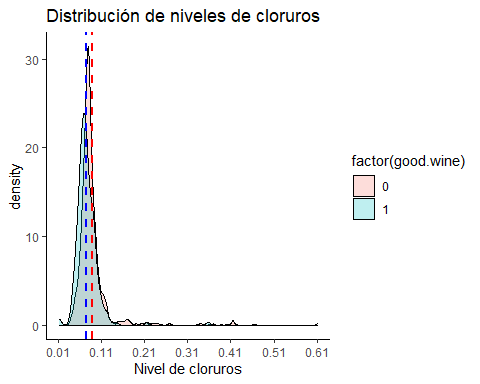
## Azúcar Residual y Calidad del Vino

ggplot(redwine,aes(x=residual.sugar,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(residual.sugar[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(residual.sugar[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0.5,15.5,1))+  
 xlab(label = "Nivel de azúcar residual")+  
 ggtitle("Distribución de los niveles de azúcar residual")+  
 theme\_classic()

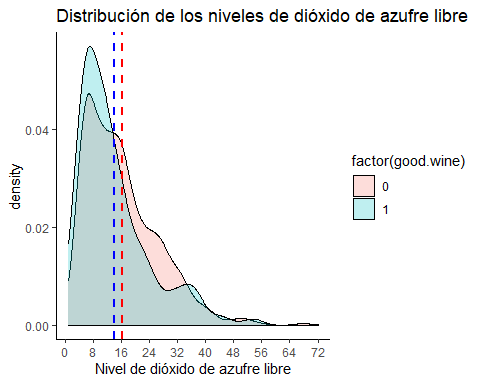


## Cloruros y calidad del vino

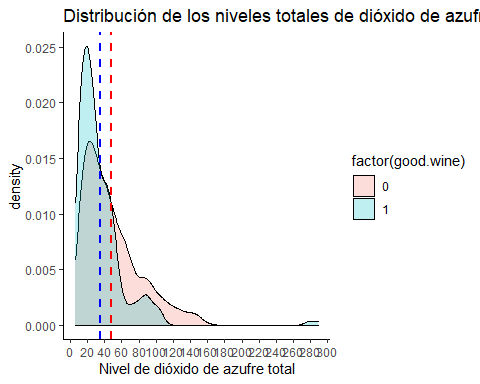
ggplot(redwine,aes(x=chlorides,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(chlorides[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(chlorides[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0.01,0.62,0.1))+  
 xlab(label = "Nivel de cloruros")+  
 ggtitle("Distribución de niveles de cloruros")+  
 theme\_classic()

 ## Dióxido de azufre libre y calidad del vino

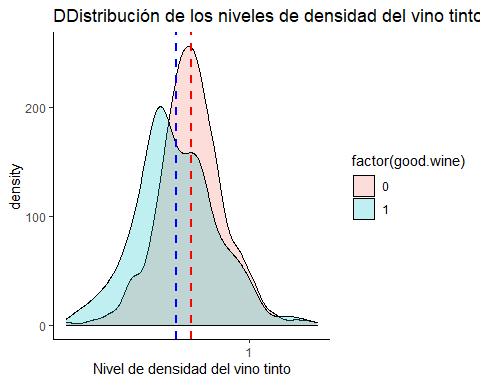
ggplot(redwine,aes(x=free.sulfur.dioxide,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(free.sulfur.dioxide[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(free.sulfur.dioxide[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0,72,8))+  
 xlab(label = "Nivel de dióxido de azufre libre")+  
 ggtitle("Distribución de los niveles de dióxido de azufre libre")+  
 theme\_classic()

 ## Dióxido de azufre total y calidad del vino

ggplot(redwine,aes(x=total.sulfur.dioxide,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(total.sulfur.dioxide[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(total.sulfur.dioxide[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0,300,20))+  
 xlab(label = "Nivel de dióxido de azufre total") +  
 ggtitle("Distribución de los niveles totales de dióxido de azufre")+  
 theme\_classic()

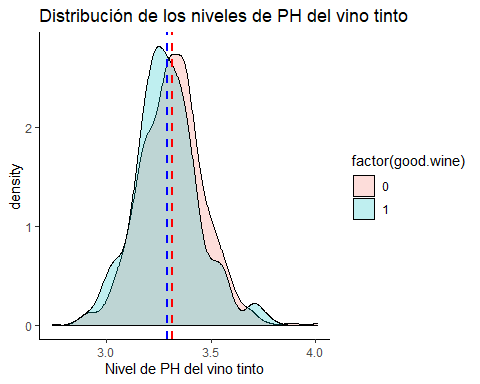
 ## Densidad y calidad del vino

ggplot(redwine,aes(x=density,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(density[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(density[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0.9,1.1,0.05))+  
 xlab(label = "Nivel de densidad del vino tinto")+  
 ggtitle("DDistribución de los niveles de densidad del vino tinto")+  
 theme\_classic()



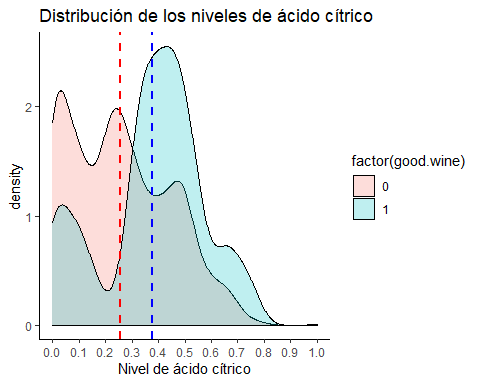
## PH y calidad del vino

ggplot(redwine,aes(x=pH,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(pH[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(pH[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(2.5,5,0.5))+  
 xlab(label = "Nivel de PH del vino tinto")+  
 ggtitle("Distribución de los niveles de PH del vino tinto")+  
 theme\_classic()



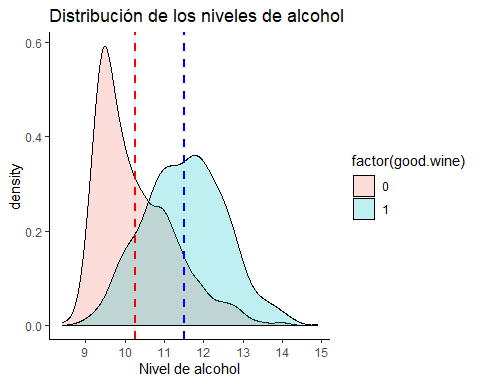
## Sulfatos y calidad del vino

ggplot(redwine,aes(x=citric.acid,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(citric.acid[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(citric.acid[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(0,1,0.1))+  
 xlab(label = "Nivel de ácido cítrico")+  
 ggtitle("Distribución de los niveles de ácido cítrico")+  
 theme\_classic()



#Alcohol y calidad del vino

ggplot(redwine,aes(x=alcohol,fill=factor(good.wine)))+geom\_density(alpha=0.25)+  
 geom\_vline(aes(xintercept=mean(alcohol[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+  
 geom\_vline(aes(xintercept=mean(alcohol[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+  
 scale\_x\_continuous(breaks = seq(8,15,1))+  
 xlab(label = "Nivel de alcohol")+  
 ggtitle("Distribución de los niveles de alcohol")+  
 theme\_classic()



# Con los graficos, me dado cuenta que lso vinos tintos,ya sean buenso o malos

# tienen una distribucion muy en sus propiedades fisioquimicas

# Los mas llamtivos en cuanto a salirse seria los niveles de sulfatos y el alcohol

# Modelo predictivo (clasificación binaria)

# utilizaríamos un random forest como nuestro modelo de referencia para predecir la calidad de un vino.

# No utilizaremos ningún ajuste de hyper-parámetro y nos quedaremos con el valor predeterminado

# de la función randomForest

#Modelo de bosque aleatorio de referencia

library(randomForest)  
#randomForest(formula = factor(good.wine) ~ . - quality, data = redwine, ntree = 150)  
redwineRF<-randomForest(factor(good.wine)~.-redwine$quality,redwine,ntree=150)  
redwineRF

##   
## Call:  
## randomForest(formula = factor(good.wine) ~ . - redwine$quality, data = redwine, ntree = 150)   
## Type of random forest: classification  
## Number of trees: 150  
## No. of variables tried at each split: 3  
##   
## OOB estimate of error rate: 0%  
## Confusion matrix:  
## 0 1 class.error  
## 0 1382 0 0  
## 1 0 217 0

# La precision de nuestro modelo esta bien, sobre el 92% .

# Aunque parece as facil decir que un vino va ser malo que decir que va ser bueno

#Variable Importance # Obtener importance

#  
importance<- importance(redwineRF)  
varImportance <- data.frame(Variables = row.names(importance),   
 Importance = round(importance[ ,'MeanDecreaseGini'],2))  
importance

## MeanDecreaseGini  
## fixed.acidity 7.461630  
## volatile.acidity 14.298730  
## citric.acid 10.202018  
## residual.sugar 5.148360  
## chlorides 5.862303  
## free.sulfur.dioxide 3.999865  
## total.sulfur.dioxide 7.803138  
## density 11.238324  
## pH 4.239046  
## sulphates 14.913752  
## alcohol 26.775523  
## quality 263.204043

varImportance

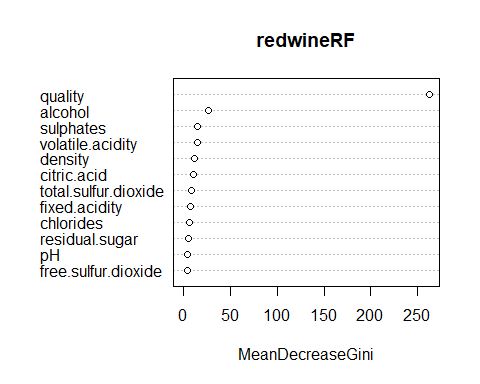
## Variables Importance  
## fixed.acidity fixed.acidity 7.46  
## volatile.acidity volatile.acidity 14.30  
## citric.acid citric.acid 10.20  
## residual.sugar residual.sugar 5.15  
## chlorides chlorides 5.86  
## free.sulfur.dioxide free.sulfur.dioxide 4.00  
## total.sulfur.dioxide total.sulfur.dioxide 7.80  
## density density 11.24  
## pH pH 4.24  
## sulphates sulphates 14.91  
## alcohol alcohol 26.78  
## quality quality 263.20

# Crear variable de rango basada en importance

# Create a rank variable based on importance  
rankImportance <- varImportance %>%  
 mutate(Rank = paste0('#',dense\_rank(desc(Importance))))

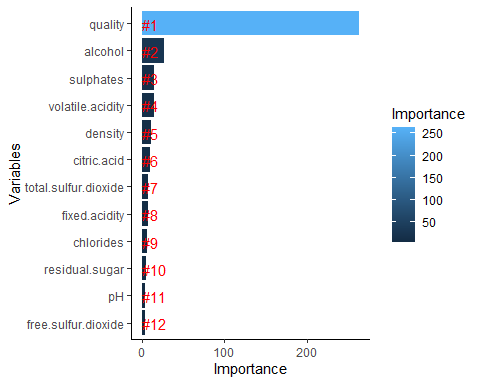
#Primer Grafico importence

varImpPlot(redwineRF)



#Segundp Grafico importence

# Use ggplot2 to visualize the relative importance of variables  
ggplot(rankImportance, aes(x = reorder(Variables, Importance),   
 y = Importance, fill = Importance)) +  
 geom\_bar(stat='identity') +   
 geom\_text(aes(x = Variables, y = 0.5, label = Rank),  
 hjust=0, vjust=0.55, size = 4, colour = 'red') +  
 labs(x = 'Variables') +  
 coord\_flip() +   
 theme\_classic()

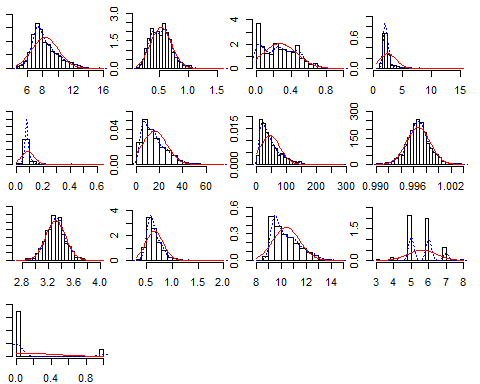


En este caso en particular, es necesario conocer el grado de influencia que tienen las variables respecto a la calidad del vino, así que se analizará como una regresión lineal

En primer lugar queremos comprobar que relacción existe entre las variables para buscar que no existe colinialidad (variables que se influyen entre ellas).Esta información es crítica para identificar las mejores variables predictoras

Para realizar dicho análisis es necesario calcular el coeficiente de correlacción de cada par de variables

multi.hist(x=redwine,dcol=c("blue", "red"),dlty = c("dotted","solid"),main="")



round(cor(x=redwine,method = "pearson"),3)

## fixed.acidity volatile.acidity citric.acid residual.sugar  
## fixed.acidity 1.000 -0.256 0.672 0.115  
## volatile.acidity -0.256 1.000 -0.552 0.002  
## citric.acid 0.672 -0.552 1.000 0.144  
## residual.sugar 0.115 0.002 0.144 1.000  
## chlorides 0.094 0.061 0.204 0.056  
## free.sulfur.dioxide -0.154 -0.011 -0.061 0.187  
## total.sulfur.dioxide -0.113 0.076 0.036 0.203  
## density 0.668 0.022 0.365 0.355  
## pH -0.683 0.235 -0.542 -0.086  
## sulphates 0.183 -0.261 0.313 0.006  
## alcohol -0.062 -0.202 0.110 0.042  
## quality 0.124 -0.391 0.226 0.014  
## good.wine 0.120 -0.271 0.215 0.048  
## chlorides free.sulfur.dioxide total.sulfur.dioxide density  
## fixed.acidity 0.094 -0.154 -0.113 0.668  
## volatile.acidity 0.061 -0.011 0.076 0.022  
## citric.acid 0.204 -0.061 0.036 0.365  
## residual.sugar 0.056 0.187 0.203 0.355  
## chlorides 1.000 0.006 0.047 0.201  
## free.sulfur.dioxide 0.006 1.000 0.668 -0.022  
## total.sulfur.dioxide 0.047 0.668 1.000 0.071  
## density 0.201 -0.022 0.071 1.000  
## pH -0.265 0.070 -0.066 -0.342  
## sulphates 0.371 0.052 0.043 0.149  
## alcohol -0.221 -0.069 -0.206 -0.496  
## quality -0.129 -0.051 -0.185 -0.175  
## good.wine -0.097 -0.072 -0.140 -0.150  
## pH sulphates alcohol quality good.wine  
## fixed.acidity -0.683 0.183 -0.062 0.124 0.120  
## volatile.acidity 0.235 -0.261 -0.202 -0.391 -0.271  
## citric.acid -0.542 0.313 0.110 0.226 0.215  
## residual.sugar -0.086 0.006 0.042 0.014 0.048  
## chlorides -0.265 0.371 -0.221 -0.129 -0.097  
## free.sulfur.dioxide 0.070 0.052 -0.069 -0.051 -0.072  
## total.sulfur.dioxide -0.066 0.043 -0.206 -0.185 -0.140  
## density -0.342 0.149 -0.496 -0.175 -0.150  
## pH 1.000 -0.197 0.206 -0.058 -0.057  
## sulphates -0.197 1.000 0.094 0.251 0.199  
## alcohol 0.206 0.094 1.000 0.476 0.407  
## quality -0.058 0.251 0.476 1.000 0.710  
## good.wine -0.057 0.199 0.407 0.710 1.000

Conclusiones: Las variables que tienen una mayor relacción lineal con la variable quality son:

* Alcohol 0.476
* Sulfatos 0.251

Comrpbamos si están relaccionadas entre ellas

cor(redwine$alcohol,redwine$sulphates)

## [1] 0.09359475

Comprobamos que no hay colinialidad entre ambas variables **(0.09359475).**

Ahora comprobamos con un modelos de regresión lineal multiple la influencia de las variables predictoras (todas menos quality) sobre la variable dependiente (quality)

calidadlm <- lm(quality ~ .,data=redwine)  
summary(calidadlm)

##   
## Call:  
## lm(formula = quality ~ ., data = redwine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.51156 -0.30857 -0.05853 0.41108 1.18315   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -24.730865 16.887574 -1.464 0.14327   
## fixed.acidity -0.019823 0.020643 -0.960 0.33707   
## volatile.acidity -0.848632 0.096403 -8.803 < 2e-16 \*\*\*  
## citric.acid -0.297001 0.116847 -2.542 0.01112 \*   
## residual.sugar -0.016876 0.011954 -1.412 0.15822   
## chlorides -1.010177 0.333909 -3.025 0.00252 \*\*   
## free.sulfur.dioxide 0.005090 0.001723 2.954 0.00319 \*\*   
## total.sulfur.dioxide -0.002387 0.000579 -4.123 3.93e-05 \*\*\*  
## density 30.456146 17.238997 1.767 0.07747 .   
## pH -0.436387 0.152038 -2.870 0.00416 \*\*   
## sulphates 0.453033 0.091986 4.925 9.32e-07 \*\*\*  
## alcohol 0.175803 0.021270 8.265 2.91e-16 \*\*\*  
## good.wine 1.317907 0.043115 30.567 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5142 on 1586 degrees of freedom  
## Multiple R-squared: 0.5976, Adjusted R-squared: 0.5946   
## F-statistic: 196.3 on 12 and 1586 DF, p-value: < 2.2e-16

Este modelo explica el 35% de los casos (variablidad) con todos los datos del dataset

Observamos que las variables independientes más influyentes en la variable dependiente quality (calidad del vino) son las siguientes:

* alcohol
* sulphates

Así creamos el siguiente modelo con solo las dos variables más predictoras (Sulfatos y Alcohol)

modelo<-lm(quality ~ sulphates+alcohol,data=redwine)  
summary(modelo)

##   
## Call:  
## lm(formula = quality ~ sulphates + alcohol, data = redwine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6685 -0.3781 -0.1005 0.4992 2.4187   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.37497 0.17745 7.748 1.64e-14 \*\*\*  
## sulphates 0.99409 0.10235 9.713 < 2e-16 \*\*\*  
## alcohol 0.34604 0.01628 21.256 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6905 on 1596 degrees of freedom  
## Multiple R-squared: 0.2699, Adjusted R-squared: 0.269   
## F-statistic: 295 on 2 and 1596 DF, p-value: < 2.2e-16

Ahora, con un modelo más simplificado, podemos explicar el 27% de la variabilida de la calidad.

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

Realizamos un análisis de inflacción de varianza de las variables prescriptoras anteriores

library(car)  
vif(modelo)

## sulphates alcohol   
## 1.008837 1.008837

sapply(redwine, ad.test)

## fixed.acidity volatile.acidity   
## statistic 28.14296 5.683075   
## p.value 3.7e-24 5.318894e-14   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## citric.acid residual.sugar   
## statistic 17.54209 188.0644   
## p.value 3.7e-24 3.7e-24   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## chlorides free.sulfur.dioxide   
## statistic 210.4492 38.60991   
## p.value 3.7e-24 3.7e-24   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## total.sulfur.dioxide density   
## statistic 52.48865 3.867595   
## p.value 3.7e-24 1.227494e-09   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## pH sulphates   
## statistic 1.864112 46.9322   
## p.value 9.245208e-05 3.7e-24   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## alcohol quality   
## statistic 34.91706 110.6328   
## p.value 3.7e-24 3.7e-24   
## method "Anderson-Darling normality test" "Anderson-Darling normality test"  
## data.name "X[[i]]" "X[[i]]"   
## good.wine   
## statistic 489.1072   
## p.value 3.7e-24   
## method "Anderson-Darling normality test"  
## data.name "X[[i]]"

Como el valor de inflación de la varianza en inferior a 5 , no se considera colinialidad: confirmamos que son buenos predictores

El valor del P-Value es inferior a 0.05 de Alcohol y Sulfatos por lo tanto no sigue una distribución normal, es más ninguna de las columnas lo siguen

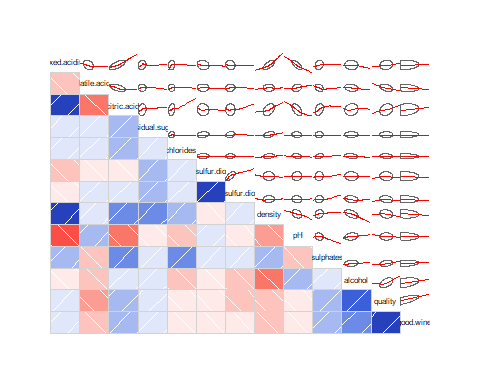
De lo que podemos asegurar que no hay correlacción lineal muy alta entre los predictores, por lo tanto son las variables que más influyen en la calidad del vino

### Aplicación de pruebas estadísticas para comparar los grupos de datos.

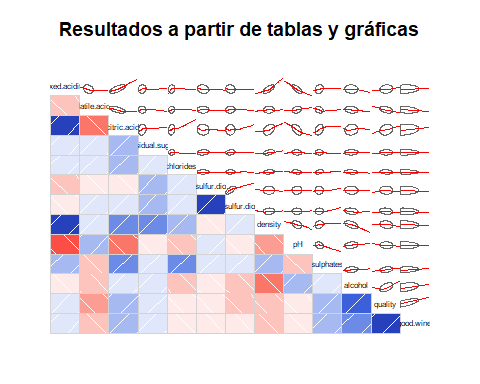
Se ha realizado la regresión lineal de arriba

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

corrgram(redwine, lower.panel=panel.shade, upper.panel=panel.ellipse)



corrgram(redwine, main="Resultados a partir de tablas y gráficas",lower.panel=panel.shade, upper.panel=panel.ellipse)



Comprobamos en la última final (quality) que los colores más azules son variables sin correlacción que más influyen en la calidad

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

Los resultados obtenidos nos demuestran que cuanto máyor sea el valor de la variable alcohol o sulfatos, mayor será la calidad del vino