Practica 2 - Tipologia y ciclo de vida de los datos

Aida Centelles Ahicart // Gonzalo Canales Nunez

May 22, 2019

```
#Leemos el fichero de datos
datos <-read.delim2("winequality-red.csv",header=TRUE,sep=",",dec=".")
n_variables<- names(datos)
```

---

1 Descripcion del dataset:
¿Por que es importante y que pregunta/problema pretende responder?

Incluye 1599 observaciones de 12 caracteristicas. 11 de ellas son variables quimicas (variables independientes), y la otra indica la calidad del vino (variable dependiente), una medida subjetiva que es la mediana de las opiniones de tres expertos en vinos. Especificamente, las caracteristicas son:

1. **Acidez fija**: la mayoria de los acidos relacionados con el vino o fijos o no volatiles (no se evaporan facilmente)
2. **Acidez volatil**: la cantidad de acido acetico en el vino, que a niveles demasiado altos puede provocar un sabor desagradable a vinagre
3. **Acido citrico**: se encuentra en pequeÃ Â±as cantidades, el acido citrico puede agregar ‘frescura’ y sabor a los vinos
4. **Azucar residual**: la cantidad de azucar que permanece en el vino despues de que se detenga la fermentacion, es infrecuente encontrar vinos con menos de 1 gramo/litro. Los vinos con mas de 45 gramos/litro se consideran dulces

5. **Cloruros**: la cantidad de sal en el vino.
6. **Dioxido de azufre libre**: la forma libre de SO2 existe en equilibrio entre el SO2 molecular (como un gas disuelto) y el ion bisulfito; Previene el crecimiento microbiano y la oxidacion del vino.
7. **Dioxido de azufre total**: cantidad de formas libres y ligadas de SO2; en bajas concentraciones, el SO2 es mayormente 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.
8. **Densidad**: la densidad del agua segÃ Âºn el porcentaje de alcohol y contenido de azÃ Âºcar
9. **pH**: describe como de acido o basico es un vino en una escala de 0 (muy acido) a 14 (muy basico); La mayoria de los vinos est?n entre 3-4 en la escala de pH.
10. **Sulfatos**: un aditivo para el vino que puede contribuir a los niveles de gas de dioxido de azufre (S02), que actÃ Âºa como un antimicrobiano y antioxidante.
11. **Alcohol**: el porcentaje de alcohol del vino
12. **Calidad**: Variable de salida (basada en datos sensoriales, con valores entre 0 y 10)

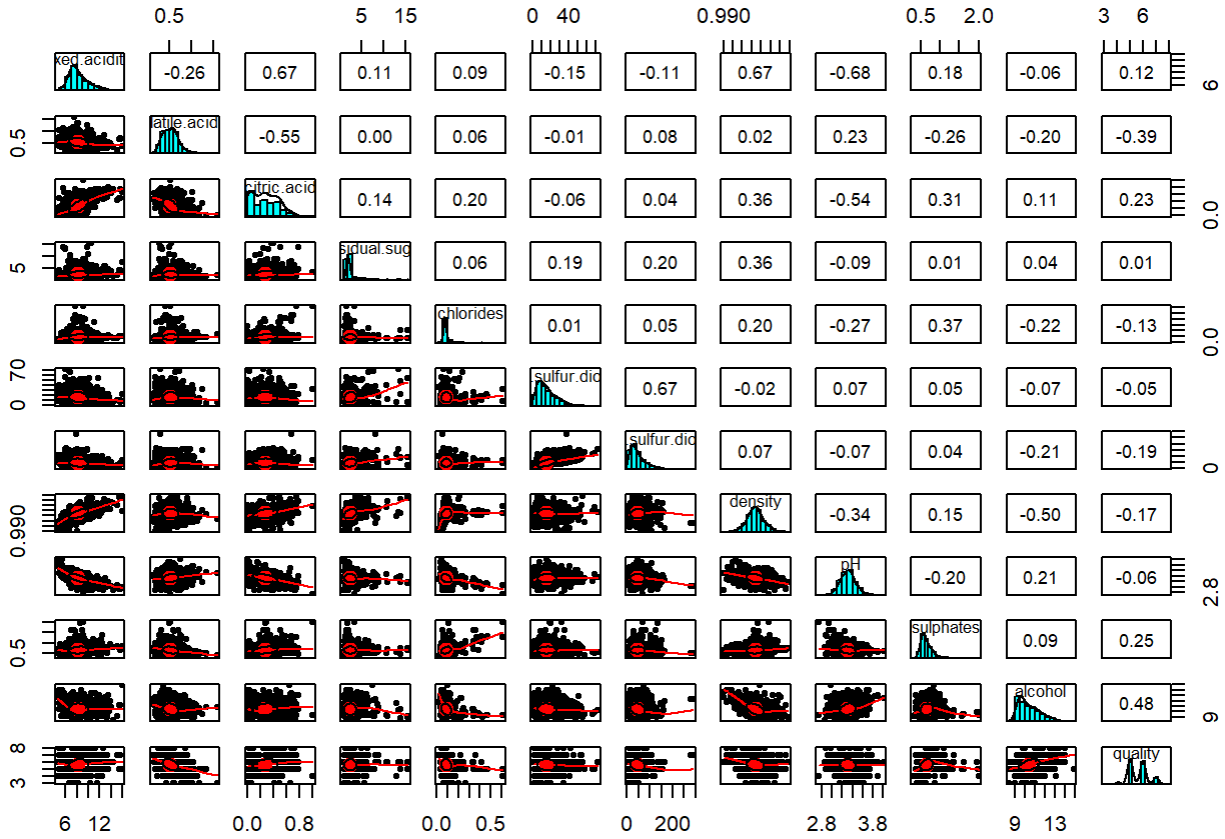
En esta practica limpiaremos los datos y trateremos de estimar un modelo que a partir de los datos nos pueda predecir la calidad de un vino.

---

2 Integracion y seleccion de los datos de interes a analizar

A priori, todas las variables van a ser usadas en el **analisis para predecir la calidad del vino**. En este caso, calidad (valor subjetivo aportado por los expertos) es la variable dependiente y el resto seran las idependientes. S

```
pairs.panels(datos)
```



3 Limpieza de los datos

Como se ha dicho anteriormete, el fichero incluye 1599 observaciones de 12 caracteristicas. 11 de ellas son variables quimicas (variables independientes), y la otra indica la calidad del vino (variable dependiente).

3.1 ¿Los datos contienen ceros o elementos vacios? ¿Como gestionarias cada uno de estos casos?

Comprobamos como se ha realizado la asignacion a cada variable.

```
res <- sapply(datos,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
fixed.acidity	numeric
volatile.acidity	numeric
citric.acid	numeric
residual.sugar	numeric
chlorides	numeric
free.sulfur.dioxide	numeric
total.sulfur.dioxide	numeric
density	numeric
pH	numeric
sulphates	numeric

alcohol numeric
quality integer

```
sapply(datos, function(x) sum(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
```

Vemos que todas las variables son numericas y que no hay valores nulos.

A continuacion, analizamos el bumero de ceros

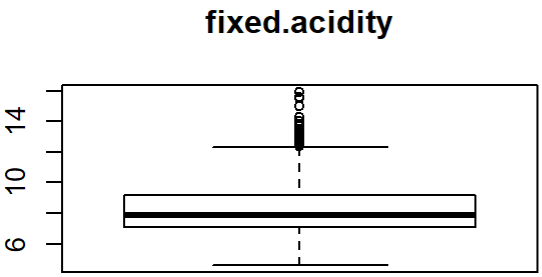
1. **Acidez fija**: Hay 0 Vinos cuya acidez fija es 0.
2. **Acidez volatil**: Hay 0 Vinos cuya acidez volatil es 0.
3. **Acido citrico**: Hay 132 Vinos cuyo nivel de acido citrico es 0.
4. **Azucar residual**: Hay 0 Vinos cuyo nivel de azucar residual es 0.
5. **Cloruros**: Hay 0 Vinos cuyo nivel de cloruros es 0.
6. **Dioxido de azufre libre**: Hay 0 Vinos cuyo nivel de di?xido de azufre libre es 0.
7. **Dioxido de azufre total**: Hay 0 Vinos cuyo nivel de di?xido de azufre total es 0.
8. **densidad**: Hay 0 Vinos cuya densidad es 0.
9. **pH**: Hay 0 Vinos cuyo ph es 0.
10. **Sulfatos**: Hay 0 Vinos cuya nivel de sulfatos es 0.
11. **alcohol**: Hay 0 Vinos cuya nivel de sulfatos es 0.
12. **quality**: Hay 0 Vinos cuya nivel de sulfatos es 0.

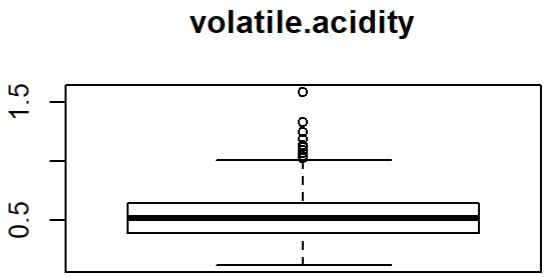
La unica variable que tiene valores cero es el nivel de acido citrico. Sin embargo, entra dentro de los parametros normales que un vino carezca de acido citrico, por tanto, se puede aceptar como un valor normal.

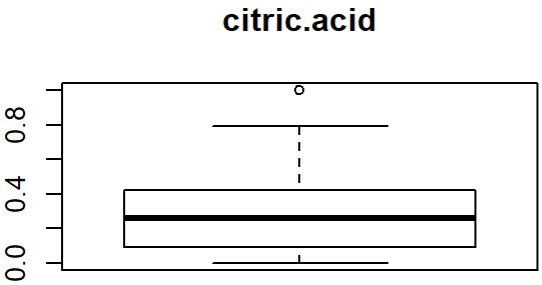
3.2 Identificacion y tratamiento de valores extremos

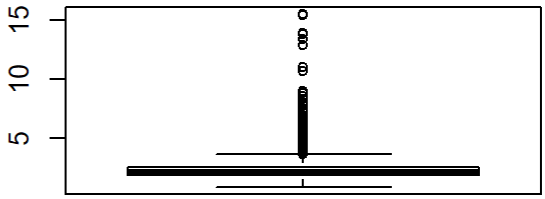
Un **outlier** es una **observacion que parece inconsistente con el resto de los valores de la muestra**, siempre teniendo en cuenta el modelo probabilistico supuesto que debe seguir la muestra. Para identificarlos, lo mas habitual es representar graficamente cada una de las variables en forma de boxplots.

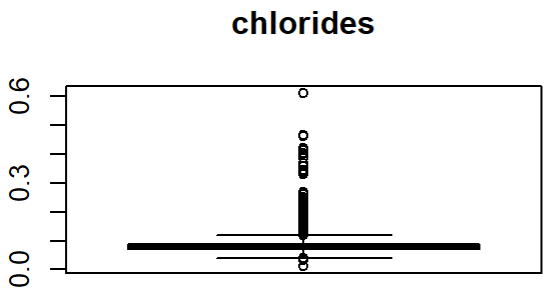
```
par(mfrow=c(2,2))
for(i in 1:ncol(datos)) {
  if (is.numeric(datos[,i])){
    boxplot(datos[,i], main = colnames(datos)[i], width = 100)
  }
}
```

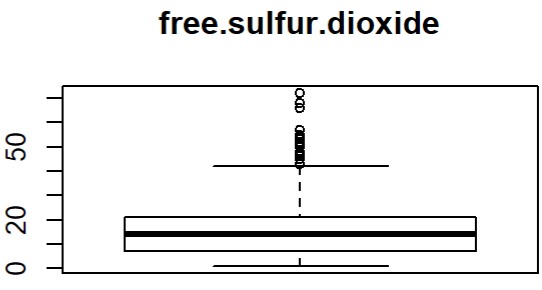

fixed.acidity
6
10
14

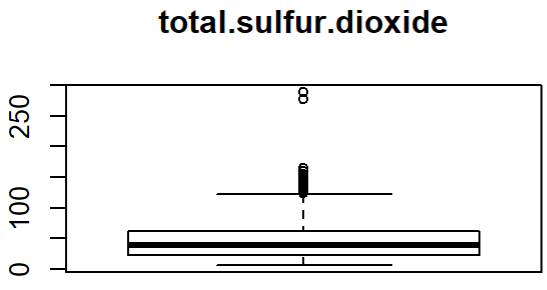

volatile.acidity
0.5
1.5

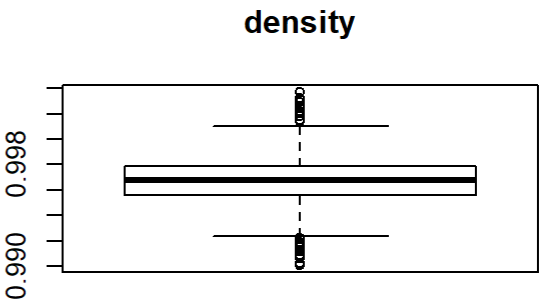

citric.acid
0.0
0.4
0.8

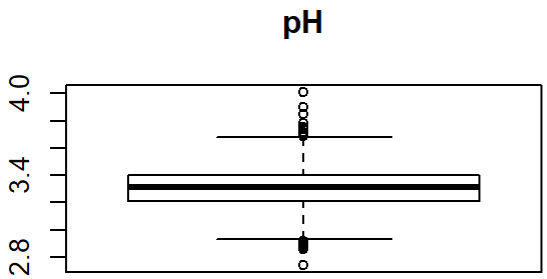

residual.sugar
5
10
15

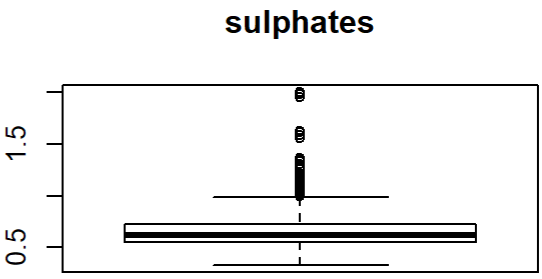

chlorides
0.0
0.3
0.6

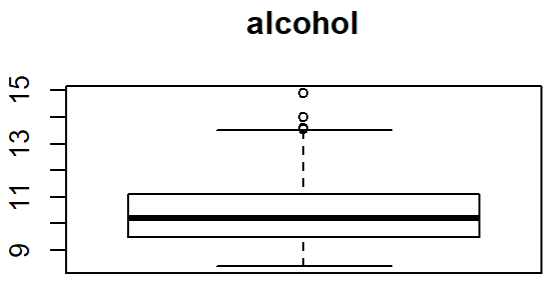

free.sulfur.dioxide
0
20
50

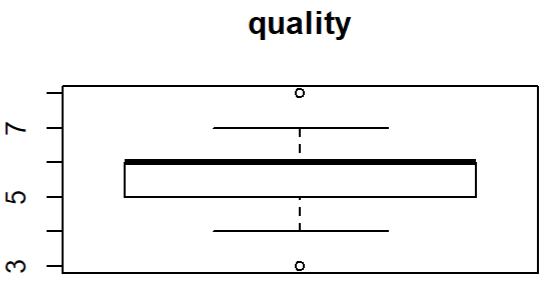

total.sulfur.dioxide
0
100
250


density
0.990
0.998


pH
2.8
3.4
4.0


sulphates
0.5
1.5


alcohol
9
11
13
15


quality
3
5
7

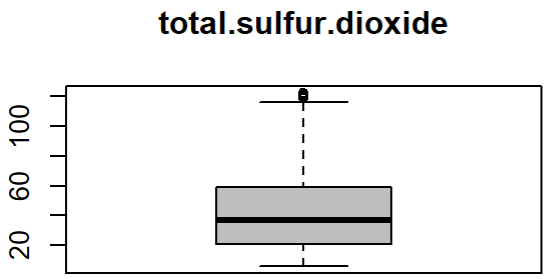
```
par(mfrow=c(1,1))
```

Como se observa en los boxplots, existen outliers en todas las variables. Vemos que la variable que tiene valores mas desviados con respecto a la media estan en la variable Dioxido de azufre total (total.sulfur.dioxide). Para evitar errores en el analisis, procedemos a eliminarlos.

```
filas_antes<-nrow(datos)

outliers_total.sulfur.dioxide <- boxplot(datos$total.sulfur.dioxide, plot=FALSE)$out

datos<- datos[-which( datos$total.sulfur.dioxide %in% outliers_total.sulfur.dioxide),]

filas_despues<-nrow(datos)
```

Tras la eliminacion de outliers hemos pasado de 1599 filas a 1544 filas. A continuacion, volvemos a representar la variable en un boxplot

```
par(mfrow=c(2,2))

boxplot(datos[,7], main=names(datos)[7],col="gray")

par(mfrow=c(1,1))
```



4 Analisis de los datos

En la etapa del analisis de datos, seleccionaremos primero los grupos de datos que queremos analizar, comprobaremos la **normalidad y homogeneidad de la varianza** y, finalmente, aplicaremos pruebas estadisticas para comparar los grupos de datos y realizar una prediccion o clasificacion.

4.1 Seleccion de los grupos de datos que se quieren analizar/comparar (planificacion de los analisis a aplicar)

Antes de empezar a planificar el analisis de los datos, es de gran ayudar hacer un resumer de los datos, observando la **estructura y distribucion de los datos**, asi como **visualizar los datos**, mediante histogramas de los atributos y diagramas de dispersion, que permiten el estudio de las relaciones entre variables.

Estructura y distribucion de los datos: Primero de todo, calculamos las estadisticas mas importantes (valor max, valor min, media, mediana, etc.) de cada atributo mediante la funcion “summary()”:

```
summary(datos)
```

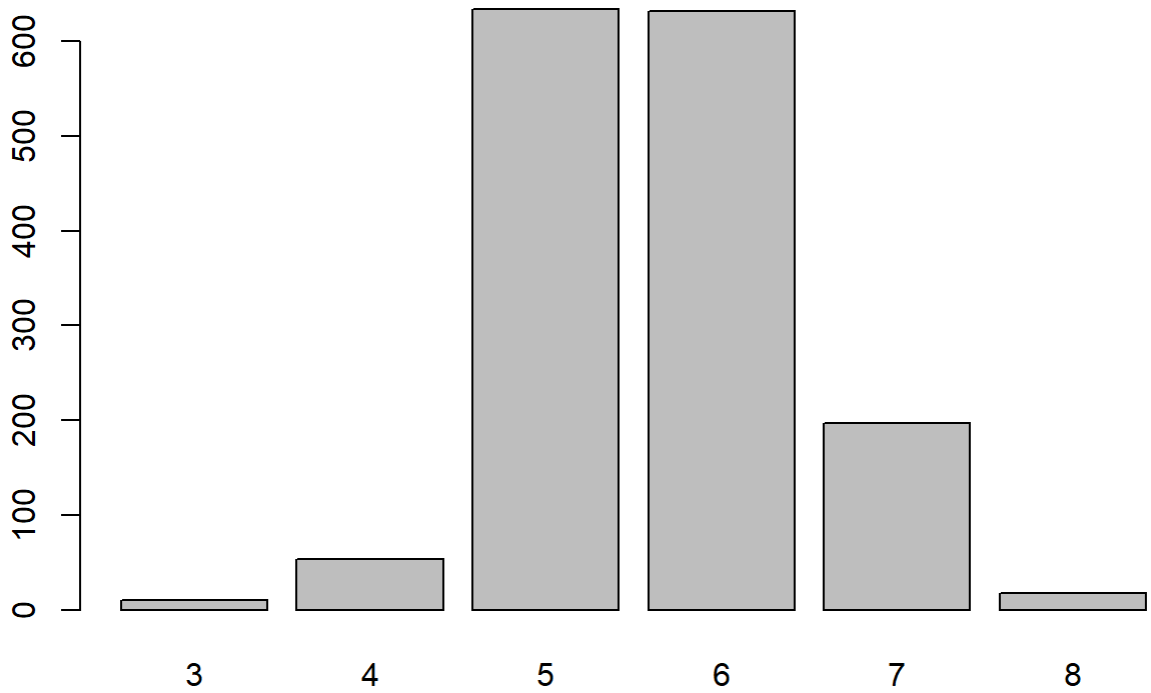
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 4.600   Min.   :0.1200   Min.   :0.0000   Min.   : 0.900
##  1st Qu.: 7.100   1st Qu.:0.3900   1st Qu.:0.0900   1st Qu.: 1.900
##  Median : 7.900   Median :0.5200   Median :0.2500   Median : 2.200
##  Mean   : 8.329   Mean   :0.5266   Mean   :0.2690   Mean   : 2.508
##  3rd Qu.: 9.300   3rd Qu.:0.6350   3rd Qu.:0.4225   3rd Qu.: 2.600
##  Max.   :15.900   Max.   :1.5800   Max.   :1.0000   Max.   :15.500
##    chlorides      free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.0120   Min.   : 1.00       Min.   :  6
##  1st Qu.:0.0700   1st Qu.: 7.00       1st Qu.: 21
##  Median :0.0790   Median :13.00       Median : 37
##  Mean   :0.0874   Mean   :15.27       Mean   : 43
##  3rd Qu.:0.0900   3rd Qu.:21.00       3rd Qu.: 59
##  Max.   :0.6110   Max.   :66.00       Max.   :122
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9967   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.314   Mean   :0.6559   Mean   :10.44
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.652
##  3rd Qu.:6.000
##  Max.   :8.000
```

La funcion summary nos ha proporcionado muchas informaciones sobre los datos de los que disponemos, como por ejemplo la mediana, los valores medios, max/min y los cuartiles.

Visualizacion de datos:

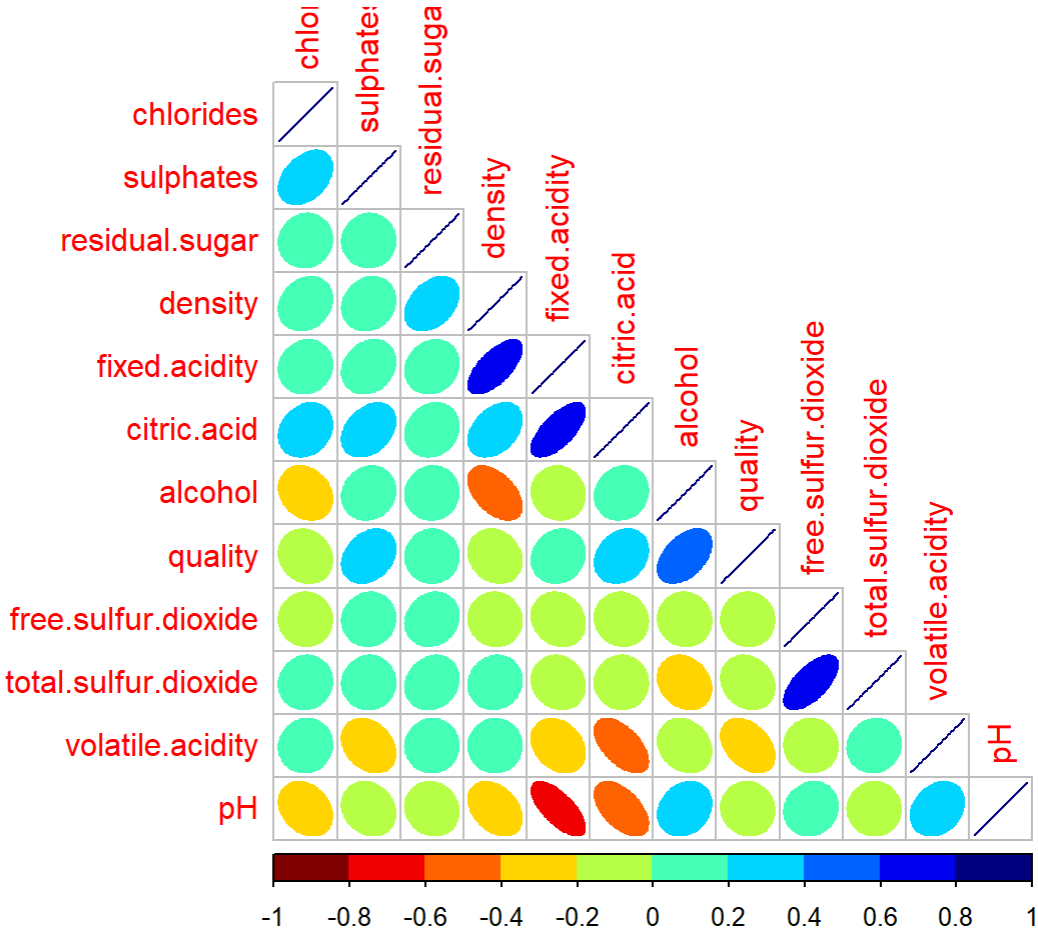
Empezaremos fijandonos en el atributo *quality*, que representa la calidad de los vinos del conjunto de datos. La calidad de la mayoria de vinos esta comprendida entre unos 5 (primer cuartil) y 6 (tercer cuartil), con una media de 5,6, sin embargo, observamos que hay vinos con una minima calidad de 3 y una maxima de 8.:

```
barplot(table(datos$quality))
```



Seleccion de variables:

Dado el gran numero de variables, representaremos de forma grafica el indice de correlacion entre variables. Para ello, se usara la libreria el metodo corrplot.



El grafico de correlacion muestra ahora existe una fuertes correlaciones positivas entre:
- dioxido de azufre total y el dioxido de azufre libre
- el acido citrico y la acided fija
- la acidez fija y la densidad
- en menor medida, entre la calidad y el alcohol

Tambien existen fuertes correlaciones negativas entre:
- el PH y la acidez fija
- en menor medida entre el alcohol y la densidad, la acidez volatil y el acido citrico, o el pH y el acido citrico

Observamos que las variables acido citrico y nivel de acidez (*citric.acid* y *fixed.acidity*), asi como los dioxidos de sulfuro (*total.sulfur.dioxide* y *free.sulfur.dioxide*) estan muy correlactionadas.

Hemos decidido por ello desestimar las variables *citric.acid* y *free.sulfur.dioxide* del conjunto de datos, para reducir el volumen de datos y la complejidad de los modelos a usar. A partir de ahora trabajaremos entonces con el conjunto de datos sin estas dos variables:

```
datos$citric.acid <- NULL
datos$free.sulfur.dioxide <- NULL
```

Analisis a realizar: Los analisis que vamos a realizar en esta practica son los siguientes:
- **Analisis de correlaciones**: para observar las correlaciones entre variables
- **Contrastes de hipotesis:** comprobar la calidad del vino con respecto al porcentaje de alcohol
- **Analisis de regresion:** regresion de la calidad con diferentes variables, como alcohol, sulfatos y acidez
- **Modelo de prediccion:** creacion de un modelo de prediccion de calidad de vinos (en calidad baja, media o alta) en funcion a las variables existentes

---

4.2 Comprobacion de la normalidad y homogeneidad de la varianza.

Comprobacion de normalidad:
El **test de Shapiro-Wilk** se considera uno de los metodos mas potentes para **contrastar la normalidad**. Asumiendo como hipotesis nula que la poblacion esta distribuida normalmente, si el p-valor es menor al nivel de significancia, generalmente = 0.05, entonces la hipotesis nula es rechazada y se concluye que los datos no cuentan con una distribucion normal.

Mediante el test de Shapiro investigamos si la distribucion de *quality* es normal:

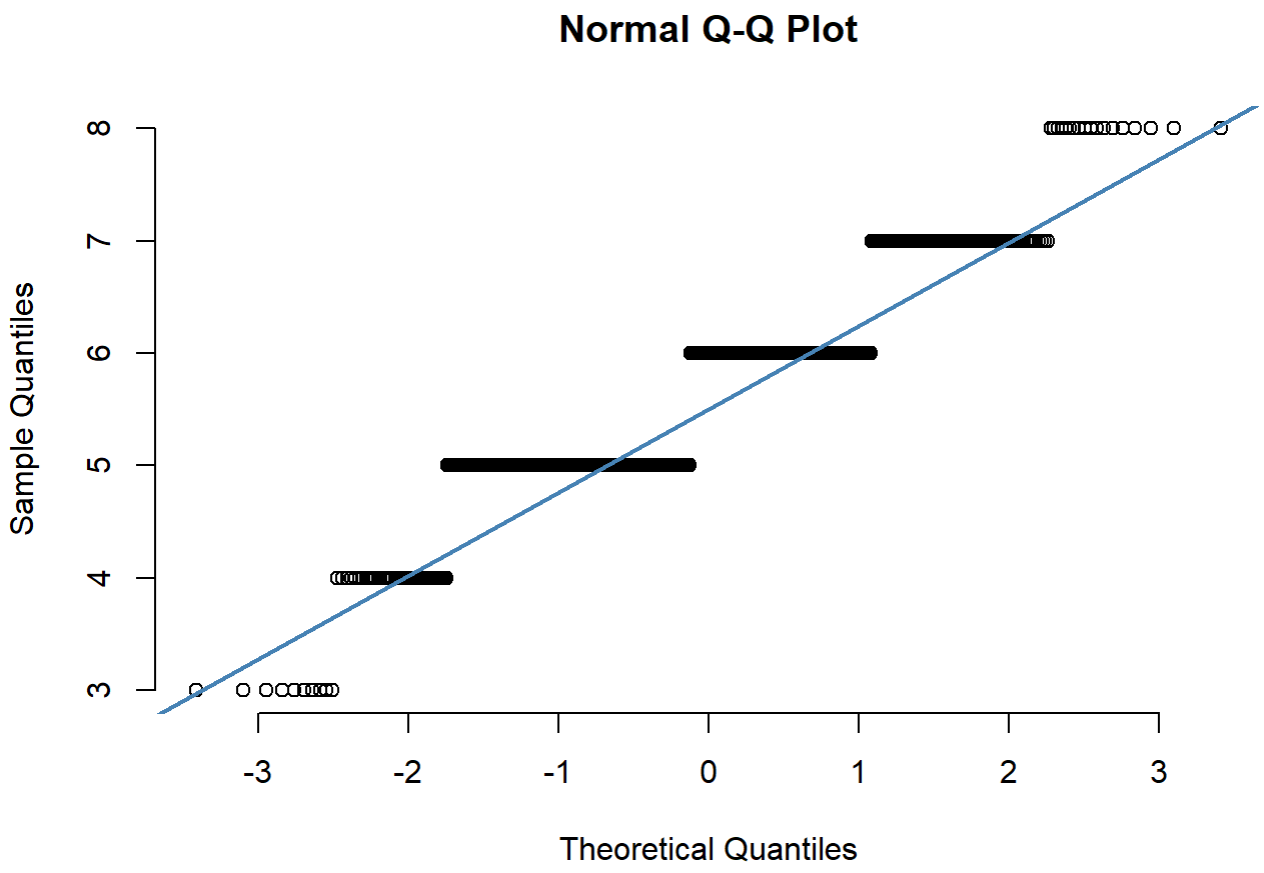
```
shapiro.test(datos$quality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$quality
## W = 0.86198, p-value < 2.2e-16
```

El valor W del test de Shapiro es W=0.969, muy cercano a 1, sin embargo p<0.05 lo que significa que debemos rechazar la hipotesis nula que indica que la distribucion es normal. Por lo tanto la distribucion de valores no sigue una distribucion normal.

Creamos tambien un grafico con qqplot(), que nos muestra los valores por cuartiles. AÃ Â±adimos una linea de referencia azul para comprobar si la distribucion de valores se aproxima a ella, para poder asumir normalidad (o como en este caso, no):

```
qqnorm(datos$quality, pch = 1, frame = FALSE)
qqline(datos$quality, col = "steelblue", lwd = 2)
```



Comprobacion de homogeneidad:
Algunas pruebas estadisticas requieren la comprobacion previa de la homocedasticidad en los datos, es decir, de la igualdad de varianzas entre los grupos que se van a comparar. Entre las pruebas mas habituales se encuentran el **test de Levene**, que se aplica cuando los datos siguen una distribucion normal, asi como el test de **Fligner-Killeen**, que se trata de la alternativa no parametrica, utilizada cuando los datos no cumplen con la condicion de normalidad. En ambas pruebas, la hipotesis nula asume igualdad de varianzas en los diferentes grupos de datos, por lo que p-valores inferiores al nivel de significancia indicaran heterocedasticidad.

En este caso nos interesa comprobar la homogeneidad de varianzas de las variables calidad y alcohol en dos subconjuntos de los datos.

Discretizamos la variable calidad:

```
table(discretize(datos$quality, "frequency", breaks=3))
```

```
##
## [3,5) [5,6) [6,8]
##    63   634   847
```

Creamos la nueva variable quality-disc como la variable discretizada, incluyendo los tres grupos de vinos. Observamos que el primer intervalo (vinos malos) se extiende de 3 a 5 en nivel de calidad y el segundo de 5 a 6 (vinos normales):

```
datos$qualitydisc <- ifelse(datos$quality <5, 'bad', ifelse(datos$quality >= 6,'good','
normal'))
datos$qualitydisc <- as.factor(datos$qualitydisc)
```

A continuacion iniciamos el test de Levene:

```
leveneTest(alcohol ~ datos$qualitydisc, data = datos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    2  69.183 < 2.2e-16 ***
##       1541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tambien usamos el test de Fligner, que es mas indicado para distribuciones no normales:

```
fligner.test(alcohol ~ datos$qualitydisc, data = datos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  alcohol by datos$qualitydisc
## Fligner-Killeen:med chi-squared = 136.67, df = 2, p-value <
## 2.2e-16
```

Dado que ambas pruebas resultan en un p-valor inferior al nivel de significancia (<0.05), se rechaza la hipotesis nula de homocedasticidad y se concluye que la variable alcohol presenta varianzas estadisticamente diferentes para los diferentes grupos de calidad de vino.

Nos deshacemos de la variable *qualitydisc* para continuar la practica:

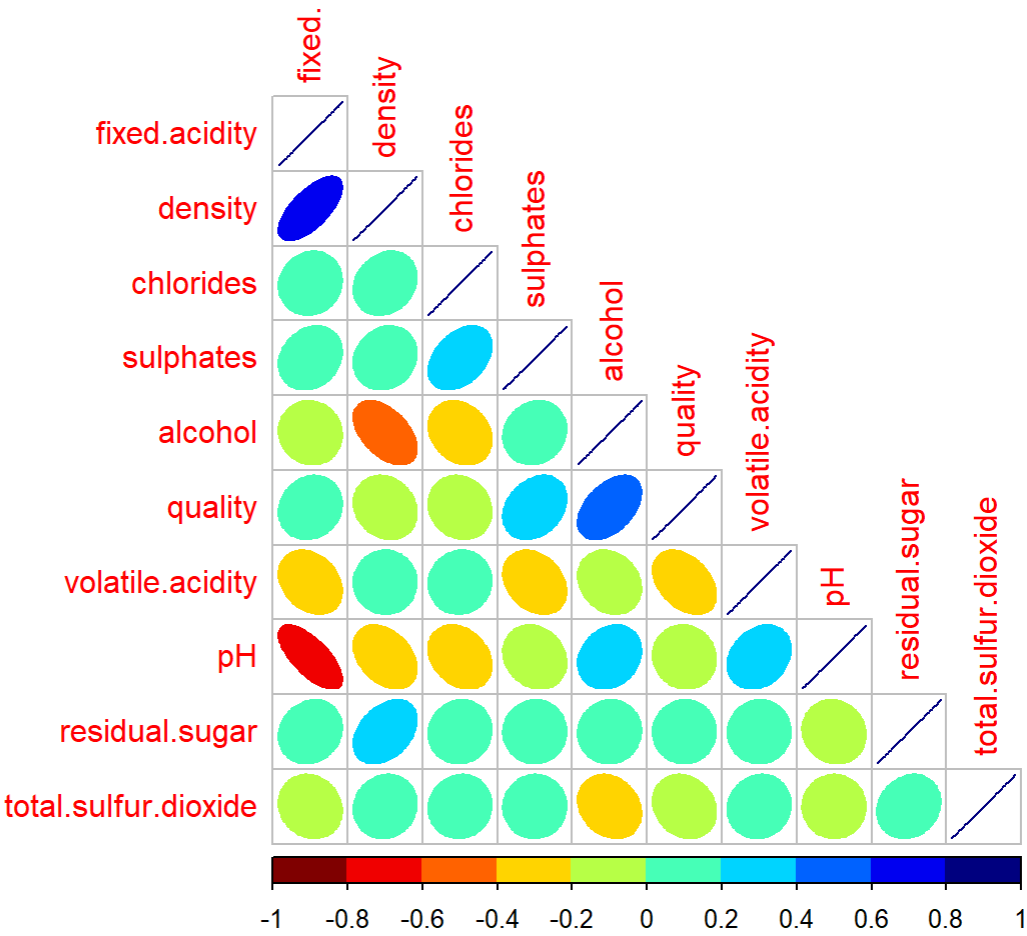
```
datos$qualitydisc <- NULL
```

4.3 Aplicacion de pruebas estadisticas para comparar los grupos de datos. En funcion de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipotesis, correlaciones, regresiones, etc. Aplicar al menos tres metodos de analisis diferentes.

4.3.1 Correlaciones

Como hemos realizado en el apartado 4.1. representaremos de forma grafica el indice de correlacion entre variables. Para ello, se usara la libreria corrplot y su metodo corrplot.

Las variables *citric.acid* y *free.sulfur.dioxide* han sido desestimadas del conjunto de datos en el apartado 4.1, por lo tanto ya no estan disponibles:



El grafico de correlacion muestra ahora existe una fuertes correlaciones positivas entre:
- la acidez fija y la densidad
- la calidad y el alcohol

Tambien existen fuertes correlaciones negativas entre:
- el PH y la acidez fija
- en menor medida entre el alcohol y la densidad

Para confirmar los resultados obtenidos en el grafico anterior, se podria calcular el indice de correlacion utilizando metodos como **Pearson** si

las variables se distribuyen de forman normal u otros tests como **Spearman**, en el caso de no seguir una distribucion normal. La normalidad de la distribucion de cada variable se puede estimar usando el test de Shapiro-Wilk para cada una de las variables.

```
shapiro.test(datos$volatile.acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$volatile.acidity
## W = 0.97459, p-value = 6.933e-16
```

```
shapiro.test(datos$fixed.acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$fixed.acidity
## W = 0.94145, p-value < 2.2e-16
```

```
shapiro.test(datos$residual.sugar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$residual.sugar
## W = 0.56421, p-value < 2.2e-16
```

```
shapiro.test(datos$chlorides)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$chlorides
## W = 0.47881, p-value < 2.2e-16
```

```
shapiro.test(datos$total.sulfur.dioxide)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$total.sulfur.dioxide
## W = 0.91438, p-value < 2.2e-16
```

```
shapiro.test(datos$density)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$density
## W = 0.9911, p-value = 4.536e-08
```

```
shapiro.test(datos$pH)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$pH
## W = 0.99334, p-value = 2.001e-06
```

```
shapiro.test(datos$sulphates)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$sulphates
## W = 0.88024, p-value < 2.2e-16
```

```
shapiro.test(datos$alcohol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$alcohol
## W = 0.93278, p-value < 2.2e-16
```

```
shapiro.test(datos$quality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$quality
## W = 0.86198, p-value < 2.2e-16
```

En este test, como ya se indico anteriormente, la hipotesis nula indica que la poblacion sigue una distribucion normal. En todos los casos anteriores, las variables tienen un p-valor inferior al nivel de signficancia de 0.05, lo cual indica que se debe rechazar la hipotesis nula. En otras palabras, **ninguna de las variables siguen una distribucion normal**.

Por tanto, para medir el indice de correlacion entre variables, deberiamos rechazar Pearson y usar otros metodos de calculo de correlacion como el de rangos de Spearman. Para confirmar los resultados obtenidos en el grafico anterior, calcularemos el indice de correlacion para

para dos pares de variables con fuerte correlacion, y para dos pares de variables, con bajo indice de correlacion.

```
## [1] 0.6262596
```

```
## [1] -0.7191249
```

```
## [1] 0.172121
```

```
## [1] -0.242061
```

4.3.2 Contraste de hipotesis

Una vez que hemos visto la correlacion entre variables, vamos a comprobar la *relacion entre calidad del vino con respecto al porcentaje de alcohol* mediatne un contraste de hipotesis.

En primer lugar vamos a clasificar la calidad de los vinos entre buenos (aquellos con una calificacion igual o superior a 6) y normales (aquellos con una calidad inferior a 6)

```
normal <-c(datos$alcohol[datos$quality < 6])
good<-c(datos$alcohol[datos$quality >= 6])
```

Asi pues vemos que hay un total de 847 vinos calificados como buenos y 697 vinos normales.

En el caso, definiremos como hipotesis nula que no varia cantidad de alcohol con respecto a la calidad del vino, y como hipotesis alternativa, que existan diferencias en la cantidad de alcohol con respecto a la calidad del vino:

- Hipotesis nula es H0:MeC=MeN
- Hipotesis alternativa es H1:MeC<>MeN

Tal y como hemos calculado anteriormente, ninguna de las variables sigue una distribucion normal. Usaremos por tanto un **test no parametrico como el de U de Mann-Whitney**.

```
#Hacemos manualmente la suma de rangos
muestraTotal <- c(good, normal)
rangosMuestra <- rank( muestraTotal)


sumaRangosGood<-sum(rangosMuestra[1:847])
sumaRangosNormal<-sum(rangosMuestra[846:1544])
n1<-length(good)
n2<-length(normal)

U1<- sumaRangosGood - n1 *(n1+1)/2
U2<- sumaRangosNormal - n2*(n2+1)/2

c(U1,U2)
```

```
## [1] 443082.5 149502.5
```

```
wilcox.test(good,normal,correct=FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  good and normal
## W = 443080, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

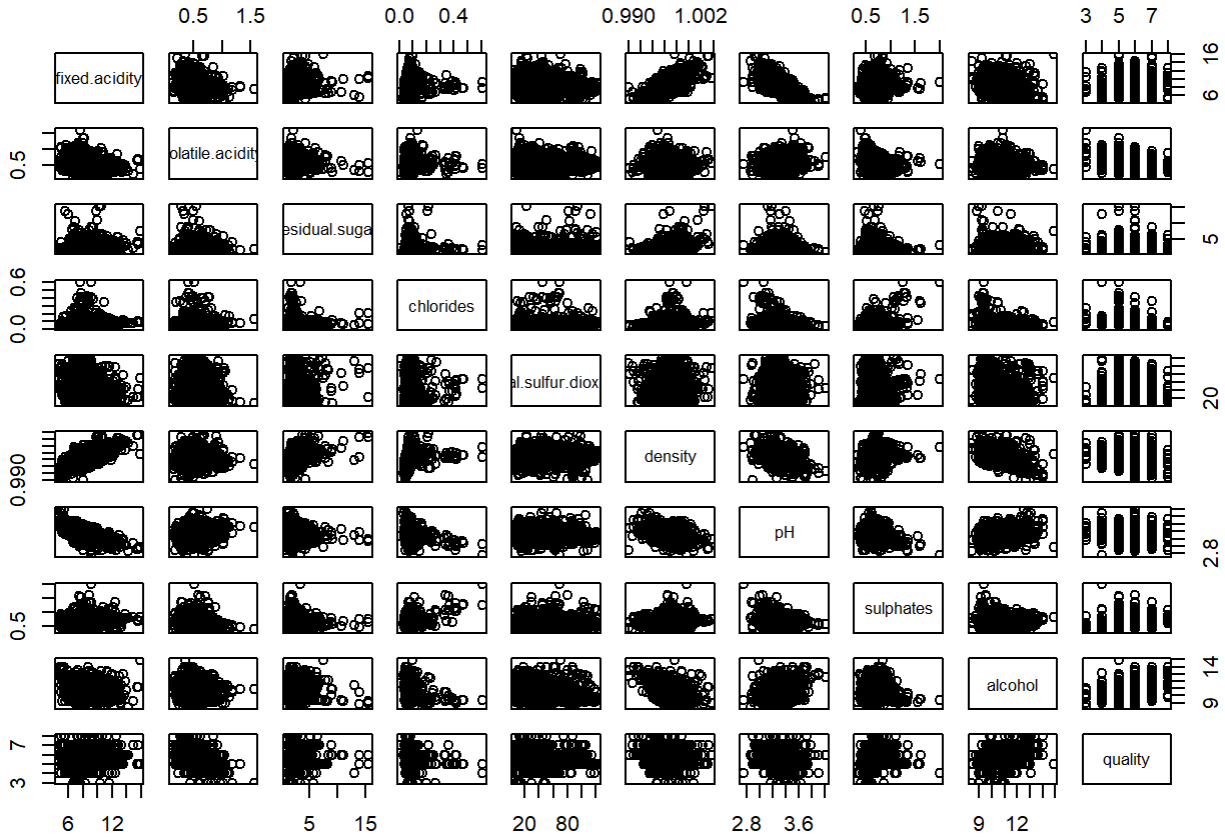
Como se puede observar, el p-valor es muy inferior al nivel de significancia (0.05) por lo tanto rechazamos la hipotesis nula. En conclusion, hay diferencias significativas en el cantidad de alcohol entre vinos buenos y vinos normales. Siendo mayor la cantidad de alcohol en los vinos buenos que en los normales (suma de rangos de los vinos buenos es mayor que la de los vinos normales).

4.3.3 Regresiones:

La **regresion lineal** es un modelo matematico que tiene como objetivo **aproximar la relacion de dependencia lineal entre una variable dependiente y una o una serie de variables independientes**.

En R, la regresion lineal se aplica mediante la funcion lm(). Esta puede ser simple o multiple en funcion de las variables independientes que se incluyan en la formula que se introduce como argumento.

```
plot(datos)
```



En el apartado “4.3.1.” hemos visto que la calidad esta relacionada con el nivel de alcohol (alto nivel de correlacion), asi como medianamente con el nivel de sulfatos y acido citrico.

Realizaremos por lo tanto una regresion de la calidad con la variable alcohol:

```
m1 = lm(quality~alcohol,data=datos)
summary(m1)
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8498 -0.3894 -0.1414  0.5194  2.5752
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9535     0.1795   10.88   <2e-16 ***
## alcohol       0.3542     0.0171   20.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7187 on 1542 degrees of freedom
```

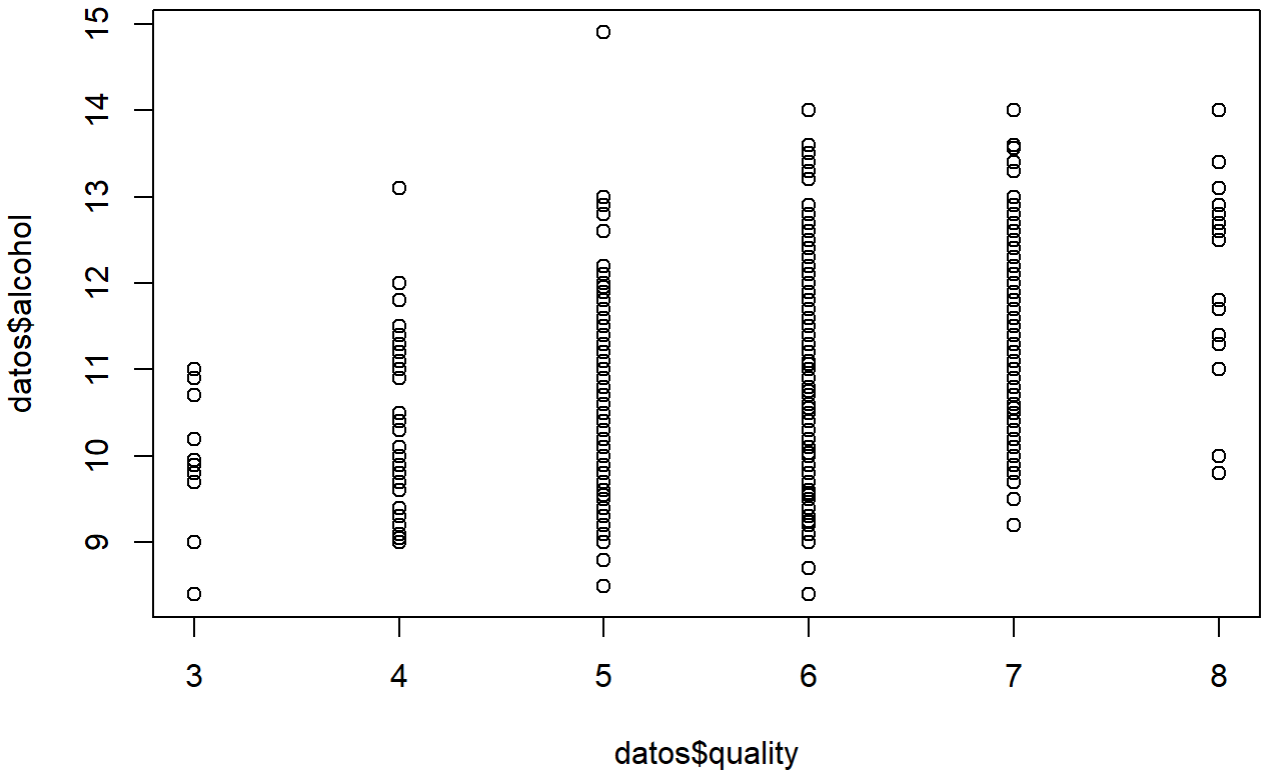
```
## Multiple R-squared:  0.2177, Adjusted R-squared:  0.2172
## F-statistic: 429.2 on 1 and 1542 DF,  p-value: < 2.2e-16
```

Siendo el coeficiente de determinacion (R-squared) una medida de calidad del modelo que toma valores entre -1 y 1, se comprueba como la calidad y el alcohol se correlacionan en parte, dando lugar a un R-squared de 0.2177.

Podemos tambien representar graficamente esta regresion: observamos que tiene una slope ligeramente positiva, lo que confirma la correlacion positiva de las dos variables:

```
reg<-lm(datos$quality~datos$alcohol , data = datos)
coeff=coefficients(reg)
```

```
plot(datos$quality, datos$alcohol)
abline(reg, col="blue")
```



Al introducir los sulfatos, este R-squared mejora hasta 0.2968 ya que tambien se correlaciona con el volumen de forma significativa, aunque en menor medida.

```
m2 = lm(quality~alcohol+sulphates,data=datos)
summary(m2)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates, data = datos)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89994 -0.38215 -0.09007  0.49339  2.36712
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.39229    0.18151   7.671 3.02e-14 ***
## alcohol      0.33341    0.01664  20.034  < 2e-16 ***
## sulphates    1.18680    0.11317  10.487  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6946 on 1541 degrees of freedom
## Multiple R-squared:  0.2698, Adjusted R-squared:  0.2689
## F-statistic: 284.8 on 2 and 1541 DF,  p-value: < 2.2e-16
```

Al introducir el nivel de acidez, baja ligeramente.

```
m3 = lm(quality~alcohol+sulphates+fixed.acidity,data=datos)
summary(m3)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + fixed.acidity, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79275 -0.36905 -0.07115  0.50556  2.16999
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.96411    0.19880   4.850 1.36e-06 ***
## alcohol        0.34058    0.01657  20.554  < 2e-16 ***
## sulphates      1.06732    0.11470   9.305  < 2e-16 ***
## fixed.acidity  0.05182    0.01020   5.082 4.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6891 on 1540 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2805
## F-statistic: 201.5 on 3 and 1540 DF,  p-value: < 2.2e-16
```

Si introducimos el nivel de azÃ Âºcar residual, vemos que el R-squared mejora algo de nuevo:

```
m4 = lm(quality~alcohol+sulphates+residual.sugar, data=datos)
summary(m4)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + residual.sugar,
##     data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89553 -0.38162 -0.08535  0.49509  2.36562
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.402461   0.183373   7.648 3.57e-14 ***
## alcohol        0.333658   0.016658  20.029  < 2e-16 ***
## sulphates      1.187147   0.113204  10.487  < 2e-16 ***
## residual.sugar -0.005179   0.013101  -0.395    0.693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6948 on 1540 degrees of freedom
## Multiple R-squared:  0.2699, Adjusted R-squared:  0.2685
## F-statistic: 189.8 on 3 and 1540 DF,  p-value: < 2.2e-16
```

La funcion lm() de Rtambien permite implementar modelos polinomicos mas complejos, como en el siguiente ejemplo:

```
m5 = lm(quality~alcohol+I(alcohol^2)+sulphates+residual.sugar, data=datos)
summary(m5)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + I(alcohol^2) + sulphates + residual.sugar,
##     data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.87755 -0.37942 -0.08079  0.49385  2.36303
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.91492    1.51132  -0.605  0.54502
## alcohol         0.76386    0.27899   2.738  0.00625 **
## I(alcohol^2)   -0.01976    0.01279  -1.545  0.12261
## sulphates       1.18580    0.11316  10.479  < 2e-16 ***
## residual.sugar -0.00404    0.01312  -0.308  0.75812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6945 on 1539 degrees of freedom
## Multiple R-squared:  0.2711, Adjusted R-squared:  0.2692
```

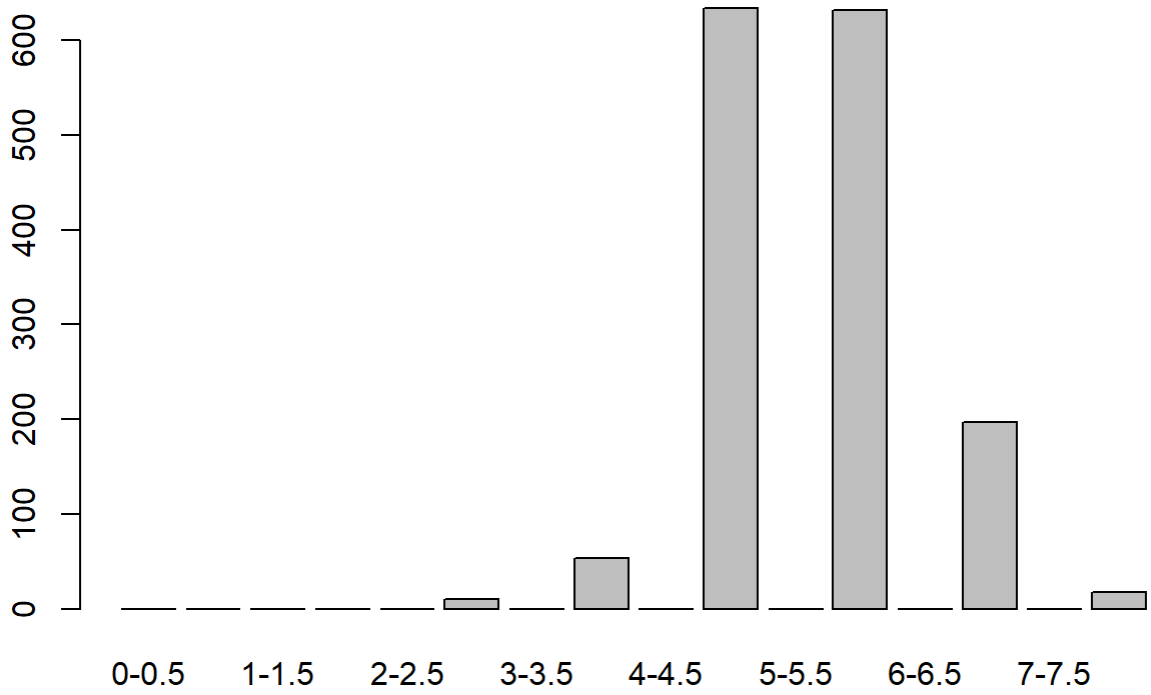
```
## F-statistic: 143.1 on 4 and 1539 DF,  p-value: < 2.2e-16
```

Se puede observar como el termino que relaciona el volumen con el alcohol de forma cuadratica resulta algo mas significativo, mejorando el R-squared. En ningun caso hemos obtenido sin embargo un R-squared muy elevado, aÃ Âºn tomando las variables mas correlacionadas.

4.3.4 Modelo de prediccion:

Existen muchos vinos con una calidad con valor de 4.5-6, muy por encima de otros niveles. Asi que en el conjunto de datos se encuentran muchos mas vinos “normales” que excelentes o muy malos.

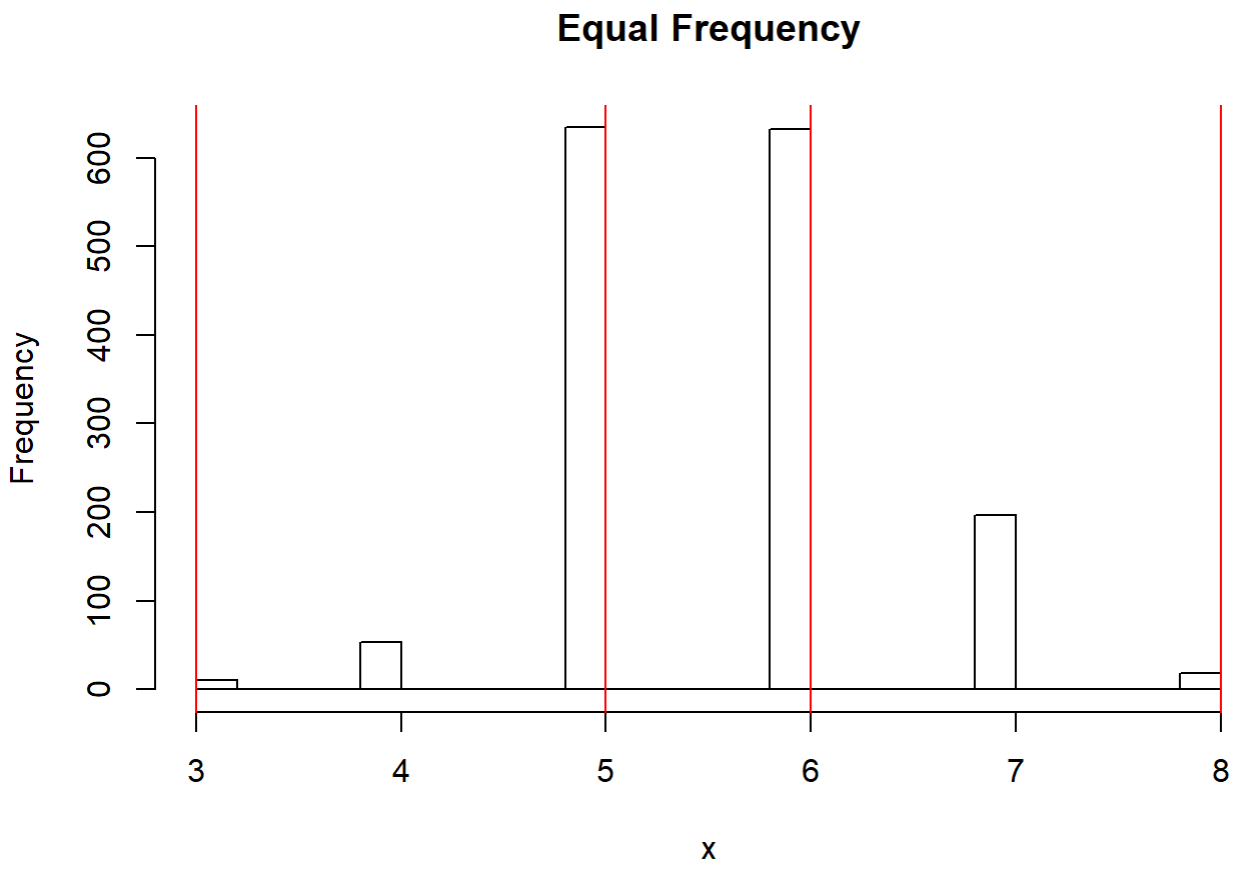
```
barplot(table(cut(datos$quality, breaks = seq(0, 8, by = 0.5),
      labels =paste(seq(0, 7.5, by =0.5), seq(0.5, 8, by = 0.5), sep="-"))))
```



Para facilitar un poco el modelo de prediccion, clasificaremos los niveles de vino en *malo*, *normal* y *bueno*.

Para saber que intervalos tomar para cada categoria utilizaremos. Para ello, utlizamos el paquete *arules* de R, que permite discretizar datos mediante la funcion discretize(). Utilizamos el modo **clusterig*, que nos devuelve tres clusters con un nÃ Âºmero de muestras igual en cada uno, automaticamente idenificando los puntos de corte para la variable calidad:

```
#equal frequency
x <- datos$quality
hist(x, breaks = 32, main = "Equal Frequency")
abline(v = discretize(x, breaks = 3, onlycuts = TRUE), col = "red")
```



Observamos que el primer intervalo (vinos malos) se extiende de 3 a 5 en nivel de calidad y el segundo de 5 a 6 (vinos normales). Como en el ejercicio 4.3.2. de contraste de hipotesis, consideramos buenos todos los vinos con un valor de 6 o mas:

```
datos$quality <- ifelse(datos$quality <5, 'bad', ifelse(datos$quality >= 6,'good','normal'))
datos$quality <- as.factor(datos$quality)
```

Esto nos **clasifica los vinos en malo, normal o bueno** dependiendo de su nivel de calidad.

Observamos la distribucion de valores por categoria, que contiene gracias a nuestra clasificacion por valores un numero mas o menos equilibrado de datos por categoria:

```
table(datos$quality)
```

```
##
##    bad   good normal
##     63    847    634
```

Antes de crear nuestro modelo, separaremos nuestro conjunto de datos en datos de test (20%) y de entrenamiento (80%), tomando la regla 80-20 para separar un conjunto de datos para test y entrenamiento:

```
set.seed(123)
samp <- sample(nrow(datos), 0.8 * nrow(datos))
train <- datos[samp, ]
test <- datos[-samp, ]
```

Creamos el modelo: Utilizaremos el modelo de mineria de datos Random Forest, y para ello necesitaremos la libreria **randomForest**:

```
library(randomForest)
```

```
model <- randomForest(quality ~ fixed.acidity + volatile.acidity + residual.sugar+chlor
ides+total.sulfur.dioxide+density+pH+sulphates+alcohol, data = train)
```

En el modelo, podemos decidir cuantos arboles (trees) queremos (por default = 500), y el nÃ Âºmero de predictores que aleatoriamente se usan en cada split to randomly sample at each

Vemos a continuacion la matriz de confusion:

```
model
```

```
##
## Call:
##  randomForest(formula = quality ~ fixed.acidity + volatile.acidity +      residual.s
ugar + chlorides + total.sulfur.dioxide + density +      pH + sulphates + alcohol, data
 = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 21.54%
## Confusion matrix:
##        bad good normal class.error
## bad      0   17     33   1.0000000
## good     1  562    104   0.1574213
## normal   1  110    407   0.2142857
```

Vemos que se han construido 500 arboles, y que el modelo aleatoriamente utiliza 3 predictores en cada split. Tambien podemos ver la matriz de confusion comparada con los datos reales, asi como el error de clasificacion de cada clase. A continucion haremos un test con los datos de test:

```
pred <- predict(model, newdata = test)
table(pred, test$quality)
```

```
##
## pred     bad good normal
##   bad      1    0      0
##   good     3  146     27
##   normal   9   34     89
```

Podemos calcular la precision del modelo asi:

```
((1+146+89)/(1+3+146+27+9+34+89))
```

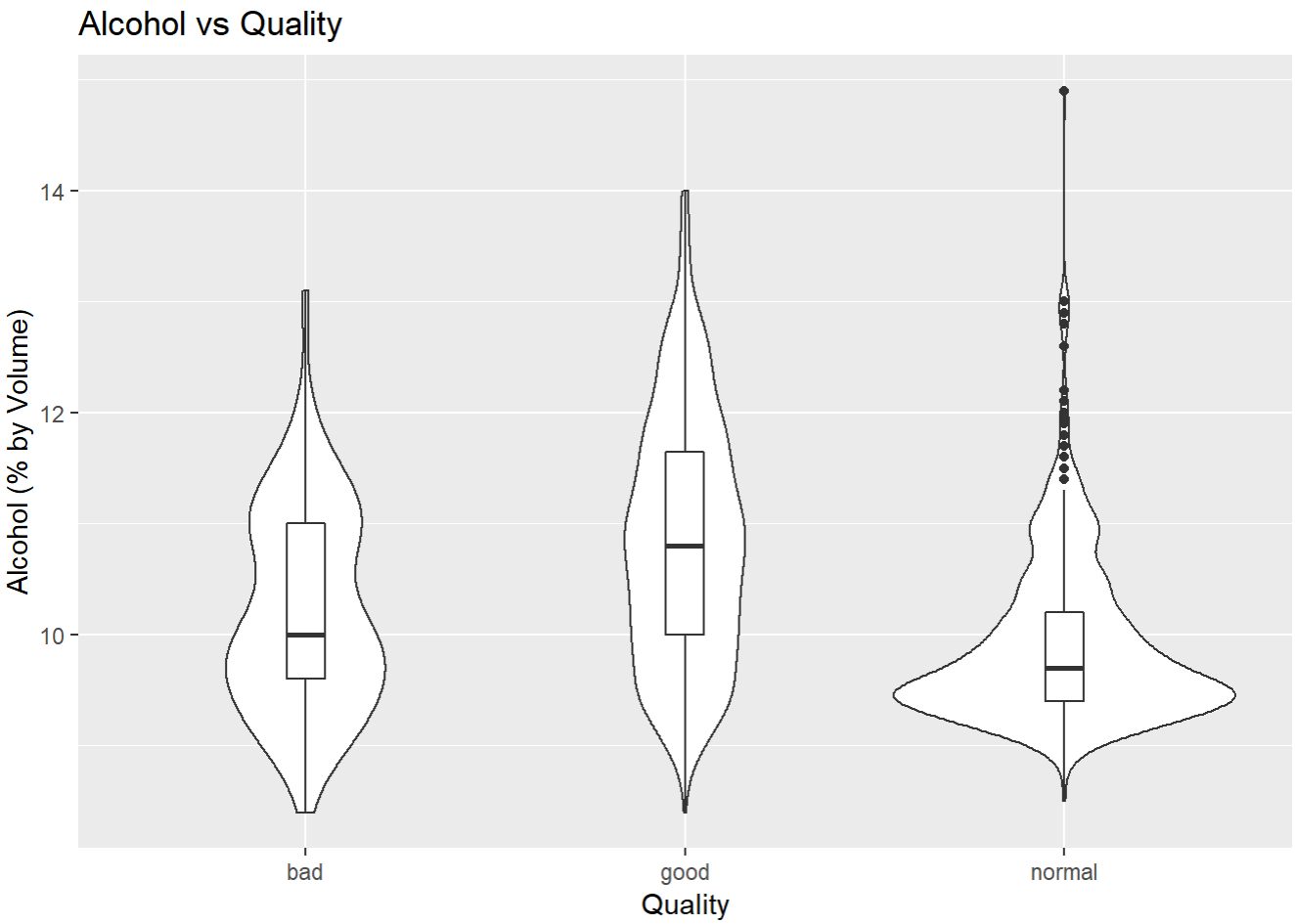
```
## [1] 0.763754
```

La precision con este modelo es bastante alta, de mas de un 70%. Podriamos mejorarla con una seleccion mas exhaustiva de features o atributos (este paso se llama *feature selection* en ingles) o usando diferente nÃ Âºmero de atributos a usar.

5 Representacion de los resultados a partir de tablas y graficas.

A continuacion se presentan una serie de graficos que representan graficamente las conclusiones extraidas durante la ejecucion del analisis.

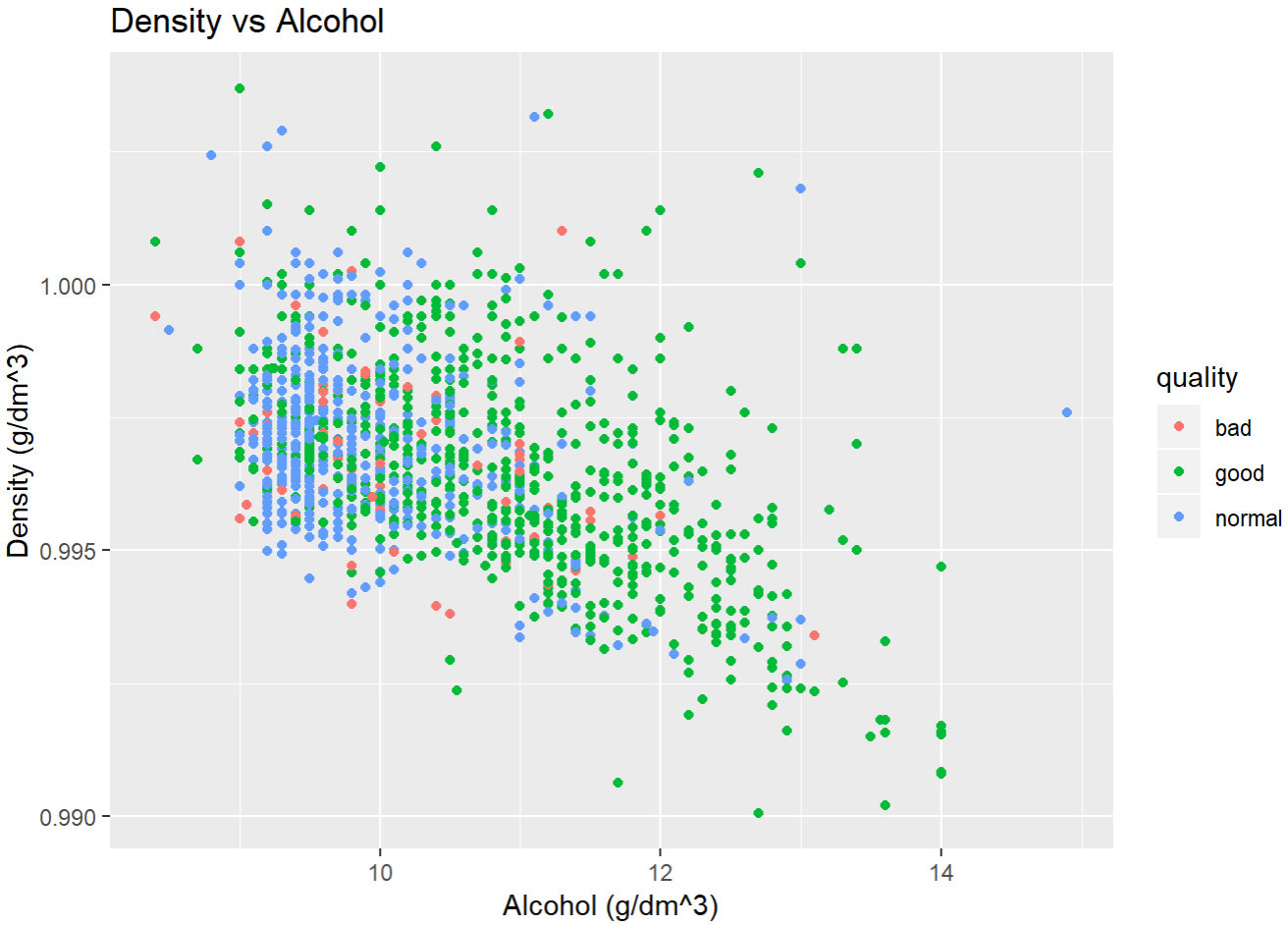
GRAFICO I



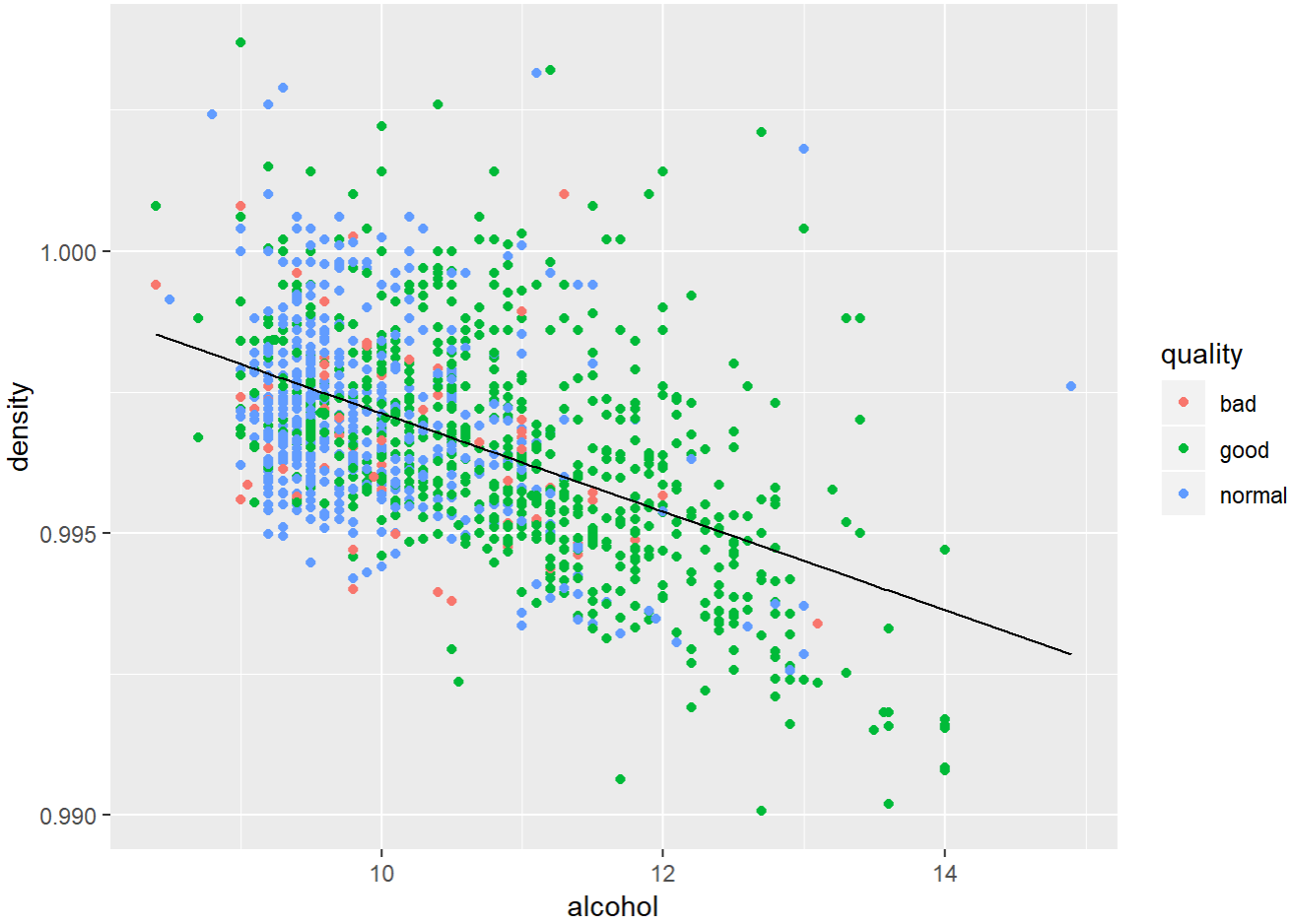
Interpretacion: El grafico de violin muestra la relacion que existe en el modelo entre la calidad y una de las variables independeinte que mas fuertemente determinan la calidad del vino, como hemos visto durante toda la practica: el alcohol. Como se aprecia, los vinos que hemos clasificado como buenos (quality> 6), tienen un mayor porcentaje de alcohol.

GRAFICO II

```
ggplot(datos, aes(x = alcohol, y = density)) +
geom_point(aes(color = quality)) +
labs(title="Density vs Alcohol") +
xlab("Alcohol (g/dm^3)") +
ylab("Density (g/dm^3)")
```



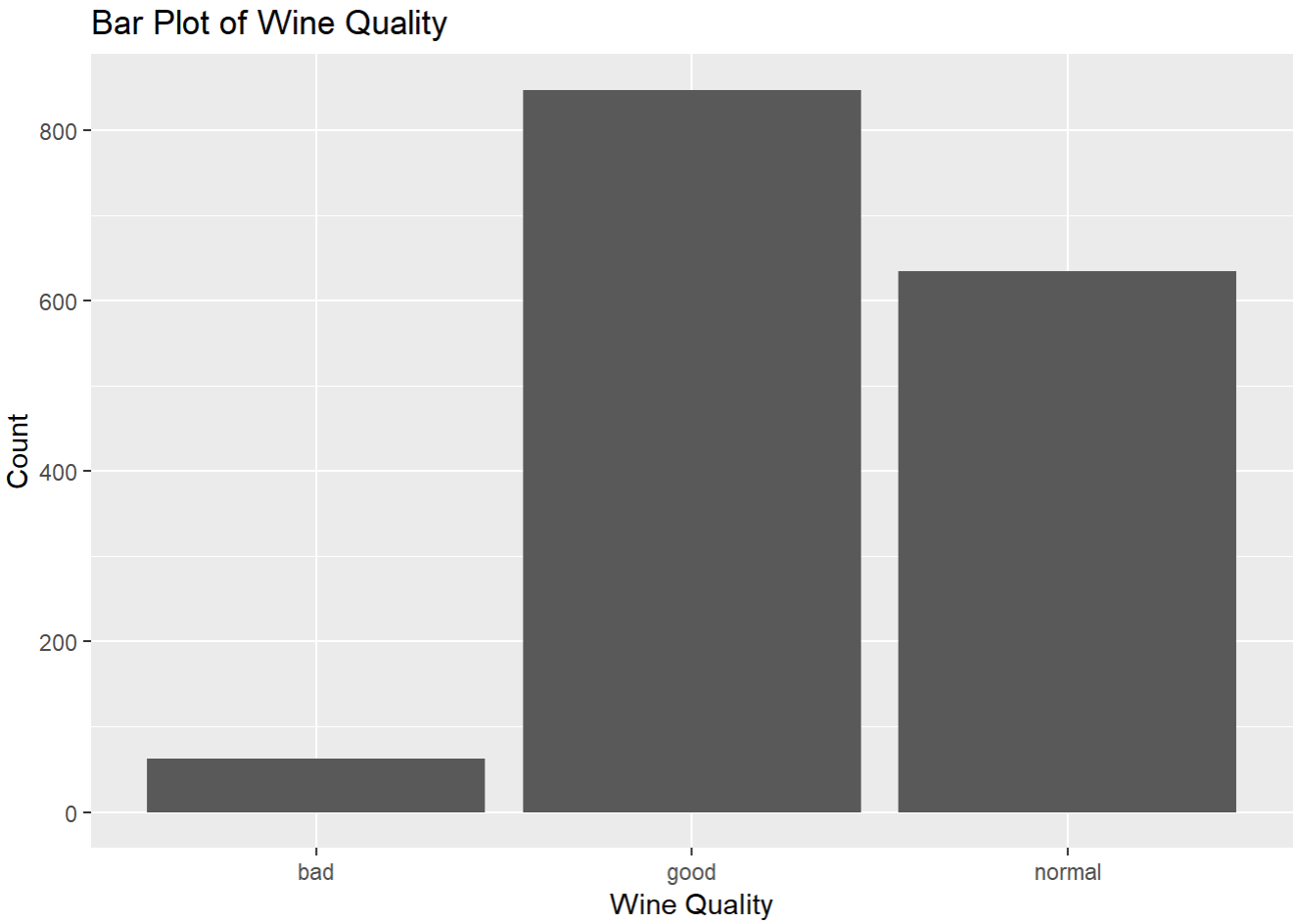

```
datos$pred.qual <- predict(lm(density ~ alcohol, data = datos))

p1 <- ggplot(datos, aes(x = alcohol, y = density))

p1 + geom_point(aes(color = quality)) +
  geom_line(aes(y = pred.qual))
```



Interpretacion: Este diagrama de dispersion muestra que la densidad y el volumen de alcohol estan correlacionadas negativamente (comprobado con un modelo de regresion). Esto indica que cuanto mayor es el porcentaje de alcohol, menor es la densidad (menor cantidad de azucares). El grafico muestra, como los vinos de mejor calidad se situan en la parte inferior derecha del grafico. En esta region se situan los vinos con mayor volumen de alchohol y menor densidad.

GRAFICO III

```
ggplot(aes(x = as.factor(quality)), data = datos)+
       geom_bar()+
       xlab("Wine Quality") + ylab("Count") +
       ggtitle("Bar Plot of Wine Quality")
```



Interpretacion: Como se explico anteriormente, existe cierta asimetria en la distribucion de la calidad. El numero de vinos calificados como malos, es bastante reducido, en relacion a los vinos calificados como normales o buenos.

---

6 Resolucion del problema. A partir de los resultados obtenidos, ¿cuales son las conclusiones? ¿los resultados permiten responder al problema?

El problema que se planteaba al iniciar la practica es que **partiamos de un dataset de vinos del que no teniamos ningun conocimiento** (correlaciones entre variables, interdependencias, etc).
Lo primero que hemos hecho es un **analisis del conjunto de datos** (volumen de datos, q, variables, tipo de variables, etc), para posterioremente **investigar si hay relaciones entre ellos**, **contrastar hipotesis**, para poder, como personas sin conocimientos especificos de enologia, **clasificar los vinos en categorias** y crear un **modelo que nos clasifique nuevos vinos** que nos encontremos en funcion a su calidad.

Para llegar a estas conclusiones, hemos procedido de la siguiente forma:
- hemos hecho un analisis de la estrucutra de los datos
- hemos hecho un analisis visual, mediante la representacion grafica
- hemos investigado las correlaciones entre los datos
- hemos desechado variables fuertemente correlacionados para evitar problemas de colinealidad
- hemos comprobado la veracidad de nuestras primeras intuiciones (alcohol y calidad van del a mano).
- Para ello hemos usado: contraste de hipotesis y tests de comprobacon de varianzas, asi como crear un modelo de regresio capaz de clasificar vinos del conjunto de test por niveles de calidad. Los resultados nos han proporcionado mas informacion sobre el dataset y nos ha permitido rechazar o confirmar hipotesis.

Finalmente, salvamos el fichero de datos, con los datos limpios y transformados.

```
write.csv2(datos, "winequality-red_out.csv", row.names = FALSE)
```

Contribuciones	Firma
Investigación previa	Aïda Centelles/Gonzalo Canales
Redacción de las respuestas	Aïda Centelles/Gonzalo Canales
Desarrollo código	Aïda Centelles/Gonzalo Canales