Práctica 2: Limpieza y validación de los datos

Hèctor Gómez Meneses

null

Índice

1. Descripción del dataset.

En esta práctica contamos con un dataset con diferentes variables referentes a vinos tintos, una de las cuales es la calidad . A partir de este conjunto de datos se propone determinar qué variables influyen más sobre la evaluación de estos vinos y se crearán modelos de regresión que permitiran predecir la calidad del vino en función de ciertos atributos.

Las diferentes variables de nuestro dataset son:

fixed acidity: variable numérica contínua con la acidez fija del vino.

volatile acidity: variable numérica contínua con la acidez volátil del vino.

citric acid: variable numérica contínua con el ácido cítrico del vino.

residual sugar: variable numérica contínua con el azúcar residual del vino.

chlorides: variable numérica contínua con los cloruros del vino.

free sulfur dioxide: variable numérica contínua con el dioxido de azufre libre del vino.

total sulfur dioxide: variable numérica contínua con el total de dioxido de azufre del vino.

density: variable numérica contínua con la densidad del vino.

pH: variable numérica contínua con el pH del vino.

sulphates: variable numérica contínua con los sulfatos del vino.

alcohol: variable numérica contínua con el alcohol del vino.

quality: variable numérica contínua con la evaluación del vino, variable objetivo.

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

Leemos los datos y mostramos un resumen del dataset que tenemos:

```
wine <- read.csv('winequality-red.csv',sep=';')
summary(wine)
```

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

3. Limpieza de los datos

3.1¿Los datos contienen ceros o elementos vacíos?

A continuación comprobamos los valores que contengan ceros y los valores nulos

```
colSums(wine==0.00)
```

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

```
colSums(is.na(wine))
```

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

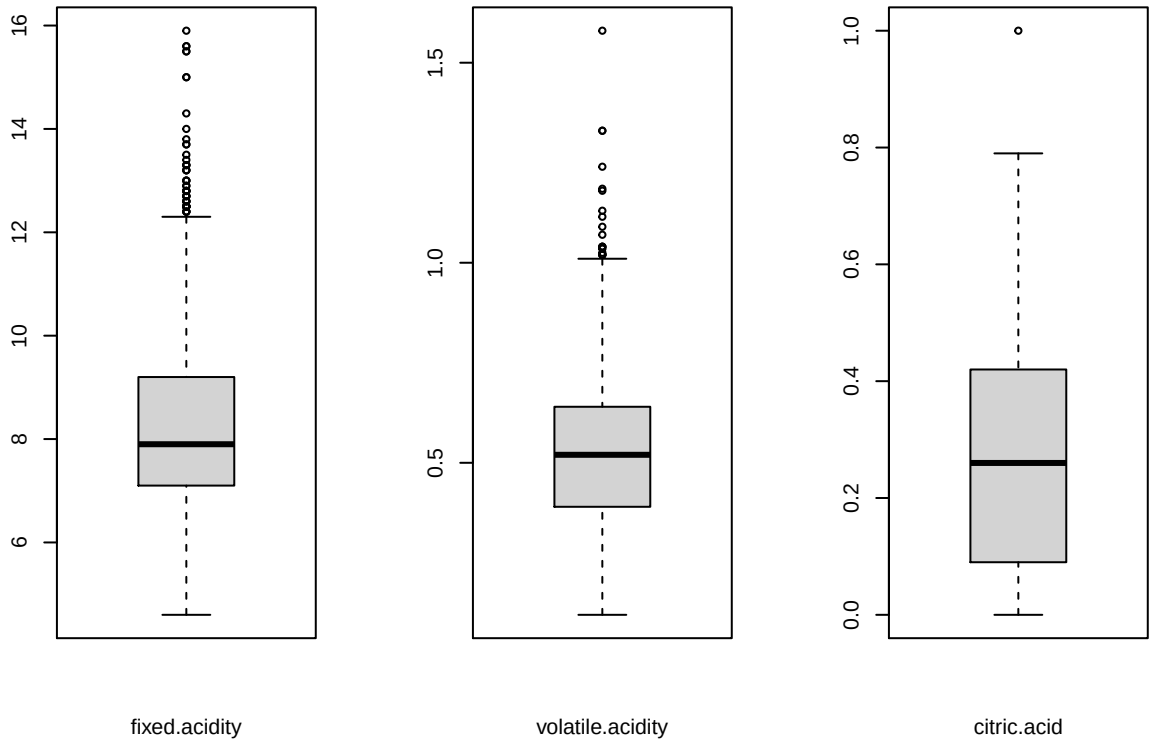
```
##                   0                   0                   0
```

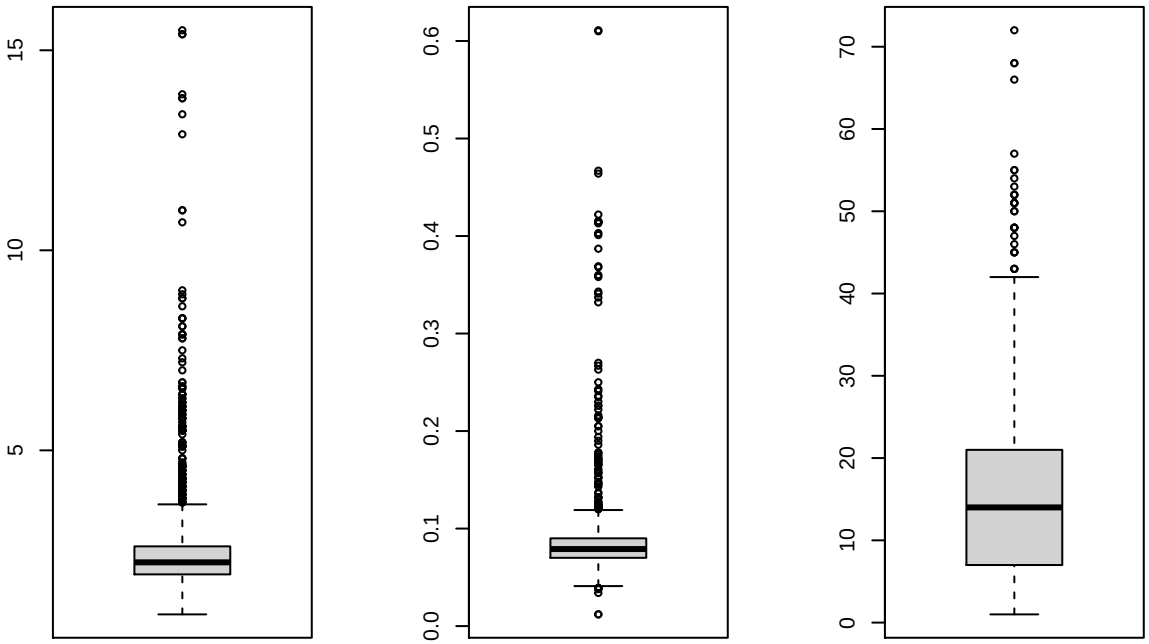
Comprobamos que no tenemos valores nulos y que los valores que contienen ceros pertenecen al acido citrico, y son valores que no son erróneos por ser 0, puede ser 0 el valor del ácido cítrico del vino, por lo que los dejaremos como estan.

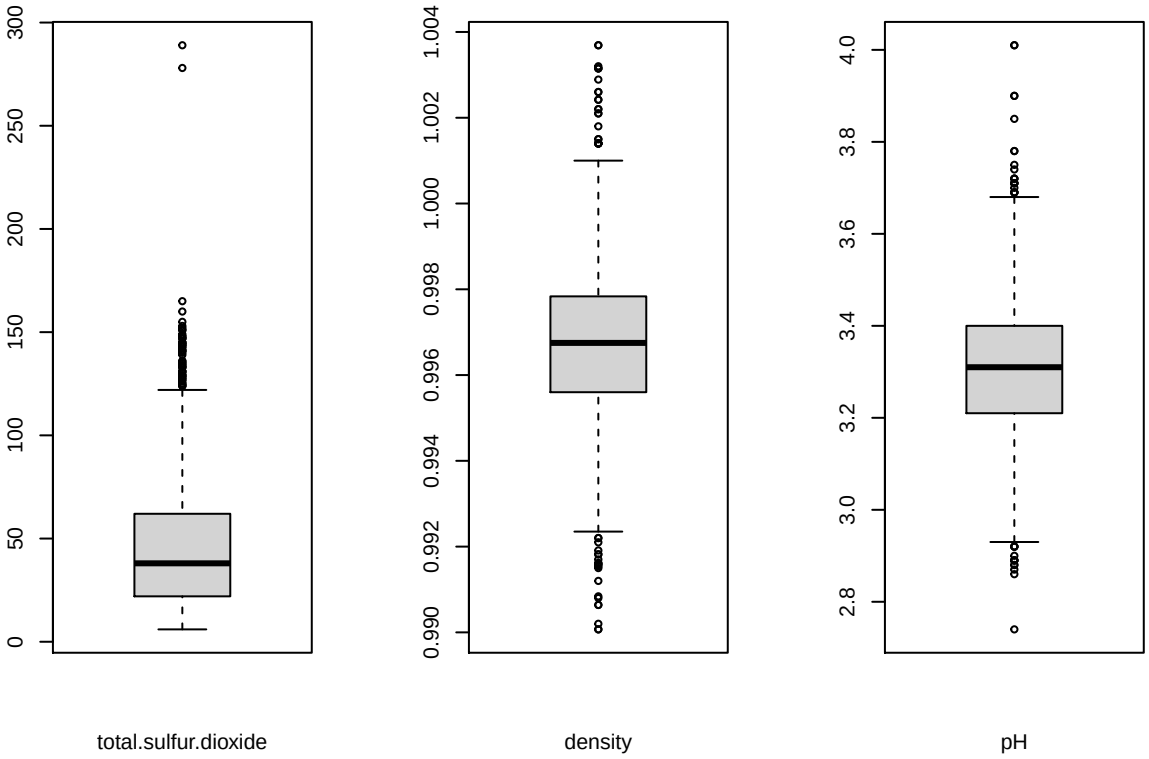
3.2 Identificación y tratamiento de valores extremos

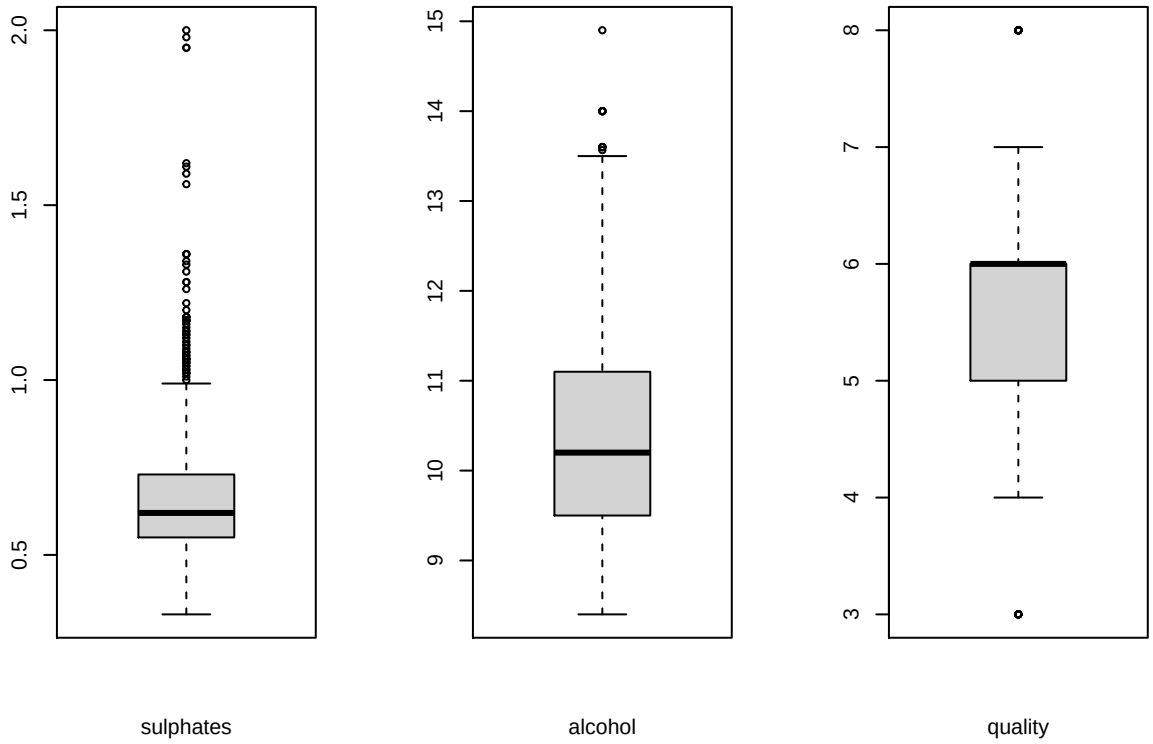
Vamos ahora con los valores extremos, para ello mostramos el boxplot de cada variable

```
i=0
par(mfrow=c(1,3))
for(x in colnames(wine)) {
  if(i==3){
    par(mfrow=c(1,3))
    i=0}
 boxplot(wine[x], xlab = x)
 i=i+1
}
```




15
10
5
residual.sugar
0.6
0.5
0.4
0.3
0.2
0.1
0.0
chlorides
70
60
50
40
30
20
10
0
free.sulfur.dioxide


0
50
100
150
200
250
300
total.sulfur.dioxide
0.990
0.992
0.994
0.996
0.998
1.000
1.002
1.004
density
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH



Observamos que, aun que tenemos valores extremos, en la mayoría de las variables, todos entran en un límite razonable de valores que podrían considerarse correctos. El único caso es en la variable total.sulfur.dioxide, donde tenemos valores cercanos a 300, y únicamete tenemos dos, el siguiente más cercano apenas supera los 150, es por eso que prescindiremos de estos dos valores, para ellos filtramos los valores mayores a 200 en esta variable y exportaremos los datos.

```
wine <- wine[wine$total.sulfur.dioxide < 200, ]
write.csv(wine, "Wine_clean.csv")
```

4. Análisis de los datos

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

En este caso, la única separación que me parece interesante hacer es entre los vinos con una puntuación mayor o igual que 5 y los vinos con una puntuación menor que 5, de forma que podamos ver las características de los vinos por separado.

```
wine$quality_factor <- wine$quality
wine <- wine %>% mutate(quality_factor = ifelse(quality_factor < 5,'Bad','Good'))
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Vamos primero a comprobar que nuestras variables sigan una distribución normal, para ello utilizamos el test AndersonDarling.

```
normal <- c()
not_normal <- c()
for(x in colnames(wine[,1:12])) {
  test <- ad.test(wine[[x]])
  if(test$p.value>0.05){
```

```
    normal<-append(normal,x)
  }
  else{
    not_normal<-append(not_normal,x)
  }
}
print(normal)
```

```
## NULL
```

```
print(not_normal)
```

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

Vemos que ninguna de nuestras variables sigue una distribución normal, vamos ahora a comprobar la homogeneidad de la varianza, para ello utilizamos el test Fligner-Killeen ya que no disponemos de datos que sigan una distribución normal. Evaluaremos los grupos formados por los que tienen una nota superior al 5 y los que la tienen inferior.

```
fligner.test(wine$alcohol~wine$quality_factor)
```

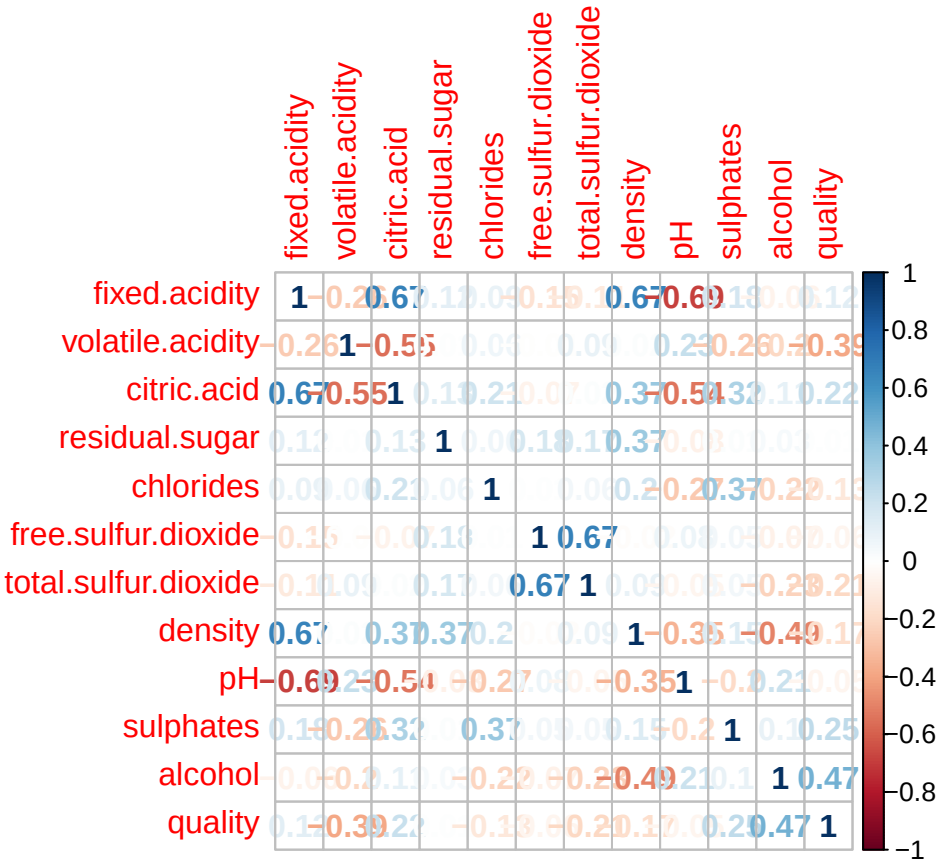
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  wine$alcohol by wine$quality_factor
## Fligner-Killeen:med chi-squared = 0.90062, df = 1, p-value = 0.3426
```

Obtenemos un valor mayor que 0.05 de p.value por lo que podemos aceptar la hipótesis de que ambas varianzas son homogéneas.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.Pruebas de contraste de hipótesis,correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes

El primer método utilizaremos será la matriz de correlaciones, lo que haremos será ver que variables tienen una realación más fuerte con la variable a predecir, la más importante, la calidad del vino.

```
wine_cor <- cor(wine[,1:12])
library(corrplot)
corrplot(wine_cor, method = "number")
```



Vemos que las variables que más correlación tienen más correlación son:

- Alcohol
- Sulphates
- Density
- Total.sulfur.dioxide
- Citric.acid
- Volatile.acidity

El segundo método de análisis será un contraste de hipótesis sobre las muestras en las que antes hemos comprobado la homogeneidad de las varianzas de forma que podamos determinar si la calidad del vino es superior dependiendo de la cantidad de alcohol que este contiene. Aplicamos la siguiente hipótesis:

H0: m1 - m2 = 0

H1: m1 - m2 > 0

Donde m1 es la media de alcohol de los vinos buenos y m2 de los malos.

```
wine_good <- wine[wine$quality_factor == 'Good',]$alcohol
wine_bad <- wine[wine$quality_factor == 'Bad',]$alcohol
t.test(wine_good, wine_bad, alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  wine_good and wine_bad
```

```
## t = 1.7935, df = 69.09, p-value = 0.03864
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0150087       Inf
## sample estimates:
## mean of x mean of y
##  10.42904  10.21587
```

Vemos que nos da un valor de p-value superior a nuestro valor de significación, por lo que rechazamos la hipótesis nula y aceptamos la hipótesis alternativa, concluyendo que los vinos catalogados como buenos tienen un porcentaje mayor de alcohol que los catalogados como malos.

El tercer y último método de análisis consiste en crear un modelo de regresión lineal múltiple, de forma que se pueda predecir la calidad del vino para nuevos juegos de datos en que contemos con los datos pero no con la calificación, crearemos un único modelo con las variables que hemos obtenido como más correlacionadas en el primer método de análisis.

```
multiple_regresion <- lm(quality~alcohol+sulphates+density+total.sulfur.dioxide+citric.acid+volatile.ac
summary(multiple_regresion)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + sulphates + density + total.sulfur.dioxide +
##     citric.acid + volatile.acidity, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73044 -0.38424 -0.05765  0.44257  2.14900
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -8.7995291 11.9700255  -0.735    0.462
## alcohol               0.3012656  0.0195856  15.382  < 2e-16 ***
## sulphates             0.7302069  0.1036419   7.045 2.74e-12 ***
## density              11.6581740 11.9294887   0.977    0.329
## total.sulfur.dioxide -0.0025671  0.0005345  -4.803 1.71e-06 ***
## citric.acid          -0.1164304  0.1202390  -0.968    0.333
## volatile.acidity     -1.2431931  0.1161833 -10.700  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6541 on 1590 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3424
## F-statistic: 139.5 on 6 and 1590 DF,  p-value: < 2.2e-16
```

Vemos que las variables density y citric.acid tienen un p-value mayor a 0.05 por lo que podríamos prescindir de ellas. Vamos a probar a predecir nuestro dataset original a ver que tal funciona nuestro modelo de regresión.

```
wine_for_prediction<-wine[c("alcohol",'sulphates','density','total.sulfur.dioxide','citric.acid','volati
wine$prediction <- predict(multiple_regresion, wine_for_prediction)

wine$error <- abs(wine$quality - wine$prediction)
mean(wine$error)
```

```
## [1] 0.506538
```

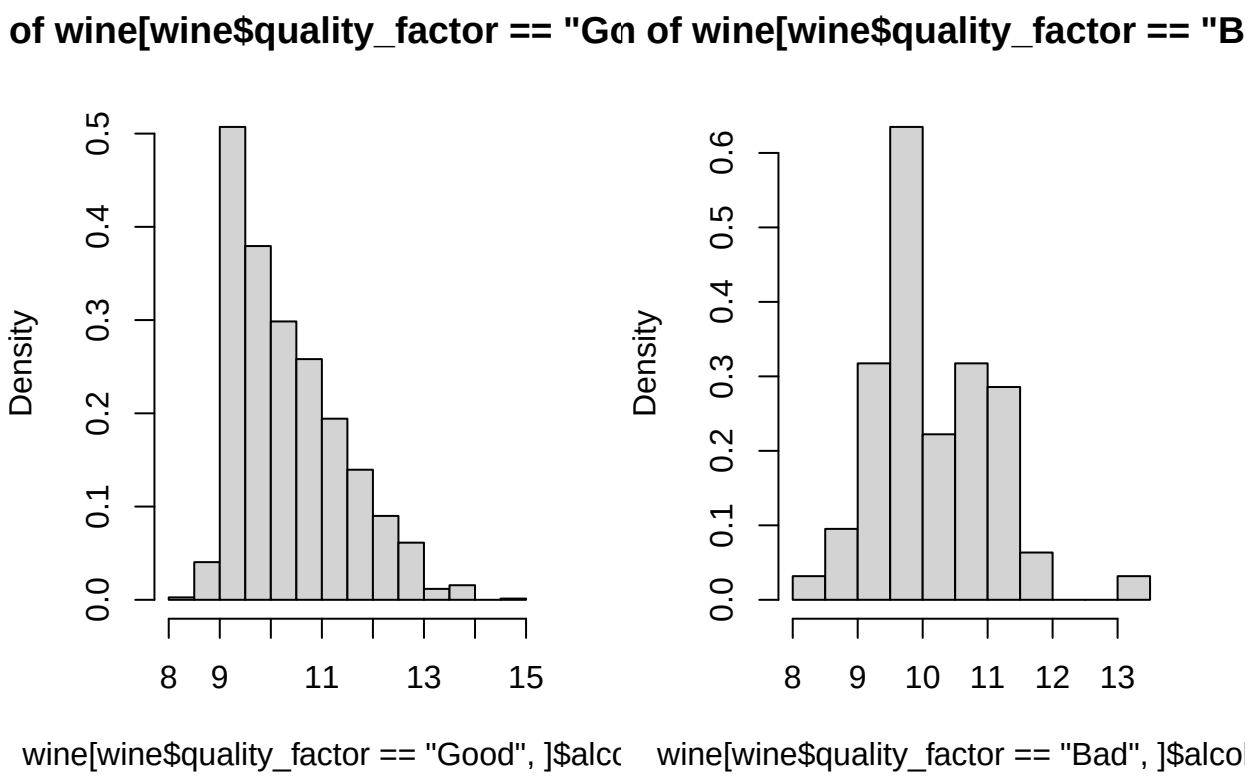
Vemos que el error medio en las predicciones de todos los datos que teníamos es de 0.507, nuestro modelo es

mejorable habiendo utilizado los mismos datos para entrenar y para testear, cosa que generalmente no se utiliza. Aún así, el modelo no es del todo malo, pero si mejorable.

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

Vamos a comprobar las distribuciones del alcohol de los vinos en función de si son catalogados como buenos o malos:

```
par(mfrow=c(1,2))
hist(wine[wine$quality_factor =='Good',]$alcohol, freq=FALSE)
hist(wine[wine$quality_factor =='Bad',]$alcohol, freq=FALSE)
```



Comprobamos efectivamente que los Good tienen más densidad alrededor del 9.5 que los Bad.

Visualizamos la matriz de correlaciones antes visualizada.

```
wine_cor
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.256785091  0.67422712    0.117248172
## volatile.acidity       -0.25678509      1.000000000 -0.55123061    0.008531062
## citric.acid             0.67422712     -0.551230612  1.00000000    0.134609755
## residual.sugar          0.11724817      0.008531062  0.13460975    1.000000000
## chlorides               0.09350529      0.060113411  0.20657072    0.060344077
## free.sulfur.dioxide    -0.15358722     -0.007234366 -0.06678113    0.178818077
## total.sulfur.dioxide   -0.11480905      0.091061805  0.01718835    0.173644280
## density                 0.66901320      0.019058755  0.37181600    0.369731554
## pH                     -0.68522731      0.232618233 -0.53954875   -0.076652612
## sulphates               0.18283587     -0.262772500  0.31609500    0.010113585
```

```
## alcohol                -0.06125771     -0.200071683  0.10576656    0.033472904
## quality                 0.12478955     -0.388954729  0.22294288    0.005146394
##                         chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity          0.093505291        -0.153587216          -0.11480905
## volatile.acidity       0.060113411        -0.007234366           0.09106181
## citric.acid            0.206570720        -0.066781129           0.01718835
## residual.sugar         0.060344077         0.178818077           0.17364428
## chlorides              1.000000000         0.007648019           0.05647956
## free.sulfur.dioxide    0.007648019         1.000000000           0.67301899
## total.sulfur.dioxide   0.056479559         0.673018994           1.00000000
## density                0.199266945        -0.017107013           0.09166396
## pH                    -0.267716681         0.075814084          -0.05067752
## sulphates              0.370713239         0.054092715           0.05260413
## alcohol               -0.219898510        -0.074314971          -0.22958904
## quality               -0.127500330        -0.055278579          -0.20758070
##                            density          pH   sulphates     alcohol
## fixed.acidity          0.66901320 -0.68522731  0.18283587 -0.06125771
## volatile.acidity       0.01905875  0.23261823 -0.26277250 -0.20007168
## citric.acid            0.37181600 -0.53954875  0.31609500  0.10576656
## residual.sugar         0.36973155 -0.07665261  0.01011358  0.03347290
## chlorides              0.19926694 -0.26771668  0.37071324 -0.21989851
## free.sulfur.dioxide   -0.01710701  0.07581408  0.05409271 -0.07431497
## total.sulfur.dioxide   0.09166396 -0.05067752  0.05260413 -0.22958904
## density                1.00000000 -0.34796081  0.14682779 -0.49406360
## pH                    -0.34796081  1.00000000 -0.19935467  0.21084985
## sulphates              0.14682779 -0.19935467  1.00000000  0.09575591
## alcohol               -0.49406360  0.21084985  0.09575591  1.00000000
## quality               -0.17159200 -0.05382786  0.25382230  0.47420776
##                            quality
## fixed.acidity          0.124789555
## volatile.acidity      -0.388954729
## citric.acid            0.222942880
## residual.sugar         0.005146394
## chlorides             -0.127500330
## free.sulfur.dioxide   -0.055278579
## total.sulfur.dioxide  -0.207580703
## density               -0.171591997
## pH                    -0.053827862
## sulphates              0.253822305
## alcohol                0.474207756
## quality                1.000000000
```

Mostramos, por último, la matriz de confusión de nuestro modelo de regresión, pasando a factor nuestra predicción:

```
wine$quality_factor_predicted <- wine$quality_factor
wine <- wine %>% mutate(quality_factor_predicted = ifelse(prediction < 5,'Bad','Good'))
table(wine$quality_factor,wine$quality_factor_predicted)
```

```
##
##         Bad Good
##   Bad    12   51
##   Good   96 1438
```

De donde volvemos a ver que tenemos un modelo que no es malo, pero tiene potencial de mejora.

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

El preprocesado de datos es igual de importante que el procesado en sí, es decir, la gestión de valores nulos, outliers, valores erróneos, etc etc

En cuanto al análisis, los resultados obtenidos nos han permitido ver cuales eran las variables que afectan más en lo que a la calificación del vino se refiere. Concretamente los resultados de los dos primeros análisis nos han permitido determinar cuales eran las que tenían más peso en la calificación, y en el tercer análisis hemos creado un modelo de regresión que nos permite predecir nuevas calificaciones para nuevos datasets. Este último modelo puede ser mejorado, aún así, como hemos visto, ha predecido razonablemente bien.