Práctica 2 - Limpieza y validación de los datos

Juan Luis Arróniz Cruz

1 de junio de 2018

Table of Contents

Descripción del dataset

El conjunto de datos recoge información sobre el vino tinto portugues “Vinho Verde”. Sólo se dispone de valores fisicoquimicos(entradas) y sensoriales (salida) de las variables disponibles, no hay datos acerca de tipos de uva, marca del vino, precio de venta, etc… La importancia de estos datos es que podemos relacionar la calidad del producto final con las distintas variables de las que disponemos. Este dataset contiene 1599 registros y 12 variables. Las variables son fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality.

Integración y selección de los datos de interés a analizar

En un primer momento, no podemos descartar ninguna variable del conjunto de datos ya que a priori no hay ninguna descartable sin un estudio previo.

Limpieza de los datos

Asignar a cada variable el tipo de dato adecuado

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

variables	class
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

No haría falta la conversión de ninguna variable

Asignación de nuevos nombres a las columnas

```
columnas <- names(datos)
columnas
```

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

```
names(datos)[1] = "Acidez_Fija"
names(datos)[2] = "Acidez_Volatil"
names(datos)[3] = "Acido_Citrico"
```

```
names(datos)[4] = "Azucar_Residual"
names(datos)[5] = "Cloruros"
names(datos)[6] = "Dioxido_de_Azufre_Libre"
names(datos)[7] = "Dioxido_de_Azufre_Total"
names(datos)[8] = "Densidad"
names(datos)[10] = "Sulfatos"
names(datos)[11] = "Alcohol"
names(datos)[12] = "Calidad"

columnas <- names(datos)
columnas
```

```
##  [1] "Acidez_Fija"             "Acidez_Volatil"
##  [3] "Acido_Citrico"           "Azucar_Residual"
##  [5] "Cloruros"                "Dioxido_de_Azufre_Libre"
##  [7] "Dioxido_de_Azufre_Total" "Densidad"
##  [9] "pH"                      "Sulfatos"
## [11] "Alcohol"                 "Calidad"
```

Ceros y elementos vacíos

```
sapply(datos, function(x) sum(is.na(x)))
```

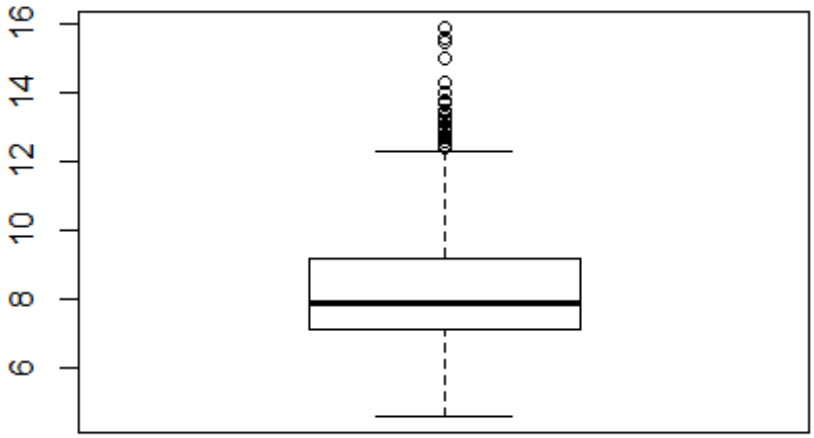
```
##             Acidez_Fija          Acidez_Volatil
Acido_Citrico
##                       0                       0
0
##         Azucar_Residual                Cloruros
Dioxido_de_Azufre_Libre
##                       0                       0
0
## Dioxido_de_Azufre_Total                Densidad
pH
##                       0                       0
0
##                Sulfatos                 Alcohol
Calidad
##                       0                       0
0
```

Observamos que no hay valores vacíos o ceros

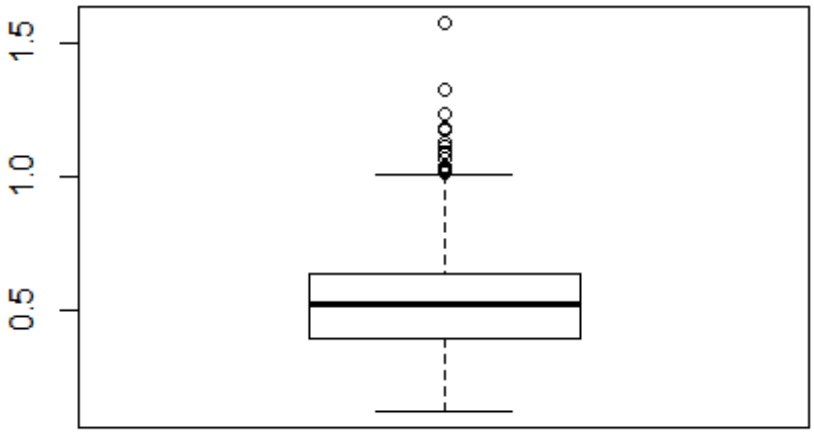
Valores extremos

Visualizamos los datos por medio de un diagrama de caja

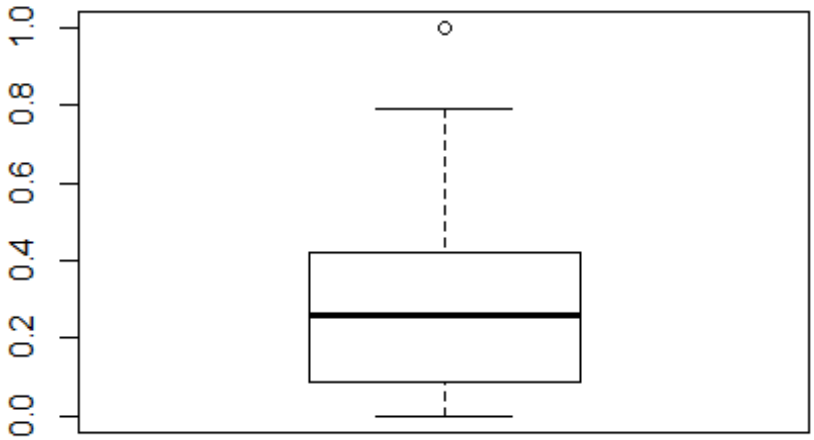
```
boxplot(datos$Acidez_Fija)
```



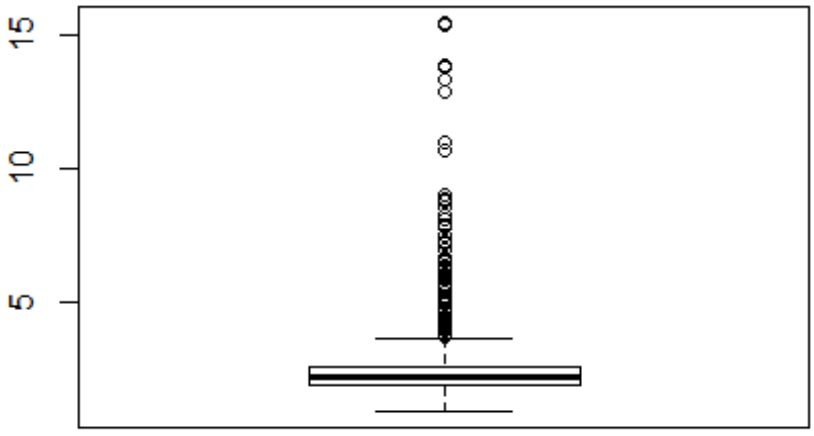

```r
boxplot(datos$Acidez_Volatil)
```



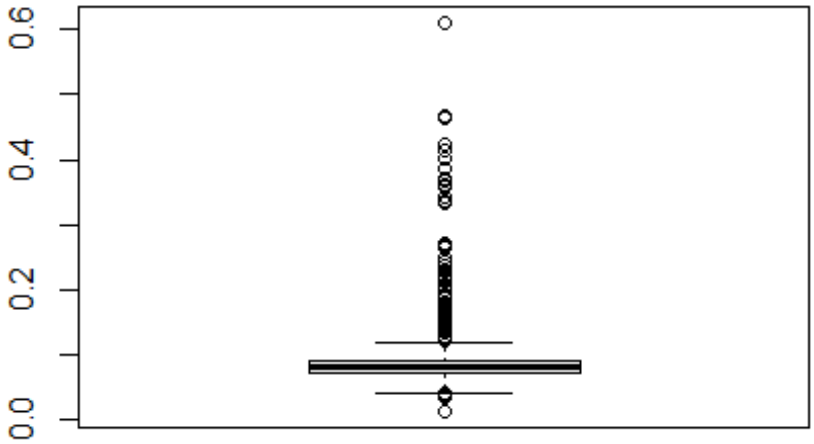

```r
boxplot(datos$Acido_Citrico)
```



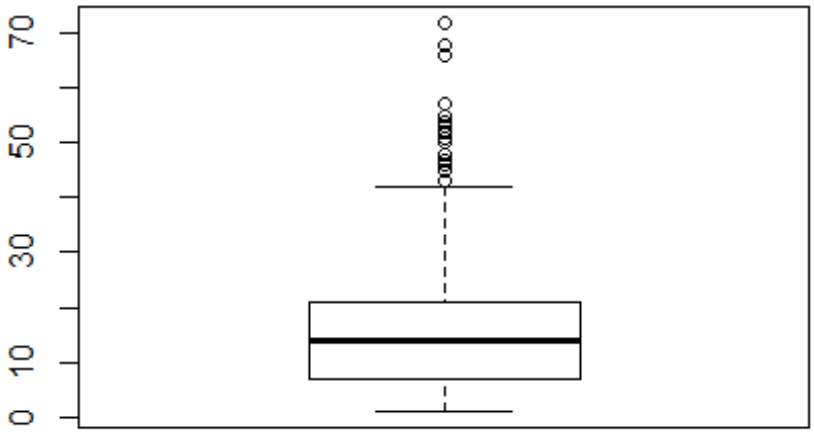

```
boxplot(datos$Azucar_Residual)
```



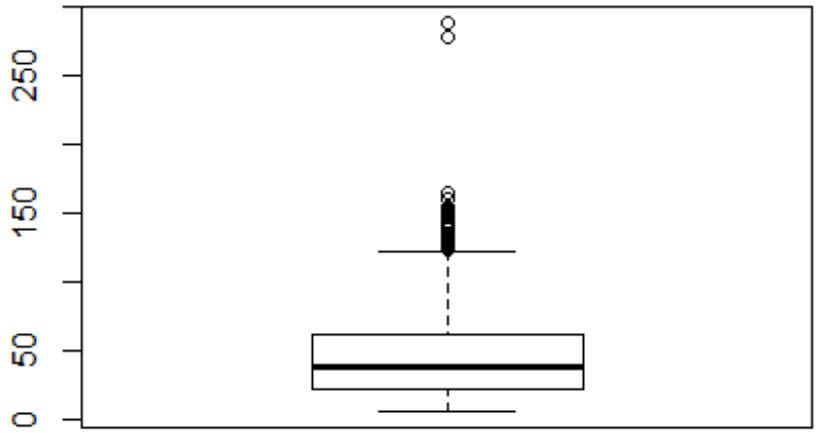

```
boxplot(datos$Cloruros)
```



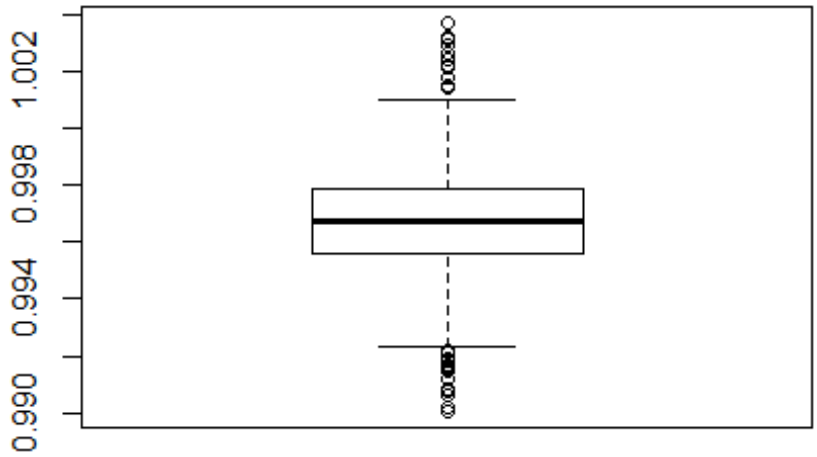

```
boxplot(datos$Dioxido_de_Azufre_Libre)
```



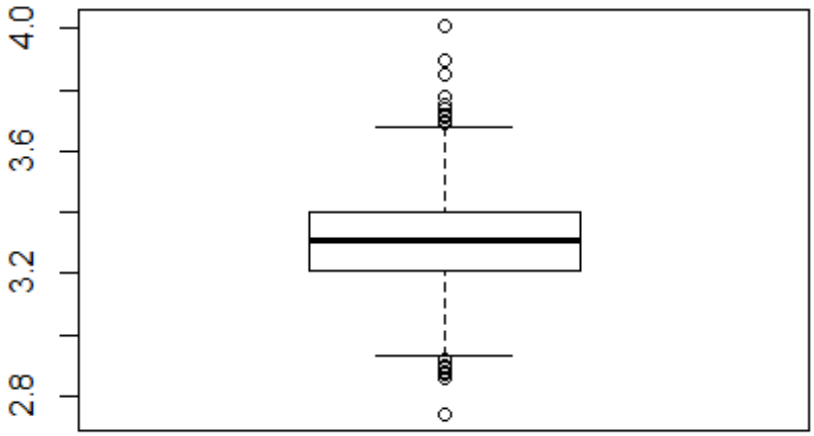

```
boxplot(datos$Dioxido_de_Azufre_Total)
```



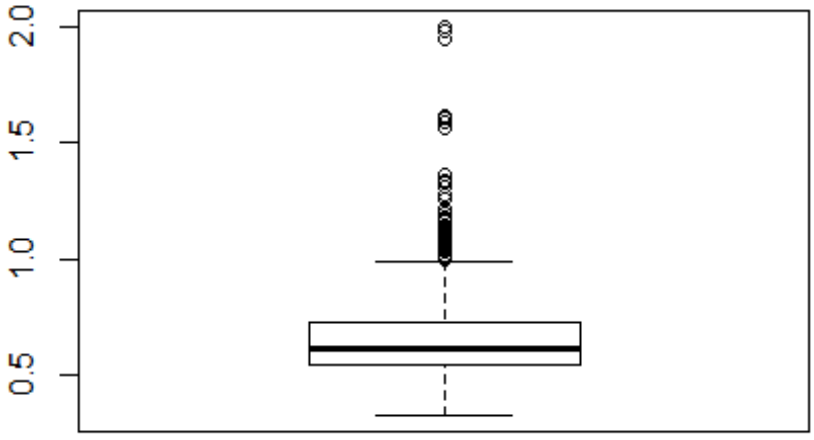

```r
boxplot(datos$Densidad)
```



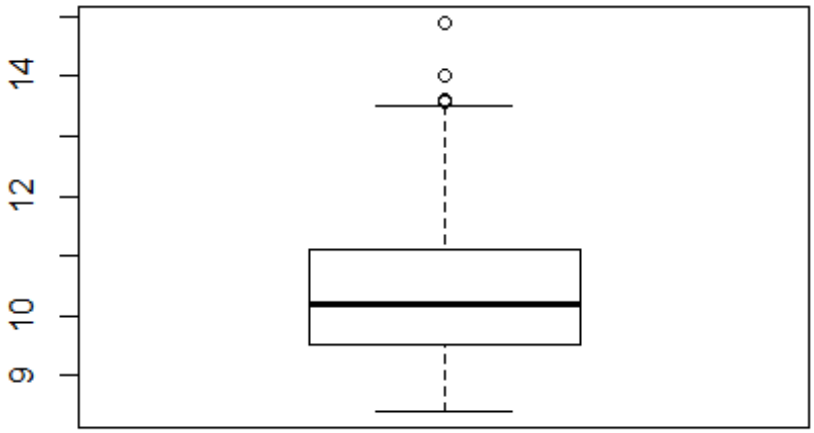

```r
boxplot(datos$pH)
```



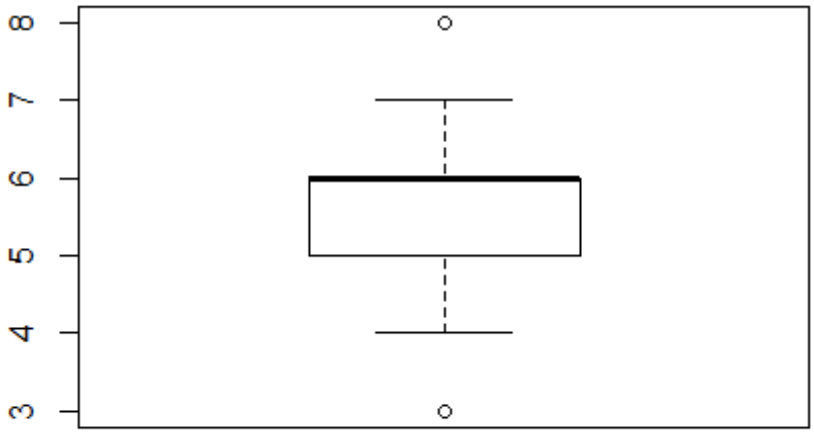

```r
boxplot(datos$Sulfatos)
```




```r
boxplot(datos$Alcohol)
```




```
boxplot(datos$Calidad)
```



Visualizamos los datos concretos

```r
boxplot.stats(datos$Acidez_Fija)$out
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5
12.8
## [15] 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3
12.4
## [29] 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7
13.2
## [43] 13.2 13.2 15.9 13.3 12.9 12.6 12.6
```

```r
boxplot.stats(datos$Acidez_Volatil)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
## [12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

```r
boxplot.stats(datos$Acido_Citrico)$out
```

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

```r
boxplot.stats(datos$Azucar_Residual)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80
5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60
4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00
11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90
3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55
4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60
4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40
4.25
##  [78]  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80
9.00
##  [89]  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90
4.00
## [100]  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50
5.50
## [111]  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90
4.30
## [122] 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40
3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20
3.75
## [144] 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90
5.10
## [155]  7.80
```

```r
boxplot.stats(datos$Cloruros)$out
```

```
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122
0.178
##  [12] 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358
0.343
##  [23] 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124
0.122
##  [34] 0.122 0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200
0.171
##  [45] 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039
0.157
##  [56] 0.422 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132
0.126
##  [67] 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161
0.120
##  [78] 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166
0.136
##  [89] 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168
0.415
## [100] 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039
0.235
## [111] 0.230 0.038
```

```
boxplot.stats(datos$Dioxido_de_Azufre_Libre)$out
```

```
##  [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43
51
## [24] 51 52 55 55 48 48 66
```

```
boxplot.stats(datos$Dioxido_de_Azufre_Total)$out
```

```
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144
127
## [18] 126 145 144 135 165 124 124 134 124 129 151 133 142 149 147 145
148
## [35] 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133
147
## [52] 147 131 131 131
```

```
boxplot.stats(datos$Densidad)$out
```

```
##  [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220
##  [9] 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140
## [17] 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260
## [25] 0.99210 0.99154 0.99064 0.99064 1.00289 0.99162 0.99007 0.99007
## [33] 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369
## [41] 1.00369 1.00242 0.99182 1.00242 0.99182
```

```
boxplot.stats(datos$pH)$out
```

```
##  [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72
2.87
## [15] 2.89 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70
```

```
3.78
## [29] 4.01 2.90 4.01 3.71 2.88 3.72 3.72
```

```
boxplot.stats(datos$Sulfatos)$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00
1.08
## [15] 1.59 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11
1.13
## [29] 1.07 1.06 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05
1.17
## [43] 1.62 1.06 1.18 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17
1.03
## [57] 1.17 1.10 1.01
```

```
boxplot.stats(datos$Alcohol)$out
```

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000
##  [8] 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
```

```
boxplot.stats(datos$Calidad)$out
```

```
##  [1] 8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8
```

Aunque en todas las variables se presentan valores extremos, vamos a dejarlos tal cual para poder estudiar como se comporta la calidad del producto final con la presencia de estos, si son influyentes o no.

Análisis de los datos

Seleccion de los grupos de datos a analizar

```
summary(datos)
```

```
##   Acidez_Fija    Acidez_Volatil   Acido_Citrico   Azucar_Residual
##  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
##     Cloruros       Dioxido_de_Azufre_Libre Dioxido_de_Azufre_Total
##  Min.   :0.01200   Min.   : 1.00           Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00           1st Qu.: 22.00
##  Median :0.07900   Median :14.00           Median : 38.00
##  Mean   :0.08747   Mean   :15.87           Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00           3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00           Max.   :289.00
```

```
##     Densidad            pH           Sulfatos         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 :0.9968   Median :3.310   Median :0.6200   Median :10.20  
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42  
##  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  
##     Calidad        
##  Min.   :3.000  
##  1st Qu.:5.000  
##  Median :6.000  
##  Mean   :5.636  
##  3rd Qu.:6.000  
##  Max.   :8.000
```

```
table(datos$Calidad)
```

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

Como a priori no pudemos agrupar por ninguna variable vamos a interpredar en base a los datos que tenemos de la calidad, agrupando como malo los que tengan una calificación 4 o menor, normal entre 5 y 6, y bueno cuando sea mayor que 7

```
vino.malo <- datos[datos$Calidad <= 4,]
vino.normal <- datos[datos$Calidad > 4 & datos$Calidad < 7,]
vino.bueno <- datos[datos$Calidad >= 7,]
```

Comprobación de la normalidad y homogeneidad de la varianza

Para saber si las variables están normalizadas aplicaremos el test de Shapiro Wilk para cada variable

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

```
## 
##  Shapiro-Wilk normality test
## 
## data:  datos$Acidez_Fija
## W = 0.94203, p-value < 2.2e-16
```

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

```
## 
##  Shapiro-Wilk normality test
## 
## data:  datos$Acidez_Volatil
## W = 0.97434, p-value = 2.693e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Acido_Citrico
## W = 0.95529, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Azucar_Residual
## W = 0.56608, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Cloruros
## W = 0.48425, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Dioxido_de_Azufre_Libre
## W = 0.90184, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Dioxido_de_Azufre_Total
## W = 0.87322, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$Densidad
## W = 0.99087, p-value = 1.936e-08
```

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

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

```
shapiro.test(datos$Sulfatos)

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

shapiro.test(datos$Alcohol)

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

shapiro.test(datos$Calidad)

##
##  Shapiro-Wilk normality test
##
## data:  datos$Calidad
## W = 0.85759, p-value < 2.2e-16
```

Se aprecia que para cada variable su p-valor es inferior a 0.05 por lo que rechazamos la hipotesis nula y entendemos que las variables no son normales.

Para estudiar la homogeneidad de la varianza aplicaremos el test de Fligner-killen, estudiaremos esta homogeneidad según los niveles de azucar frente a la calidad final de los vinos.

```
fligner.test(Azucar_Residual ~ Calidad, data = datos)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Azucar_Residual by Calidad
## Fligner-Killeen:med chi-squared = 7.9984, df = 5, p-value = 0.1563
```

Se obserta que el p-valor es superior al 0.05, por lo que aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

Pruebas estadísticas

Primeramente vamos a relizar un análisis de correlación de las distintas variables para determinar cual de ellas tienen más peso a la hora de establecer la calidad del vino. Nos basaremos en el coeficiente de correlación de Spearman puesto que no siguen una distribución normal los datos.

```
corr <- matrix(nc = 2, nr = 0)
colnames(corr) <- c("estimado","p-valor")
```

```r
for(i in 1:(ncol(datos)-1)){
  test = cor.test(datos[,i],datos[,length(datos)],method = "spearman")
  est = test$estimate
  pval = test$p.value

  par = matrix(ncol = 2, nrow = 1)
  par[1][1] = est
  par[2][1] = pval
  corr <- rbind(corr,par)
  rownames(corr)[nrow(corr)] <- colnames(datos)[i]
}
```

```
## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
```

```
=
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(datos[, i], datos[, length(datos)], method
=
## "spearman"): Cannot compute exact p-value with ties
```

```
print(corr)
```

```
##                              estimado      p-valor
## Acidez_Fija                0.11408367 4.801220e-06
## Acidez_Volatil            -0.38064651 2.734944e-56
## Acido_Citrico              0.21348091 6.158952e-18
## Azucar_Residual            0.03204817 2.002454e-01
## Cloruros                  -0.18992234 1.882858e-14
## Dioxido_de_Azufre_Libre -0.05690065 2.288322e-02
## Dioxido_de_Azufre_Total -0.19673508 2.046488e-15
## Densidad                  -0.17707407 9.918139e-13
## pH                        -0.04367193 8.084594e-02
## Sulfatos                   0.37706020 3.477695e-55
## Alcohol                    0.47853169 2.726838e-92
```

A tenor de los datos obtenidos, podríamos aventurar que el Alcohol es la variable que más peso tiene a la hora establecer la calidad del vino, ya que es el más proximo a los valores 1 o -1

Como siguiente prueba estadística vamos a realizar un contraste de hipótesis en el vamos a ver si un nivel superior de azucares determina una mayor calidad del vino,

```
summary(datos$Azucar_Residual)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

Vamos a crear dos grupos diferentes en función del nivel de azucar, que esten por debajo de la mediana y por encima

```
mediana.azucar <- median(datos$Azucar_Residual)
vino.bajo.azucar <- datos[datos$Azucar_Residual <=
mediana.azucar,]$Calidad
vino.alto.azucar <- datos[datos$Azucar_Residual >
mediana.azucar,]$Calidad
```

A partir de estas dos muestras plantearemos el contraste de hipotesis, donde definimos como hipotesis nula que la calidad no se ve afectada por la cantidad de azucar y como alternativa que a un mayor nivel de azucar la calidad del vino es mayor.

```
t.test (vino.bajo.azucar, vino.alto.azucar, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  vino.bajo.azucar and vino.alto.azucar
## t = -1.5942, df = 1524.6, p-value = 0.05555
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.002099426
## sample estimates:
## mean of x mean of y
##  5.607022  5.671788
```

Puesto que el p-valor obtenido es su perior a 0.05 aceptamos la hipótesis nula, en donde la calidad del vino se se ve afectada por la cantidad de azucar. Esto se ve reflajado en la pueba estadística anterior donde veíamos que el azucar tenia poco peso en la calidad.

Conclusiones

Se han realizado dos pruebas estadísticas sobre un conjunto de datos en donde se registran distintas variables que determinan la calidad del vino tinto portugues “Vinho Verde”, este siendo determinado por expertos.

El análisis de correlación nos a llevado a determinar que la calidad del vino se ve más influenciada por el Alcohol del mismo en deprimento del resto de variables, además hemos determinado por contraste de hipótesis que el azucar no influye en la caliadd del mismo, esto también corroborado por el por el análisis de correlación al tener este bajo peso de influencia.

El preprocesado previo de los datos a sido mínimo ya que no se observaron valores perdidos ni ceros, por último los valores extremos se ha decidido dejarlos, ya que no se observar valores extraños sobre todo en la nivel de Alcohol que es el que más conocimiento se puede tener.