

# Práctica 2: Limpieza y validación de los datos

Waziri Ajibola Lawal, David Fernández González

1/3/2021

---

## Realización de la práctica

---

### Procesos de limpieza del conjunto de datos

Primer contacto con el conjunto de datos, visualizamos su estructura.

```
# Cargamos los paquetes R que vamos a usar
library(ggplot2)
library(dplyr)
library(corrplot)
library(psych)

if(!require(corrplot)){
  install.packages('corrplot', repos='http://cran.us.r-project.org')
  library(corrplot)
}

# Cargamos el fichero de datos
redWineData <- read.csv('winequality-red.csv', stringsAsFactors = FALSE, header = TRUE)
#filas=dim(redWineData)[1]

attach(redWineData)

# Verificamos la dimension del conjunto de datos
dim(redWineData)

## [1] 1599    12

# Verificamos la estructura del conjunto de datos
sapply(redWineData, class)

##          fixed.acidity      volatile.acidity       citric.acid
##             "numeric"           "numeric"           "numeric"
##          residual.sugar      chlorides   free.sulfur.dioxide
```

```

##          "numeric"      "numeric"      "numeric"
## total.sulfur.dioxide      density          pH
##          "numeric"      "numeric"      "numeric"
##      sulphates      alcohol       quality
##          "numeric"      "numeric"      "integer"

# Verificamos la estructura del conjunto de datos
str(redWineData)

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

# Verificamos la distribución de los datos
head(redWineData)

##   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

# Estadísticas básicas
summary(redWineData)

##   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

```

```
describe(redWineData)
```

	vars	n	mean	sd	median	trimmed	mad	min	max
## fixed.acidity	1	1599	8.32	1.74	7.90	8.15	1.48	4.60	15.90
## volatile.acidity	2	1599	0.53	0.18	0.52	0.52	0.18	0.12	1.58
## citric.acid	3	1599	0.27	0.19	0.26	0.26	0.25	0.00	1.00
## residual.sugar	4	1599	2.54	1.41	2.20	2.26	0.44	0.90	15.50
## chlorides	5	1599	0.09	0.05	0.08	0.08	0.01	0.01	0.61
## free.sulfur.dioxide	6	1599	15.87	10.46	14.00	14.58	10.38	1.00	72.00
## total.sulfur.dioxide	7	1599	46.47	32.90	38.00	41.84	26.69	6.00	289.00
## density	8	1599	1.00	0.00	1.00	1.00	0.00	0.99	1.00
## pH	9	1599	3.31	0.15	3.31	3.31	0.15	2.74	4.01
## sulphates	10	1599	0.66	0.17	0.62	0.64	0.12	0.33	2.00
## alcohol	11	1599	10.42	1.07	10.20	10.31	1.04	8.40	14.90
## quality	12	1599	5.64	0.81	6.00	5.59	1.48	3.00	8.00
			range	skew	kurtosis	se			
## fixed.acidity			11.30	0.98	1.12	0.04			
## volatile.acidity			1.46	0.67	1.21	0.00			
## citric.acid			1.00	0.32	-0.79	0.00			
## residual.sugar			14.60	4.53	28.49	0.04			
## chlorides			0.60	5.67	41.53	0.00			
## free.sulfur.dioxide			71.00	1.25	2.01	0.26			
## total.sulfur.dioxide			283.00	1.51	3.79	0.82			
## density			0.01	0.07	0.92	0.00			
## pH			1.27	0.19	0.80	0.00			
## sulphates			1.67	2.42	11.66	0.00			
## alcohol			6.50	0.86	0.19	0.03			
## quality			5.00	0.22	0.29	0.02			

```
# Verificamos si existen valores vacíos en el conjunto de datos
colSums(is.na(redWineData))
```

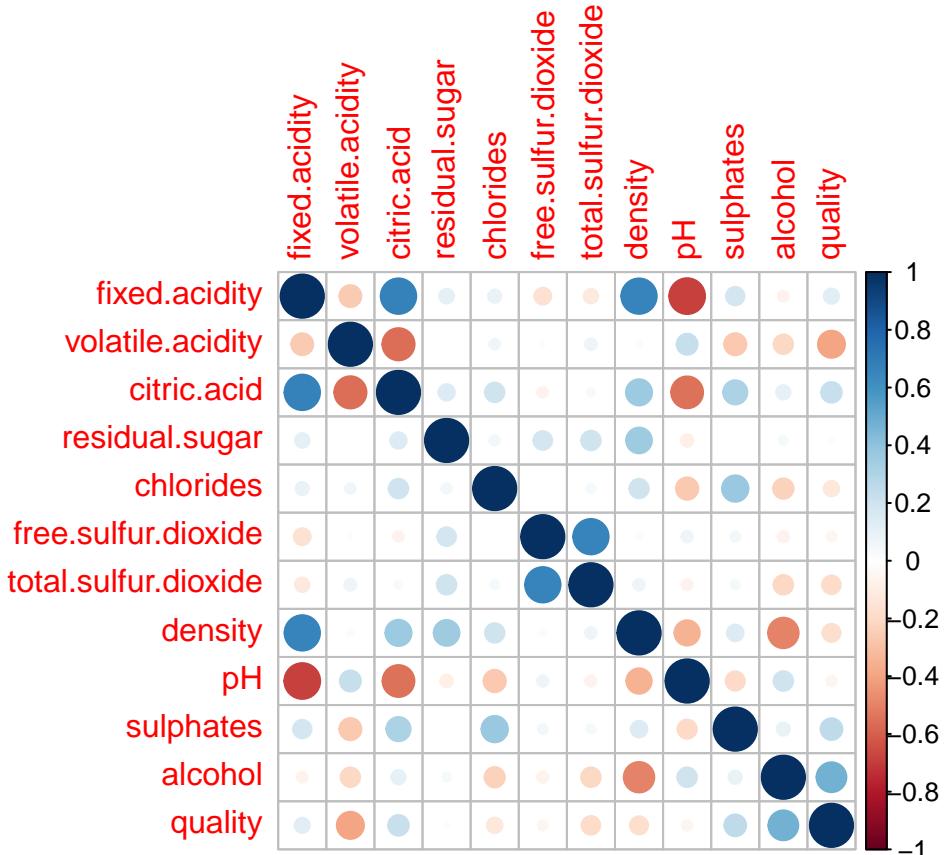
```
##      fixed.acidity      volatile.acidity      citric.acid
##                0                 0                  0
```

```

##          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

#Correlation Heatmap of Variables
corrplot(cor(redWineData))

```



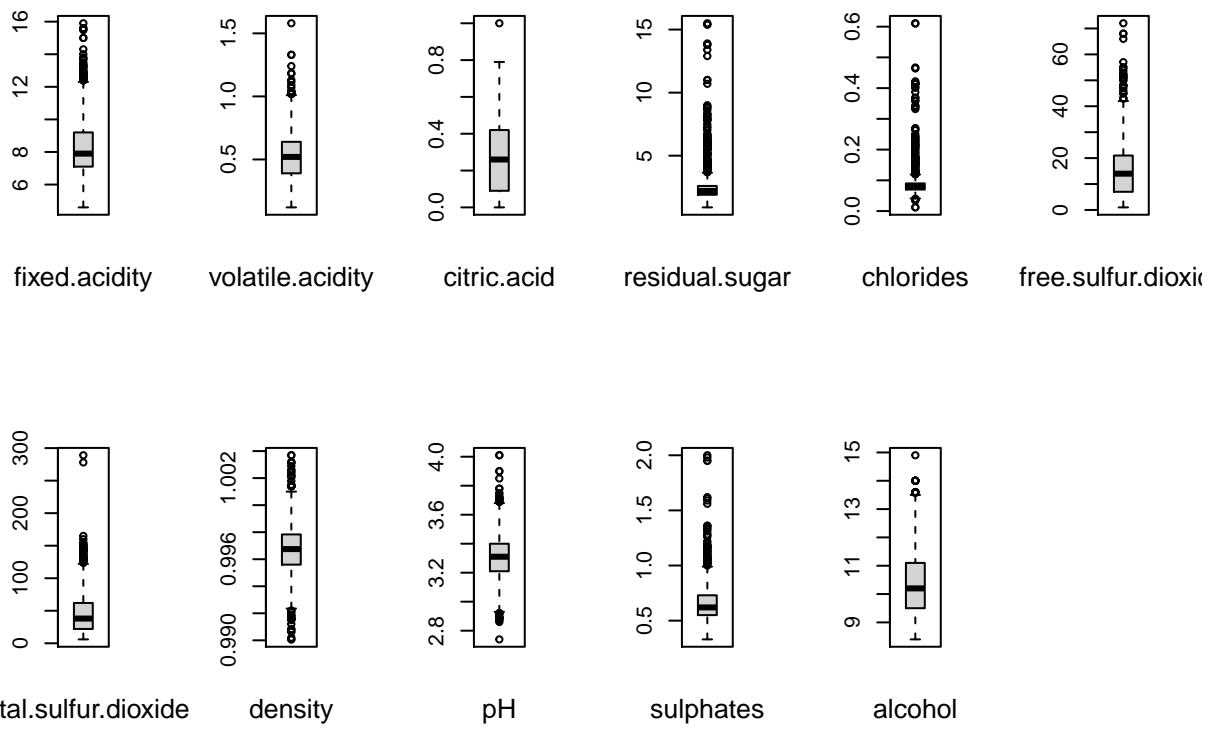
El conjunto de datos de vino tinto contiene 1599 observaciones, 11 predictores y 1 valor categórico que indica la calidad del vino. Todos los predictores son valores numéricos, los resultados son enteros. Como podemos observar, no existen valores vacíos en el conjunto de datos.

Las estadísticas resumidas muestran que la mayoría de las variables tienen un rango amplio en comparación con el rango intercuartil, lo que puede indicar una dispersión en los datos y la presencia de valores atípicos. Investigamos más a fondo produciendo diagramas de caja para cada una de las variables:

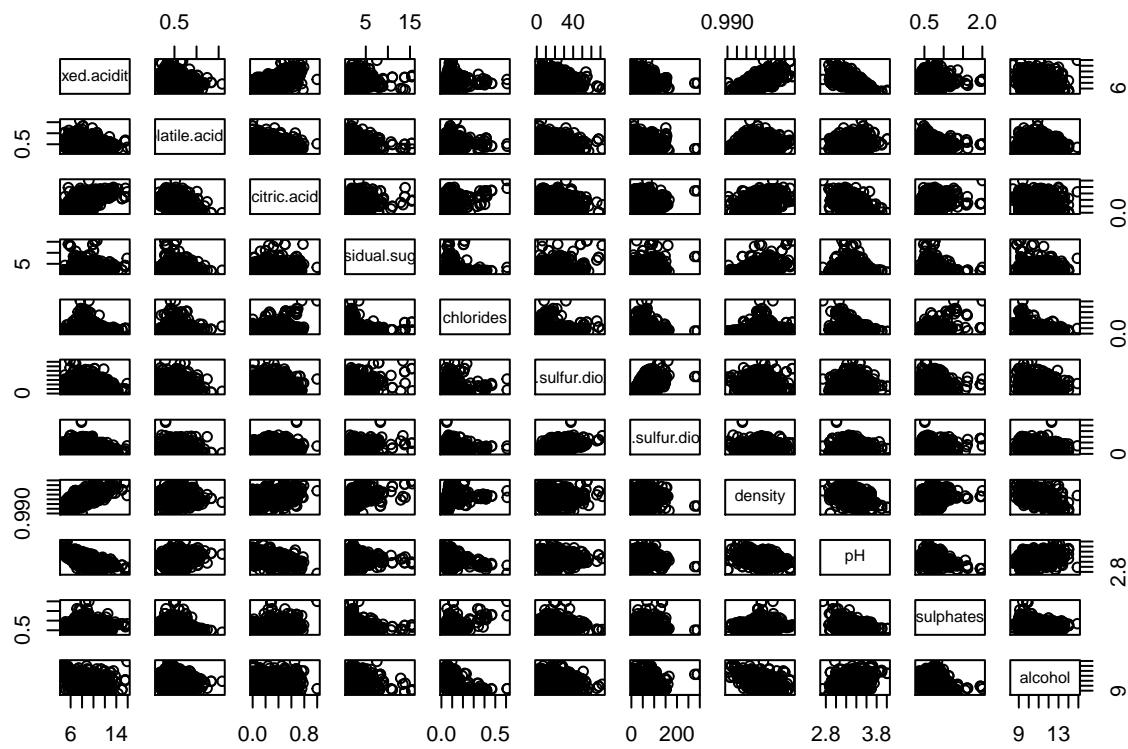
```

oldpar = par(mfrow = c(2,6))
for ( i in 1:11 ) {
  boxplot(redWineData[[i]])
  mtext(names(redWineData)[i], cex = 0.8, side = 1, line = 2)
}
par(oldpar)

```



```
pairs(redWineData[, -grep("quality", colnames(redWineData))])
```

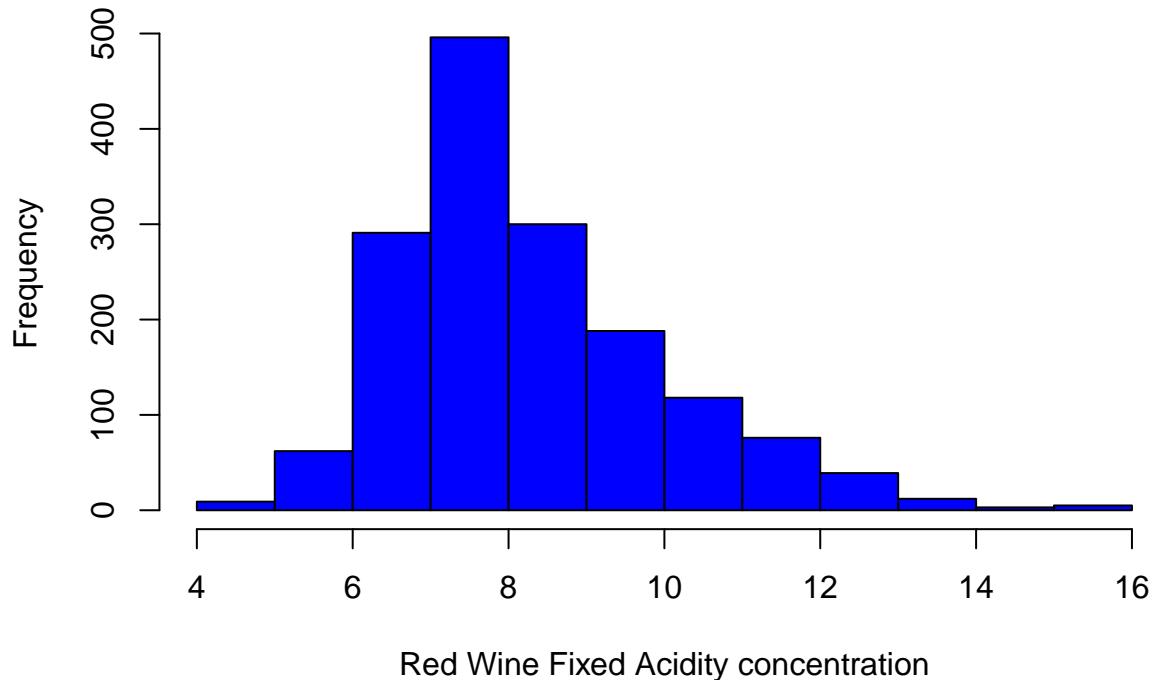


Podemos ver que todas las variables contienen valores atípicos. Estos valores atípicos se encuentran en los extremos superiores.

Procedemos a la generación de histogramas para entender la distribución de cada variable (predictor).

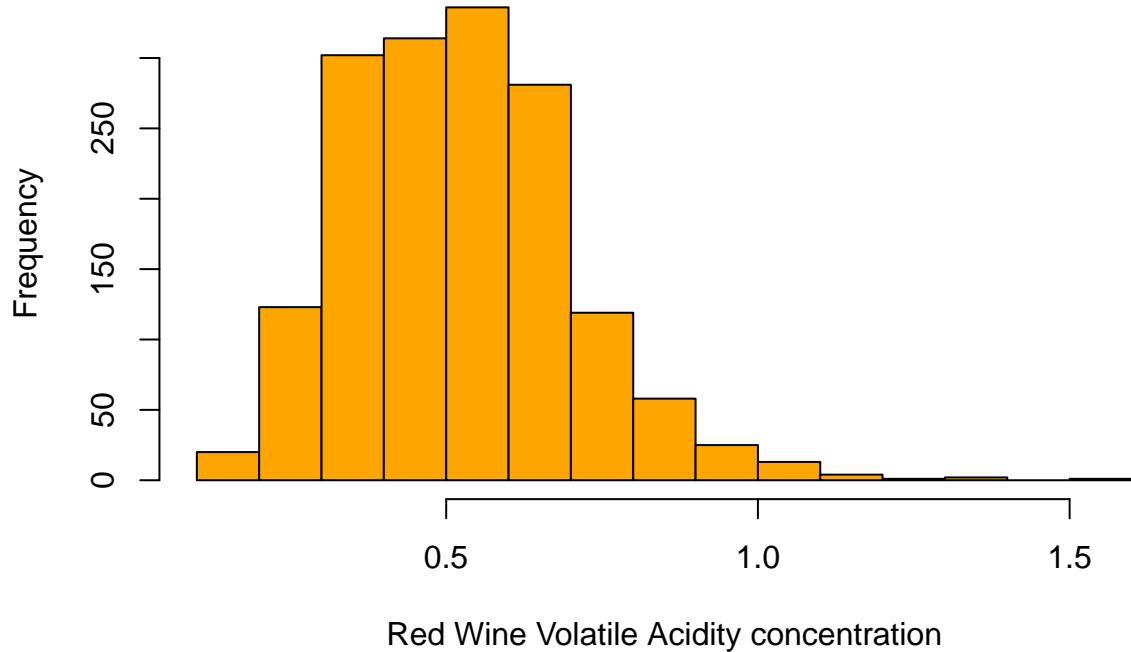
```
# Histograma de la variable Fixed Acidity
hist(fixed.acidity, main = "Red Wine fixed acidity", xlab="Red Wine Fixed Acidity concentration", col="red")
```

## Red Wine fixed acidity



```
# Histograma de la variable Volatile Acidity
hist(volatile.acidity, main = "Red Wine volatile Acidity distribution", xlab="Red Wine Volatile Acidity")
```

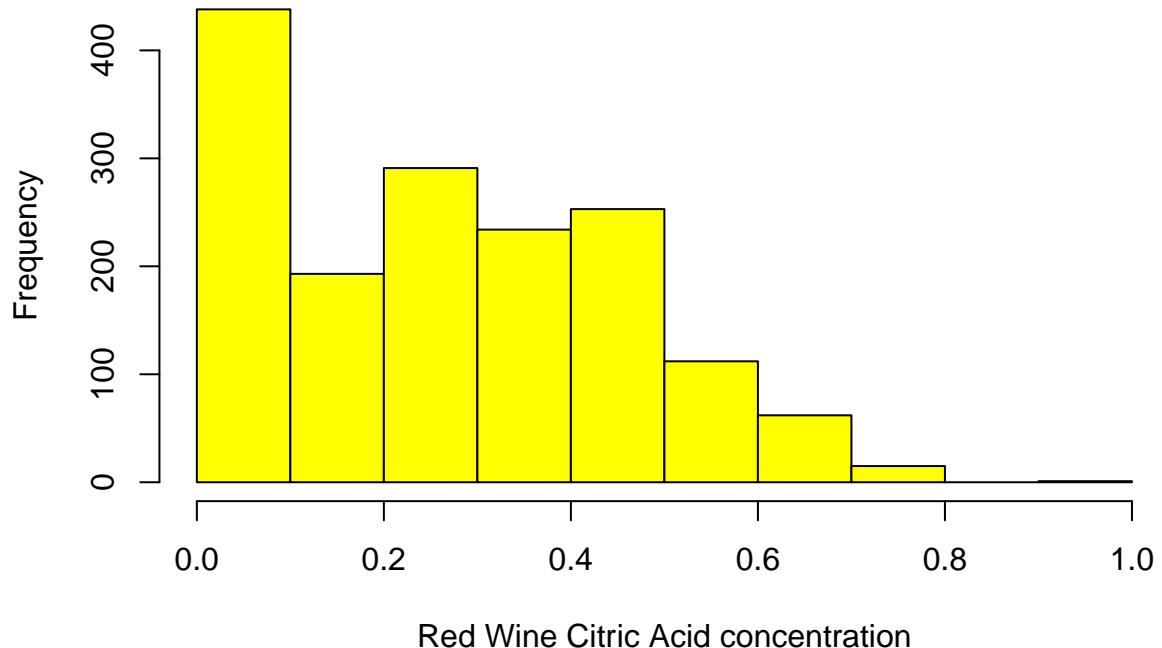
## Red Wine volatile Acidity distribution



```
# Histograma de la variable Citric Acid
```

```
hist(citric.acid, main = "Red Wine Citric acid distribution", xlab="Red Wine Citric Acid concentration")
```

## Red Wine Citric acid distribution



```
# Histograma de la variable Residual Sugar
hist(residual.sugar, main = "Red Wine Residual Sugar distribution", xlab="Red Wine Residual Sugar concen
```

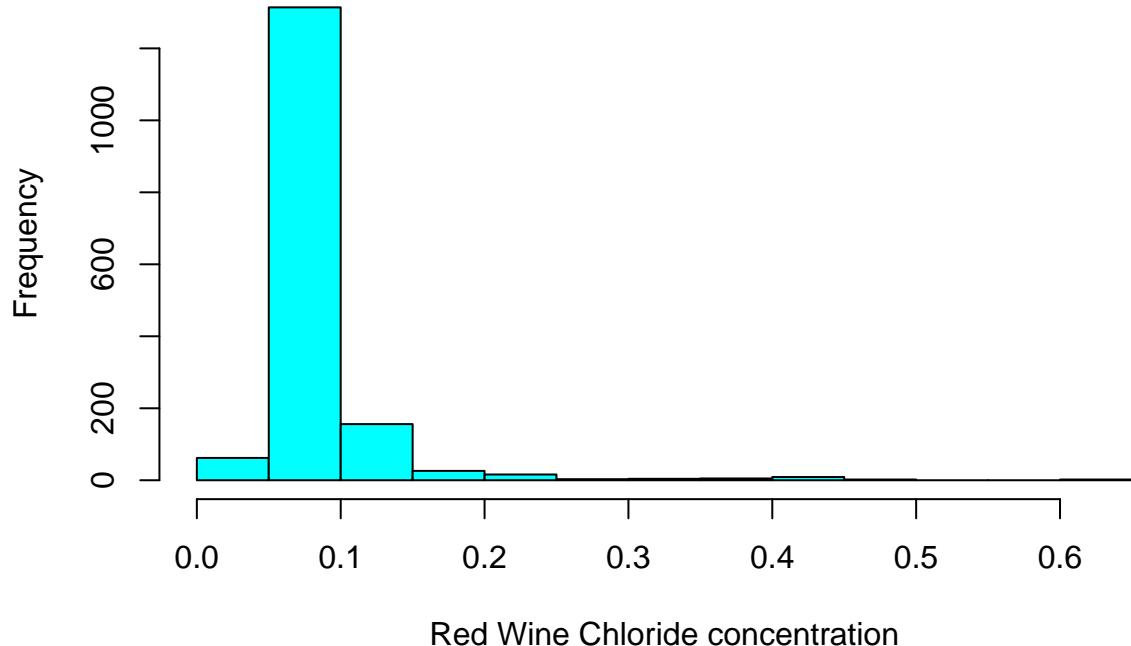
## Red Wine Residual Sugar distribution



```
# Histograma de la variable Chlorides
```

```
hist(chlorides, main = "Red Wine Chloride distribution", xlab="Red Wine Chloride concentration", col="Cyan")
```

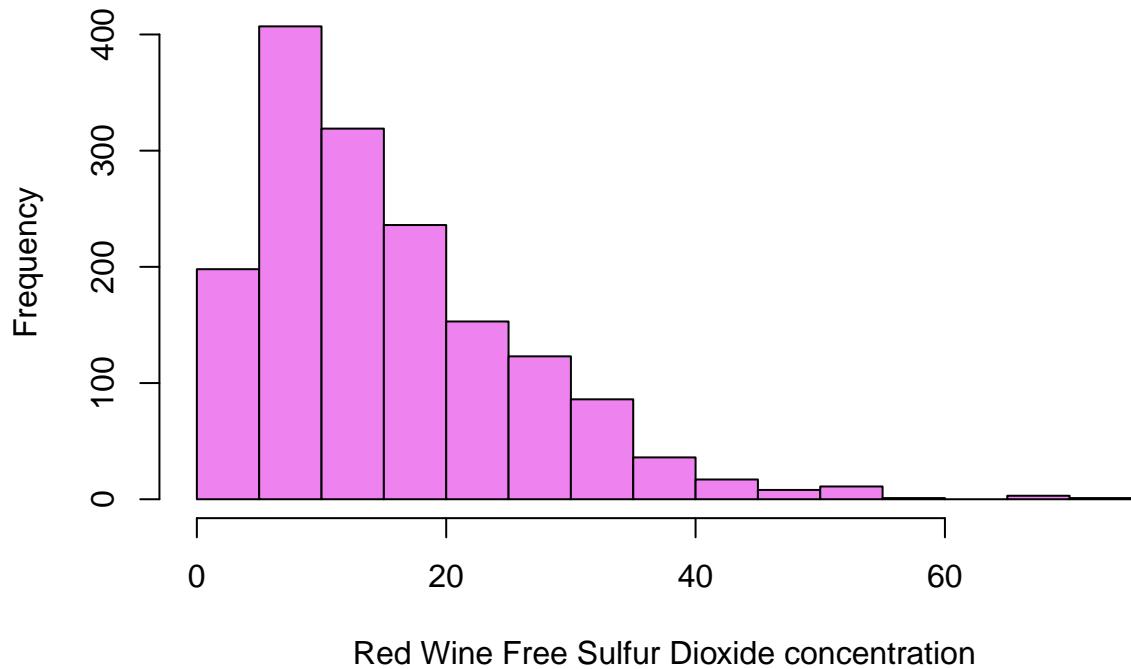
## Red Wine Chloride distribution



```
# Histograma de la variable Free Sulfur Dioxide
```

```
hist(free.sulfur.dioxide, main = "Red Wine Free Sulfur Dioxide distribution", xlab="Red Wine Free Sulfur Dioxide")
```

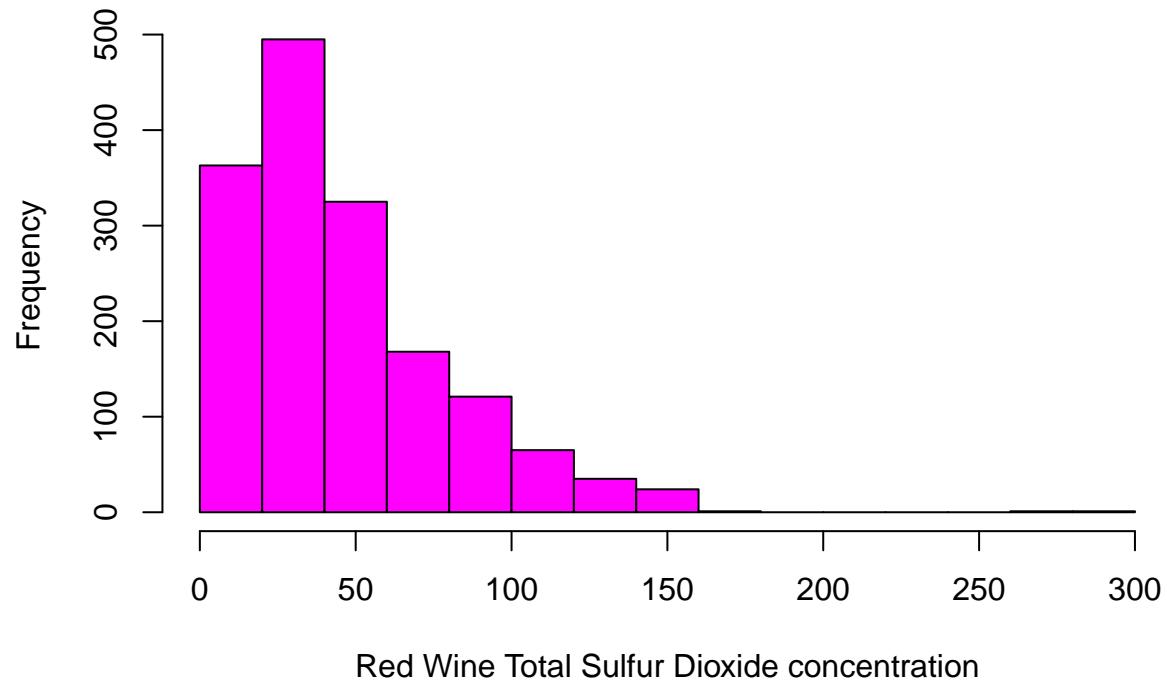
## Red Wine Free Sulfur Dioxide distribution



```
# Histograma de la variable Total Sulfur Dioxide
```

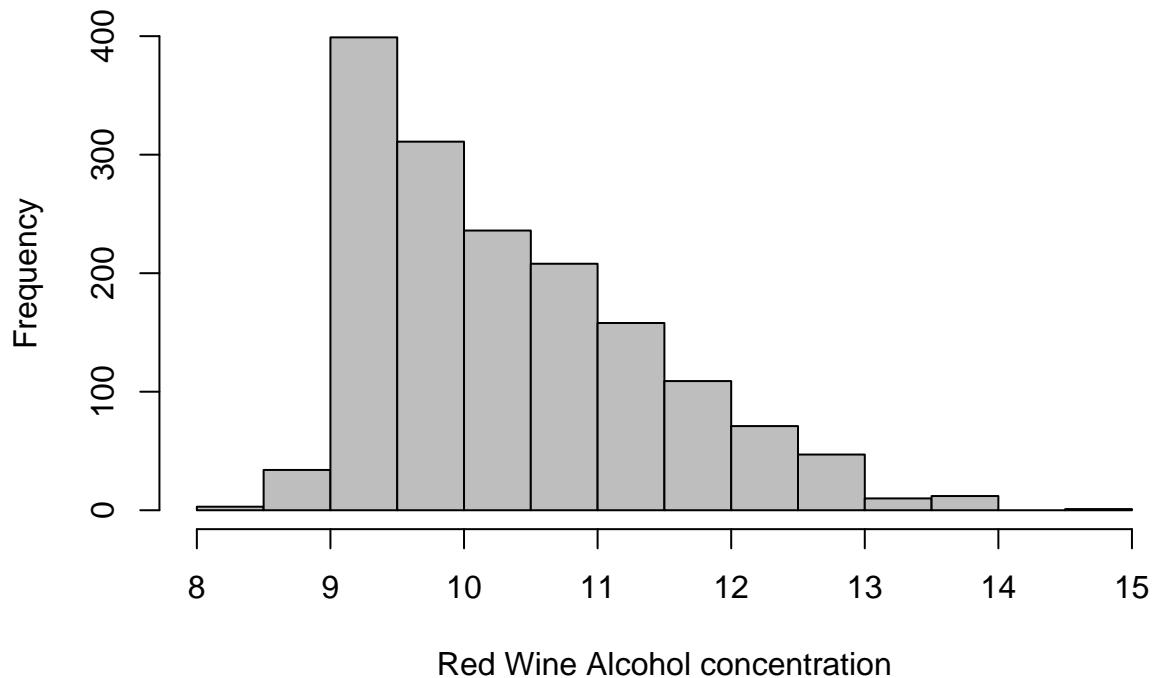
```
hist(total.sulfur.dioxide, main = "Red Wine Total Sulfur Dioxide distribution", xlab="Red Wine Total Su")
```

## Red Wine Total Sulfur Dioxide distribution



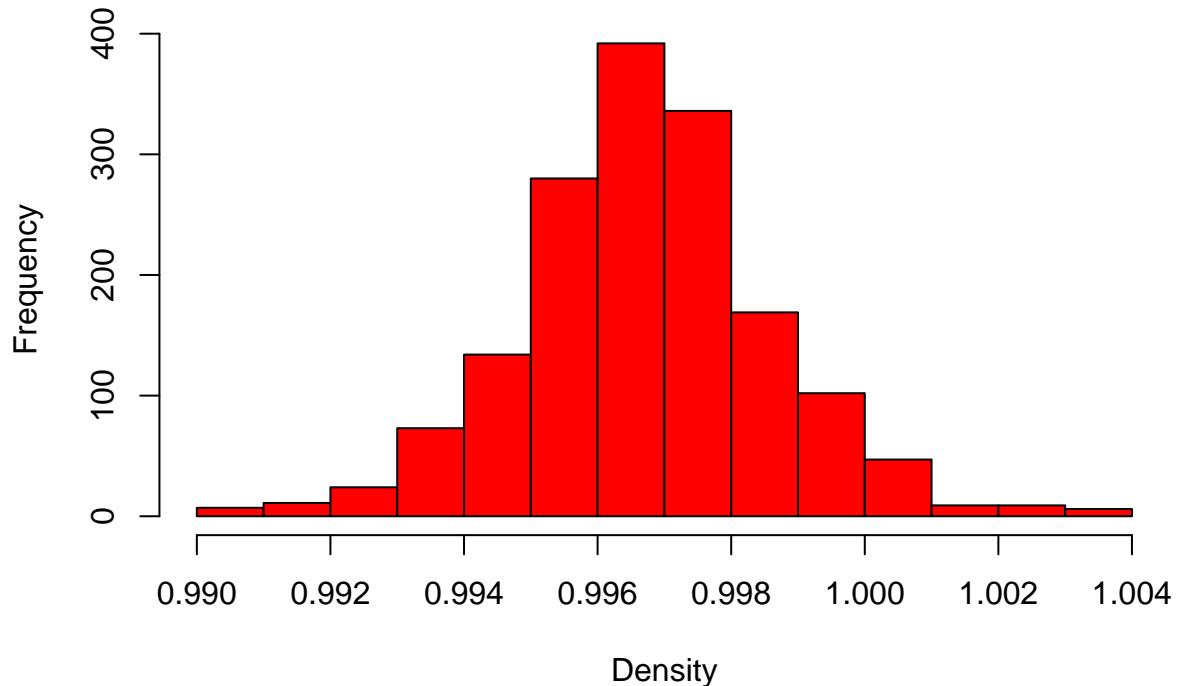
```
# Histograma de la variable Alcohol
hist(alcohol, main = "Red Wine Alcohol distribution", xlab="Red Wine Alcohol concentration", col="Grey")
```

## Red Wine Alcohol distribution



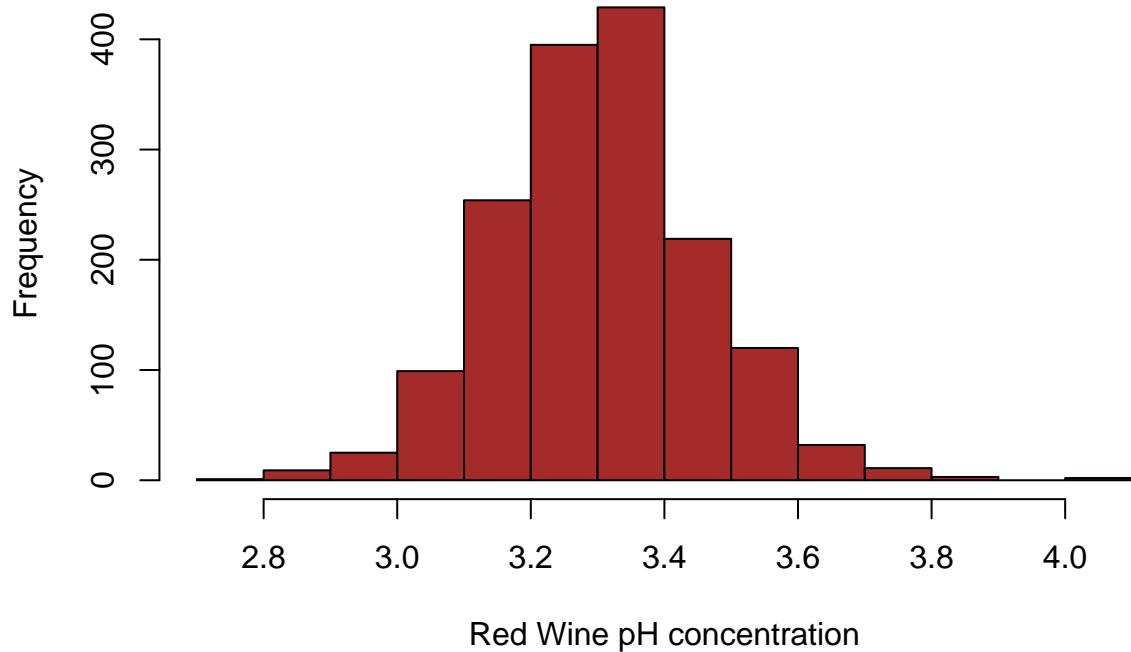
```
# Histograma de la variable Density
hist(density, main = "Red Wine Density distribution", xlab="Density", col="Red")
```

## Red Wine Density distribution



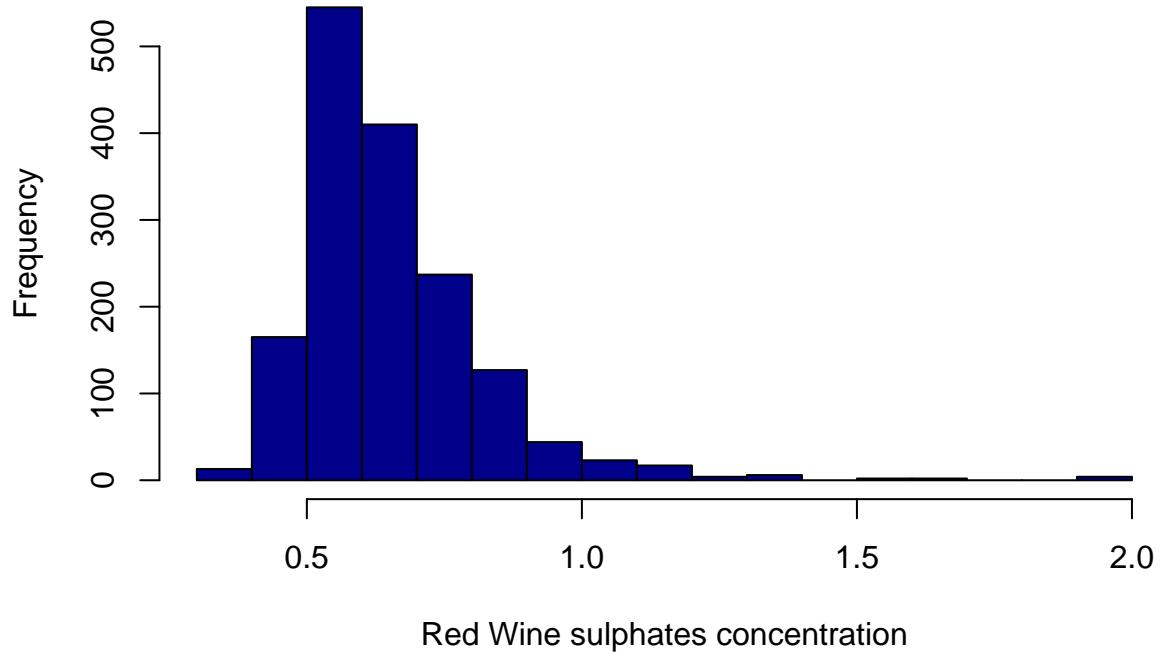
```
# Histograma de la variable pH
hist(pH, main = "Red Wine pH distribution", xlab="Red Wine pH concentration", col="Brown")
```

## Red Wine pH distribution

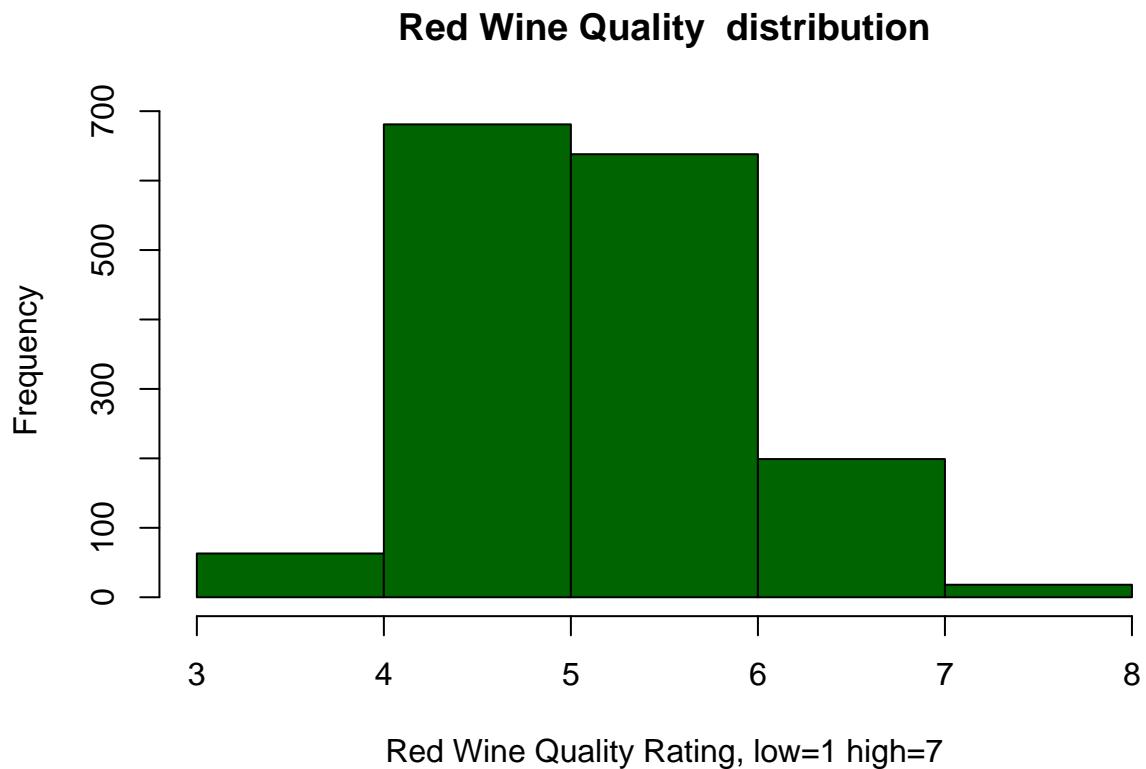


```
# Histograma de la variable Sulphates
hist(sulphates, main = "Red Wine sulphates distribution", xlab="Red Wine sulphates concentration", col=
```

## Red Wine sulphates distribution



```
# Histograma de la variable quality
hist(quality, breaks=6, col="Dark Green", xlab="Red Wine Quality Rating, low=1 high=7", main="Red Wine Q")
```



Mediante los histogramas podemos observar que casi todas las distribuciones están sesgadas positivamente. Las variables pH, density y quality tiene una distribución aproximadamente normal.

```

eliminated <- redWineData
for (i in 1:11) {
  # Q <- quantile(redWineData[[i]], probs=c(.25, .75), na.rm = FALSE)
  # iqr <- IQR(redWineData[[i]])
  # up <- Q[2]+1.5*iqr # Upper Range
  # low <- Q[1]-1.5*iqr # Lower Range
  # eliminated <- subset(redWineData, redWineData[[i]] > (Q[1] - 1.5*iqr) & redWineData[[i]] < (Q[2]+1.5*iqr))
  # ggbetweenstats(eliminated, quality, redWineData[[i]], outlier.tagging = TRUE)
  boxplot(redWineData[[i]], plot = FALSE)$out
  outliers <- boxplot(redWineData[[i]], plot = FALSE)$out
  eliminated <- eliminated[-which(eliminated[[i]] %in% outliers), ]
}

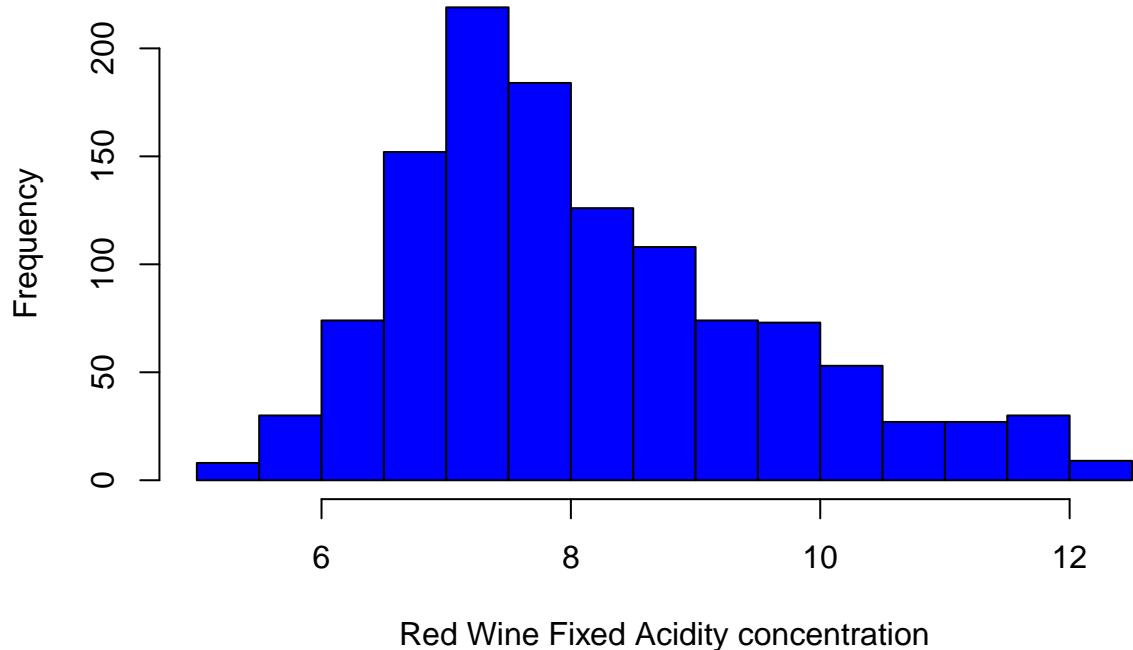
dim(eliminated)

## [1] 1194   12

# Histograma de la variable Fixed Acidity
hist(eliminated$fixed.acidity, main = "Red Wine fixed acidity", xlab="Red Wine Fixed Acidity concentration")

```

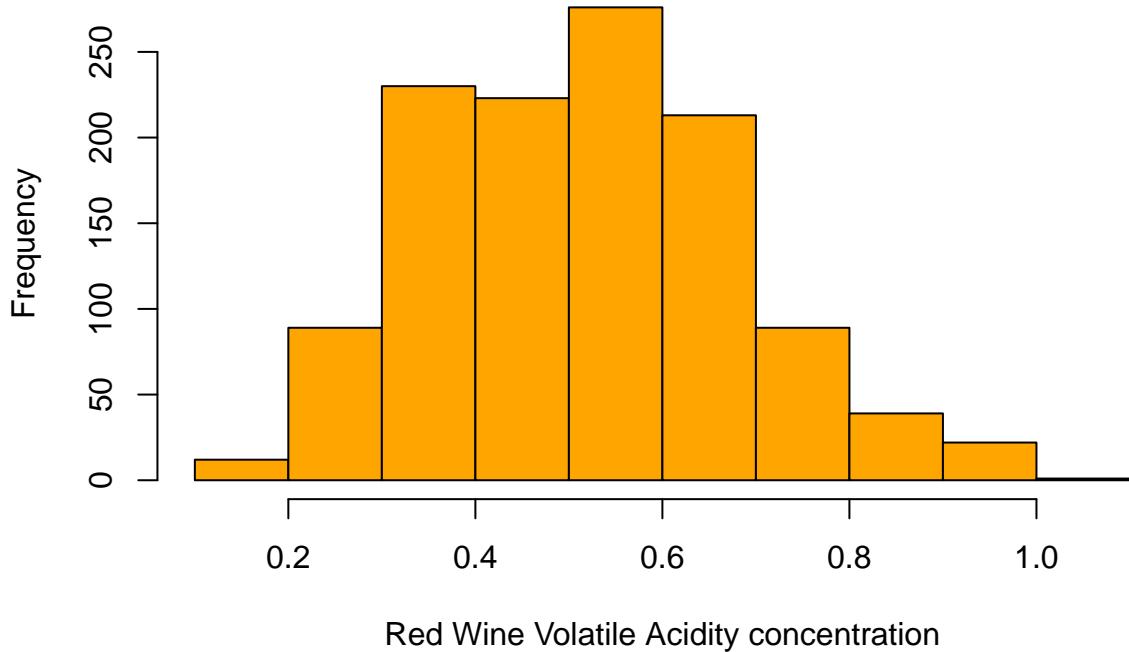
## Red Wine fixed acidity



```
# Histograma de la variable Volatile Acidity
```

```
hist(eliminated$volatile.acidity, main = "Red Wine volatile Acidity distribution", xlab="Red Wine Volatil")
```

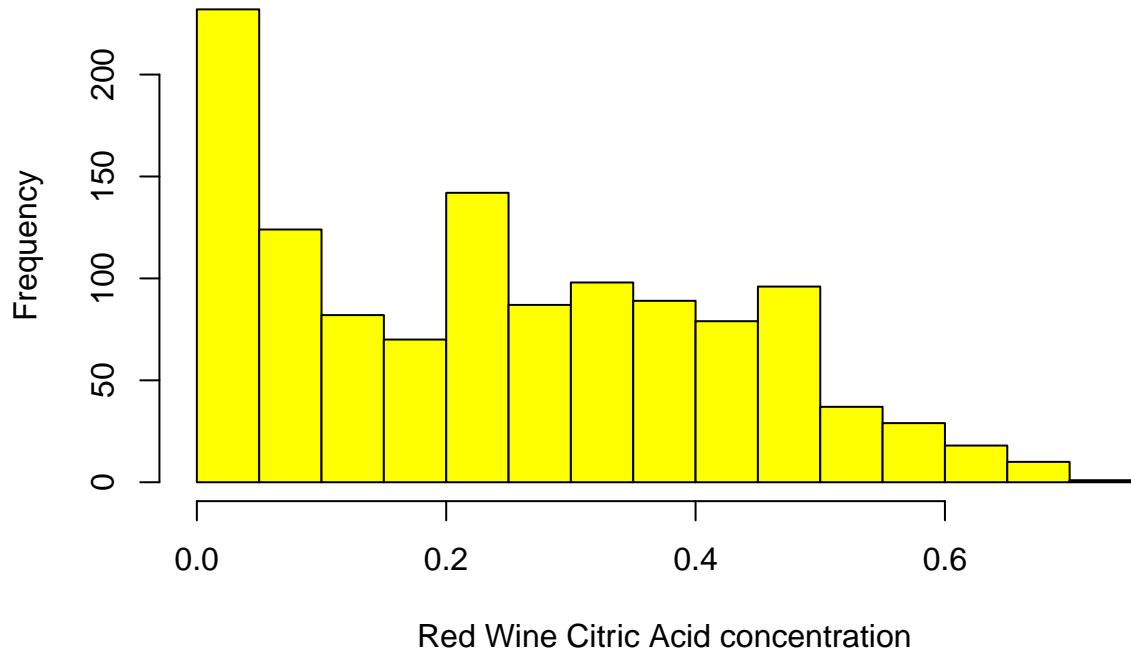
## Red Wine volatile Acidity distribution



```
# Histograma de la variable Citric Acid
```

```
hist(eliminated$citric.acid, main = "Red Wine Citric acid distribution", xlab="Red Wine Citric Acid con-
```

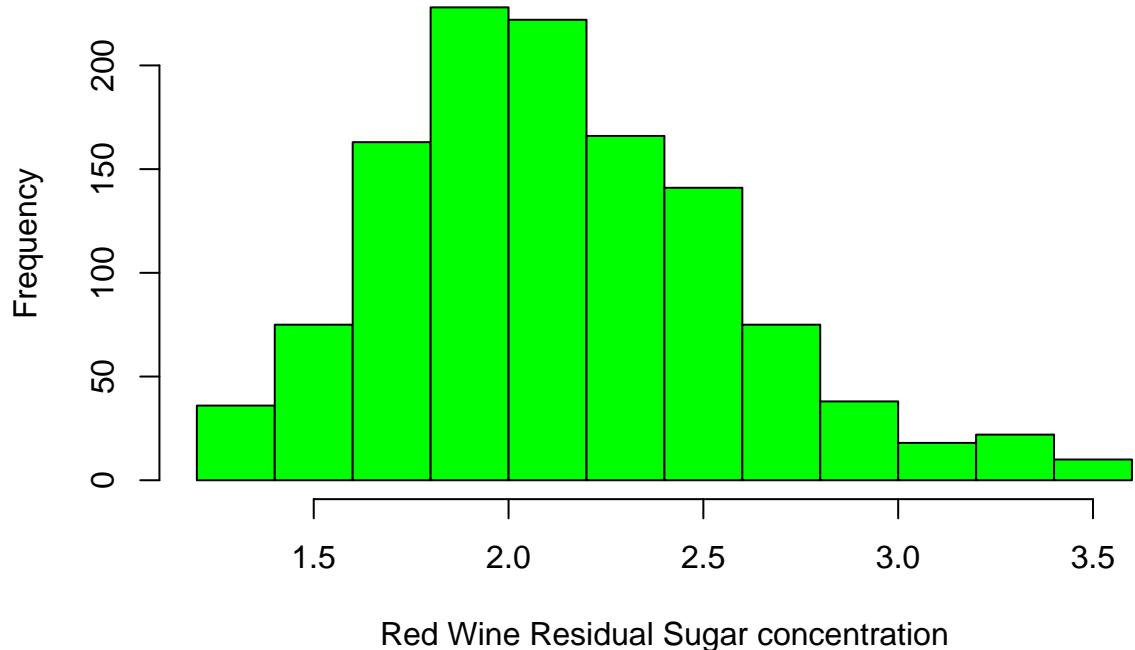
## Red Wine Citric acid distribution



```
# Histograma de la variable Residual Sugar
```

```
hist(eliminated$residual.sugar, main = "Red Wine Residual Sugar distribution", xlab="Red Wine Residual Sugar")
```

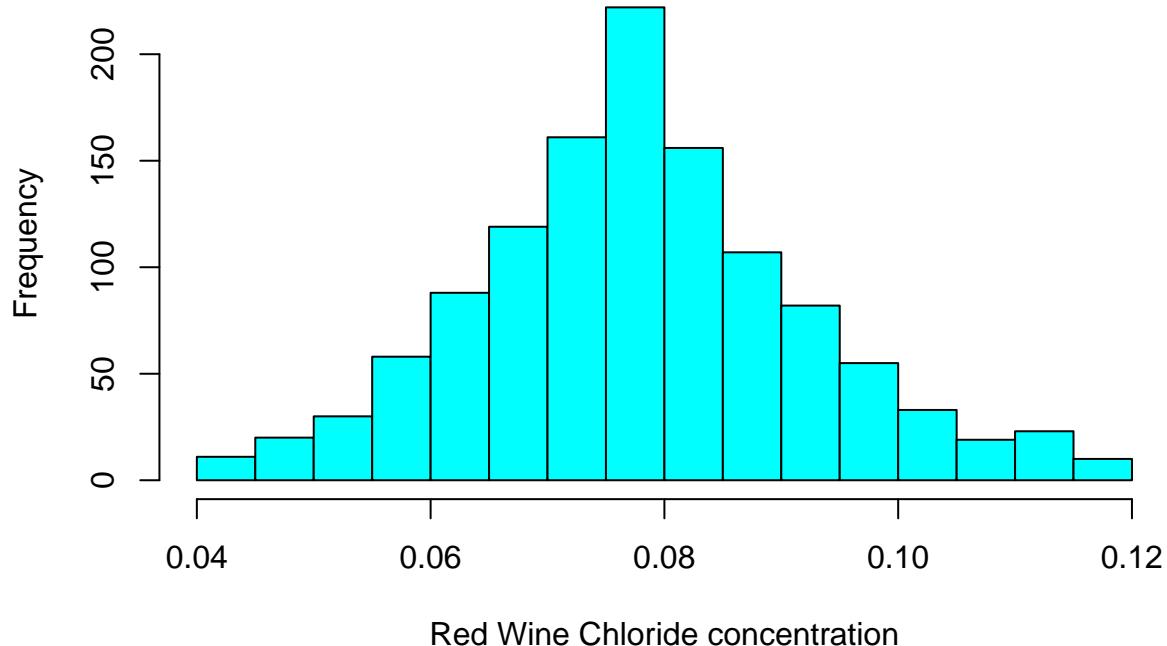
## Red Wine Residual Sugar distribution



```
# Histograma de la variable Chlorides
```

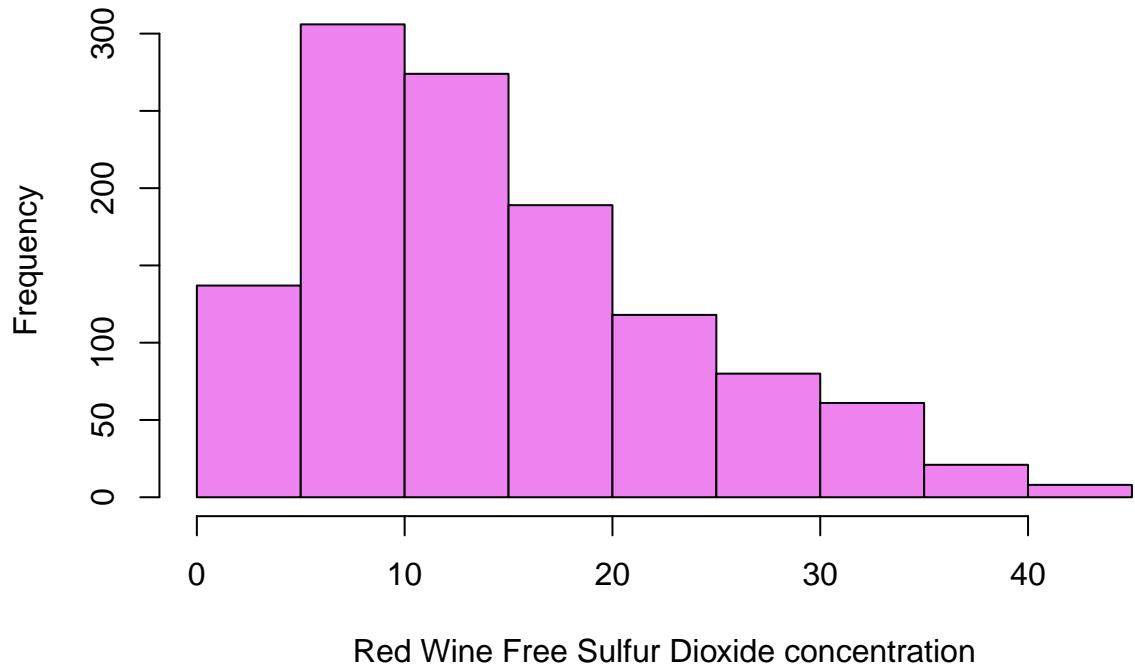
```
hist(eliminated$chlorides, main = "Red Wine Chloride distribution", xlab="Red Wine Chloride concentration")
```

## Red Wine Chloride distribution



```
# Histograma de la variable Free Sulfur Dioxide
hist(eliminated$free.sulfur.dioxide, main = "Red Wine Free Sulfur Dioxide distribution", xlab="Red Wine
```

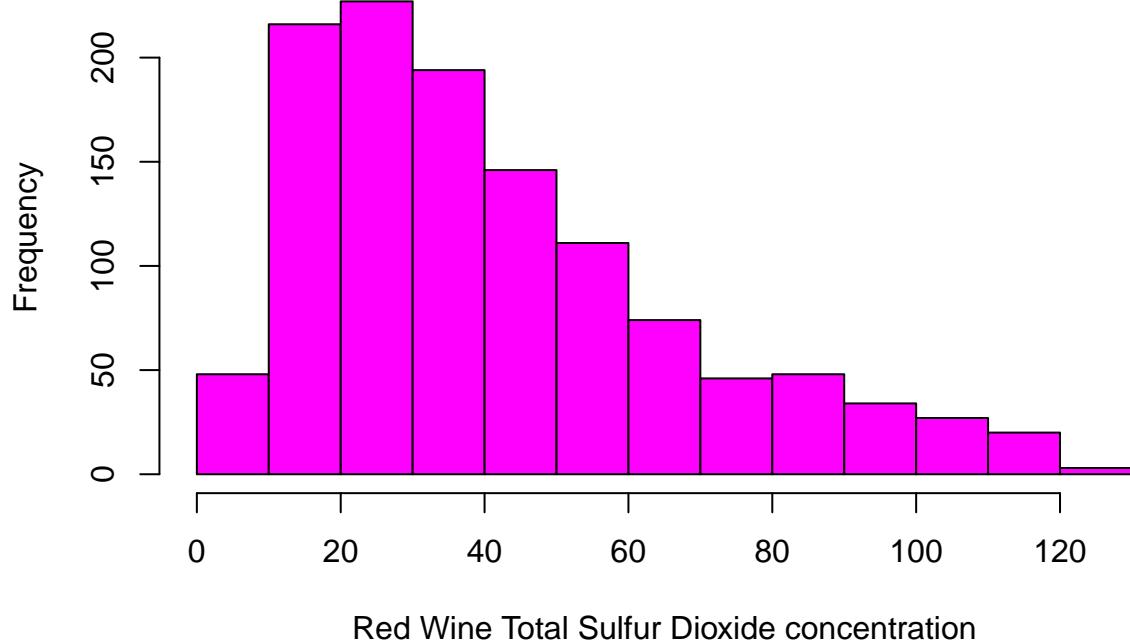
## Red Wine Free Sulfur Dioxide distribution



```
# Histograma de la variable Total Sulfur Dioxide
```

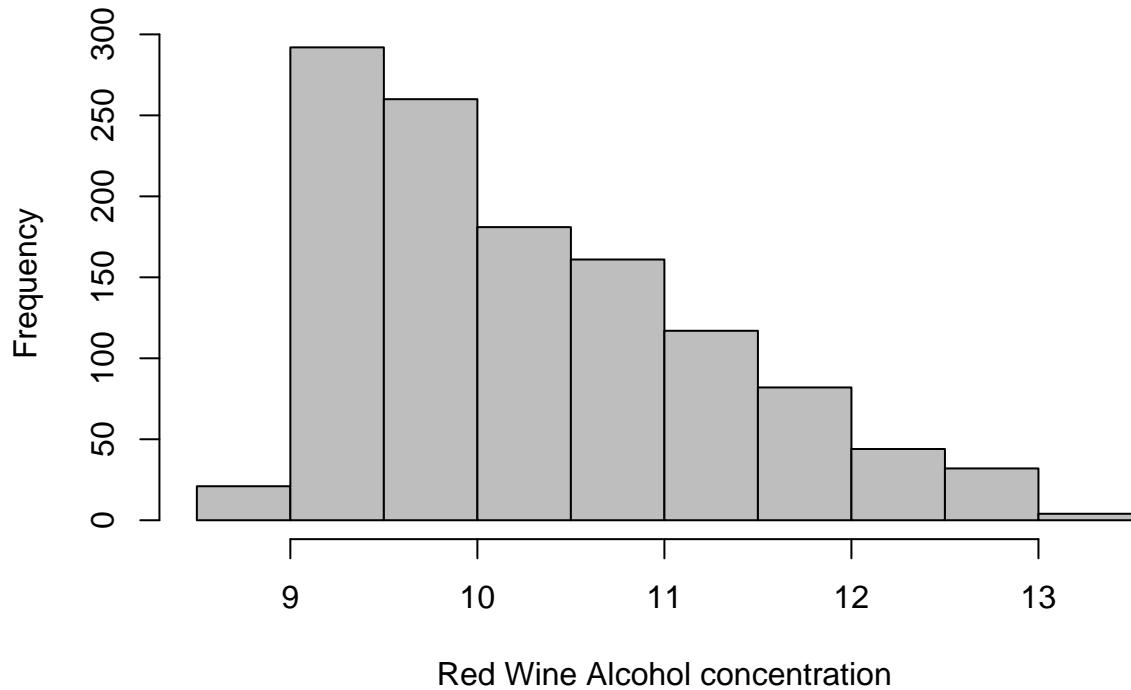
```
hist(eliminated$total.sulfur.dioxide, main = "Red Wine Total Sulfur Dioxide distribution", xlab="Red Wine Total Sulfur Dioxide")
```

## Red Wine Total Sulfur Dioxide distribution



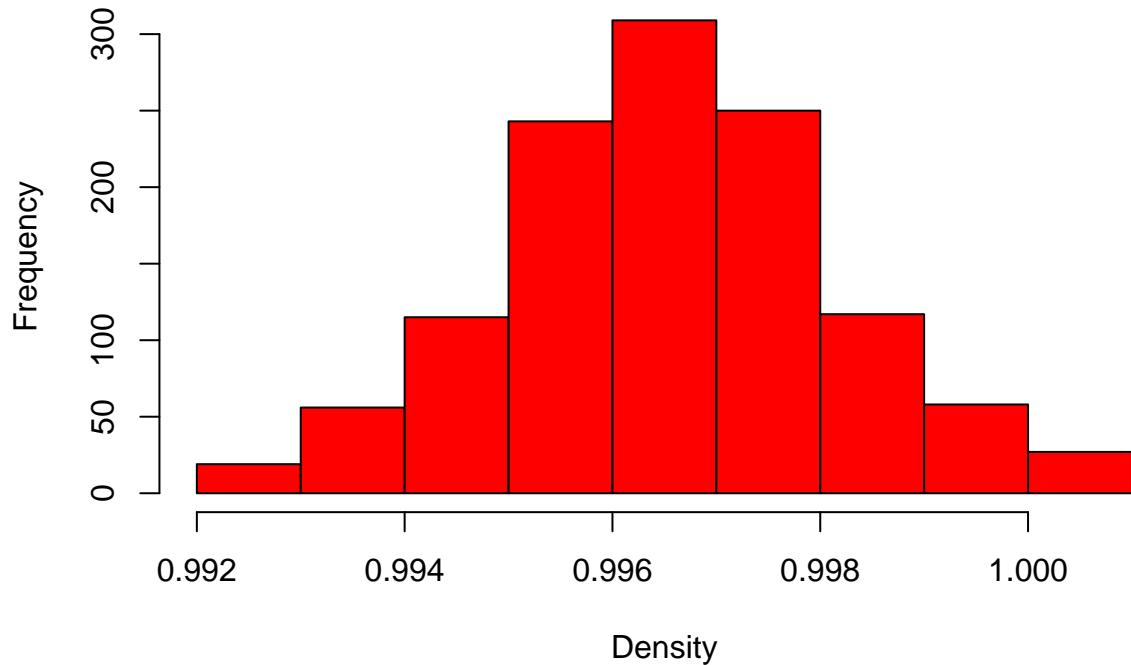
```
# Histograma de la variable Alcohol
hist(eliminated$alcohol, main = "Red Wine Alcohol distribution", xlab="Red Wine Alcohol concentration",
```

## Red Wine Alcohol distribution



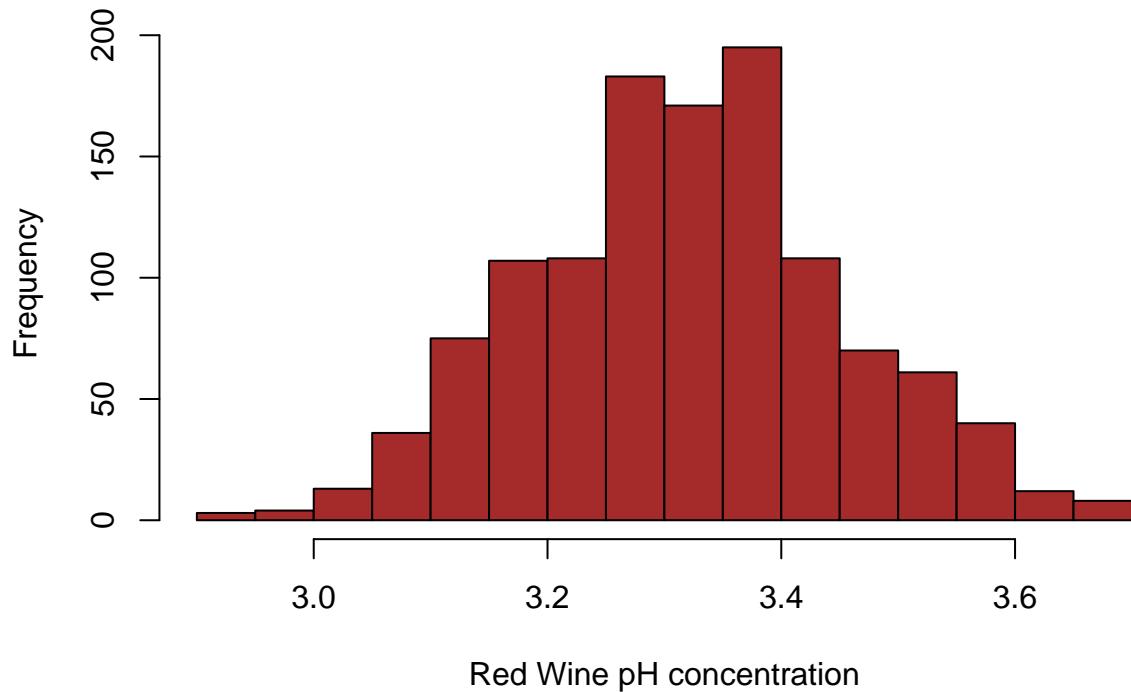
```
# Histograma de la variable Density
hist(eliminated$density, main = "Red Wine Density distribution", xlab="Density", col="Red")
```

## Red Wine Density distribution



```
# Histograma de la variable pH
hist(eliminated$pH, main = "Red Wine pH distribution", xlab="Red Wine pH concentration", col="Brown")
```

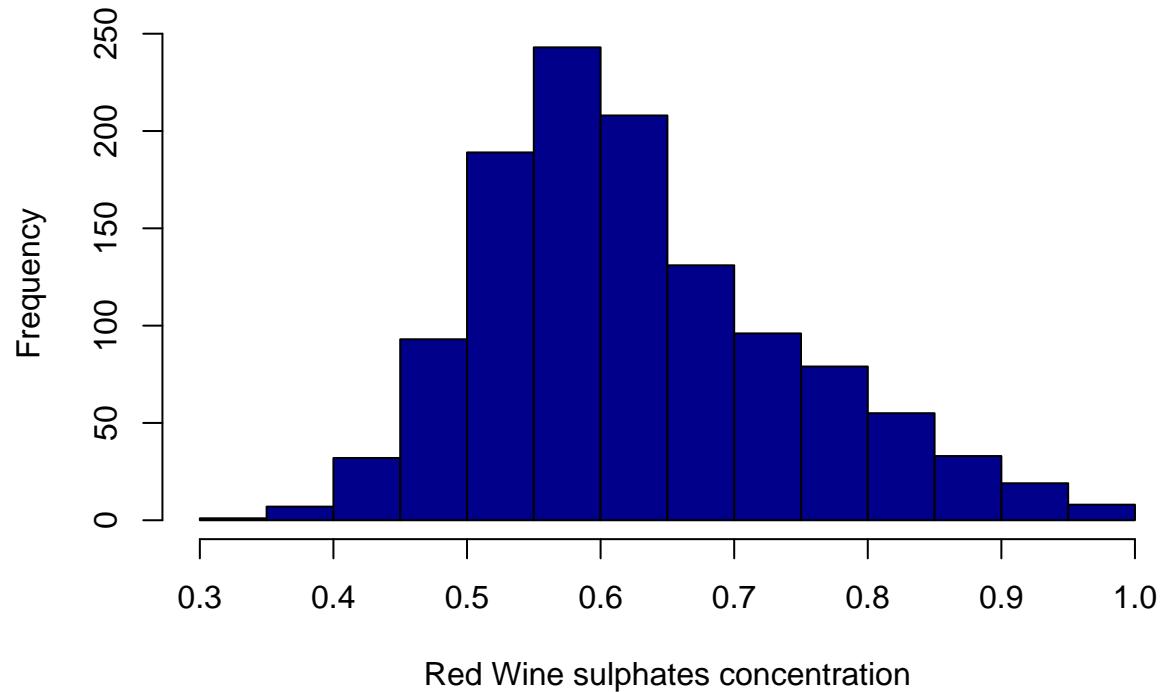
## Red Wine pH distribution



```
# Histograma de la variable Sulphates
```

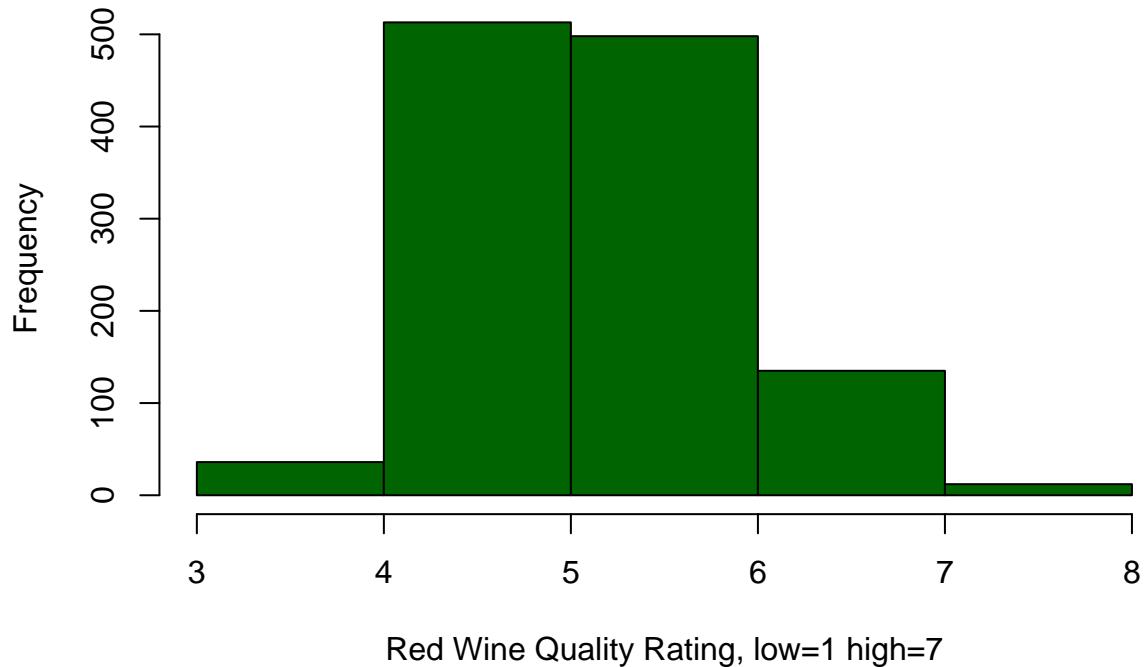
```
hist(eliminated$sulphates, main = "Red Wine sulphates distribution", xlab="Red Wine sulphates concentrat
```

## Red Wine sulphates distribution



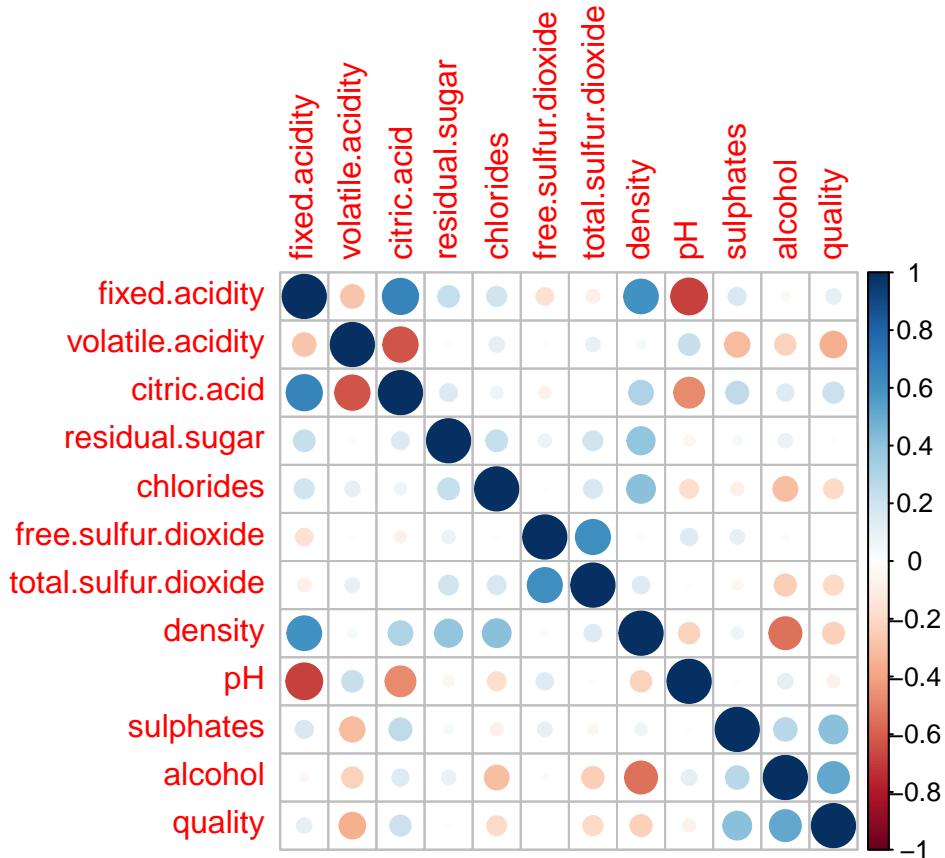
```
# Histograma de la variable quality
hist(eliminated$quality, breaks=6, col="Dark Green", xlab="Red Wine Quality Rating, low=1 high=7", main=
```

## Red Wine Quality distribution



## Correlaciones

```
corrplot(cor(eliminated), method = "circle")
```



Podemos ver que la calidad de los vinos está relacionada en gran medida con los parámetros: volatile.acidity, citric.acid, sulphates y alcohol. Estas correlaciones se pueden verificar mediante las pruebas de correlaciones de la siguiente manera:

```
cor.test(eliminated$volatile.acidity, eliminated$quality)
```

```
##
## Pearson's product-moment correlation
##
## data: eliminated$volatile.acidity and eliminated$quality
## t = -12.993, df = 1192, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.4009349 -0.3015075
## sample estimates:
##       cor
## -0.3522146
```

```
cor.test(eliminated$citric.acid, eliminated$quality)
```

```
##
## Pearson's product-moment correlation
##
## data: eliminated$citric.acid and eliminated$quality
## t = 7.7825, df = 1192, p-value = 1.534e-14
```

```

## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1652268 0.2732205
## sample estimates:
##       cor
## 0.2198973

cor.test(eliminated$sulphates, eliminated$quality)

## 
## Pearson's product-moment correlation
##
## data: eliminated$sulphates and eliminated$quality
## t = 15.764, df = 1192, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3672787 0.4612195
## sample estimates:
##       cor
## 0.4153559

cor.test(eliminated$alcohol, eliminated$quality)

## 
## Pearson's product-moment correlation
##
## data: eliminated$alcohol and eliminated$quality
## t = 20.545, df = 1192, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4682257 0.5520887
## sample estimates:
##       cor
## 0.5113737

library(nortest)
alpha = 0.05
col.names = colnames(eliminated)
for (i in 1:ncol(eliminated)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(eliminated[,i]) | is.numeric(eliminated[,i])) {
    p_val = ad.test(eliminated[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(eliminated) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}

## Variables que no siguen una distribución normal:

```

```
## fixed.acidity, volatile.acidity, citric.acid,  
## residual.sugar, chlorides, free.sulfur.dioxide,  
## total.sulfur.dioxide, density, pH,  
## sulphates, alcoholquality
```