

Práctica 2: Limpieza y validación de los datos

Waziri Ajibola Lawal, David Fernández González

1/3/2021

Detalles de la actividad

Descripción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
 - Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
 - Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
 - Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
 - Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
 - Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
 - Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.
-

Realización de la práctica

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos que se va a analizar es el de Red Wine Quality y se ha obtenido a partir de este enlace en Kaggle (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>). El conjunto de datos de vino tinto contiene 1599 observaciones, 11 predictores y 1 valor categórico que indica la calidad del vino. Entre los campos de este conjunto de datos, encontramos los siguientes:

- fixed acidity: la mayoría de los ácidos involucrados con el vino, o fijos o no volátiles (no se evaporan fácilmente).
- volatile acidity: cantidad de ácido acético en el vino.
- citric acid: cantidad de ácido cítrico en el vino.
- residual sugar: cantidad de azúcar residual en el vino.
- chlorides: cantidad de sal de potasio en el vino.
- free sulfur dioxide: El SO₂ existe en equilibrio, demasiado afectará la salud.
- total sulfur dioxide: cantidad de total de SO₂ en el vino.
- density: la densidad del vino se acerca a la del agua dependiendo del porcentaje de alcohol y del contenido de azúcar.
- pH: describe qué tan ácido o básico es un vino.
- sulphates: un aditivo para el vino que puede contribuir a los niveles de dióxido de azufre (SO₂), que actúa como antimicrobiano y antioxidante.
- alcohol: el porcentaje de contenido de alcohol del vino.
- quality: valor que describe la calidad del vino (basada en datos sensoriales, puntuación entre 0 y 10).

El objetivo principal es encontrar qué variables ofrecen más información sobre la calidad del vino. También intentaremos hacer predicciones de la calidad de un vino, y comprobar si se corresponde con su calidad real.

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

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

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

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

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

# 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, verificamos algunas métricas sobre las variables

```
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
```

```
summary(redWineData$quality)
```

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

Hay 1599 observaciones de 12 variables numéricas.

Quantity es una variable categórica y discreta, con una escala de 0 a 10. Los valores varían sólo de 3 a 8, con una media de 5,6 y una mediana de 6. Todas las demás variables parecen ser cantidades continuas (con la excepción de los sufijos .sulfur.dioxide).

Todos los predictores son valores numéricos, los resultados son enteros.

3. Limpieza de los datos.

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

Como podemos observar, no existen valores vacíos en nuestro conjunto de datos.

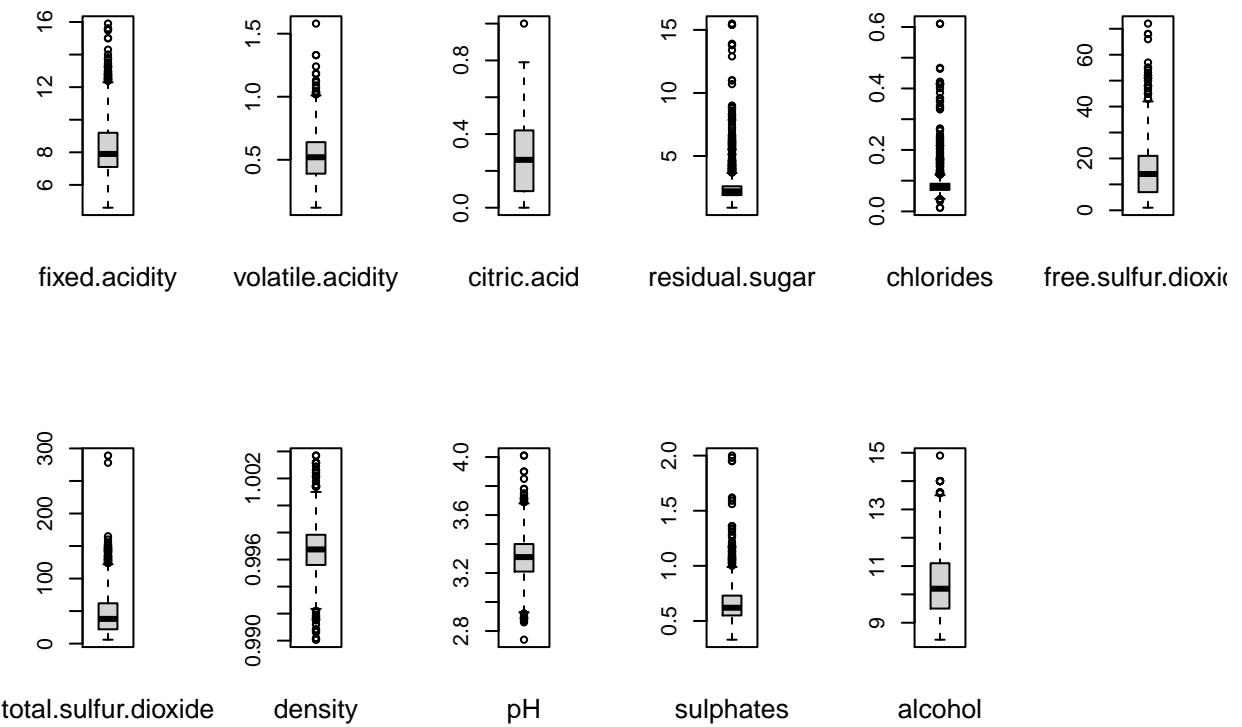
```
# Verificamos si existen valores vacíos en el conjunto de datos  
colSums(is.na(redWineData))
```

```
##      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
```

3.2. Identificación y tratamiento de valores extremos.

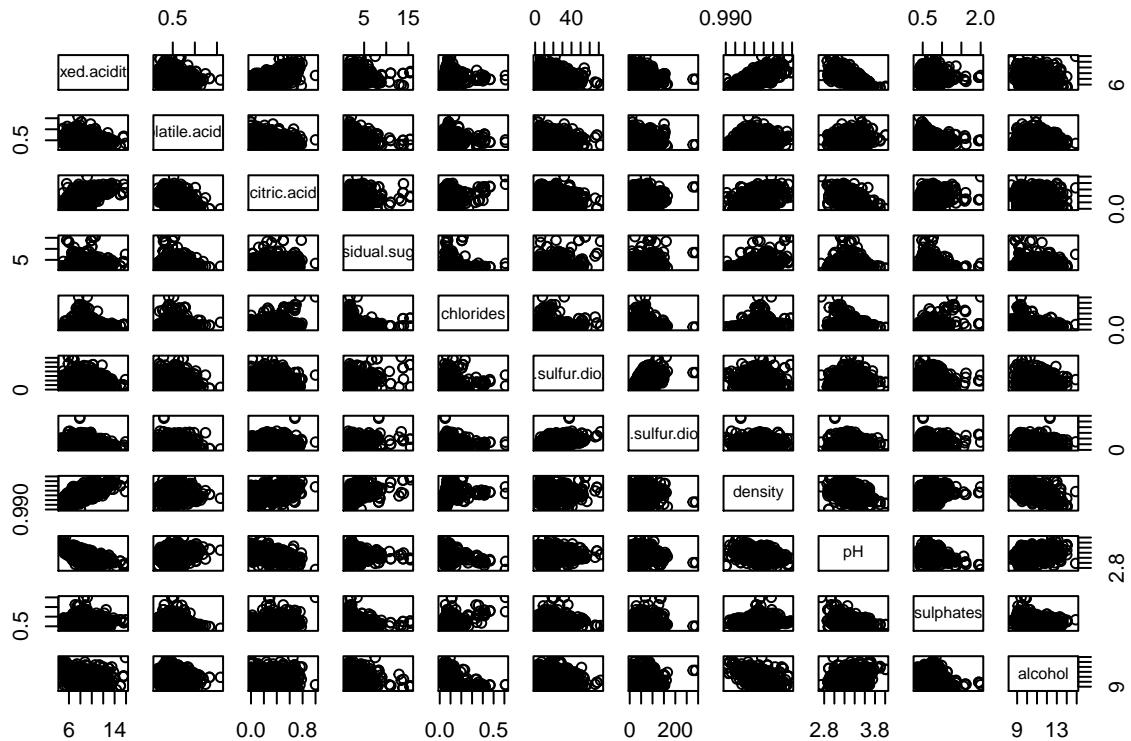
Las métricas del conjunto de datos nos muestran que la mayoría de las variables tienen un rango amplio en comparación con el rango intercuartil, lo que puede indicar una dispersión en los datos y la presencia de valores atípicos. Investigamos más a fondo produciendo diagramas de caja para cada una de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



Para obtener mas información sobre la posición de los valores atípicos, podemos usar la función pairs() con la que obtendremos una matriz de gráfico de dispersión.

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



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

```

eliminated_outliers <- 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_outliers <- subset(redWineData, redWineData[[i]] > (Q[1] - 1.5*iqr) & redWineData[[i]] <
  # ggbetweenstats(eliminated_outliers, quality, redWineData[[i]], outlier.tagging = TRUE)
  boxplot(redWineData[[i]], plot = FALSE)$out
  outliers <- boxplot(redWineData[[i]], plot = FALSE)$out
  eliminated_outliers <- eliminated_outliers[-which(eliminated_outliers[[i]] %in% outliers), ]
}

dim(eliminated_outliers)

## [1] 1194   12

```

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

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

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

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

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

Tabla de contribuciones al trabajo

```
contribuciones <- matrix(c("Investigación previa","WAjibolaL,DFdezGlez","Redacción de las respuestas","WAjibolaL,DFdezGlez"), nrow=1, byrow=TRUE)
colnames(contribuciones) <- c("Contribuciones","Firmas")
rownames(contribuciones) <- c("", "", "")
contribuciones <- as.table(contribuciones)
contribuciones

## Contribuciones           Firmas
## Investigación previa   WAjibolaL,DFdezGlez
## Redacción de las respuestas WAjibolaL,DFdezGlez
## Desarrollo código       WAjibolaL,DFdezGlez
```

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="Grey")
#
# # Histograma de la variable Volatile Acidity
# hist(volatile.acidity, main = "Red Wine volatile Acidity distribution", xlab="Red Wine Volatile Acidity concentration", col="Dark Green")
#
# # Histograma de la variable Citric Acid
# hist(citric.acid, main = "Red Wine Citric acid distribution", xlab="Red Wine Citric Acid concentration", col="Blue")
#
# # Histograma de la variable Residual Sugar
# hist(residual.sugar, main = "Red Wine Residual Sugar distribution", xlab="Red Wine Residual Sugar concentration", col="Red")
#
# # Histograma de la variable Chlorides
# hist(chlorides, main = "Red Wine Chloride distribution", xlab="Red Wine Chloride concentration", col="Green")
#
# # Histograma de la variable Free Sulfur Dioxide
# hist(free.sulfur.dioxide, main = "Red Wine Free Sulfur Dioxide distribution", xlab="Red Wine Free Sulfur Dioxide concentration", col="Brown")
#
# # Histograma de la variable Total Sulfur Dioxide
# hist(total.sulfur.dioxide, main = "Red Wine Total Sulfur Dioxide distribution", xlab="Red Wine Total Sulfur Dioxide concentration", col="Dark Green")
#
# # Histograma de la variable Alcohol
# hist(alcohol, main = "Red Wine Alcohol distribution", xlab="Red Wine Alcohol concentration", col="Grey")
#
# # Histograma de la variable Density
# hist(density, main = "Red Wine Density distribution", xlab="Density", col="Red")
#
# # Histograma de la variable pH
# hist(pH, main = "Red Wine pH distribution", xlab="Red Wine pH concentration", col="Brown")
#
# # Histograma de la variable Sulphates
# hist(sulphates, main = "Red Wine sulphates distribution", xlab="Red Wine sulphates concentration", col="Dark Green")
#
# # Histograma de la variable quality
# hist(quality, breaks=6, col="Dark Green", xlab="Red Wine Quality Rating, low=1 high=7", main="Red Wine Quality Rating")
```



```
for (i in 1:12) {
```

```

## Have a look at the densities
# plot(density(redWineData[[i]]), main = names(redWineData)[i], xlab = "");

## Perform the test
# shapiro.test(redWineData[[i]]);

## Plot using a qqplot
# qqnorm(redWineData[[i]]); qqline(redWineData[[i]], col = 2)
}

```

Podemos observar que casi todas las distribuciones están sesgadas positivamente. Las variables pH, density y quality tienen una distribución aproximadamente normal.

```

# # Histograma de la variable Fixed Acidity
# hist(eliminated_outliers$fixed.acidity, main = "Red Wine fixed acidity", xlab="Red Wine Fixed Acidity"
#
# # Histograma de la variable Volatile Acidity
# hist(eliminated_outliers$volatile.acidity, main = "Red Wine volatile Acidity distribution", xlab="Red
#
# # Histograma de la variable Citric Acid
# hist(eliminated_outliers$citric.acid, main = "Red Wine Citric acid distribution", xlab="Red Wine Citr
#
# # Histograma de la variable Residual Sugar
# hist(eliminated_outliers$residual.sugar, main = "Red Wine Residual Sugar distribution", xlab="Red Win
#
# # Histograma de la variable Chlorides
# hist(eliminated_outliers$chlorides, main = "Red Wine Chloride distribution", xlab="Red Wine Chloride
#
# # Histograma de la variable Free Sulfur Dioxide
# hist(eliminated_outliers$free.sulfur.dioxide, main = "Red Wine Free Sulfur Dioxide distribution", xla
#
# # Histograma de la variable Total Sulfur Dioxide
# hist(eliminated_outliers$total.sulfur.dioxide, main = "Red Wine Total Sulfur Dioxide distribution", x
#
# # Histograma de la variable Alcohol
# hist(eliminated_outliers$alcohol, main = "Red Wine Alcohol distribution", xlab="Red Wine Alcohol conc
#
# # Histograma de la variable Density
# hist(eliminated_outliers$density, main = "Red Wine Density distribution", xlab="Density", col="Red")
#
# # Histograma de la variable pH
# hist(eliminated_outliers$pH, main = "Red Wine pH distribution", xlab="Red Wine pH concentration", col=
#
# # Histograma de la variable Sulphates
# hist(eliminated_outliers$sulphates, main = "Red Wine sulphates distribution", xlab="Red Wine sulphate
#
# # Histograma de la variable quality
# hist(eliminated_outliers$quality, breaks=6, col="Dark Green",xlab="Red Wine Quality Rating, low=1 high
#
for (i in 1:12) {
  ## Have a look at the densities
  # plot(density(eliminated_outliers[[i]]), main = names(eliminated_outliers)[i]);

```

```

## Perform the test
# shapiro.test(eliminated_outliers[[i]]);

## Plot using a qqplot
# qqnorm(eliminated_outliers[[i]]);qqline(eliminated_outliers[[i]], col = 2)
}

```

Exportación de los datos preprocesados

```

# Exportación de los datos limpios en .csv
write.csv(eliminated_outliers, "winequality-red_data_clean.csv")

```

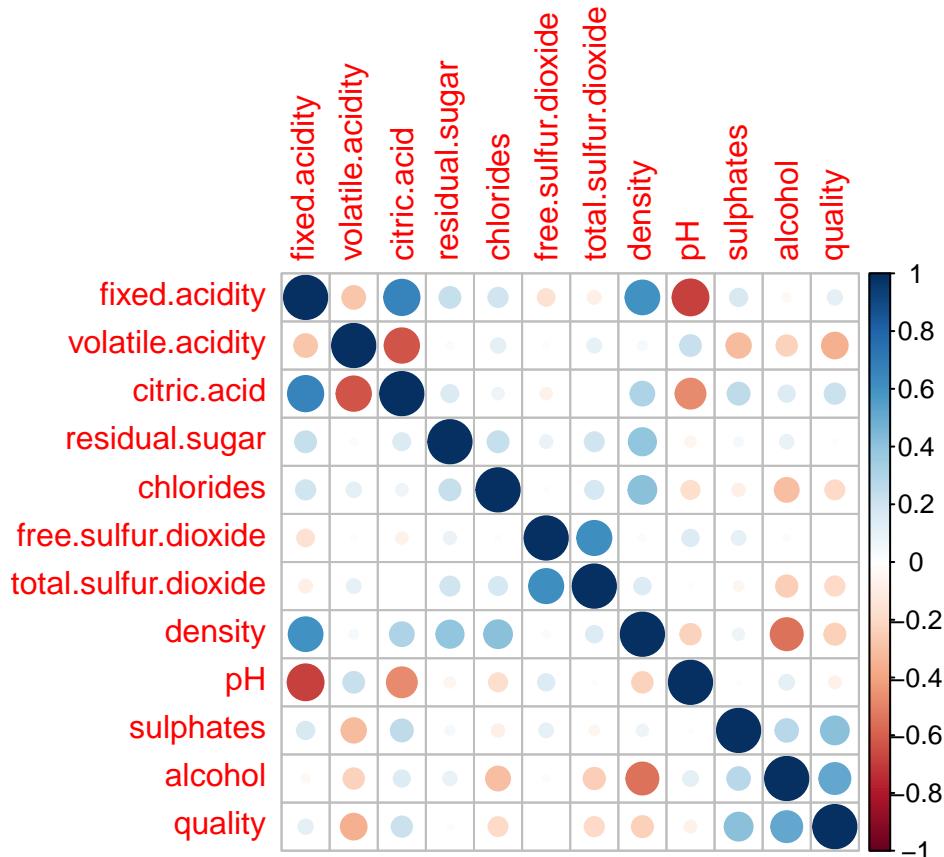
Pruebas estadísticas

Correlaciones

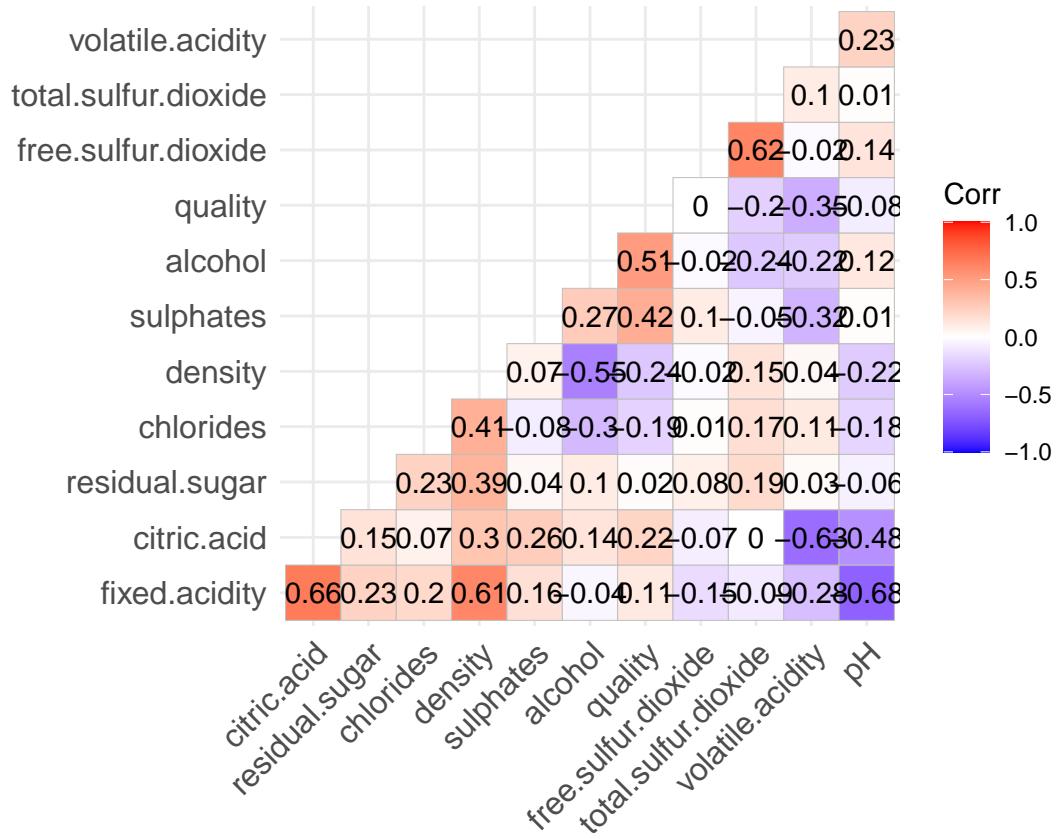
Mediante la matriz de correlación, podemos ver que la calidad de los vinos está relacionada en gran medida con las variables sulphates y alcohol.

Tambien podemos observar las correlaciones entre las variables fixed.acidity - density (correlación fuerte), fixed.acidity - pH (correlación fuerte), fixed.acidity - citric.acid, volatlie.acidity - citric.acid y total.sulfur.dioxide - free.sulfur.dioxide.

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



```
ggcorrplot(cor(eliminated_outliers), hc.order = TRUE, type = "lower", lab = TRUE, insig = "blank")
```



Estas correlaciones se pueden verificar mediante las pruebas de correlaciones de la siguiente manera:

```
# cor.test(eliminated_outliers$sulphates, eliminated_outliers$quality)

# cor.test(eliminated_outliers$alcohol, eliminated_outliers$quality)

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

```
cor(x=eliminated_outliers[1:11], y=eliminated_outliers$quality)
```

```
## [1]
## fixed.acidity      0.11067147
## volatile.acidity   -0.35221458
## citric.acid        0.21989725
## residual.sugar     0.01543850
## chlorides          -0.19227105
## free.sulfur.dioxide -0.00278825
## total.sulfur.dioxide -0.19853858
## density            -0.23633307
## pH                 -0.07572362
## sulphates          0.41535589
## alcohol             0.51137367
```

Podemos ver que las variables están correlacionadas con quality de la siguiente manera From previous we see that the following variables are correlated with quality:

- alcohol (+++)
- sulphates (+++)
- volatile.acidity (-)
- citric.acid (++)
- fixed.acidity (+)
- total.sulfur.dioxide (-)
- density (-)
- chlorides (-)