PRAC2. Limpieza y validación de los datos

Mónica Arrúe Gabaráin

Ejercicio 1

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

Este dataset contiene métricas tomadas a diferentes muestras del vino portugués “Vinho Verde”. De cada una de las muestras se tiene la siguiente información:

- fixed acidity: cantidad de acidez fija del vino
- volatile acidity: cantidad de acidez volatil del vino
- citric acid: cantidad de ácido cítrico
- residual sugar: cantidad de azúcar encontrado cuando la fermentación termina
- chlorides: cantidad de cloruros
- free sulfur dioxide: cantidad de SO2 libre
- total sulfur dioxide: cantidad total de SO2 libre y unido
- density: densidad
- pH: pH del vino (cuan ácido o básico es)
- sulphates: cantidad de sulfatos
- alcohol: porcentaje de alcohol
- quality: calidad del vino. Puede tomar un valor entre 0 y 10.

Este dataset se ha extraído del siguiente link: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Como se puede apreciar, con la información que se dispone y llevando a cabo las analíticas adecuadas, se puede extraer información de gran interés para los fabricantes del vino con la finalidad de mejorar el producto que ofrecen, detectar posibles errores, etc. Mediante estas analíticas se pueden responder preguntas tales como:

- ¿Que sustancia o combinación se sustancias influye más en la calidad del vino?
- ¿Hay algún vino que no se encuentre en buen estado?
- ¿Existe alguna relación entre las sustancias del vino?

Ejercicio 2

Enunciado: Integración y selección de los datos de interés a analizar

En primer lugar, vamos a cargar el dataset original y a mostrar un resumen de los datos para familiarizarnos con ellos y detectar qué variables son necesarias y cuáles no para las posteriores analíticas.

```
# Cargamos los datos
data <- read.csv(file = "winequality-red.csv", header = TRUE, sep = ",")

# Mostramos un resumen de los datos
summary(data)
```

```
##  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
##  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
##     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 :0.9968   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
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

Analizando la tipología de los datos y la información que contiene cada uno de los atributos, se ha decidido que se van a utilizar todos los atributos para llevar a cabo las analíticas.

Ejercicio 3

Enunciado: Limpieza de los datos

Ejercicio 3.1

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

```
# Contamos el número de ceros (si los hay) por columnas
sapply(data, function(x) sum(x==0))
```

```
##        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
```

```
# Contamos el número de elementos vacíos (si los hay) por columnas
sapply(data, 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
```

Como se puede apreciar, este dataset contiene 132 ceros en el atributo “citric.acid”. Analizando la descripción de la variable concluimos que son valores válidos ya que una muestra de vino puede no tener nada de ácido cítrico y que, por lo tanto, ese atributo tome un valor de cero sin deberse a ningún error.

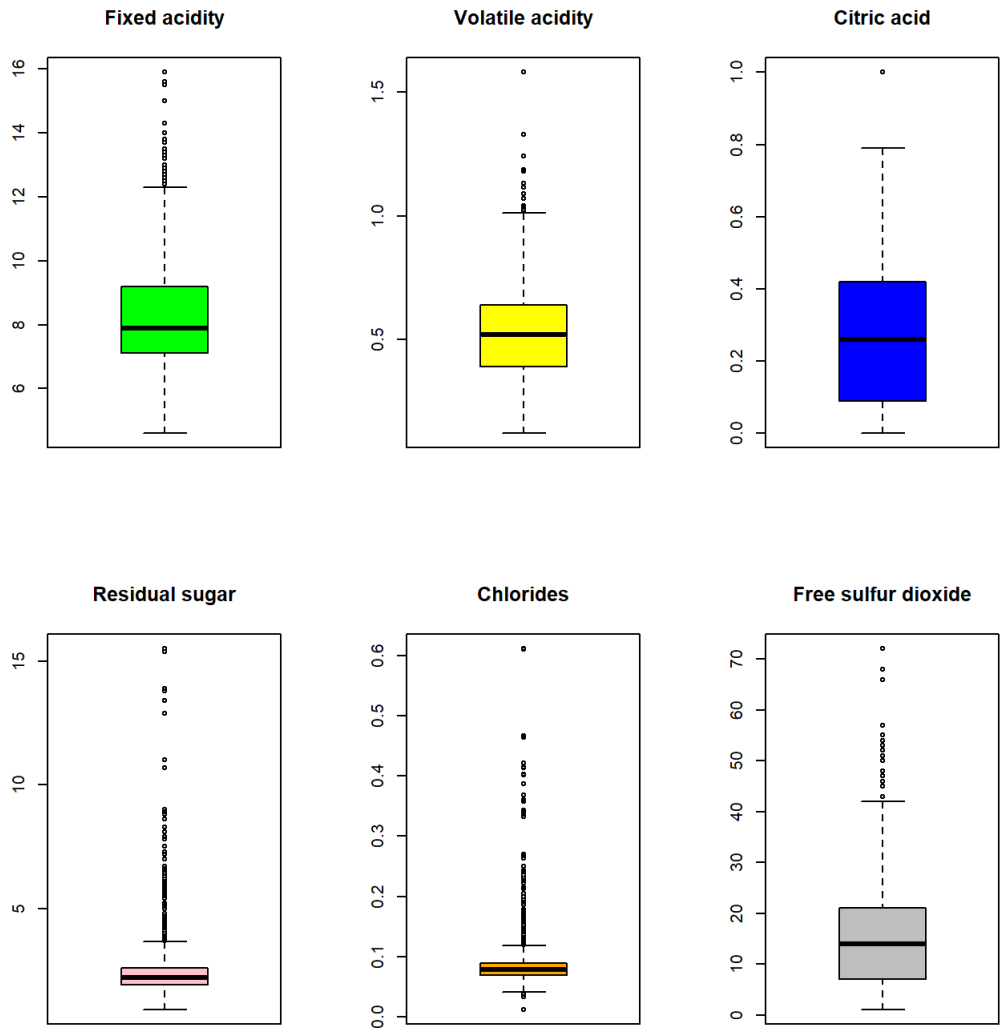
En cuanto a los elementos vacíos, este dataset no tiene ninguno. En caso de que hubiese elementos vacíos, dependería de cuantos hubiese y en cuantos atributos. En el caso de que hubiese muy pocos valores null y en pocos atributos, la opción más adecuada sería ignorar las filas que tuviesen algun valor nulo. Sin embargo, en el caso de que hubiese un cantidad considerable de valores vacíos y en bastantes atributos, si se tomase la medida anterior, nos quedaríamos prácticamente sin datos para llevar a cabo las analíticas posteriores. Por lo que en este caso la mejor opción sería completar estos valores vacíos. Estos valores vacios no se pueden rellenar con cualquier valor ya que esto podría distorsionar el resultado obtenido en las analíticas, especialmente si la cantidad de datos vacíos es grande. Por lo tanto, en este caso, si el valor vacío se encontrase en alguna de las métricas del vino, añadiría como sustituto del dato vacío la media de esa métrica para los vinos de la misma calidad del vino en cuestión. Sin embargo, si el dato vacío se encontrase en la columna de calidad del vino, estudiaría qué calidad de vino puede tener un vino con esas métricas mediante técnicas de machine learning y le añadiría la calidad correspondiente.

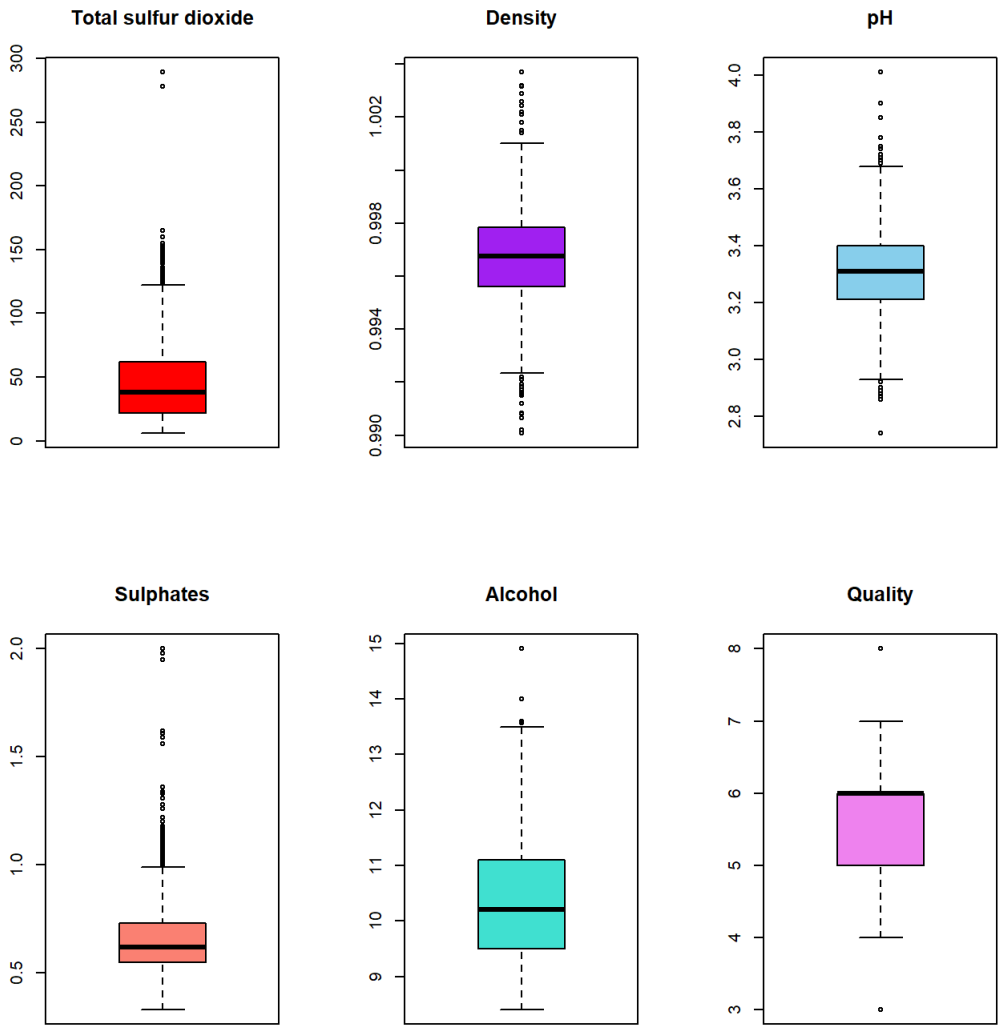
Ejercicio 3.2

Enunciado: Identificación y tratamiento de valores extremos.

Generalmente, se considera que un valor es extremo si se encuentra fuera del rango [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] donde IQR es el rango intercuartílico (Q3 - Q1). Para detectarlos, vamos a mostrar en primer lugar un boxplot por cada una de las variables numéricas. Este tipo de gráfica representa los valores extremos (según la definición anterior) mediante unos puntos por encima y por debajo de la los valores extremos que toma la variable (representados mediante una linea horizontal arriba y abajo de la caja).

```
par(mfrow=c(4,3))
boxplot(data$fixed.acidity, main = "Fixed acidity", col = "green")
boxplot(data$volatile.acidity, main = "Volatile acidity", col = "yellow")
boxplot(data$citric.acid, main = "Citric acid", col = "blue")
boxplot(data$residual.sugar, main = "Residual sugar", col = "pink")
boxplot(data$chlorides, main = "Chlorides", col = "orange")
boxplot(data$free.sulfur.dioxide, main = "Free sulfur dioxide", col = "grey")
boxplot(data$total.sulfur.dioxide, main = "Total sulfur dioxide", col = "red")
boxplot(data$density, main = "Density", col = "purple")
boxplot(data$pH, main = "pH", col = "skyblue")
boxplot(data$sulphates, main = "Sulphates", col = "salmon")
boxplot(data$alcohol, main = "Alcohol", col = "turquoise")
boxplot(data$quality, main = "Quality", col = "violet")
```





Como se puede apreciar en los boxplots, se identifican valores extremos en todos los atributos. Analizando los atributos uno por uno, podemos apreciar que en varios casos los valores extremos se encuentran muy cerca del valor máximo o mínimo que toma el atributo y no solo hay uno, sino que existen muchos valores muy seguidos. Teniendo en cuenta las definiciones de cada atributo, aunque se hayan detectado estos valores como outliers, una muestra de vino puede tener perfectamente esos valores. Además, según es indica en la fuente de origen, este dataset ha sido limpiado antes de subirlo a la plataforma, por lo que dudo que los outliers detectados se deban a errores. El único caso en el que es probable que se deba a un outlier es en el caso del atributo “total.sulfur.dioxide”, ya que hay un par de valores que se separan mucho del resto. Finalmente, en el caso del atributo “quality”, podemos ver que detecta dos outliers, uno que toma el valor de 3 y otro el valor de 8. Aunque los detecta como outliers en realidad no lo son, ya que es un parámetro que mide la calidad del vino mediante un rango entre 0 y 10 y es perfectamente posible que existan dos muestras de vinos con esa calificación.

A continuación, vamos a proceder a eliminar los valores extremos del atributo “total.sulfur.dioxide” (puesto que son dos muestras únicamente optamos por eliminarlas). Observando el boxplot podemos ver que son dos valores que se encuentran entre los valores 250 y 300. En primer lugar vamos a mostrar el listado de los outliers detectados para, a continuación, eliminar las muestras que contienen los dos valores mayores.

```
boxplot.stats(data$total.sulfur.dioxide)$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
```

Los valores se los outliers que hemos detectado son “278” y “289”, por lo que vamos a eliminar esas muestras.

```
data_no_outliers <- data[!(data$total.sulfur.dioxide==278 | data$total.sulfur.dioxide==289),]
```

Numero de filas del dataset original: 1599

Numero de filas del dataset sin outliers: 1597

Ejercicio 4

Ejercicio 4.1

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

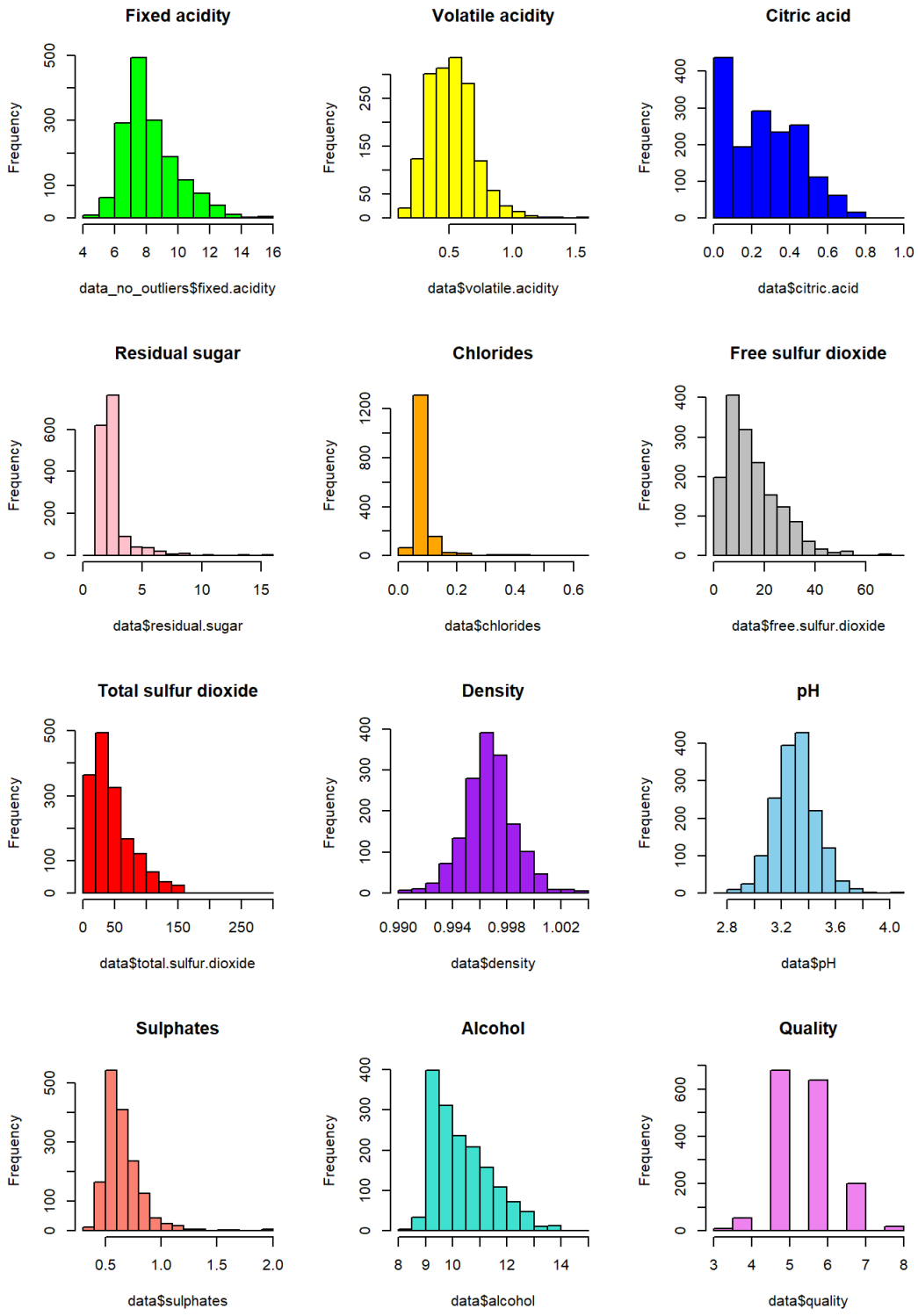
Dado los datos que disponemos, una analítica interesante de llevar a cabo es estudiar cuáles son las cualidades que más afectan en la calidad del vino. Para llevar a cabo esta analítica se van a utilizar todas las variables del dataset.

Ejercicio 4.2

Enunciado: Comprobación de la normalidad y homogeneidad de la varianza

En primer lugar vamos a comprobar la normalidad de los datos. Para ello, vamos a visualizar las variables mediante histogramas para hacernos una idea de la forma que toman para a continuación aplicar el test de Shapiro Wilk a cada una de las variables para analizar si están normalizadas o no.

```
par(mfrow=c(4,3))
hist(data_no_outliers$fixed.acidity, main = "Fixed acidity", col = "green")
hist(data$volatile.acidity, main = "Volatile acidity", col = "yellow")
hist(data$citric.acid, main = "Citric acid", col = "blue")
hist(data$residual.sugar, main = "Residual sugar", col = "pink")
hist(data$chlorides, main = "Chlorides", col = "orange")
hist(data$free.sulfur.dioxide, main = "Free sulfur dioxide", col = "grey")
hist(data$total.sulfur.dioxide, main = "Total sulfur dioxide", col = "red")
hist(data$density, main = "Density", col = "purple")
hist(data$pH, main = "pH", col = "skyblue")
hist(data$sulphates, main = "Sulphates", col = "salmon")
hist(data$alcohol, main = "Alcohol", col = "turquoise")
hist(data$quality, main = "Quality", col = "violet")
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$fixed.acidity
## W = 0.94214, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$volatile.acidity
## W = 0.97444, p-value = 3.005e-16
```

```
shapiro.test(data_no_outliers$citric.acid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$citric.acid
## W = 0.95532, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$residual.sugar
## W = 0.56499, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$chlorides
## W = 0.48377, p-value < 2.2e-16
```

```
shapiro.test(data_no_outliers$free.sulfur.dioxide)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$free.sulfur.dioxide
## W = 0.90156, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$total.sulfur.dioxide
## W = 0.8901, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$density
## W = 0.99069, p-value = 1.491e-08
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$pH
## W = 0.99336, p-value = 1.362e-06
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$sulphates
## W = 0.83313, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$alcohol
## W = 0.92868, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data_no_outliers$quality
## W = 0.85728, p-value < 2.2e-16
```

Como podemos observar, ninguno de los test tiene un p-valor mayor que 0.05, por lo tanto concluímos que ninguna variable de las variables está normalizada. Sin embargo, siguiendo el teorema del límite central, como tenemos más de 30 muestras podemos aproximar estas variables como una distribución normal de media 0 y desviación estándar 1.

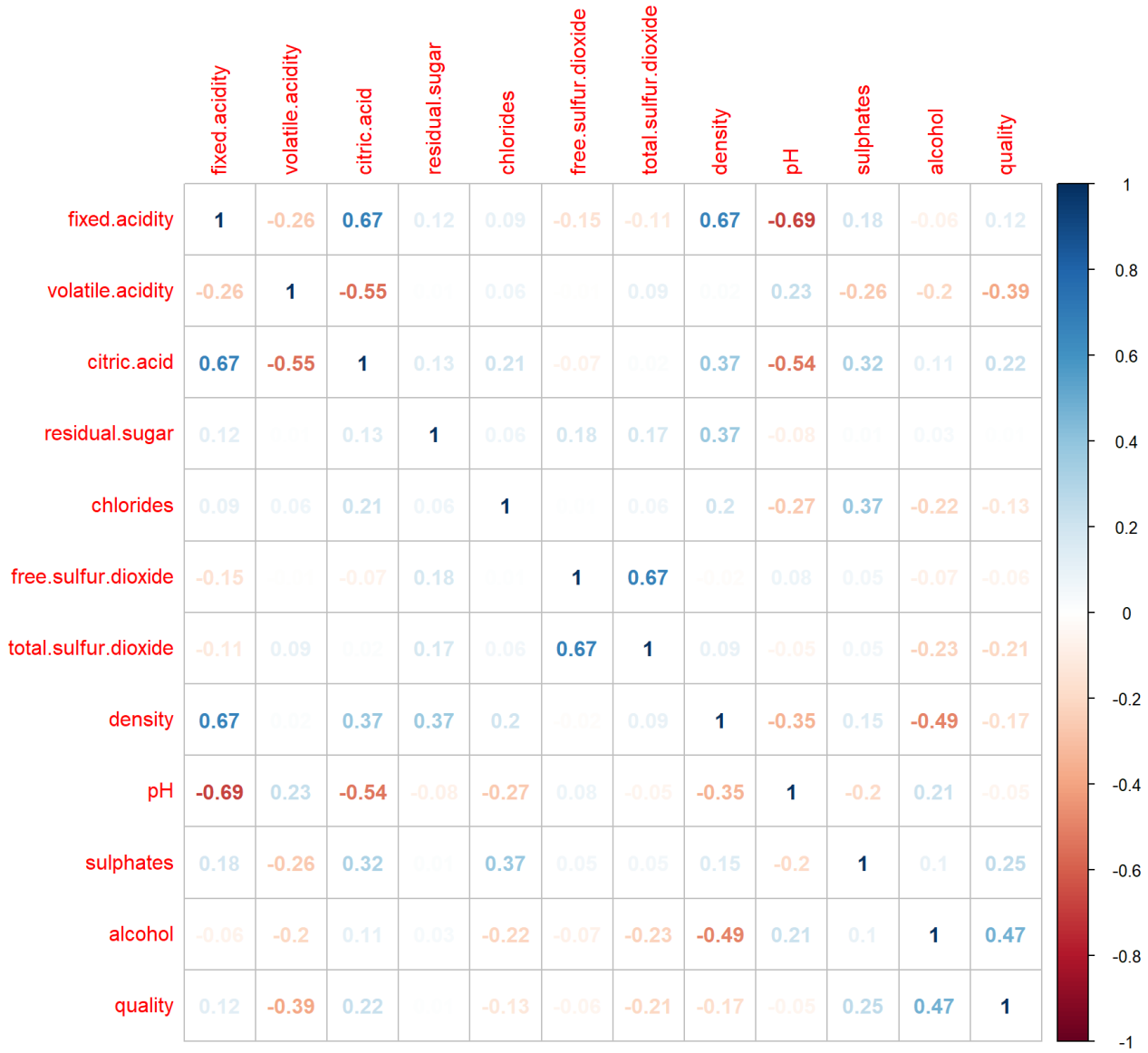
Puesto que tenemos una única muestra de la población no se va a realizar la prueba de la homogeneidad de varianzas.

Ejercicio 4.3

Enunciado: Aplicación de las pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo de estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

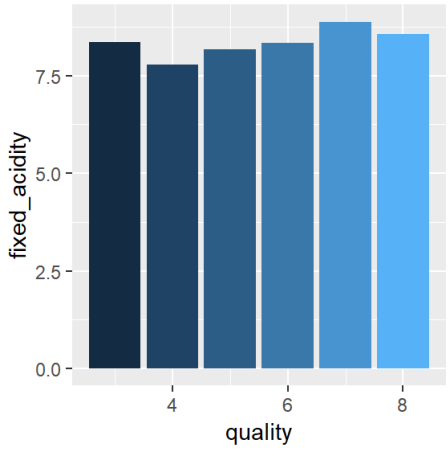
El objetivo principal de este estudio es analizar cuáles son las sustancias o características que más influyen a la calidad del vino y cómo. Para ello, en primer lugar vamos a crear una matriz de correlaciones que nos permita analizar la relación entre los atributos del dataset. A continuación, vamos a crear varias gráficas que comparen la calidad del vino con los atributos más correlacionados de tal forma que nos permitan analizar visualmente la relación entre estas variables. Finalmente, crearemos un modelo de regresión múltiple con los atributos que parezcan que más influencia tienen en la calidad del vino.

```
correlations <- cor(data_no_outliers)
corrplot(correlations, method = 'number')
```

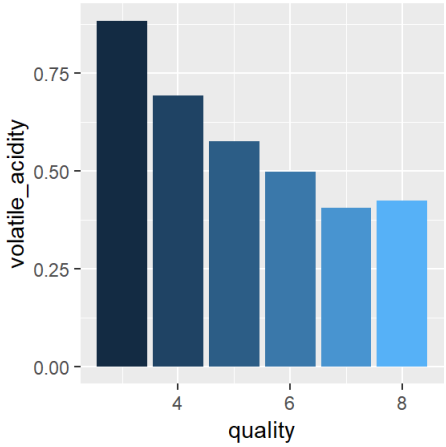


Como podemos apreciar, los atributos “residual.sugar”, “free.sulfur.dioxide” y “pH” son cláramente los que menos correlacionados se encuentran con el atributo de la calidad del vino. A continuación, vamos a crear histogramas que nos permitan analizar visualmente relación que tienen la calidad del vino y los atributos más correlacionados.

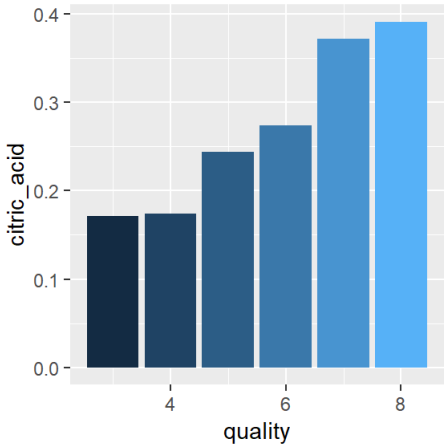
```
data_grouped_1 <- data_no_outliers %>% group_by(quality) %>% summarise(fixed_acidity = mean(fixed.acidity))
ggplot(aes(x = quality, y = fixed_acidity, fill = quality), data = data_grouped_1) + geom_bar(stat = "identi
ty") + theme(legend.position="none")
```



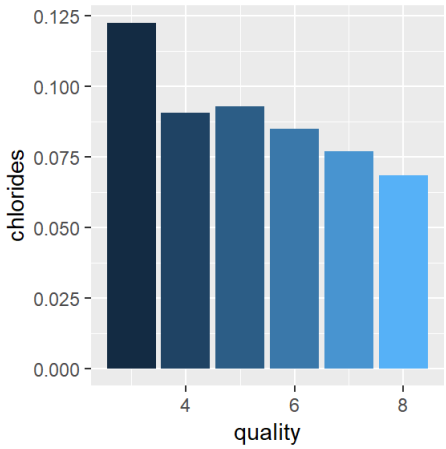
```
data_grouped_2 <- data_no_outliers %>% group_by(quality) %>% summarise(volatile_acidity = mean(volatile.acid
ity))
ggplot(aes(x = quality, y = volatile_acidity, fill = quality), data = data_grouped_2) + geom_bar(stat = "ide
ntity") + theme(legend.position="none")
```



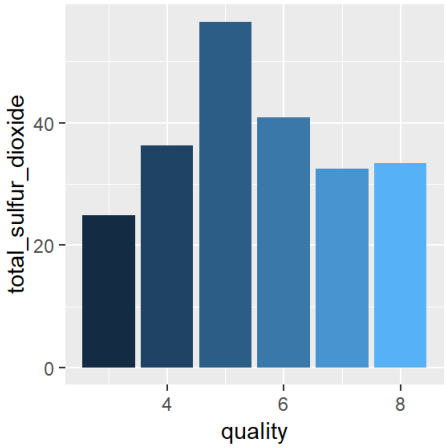

```
data_grouped_3 <- data_no_outliers %>% group_by(quality) %>% summarise(citric_acid = mean(citric.acid))
ggplot(aes(x = quality, y = citric_acid, fill = quality), data = data_grouped_3) + geom_bar(stat = "identity
") + theme(legend.position="none")
```



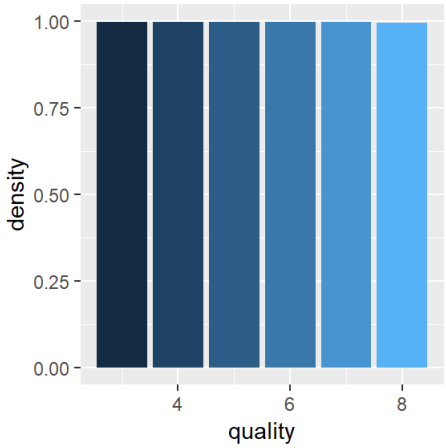

```
data_grouped_4 <- data_no_outliers %>% group_by(quality) %>% summarise(chlorides = mean(chlorides))
ggplot(aes(x = quality, y = chlorides, fill = quality), data = data_grouped_4) + geom_bar(stat = "identity")
+ theme(legend.position="none")
```



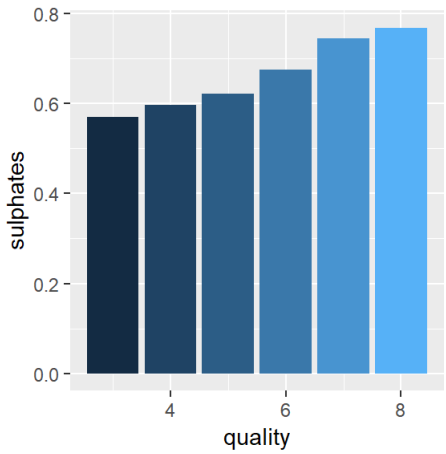

```
data_grouped_5 <- data_no_outliers %>% group_by(quality) %>% summarise(total_sulfur_dioxide = mean(total.sul
fur.dioxide))
ggplot(aes(x = quality, y = total_sulfur_dioxide, fill = quality), data = data_grouped_5) + geom_bar(stat =
"identity") + theme(legend.position="none")
```



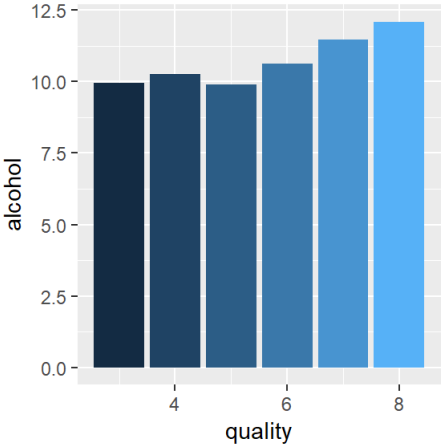

```
data_grouped_6 <- data_no_outliers %>% group_by(quality) %>% summarise(density = mean(density))
ggplot(aes(x = quality, y = density, fill = quality), data = data_grouped_6) + geom_bar(stat = "identity") +
theme(legend.position="none")
```




```
data_grouped_7 <- data_no_outliers %>% group_by(quality) %>% summarise(sulphates = mean(sulphates))
ggplot(aes(x = quality, y = sulphates, fill = quality), data = data_grouped_7) + geom_bar(stat = "identity")
+ theme(legend.position="none")
```




```
data_grouped_8 <- data_no_outliers %>% group_by(quality) %>% summarise(alcohol = mean(alcohol))
ggplot(aes(x = quality, y = alcohol, fill = quality), data = data_grouped_8) + geom_bar(stat = "identity") +
theme(legend.position="none")
```



A continuación, vamos a crear el modelo de regresión lineal con todos los atributos menos los atributos que menos relación tienen con la calidad del vino: “residual.sugar”, “free.sulfur.dioxide” y “pH”.

```
reg_model_1 <- lm(quality~fixed.acidity+volatile.acidity+citric.acid+chlorides+total.sulfur.dioxide+density+
sulphates+alcohol, data=data_no_outliers)

summary(reg_model_1)
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     chlorides + total.sulfur.dioxide + density + sulphates +
##     alcohol, data = data_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73155 -0.37089 -0.06352  0.44872  1.98177
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.549e+01  1.509e+01   1.689 0.091447 .
## fixed.acidity         5.293e-02  1.727e-02   3.064 0.002218 **
## volatile.acidity     -1.133e+00  1.194e-01  -9.485  < 2e-16 ***
## citric.acid          -2.129e-01  1.451e-01  -1.468 0.142390
## chlorides            -1.587e+00  4.072e-01  -3.897 0.000101 ***
## total.sulfur.dioxide -2.295e-03  5.484e-04  -4.185 3.01e-05 ***
## density              -2.285e+01  1.514e+01  -1.510 0.131286
## sulphates             9.349e-01  1.121e-01   8.341  < 2e-16 ***
## alcohol               2.653e-01  2.058e-02  12.889  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6488 on 1588 degrees of freedom
## Multiple R-squared:  0.3562, Adjusted R-squared:  0.353
## F-statistic: 109.8 on 8 and 1588 DF,  p-value: < 2.2e-16
```

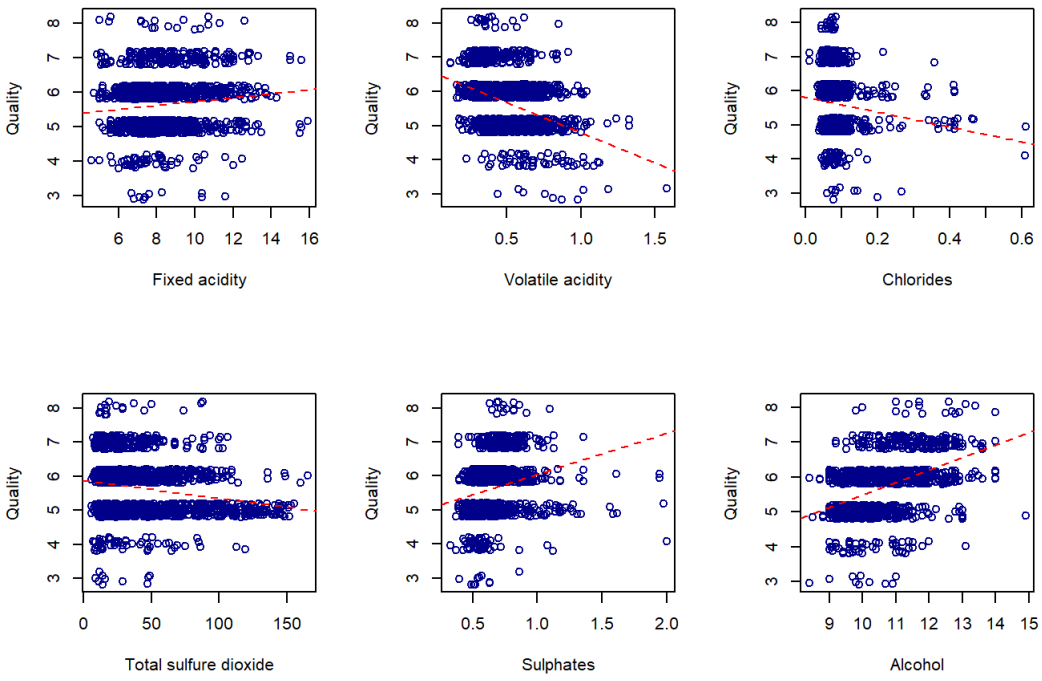
Como podemos apreciar, los atributos más significativos a la hora de definir la calidad del vino son “volatile.acidity”, “sulphates” y “alcohol”. También son significativos los atributos “fixed.acidity”, chlorides" y “total.sulfure.dioxide”, aunque en menor cantidad.

Ejercicio 5

Enunciado: Representación de los resultados a partir de tablas y gráficas.

A continuación, vamos a representar gráficamente la relación de los atributos más correlacionados con la calidad del vino mediante un scatter plot y una línea que nos muestre la regresión lineal del atributo con respecto a la calidad del vino.

```
par(mfrow=c(2,3))
plot(data_no_outliers$fixed.acidity, jitter(data_no_outliers$quality), col = "blue4", xlab = "Fixed acidity"
, ylab = "Quality")
abline(lm(data_no_outliers[,"quality"] ~ data_no_outliers[ ,"fixed.acidity"]), col = "red", lty = 2)

plot(data_no_outliers$volatile.acidity, jitter(data_no_outliers$quality), col = "blue4", xlab = "Volatile ac
idity", ylab = "Quality")
abline(lm(data_no_outliers[,"quality"] ~ data_no_outliers[ ,"volatile.acidity"]), col = "red", lty = 2)

plot(data_no_outliers$chlorides, jitter(data_no_outliers$quality), col = "blue4", xlab = "Chlorides", ylab =
"Quality")
abline(lm(data_no_outliers[,"quality"] ~ data_no_outliers[ ,"chlorides"]), col = "red", lty = 2)

plot(data_no_outliers$total.sulfur.dioxide, jitter(data_no_outliers$quality), col = "blue4", xlab = "Total s
ulfure dioxide", ylab = "Quality")
abline(lm(data_no_outliers[,"quality"] ~ data_no_outliers[ ,"total.sulfur.dioxide"]), col = "red", lty = 2)

plot(data_no_outliers$sulphates, jitter(data_no_outliers$quality), col = "blue4", xlab = "Sulphates", ylab =
"Quality")
abline(lm(data_no_outliers[,"quality"] ~ data_no_outliers[ ,"sulphates"]), col = "red", lty = 2)

plot(data_no_outliers$alcohol, jitter(data_no_outliers$quality), col = "blue4", xlab = "Alcohol", ylab = "Qu
ality")
abline(lm(data_no_outliers[,"quality"] ~ data_no_outliers[ ,"alcohol"]), col = "red", lty = 2)
```



Ejercicio 6

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

Analizando los resultados obtenidos, podemos concluir que las características del vino que más influyen en la calidad del vino son el alcohol, la acidez volatil y la cantidad del sulfatos. En el caso del alcohol, podemos concluir que a mayor cantidad de alcohol los usuarios valoran mejor el vino. En el caso de la acidez volatil y la cantidad de sulfatos podemos concluir que a mayor cantidad de sulfatos y menor acidez volátil el vino es mejor valorado. Aunque menos, las características de la acidez fija, la cantidad de cloruros y la cantidad total de dioxido de sulfuro también influyen en la calidad del vino de la siguiente manera: a mayor acidez fija, de mayor calidad es el vino y a menor cantidad de cloruros y menor cantidad total de dioxido de sulfuro, la calidad del vino aumenta.