Tipología y ciclo de vida de los datos: PRA2 - Selección y preparación de un juego de datos

Autor: Manuel Taberner Llorca y Andrés Pérez Santano

Diciembre 2020

Contents

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

El propósito de este proyecto es realizar un análisis de datos exploratorios para descubrir distribuciones, valores atípicos, relaciones y cualquier otro suceso sorprendente mediante la exploración de datos de una variable a múltiples variables. El objetivo de este proyecto es encontrar variables importantes que influyan en la calidad del vino tinto.

Dataset

El dataset seleccionado es el siguiente:

UCI: Wine Quality Data Set - Red Wine

En la página se incluyen dos dataset diferentes uno para vinos blancos y otro para vinos tintos y se ha seleccionado el segundo. Los datasets son de dominio público pero el crédito de juntar toda la información es de los siguientes autores.

- Paulo Cortez, University of Minho, Guimarães, Portugal.
- A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal

Descripción de variables

- **Fixed acidity**: la mayoría de los ácidos involucrados con el vino son fijos o no volátiles (no se evaporan fácilmente).
- **Volatile acidity**: la cantidad de ácido acético en el vino, que en grandes niveles puede provocar un sabor desagradable a vinagre.
- **Citric acid**: se encuentra en pequeñas cantidades, el ácido cítrico puede agregar 'frescura' y sabor a vinos.
- **Residual sugar**: la cantidad de azúcar que queda después de que se para la fermentación, es raro encontrar vinos con menos de 1 gramo/litro y vinos con más de 45 gramos/litro se consideran dulces.
- **Chlorides**: la cantidad de sal en el vino
- **Free sulfur dioxide**: la forma libre del SO2 que existe en equilibrio entre SO2 molecular (como gas disuelto) e ion bisulfito
- **Total sulfur dioxide**: cantidad de formas libres y ligadas del SO2 en bajas concentraciones, el SO2 es mayormente indetectable en el vino, pero en SO2 libre concentraciones superiores a 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino
- **Density**: la densidad del vino es cercana a la del agua dependiendo de el porcentaje de contenido de alcohol y azúcar
- **pH**: describe cómo de ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico), la mayoría de los vinos tienen entre 3-4 en la escala de pH
- **Sulphates**: un aditivo del vino que puede contribuir al gas de dióxido de azufre(S02), que actúa como antimicrobiano y antioxidante
- **Alcohol**: el porcentaje de contenido de alcohol del vino
- **Quality**: variable de salida (basada en datos sensoriales, puntuación entre 0 y 10)

Información de las variables

Variables de entrada (basadas en tests fisicoquímicos):

- fixed acidity (tartaric acid - g/dmˆ3)
- volatile acidity (acetic acid - g/dmˆ3)
- citric acid (g/dmˆ3)
- residual sugar (g/dmˆ3)
- chlorides (sodium chloride - g/dmˆ3

- free sulfur dioxide (mg/dm^3)
- total sulfur dioxide (mg/dm^3)
- density (g/cm^3)
- pH
- sulphates (potassium sulphate - g/dm3)
- alcohol (% de volumen)

Variable de salida (basada en datos sensoriales de personas):

- quality (puntuación entre 0 y 10)

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

Carga de las librerias para el proyecto

- ggplot2 : se utiliza en la creación de gráficas
- dplyr: se utiliza para manipular funciones de los datos
- gridExtra: permite colcoar varias gráficas en el mismo grid
- reshape: se usa para funciones de agregación de datos
- RColorBrewer: libreria de colores de palettes
- lattice: gráficos
- scales: métodos genericos de esclado de las gráficas
- memisc: herramientas que facilitan el traba
- reshape: se usa para funciones de agregación de datos
- sandwich: Construcción de estimadores de matrices de covarianza de sandwich multiplicando matrices de bread y meat
- graphics: funciones para gráficos
- ggbiplot: crea una gráfica especial para el PCA

```
library(ggplot2)
library(ggbiplot) # Se ha instalado de una forma especial para poder realizar la gráfica de abajo
# install.package('devtools')
# library(devtools)
# install_github("vqv/ggbiplot", force = TRUE)
library(dplyr)
library(gridExtra)
library(RColorBrewer)
library(lattice)
library(scales)
library(memisc)
library(reshape)
library(sandwich)
library(GGally)
library(graphics)
```

Toma de contacto con el dataset

```
# Carga de el dataset de RED WINE desde el repositorio de UCI
vinos <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-r
# Observación de las primeras 5 lineas del dataset
head(vinos)
```

```
##   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
```

```
# Resumen del dataset y las variables
str(vinos)
```

```
## '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 ...
```

Observamos que nuestro dataset está compuesto por 1599 observaciones y 12 variables.

```
# Resumen de estadísticas básicas sobre cada variable
summary(vinos)
```

```
##  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
```

Vamos a realizar un primer análisis sobre las estadísitcas básicas. * No se observa ningún vino que tenga calidad superior a 8 * Todas las variables parece que son continuas * Quality es una variable categórica ordinal, cómo es la variable sobre la que estamos interesados la vamos a modificar para realizar el estudio.

3. Limpieza de los datos

```
# Estadísticas de valores vacíos
colSums(is.na(vinos))
```

```
##        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
```

```
# Estadísticas de valores vacíos
colSums(vinos=="")
```

```
##        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
```

Cómo se puede observar nuestro dataset no contiene ningún valor nulo o vacío.

4. Análisis de los datos

En esta sección vamos a realizar una exploración inicial de los datos para observar como se comporta cada variable de manera indivudual y entender su estructura.

Creación de funciones que agilizan la escritura de código

```
# Creación de funciones para agilizar la escritura de código
# Crea la línea de separación de los Outliers, que es 3.5 desviaciones estandard mmás que la mediana de
outlier_line <- function(variable) {
  return (geom_hline(yintercept = (median(variable) + sd(variable)*3.5),
                     alpha = 1/3, linetype = 2))
}
# Encuentra la mediana que se dibuja en el histograma
median_hist <- function(variable) {
```

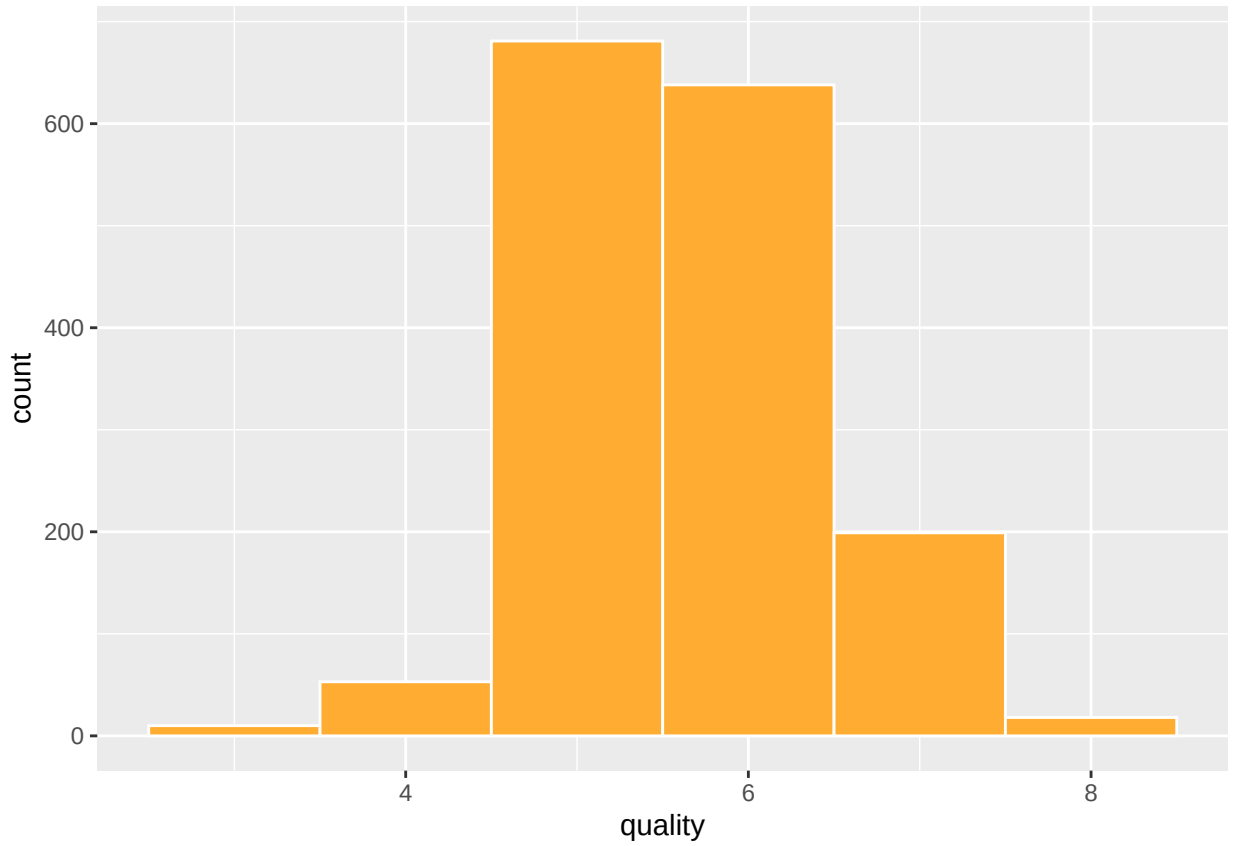
```
  return(geom_vline(xintercept = median(variable),
                    color = "red", alpha = 1/3))
}
# Crea un boxplot y dos histogramas para observar el comportamiento de la variable (normal, log10)
# 'variable_name' se usa para nombrar el Eje X.
boxplot_hist <- function(variable, variable_name) {
  return (grid.arrange(
    # boxplot de la variable para visualizar su distribución
    ggplot(aes(x = 1, y = variable), data = vinos) +
      geom_jitter(alpha = 0.1) +
      geom_boxplot(alpha = 0.2, color = 'blue') +
      stat_summary(fun=mean, shape=1, col = 'red', geom = 'point') +
      outlier_line(variable) +
      ylab(variable_name),
    # Histograma simple
    ggplot(aes(variable), data = vinos) +
      geom_histogram(bins=30, color = 'white', fill = '#FFAC33') +
      median_hist(variable) +
      labs(x = variable_name),
    # Histograma en log10 para observar si es una distribución normal
    ggplot(aes(log10(variable)), data = vinos) +
      geom_histogram(bins=30, color = 'white', fill = '#FF8633') +
      labs(x = paste("log10", variable_name)),
    ncol=3))
}
```

##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. En los siguientes apartados se realizan análisis que aportan la información solicitada por las cuestiones propuestas en los puntos 4.1 y 4.2.

Quality

El objetivo del estudio es intentar entender que variables son las que tienen mayor relación con al calidad del vino. Lo más adecuado es comenzar observando como se comporta la variable quality.

```
# Histogrma de quality de los vinos
ggplot(data = vinos, aes(x = quality)) +
      geom_bar(width = 1, color = 'white', fill = '#FFAC33')
```




```
# Estadísitcas básicas sobre la variable Quality
summary(vinos$quality)
```

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

Observamos que hay vinos con calidad máxima de 8 y mínima de 3. La media es de 5.636 y la mediana es de 6.

Discretizamos la variable quality Creamos una nueva variable a partir de la variable quality que se va a llamar **rating**, se trata de otra forma de puntuar la calidad de cada vino que los separa en solamente tres grupos.

- **low**: los vinos con calidad inferior a 5
- **medium**: los vinos con calidad inferior a 7
- **high**: los vinos con calidad superior o igual a 7

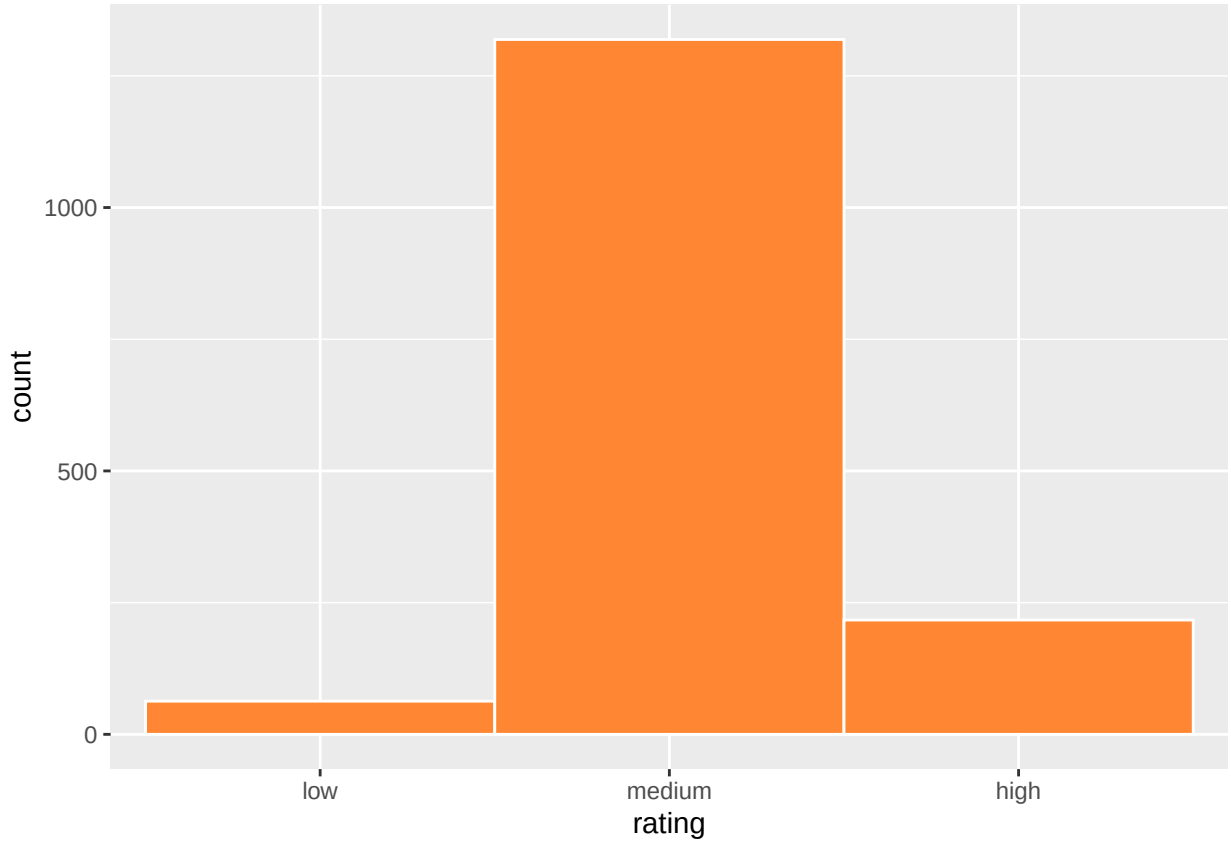
```
# Transformamos la variable calidad de un entero a un factor
calidad <- factor(vinos$quality, ordered = T)

# Creación de una variable factorizada llamada Rating
vinos$rating <- ifelse(calidad < 5, 'low', ifelse(
  calidad < 7, 'medium', 'high'))
vinos$rating <- ordered(vinos$rating,
                        levels = c('low', 'medium', 'high'))
head(vinos)
```

```
##   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 rating
## 1       5 medium
## 2       5 medium
## 3       5 medium
## 4       6 medium
## 5       5 medium
## 6       5 medium
```

```
# Histograma de rating
ggplot(data = vinos, aes(x = rating)) +
    geom_bar(width = 1, color = 'white', fill = '#FF8633')
```




```
# Estadísitcas básicas sobre la variable Rating
summary(vinos$rating)
```

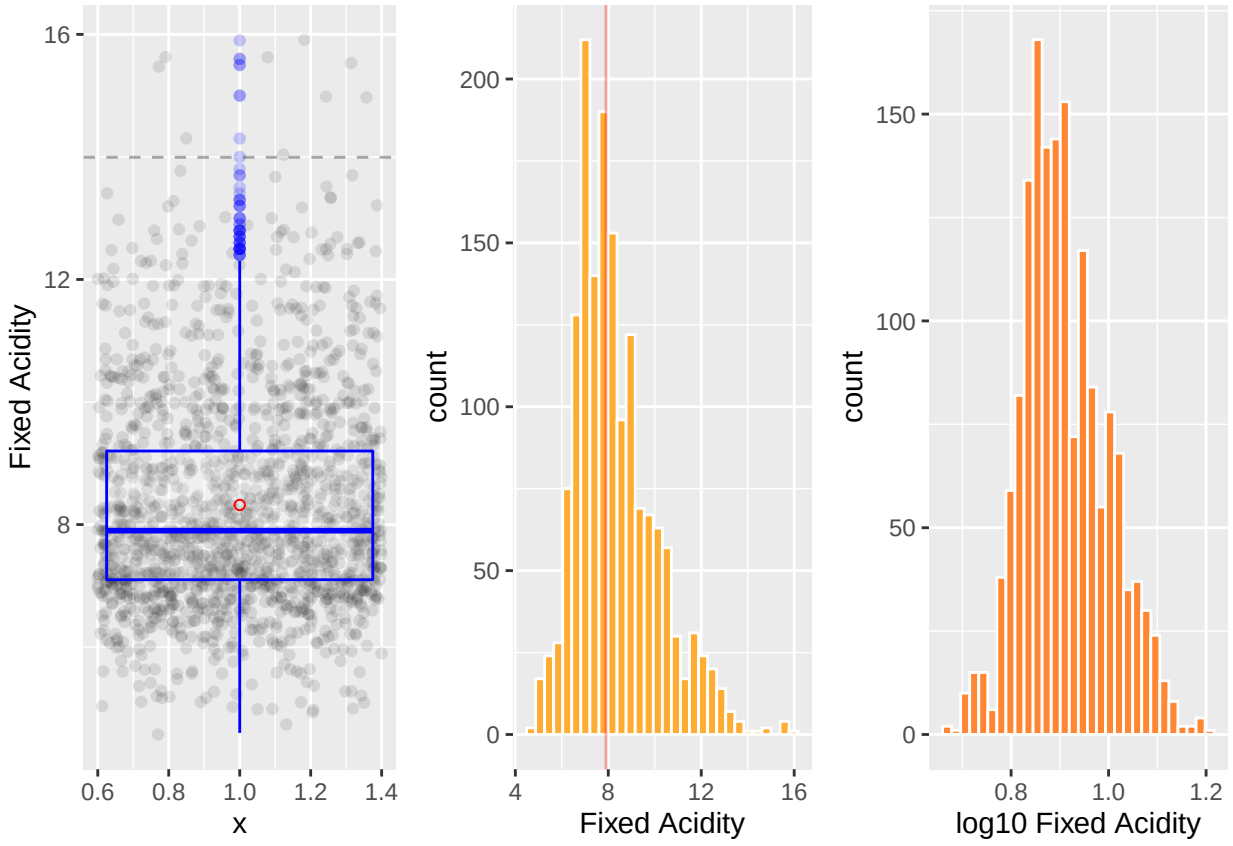
```
##    low medium   high
```

```
##      63    1319     217
```

Observaciones sobre la variable quality y la nueva variable introducida rating: * La mayoría de los vinos tienen puntuaciones de 5 y 6. * Más del 50% de los vinos son de calidad media.

Fixed Acidity

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$fixed.acidity, "Fixed Acidity")
```



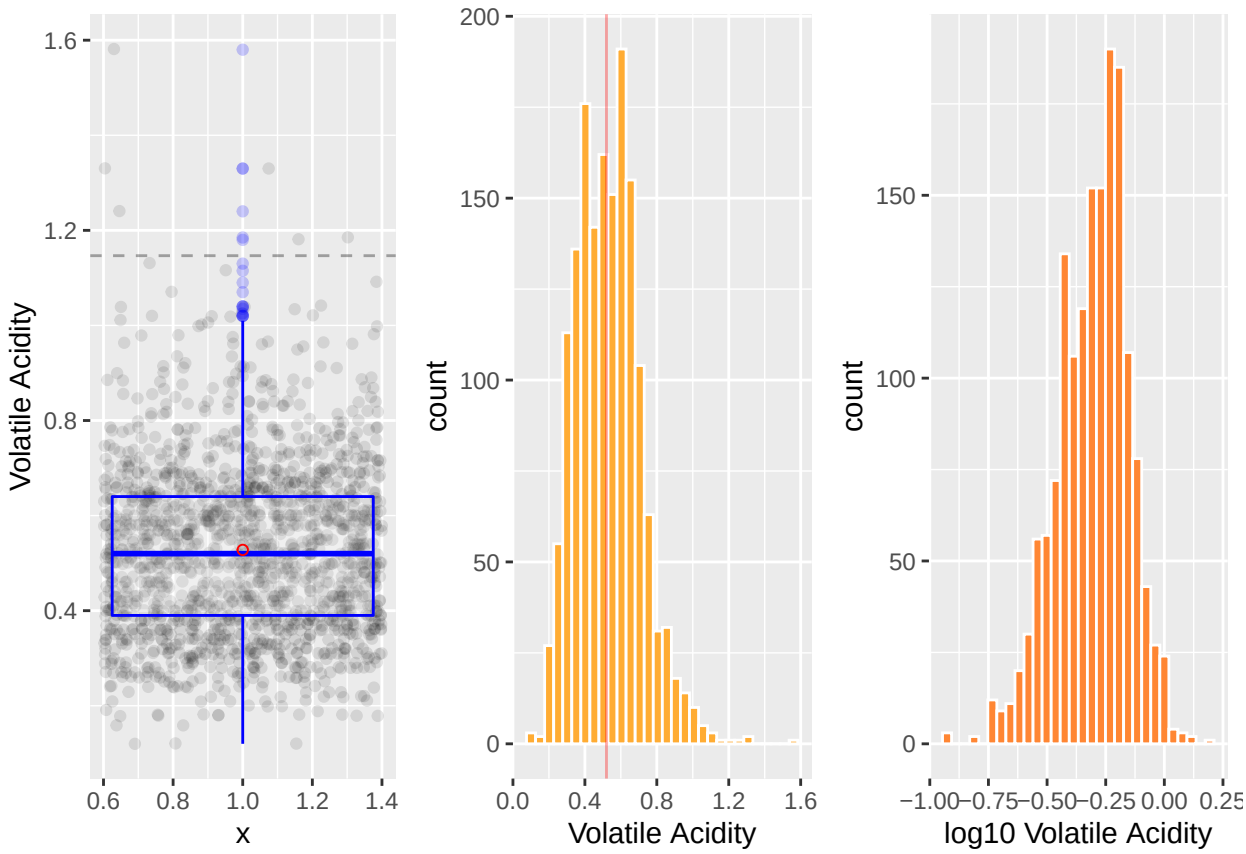

```
# Resumen estadísitcas básicas
summary(vinos$fixed.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

- Fixed acidity tiene una cola larga en la distribución.
- La gráfica log10 normaliza la distribución de la variable.
- Observamos un pico alrededor de 7, mediana de 7.90 ay los valores varian entre 4.60 y 15.90

Volatile Acidity

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$volatile.acidity, "Volatile Acidity")
```




```
# Resumen estadísitcas básicas
summary(vinos$volatile.acidity)
```

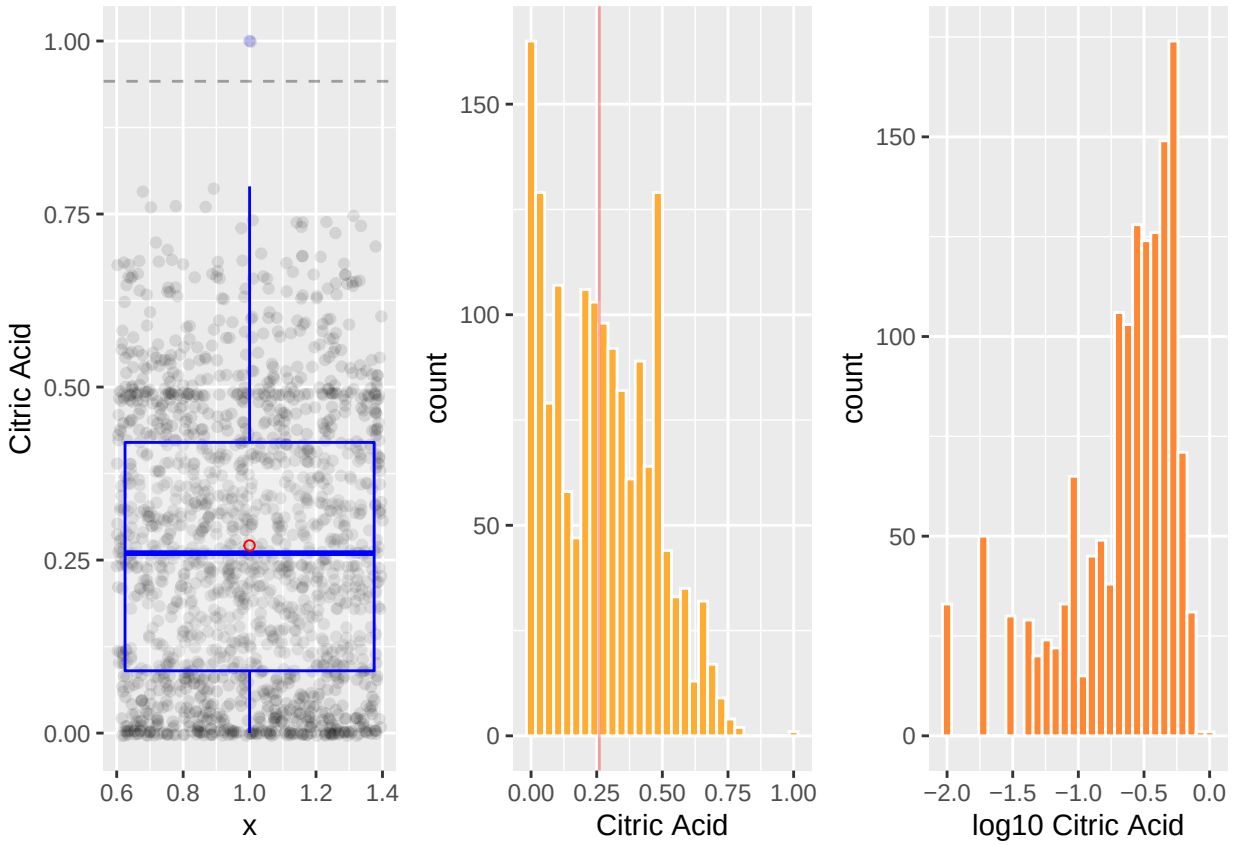
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

- Gráficas similares a fixed acidity, volatile acidity también tiene un cola larga en la distribución.
- Distribución normalizada con picos aproximademente en 0.4 y 0.7, y mediana de 0.52

Citric Acid

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$citric.acid, "Citric Acid")
```

```
## Warning: Removed 132 rows containing non-finite values (stat_bin).
```



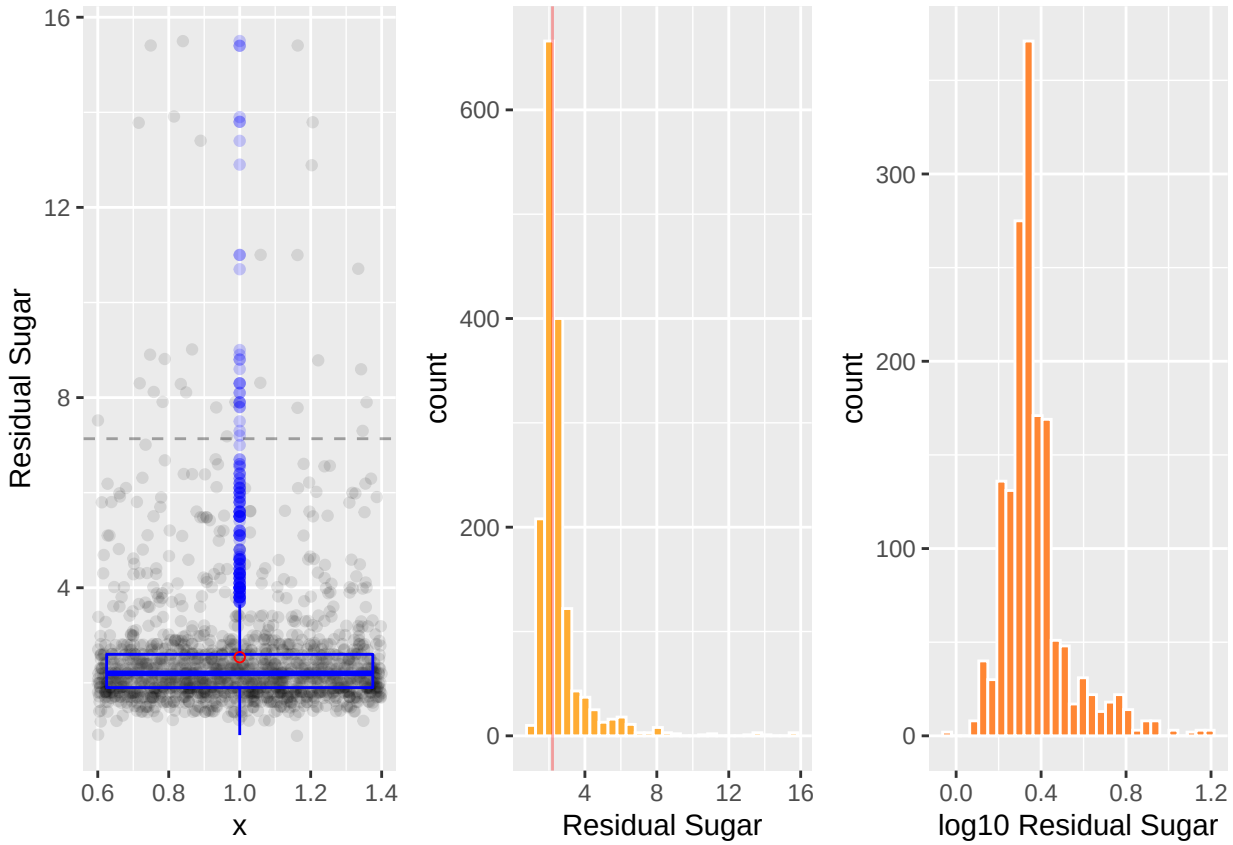

```
# Resumen estadísitcas básicas
summary(vinos$citric.acid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

- 132 vinos tienen 0.0 de citric acid, es decir, tenemos 132 vinos sin ácido cítrico
- Citric acid no parece que cumpla una distribución normal, como si hemos observado en las variables anteriores
- Interesente observar que para la gráfica de log10 la distribución cambia hacia la derecha
- Tiene sentido que haya vinos sin ácido cítrico ya que se añade como un refrescante al vino y además **NO** está permitido en la unión Europea

Residual Sugar

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$residual.sugar, "Residual Sugar")
```



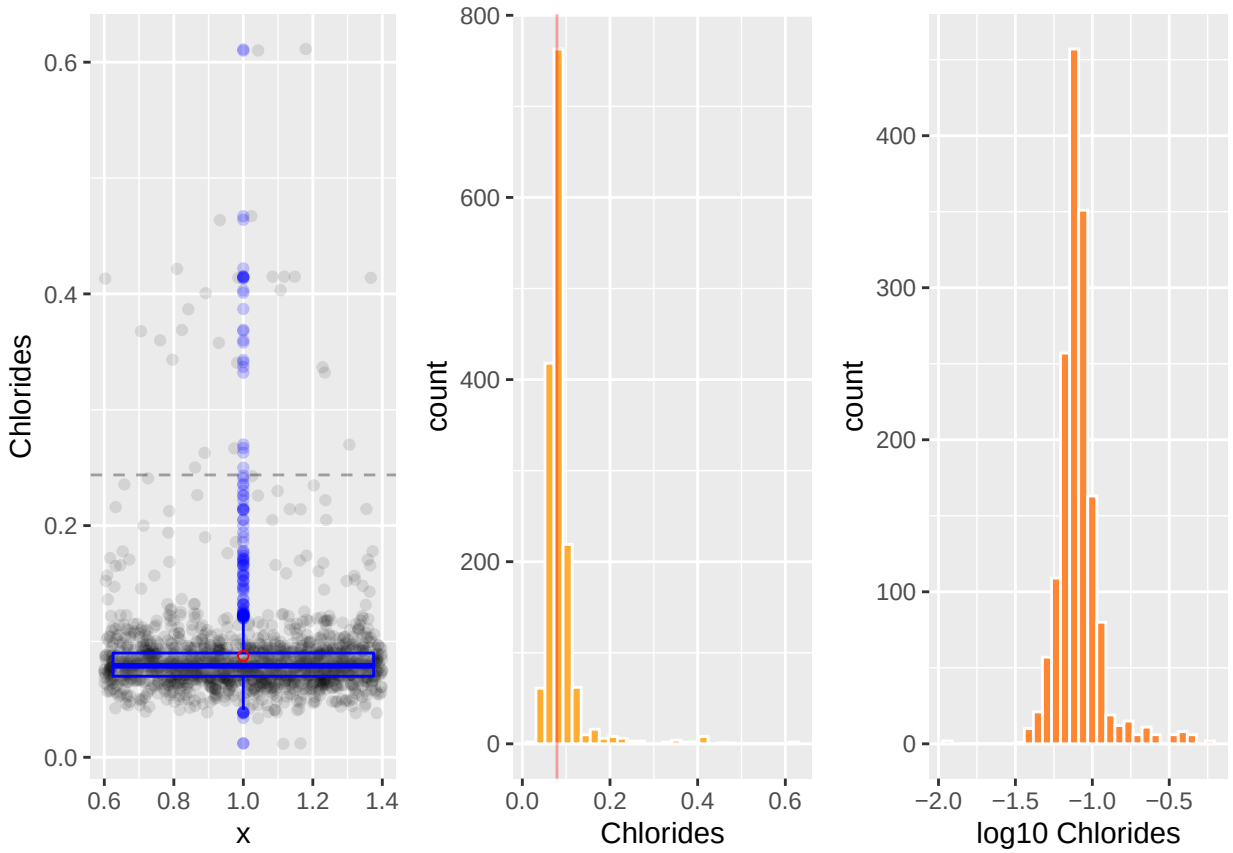

```
# Resumen estadísitcas básicas
summary(vinos$residual.sugar)
```

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

- Residual sugar tiene una cola muy larga en la distribución y hay muchos valores que són outliers. Algunos de ellos con desviaciones estandard de más de 9.
- La gráfica de log10 los valores siguen pegados a la izquierda per ya se asemeja más a una distribución normal
- Picos de variables alrededor de 2.3 con muchos outliers presentes en valores altos de residual sugar.

Chlorides

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$chlorides, "Chlorides")
```



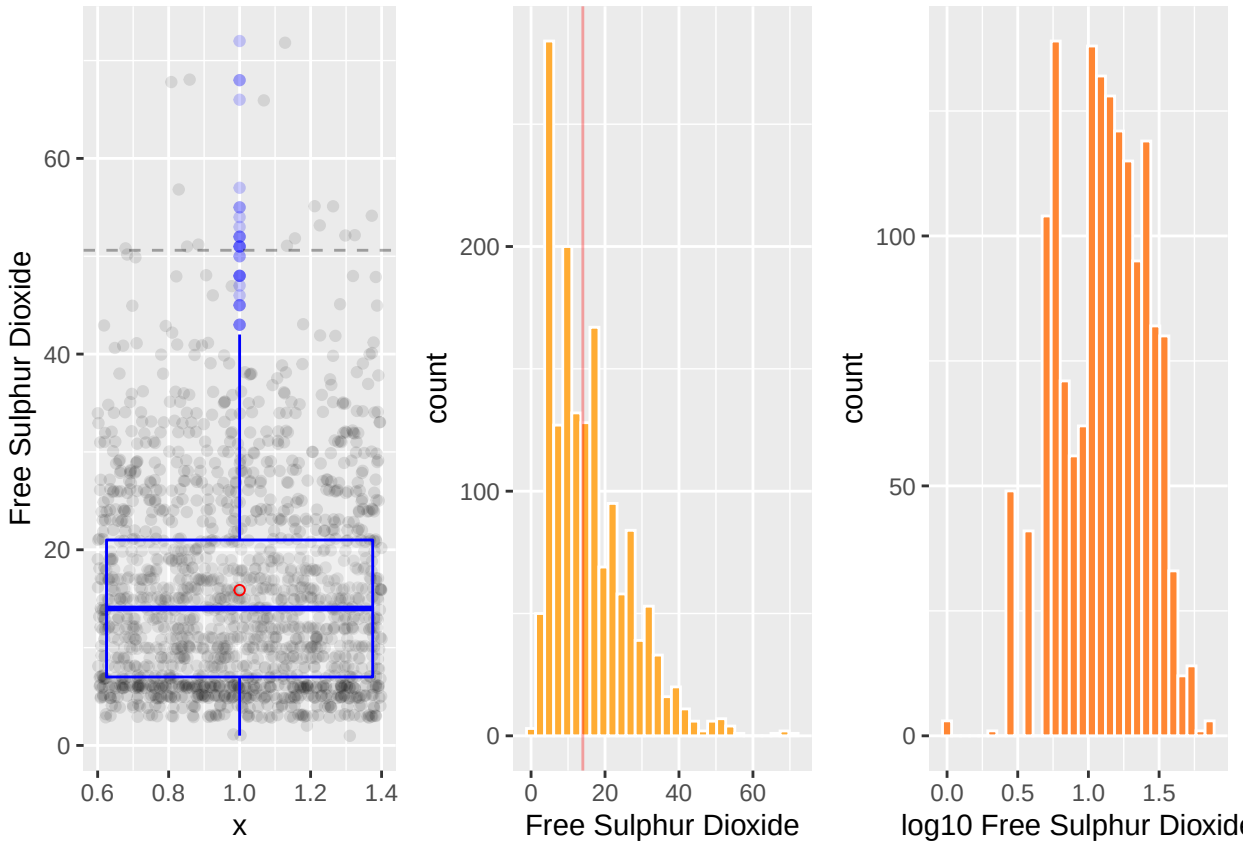

```
# Resumen estadísitcas básicas
summary(vinos$chlorides)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

- Muy similar a residual sugar, concentración de puntos cerca de la mediana y muchos outliers.
- Alguno de los outliers está muy lejos de la desviación estandard
- La gráfica log10 nos permite observar que a pesar de estar bastante sesgados los datos se asimila bastante a una distribución normal

Free Sulphur Dioxide

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$free.sulfur.dioxide, "Free Sulphur Dioxide")
```



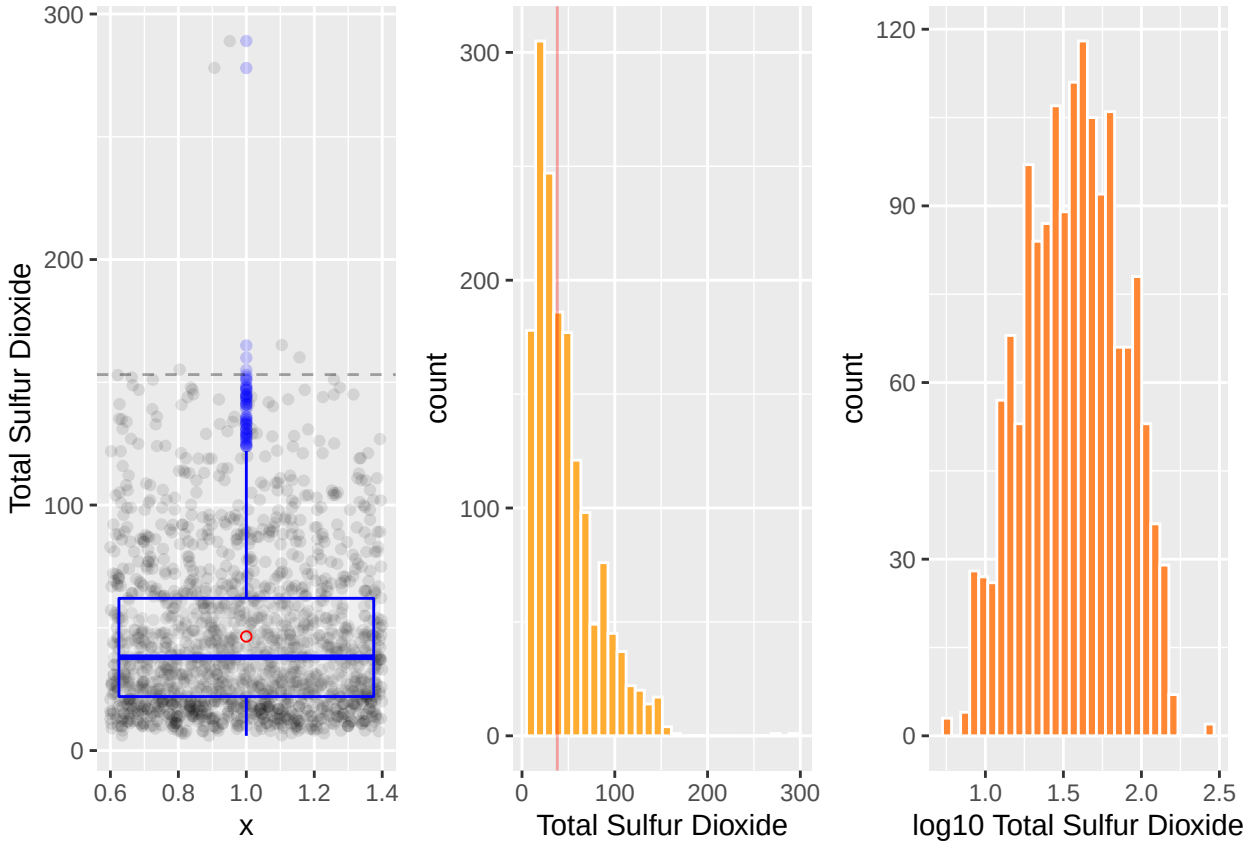

```
# Resumen estadísitcas básicas
summary(vinos$free.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

- Observamos que en el Free Sulphur Dioxide que hay un pico alrededor de 7 pero aún así sigue una distribución normal como estamos observando en la mayoría de las variables
- La mediana es 14 y 75% de la concentración del free sulfur es menos de 21 aunque el valor máximo es de 72.
- En la gráfica log10 observamos una distribución más dispersa.

Total Sulfur Dioxide

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$total.sulfur.dioxide, "Total Sulfur Dioxide")
```



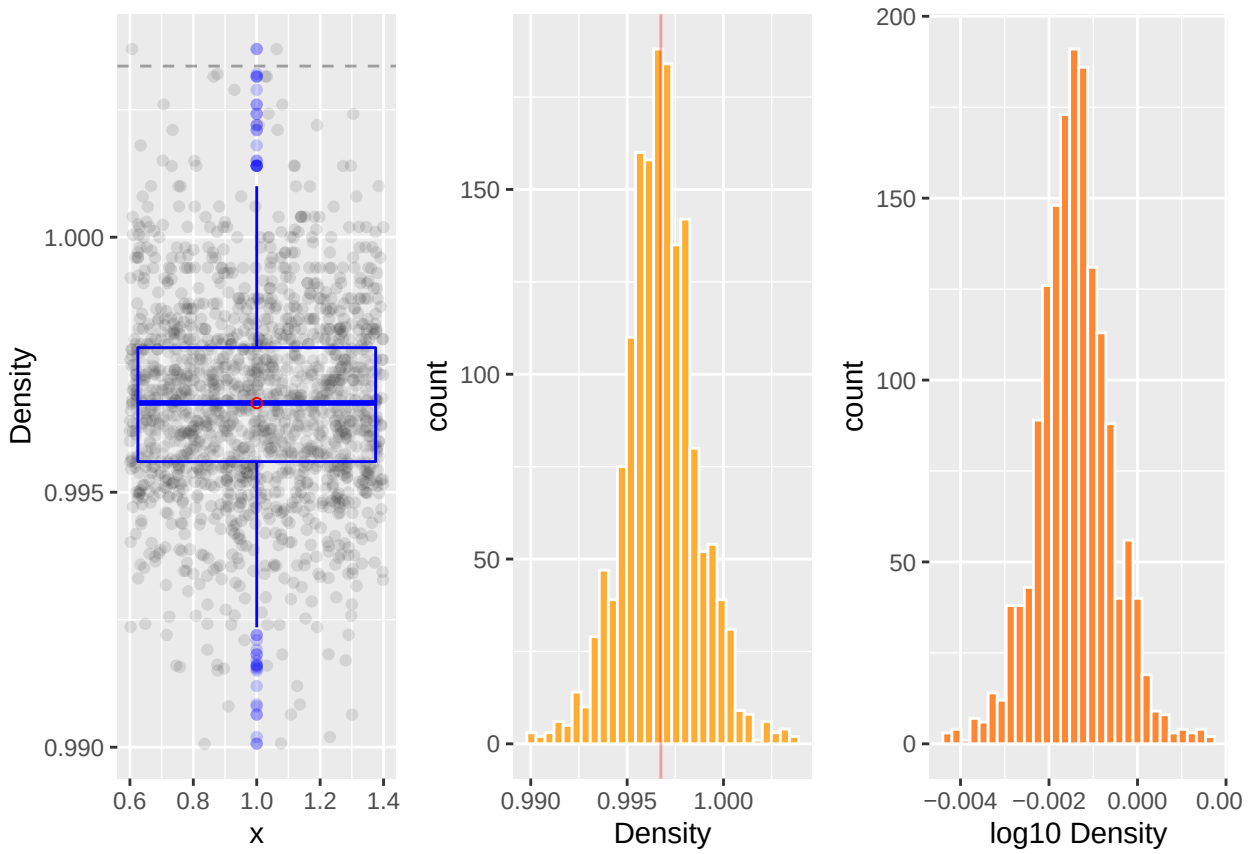

```
# Resumen estadísitcas básicas
summary(vinos$total.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

- Total sulfur dioxide no parece tan sipersa como la anterior porque su rango entre cuartiles no es tan elevado.
- La gráfica de log10 parece una distirbución normal.
- La mediana es de 38 y 75% de los vinos tienen una concentración menor de 62.

Density

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$density, "Density")
```



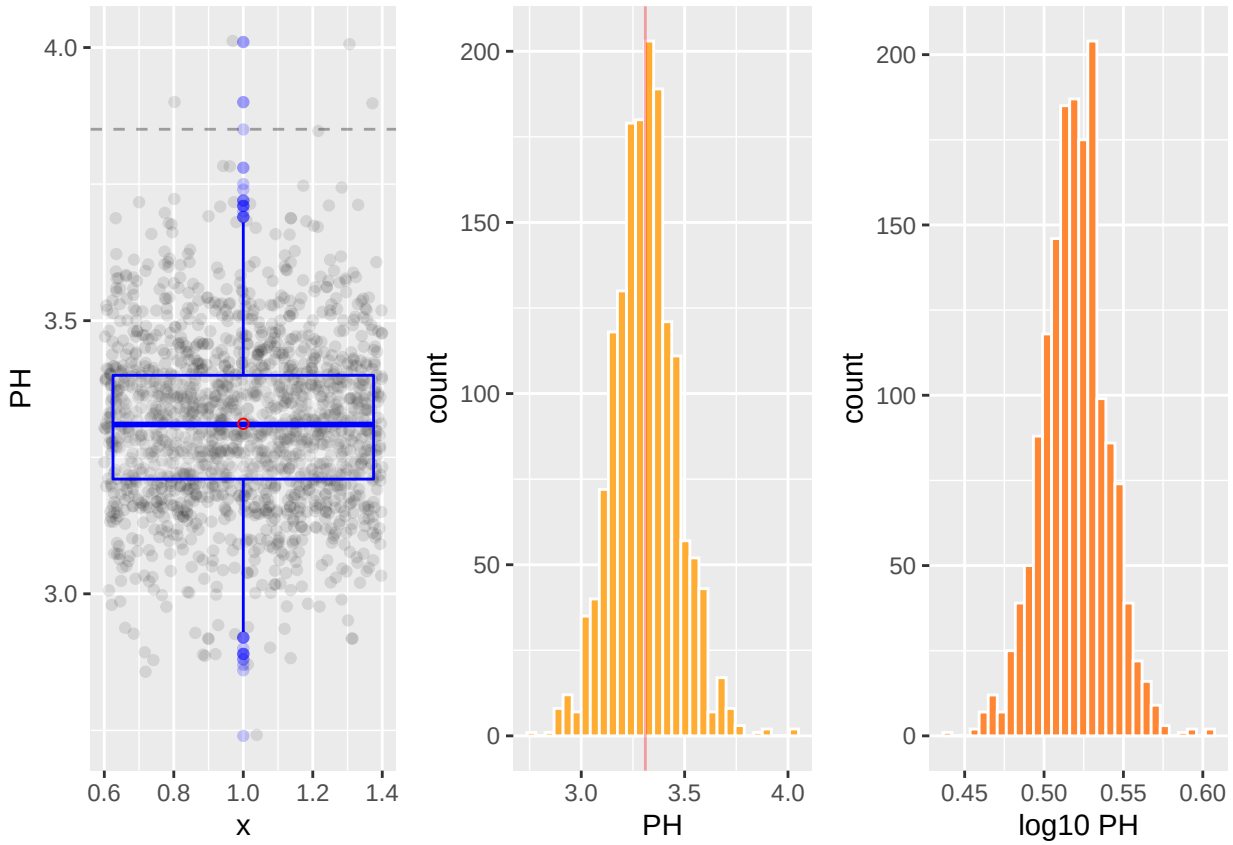

```
# Resumen estadísitcas básicas
summary(vinos$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

- La densidad varia entre 0.99 a 1. Es una variación muy pequeña.
- COmo observamos en ambas gráficas la distribución es normal..

PH

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$pH, "PH")
```



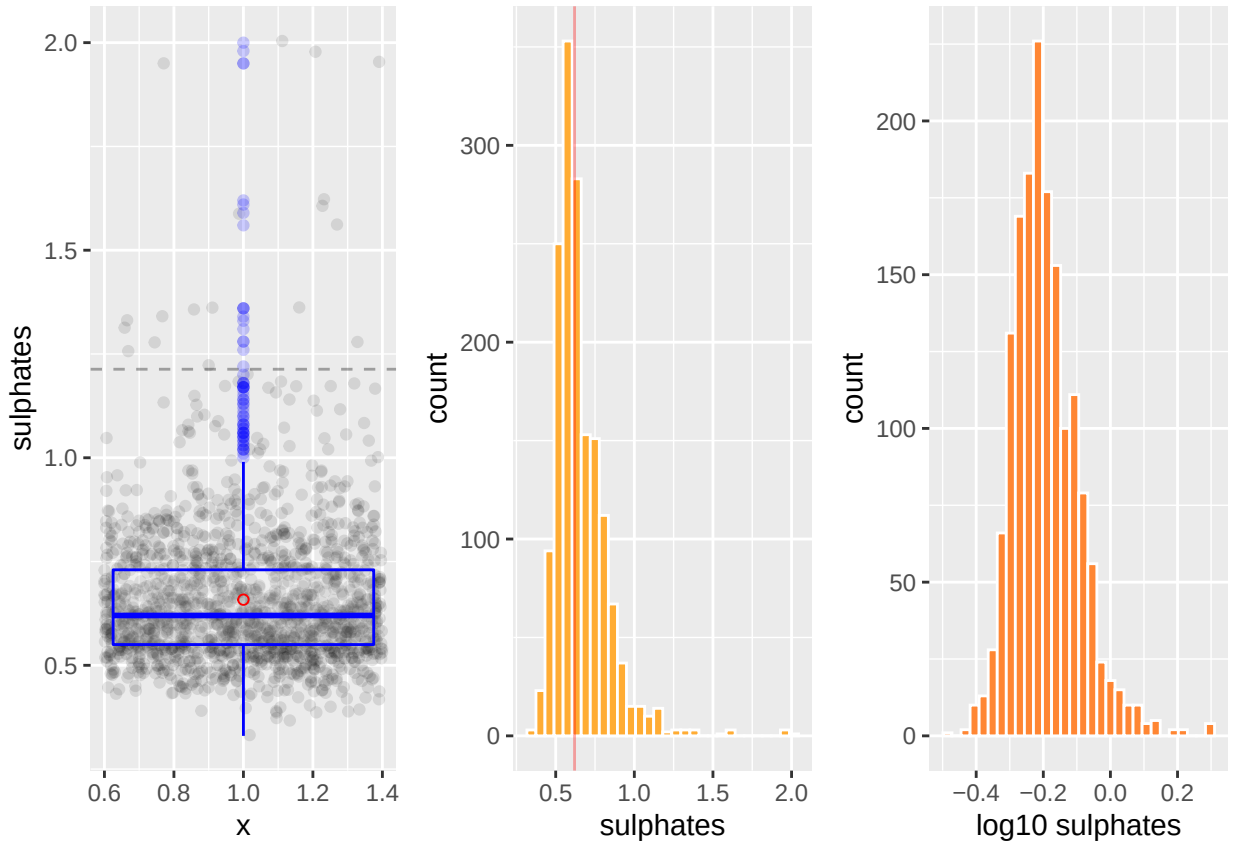

```
# Resumen estadísitcas básicas
summary(vinos$pH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

- Distribucioón totalmente normal, con la mayorá de los valores entre 3.1 y 3.5 y una mediana de 3.310

Sulphates

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$sulphates, "sulphates")
```



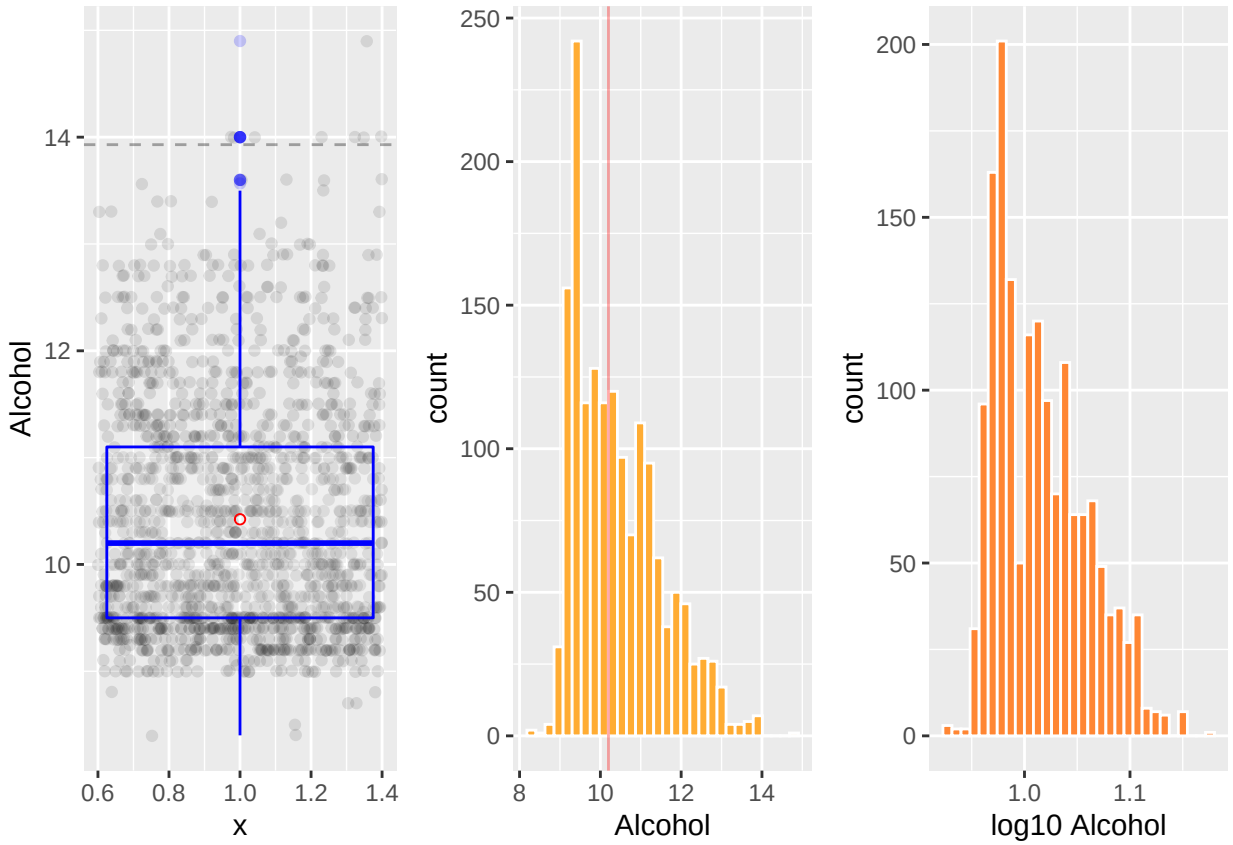

```
# Resumen estadísitcas básicas
summary(vinos$sulphates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

- La distribución de los sulfatos esta sesgada hacia la derecha y tiene muchos outliers elevados
- En la gráfica de log10 observamos una distribución normal.
- La mayoría de valores se encuentran alrededor de 0.6

Alcohol

```
# Gráficas para análisis de la variable
boxplot_hist(vinos$alcohol, "Alcohol")
```




```
# Resumen estadísitcas básicas
summary(vinos$alcohol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

- La mayoría de los vinos tienen menos de 11% de alcohol.
- La distribución de los valores está sesgada hacia la dercha con agunos picos y una media de 10.2
- Hay muy pocos valores por debajo de 9% y por ecnima de 13% de alcohol

Conclusiones del análisis indivudual de las variables

Mediante el análisis individual de cada variable hemos observado y lo siguiente:

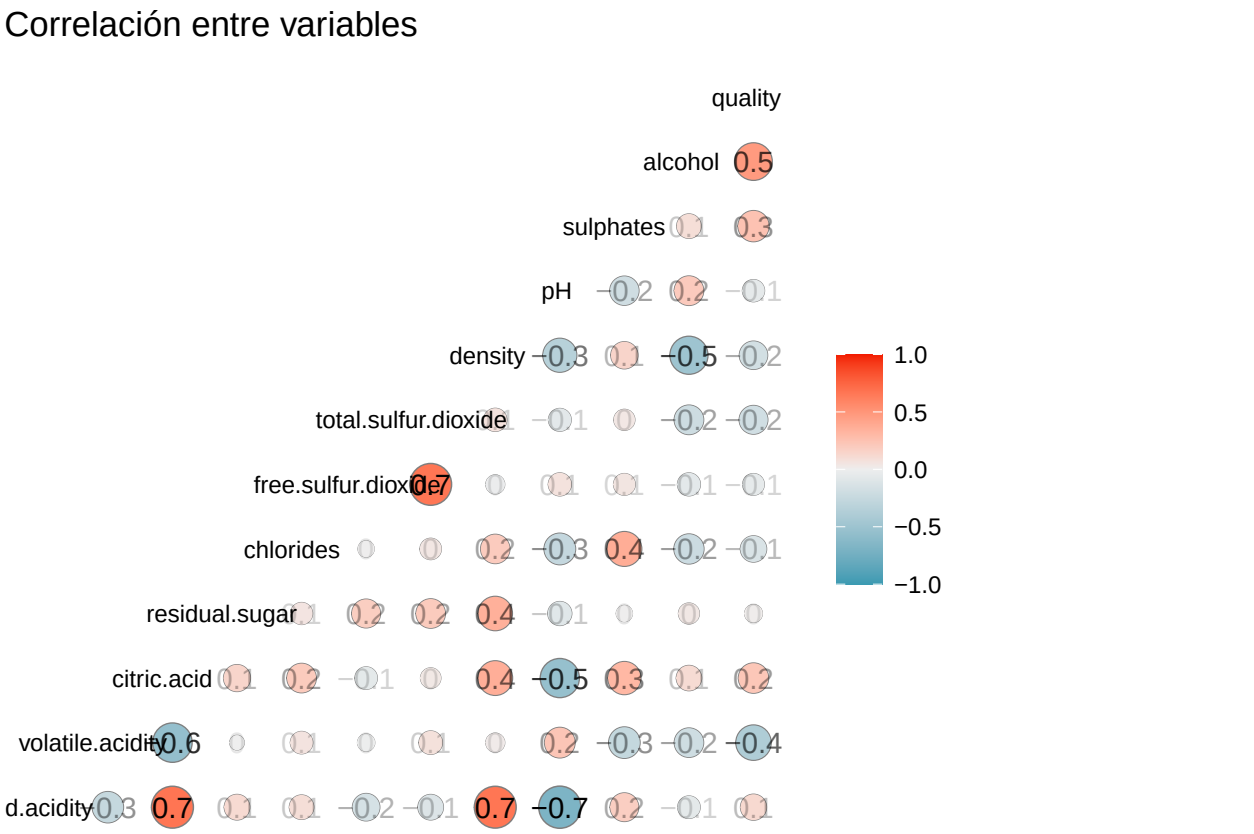
- 82.5% de los vinos tienen una calidad de 5 o 6
- La mayoría de los vinos tienen un contenido de alcohol menor al 11%
- 34% fixed.acidity valores entre 7 y 8 and
- 78% Citric.acid vvalores por debajo de 0.5
- pH y density son las variables que tienen una mayor distribución normal
- Residual sugar, chlorides, sulphates tienen outiers muy elevados
- Citric acid tiene una distribución bastante diferente a las demás variables numéricas, tiene muy pocos ouliers

  Se ha cambiado la variable **quality** a un factor y hemos creado una nueva variable **rating** para clasificar los vinos en tres categorias.

Estudio de las variables en conjuntos de dos

A continuación vamos a realizar análisis comparando las variables por parejas y comprobar la homogeneidad y varianza entre ellas. La mejor manera de empezar este análisis es crear una gráfica de correlación entre todas las variables como se muestra a continuación, esto permite observar que parejas de variables vamos a analizar y cuales no hace falta que prestemos atención.

```
# Cambio de la calidad a numeric para que no sea ignorada en la gráfica de correlación
vinos$quality<- as.numeric(vinos$quality)
ggcorr(vinos %>%
         dplyr::select(-rating), # quitamos la comuna rating ya que solo queremos los valores númericos
  hjust = 0.60,
  size = 3,
  label = TRUE,
  label_alpha = TRUE,
  geom = "circle",
  max_size = 7,
  size = 3,
  hjust = 0.75,
  angle = 0,
  palette = "viridis")+
  ggplot2::labs(title = "Correlación entre variables")
```



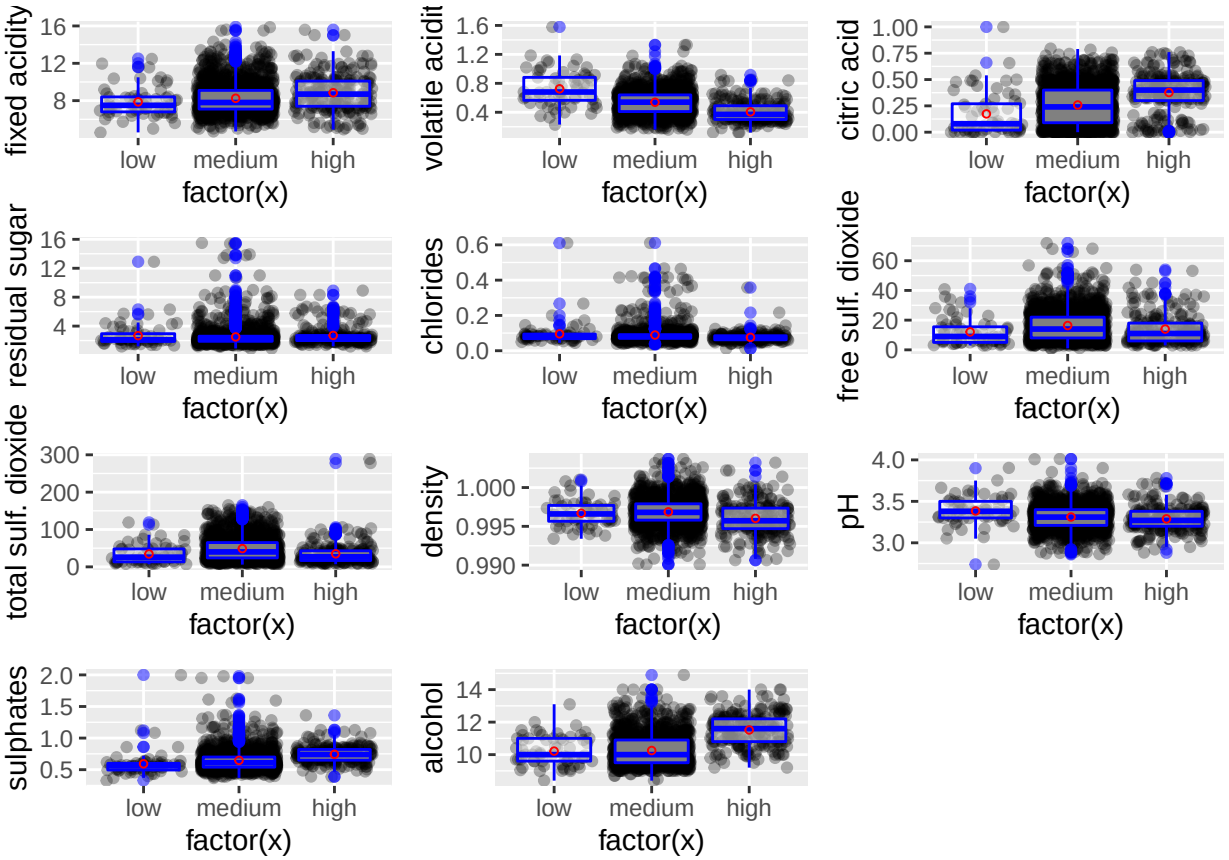
En el gráfico anteriro podemos observar las correlaciones entre las variables del dataset, las correlaciones que tienen una tonalidad roja son las que están relacionadas positivamente y las correlaciones de azul de manera negativa, cuanto más vibrante es el color mayor es la correlación. Las más interesantes son las siguientes:

- Este gráfico muestra que la calidad está relacionada positivamente de manera fuerte con alcohol y sulfates, por otro lado está correlacionada negativamente con volatile.acidity

- Residual.sugar no está nada relacionado con calidad
- Density y fixed.acidity tienen una correlación positiva fuerte.
- Se observa una correlacion negativa fuerte entre pH fixed/citric.acid, y también entre alcohol y density.
- Volatile.acidity tiene relación positiva con pH, esto se debe a que cuando aumentas el PH disminuye la acidez.

En la siguiente gráfica se puede observar de forma rápida como afectan cada variable a quality

```
# Creación de una función para crear gráficas de dos variables
twovar_boxplot <- function(x, y, ylab) {
  return(ggplot(aes(factor(x), y), data = vinos) +
           geom_jitter( alpha = .3)  +
           geom_boxplot( alpha = .5,color = 'blue')+
           stat_summary(fun=mean, shape=1, col = 'red',
                        geom = 'point', size = 1) +
           ylab(ylab))
}
# Crea las gráficas y las coloca de manera ordenada
grid.arrange(twovar_boxplot(vinos$rating, vinos$fixed.acidity,
                            'fixed acidity'),
             twovar_boxplot(vinos$rating, vinos$volatile.acidity,
                            'volatile acidity'),
             twovar_boxplot(vinos$rating, vinos$citric.acid,
                            'citric acid'),
             twovar_boxplot(vinos$rating, vinos$residual.sugar,
                            'residual sugar'),
             twovar_boxplot(vinos$rating, vinos$chlorides,
                            'chlorides'),
             twovar_boxplot(vinos$rating, vinos$free.sulfur.dioxide,
                            'free sulf. dioxide'),
             twovar_boxplot(vinos$rating,
                            vinos$total.sulfur.dioxide,
                            'total sulf. dioxide'),
             twovar_boxplot(vinos$rating, vinos$density,
                            'density'),
             twovar_boxplot(vinos$rating, vinos$pH,
                            'pH'),
             twovar_boxplot(vinos$rating, vinos$sulphates,
                            'sulphates'),
             twovar_boxplot(vinos$rating, vinos$alcohol,
                            'alcohol'),
             ncol = 3)
```



Como podemos observar en las gráficas anteriores, un *buen* vino normalmente sigue los siguientes patrones:

- Mayor fixed.acidity y citric.acid y poca volatile.acidity
- pH bajo
- Alta cantidad de sulphates
- Alto porcentaje de alcohol

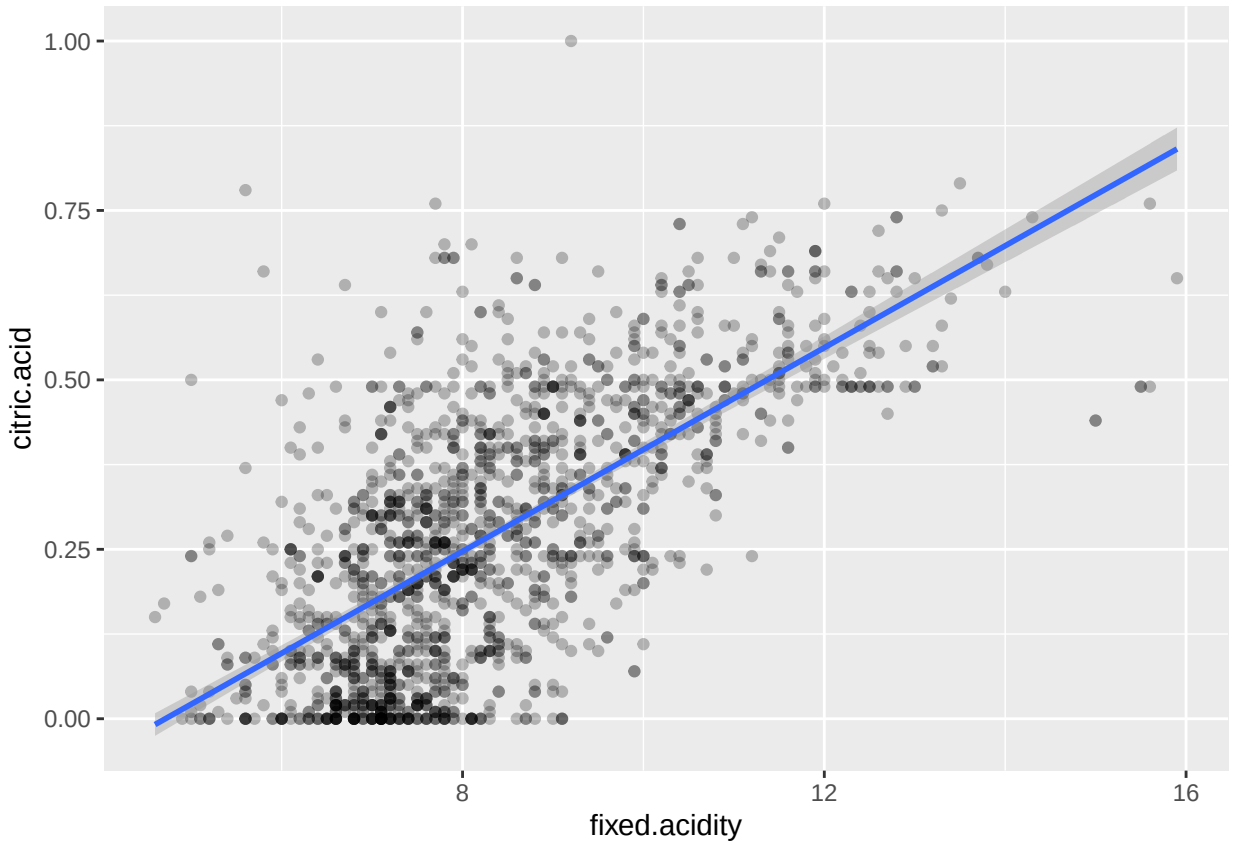
Correlaciones de variables

Gracias a la gráfica de correlaciones a continuación vamos a observar las que se considera son más fuertes ya sea negativas o positivas.

```
# Creación gráfica scatter para comparar dos variables
ggplot(data = vinos, aes(x = fixed.acidity, y = citric.acid)) +
  geom_point(alpha = 1/4) +
  geom_smooth(method = "lm")
```

Correlación entre Citric Acid y Fixed Acidity

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Test de correlación y estadísticas básicas
cor.test(vinos$fixed.acidity, vinos$citric.acid)
```

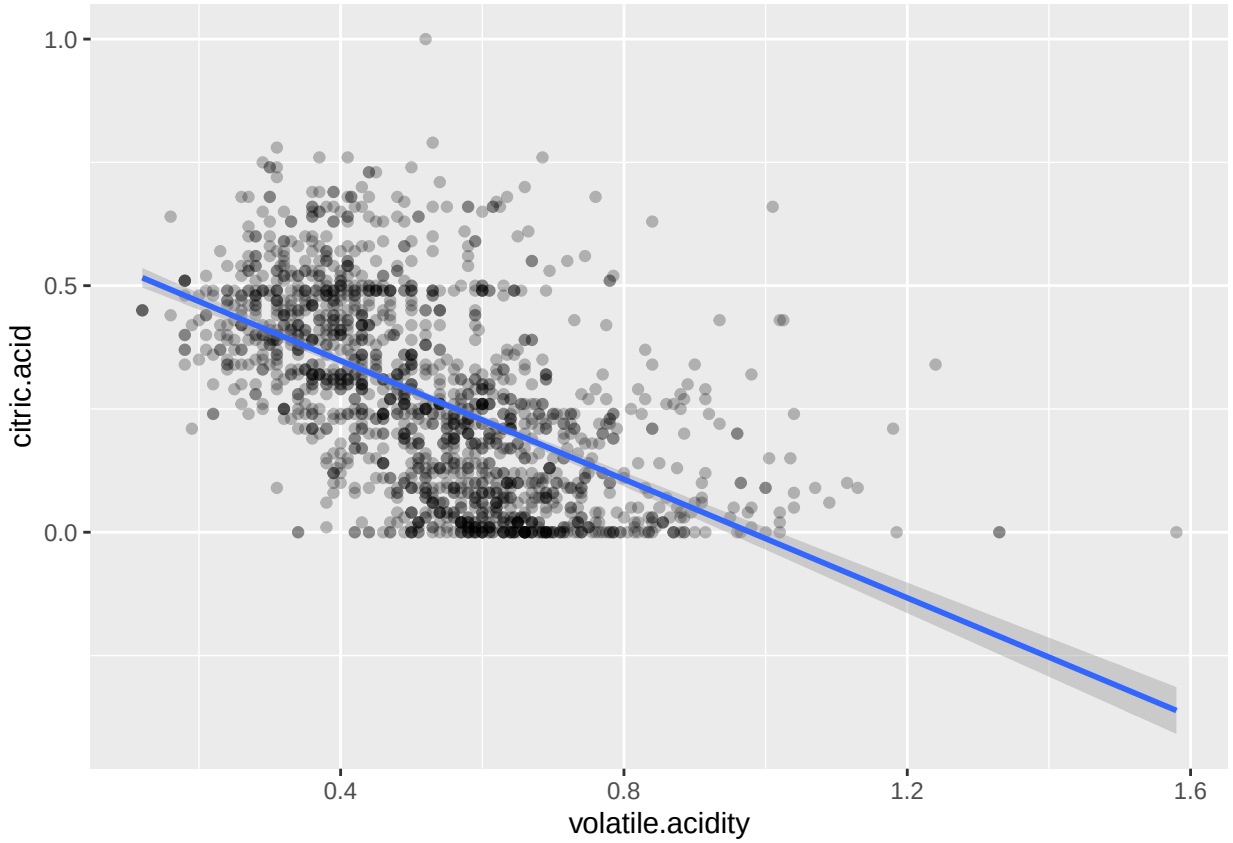
```
##
##  Pearson's product-moment correlation
##
## data:  vinos$fixed.acidity and vinos$citric.acid
## t = 36.234, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6438839 0.6977493
## sample estimates:
##       cor
## 0.6717034
```

- Citric acid y fixed acidity tienen una correlación positiva de 0.67
- Citric acid tiene una relación positiva débil de 0.23 con quality además fixed acidity tiene una relación muy débil con quality de 0.12.
- Ambas variables no afectan mucho al resultado de la calidad del vino

```
# Creación gráfica scatter para comparar dos variables
ggplot(data = vinos, aes(x = volatile.acidity, y = citric.acid)) +
  geom_point(alpha = 1/4) +
  geom_smooth(method = "lm")
```

Correlación entre Citric Acid y Volatile Acidity

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Test de correlación y estadísticas básicas
cor.test(vinos$volatile.acidity, vinos$citric.acid)
```

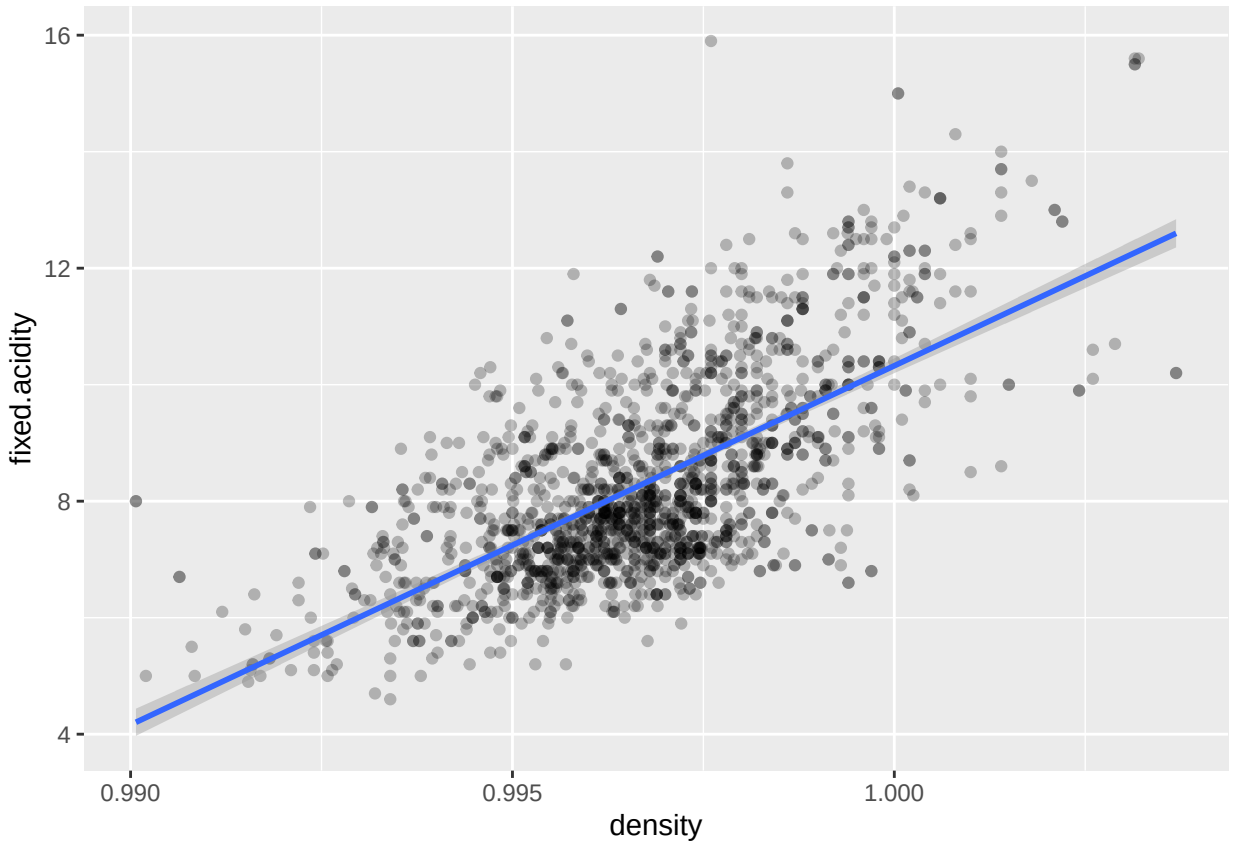
```
##
##  Pearson's product-moment correlation
##
## data:  vinos$volatile.acidity and vinos$citric.acid
## t = -26.489, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##       cor
## -0.5524957
```

- Citric acid y volatile acidity tienen una correlación negativa de -0.55
- No afectan en gran medida a quality

```
# Creación gráfica scatter para comparar dos variables
ggplot(data = vinos, aes(x = density, y = fixed.acidity)) +
  geom_point(alpha = 1/4) +
  geom_smooth(method = "lm")
```

Correlación entre Density y Fixed Acidity

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Test de correlación y estadísticas básicas
cor.test(vinos$density, vinos$fixed.acidity)
```

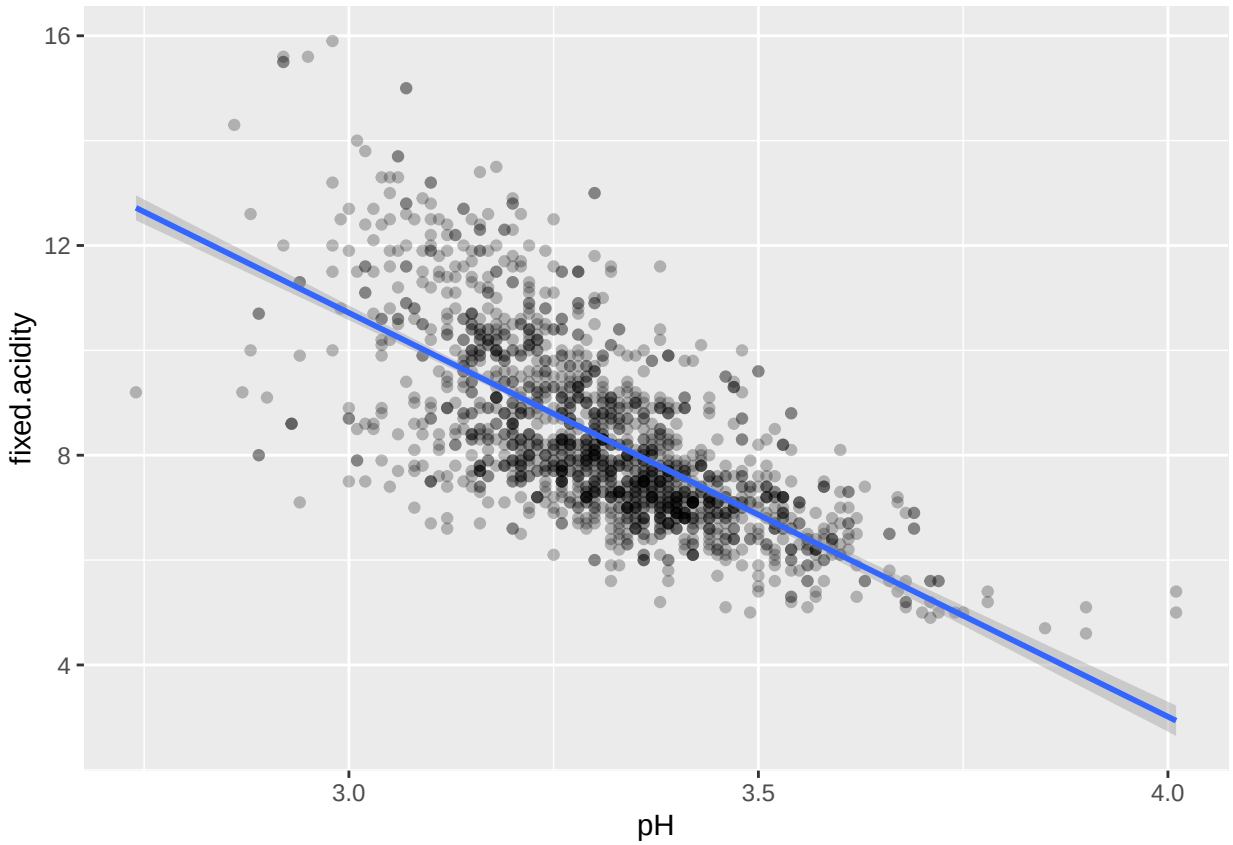
```
##
##  Pearson's product-moment correlation
##
## data:  vinos$density and vinos$fixed.acidity
## t = 35.877, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6399847 0.6943302
## sample estimates:
##       cor
## 0.6680473
```

- Cuando mayor es la cantidad de fixed acidity mayor es la density, es decir el vino es más denso
- La correlación entre estas variables es 0.668

```
# Creación gráfica scatter para comparar dos variables
ggplot(data = vinos, aes(x = pH, y = fixed.acidity)) +
  geom_point(alpha = 1/4) +
  geom_smooth(method = "lm")
```

Correlación entre PH y Fixed Acidity

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Test de correlación y estadísticas básicas
cor.test(vinos$pH, vinos$fixed.acidity)
```

```
##
##  Pearson's product-moment correlation
##
## data:  vinos$pH and vinos$fixed.acidity
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##        cor
## -0.6829782
```

- Observamos que la cantidad de pH es inversamente proporcional a fixed acidity
- La correalción entre ambas varaibles es de 0.0682

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.

Conclusión del análisis de variables por parejas

La gráfica de correlación ayuda a comprender la correlación entre diferentes características.La calidad está fuertemente correlacionada positivamente con el alcohol y los sulfatos, y negativamente con la acidez volátil. Los buenos vinos tienen valores de pH más bajos, que está relacionado con tener más ácido cítrico y fijo.

- El ácido cítrico y la acidez fija tienen una fuerte correlación positiva de 0,7, mientras que el ácido cítrico y la acidez volátil tienen una correlación negativa moderada de -0,6
- La densidad y la acidez fija son dos características con una fuerte correlación positiva de 0,7

- Correlación negativa entre alcohol y densidad.
- Se espera una fuerte correlación negativa entre el pH y el ácido cítrico y fijo.
- Una correlación positiva sorprendente entre el pH y la acidez volátil, ya que un valor de pH más alto significa menos acidez, pero una acidez volátil más alta significa más acidez.

```
# Funcion que devuelve la correlación de las variables
cor_test <- function(x, y) {
  return(cor(as.numeric(x), as.numeric(y)))
}
# Calculo de las correlacines normales
correlations <- c(
  cor_test(vinos$fixed.acidity, vinos$quality),
  cor_test(vinos$volatile.acidity, vinos$quality),
  cor_test(vinos$citric.acid, vinos$quality),
  cor_test(vinos$residual.sugar, vinos$quality),
  cor_test(vinos$chlorides, vinos$quality),
  cor_test(vinos$free.sulfur.dioxide, vinos$quality),
  cor_test(vinos$total.sulfur.dioxide, vinos$quality),
  cor_test(vinos$density, vinos$quality),
  cor_test(vinos$pH, vinos$quality),
  cor_test(vinos$sulphates, vinos$quality),
  cor_test(vinos$alcohol, vinos$quality))
names(correlations) <- c('fixed.acidity', 'volatile.acidity', 'citric.acid',
                         'residual.sugar', 'chlordies', 'free.sulfur.dioxide',
                         'total.sulfur.dioxide', 'density', 'pH',
                         'sulphates', 'alcohol')
# Cálculo de las correlaciones en log10
correlations_log10 <- c(
  cor_test(log10(vinos$fixed.acidity), vinos$quality),
  cor_test(log10(vinos$volatile.acidity), vinos$quality),
  cor_test(log10(vinos$citric.acid), vinos$quality),
  cor_test(log10(vinos$residual.sugar), vinos$quality),
  cor_test(log10(vinos$chlorides), vinos$quality),
  cor_test(log10(vinos$free.sulfur.dioxide), vinos$quality),
  cor_test(log10(vinos$total.sulfur.dioxide), vinos$quality),
  cor_test(log10(vinos$density), vinos$quality),
  cor_test(log10(vinos$pH), vinos$quality),
  cor_test(log10(vinos$sulphates), vinos$quality),
  cor_test(log10(vinos$alcohol), vinos$quality))
names(correlations_log10) <- c('fixed.acidity', 'volatile.acidity', 'citric.acid',
                         'residual.sugar', 'chlordies', 'free.sulfur.dioxide',
                         'total.sulfur.dioxide', 'density', 'pH',
                         'sulphates', 'alcohol')
```

Correlaciones normales entre quality y otras variables

```
correlations
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
##       residual.sugar            chlordies  free.sulfur.dioxide
##           0.01373164          -0.12890656          -0.05065606
## total.sulfur.dioxide               density                   pH
##          -0.18510029          -0.17491923          -0.05773139
##            sulphates               alcohol
```

```
##           0.25139708           0.47616632
```

Correlaciones log10 entre quality y otras variables

```
correlations_log10
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.11423756          -0.39124918                  NaN
##       residual.sugar            chlordies  free.sulfur.dioxide
##           0.02353331          -0.17613996          -0.05008749
## total.sulfur.dioxide              density                   pH
##          -0.17014272          -0.17517368          -0.05757386
##            sulphates              alcohol
##           0.30864193           0.47698109
```

Podemos decir que las siguientes variables tienen correlaciones relativamente más altas con la calidad del vino:

- **alcohol**
- **volatile acidity**
- **sulphates (log10)**
- **citric acid**

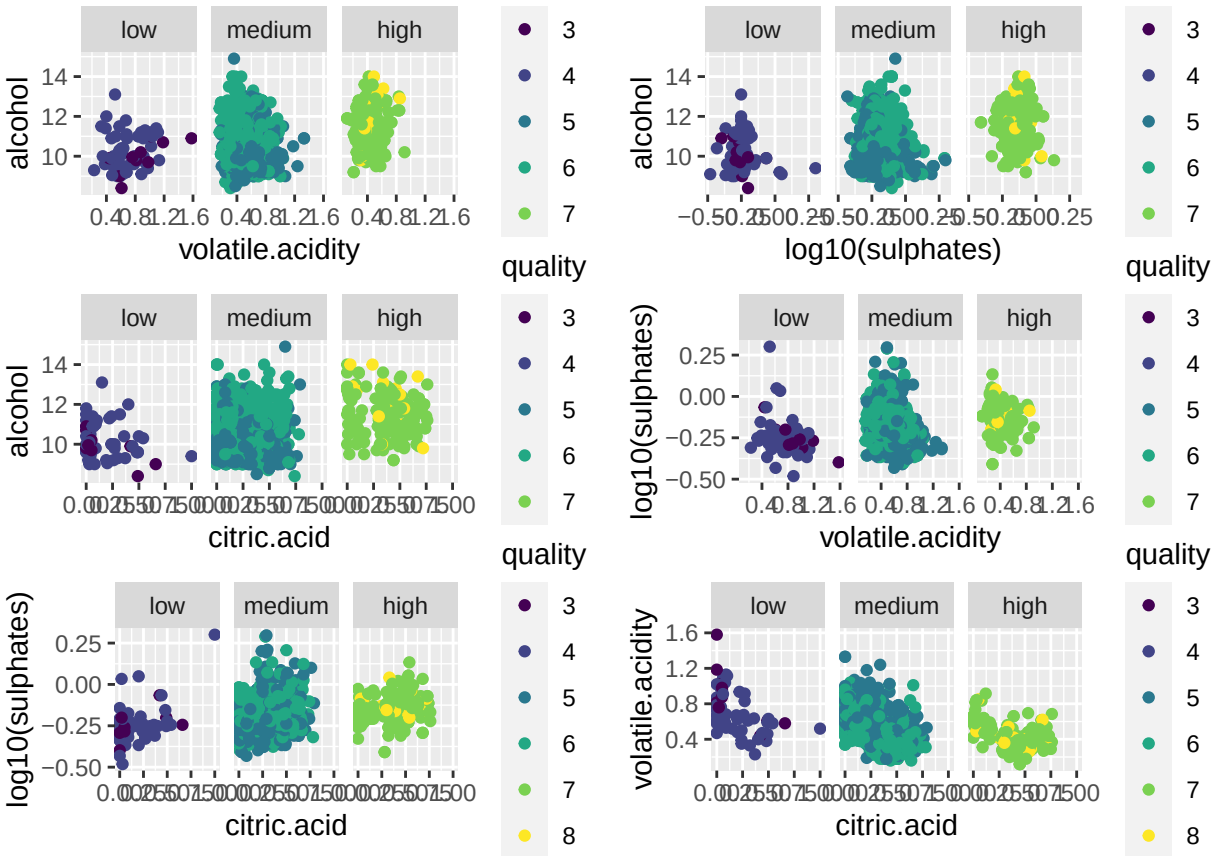
Podemos decir que las siguientes variables tienen correlaciones relativamente más altas con la calidad del vino:

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

Gráficas multicomparativas

En esta sección vamos a crear algunas gráficas de diferentes variables para investigar interacciones más complejas entre las variables que están más relacionadas con la calidad del vino.

```
vinos$quality <- factor(vinos$quality, ordered = T)
grid.arrange(
  ggplot(data = vinos, aes(x = volatile.acidity, y = alcohol)) +
    facet_wrap(~rating) +
    geom_point(aes(color = quality)),
  ggplot(data = vinos, aes(x = log10(sulphates), y = alcohol)) +
    facet_wrap(~rating) +
    geom_point(aes(color = quality)),
  ggplot(data = vinos, aes(x = citric.acid, y = alcohol)) +
    facet_wrap(~rating) +
    geom_point(aes(color = quality)),
  ggplot(data = vinos, aes(x = volatile.acidity, y = log10(sulphates))) +
    facet_wrap(~rating) +
    geom_point(aes(color = quality)),
  ggplot(data = vinos, aes(x = citric.acid, y = log10(sulphates))) +
    facet_wrap(~rating) +
    geom_point(aes(color = quality)),
  ggplot(data = vinos, aes(x = citric.acid, y = volatile.acidity)) +
    facet_wrap(~rating) +
    geom_point(aes(color = quality)),
  ncol = 2)
```



Estos diagramas de dispersión estaban abarrotados ya que más del 80% tienen una **quality** promedio, por lo que los clasificamos por **rating**. Ahora está más claro que los vinos de mayor calidad tienden a ser más altos en alcohol, ácido cítrico y sulfatos. Por otro lado, los vinos de mayor calidad tienden a tener menor acidez volátil.

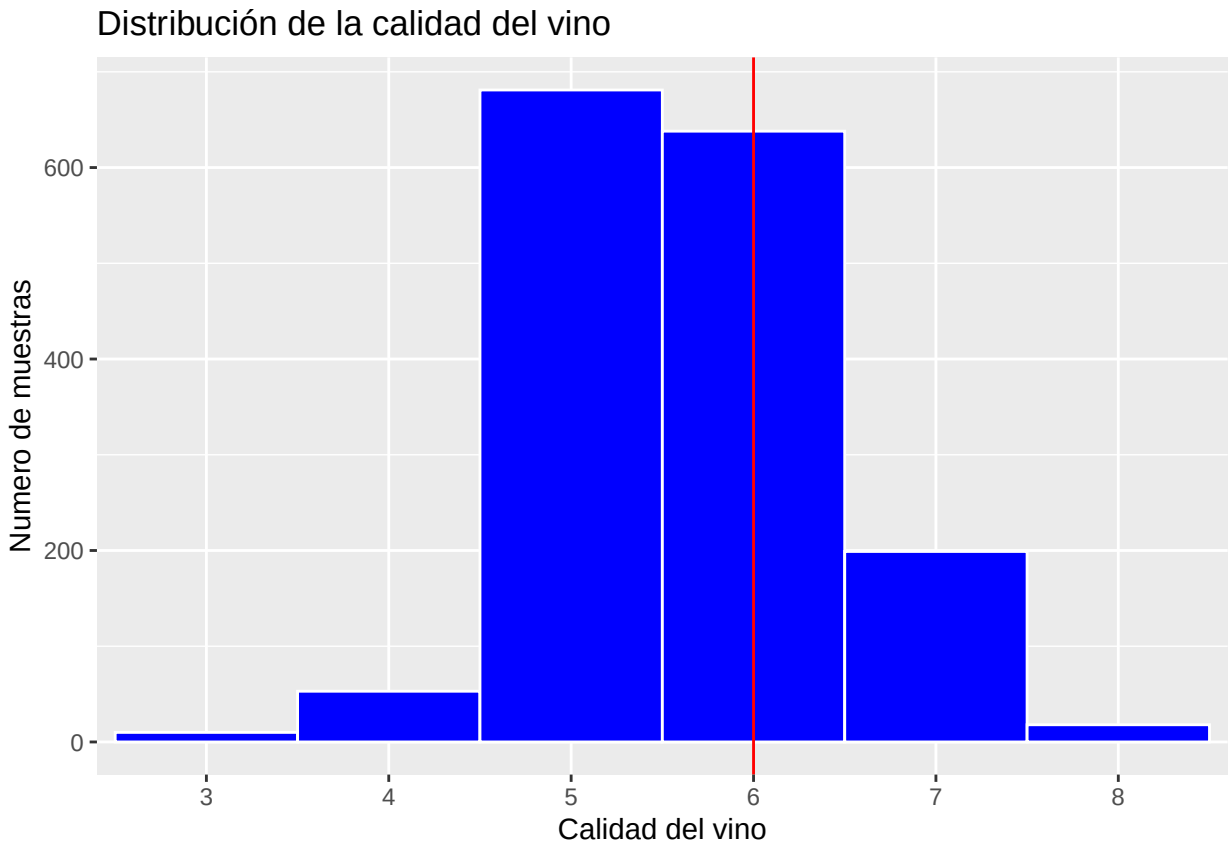
Conclusión del estudio de variables multiples

El alto contenido de alcohol contribuye a la buena calidad del vino, la adición de sulfatos o ácido cítrico influirá positivamente en la calidad del vino, mientras que la adición de ácido volátil influirá negativamente en la calidad del vino.

La gráfica de correlación mostraba que el ácido cítrico influye en la calidad del vino, pero de las gráficas anteriores, podemos observar que el ácido cítrico por sí solo no influye tanto en la calidad.

Quality of wine

```
ggplot(data = vinos, aes(x = quality)) +
  geom_bar(width = 1, color = 'white', fill = 'blue') +
  geom_vline(xintercept = median(as.numeric(vinos$quality)), color = "red") +
  labs(x = "Calidad del vino",
       y = "Numero de muestras",
       title = "Distribución de la calidad del vino")
```



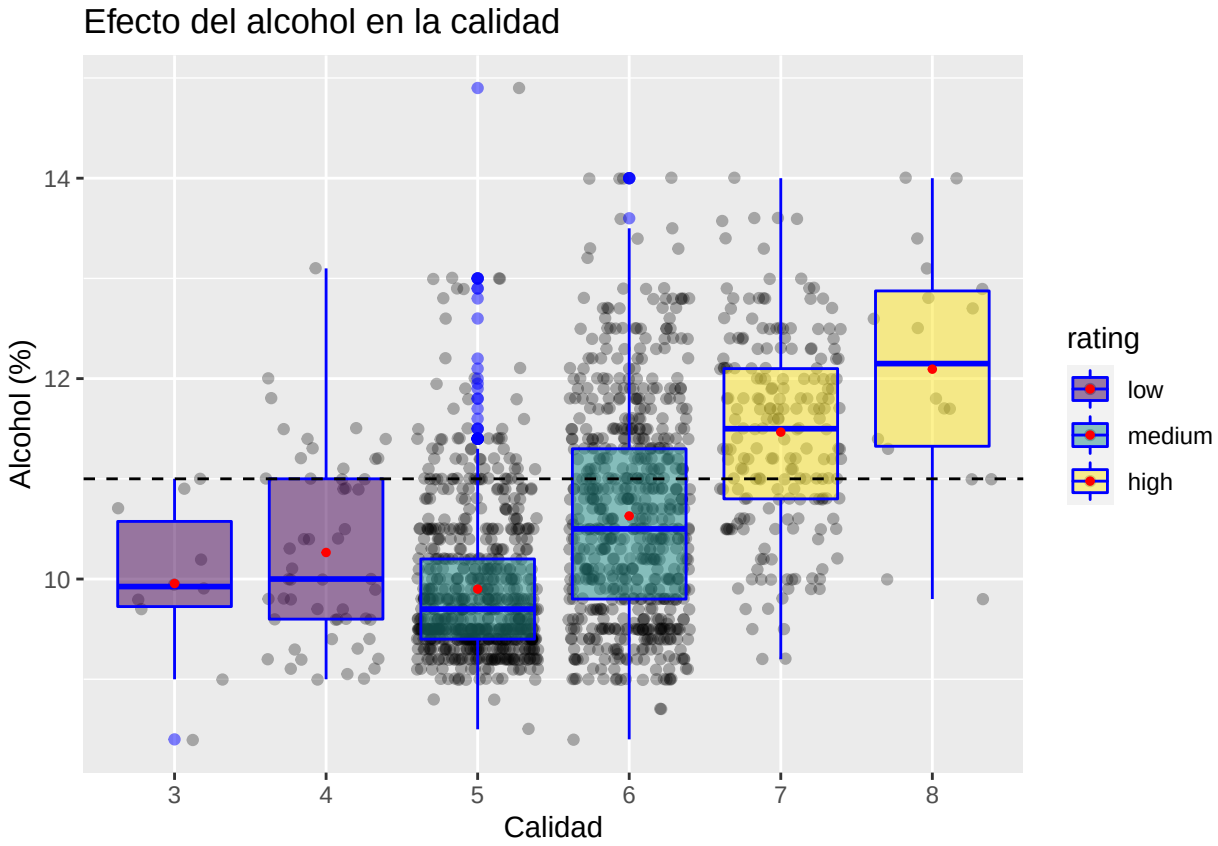

```
summary(vinos$quality)
```

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

Este gráfico explica que el 82,5% de los vinos en el conjunto de datos son de calidad 5 y 6. Como tenemos que encontrar la influencia de otras métricas en la calidad del vino, un conjunto de datos tan limitado hace que sea difícil entender qué hace que un buen vino sea bueno.

Efecto del Alcohol

```
ggplot(data = vinos, aes(x = factor(quality), y = alcohol, fill = rating)) +
  geom_jitter(alpha = .3)  +
  geom_boxplot(alpha = .5, color = 'blue')+
  stat_summary(fun = "mean",
               geom = "point",
               color = "red",
               size = 1) +
  geom_hline(yintercept = 11, linetype="dashed") +
  labs(x = "Calidad",
       y = "Alcohol (%)",
       title = "Efecto del alcohol en la calidad")
```




```
cor.test(vinos$alcohol, as.numeric(vinos$quality))
```

```
##
##  Pearson's product-moment correlation
##
## data:  vinos$alcohol and as.numeric(vinos$quality)
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

```
by(vinos$alcohol, vinos$rating, summary)
```

```
## vinos$rating: low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.60   10.00   10.22   11.00   13.10
## ------------------------------------------------------------
## vinos$rating: medium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.00   10.25   10.90   14.90
## ------------------------------------------------------------
## vinos$rating: high
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.60   11.52   12.20   14.00
```

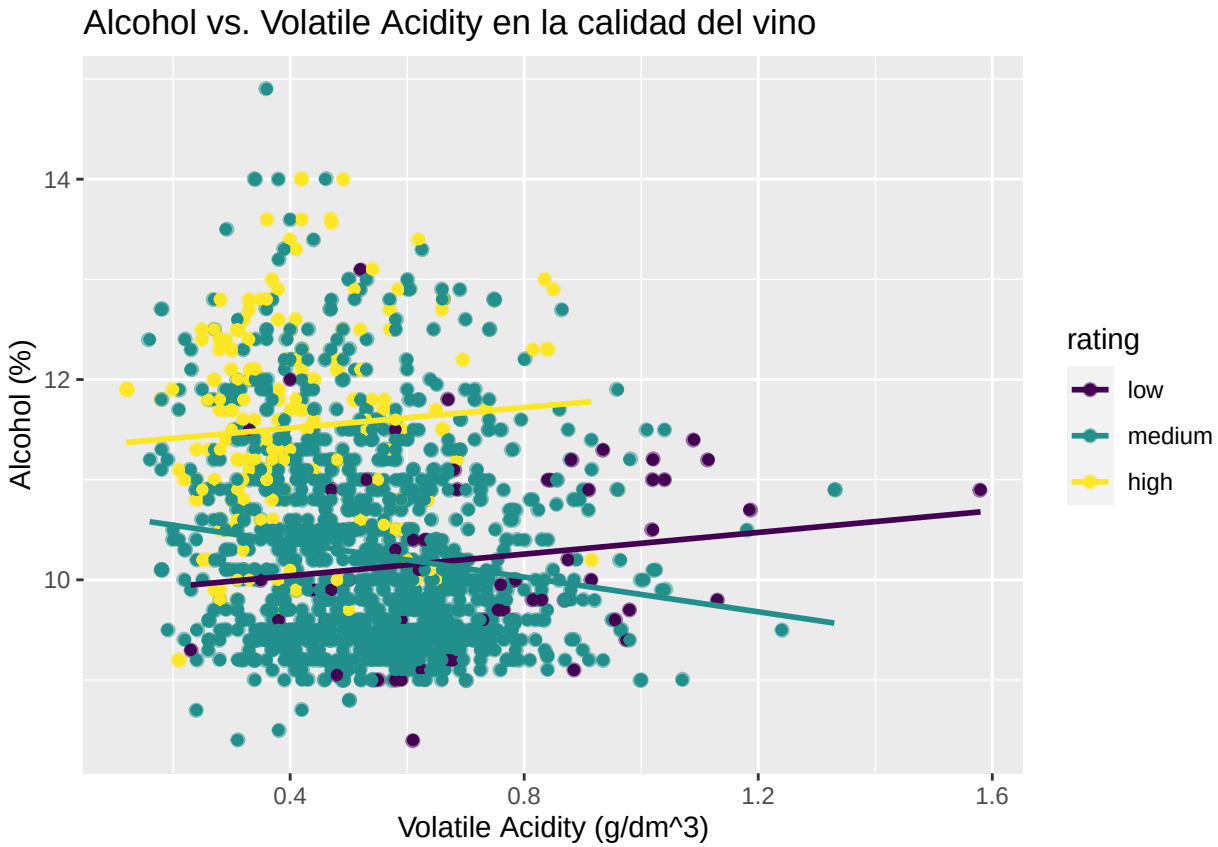
El alcohol tiene la correlación más fuerte con la calidad. A medida que aumenta el contenido alcohólico,

normalmente también aumenta la calidad del vino. El 75% de los buenos vinos contienen más del 11% de alcohol, mientras que el 75% de los de calidad media y mala tienen un porcentaje de alcohol inferior al 11%.

Alcohol y volatile acidity

```
ggplot(data = subset(vinos, rating != 'average'),
       aes(x = volatile.acidity, y = alcohol, color = rating)) +
  geom_jitter(size = 2, alpha = 1/2) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 1) +
  labs(x = "Volatile Acidity (g/dm^3)",
       y = "Alcohol (%)",
       title = "Alcohol vs. Volatile Acidity en la calidad del vino")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Creamos subconjuntos de los datos para quitar los vinos 'promedio'. La alta acidez volátil, con pocas excepciones, hace que la calidad del vino baje. Notamos que las líneas dejan ver con más claridad la relación alcohol y acidez volátil por calificación. El vino de alta calidad tiene una combinación de alto contenido de alcohol y baja acidez volátil.

6. Reflexión

A través de este análisis de datos exploratorio del conjunto de datos de vino tinto, podemos observar cómo el conocimiento del dominio es útil durante el proceso. Hemos revelado los facotres clave que afectan a la calidad del vino, principalmente: alcohol, sulfatos y acidez volátil, aunque los datos están limitados de 1599 observaciones. En ese conjunto de datos, el 82% de los vinos son de calidad media entre 5 y 6. Si pudiéramos

tener un conjunto de datos de más observaciones y una calidad uniforme de vinos sería posible realizar un mejor análisis.

```
# Guardando el fichero
write.csv(vinos ,"/Users/manutaberner/Google Drive/UOC/Tipologia y ciclos/PRAC2/Final/vinos_final.csv",
```

7. Enlace a Github y contribuciones.

Pincha aquí para acceder a Github

Contribuciones	Firma
Investigación previa	Manuel Taberner y Andrés Pérez
Redacción de las respuestas	Manuel Taberner y Andrés Pérez
Desarrollo código	Manuel Taberner y Andrés Pérez