TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS – PRÁCTICA 2

Carlos Giner Baixauli

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

El dataset seleccionado es winequality-red.csv y contiene información acerca de ciertos vinos tintos de Portugal. El dataset está disponible en https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009. Está estructurado en 1599 observaciones de 12 variables, que son las siguientes:

- fixed_acidity: acidez fija del vino
- volatile_acidity: acidez volátil del vino
- citric_acid: concentración de ácido cítrico
- residual_sugar: cantidad de azúcar residual
- chlorides: cantidad de cloruros
- free_sulfur_dioxide: cantidad de dióxido de azufre libre
- total_sulfur_dioxide: cantidad total de dióxido de azufre
- density: densidad del vino
- pH: pH del vino
- sulphates: cantidad de sulfatos
- alcohol: concentración de alcohol
- quality: calidad del vino valorada de 1 a 10

Todas las variables descritas son numéricas y la variable “quality” está estructurada en factores.

Por las variables que contiene este dataset, puede ser particularmente interesante enfocar el análisis de manera que podamos determinar qué variables son determinantes y en qué medida están relacionadas con la calidad del vino. Además, también nos permitirá plantear contrastes de hipótesis y construir un modelo de regresión lineal para predecir la calidad de un vino en función de sus características. Estos análisis pueden ser de gran relevancia en el sector enológico para optimizar la producción de vinos de alta calidad.

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

Para leer el dataset winequality-red.csv lo descargamos previamente de Kaggle y lo cargamos mediante la función read_csv de la librería readr. Para facilitar las operaciones renombramos las variables. En este caso concreto no es necesario realizar ninguna integración puesto que toda la información está integrada en un único fichero csv.

Obtenemos la estructura y la distribución de las variables con las funciones str y summary.

```
> str(data)
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and '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             : num  5 5 5 6 5 5 5 7 7 5 ...

> summary(data)
 fixed_acidity   volatile_acidity  citric_acid    residual_sugar     chlorides       free_sulfur_dioxide
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00
 total_sulfur_dioxide    density             pH          sulphates         alcohol         quality
 Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000
 Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Por el volumen del dataset consideramos que no es necesario seleccionar las variables a priori ni tomar muestras, ya que tan solo hay 1599 observaciones de 12 variables.

3. Limpieza de los datos.

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

Mediante el conteo de los valores que toma cada variable en el dataset con la función table(data$variable, useNA = "always") observamos que no existen elementos no informados, por lo que no tenemos que hacer la imputación de los mismos. De haber sido necesario hubiéramos optado por un método de imputación como kNN, más adecuado para este caso que otros métodos como la imputación con la media (que no tiene en cuenta la dispersión) o la supresión del registro (no recomendable en datasets con pocos registros).

No obstante, sí se da el valor cero en la variable “citric_acid”, pero este valor tiene sentido según la definición puesto que esta variable mide la concentración y pueden existir vinos sin ácido cítrico. Por este motivo dejaremos estos ceros como valores informados.

3.2. Identificación y tratamiento de valores extremos.

Si obtenemos los outliers de las variables con boxplot.stats(data$variable)$out, observamos la presencia de algunos valores de este tipo. No obstante, comprobamos que se trata de valores que, aunque están algo alejados son valores válidos, por lo que es conveniente dejarlos en el dataset para la realización del análisis.

4. Análisis de los datos.

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

Para seleccionar las variables más relevantes haremos primero un análisis de la correlación con las funciones cor y corrplot (ver apartado 5 de la práctica).

Seleccionamos las variables “citric_acid”, “residual_sugar”, “total_sulfur_dioxide”, “sulphates” y “alcohol” ya que al estar algunas de ellas fuertemente correlacionadas con las restantes, no hay redundancias ni mucha pérdida de información por eliminarlas. También seleccionamos “quality” puesto que es la variable objetivo del análisis.

En los siguientes apartados analizaremos si la calidad del vino es más alta cuando aumenta la concentración de alcohol, averiguaremos en qué medida están correlacionadas las variables con la calidad y construiremos un modelo de regresión lineal para predecir la calidad conociendo las variables más relevantes.

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

Para realizar las pruebas de normalidad planteamos un contraste de hipótesis en el que la hipótesis nula es que la muestra proviene de una distribución normal y la hipótesis alternativa es que la muestra no proviene de una distribución normal, por lo que consideraremos que existe normalidad si el p-valor es mayor a 0.05 (rechazamos la hipótesis nula).

Utilizamos el test de normalidad de Shapiro-Wilk con la función shapiro.test y obtenemos que las variables analizadas no provienen de una distribución normal por tener p-valores próximos a cero.

A continuación, vamos a comprobar la homogeneidad de la varianza. Puesto que no se cumple la normalidad, utilizamos el test no paramétrico de Fligner-Killeen con la función fligner.test aplicada a la calidad y a concentraciones de alcohol de menos de 10 y más de 10 grados. En este caso la hipótesis nula es que las varianzas de los grupos son homogéneas. Puesto que el p-valor obtenido es muy cercano a 0, concluimos que las varianzas de los dos grupos no son homogéneas.

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.

En primer lugar, vamos a analizar si la calidad del vino es más alta cuando la concentración de alcohol supera 10 grados. Como hemos visto en el apartado anterior, no se cumple la propiedad

de homocedasticidad de la varianza, por lo que aplicaremos el test no paramétrico de Kruskal-Wallis con la función kruskal.test. Planteamos la hipótesis nula de que las medias de los grupos son iguales y la hipótesis alternativa de que son diferentes. El p-valor obtenido es mucho menor que 0.05, por lo que hay evidencias para suponer que existen diferencias entre las medias de los grupos.

En el siguiente test vamos a averiguar en qué medida están correlacionadas las variables con la calidad. Para ello calculamos la matriz de correlación con la función cor y obtenemos los siguientes resultados:

```
                         quality
citric_acid           0.22637251
residual_sugar        0.01373164
total_sulfur_dioxide -0.18510029
sulphates             0.25139708
alcohol               0.47616632
quality               1.00000000
```

También comprobamos que, excepto "residual_sugar", estos valores presentan p-valores muy próximos a cero con la función rcorr de la librería Hmisc

Así comprobamos que algunas variables como la concentración de alcohol y en menor medida otras como la concentración de ácido cítrico o la cantidad de sulfatos están correlacionadas con la calidad. También observamos que la cantidad total de dióxido de azufre presenta una correlación inversa con la calidad.

Finalmente vamos a construir un modelo lineal con la función lm, con las variables seleccionadas. Aunque no todas están correlacionadas con la calidad las vamos a incluir puesto que pueden estarlo en combinación con otras de las variables.

```
Call:
lm(formula = quality ~ citric_acid + residual_sugar + total_sulfur_dio
xide +
    sulphates + alcohol, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.76283 -0.36450 -0.06842  0.49016  2.14048

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.7253318  0.1843294   9.360  < 2e-16 ***
citric_acid           0.5341553  0.0931500   5.734 1.17e-08 ***
residual_sugar       -0.0006866  0.0124765  -0.055    0.956
total_sulfur_dioxide -0.0027011  0.0005417  -4.987 6.81e-07 ***
sulphates             0.8396698  0.1059946   7.922 4.36e-15 ***
alcohol               0.3205012  0.0164956  19.429  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.679 on 1593 degrees of freedom
Multiple R-squared:  0.2952,   Adjusted R-squared:  0.293
F-statistic: 133.5 on 5 and 1593 DF,  p-value: < 2.2e-16
```

Como observamos en los resultados, "residual_sugar" no es significativo en el modelo, por lo que vamos a construir un segundo modelo sin esta variable.

```
Call:
lm(formula = quality ~ citric_acid + total_sulfur_dioxide + sulphates
+
```

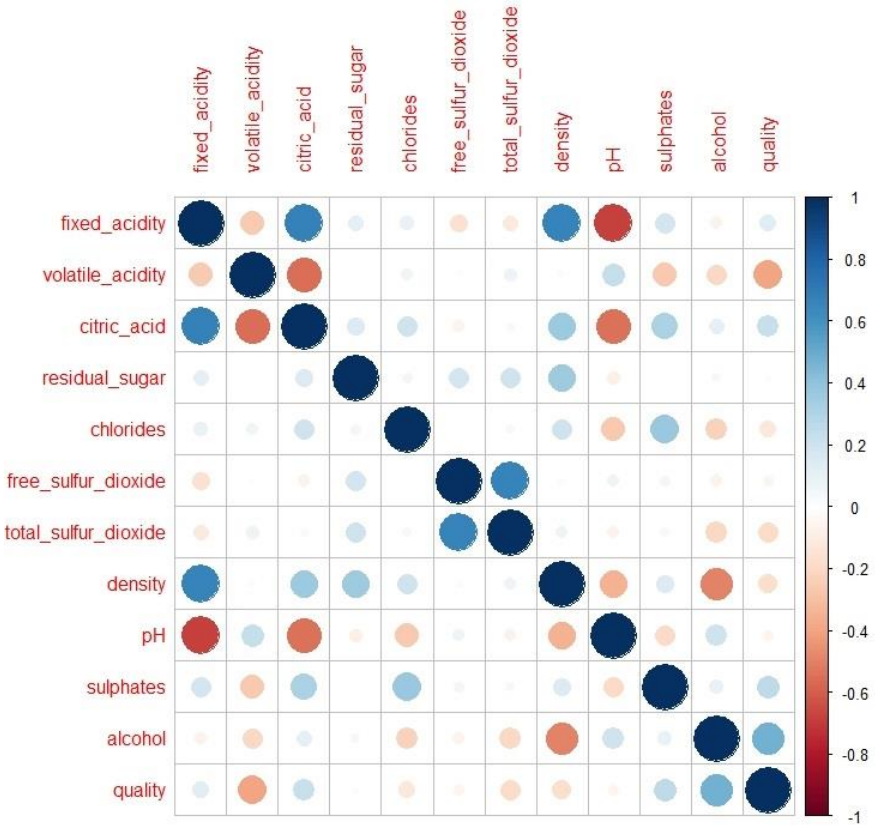
```
    alcohol, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.76325 -0.36539 -0.06773  0.48909  2.14083

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.724585   0.183772   9.384  < 2e-16 ***
citric_acid           0.533434   0.092194   5.786 8.66e-09 ***
total_sulfur_dioxide -0.002707   0.000529  -5.118 3.47e-07 ***
sulphates             0.839991   0.105801   7.939 3.80e-15 ***
alcohol               0.320432   0.016443  19.487  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6788 on 1594 degrees of freedom
Multiple R-squared:  0.2952,   Adjusted R-squared:  0.2934
F-statistic: 166.9 on 4 and 1594 DF,  p-value: < 2.2e-16
```

Hemos conseguido un ligero aumento de la precisión, aunque por la tipología de los datos tal vez sería más adecuado utilizar modelos de regresión más complejos.

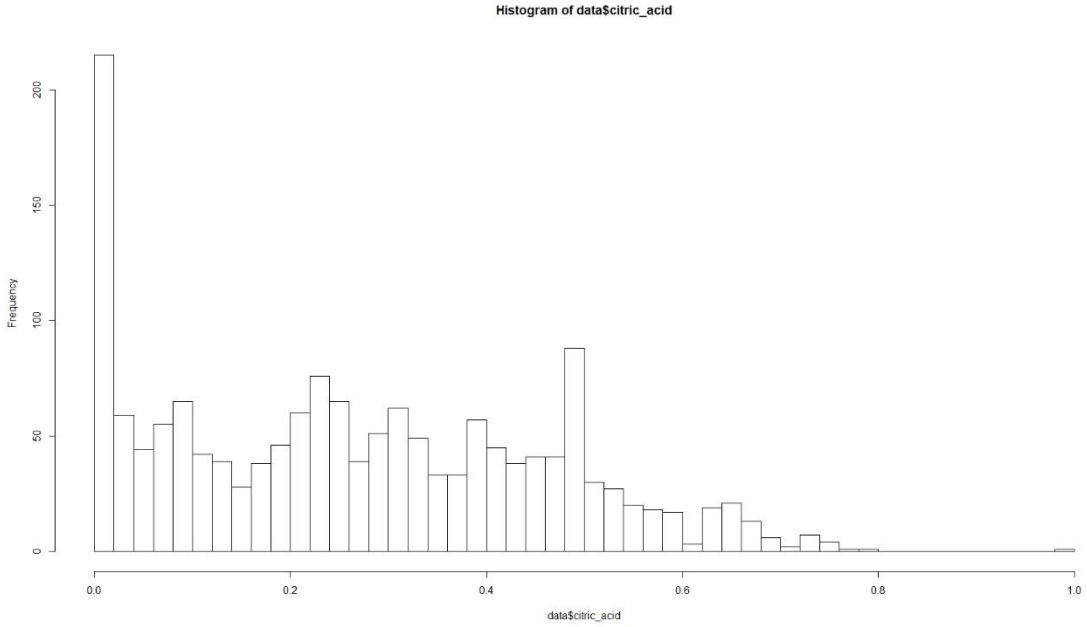
5. Representación de los resultados a partir de tablas y gráficas.

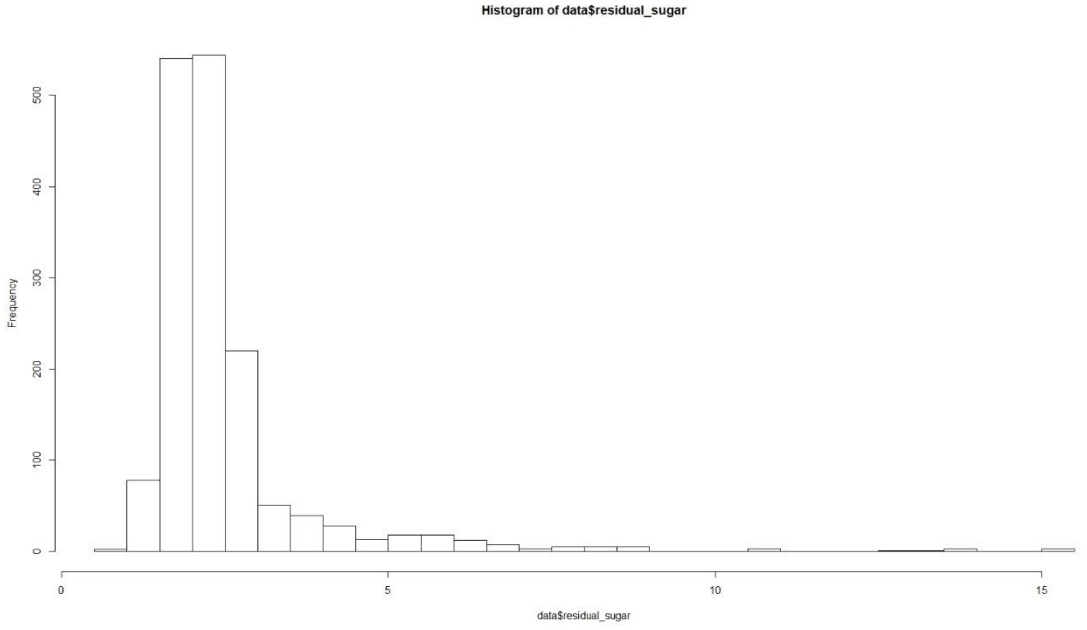
Mostramos los gráficos obtenidos en los análisis anteriores:

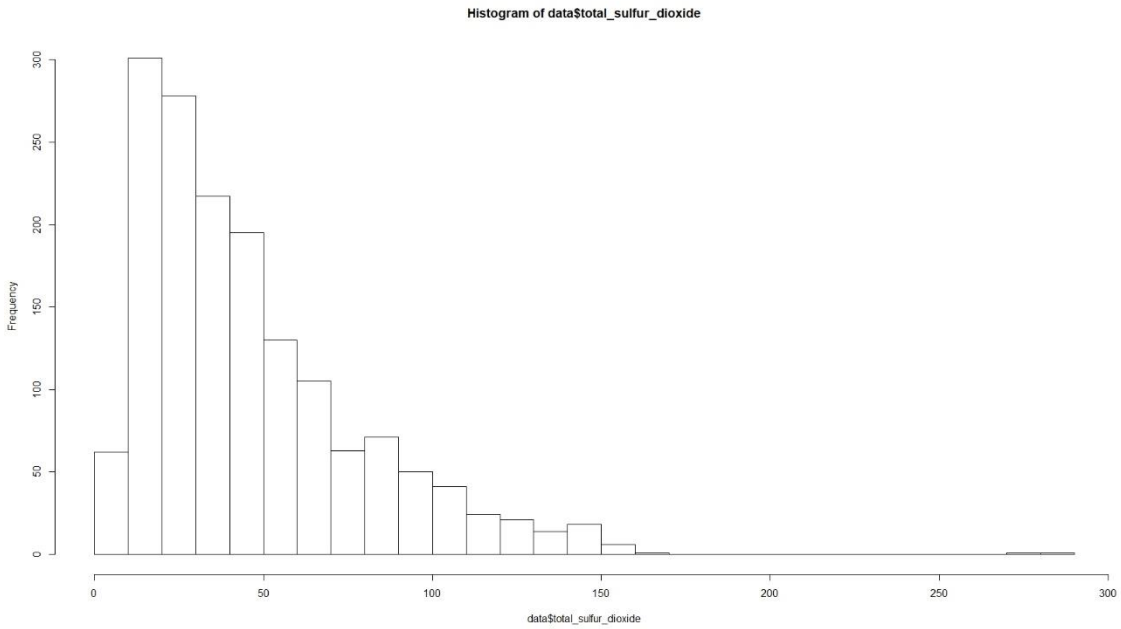
Análisis de correlación entre las variables del dataset:

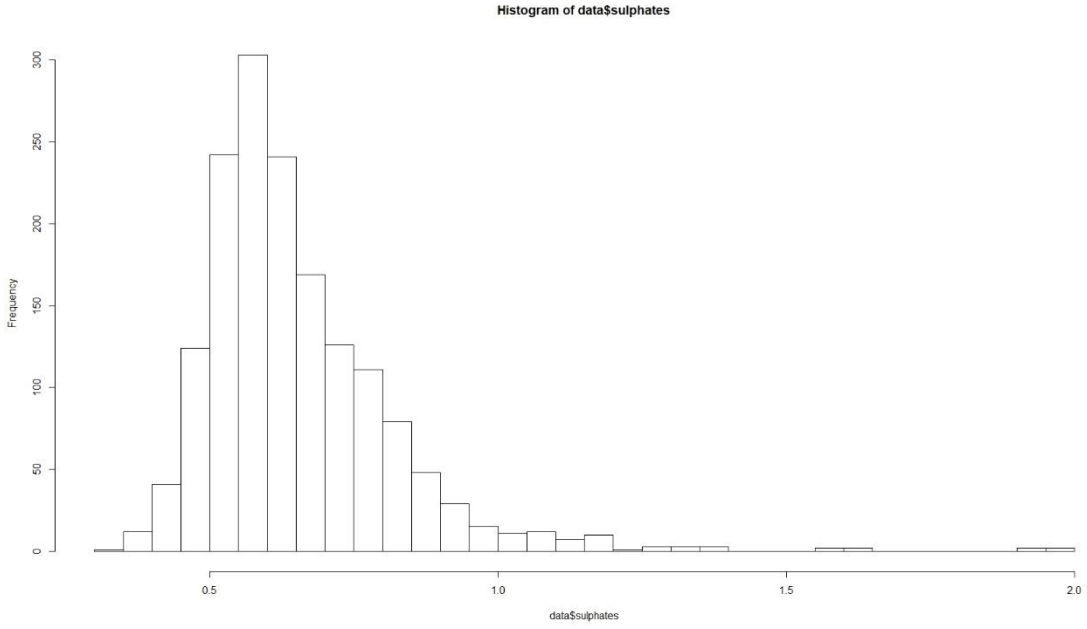


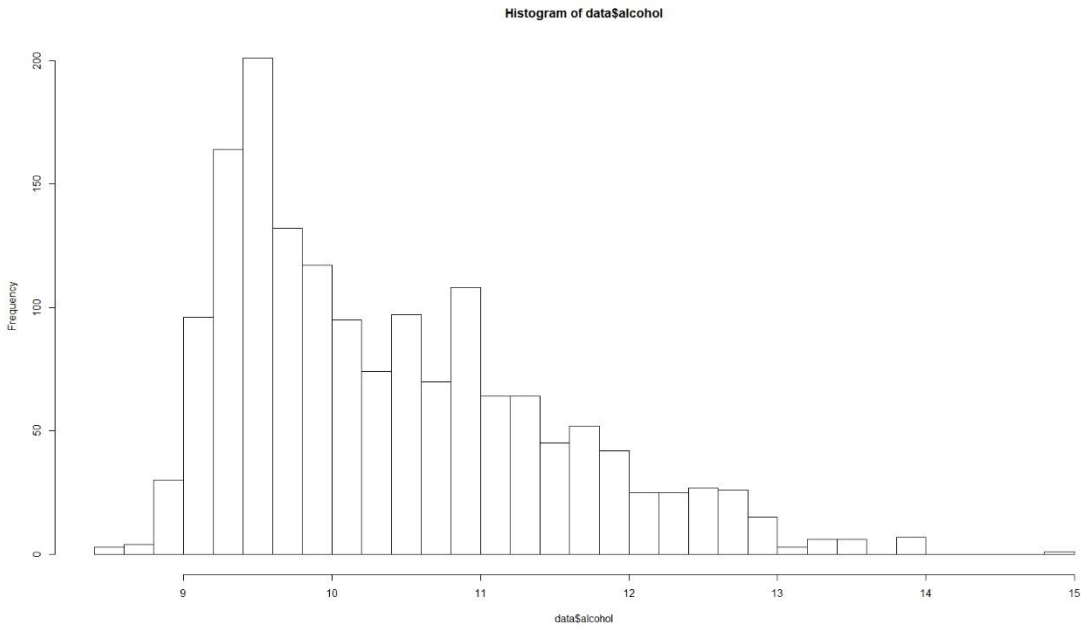
Análisis visual de la normalidad de las variables seleccionadas:






Histogram of data$total_sulfur_dioxide
Frequency
0
50
100
150
200
250
300
data$total_sulfur_dioxide


Histogram of data$sulphates
Frequency
0
50
100
150
200
250
300
0.5
1.0
1.5
2.0
data$sulphates


Histogram of data$alcohol
Frequency
0
50
100
150
200
9
10
11
12
13
14
15
data$alcohol

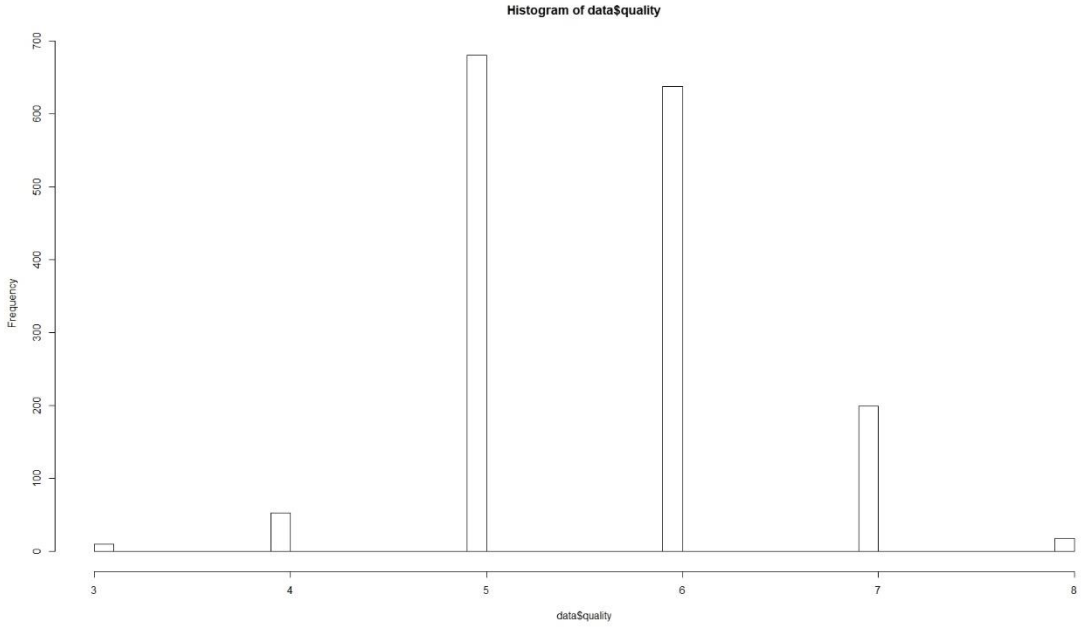
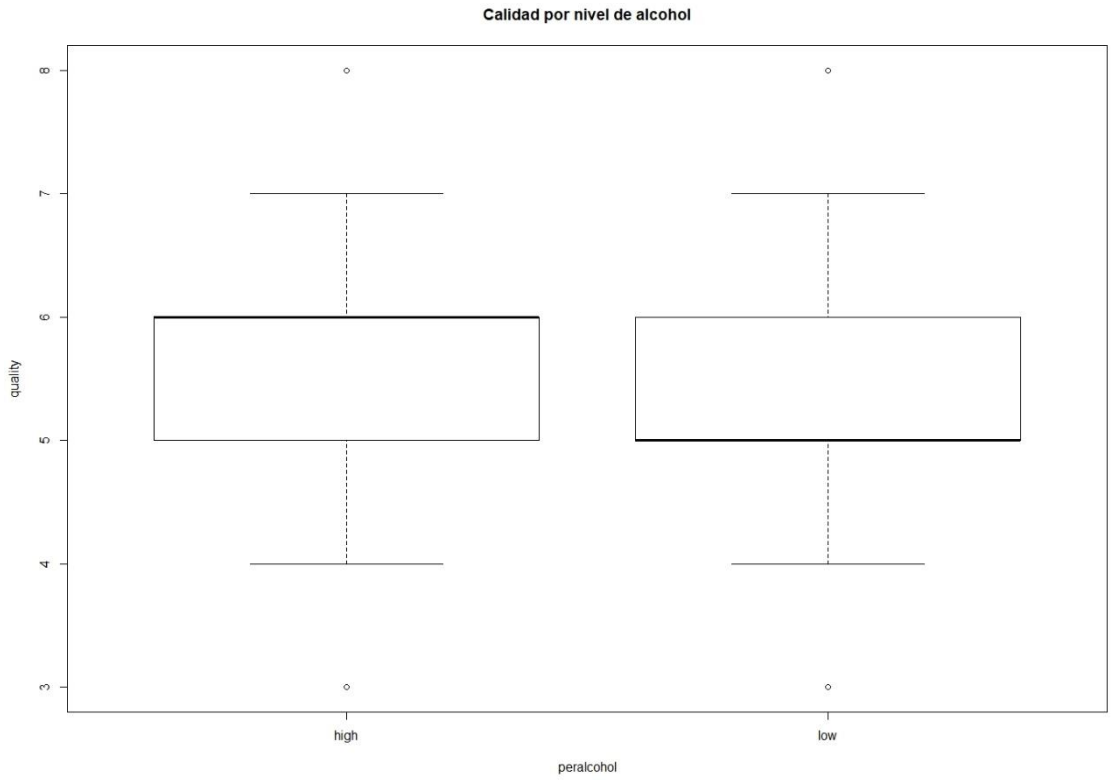

Histogram of data$quality
Frequency
0
100
200
300
400
500
600
700
3
4
5
6
7
8
data$quality

Diagrama de cajas con la calidad por grupos de nivel de alcohol (bajo si es menor de 10 grados y alto si es mayor).



También recopilamos aquí las tablas obtenidas durante el análisis:

Correlación de la calidad con el resto de variables seleccionadas:

```
                        quality
citric_acid           0.22637251
residual_sugar        0.01373164
total_sulfur_dioxide -0.18510029
sulphates             0.25139708
alcohol               0.47616632
quality               1.00000000
```

Resultados del primer modelo de regresión lineal generado:

```
Call:
lm(formula = quality ~ citric_acid + residual_sugar + total_sulfur_dio
xide +
    sulphates + alcohol, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.76283 -0.36450 -0.06842  0.49016  2.14048

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.7253318  0.1843294   9.360  < 2e-16 ***
citric_acid           0.5341553  0.0931500   5.734 1.17e-08 ***
residual_sugar       -0.0006866  0.0124765  -0.055    0.956
total_sulfur_dioxide -0.0027011  0.0005417  -4.987 6.81e-07 ***
sulphates             0.8396698  0.1059946   7.922 4.36e-15 ***
alcohol               0.3205012  0.0164956  19.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.679 on 1593 degrees of freedom
Multiple R-squared:  0.2952,  Adjusted R-squared:  0.293
F-statistic: 133.5 on 5 and 1593 DF,  p-value: < 2.2e-16
```

Resultados del primer modelo de regresión lineal generado:

```
Call:
lm(formula = quality ~ citric_acid + total_sulfur_dioxide + sulphates
+
    alcohol, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.76325 -0.36539 -0.06773  0.48909  2.14083

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.724585   0.183772   9.384  < 2e-16 ***
citric_acid           0.533434   0.092194   5.786 8.66e-09 ***
total_sulfur_dioxide -0.002707   0.000529  -5.118 3.47e-07 ***
sulphates             0.839991   0.105801   7.939 3.80e-15 ***
alcohol               0.320432   0.016443  19.487  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6788 on 1594 degrees of freedom
Multiple R-squared:  0.2952,  Adjusted R-squared:  0.2934
F-statistic: 166.9 on 4 and 1594 DF,  p-value: < 2.2e-16
```

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

Como conclusión, el dataset analizado no presentaba ceros ni valores vacíos, no obstante, los datos no pertenecen a una población con distribución normal ni tampoco hay homogeneidad en la varianza.

En primer lugar, hemos analizado si la calidad del vino es más alta cuando la concentración de alcohol supera 10 grados. Mediante el test de Kruskal-Wallis concluimos que existen diferencias significativas entre las medias de los grupos y por tanto la calidad del vino es mayor al aumentar la concentración de alcohol.

En el test de correlación hemos comprobado que, excepto “residual_sugar”, las demás variables presentan cierta correlación con la calidad. Algunas variables como la concentración de alcohol y en menor medida otras como la concentración de ácido cítrico o la cantidad de sulfatos presentan esta correlación. También observamos que la cantidad total de dióxido de azufre presenta una correlación inversa con la calidad.

Finalmente hemos construido un modelo lineal con las variables seleccionadas. Puesto que, “residual_sugar” no es significativo en el modelo, construimos un segundo modelo sin esta variable en el que conseguimos un ligero aumento de la precisión, aunque pensamos que por la tipología de los datos tal vez sería más adecuado utilizar modelos de regresión más complejos.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

```
library(readr)

library(corrplot)

library(Hmisc)


data <- read_csv("C:/Users/Carlos/Desktop/Tipología y ciclo de vida de los datos/PRA2/winequality-red.csv")


colnames(data) <- c("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar",

                    "chlorides", "free_sulfur_dioxide", "total_sulfur_dioxide", "density",

                    "pH", "sulphates", "alcohol", "quality")



# Descripción del dataset

dim(data)

str(data)

summary(data)



# Comprobamos que no hay NA

table(data$fixed_acidity, useNA = "always")

table(data$volatile_acidity, useNA = "always")

table(data$citric_acid, useNA = "always")

table(data$residual_sugar, useNA = "always")

table(data$chlorides, useNA = "always")

table(data$free_sugar_dioxide, useNA = "always")

table(data$total_sulfur_dioxide, useNA = "always")

table(data$density, useNA = "always")

table(data$pH, useNA = "always")

table(data$sulphates, useNA = "always")
```

```
table(data$alcohol, useNA = "always")
table(data$quality, useNA = "always")

# Comprobación de outliers
boxplot.stats(data$fixed_acidity)$out
boxplot.stats(data$volatile_acidity)$out
boxplot.stats(data$citric_acid)$out
boxplot.stats(data$residual_sugar)$out
boxplot.stats(data$chlorides)$out
boxplot.stats(data$free_sugar_dioxide)$out
boxplot.stats(data$total_sulfur_dioxide)$out
boxplot.stats(data$density)$out
boxplot.stats(data$pH)$out
boxplot.stats(data$sulphates)$out
boxplot.stats(data$alcohol)$out
boxplot.stats(data$quality)$out

# Análisis de la correlación
M <- cor(data)
corrplot(M, method="circle")

# Selección de variables
data <- data[,c("citric_acid", "residual_sugar",
"total_sulfur_dioxide", "sulphates", "alcohol", "quality")]

# Pruebas de normalidad
shapiro.test(data$citric_acid)
shapiro.test(data$residual_sugar)
shapiro.test(data$total_sulfur_dioxide)
shapiro.test(data$sulphates)
shapiro.test(data$alcohol)
shapiro.test(data$quality)
```

```
# Pruebas de homogeneidad

data$peralcohol <- ifelse(data$alcohol >= 10, "high", "low")

fligner.test(quality ~ peralcohol, data)


# Test de calidad por grado de alcohol

kruskal.test(quality ~ peralcohol, data)

data$peralcohol <- NULL


# Test de correlación

cor(data)

mcor <- rcorr(as.matrix(data), type="pearson")

View(mcor$P)


# Modelo de regresión lineal

model1 <- lm(quality ~ citric_acid + residual_sugar +
total_sulfur_dioxide + sulphates + alcohol, data)

summary(model1)

model2 <- lm(quality ~ citric_acid + total_sulfur_dioxide +
sulphates + alcohol, data)

summary(model2)


# Representación gráfica de los resultados

corrplot(M, method="circle")


data$peralcohol <- ifelse(data$alcohol >=10, "high", "low")

boxplot(quality ~ peralcohol, data, main="Calidad por nivel de
alcohol")


hist(data$citric_acid, breaks = 40)

hist(data$residual_sugar, breaks = 40)

hist(data$total_sulfur_dioxide, breaks = 40)

hist(data$sulphates, breaks = 40)

hist(data$alcohol, breaks = 40)

hist(data$quality, breaks = 40)
```

Contribuciones	Firma
Investigación previa	CGB
Redacción de las respuestas	CGB
Desarrollo código	CGB