Tipología y ciclo de vida de los datos

Práctica 2

Mar Bonora Ortega

Máster Ciencia de Datos

Universitat Oberta de Catalunya

1. Descripción del dataset

El dataset elegido es "Red Wine Quality", disponible en Kaggle. (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Dicho conjunto de datos contiene información respecto a diferentes tipos de vinos, como la densidad, el pH, el azúcar residual, etc. Además, también contiene un valor para la calidad, con lo cual, lo que se pretende con el estudio de este conjunto de datos, es intentar clasificar los diferentes vinos en función de sus características, así como poder predecir la calidad de otros vinos que no forman parte del conjunto en la actualidad, en función de dichas características.

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

Como se puede observar en el archivo R markdown donde se desarrolla el estudio de datos, se ha creado un dataframe que almacena los datos sobre la calidad de vino tinto, por medio de la carga del archivo csv que nos hemos descargado de Kaggle.

A continuación podemos visualizar las primeras filas del dataframe obtenido tras cargar los datos.

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>	free.sulfur.dioxide <dbl>
1	7.4	0.70	0.00	1.9	0.076	11
2	7.8	0.88	0.00	2.6	0.098	25
3	7.8	0.76	0.04	2.3	0.092	15
4	11.2	0.28	0.56	1.9	0.075	17
5	7.4	0.70	0.00	1.9	0.076	11
6	7.4	0.66	0.00	1.8	0.075	13

6 rows | 1-7 of 12 columns

3. Limpieza de los datos.

El siguiente paso consiste en explorar los datos, de forma que podamos detectar incongruencias o errores y limpiarlos.

Disponemos de 1599 registros o filas y 12 variables o columnas. Todas las variables son numéricas como podemos ver en la captura siguiente.

```
'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 ...
```

La lista de variables que componen el conjunto de datos es la siguiente:

- Fixed acidity (num): acidez fija del vino, es decir, ácidos que no se evaporan rápidamente.
- Volatile acidity (num): acidez volátil del vino, que en grandes cantidades puede dar sabor avinagrado
- Citric acid (num): ácido cítrico, que en pequeñas cantidades puede añadir "frescura" y sabor al vino
- Residual sugar (num): cantidad de azúcar después de la fermentación
- Chlorides (num): cantidad de sal en el vino
- Free sulfur dioxide (num): cantidad de forma libre de SO2 en el vino
- Total sulfur dioxide (num): cantidad de formas libres y ligadas de SO2 en el vino
- Density (num): densidad del vino
- pH (num): nivel de pH, cuán ácido o básico es el vino
- Sulphates (num): cantidad de sulfatos en el vino
- Alcohol (num): porcentaje de alcohol en el vino
- Quality (int): puntuación del vino (entre 0 y 10), se considera la variable de salida

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

Por medio de las funciones summary() y missing() hemos podido comprobar que no hay valores nulos ni missing en el dataset. Es por ello que no hemos tenido que gestionar esta situación, que normalmente se lleva a cabo por medio de impugnación de valores.

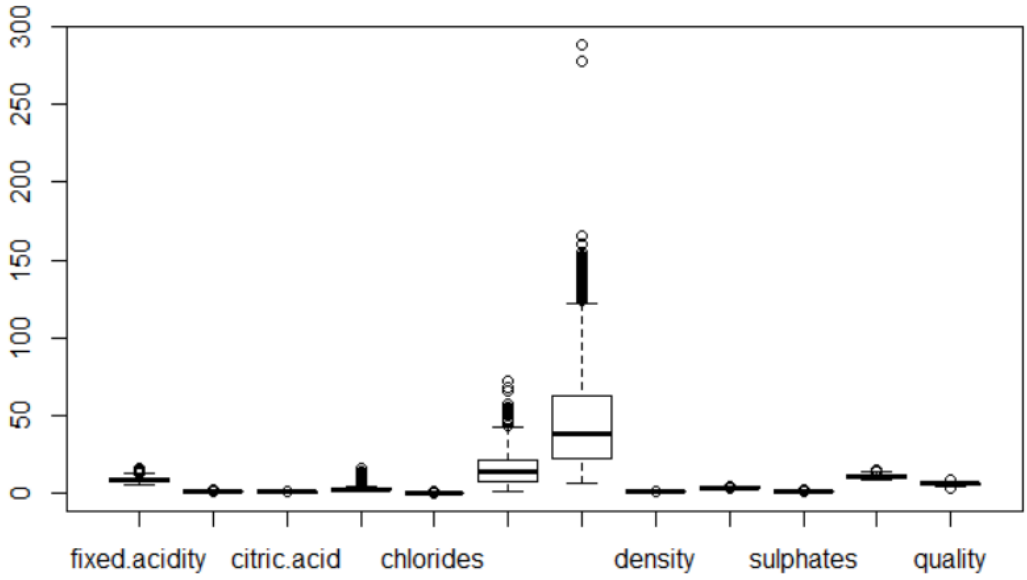
Podemos confirmar esta información visualizando el resultado obtenido.

```
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
```

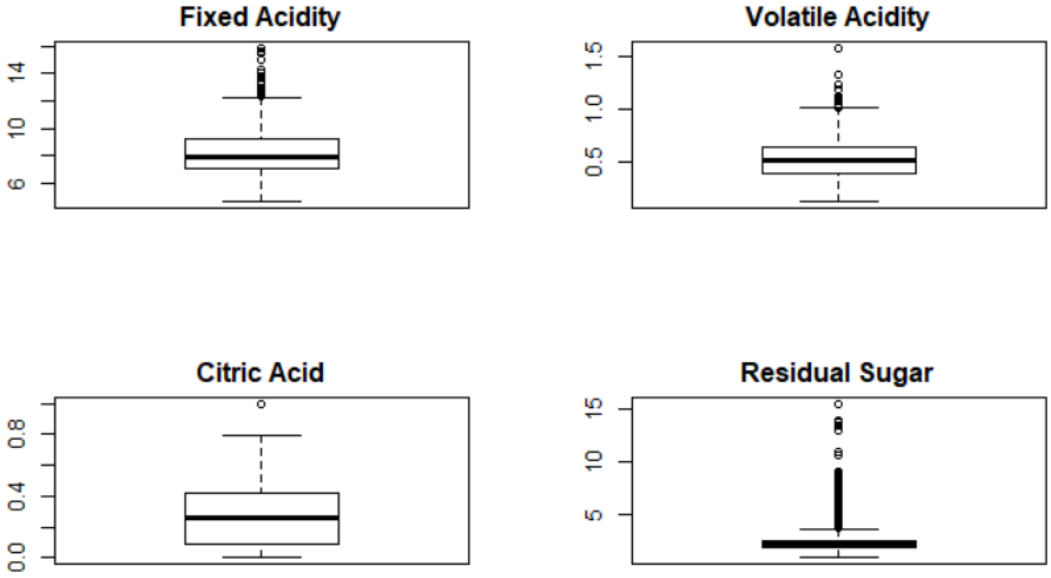
3.2. Identificación y tratamiento de valores extremos.

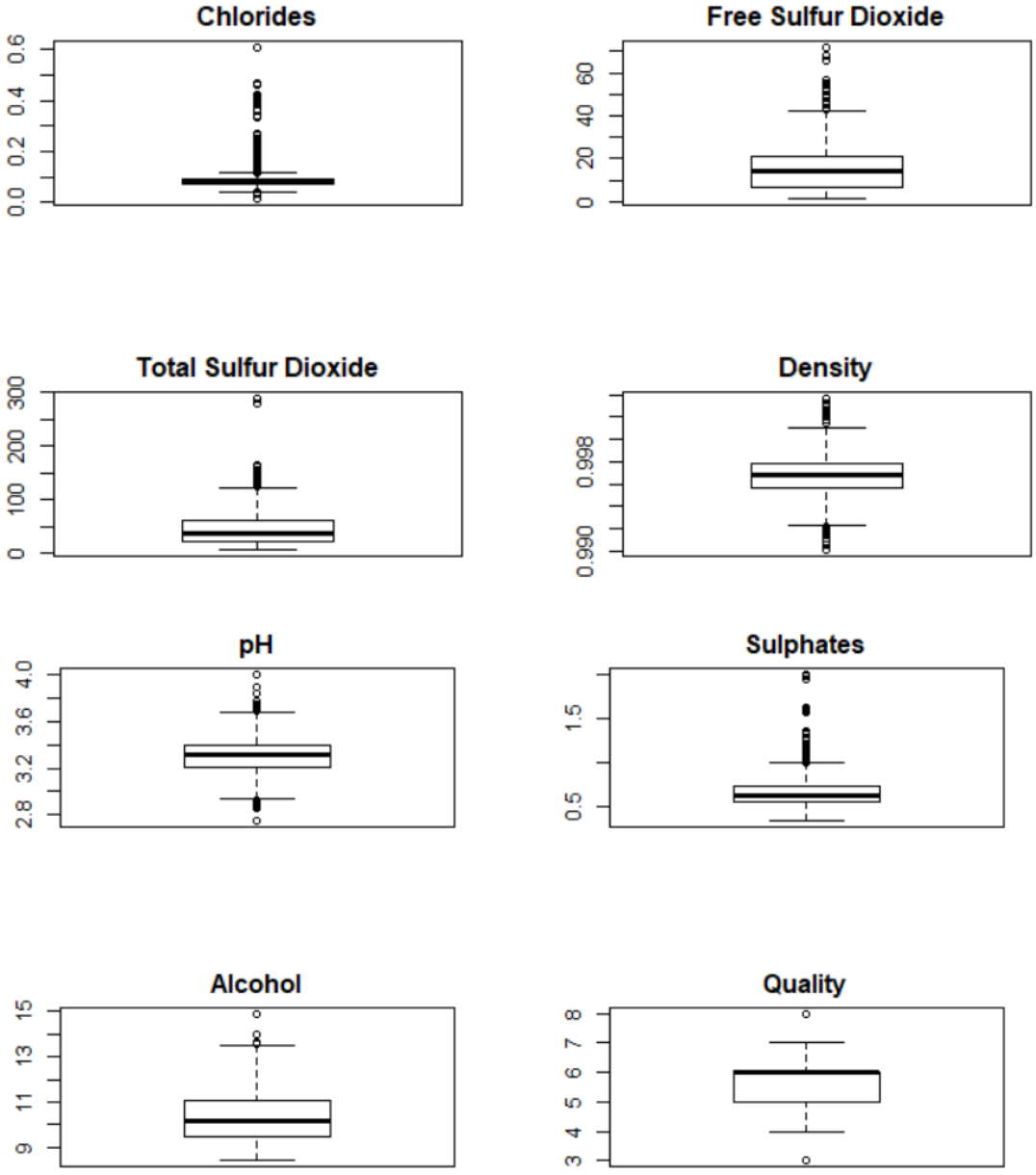
De la misma forma, gracias a la generación de gráficos de cajas y bigotes, hemos podido observar visualmente si existían valores extremos. Primero hemos visualizado todas las

variables de forma conjunta pero, como podremos ver a continuación, es complicado obtener información de este gráfico.



Por ello hemos generado un gráfico individual para cada variable.

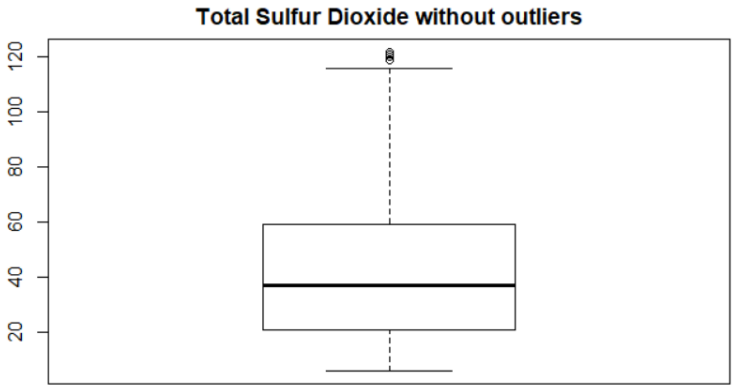


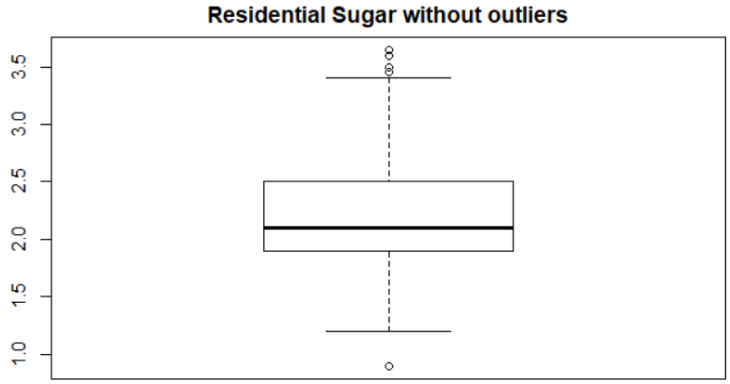


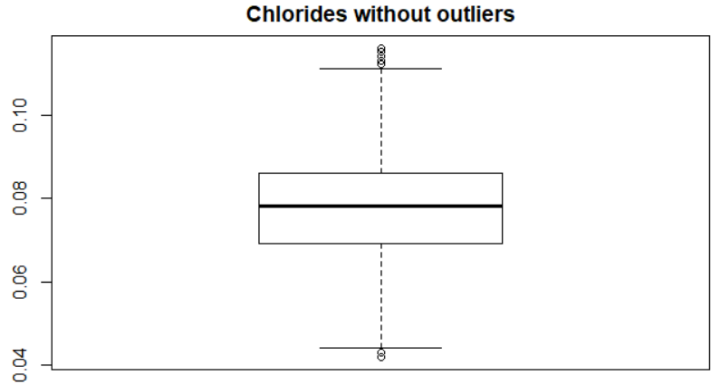
Aunque en la mayoría de variables había valores por fuera de los bigotes, con lo cual serían valores extremos, solo hemos desechado los valores extremos de las siguientes variables: *total sulfur dioxide, residential sugar, chlorides* y *sulphates*. Hemos elegido estas variables para tratar los outliers, ya que hemos considerado que los valores de dichos *outliers* sí que distaban mucho de los valores centrales y por lo tanto iban a ser significativos para el desarrollo del estudio de los datos.

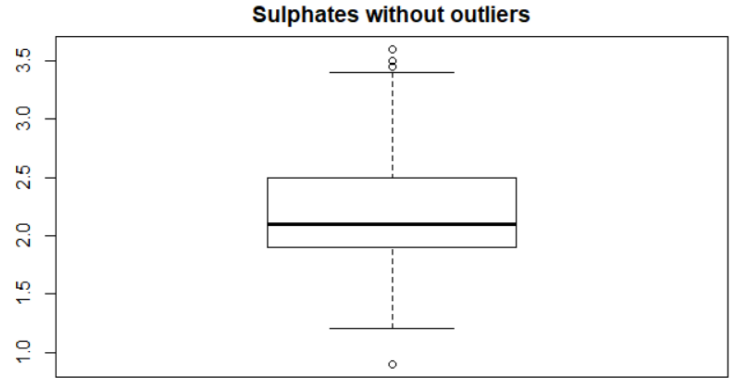
Una vez hechas estas modificaciones, podemos decir que tenemos un dataframe de 12 variables numéricas sin valores nulo, sin valores missing y sin valores extremos de importancia.

A continuación veremos el resultado de las correcciones realizadas.









4. Análisis de los datos.

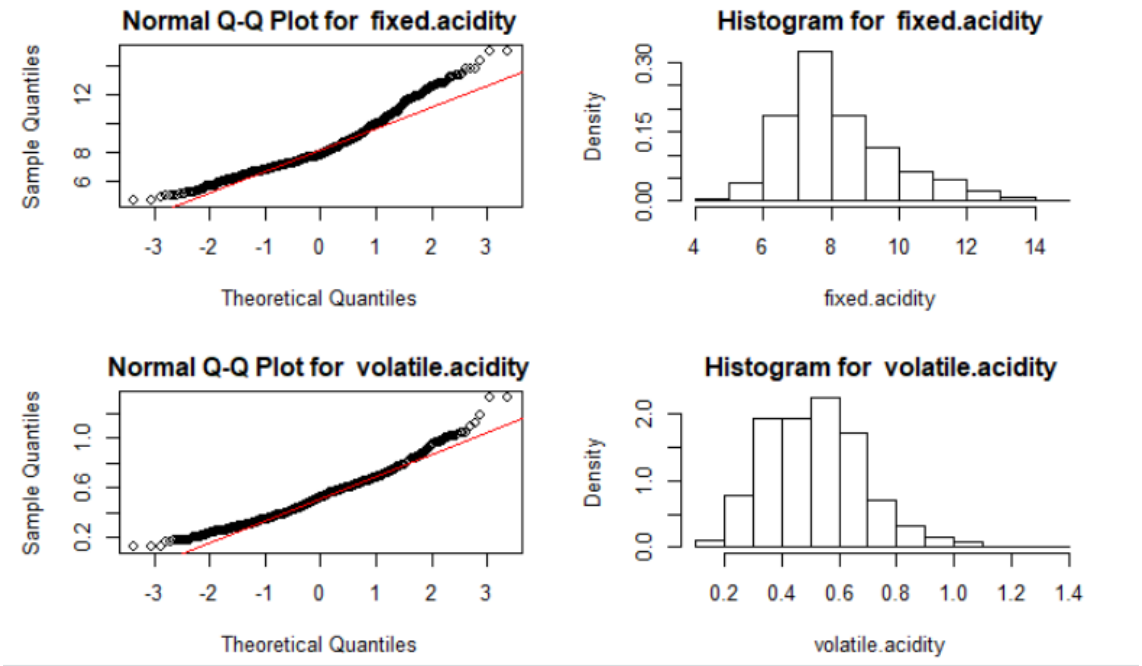
Una vez hemos explorado los datos para poder limpiarlos, vamos a pasar al análisis de los mismos.

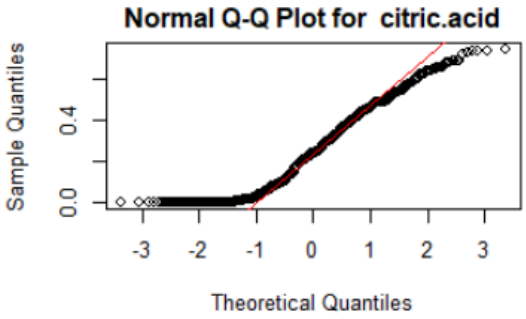
4.1. Selección de los grupos de datos que se quieren analizar.

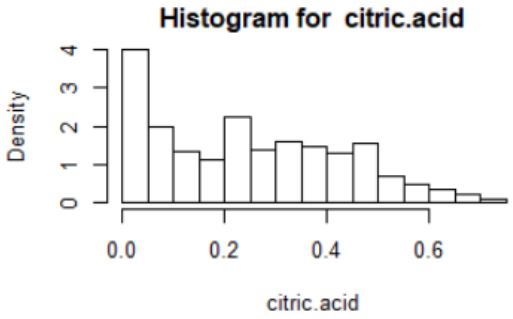
Vamos a utilizar todas las variables del conjunto de datos que tenemos, ya que tenemos, por un lado 11 variables que nos proporcionan información de las características del vino en cuestión, y otra variable que sería la "target" (variable quality) que es la puntuación final del vino.

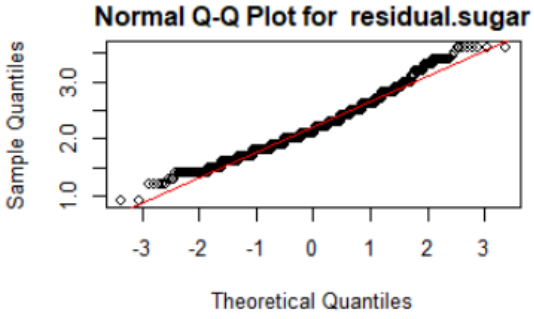
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

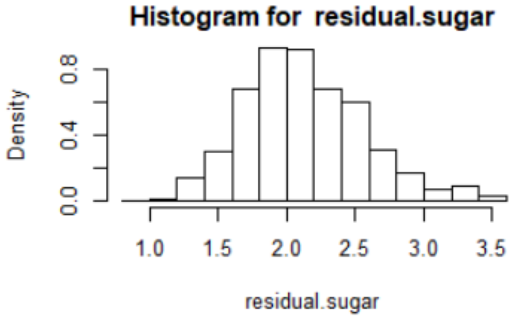
En primer lugar, nos interesa comprobar si las variables son candidatas a la normalización. Para ello, utilizaremos las gráficas de quantile-quantile e histogramas.

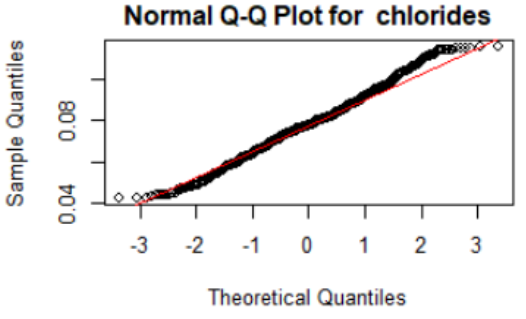


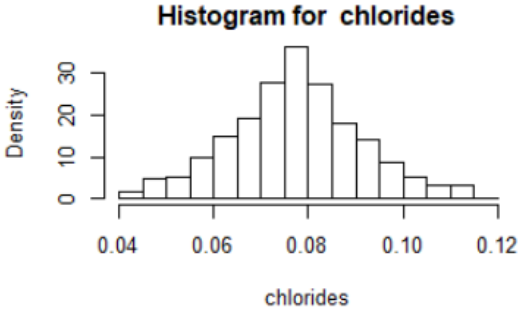

Normal Q-Q Plot for citric.acid
Sample Quantiles
0.0
0.4
-3
-2
-1
0
1
2
3
Theoretical Quantiles

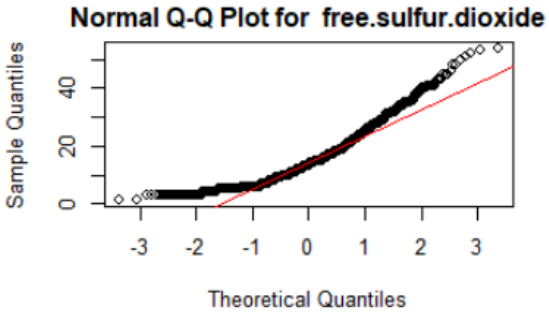

Histogram for citric.acid
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
citric.acid

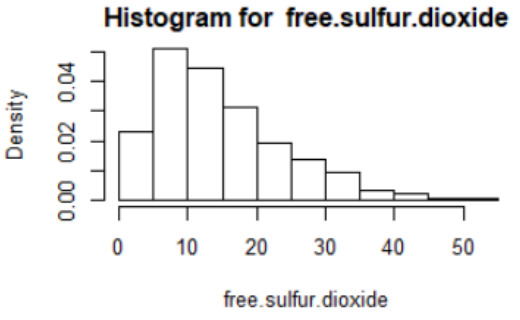

Normal Q-Q Plot for residual.sugar
Sample Quantiles
1.0
2.0
3.0
-3
-2
-1
0
1
2
3
Theoretical Quantiles

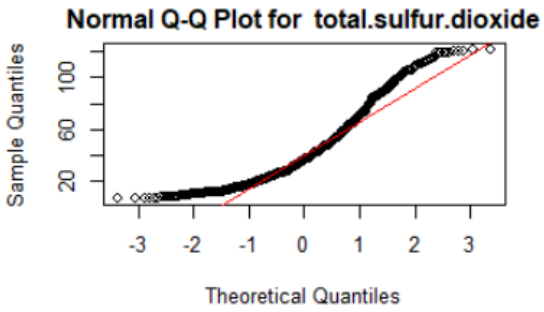

Histogram for residual.sugar
Density
0.0
0.4
0.8
1.0
1.5
2.0
2.5
3.0
3.5
residual.sugar

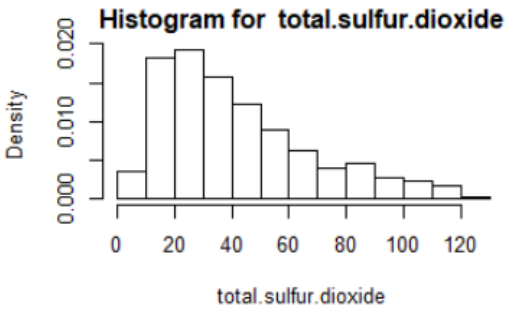

Normal Q-Q Plot for chlorides
Sample Quantiles
0.04
0.08
-3
-2
-1
0
1
2
3
Theoretical Quantiles

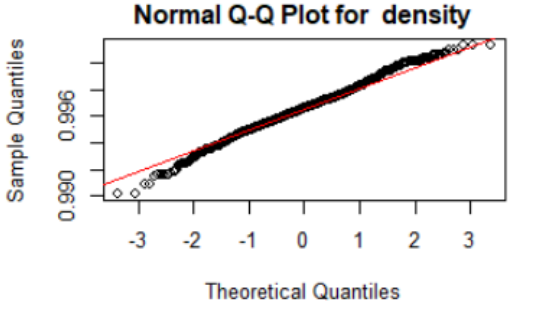

Histogram for chlorides
Density
0
10
20
30
0.04
0.06
0.08
0.10
0.12
chlorides

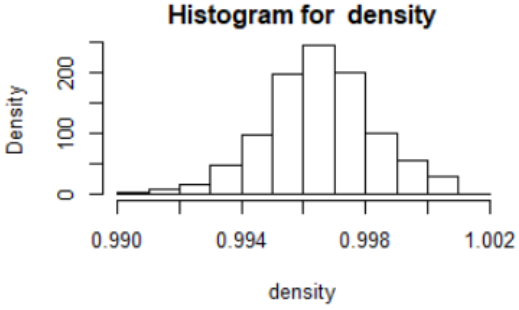

Normal Q-Q Plot for free.sulfur.dioxide
Sample Quantiles
0
20
40
-3
-2
-1
0
1
2
3
Theoretical Quantiles

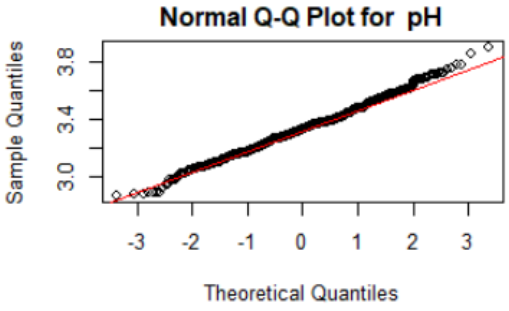

Histogram for free.sulfur.dioxide
Density
0.00
0.02
0.04
0
10
20
30
40
50
free.sulfur.dioxide

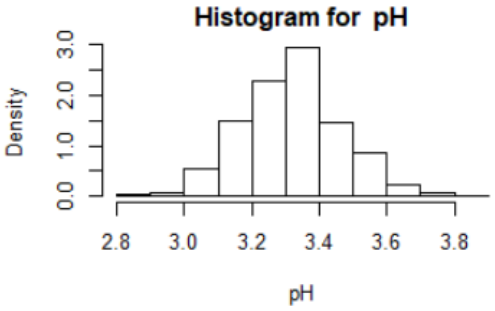

Normal Q-Q Plot for total.sulfur.dioxide
Sample Quantiles
20
60
100
-3
-2
-1
0
1
2
3
Theoretical Quantiles

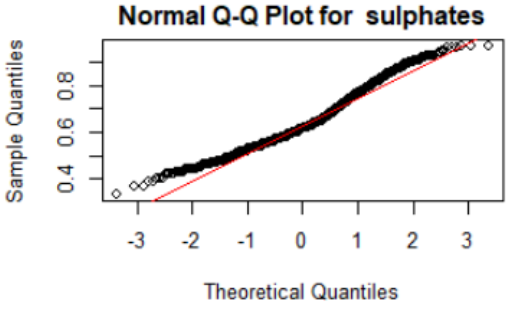

Histogram for total.sulfur.dioxide
Density
0.000
0.010
0.020
0
20
40
60
80
100
120
total.sulfur.dioxide

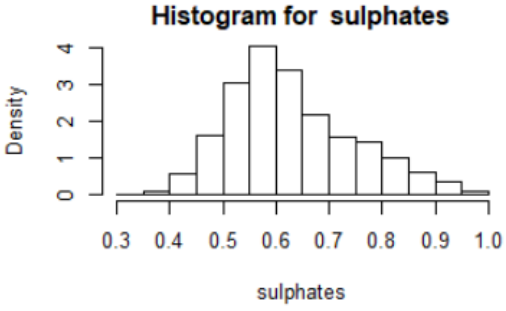

Normal Q-Q Plot for density
Sample Quantiles
0.990
0.996
-3
-2
-1
0
1
2
3
Theoretical Quantiles

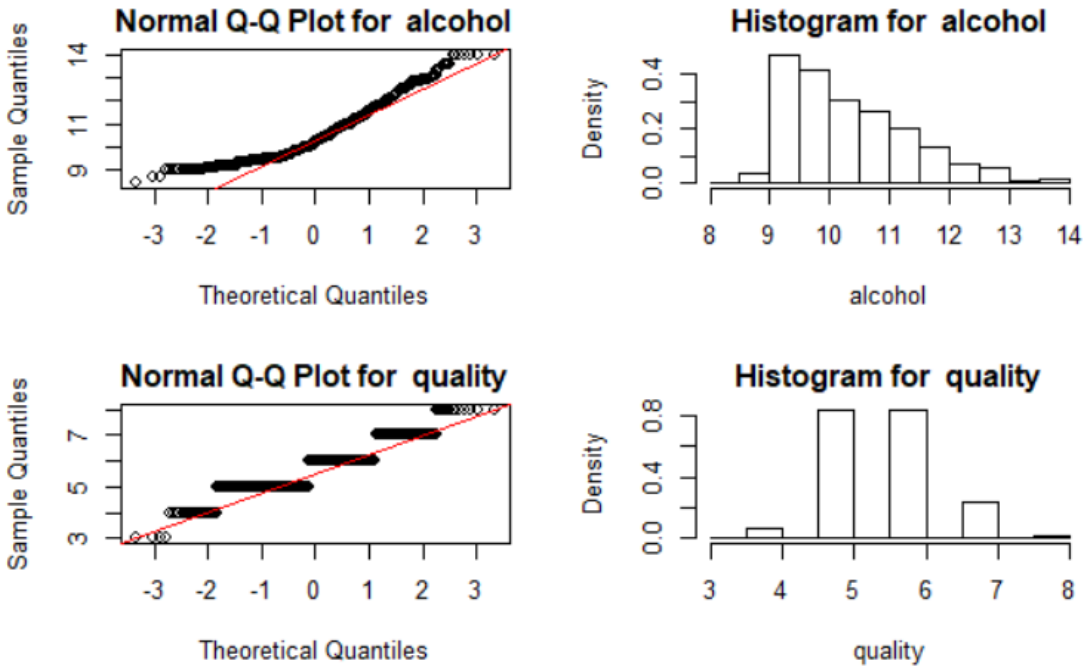

Histogram for density
Density
0
100
200
0.990
0.994
0.998
1.002
density


Normal Q-Q Plot for pH
Sample Quantiles
3.0
3.4
3.8
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Histogram for pH
Density
0.0
1.0
2.0
3.0
2.8
3.0
3.2
3.4
3.6
3.8
pH


Normal Q-Q Plot for sulphates
Sample Quantiles
0.4
0.6
0.8
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Histogram for sulphates
Density
0
1
2
3
4
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
sulphates



Las variables sí pueden normalizarse en caso de ser necesario, como podemos observar en las gráficas QQ. Aplicamos el test de Shapiro Wilk en todas las variables, ya que todas son numérica, para comprobar si están normalizadas.

```
Shapiro-Wilk normality test

data:  data_wine$fixed.acidity
W = 0.94393, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  data_wine$volatile.acidity
W = 0.97926, p-value = 1.447e-12


        Shapiro-Wilk normality test

data:  data_wine$citric.acid
W = 0.94933, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  data_wine$residual.sugar
W = 0.97521, p-value = 5.481e-14


        Shapiro-Wilk normality test

data:  data_wine$chlorides
W = 0.99386, p-value = 4.121e-05


        Shapiro-Wilk normality test

data:  data_wine$free.sulfur.dioxide
W = 0.91919, p-value < 2.2e-16


        Shapiro-Wilk normality test
```

```
data:  data_wine$total.sulfur.dioxide
W = 0.91472, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  data_wine$density
W = 0.99485, p-value = 0.0002319


        Shapiro-Wilk normality test

data:  data_wine$pH
W = 0.99607, p-value = 0.002385


        Shapiro-Wilk normality test

data:  data_wine$sulphates
W = 0.96975, p-value = 1.121e-15


        Shapiro-Wilk normality test

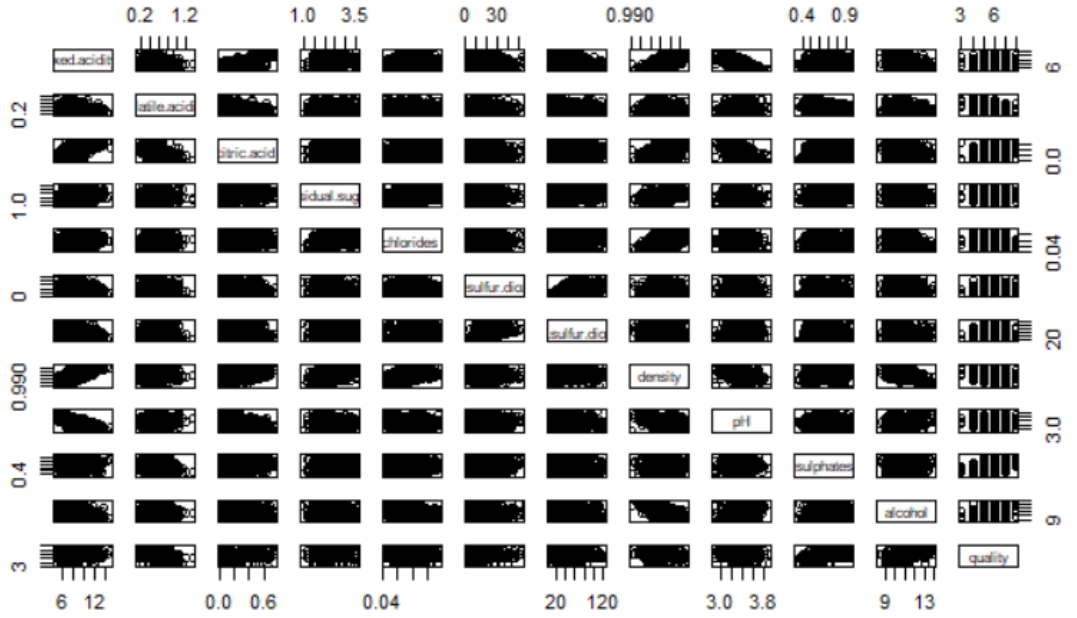
data:  data_wine$alcohol
W = 0.92685, p-value < 2.2e-16
```

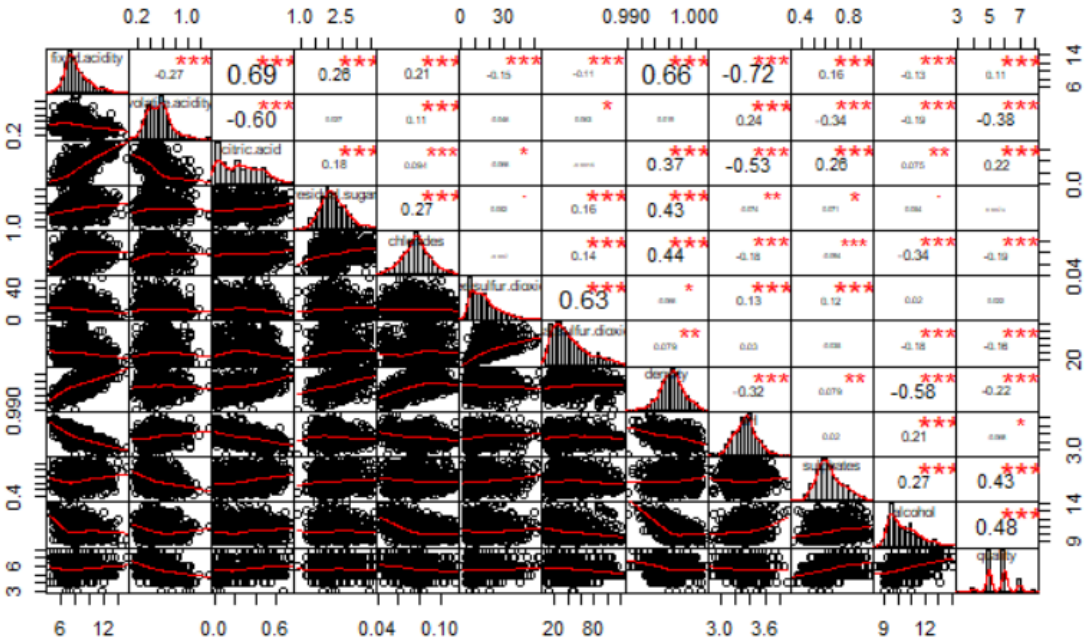
Nos fijamos en que para todas las variables, el valor de p-value es inferior a 0.05. Esto nos confirma que ninguna de las variables está normalizada; rechazamos la hipótesis nula del Shapiro Wilk normality test.

Este hecho no supone nigún problema. Según el Teorema del Límite Central, cuando tenemos un conjunto de datos "lo suficientemente grande" como en nuestro caso, podemos aproximar como una distribución normal de media 0 y distribución estándar 1.

4.3. Aplicación de pruebas estadísticas.

Vamos a estudiar la correlación entre variables para obtener las que más relación tengan con nuestra target. Nos interesa esta información para desarrollar un modelo de regresión.





Nos interesa detectar las variables que tengan una relación al menos moderadamente fuerte con la variable target *quality*, ya sea positiva o negativa. De la segunda tabla podemos ver lo siguiente:

- Con Alcohol el coeficiente de correlación es 0.48.

- Con Sulphates el coeficiente de correlación es 0.43.

- Con Volatile.acidity el coeficiente de correlación es -0.38.

Vamos a probar a generar un modelo de regresión lineal que pretenda explicar la puntuación de calidad del vino, utilizando estas variables. Comenzaremos con un modelo de regresión lineal simple que utilice la variable **alcohol** como explicativa.

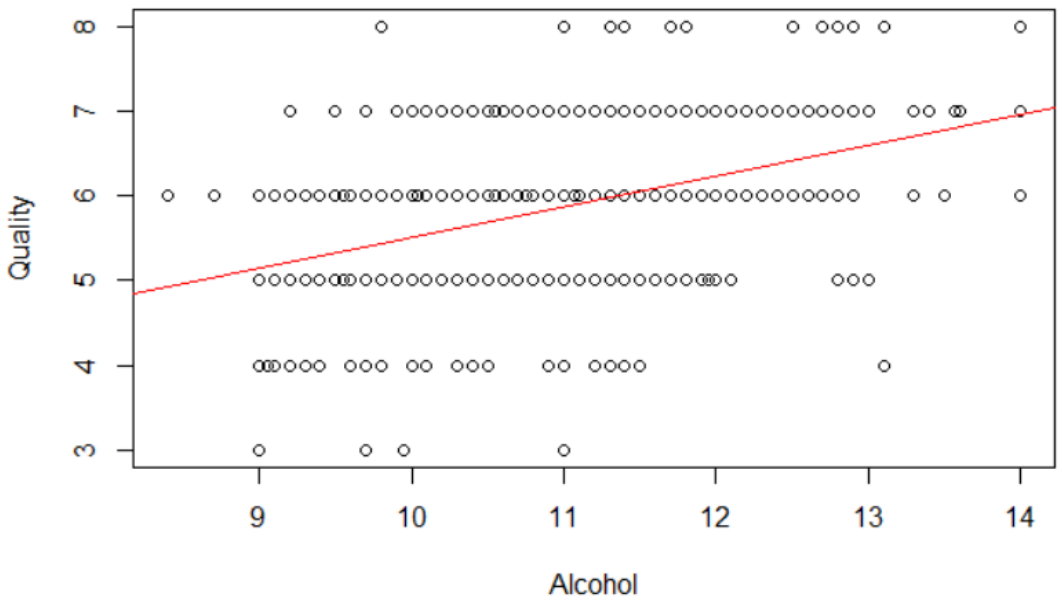
```
Call:
lm(formula = quality ~ alcohol, data = data_wine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8630 -0.4068 -0.1530  0.4995  2.5721

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.87491    0.19389    9.67   <2e-16 ***
alcohol      0.36255    0.01852   19.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6885 on 1273 degrees of freedom
Multiple R-squared:  0.2315,    Adjusted R-squared:  0.2309
F-statistic: 383.4 on 1 and 1273 DF,  p-value: < 2.2e-16
```

A continuación representamos la nube de puntos y la recta de mínimos cuadrados (en rojo).



Queremos evaluar la bondad del ajuste, que es el coeficiente de determinación de R2. Nos indica el grado de ajuste de la recta a los valores de muestra, y se define como la proporción de la varianza explicada por la recta de regresión.

Este valor lo podemos ver del modelo obtenido, en concreto es el valor Multiple R-squared: 0.2315. El valor se acerca mucho a 0, lo cual indica que el modelo no explica ninguna porción de variabilidad de los datos de respuesta en torno a su media.

Probaremos a construir un modelo de regresión lineal múltiple para explicar la calidad del vino. Vamos a ir añadiendo variables explicativas una a una para comprobar que efectivamente el modelo va mejorando según lo esperado, y que no estamos utilizando variables redundantes que no aportan valor.

En primer lugar, utilizaremos como variables explicativas alcohol, con un coeficiente de correlación con quality de 0.48, y sulphates de 0.43.

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.50683 -0.36830 -0.07426  0.45894  2.14089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1668     0.1877   6.216 6.91e-10 ***
alcohol       0.2965     0.0179  16.565  < 2e-16 ***
sulphates     2.2018     0.1582  13.919  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6416 on 1272 degrees of freedom
Multiple R-squared:  0.333,     Adjusted R-squared:  0.332
F-statistic: 317.6 on 2 and 1272 DF,  p-value: < 2.2e-16
```

Una vez generado el modelo, podemos ver que la bondad de ajuste ha mejorado respecto al modelo lineal. Mientras antes tenía un valor de 0.2315, ahora su valor ha subido a 0.333 Aún así, sigue siendo un valor bajo y por lo tanto el modelo no es explicativo.

Vamos a añadirle al modelo la variable explicativa volatile.acidity que, como hemos visto, tiene un coeficiente de correlación con quality de -0.38.

```
Call:
lm(formula = quality ~ alcohol + sulphates + volatile.acidity,
    data = data_wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28431 -0.39224 -0.06183  0.45731  1.98926

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.13967    0.21119  10.131   <2e-16 ***
alcohol            0.27936    0.01745  16.007   <2e-16 ***
sulphates          1.75071    0.16120  10.860   <2e-16 ***
volatile.acidity -0.97094    0.10696  -9.078   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.622 on 1271 degrees of freedom
Multiple R-squared:  0.3737,    Adjusted R-squared:  0.3722
F-statistic: 252.7 on 3 and 1271 DF,  p-value: < 2.2e-16
```

Observamos que la bondad de ajuste ha mejorado respecto al modelo lineal múltiple anterior. Mientras antes tenía un valor de 0.333, ahora su valor ha subido a 0.3737 Aún así, sigue siendo un valor bajo y por lo tanto el modelo no es explicativo.

En la tabla de correlaciones también tenemos dos variables que tienen un coeficiente de correlación con quality de 0.22. Dichas variable son citric.acid y density. Vamos a probar a añadirlas al modelo.

```
Call:
lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
    citric.acid, data = data_wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28673 -0.39369 -0.06024  0.45489  1.99628

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.15599    0.22274   9.679  < 2e-16 ***
alcohol            0.27909    0.01750  15.950  < 2e-16 ***
sulphates          1.75409    0.16192  10.833  < 2e-16 ***
volatile.acidity -0.98781    0.12949  -7.629 4.63e-14 ***
citric.acid      -0.02705    0.11695  -0.231    0.817
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6222 on 1270 degrees of freedom
Multiple R-squared:  0.3737,    Adjusted R-squared:  0.3717
F-statistic: 189.4 on 4 and 1270 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
    density, data = data_wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29825 -0.38895 -0.05986  0.45520  2.00522

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       15.78287   12.79818   1.233    0.218
alcohol            0.26432    0.02244  11.780   <2e-16 ***
sulphates          1.80054    0.16783  10.728   <2e-16 ***
volatile.acidity  -0.97378    0.10698  -9.102   <2e-16 ***
density          -13.56315   12.72134  -1.066    0.287
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.622 on 1270 degrees of freedom
Multiple R-squared:  0.3742,    Adjusted R-squared:  0.3722
F-statistic: 189.9 on 4 and 1270 DF,  p-value: < 2.2e-16
```

Observamos que la bondad de ajuste NO ha mejorado respecto al modelo lineal múltiple anterior. Si añadimos la variable explicativa citric.acid su valor no aumenta en absoluto. Si añadimos density, su valor aumenta de 0.3737 a 0.3742, lo cual NO es significativo.

La capacidad explicativa del modelo elegido, que es el de tres variables explicativas, no es satisfactoria, ya que R2 tiene un valor de 0.3737.

Aun así, podemos hacer un ejemplo de cómo se llevaría a cabo la predicción de un nuevo dato.

```
```{r}
predict(regresion_multiple_wine2, newdata = data.frame(alcohol=9.4, sulphates=0.56, volatile.acidity=0.70))
```

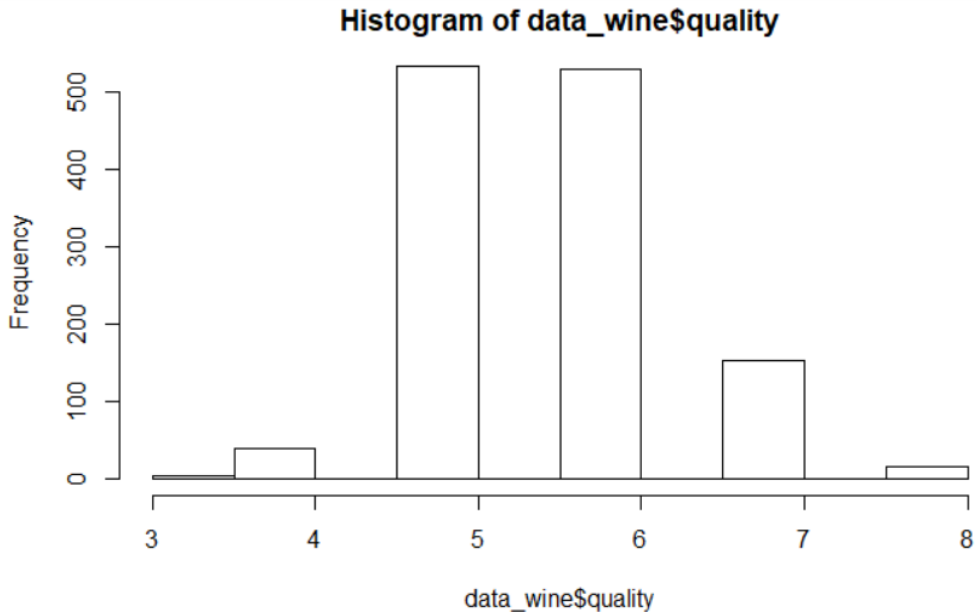
       1
5.066373
```

La puntuación esperada para la calidad es de 5, así que ha acertado, aunque no nos fiamos del modelo para utilizarlo para otras predicciones.

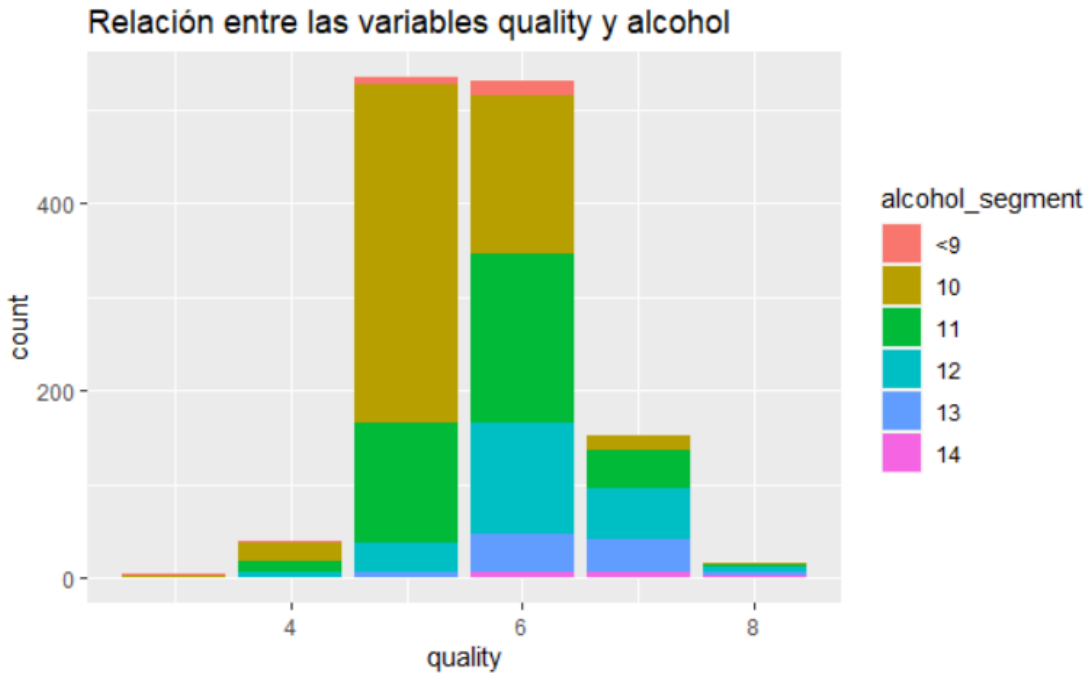
5. Representación de los resultados a partir de tablas y gráficas.

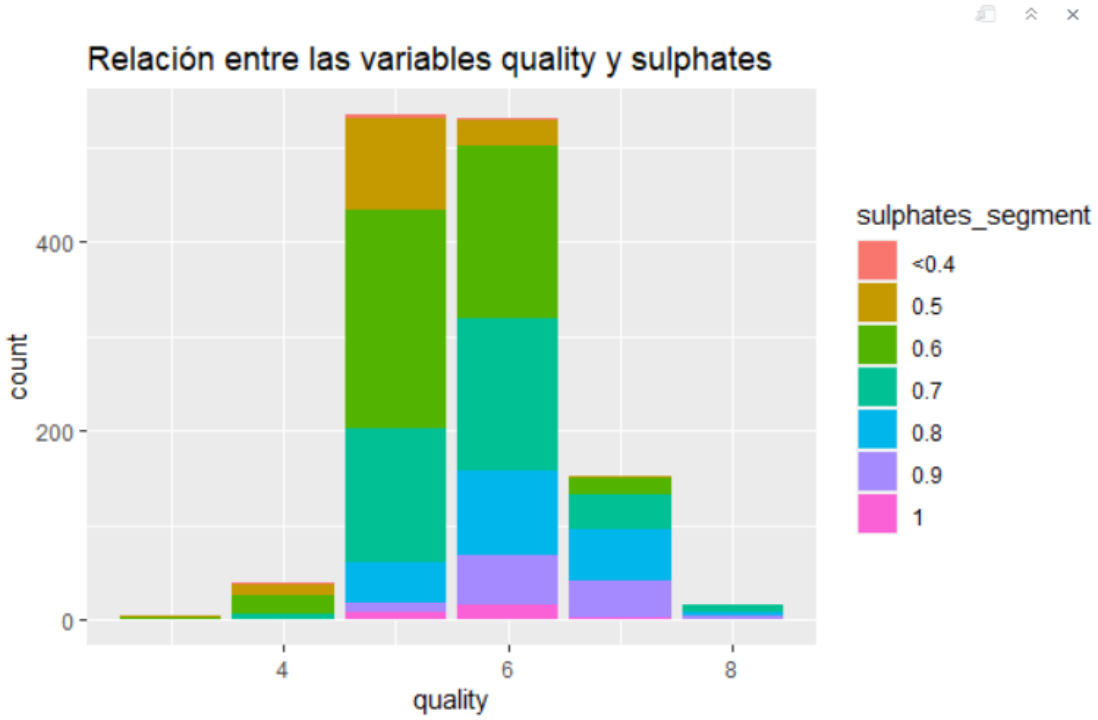
A lo largo del estudio hemos ido mostrando tablas de datos y gráficas, para las relaciones entre variables, por ejemplo. De todas formas, vamos a concluir con algunos gráficos que pueden mostrarnos más información sobre el conjunto de datos y el estudio que hemos realizado.

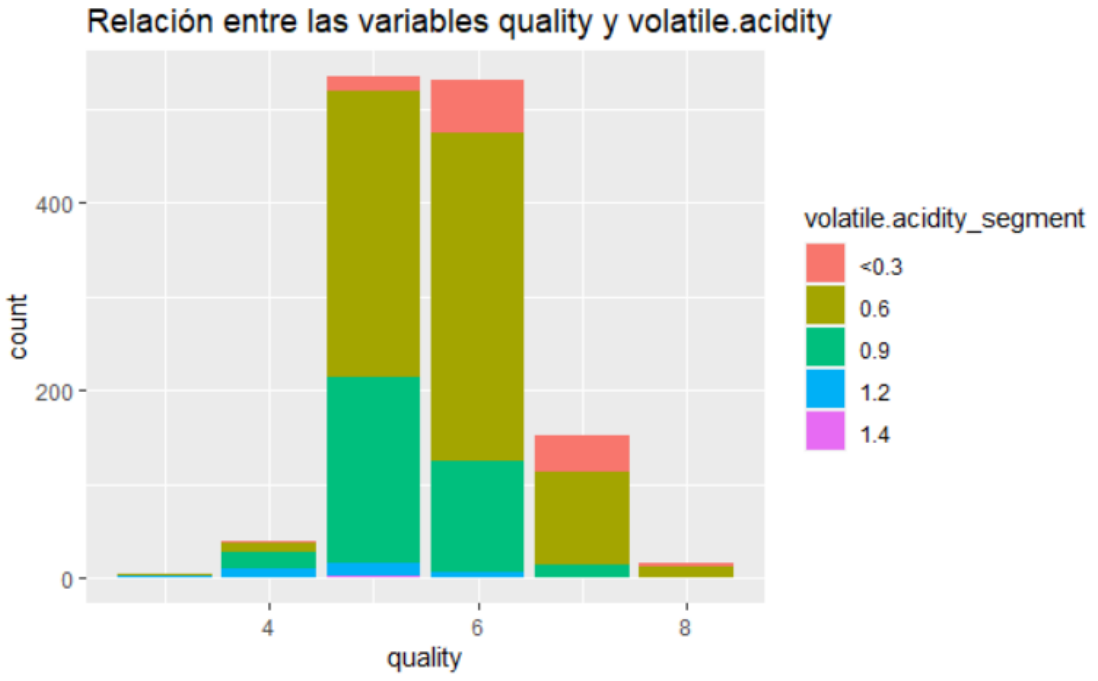
Por un lado, vamos a volver a visualizar la frecuencia de puntuaciones que nos ofrece la variable target, que es quality.



Vamos a ver de forma visual la forma en que están distribuidos los tipos de vino según las tres variables explicativas más fuertes. Para ello las discretizaremos.







6. Resolución del problema.

En resumen, se ha realizado un estudio lo suficientemente exhaustivo sobre los datos que tenemos como para poder hacernos una idea profunda de los mismos. Hemos intentado construir un modelo de regresión lineal múltiple utilizando como variables explicativas, las que están más fuertemente correlacionadas con la variable que nos informa sobre la calidad. Aún

así, el modelo construido no es suficientemente explicativo como para poder utilizarlo para realizar predicciones sobre nuevos datos reales.

Utilizando esta información que ahora tenemos, aunque podemos ver qué variables influyen más en la puntuación del vino, seguimos sin poder utilizarlas realmente para predecir. Por ello, podemos intuir que la calidad del vino tiene un componente subjetivo de la persona que lo califica, que no puede ser reflejado en los datos que se nos proporcionan sobre dicho vino.

7. Código

El código que se ha ido desarrollando a lo largo del estudio de datos ha sido en R, utilizando RStudio.

8. Integrantes del grupo

Por motivos personales, he decidido que sería mejor desarrollar la práctica por mi cuenta en lugar de formar grupo con otro compañero o compañera.

Contribuciones	**Firma**
Investigación previa	Mar Bonora Ortega
Redacción de respuestas	Mar Bonora Ortega
Desarrollo código	Mar Bonora Ortega