TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Montse Rodríguez Capece

MASTER DATA SCIENCE - UOC

Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder?

Disponemos de una base de datos *Red Wine Quality* obtenida del repositorio *Kaggle*. Dicha base contiene información sobre distintas propiedades fisicoquímicas que intervienen en la elaboración del vino.

El objetivo principal es poder explicar la calidad del vino con la información en dicho dataset, de esta forma también se podrá predecir su calidad, sabiendo las propiedad fisicoquímicas.

Análisis exploratorio

Disponemos de una base de datos con un total de 4898 registros, con 12 variables, de las cuales una corresponde con la variable objetivo, *quality*, que determina la calidad del vino, y las 11 restantes son propiedades fisicoquímicas del vino.

VARIABLE	DESCRIPCIÓN
FIXED.ACIDITY	Acidez fija: la mayoría de los ácidos relacionados con el vino o fijos o no volátiles (no se evaporan fácilmente)
VOLATILE.ACIDITY	Acidez volátil: la cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable, a vinagre.
CITRIC.ACID	Ácido cítrico: encontrado en pequeñas cantidades, el ácido cítrico puede agregar 'frescura' y sabor a los vinos
RESIDUAL.SUGAR	Azúcar residual: la cantidad de azúcar restante después de que se detiene la fermentación, es raro encontrar vinos con menos de 1 gramo / litro y vinos con más de 45 gramos / litro se consideran dulces
CHLORIDES	Cloruro: la cantidad de sal en el vino.
FREE.SULF.DIOXIDE	Dióxido de azufre libre: la forma libre de SO2 existe en equilibrio entre el SO2 molecular (como un gas disuelto) y el ion bisulfito; previene el crecimiento microbiano y la oxidación del vino.
TOTAL.SULFUR.DIOXIDE	Cantidad total de dióxido de azufre: de formas libres y unidas de SO2; en bajas concentraciones, el SO2 es mayormente indetectable en el vino, pero a concentraciones de SO2 libres de más de 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino.
DENSITY	Densidad: del agua, es cercana a la del agua dependiendo del porcentaje de alcohol y de contenido de azúcar.
PH	pH: describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); La mayoría de los vinos están entre 3-4 en la escala de pH.
SULPHATES	Sulfatos: aditivo de vino que puede contribuir a los niveles de gas de dióxido de azufre (S02), que actúa como antimicrobiano y antioxidante.
ALCOHOL	Alcohol: el porcentaje de alcohol del vino
QUALITY	Calidad: variable de salida (basada en datos sensoriales, puntuación entre 0 y 10)

```
> str(wine)
'data.frame':   4898 obs. of  12 variables:
 $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
 $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
 $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
 $ residual.sugar      : num  191 19 259 287 287 259 262 191 19 17 ...
 $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
 $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 NA ...
 $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
 $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
 $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
 $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
 $ alcohol             : num  88 95 2 101 101 2 97 88 95 22 ...
 $ quality             : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4 4 4 4 4 4 4 4 ...
> head (wine)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
1           7.0             0.27        0.36            191     0.045                  45                  170  1.0010 3.00
2           6.3             0.30        0.34             19     0.049                  14                  132  0.9940 3.30
3           8.1             0.28        0.40            259     0.050                  30                   97  0.9951 3.26
4           7.2             0.23        0.32            287     0.058                  47                  186  0.9956 3.19
5           7.2             0.23        0.32            287     0.058                  47                  186  0.9956 3.19
6           8.1             0.28        0.40            259     0.050                  30                   97  0.9951 3.26
  sulphates alcohol quality
1      0.45      88       6
2      0.49      95       6
3      0.44       2       6
4      0.40     101       6
5      0.40     101       6
6      0.44       2       6
```

Integración y selección de los datos

Integración y selección de los datos de interés a analizar

Tras una primera presentación del dataset, y dadas las dimensiones del mismo, se decide descarga el fichero de la web de *Kaggle* directamente en un csv, el cuál se cargará, en su totalidad, en R para su explotación.

La única modificación que se hace en la carga es la creación de una nueva variable, en esta caso una nueva variable respuesta, llamada *rating*, que se define como una variable categórica que pretende clasificar, según la putuación *quality*, la calidad del vino en 3 categorías diferentes: *good-bueno*, *average-medio* o *bad-mal*. La definición se detalla a continuación:

- *quality* >= 7 → *rating* = 'good'
- 5 <= *quality* < 7 → *rating* = 'average'
- *quality* < 5 → *rating* = 'bad'

NOTA: en este caso se decide realizar, de manera totalmente arbitraria, un target categórico con estas puntuaciones como cortes. Se podrían implementar diferentes algoritmos y/o métodos para decidir el número de categorías y sus puntos de corte, así como determinar, por decisiones de empresa, una puntuación a partir de la cual se considere de calidad o no el vino, generando una variable binaria.

Limpieza de los datos

Identificación y tratamiento de valores que generan ruido, son missing o outliers. Resolver inconsistencias.

Los datos de baja calidad conducirán a resultados baja calidad.

Las técnicas de preprocesamiento de datos pueden mejorar la calidad de los datos, lo que ayuda a mejorar la precisión y la eficiencia del proceso de análisis posterior.

El objetivo de este punto es el conocimiento en detalle de los datos a analizar sobre los cuales posteriormente aplicaremos una serie de métodos para poder dar respuesta a nuestro objetivo.

Estudiar y conocer los datos nos permitirá reconocer posibles relaciones e identificar la presencia de valores atípicos y valores *missing* y así poder actuar en consecuencia, resolviendo dichas inconsistencias y consiguiendo un análisis posterior robusto.

Para ello, realizamos una exploración de los datos a través de las variables, mediante descriptivas numéricas y gráficas, donde graficamos cada una de las variables, denominado análisis univariante, y graficamos relaciones entre la variable respuesta y las variables regresoras, denominado análisis bivariante.

Cabe destacar que esta tarea es iterativa, corrigiendo y validando los datos constantemente hasta obtener una dataset óptimo para su tratamiento y explotación.

Valores missing, ceros y outliers:

Valores NA's:

```
> colSums(is.na(wine))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide 
                   0                    1                    0                    0                    0                   11 
total.sulfur.dioxide              density                   pH            sulphates              alcohol              quality 
                   0                    2                    0                    6                    0                    0 
              rating 
                   0 
```

Ceros:

```
> colSums(wine=="")
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides 
                   0                   NA                    0                    0                    0 
 free.sulfur.dioxide total.sulfur.dioxide              density                   pH            sulphates 
                  NA                    0                   NA                    0                   NA 
             alcohol              quality               rating 
                   0                    0                    0 
```

Valores extremos:

Fixed acidity valores extremos:

Se puede ver el valor más alto marcado, podría ser un outlier, como es un solo registro, lo eliminaremos cuando lo tratemos.

```
> boxplot.stats(wine$fixed.acidity)$out
  [1]  9.8  9.8 10.2  9.1 10.0  9.2  9.2  9.0  9.1  9.2 10.3  9.4  9.2  9.8  9.6  9.2  9.0  9.3  9.2  9.1  8.9
 [22]  9.8  8.9  9.2  9.7  9.4 10.3  9.6  9.0  9.7  9.2  9.4  9.6  9.2  9.0  9.2 10.7 10.7  9.0  9.2  9.8  9.2
 [43] 14.2  8.9  8.9  9.1  9.1  9.8  9.0  9.3  8.9  9.0  9.0  8.9  9.0  9.3  9.2  9.6  9.4  9.4 10.0  8.9  8.9
 [64] 10.0  9.2  9.2  9.2  9.9  9.5  9.0  9.0  8.9  9.5 11.8  9.4  9.1  9.8  9.9  9.2  8.9  9.2  9.4  9.4  9.4
 [85]  4.6  8.9  9.4  9.2  9.2  9.8  9.0  9.0  9.0  8.9  8.9  4.5  9.2  9.6  4.2  9.7  9.7  9.0  4.2  9.4  8.9
[106]  8.9  8.9  4.7  4.7  3.8  4.4  4.7  9.0  9.0  4.7  4.4  3.9  4.7  4.4
```

Volatile-Acidity valores extremos:

Se puede ver el valor más alto marcado, no es un valor outlier.

```
> boxplot.stats(wine$volatile.acidity)$out
  [1] 0.660 0.660 0.670 0.540 0.595 0.670 0.530 0.540 0.570 0.685 0.495 0.640 0.520 0.580 0.585 0.590 0.600 0.580
 [19] 0.590 0.550 0.905 0.550 0.490 0.550 0.520 0.600 0.550 0.510 0.620 0.510 0.560 0.570 0.670 0.500 0.560 0.560
 [37] 0.655 0.595 0.705 0.520 0.550 0.600 0.640 0.680 0.490 0.510 0.550 0.520 0.500 0.550 0.600 0.610 0.610 0.610
 [55] 0.660 0.570 0.500 0.500 0.590 0.580 0.540 0.580 0.570 0.640 0.560 0.490 0.490 0.670 0.550 0.560 0.520 0.520
 [73] 0.850 0.510 0.620 0.510 0.530 0.640 0.550 0.490 0.490 0.610 0.545 0.620 0.490 0.500 0.490 0.490 0.550 0.490
 [91] 0.910 0.530 0.490 0.710 1.005 0.490 0.550 0.550 0.760 0.500 0.930 0.490 0.495 0.695 0.705 0.815 0.560 0.560
[109] 0.560 0.510 0.540 0.540 0.500 0.615 0.500 0.520 0.600 0.680 0.655 0.510 0.510 0.615 0.615 0.965 0.740 0.530
[127] 0.780 0.680 0.640 0.540 0.750 0.640 0.640 0.655 0.580 0.520 0.530 0.600 0.530 0.580 0.670 0.610 0.730 0.650
[145] 0.580 1.100 0.500 0.500 0.500 0.650 0.520 0.550 0.585 0.560 0.555 0.555 0.540 0.610 0.550 0.530 0.660 0.615
[163] 0.500 0.620 0.500 0.490 0.510 0.510 0.540 0.610 0.695 0.695 0.630 0.630 0.690 0.690 0.590 0.620 0.785 0.760
[181] 0.500 0.540 0.520 0.600 0.540 0.530
```

Citrix valores extremos:

Se puede ver el valor más alto marcado, podría ser un outlier, como es un caso tan cercano, no vamos a eliminarlo.

```
> boxplot.stats(wine$citric.acid)$out
  [1] 0.62 0.04 0.59 0.07 0.03 0.61 0.62 0.63 0.61 0.62 0.63 0.66 0.66 0.00 0.04 0.67 0.67 0.04 0.04 0.07 0.88
 [22] 0.08 0.59 0.07 0.07 0.07 0.07 0.58 0.70 0.00 0.00 0.60 0.07 0.09 0.04 0.62 0.58 0.62 0.70 0.62 0.62 0.58
 [43] 0.02 0.65 0.65 0.71 0.66 0.66 0.07 0.06 0.07 0.06 0.68 0.68 0.68 0.68 0.06 0.72 0.69 0.58 0.70 1.66 0.04
 [64] 0.63 0.60 0.00 0.08 0.58 0.58 0.05 0.58 0.00 0.00 0.65 0.58 0.00 0.05 0.05 0.62 0.62 0.58 0.58 1.00 0.09
 [85] 0.01 0.71 0.71 0.60 0.06 0.74 0.81 0.69 0.58 0.69 0.00 0.07 0.64 0.72 0.73 0.65 0.68 0.65 0.74 0.71 0.59
[106] 0.68 0.08 0.72 0.64 0.02 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74
[127] 0.74 0.74 0.74 0.74 0.74 0.99 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.74 0.01
[148] 0.74 0.01 0.74 0.74 1.00 0.04 0.58 0.07 1.00 0.00 0.58 0.61 0.61 0.61 0.02 0.67 0.67 0.67 0.58 0.65 0.58
[169] 0.09 0.08 0.71 0.04 0.03 0.05 0.64 0.64 0.58 0.58 0.81 0.58 0.61 0.62 0.59 0.00 0.04 0.63 0.73 0.68 0.09
[190] 0.78 0.79 0.09 0.64 0.65 0.65 0.00 0.73 0.73 0.64 0.60 0.71 0.72 0.82 0.07 0.58 0.58 1.00 0.66 0.80 0.80
[211] 1.23 0.59 0.02 0.00 1.00 0.62 0.00 0.71 0.71 0.71 0.61 0.61 0.00 0.60 0.58 0.09 0.09 0.72 0.62 0.62 0.79
[232] 0.82 0.67 0.01 0.01 0.86 0.61 0.02 0.05 0.00 0.69 0.69 0.59 0.01 0.66 0.66 0.78 0.00 0.04 0.91 0.91 0.06
[253] 0.06 0.04 0.04 0.74 0.09 0.09 0.60 0.62 0.73 0.00 0.09 0.00 0.09 0.67 0.01 0.09 0.00 0.02
```

Chrorides valores extremos:

Se puede ver el valor más alto marcado, no es un valor outlier.

```
> boxplot.stats(wine$chlorides)$out
  [1] 0.074 0.080 0.172 0.173 0.147 0.092 0.082 0.092 0.200 0.197 0.197 0.074 0.132 0.089 0.108 0.081 0.073 0.346
 [19] 0.090 0.114 0.186 0.180 0.084 0.083 0.096 0.094 0.240 0.290 0.185 0.110 0.078 0.130 0.135 0.115 0.072 0.170
 [37] 0.080 0.119 0.126 0.150 0.152 0.088 0.244 0.137 0.093 0.077 0.079 0.073 0.072 0.076 0.201 0.201 0.074 0.074
 [55] 0.301 0.138 0.169 0.083 0.093 0.168 0.122 0.172 0.167 0.239 0.076 0.138 0.137 0.123 0.123 0.133 0.073 0.073
 [73] 0.211 0.123 0.123 0.255 0.204 0.208 0.083 0.080 0.076 0.086 0.084 0.084 0.168 0.160 0.179 0.076 0.076 0.087
 [91] 0.217 0.094 0.157 0.157 0.148 0.158 0.157 0.168 0.157 0.092 0.099 0.084 0.085 0.091 0.093 0.080 0.095 0.096
[109] 0.096 0.147 0.142 0.079 0.074 0.075 0.074 0.121 0.121 0.079 0.079 0.014 0.156 0.012 0.119 0.119 0.081 0.170
[127] 0.171 0.082 0.074 0.083 0.083 0.152 0.169 0.073 0.014 0.078 0.112 0.154 0.126 0.126 0.104 0.142 0.102 0.184
[145] 0.184 0.096 0.076 0.146 0.117 0.117 0.118 0.014 0.085 0.087 0.085 0.087 0.076 0.088 0.160 0.167 0.014 0.009
[163] 0.098 0.098 0.086 0.086 0.194 0.094 0.013 0.144 0.149 0.185 0.084 0.175 0.090 0.098 0.110 0.110 0.095 0.174
[181] 0.097 0.142 0.145 0.208 0.209 0.105 0.086 0.176 0.176 0.108 0.096 0.271 0.120 0.212 0.094 0.094 0.117 0.173
[199] 0.074 0.076 0.076 0.175 0.174 0.075 0.127 0.127 0.096 0.136
```

Free Sulfur Dioxide valores extremos:

El valor parece más alto parece ser un outlier, lo eliminaremos a continuación.

```
> boxplot.stats(wine$free.sulfur.dioxide)$out
  [1]  81.0  82.0 131.0  82.5  87.0  87.0  83.0 122.5  83.0  81.0  88.0  82.0 118.5  81.0  96.0  83.0  83.0 146.5
 [19] 128.0 110.0  85.0  89.0  86.0  86.0  96.0  96.0  93.0  85.0  81.0 138.5  95.0 124.0  87.0  87.0 105.0 105.0
 [37] 101.0 101.0 108.0 108.0  98.0  98.0 112.0 108.0  98.0  81.0  81.0  81.0 289.0  97.0
```

Total Sulfur Dioxide valores extremos:

```
> boxplot.stats(wine$total.sulfur.dioxide)$out
 [1] 272.0 313.0 260.0  19.0 366.5 307.5 256.0 256.0 344.0 282.0 303.0 272.0  18.0  18.0 294.0   9.0  10.0 259.0
[19] 440.0
```

Se puede ver el valor más alto marcado, podría ser un outlier, no lo vamos a eliminar.

Density valores extremos:

El valor más alto está marcado, es un outlier, de hecho, lo son todos los valores por encima de 2 en este campo y se elimarán al tratarlos

```
> boxplot.stats(wine$density)$out
 [1]  10.002  10.002 100.055  10.006  10.006  10.002  10.002  10.004  10.006  10.003  10.003  10.003  10.004
[14]  10.001  10.005  10.012  10.004  10.004  10.024  10.001  10.103  10.103  10.004  10.008  10.002  10.008
[27]  10.008  10.007  10.001  10.001  10.017  10.017  10.011  10.011  10.006  10.004  10.004  10.004  10.002
[40]  10.004  10.001  10.001  10.001  10.005  10.001  10.001  10.001  10.001 100.182 100.047 100.241 100.098
[53] 100.016 100.051 100.118 100.014  10.002 100.013 100.013 103.898 100.014 100.196 100.037 100.037 100.295
[66] 100.295 100.044 100.044 100.022 100.038 100.038
```

pH valores extremos:

El valor más alto está marcado y no parece un outlier, está dentro del rango posible de valores.

```
> boxplot.stats(wine$pH)$out
 [1] 3.69 3.63 3.72 3.61 3.64 3.64 3.72 3.72 3.58 3.58 3.66 3.59 2.74 3.82 3.81 3.65 3.65 3.59 3.77 3.62 3.63
[22] 3.58 3.58 3.65 3.74 2.80 3.60 3.60 2.72 3.60 2.79 2.79 3.57 3.80 3.60 3.60 3.68 3.63 3.63 2.77 3.63 3.60
[43] 3.60 3.61 3.61 3.59 3.79 3.59 3.68 3.59 3.66 3.70 3.74 3.80 3.57 3.57 3.57 3.65 3.58 2.80 3.77 3.76 3.69
[64] 3.66 3.59 2.79 3.75 3.63 3.75 3.76 3.66 3.66 2.80 3.67 3.57
```

Sulphates valores extremos:

El valor marcado es el más alto y no es un outlier.

```
> boxplot.stats(wine$sulphates)$out
  [1] 0.77 0.84 0.77 0.79 0.85 0.78 0.79 0.79 0.79 0.77 0.78 0.85 0.96 0.97 0.82 0.82 0.77 0.95 0.95 0.77 0.95
 [22] 0.82 0.82 0.90 0.88 0.88 0.79 0.80 0.80 0.78 0.78 0.87 0.86 0.90 0.90 0.78 0.79 0.81 0.81 0.77 0.82 0.79
 [43] 0.79 0.77 0.82 0.92 0.79 0.79 0.82 0.82 0.82 0.82 0.82 0.79 0.78 0.79 0.77 0.77 0.77 0.98 1.06 0.88 0.88
 [64] 0.88 0.80 0.78 1.00 0.80 0.90 0.90 0.89 0.94 0.99 0.86 0.84 0.95 0.84 0.84 0.81 0.80 0.87 0.82 0.78 0.78
 [85] 0.78 0.78 0.78 0.77 0.85 0.78 0.78 0.88 0.88 0.78 0.78 0.78 0.78 0.79 0.77 0.77 0.83 0.83 0.81 0.81 0.98
[106] 0.98 0.98 0.98 0.79 0.79 0.78 0.82 0.98 0.77 0.96 1.01 0.77 0.96 0.77 0.92 0.94 0.95 1.08 0.79
```

Tratamiento de missing y outliers:

Como ya hemos visto en los análisis anteriores, nuestros datos presentan tanto outliers como missing.

```
> colSums(is.na(wine))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide
                   0                    1                    0                    0                    0                   11
total.sulfur.dioxide              density                   pH            sulphates              alcohol              quality
                   0                    2                    0                    6                    0                    0
              rating
                   0
```

En el caso de los *missing*, dado que el volumen es pequeño y la muestra de análisis es elevada, procederemos a asignarles el valor medio.

Tras aplicar el algoritmo kNN que nos buscará los objetos vecinos más cercanos y nos asignará el valor más adecuado en cada caso.

```
> colSums(is.na(wine))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
                   0                    0                    0                    0                    0
 free.sulfur.dioxide total.sulfur.dioxide              density                   pH            sulphates
                   0                    0                    0                    0                    0
             alcohol              quality               rating
                   0                    0                    0
> |
```

Respecto a los *outliers*, están presentan en prácticamente todas las variables y procederemos a su borrado una vez se estudien la coherencia de estos y no tenerlos en cuenta con el objetivo de conseguir una base de datos capaz de proporcionarnos resultados robustos.

Fixed Acidity tratada:

```
> boxplot.stats(wine$fixed.acidity)$out
  [1]  9.8  9.8 10.2  9.1 10.0  9.2  9.2  9.0  9.1  9.2 10.3  9.4  9.2  9.8  9.6  9.2  9.0  9.3  9.2  9.1  8.9  9.8  8.9
 [24]  9.2  9.7  9.4 10.3  9.6  9.0  9.7  9.2  9.4  9.6  9.2  9.0  9.2 10.7 10.7  9.0  9.2  9.8  9.2  8.9  8.9  9.1  9.1
 [47]  9.0  9.3  8.9  9.0  9.0  8.9  9.0  9.3  9.2  9.6  9.4  9.4 10.0  8.9  8.9 10.0  9.2  9.2  9.2  9.9  9.5  9.0  9.0
 [70]  8.9  9.5 11.8  9.4  9.1  9.8  9.9  9.2  8.9  9.4  9.4  9.4  4.6  8.9  9.2  9.2  9.8  9.0  9.0  9.0  8.9  8.9  4.5
 [93]  9.2  9.6  4.2  9.7  9.7  9.0  4.2  9.4  8.9  8.9  8.9  4.7  4.7  3.8  4.4  4.7  9.0  9.0  4.7  4.4  3.9  4.7  4.4
```

Density tratada:

```
> boxplot.stats(wine$density)$out
numeric(0)
```

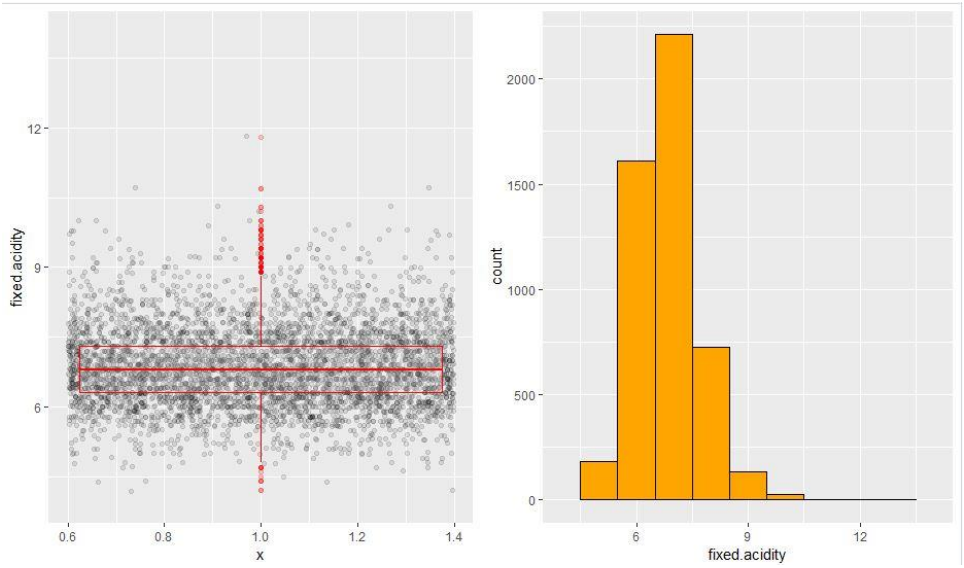
Free Sulfur Dioxide tratada:

```
> boxplot.stats(wine$free.sulfur.dioxide)$out
 [1]  81.0  82.0 131.0  82.5  87.0  87.0  83.0 122.5  83.0  81.0  88.0  82.0 118.5  81.0  96.0  83.0  83.0 146.5 110.0
[20]  85.0  89.0  86.0  86.0  96.0  96.0  93.0  85.0  81.0 138.5  95.0 124.0  87.0  87.0 105.0 105.0 101.0 101.0 108.0
[39] 108.0  98.0  98.0 112.0 108.0  98.0  81.0  81.0  81.0  97.0
> |
```

Análisis univariante

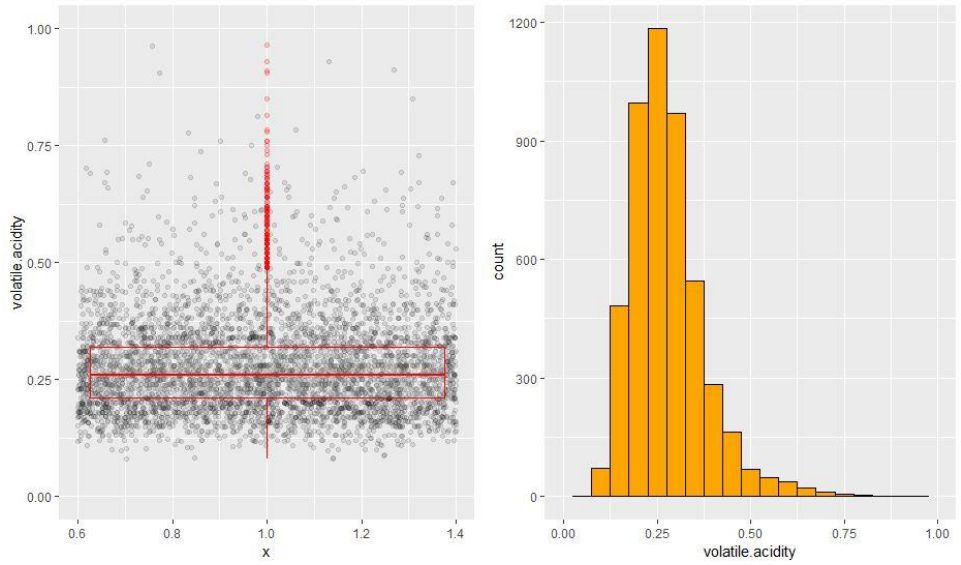
En este caso, como ya se ha dicho antes, conocemos las variables que intervienen, su forma y la presencia de valores atípicos.

Fixed-Acidity Plot:



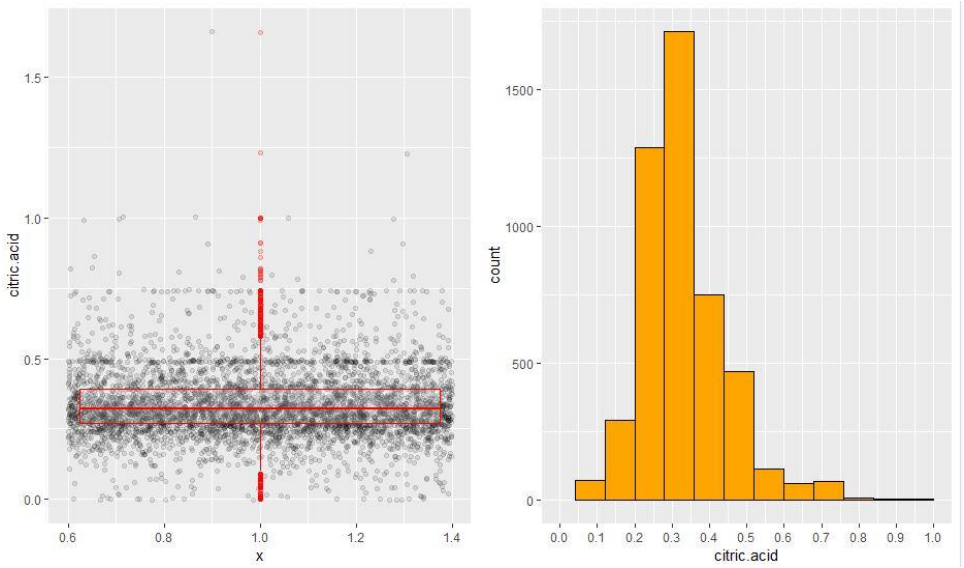

```
summary(wine$fixed.acidity)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.800   6.300   6.800   6.855   7.300  14.200
```

Volatile-Acidity Plot:



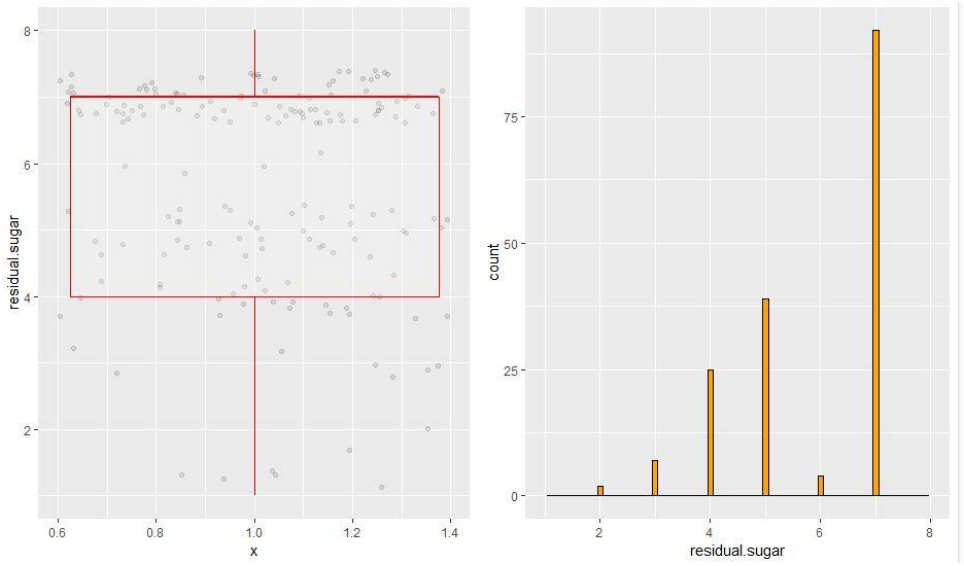

```
summary(wine$volatile.acidity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.0800  0.2100  0.2600  0.2782  0.3200  1.1000       1
```

Citric Plot:



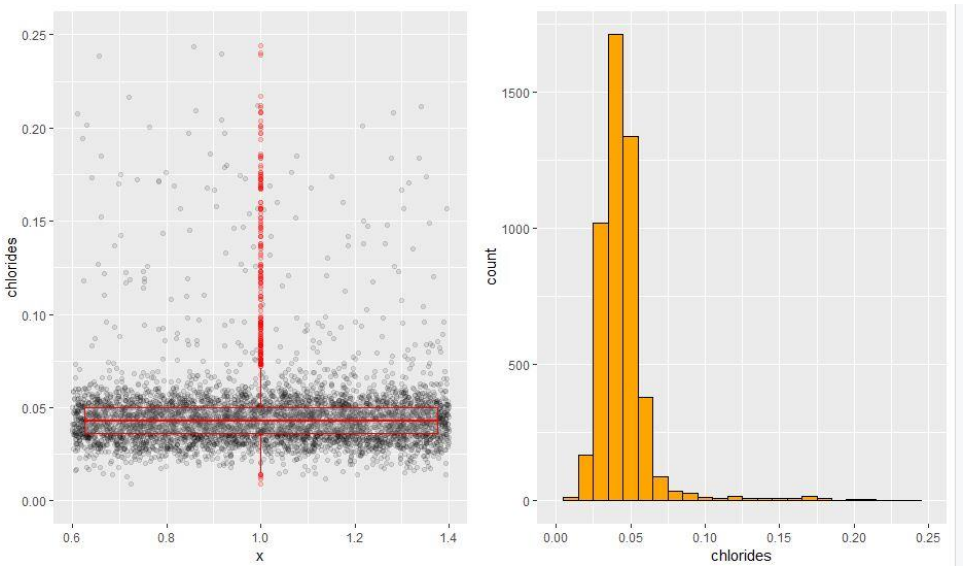

```
> summary(wine$citric.acid)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

Residual Sugar Plot:



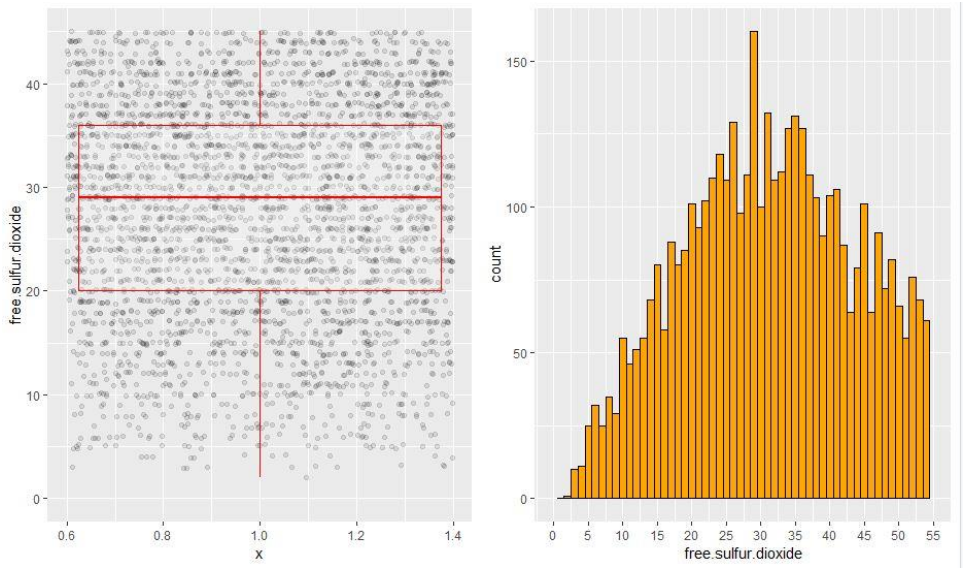

```
> summary(wine$residual.sugar)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    21.0   120.5   133.3   231.0   311.0
```

Chrorides Plot:



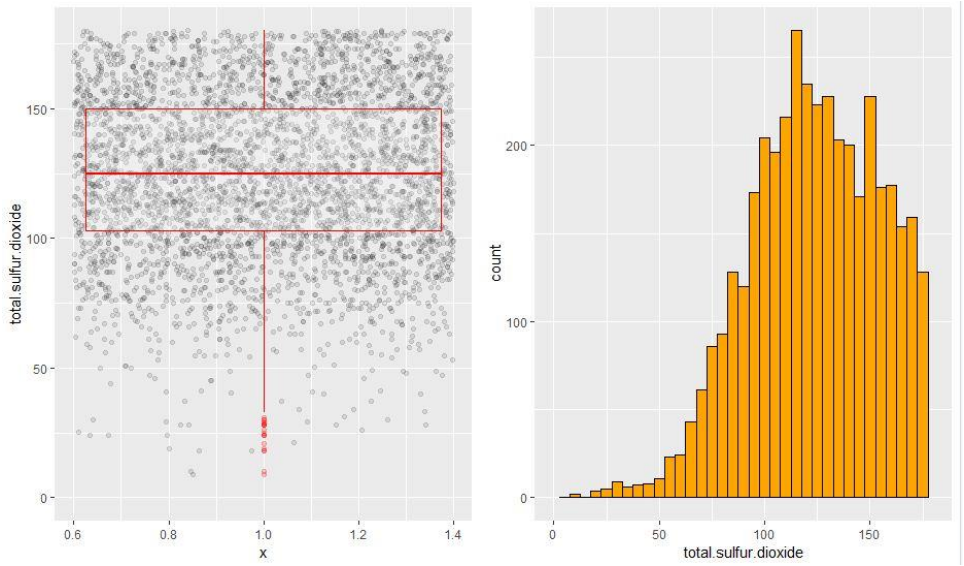

```
> summary(wine$chlorides)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

Free Sulfur Dioxide Plot:




```
summary(wine$free.sulfur.dioxide)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 2.00   23.00   34.00   35.31   46.00  289.00      11
```

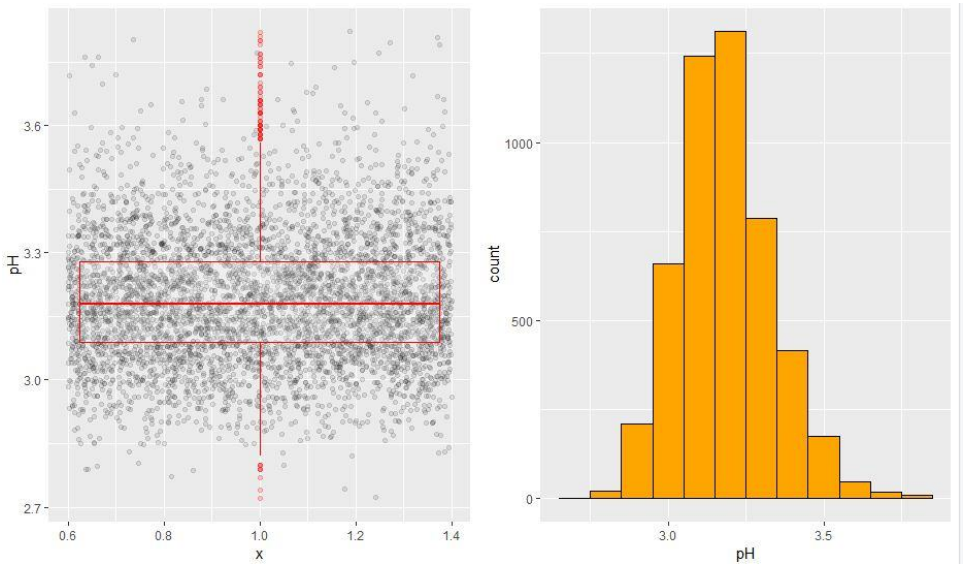
Total Sulfur Dioxide Plot:




```
> summary(wine$total.sulfur.dioxide)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    9.0   108.0   134.0   138.4   167.0   440.0
```

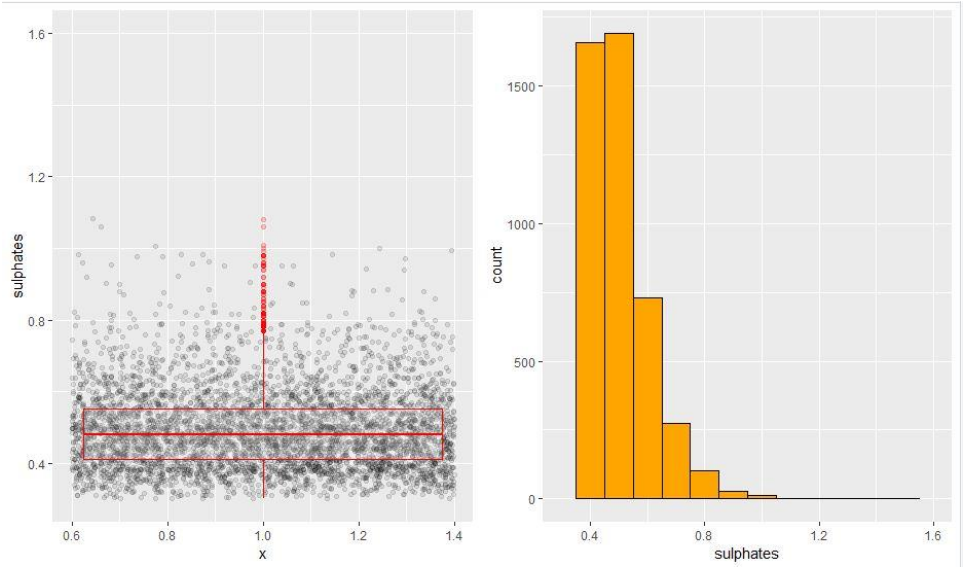
```
summary(wine$density)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
0.9871   0.9917   0.9937   1.5486   0.9961 103.8980        2
```

pH Plot:



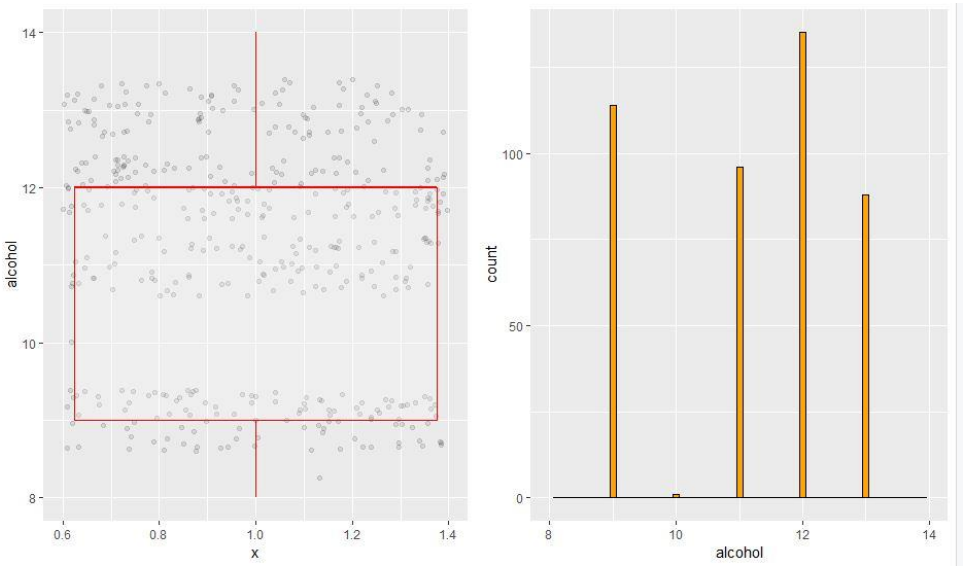

```
summary(wine$pH)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.720   3.090   3.180   3.188   3.280   3.820
```

Sulphates Plot:




```
> summary(wine$sulphates)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.2200  0.4100  0.4700  0.4898  0.5500  1.0800       6
```

Alcohol Plot:




```
summary(wine$alcohol)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   13.00   55.00   53.26   92.00  104.00
```

De este primer análisis se concluye:

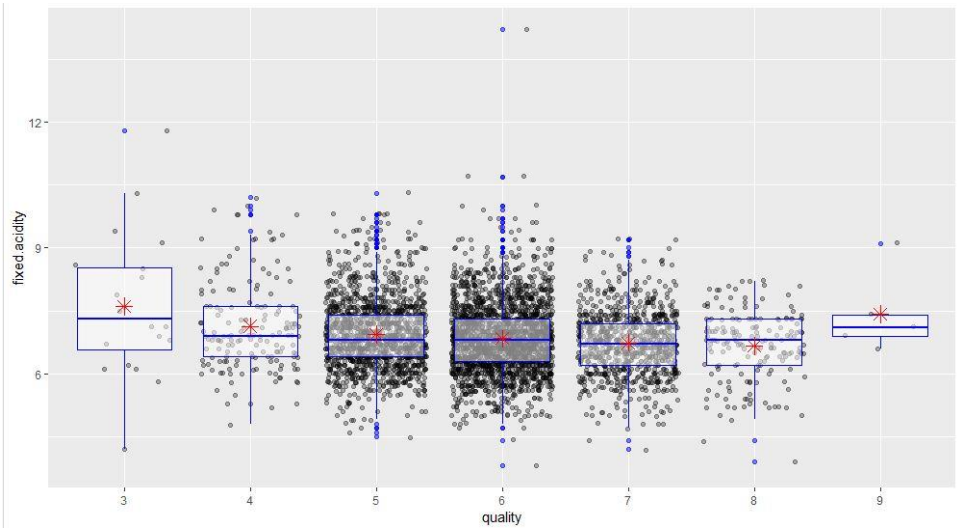
- Prácticamente todas las variables presentan un considerable número de outliers
- Las variables *density* y *pH* presentan una distribución normal, según el histograma.

Análisis bivariante

A continuación vamos a comparar cada una de las variables con la variable response, en nuestro caso quality. Si tenemos en cuenta que queremos responder a la pregunta de cuales son las cualidades que hacen el mejor vino:

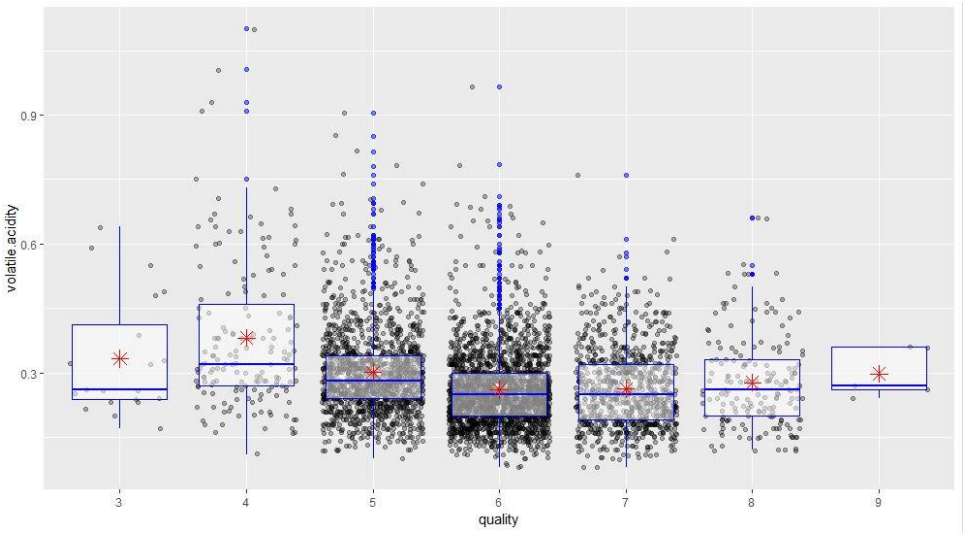
Fixed Acidity – Quality Plot:

Vemos que hay un rango entre 5-8 que nos da una calidad alta y esta variable podría ser significativa para determinar un buen vino.



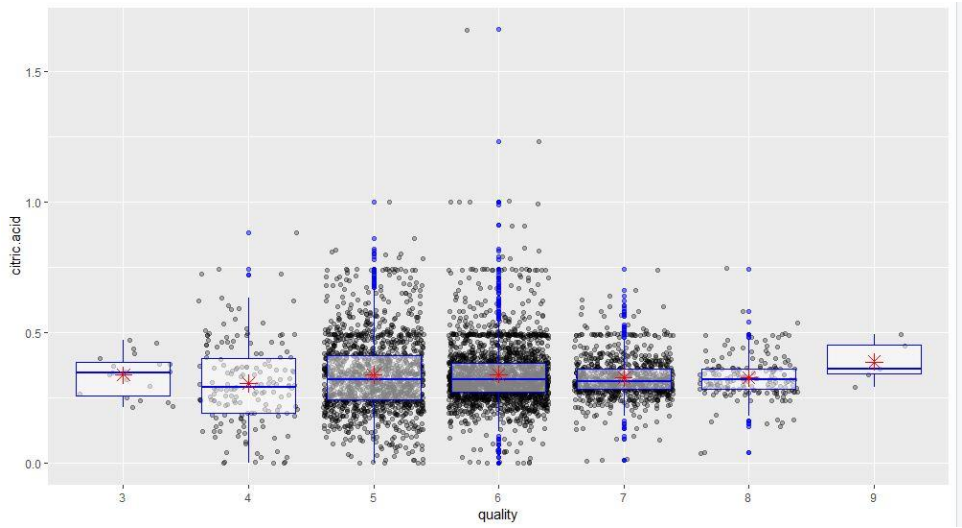
Volatile Acidity – Quality Plot:

Podemos observar como aquí la mayoría de puntos están en el mismo rango, por tanto si nos mantenemos dentro del rango, esta variable no es decisiva para determinar un buen vino.



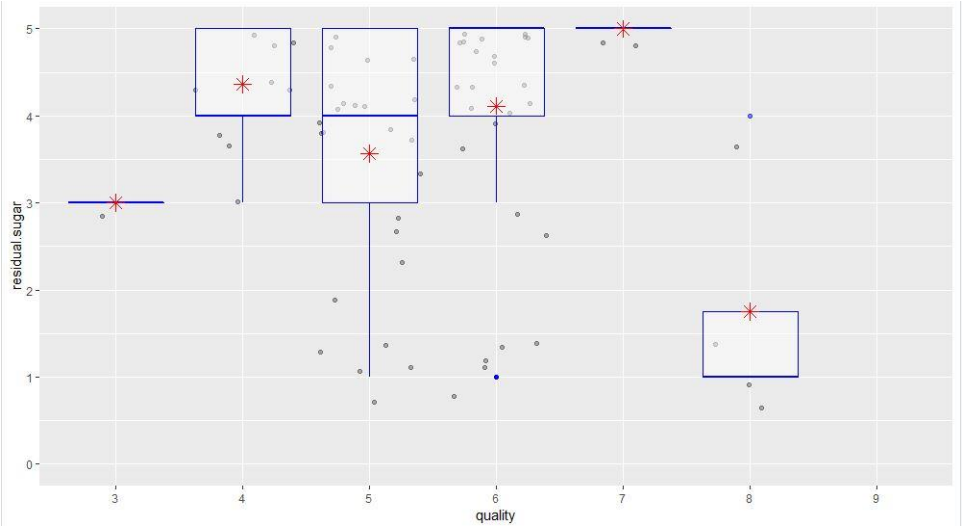
Citric – Quality Plot:

Si observamos los valores obtenidos, parecen estar concentrado en un mismo rango, parece que los de mayor calidad son aquellos que estan en la parte alta de este rango por ello esta variable podría ser significativa en la calidad final del vino.



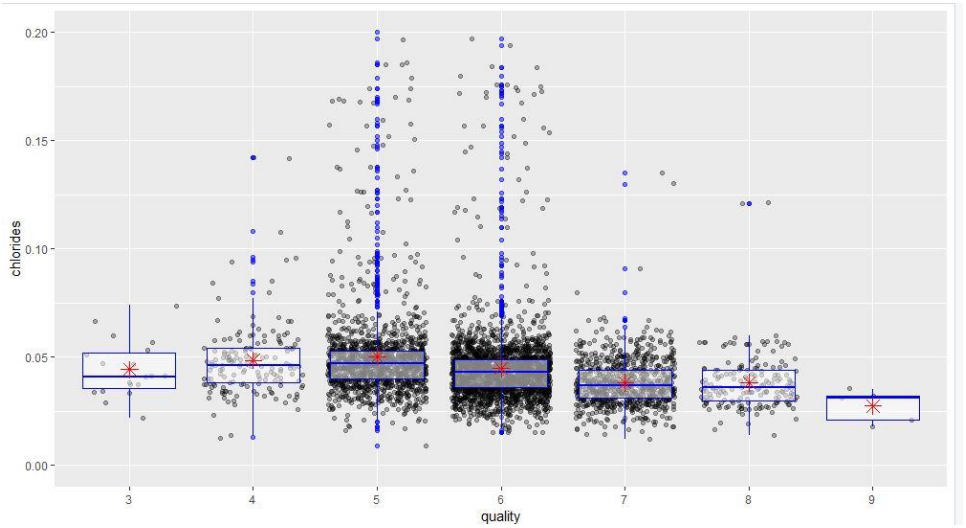
Residual Sugar – Quality Plot:

Es una calidad que obtiene valores muy dispersos, parece que para ser buenos no deben ser valores pequeños pero hay muy pocas muestras pero podría ser significativa para determinar un vino de calidad.



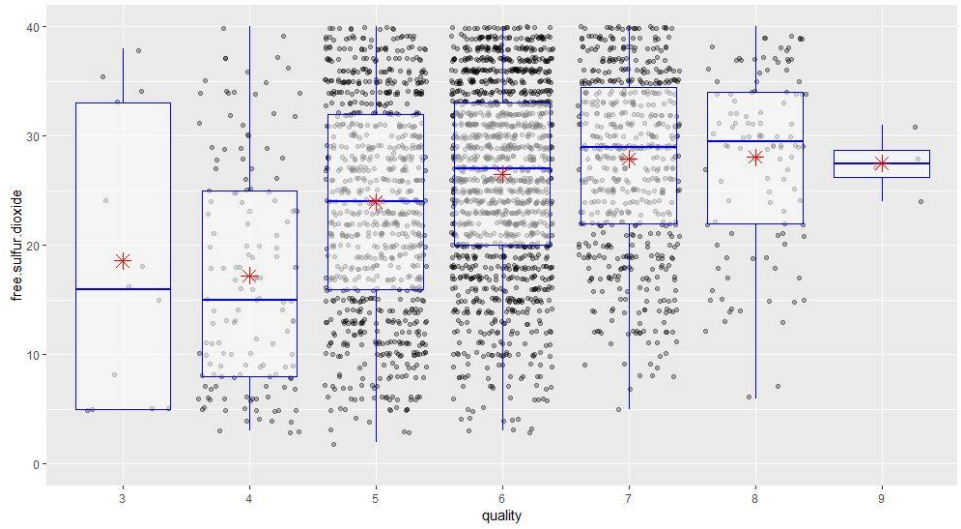
Chrorides – Quality Plot:

Se puede observar que la concentracion de puntos en los vinos de mayor calidad pertenecen a un rango pequeño donde estan todos los demas, es por ello que esta variable pueda no ser decisiva en la calidad final del vino.



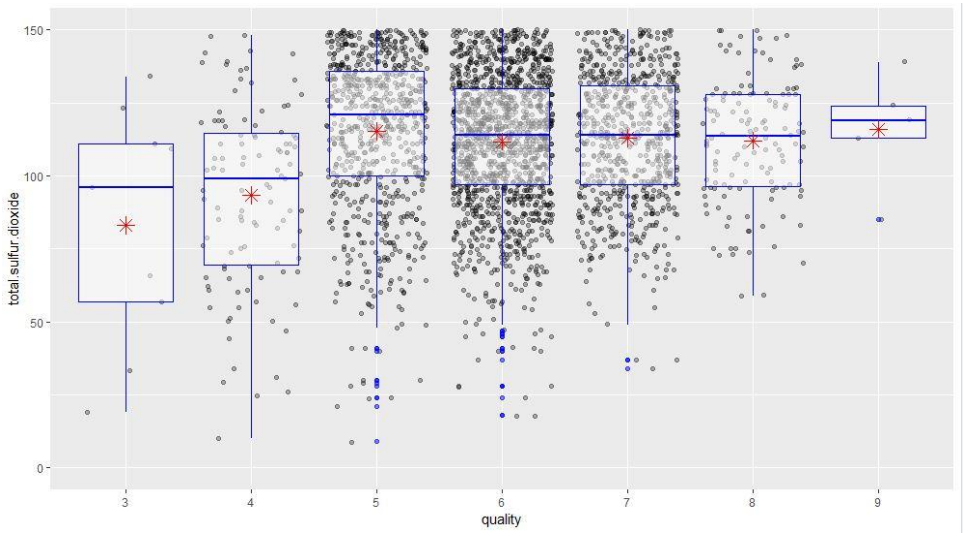
Free Sulfur Dioxide – Quality Plot:

Los valores que se obtienen son muy dispersos y no parece ser significativa para decidir si un vino es bueno o no.



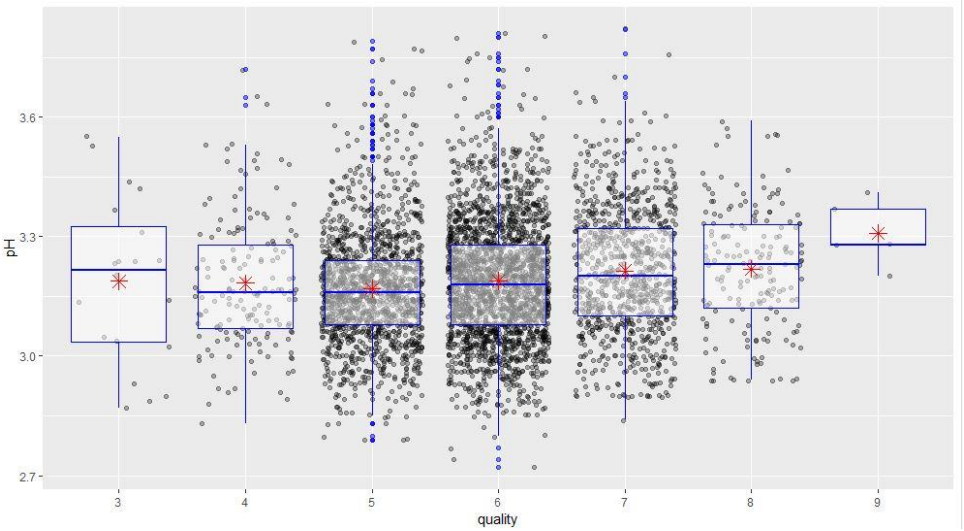
Total Sulfur Dioxide – Quality Plot:

En este caso, hay una gran concentración de valores en un rango amplio pero definido dentro del cual se encuentra dispersos los valores de los vinos clasificados con mejor puntuación.



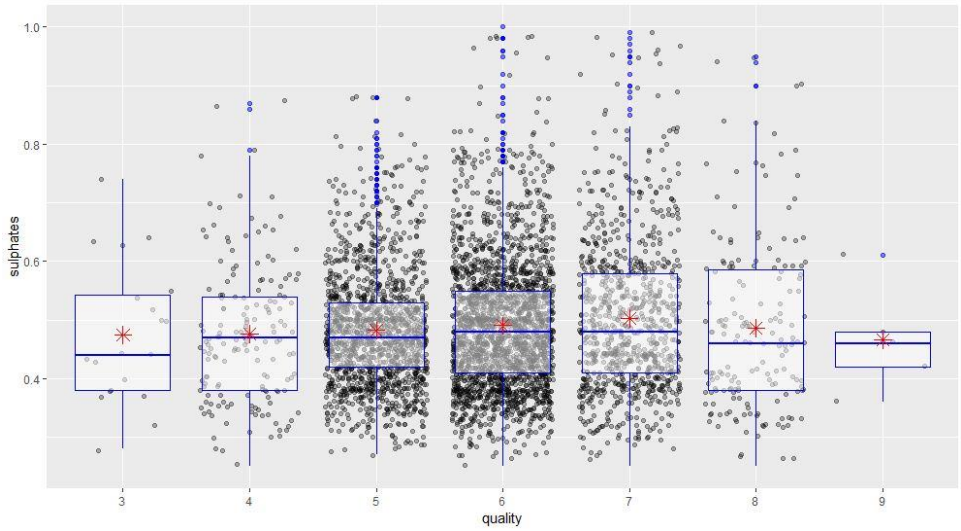
pH – Quality Plot:

Los valores de los vinos mejor clasificados estan bien delimitados alrededor de 3.3 pero este indicador tambien tiene muchos valores en ese pequeño rango pero podría ser significativa para determinar un vino de calidad alta.



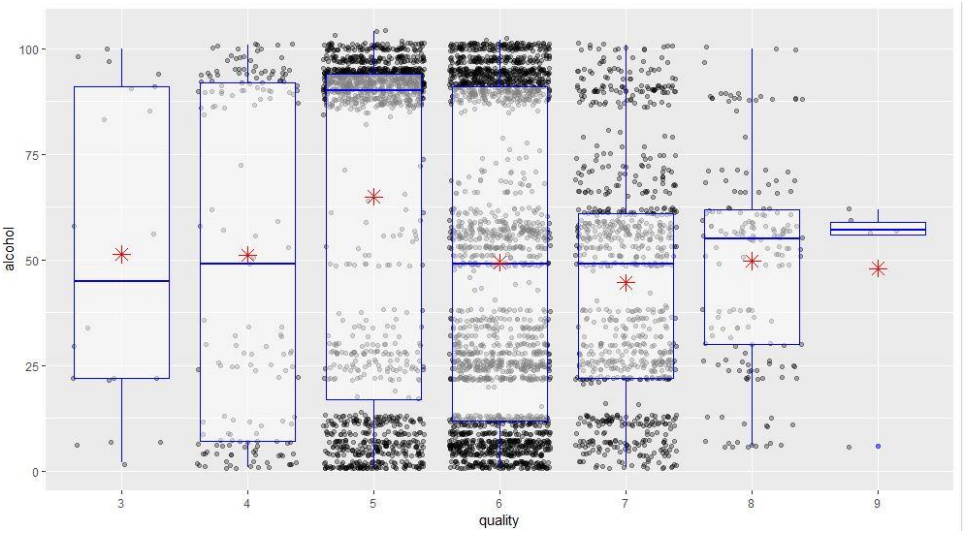
Sulphates – Quality Plot:

Los puntos que representan los vinos mejor puntuados estan repartidos al igual que los demás valores y no representa una variable significativa.



Alcohol – Quality Plot:

Se pueden observar datos concentrados en diversos rangos, podría ser significativa para determinar un vino de calidad.



Las conclusiones principales son:

- Los mejores vinos parecen tener una mayor concentración de citric acid, total.sulfur.dioxide, y un porcentaje menor de chlorides, free.sulfur.dioxide y Ph .
- El fixed acidity, volátil acidity y sulphates parece no tener efecto en la calidad del vino, los valores medios parecen constantes en todas las puntuaciones de la calidad del vino.
- Las variables residual sugar y densitys no son concluyentes dada la naturaleza de su propio valor.
- Valores más bajo de pH implica que el vino sea más ácido.

Análisis de los datos

En función de los datos y el objetivo del estudio aplicar diferentes métodos de análisis.

El objetivo es crear un modelo lineal que explique la calidad del vino o determine si es o no de calidad, a partir de los datos y las variables que disponemos en la base de datos.

Debemos analizar si todas las variables son relevantes y explican nuestras variables objetivo, nuestros *targets.*

Para ello, en primera instancia debemos hacer una serie de comprobaciones que nos ayudarán a determinar los métodos a implementar posteriormente.

Test de normalidad

El objetivo con este test es analizar cuánto difiere la distribución de los datos observados respecto a lo esperado si procediesen de una distribución normal con la misma media y desviación típica.

Implementamos la estrategia gráfica combinada con un contraste de hipótesis.

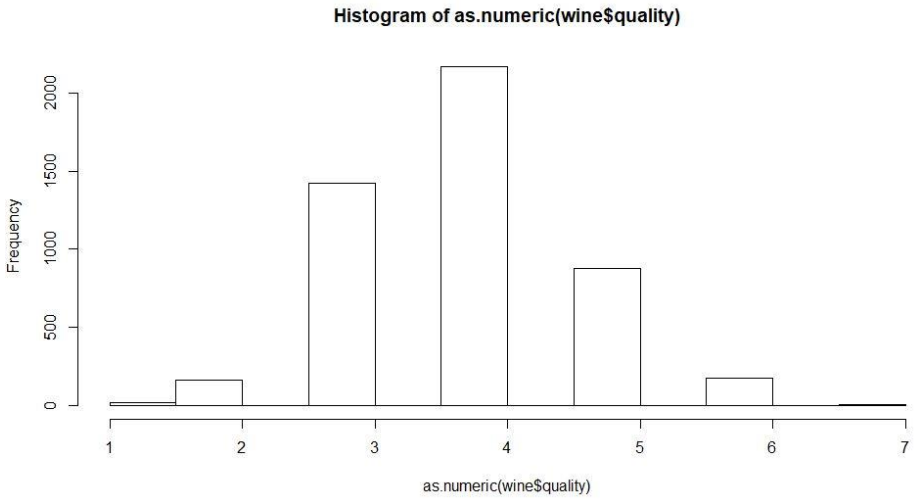
Se considera como hipótesis nula que los datos sí proceden de una distribución normal y como hipótesis alternativa que no lo hacen. El *p-value* de estos test indica la probabilidad de obtener una distribución como la observada si los datos proceden realmente de una población con una distribución normal.

El primer método que utilizaremos para comprobarlo será el algoritmo de Anderson Darling con el obtendremos el p-value para cada variable y así saber si superan el umbral de 0.05 de nivel de significación alpha.

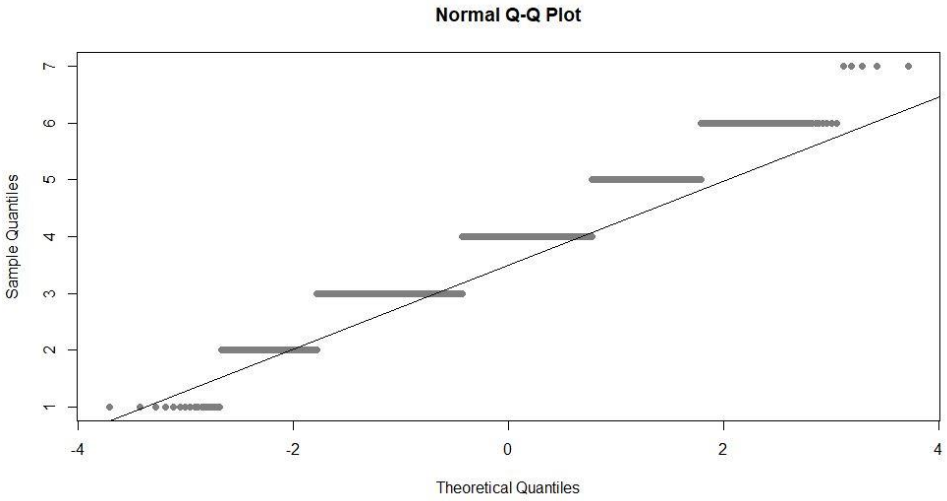
```
> # Algoritmo Anderson Darling:
> alpha = 0.05
>
> col.names = colnames(wine)
>
> for (i in 1:ncol(wine)){
+   if(i == 1) cat("No sigue una distribución normal:\n")
+   if(is.integer(wine[,i]) | is.numeric(wine[,i])){
+     valor = ad.test(wine[,i])$p.value
+     if(valor < alpha){
+       cat(col.names[i])
+       cat("\n")
+
+       if(i < ncol(wine)-1) cat (", ")
+       if( i%% 3 == 0) cat("\n")
+     }
+   }
+ }
No sigue una distribución normal:
fixed.acidity
, volatile.acidity
, citric.acid
,
residual.sugar
, chlorides
, free.sulfur.dioxide
,
total.sulfur.dioxide
, density
, pH
,
sulphates
, alcohol
,
```

Normalidad en la variable respuesta

Histograma Quality:



Q-Qplot Quality:



Método analítico:

```
> jarque.bera.test(x = as.numeric(wine$quality))

        Jarque Bera Test

data:  as.numeric(wine$quality)
X-squared = NaN, df = 2, p-value = NA
```

Los gráficos aparentan seguir una distribución normal pero sin ser del claros en su interpretación.

Por el contrario, las pruebas estadísticas utilizadas presentan p-valores < 0.05, por lo que rechazamos la hipótesis nula, nuestros datos no siguen una distribución normal.

Correlación

Empleamos el test de correlación de Spearman como una alternativa no paramétrica que mide el grado de dependencia entre dos variables.

```
> print (corr_matrix)
                       estimate       p-value
fixed.acidity        -0.08448545  3.183308e-09
volatile.acidity     -0.19654591  7.686512e-44
citric.acid           0.01833273  1.995589e-01
residual.sugar        0.02553427  7.395863e-02
chlorides            -0.31448848 6.907550e-113
free.sulfur.dioxide   0.02358235  9.927460e-02
total.sulfur.dioxide -0.19668029  6.582657e-44
density              -0.34811581 1.707165e-139
pH                    0.10936208  1.656016e-14
sulphates             0.03312253  2.051860e-02
alcohol              -0.16377001  8.626263e-31
```

De este análisis podemos concluir que las variable que guardan una mayor relación con la calidad del vino son volatil acidity, chlorides, total.sulfur.dioxide, sulphates y alcohol.

Modelo de regresión lineal

La regresión lineal es un modelo matemático que tiene como objetivo aproximar la relación de dependencia lineal entre una variable dependiente y una o una serie de variables independientes.

Mediante el método Backward se procede a la estimación del modelo.

Se hace de forma manual por lo que se ajusta el modelo completo y mediante la comparación del modelo completo contra el modelo sin cada una de las variables se decide qué variable se elimina de forma progresiva.

Luego de varios ajustes[1] y análisis, se concluye que la expresión del modelo que mejor explica calidad del vino, con un R^2 ajustado de 0.8303, es el siguiente:

```
> summary(m4 <- lm(as.numeric(quality)~(.-density-pH-sulphates-rating),wine))

Call:
lm(formula = as.numeric(quality) ~ (. - density - pH - sulphates -
    rating), data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5699 -0.6172 -0.0319  0.5021  3.2033

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.4960663  0.1109050  49.557  < 2e-16 ***
fixed.acidity        -0.1025738  0.0149381  -6.867 7.40e-12 ***
volatile.acidity     -1.3069853  0.1235119 -10.582  < 2e-16 ***
citric.acid           0.2505327  0.1052917   2.379   0.0174 *
residual.sugar        0.0006130  0.0001157   5.296 1.24e-07 ***
chlorides            -6.3801817  0.5684413 -11.224  < 2e-16 ***
free.sulfur.dioxide   0.0067716  0.0009141   7.408 1.50e-13 ***
total.sulfur.dioxide -0.0040471  0.0003809 -10.624  < 2e-16 ***
alcohol              -0.0019538  0.0003464  -5.640 1.79e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8303 on 4869 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1236,    Adjusted R-squared:  0.1222
F-statistic: 85.86 on 8 and 4869 DF,  p-value: < 2.2e-16
```

[1] En el anexo se adjuntan todas la salidas de los modelos intermedios

```
> Anova(m4)
Anova Table (Type II tests)

Response: as.numeric(quality)
                      Sum Sq   Df  F value    Pr(>F)
fixed.acidity           32.5    1  47.1504 7.399e-12 ***
volatile.acidity        77.2    1 111.9758 < 2.2e-16 ***
citric.acid              3.9    1   5.6616   0.01738 *
residual.sugar          19.3    1  28.0457 1.237e-07 ***
chlorides               86.9    1 125.9781 < 2.2e-16 ***
free.sulfur.dioxide     37.8    1  54.8785 1.503e-13 ***
total.sulfur.dioxide    77.8    1 112.8784 < 2.2e-16 ***
alcohol                 21.9    1  31.8139 1.793e-08 ***
Residuals             3356.9 4869
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Una vez tenemos estimado el modelo, analizamos la correlación entre las regresoras para detectar posibles problemas de multicolinealidad.

```
> mfinal <- m4
> # DETECCIÓN DE MULTICOLINEALIDAD
> ################################
> vif(mfinal)
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
            1.124325             1.097552             1.149142             1.039778             1.089804
 free.sulfur.dioxide total.sulfur.dioxide              alcohol
            1.712048             1.855010             1.130691
```

Tabla de Factor de Incremento de la Varianza

La causa de esta elevada correlación es que, como podemos apreciar, tenemos dos variables que recogen información muy similar, estamos hablando de las variables de *free.sulfur.dioxide* y *total.sulfur.dioxide* ya que tienen un valor VIF[2] > 1.

Ante la presencia de un problema de multicolinealidad, decidimos eliminar una de las variables. Sabemos que sacrificaremos el R^2, pero a cambio no hay redundancia y la interpretación es mejor. Para decidir qué variable se elimina, se crean dos nuevos modelos, donde en cada uno de los modelos se elimina una de las variables y se comparan ambos modelos. Una vez comparada la validación del modelo, observamos que el resultado del R^2ajustado más elevado, implicando un mejor ajuste es quitando del modelo la variable total.sulfur.dioxide, con un R^2 ajustado de 0.8398, es:

[2] VIF = 1: Ausencia total de colinealidad
1 < VIF < 5: La regresión puede verse afectada por cierta colinealidad.
5 < VIF < 10: Causa de preocupación
El termino tolerancia es 1/VIF por lo que los límites recomendables están entre 1 y 0.1.

```
> summary(mt <- lm(as.numeric(quality)~(.-density-pH-sulphates-total.sulfur.dioxide-rating),wine))

Call:
lm(formula = as.numeric(quality) ~ (. - density - pH - sulphates -
    total.sulfur.dioxide - rating), data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3384 -0.6449 -0.0138  0.4399  3.2207

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.4737852  0.1121517  48.807  < 2e-16 ***
fixed.acidity       -0.1261102  0.0149416  -8.440  < 2e-16 ***
volatile.acidity    -1.5109736  0.1234039 -12.244  < 2e-16 ***
citric.acid          0.1964994  0.1063700   1.847 0.064761 .
residual.sugar       0.0004499  0.0001160   3.877 0.000107 ***
chlorides           -7.1396266  0.5703698 -12.518  < 2e-16 ***
free.sulfur.dioxide  0.0007996  0.0007291   1.097 0.272822
alcohol             -0.0025915  0.0003450  -7.511 6.96e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8398 on 4870 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1033,    Adjusted R-squared:  0.102
F-statistic: 80.16 on 7 and 4870 DF,  p-value: < 2.2e-16


> Anova(mt)
Anova Table (Type II tests)

Response: as.numeric(quality)
                     Sum Sq   Df  F value    Pr(>F)
fixed.acidity          50.2    1  71.2370 < 2.2e-16 ***
volatile.acidity      105.7    1 149.9187 < 2.2e-16 ***
citric.acid             2.4    1   3.4126 0.0647614 .
residual.sugar         10.6    1  15.0327 0.0001071 ***
chlorides             110.5    1 156.6888 < 2.2e-16 ***
free.sulfur.dioxide     0.8    1   1.2028 0.2728218
alcohol                39.8    1  56.4103 6.957e-14 ***
Residuals            3434.7 4870
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusiones

Como hemos ido viendo a lo largo del análisis, existen ciertas variables que mejor explican la calidad del vino, siendo la mejor forma de predecir la calidad del vino siguiendo el modelo generado, donde intervienen las variables de fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides y free.sulfur.dioxide

Contribuciones

El desarrollo de este trabajo se hizo en equipo. Los intergrandes somos Carlos Herrero y Montse Rodríguez, estudiantes del máster Data Science.

Si bien se realizó en la modalidad ‘a distancia’, hemos llevado a cabo distintas videollamadas e intercambio tanto de información como de contenido via whats upp y mail.

Contribuciones	Firma
Investigación previa	Carlos Herrero, Montse Rodriguez
Redacción de las respuestas	Carlos Herrero, Montse Rodriguez
Desarrollo código	Carlos Herrero, Montse Rodriguez

Anexo

```
> summary(m1 <- lm(as.numeric(quality)~(.-rating),wine))


Call:

lm(formula = as.numeric(quality) ~ (. - rating), data = wine)


Residuals:

    Min      1Q  Median      3Q     Max

-3.6445 -0.6022 -0.0272  0.5079  3.1903


Coefficients:

                       Estimate Std. Error t value Pr(>|t|)

(Intercept)           4.537e+00  3.550e-01  12.780  < 2e-16 ***

fixed.acidity        -8.386e-02  1.649e-02  -5.085 3.82e-07 ***

volatile.acidity     -1.273e+00  1.234e-01 -10.322  < 2e-16 ***

citric.acid           2.343e-01  1.053e-01   2.225   0.0262 *

residual.sugar        6.571e-04  1.157e-04   5.680 1.43e-08 ***

chlorides            -6.290e+00  5.680e-01 -11.073  < 2e-16 ***

free.sulfur.dioxide   7.132e-03  9.151e-04   7.794 7.89e-15 ***

total.sulfur.dioxide -4.439e-03  3.878e-04 -11.446  < 2e-16 ***

density              -5.622e-05  1.827e-03  -0.031   0.9755

pH                    1.868e-01  9.046e-02   2.066   0.0389 *

sulphates             5.092e-01  1.072e-01   4.752 2.08e-06 ***

alcohol              -1.709e-03  3.522e-04  -4.851 1.27e-06 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.828 on 4866 degrees of freedom

  (20 observations deleted due to missingness)

Multiple R-squared:  0.1291,  Adjusted R-squared:  0.1271

F-statistic: 65.56 on 11 and 4866 DF,  p-value: < 2.2e-16
```

```
> Anova(m1)
Anova Table (Type II tests)

Response: as.numeric(quality)
                      Sum Sq   Df  F value    Pr(>F)
fixed.acidity           17.7    1  25.8552 3.818e-07 ***
volatile.acidity        73.0    1 106.5466 < 2.2e-16 ***
citric.acid              3.4    1   4.9484   0.02616 *
residual.sugar          22.1    1  32.2576 1.429e-08 ***
chlorides               84.1    1 122.6208 < 2.2e-16 ***
free.sulfur.dioxide     41.6    1  60.7442 7.892e-15 ***
total.sulfur.dioxide    89.8    1 131.0153 < 2.2e-16 ***
density                  0.0    1   0.0009   0.97545
pH                       2.9    1   4.2663   0.03893 *
sulphates               15.5    1  22.5787 2.075e-06 ***
alcohol                 16.1    1  23.5317 1.267e-06 ***
Residuals             3336.0 4866
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(m2 <- lm(as.numeric(quality)~(.-density-rating),wine))

Call:
lm(formula = as.numeric(quality) ~ (. - density - rating), data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6444 -0.6023 -0.0272  0.5079  3.1904

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.5374350  0.3549287  12.784  < 2e-16 ***
fixed.acidity        -0.0838877  0.0164690  -5.094 3.64e-07 ***
```

```
volatile.acidity     -1.2732619  0.1233396 -10.323  < 2e-16 ***
citric.acid           0.2341848  0.1052703   2.225   0.0262 *
residual.sugar        0.0006571  0.0001157   5.680 1.42e-08 ***
chlorides            -6.2898394  0.5679509 -11.075  < 2e-16 ***
free.sulfur.dioxide   0.0071325  0.0009150   7.796 7.79e-15 ***
total.sulfur.dioxide -0.0044395  0.0003871 -11.470  < 2e-16 ***
pH                    0.1868672  0.0904409   2.066   0.0389 *
sulphates             0.5090518  0.1070270   4.756 2.03e-06 ***
alcohol              -0.0017090  0.0003519  -4.857 1.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8279 on 4867 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1291,  Adjusted R-squared:  0.1273
F-statistic: 72.13 on 10 and 4867 DF,  p-value: < 2.2e-16

> Anova(m2)
Anova Table (Type II tests)

Response: as.numeric(quality)
                     Sum Sq   Df  F value    Pr(>F)
fixed.acidity          17.8    1  25.9456 3.644e-07 ***
volatile.acidity       73.0    1 106.5689 < 2.2e-16 ***
citric.acid             3.4    1   4.9489   0.02615 *
residual.sugar         22.1    1  32.2634 1.425e-08 ***
chlorides              84.1    1 122.6472 < 2.2e-16 ***
free.sulfur.dioxide    41.7    1  60.7699 7.791e-15 ***
total.sulfur.dioxide   90.2    1 131.5562 < 2.2e-16 ***
pH                      2.9    1   4.2691   0.03886 *
sulphates              15.5    1  22.6223 2.029e-06 ***
alcohol                16.2    1  23.5875 1.231e-06 ***
```

```
Residuals            3336.0 4867
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(m3 <- lm(as.numeric(quality)~(.-density-pH-rating),wine))

Call:
lm(formula = as.numeric(quality) ~ (. - density - pH - rating),
    data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6140 -0.6050 -0.0262  0.5035  3.2198

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.2256661  0.1226104  42.620  < 2e-16 ***
fixed.acidity        -0.0983036  0.0149230  -6.587 4.95e-11 ***
volatile.acidity     -1.2836342  0.1232788 -10.412  < 2e-16 ***
citric.acid           0.2227866  0.1051610   2.119   0.0342 *
residual.sugar        0.0006498  0.0001157   5.618 2.04e-08 ***
chlorides            -6.3646058  0.5669873 -11.225  < 2e-16 ***
free.sulfur.dioxide   0.0069930  0.0009128   7.661 2.21e-14 ***
total.sulfur.dioxide -0.0043418  0.0003843 -11.298  < 2e-16 ***
sulphates             0.5415076  0.1059035   5.113 3.29e-07 ***
alcohol              -0.0018403  0.0003462  -5.316 1.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8282 on 4868 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1283,  Adjusted R-squared:  0.1267
F-statistic: 79.62 on 9 and 4868 DF,  p-value: < 2.2e-16
```

```
> Anova(m3)
Anova Table (Type II tests)

Response: as.numeric(quality)
                      Sum Sq   Df  F value    Pr(>F)
fixed.acidity           29.8    1  43.3936 4.950e-11 ***
volatile.acidity        74.4    1 108.4191 < 2.2e-16 ***
citric.acid              3.1    1   4.4882   0.03418 *
residual.sugar          21.6    1  31.5601 2.041e-08 ***
chlorides               86.4    1 126.0075 < 2.2e-16 ***
free.sulfur.dioxide     40.3    1  58.6965 2.206e-14 ***
total.sulfur.dioxide    87.6    1 127.6493 < 2.2e-16 ***
sulphates               17.9    1  26.1450 3.289e-07 ***
alcohol                 19.4    1  28.2572 1.110e-07 ***
Residuals             3338.9 4868
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(m4 <- lm(as.numeric(quality)~(.-density-pH-sulphates-rating),wine))

Call:
lm(formula = as.numeric(quality) ~ (. - density - pH - sulphates -
    rating), data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5699 -0.6172 -0.0319  0.5021  3.2033

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.4960663  0.1109050  49.557  < 2e-16 ***
fixed.acidity        -0.1025738  0.0149381  -6.867 7.40e-12 ***
```

```
volatile.acidity     -1.3069853  0.1235119 -10.582  < 2e-16 ***
citric.acid           0.2505327  0.1052917   2.379   0.0174 *
residual.sugar        0.0006130  0.0001157   5.296 1.24e-07 ***
chlorides            -6.3801817  0.5684413 -11.224  < 2e-16 ***
free.sulfur.dioxide   0.0067716  0.0009141   7.408 1.50e-13 ***
total.sulfur.dioxide -0.0040471  0.0003809 -10.624  < 2e-16 ***
alcohol              -0.0019538  0.0003464  -5.640 1.79e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8303 on 4869 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1236,  Adjusted R-squared:  0.1222
F-statistic: 85.86 on 8 and 4869 DF,  p-value: < 2.2e-16

> Anova(m4)
Anova Table (Type II tests)

Response: as.numeric(quality)
                     Sum Sq   Df  F value    Pr(>F)
fixed.acidity          32.5    1  47.1504 7.399e-12 ***
volatile.acidity       77.2    1 111.9758 < 2.2e-16 ***
citric.acid             3.9    1   5.6616   0.01738 *
residual.sugar         19.3    1  28.0457 1.237e-07 ***
chlorides              86.9    1 125.9781 < 2.2e-16 ***
free.sulfur.dioxide    37.8    1  54.8785 1.503e-13 ***
total.sulfur.dioxide   77.8    1 112.8784 < 2.2e-16 ***
alcohol                21.9    1  31.8139 1.793e-08 ***
Residuals            3356.9 4869
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```