Practica 2 - Tipología y Ciclo de vida de datos

Nacho Serrano

5 de enero de 2019

Índice

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

El dataset elegido para el análisis ha sido obtenido a través de este enlace de Kaggle. Los datos que se han recolectado en este dataset son las características químicas de 1599 vinos, concretamente sobre el vino verde de Portugal. El dataset contiene 12 columnas que corresponden con las siguientes características:

- **fixed acidity**: la acidez relacionada con el vino que puede ser fija o no volátil, es decir, que no se evapora fácilmente.
- **volatile acidity**: la cantidad de ácido acético que hay en el vino, en niveles demasiado altos puede provocar un sabor desagradable a vinagre.
- **citric acid**: la cantidad de ácido cítrico que hay en el vino. Si se encuentra en pequeñas cantidades, puede agregar 'frescura' y sabor a los vinos.
- **residual sugar**: 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**: la cantidad de sal que hay en el vino.
- **free sulfur dioxide**: 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 superiores a 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino.
- **Density**: la densidad del agua es cercana a la del agua según el porcentaje de alcohol y contenido de azúcar.
- **pH**: describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico).
- **sulphates**: Muestra la cantidad de sulfatos que hay en el vino. Los sulfatos son un aditivo para el vino que puede contribuir a los niveles de dióxido de azufre (S02), que actúa como antimicrobiano y antioxidante.
- **alcohol**: el porcentaje de alcohol del vino.
- **quality**: Calidad del vino determinada por etnólogos, puntuación marcada a través de paramétros sensoriales y no químicos.

El objetivo es comparar las características *volatile acidity*, *Residual sugar*, *Chlorides*, *Density* y *pH*, para determinar la calidad del vino.

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

```
wines<- read.csv("~/Master - Data Science/Tipología y ciclo de vida de los datos/Practica 2/winequality-
head(wines)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
```

```
## 2                 25                  67  0.9968 3.20      0.68     9.8
## 3                 15                  54  0.9970 3.26      0.65     9.8
## 4                 17                  60  0.9980 3.16      0.58     9.8
## 5                 11                  34  0.9978 3.51      0.56     9.4
## 6                 13                  40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

Como he comentado anteriormente no nos interesa todas las características del dataset. Procedemos a quedarnos con las características: *volatile acidity*, *Residual sugar*, *Chlorides*, *Density* y *pH*

```
new_wines<-data.frame(wines[,c("volatile.acidity","residual.sugar","chlorides","density","pH")])
head(new_wines)
```

```
##   volatile.acidity residual.sugar chlorides density   pH
## 1             0.70            1.9     0.076  0.9978 3.51
## 2             0.88            2.6     0.098  0.9968 3.20
## 3             0.76            2.3     0.092  0.9970 3.26
## 4             0.28            1.9     0.075  0.9980 3.16
## 5             0.70            1.9     0.076  0.9978 3.51
## 6             0.66            1.8     0.075  0.9978 3.51
```

Ahora en la variable *new_wines* contiene las características que queremos analizar para determinar la calidad del vino.

3. Limpieza de los datos.

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

```
#Para saber si hay elementos vacios
sapply(new_wines, function(x) is.null(x))
```

```
## volatile.acidity   residual.sugar        chlorides          density
##            FALSE            FALSE            FALSE            FALSE
##               pH
##            FALSE
```

El valor FALSE en todas las columnas nos indica que no hay elementos vacios.

```
#Para saber si hay Not Available
sapply(new_wines, function(x) sum(is.na(x)))
```

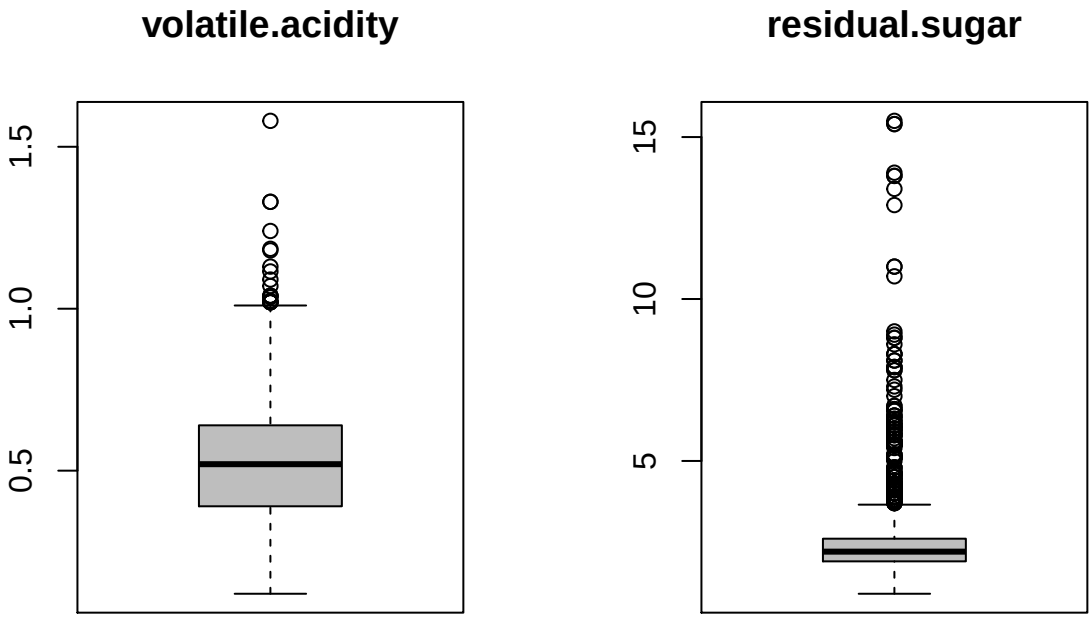
```
## volatile.acidity   residual.sugar        chlorides          density
##                0                0                0                0
##               pH
##                0
```

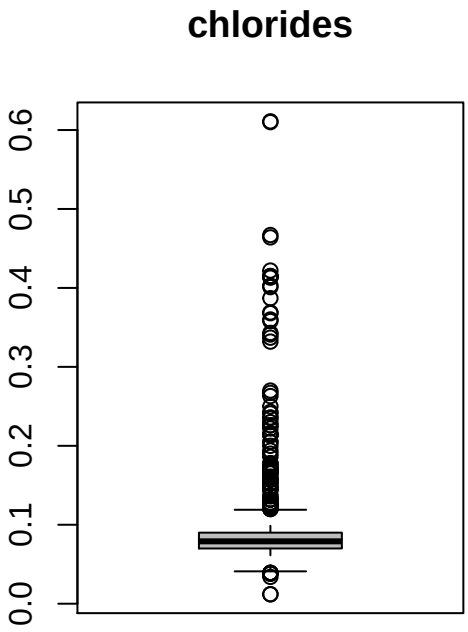
Los resultados de la comprobación nos indica que no hay ningun valor Na.

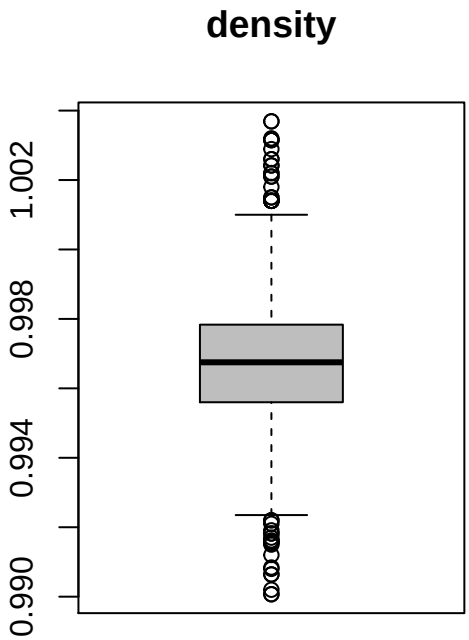
En este caso, el dataset no tiene ni valores vacios ni missing.

3.2. Identificación y tratamiento de valores extremos.

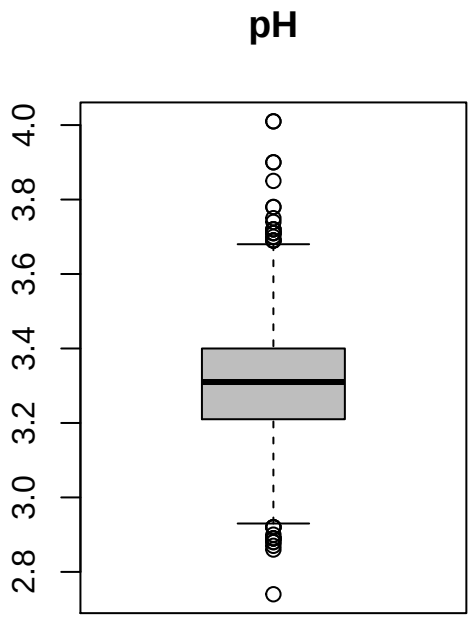
```
par(mfrow=c(1,2))
for (i in 1:5){
  boxplot(new_wines[,i], main=names(new_wines)[i],col="gray")
}
```








```
par(mfrow=c(1,1))
```



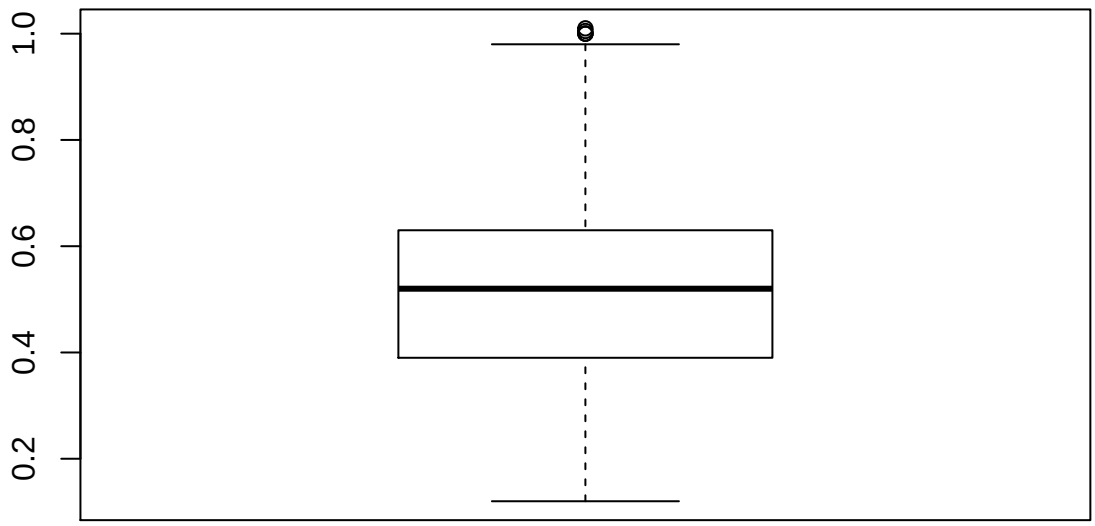
Se observa que en todas las variables existen valores fuera de rango. Dado que tenemos una muestra de 1599 (es muy grande) vamos a eliminar los valores fuera de rango.

Volatile Acidity

```
#Los valores fuera de rango son:
boxplot.stats(new_wines$volatile.acidity)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020
## [12] 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

```
#Eliminamos del recordset los valores outliers
new_wines<-data.frame(new_wines[!new_wines$volatile.acidity %in% boxplot.stats(new_wines$volatile.acidi
boxplot(new_wines$volatile.acidity)
```



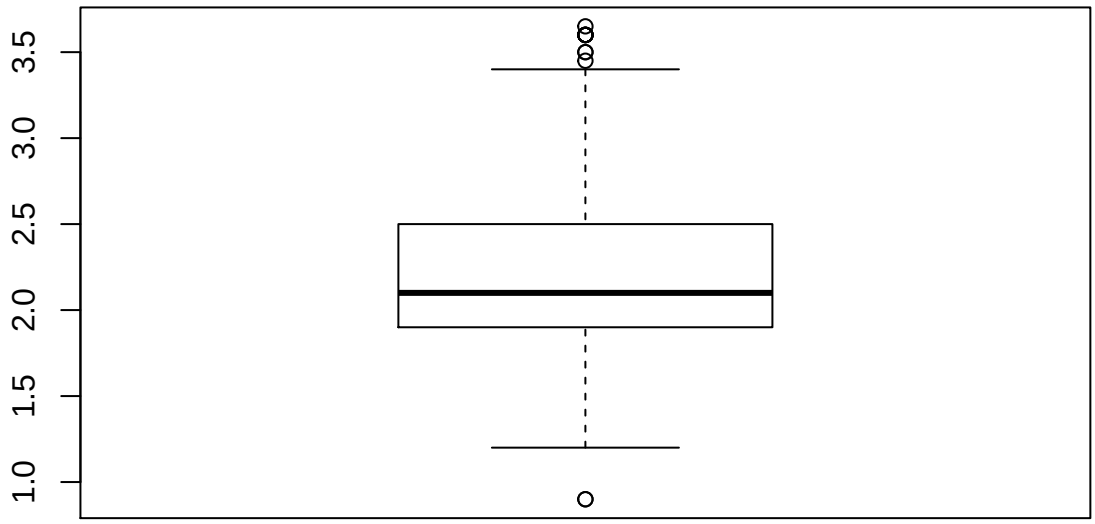
Se observa que despues de eliminar los valores fuera de rango, con los nuevos valores aparecen nuevos valores fuera de rango. En este caso, no voy a eliminarlos porque son valores que, antes de eliminar los outliers anteriores, estaban dentro del rango.

Residual Sugar

```
#Los valores fuera de rango son:
boxplot.stats(new_wines$residual.sugar)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10
##  [12]  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00
##  [23]  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00
##  [34]  4.50  4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70
##  [45]  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60
##  [56]  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30
##  [67]  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  3.90
##  [78]  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00  4.60  8.80
##  [89]  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00  3.90  4.00
## [100]  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50  4.30  5.50
## [111]  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90  4.30 13.40  4.80
## [122]  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80  6.10  3.90
## [133]  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75 13.80 13.80
## [144]  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10  7.80
```

```
#Eliminamos del recordset los valores outliers
new_wines<-data.frame(new_wines[!new_wines$residual.sugar %in% boxplot.stats(new_wines$residual.sugar)$
boxplot(new_wines$residual.sugar)
```



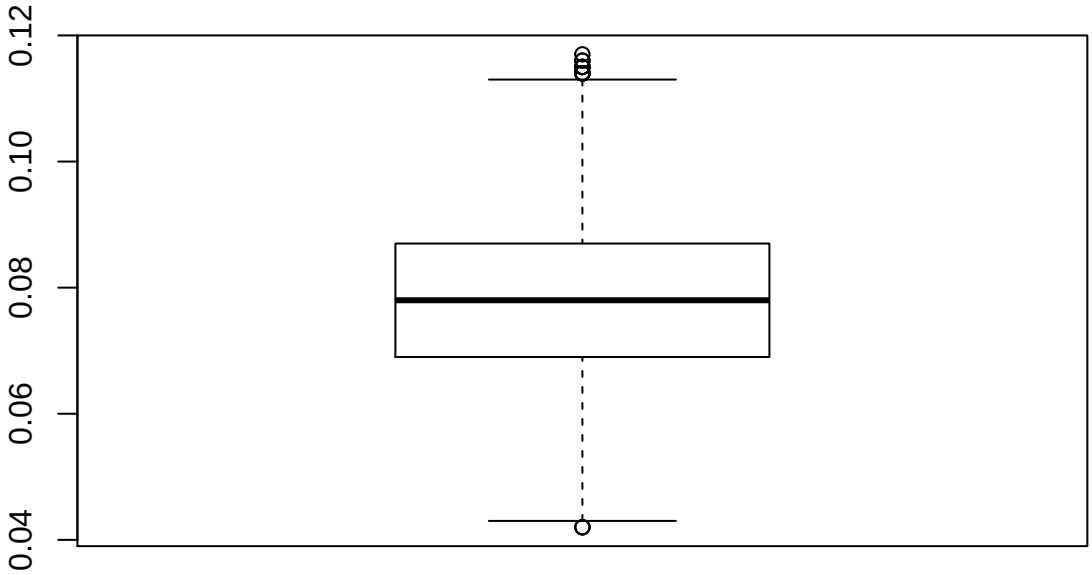
Igual que en la anterior característica, se observa que aparecen nuevos valores fuera de rango. Por el mismo motivo que en la anterior característica no los voy a eliminar.

Chlorides

```
#Los valores fuera de rango son:
boxplot.stats(new_wines$chlorides)$out
```

```
##   [1] 0.368 0.341 0.332 0.464 0.401 0.467 0.122 0.119 0.119 0.146 0.118
##  [12] 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343 0.186
##  [23] 0.213 0.214 0.121 0.122 0.122 0.128 0.118 0.118 0.159 0.121 0.127
##  [34] 0.413 0.152 0.152 0.125 0.200 0.171 0.226 0.226 0.250 0.148 0.122
##  [45] 0.143 0.222 0.039 0.157 0.422 0.034 0.387 0.415 0.157 0.157 0.243
##  [56] 0.241 0.190 0.132 0.126 0.038 0.041 0.165 0.145 0.147 0.012 0.012
##  [67] 0.119 0.039 0.194 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171
##  [78] 0.178 0.118 0.118 0.369 0.041 0.166 0.166 0.136 0.132 0.132 0.123
##  [89] 0.123 0.123 0.403 0.041 0.414 0.041 0.166 0.415 0.153 0.415 0.267
## [100] 0.169 0.039 0.230 0.038 0.118
```

```
#Eliminamos del recordset los valores outliers
new_wines<-data.frame(new_wines[!new_wines$chlorides %in% boxplot.stats(new_wines$chlorides)$out,])
boxplot(new_wines$chlorides)
```

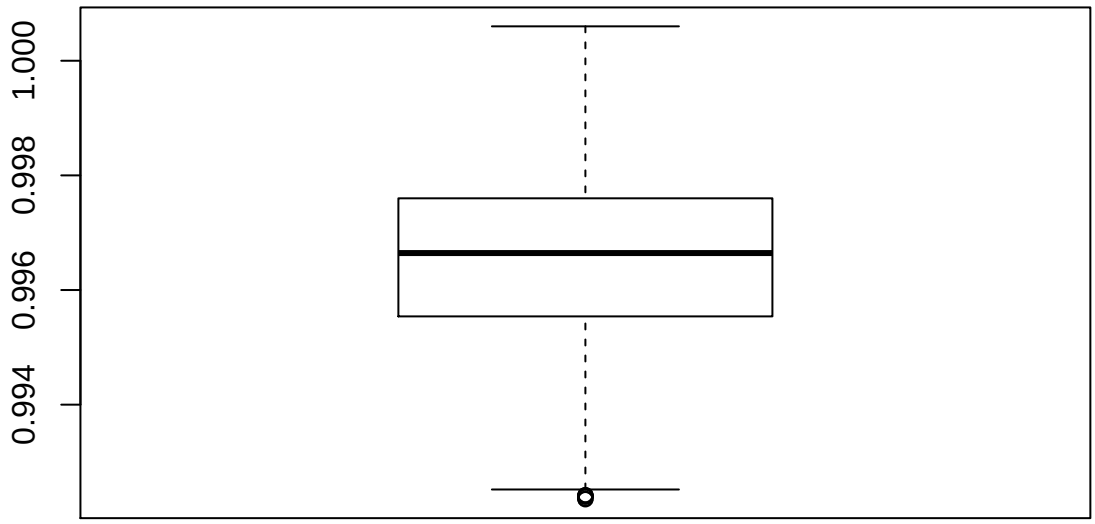


Density

```
#Los valores fuera de rango son:
boxplot.stats(new_wines$density)$out
```

```
##  [1] 0.99160 0.99160 1.00100 1.00140 1.00080 0.99120 1.00100 1.00140
##  [9] 1.00140 1.00080 1.00080 0.99170 0.99220 1.00100 1.00100 0.99154
## [17] 0.99162 0.99007 0.99007 0.99220 0.99150 0.99157 0.99080 0.99084
## [25] 0.99191 0.99182 0.99182
```

```
#Eliminamos del recordset los valores outliers
new_wines<-data.frame(new_wines[!new_wines$density %in% boxplot.stats(new_wines$density)$out,])
boxplot(new_wines$density)
```

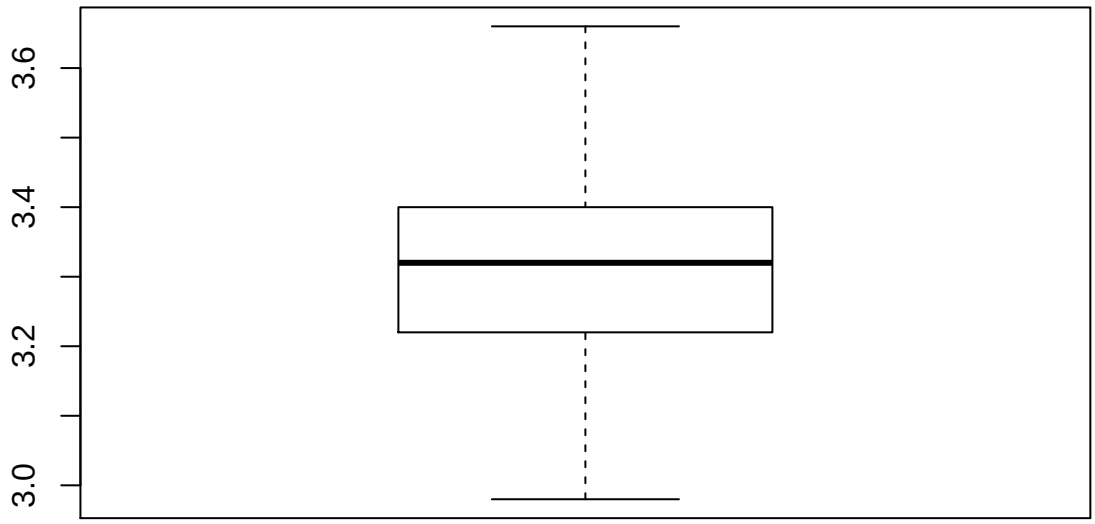


pH

```
#Los valores fuera de rango son:
boxplot.stats(new_wines$pH)$out
```

```
##  [1] 3.90 2.93 2.93 2.93 3.85 3.69 3.69 3.67 3.67 3.68 2.88 2.89 2.89 2.92
## [15] 2.94 2.94 2.94 3.69 3.69 3.71 3.71 3.78 2.94 3.78 3.71 2.88 3.72 3.72
## [29] 3.67
```

```
#Eliminamos del recordset los valores outliers
new_wines<-data.frame(new_wines[!new_wines$pH %in% boxplot.stats(new_wines$pH)$out,])
boxplot(new_wines$pH)
```



Como se ha podido observer después de eliminar los valores fuera de rango, con los nuevos datos, quedan nuevos valores fuera de rango, que como ya he explicado anteriormente, no los elimino porque serían valores que quedaban dentro del rango en primer lugar.

4. Análisis de los datos.

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

El objetivo es determinar si un vino es de buena calidad o no. Para que un vino sea de buena calidad, se tiene que cumplir las siguientes condiciones:

- *volatile acidity* ≤ 0.5
- *Residual sugar* ≥ 3
- *Chlorides* ≤ 0.1
- *Destiny* ≤ 1
- *pH* ≤ 3.5

El siguiente paso es obtener un nuevo dataset con los registros que cumplan estas condiciones. Sobre este nuevo dataset será sobre el que comprobaremos que nuesta hipotesis es cierta.

```
good_wines<-data.frame(new_wines[new_wines$volatile.acidity <= 0.5,])
good_wines<-data.frame(good_wines[good_wines$residual.sugar >= 3,])
good_wines<-data.frame(good_wines[good_wines$chlorides <= 0.1,])
good_wines<-data.frame(good_wines[good_wines$density <= 1,])
good_wines<-data.frame(good_wines[good_wines$pH <= 3.5,])
head(good_wines)
```

```
##     volatile.acidity residual.sugar chlorides density   pH
## 54              0.38            3.0     0.081  0.9970 3.20
## 57              0.42            3.4     0.070  0.9971 3.04
## 253             0.35            3.1     0.090  0.9986 3.17
## 268             0.35            3.6     0.078  0.9973 3.35
## 281             0.26            3.6     0.071  0.9986 3.12
## 288             0.40            3.0     0.092  0.9967 3.37
```

La variable *good_wines* contiene 26 registros que cumplen con las condiciones que he determinado para considerar un buen vino.

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

Para revisar si las variables estan normalizadas voy a aplicar el test de Shapiro Wilk a cada variable.

```
#Para Volatile Acidity
shapiro.test(good_wines$volatile.acidity)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  good_wines$volatile.acidity
## W = 0.94472, p-value = 0.174
```

```
#Para Residual Sugar
shapiro.test(good_wines$residual.sugar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  good_wines$residual.sugar
## W = 0.86514, p-value = 0.002838
```

```
#Para Chlorides
shapiro.test(good_wines$chlorides)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  good_wines$chlorides
## W = 0.9408, p-value = 0.1404
```

```
#Density
shapiro.test(good_wines$density)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  good_wines$density
## W = 0.90636, p-value = 0.02178
```

```
#pH
shapiro.test(good_wines$pH)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  good_wines$pH
## W = 0.94035, p-value = 0.1369
```

Se observa que la variable *Volatile Acidity*, *Chlorides* y *pH* estan normalizadas porque tienen un p-valor superior al 0.05. Sin embargo, las variables *Residual Sugar* y *Density* no estan normalizadas porque su p-valor es inferior al 0.05.

Seguidamente comprobamos la homogeneidad de la varianza, para ello utilizaré el test de Fligner-Killeen. Usaré este test porque es una alternativa cuando no se cumple la condición de normalidad en las muestras.

```
fligner.test(good_wines)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  good_wines
## Fligner-Killeen:med chi-squared = 87.367, df = 4, p-value <
## 2.2e-16
```

Como se observa, el p-valor es inferior a 0.05 y por tanto no podemos aceptar la hipótesis de que la varianza de las variables sean 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.

La prueba estadística que voy a realizar es la regresión lineal. Para ello primero necestio introducir la característica *quality*. Como explicaba anteriormente, esta característica es la puntuación que los etnólogos han puesto a cada vino utilizando paramentros no químicos, es decir, a través del olfato, el gusto, etc.
Necesito introducir esta variable porque para realizar el modelo de regresión lineal voy a comparar la variable *quality* con el resto de características.

El resultado del coeficiente de determinación de la regresion lineal es $R^2 = 0.2054366$. El valor obtenido esta muy cerca de cero por tanto nos indica que no es fiable la predicción de la calidad del vino con este modelo de regresión.

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

Antes de mostrar los resultados voy a exportar el dataset resultado.

```
write.csv(good_wines,"~/Master - Data Science/Tipología y ciclo de vida de los datos/Practica 2/winequal
```

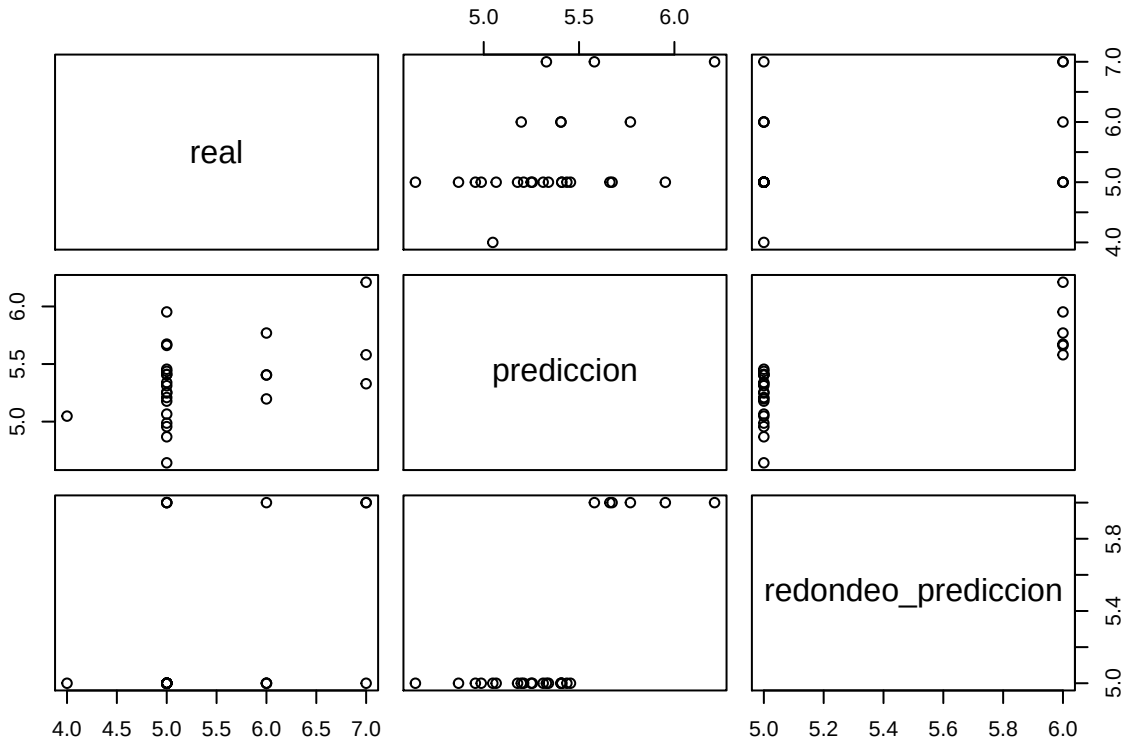
A continuación voy a mostrar en una tabla los valores reales (los que venían en el dataset), con los valores que se obtienen con el modelo de regresión.

```
prob_q<-predict(model_wines,good_wines, type = "response")
tbl<-data.frame(
  real = good_wines$quality
  , prediccion=prob_q
  , redondeo_prediccion = round(prob_q,digits = 0)
)
kable(tbl)
```

real	prediccion	redondeo_prediccion
5	5.454879	5
5	5.952470	6
5	5.673196	6
6	5.196915	5
5	5.661648	6

real	prediccion	redondeo_prediccion
5	5.177558	5
5	4.986940	5
7	6.210593	6
7	5.580317	6
5	5.409098	5
5	5.209285	5
5	5.409098	5
5	5.339444	5
5	5.248598	5
5	5.436117	5
5	5.255017	5
7	5.328712	5
5	4.867901	5
4	5.047638	5
6	5.405032	5
6	5.405032	5
5	5.065328	5
5	5.311782	5
5	4.641932	5
6	5.769206	6
5	4.956264	5

```
pairs(tbl)
```



Como se puede observar la predicción no se corresponde. Además el valor de calidad que venía en el dataset es entero y el obtenido en el modelo de regresión es decimal, al redondear lo que se hace más evidente es que se diferencian.

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

La conclusión a la que he llegado es que a partir del dataset original donde venía recogida características químicas de vinos de Portugal, yo había marcado que dado cinco características (*volatile Acidity*, *Residual Sugar*, *Chlorides*, *Density* y *pH*) y habíendo marcado a partir de que valores consideraba que esas caracteríasticas haría un buen vino, quería comprobar que era cierto. Despues de realizar el análisis estadístico he comprobado que no he decidido bien en escoger las características y los valores. En primer lugar no todas las variables estaban normalizadas y tampoco había una homogeneidad en la varianza, lo que ya me hacía darme cuenta de que no iba bien. Posteriormente, cuando he obtenido en el modelo de regresión que $R^2 = 0.2054366$, he comprobado que dado que el valor se acerca más a 0 que a 1 no podía aceptar la hipotesis de que estas variables me indicarían si un vino era de buena calidad o no. Una vez que tenía el modelo de regresión he comprobado los resultados reales con los predictivos y he visto que hay diferencias entre los resultados, dejando patente que no puedo resolver el problema orignal que es determinar si un vino es de buena calidad o no dependiendo de las características que había elegido.