Practica - 2

Student:

Carlos Rea Nogales & Yago Novoa

Course:

Tipologia y Ciclo de vida de los datos

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

El dataset se basa en la variante de vino tinto de marca portuguesa “vinho verde”, el dataset está disponible en el repositorio de UCI ML (https://archive.ics.uci.edu/ml/datasets/wine+quality). El objetivo del proyecto es analizar la relación entre la calidad del vino y sus características analizadas. La calidad está dada como una nota del 0 (peor) al 10 (mejor)

El dataset consta de 1599 observaciones 11 variables de independientes y 1 variable dependiente. Todas las variables independientes son variables numéricas.

- Acidez fija
- Acidez volátil
- Acido citrico
- Azúcar residual
- Cloritos
- Dióxido de sulfuro libre
- Dióxido de sulfuro total
- Densidad
- pH
- Sulfatos
- Alcohol
- Calidad: Variable objetivo

En este dataset se plantea determinar que variables influyen más sobre la calidad del vino, también se procederá a crear modelos de regresión para predecir si el vino será bueno o no dependiendo de sus propiedades base y se generaran contrastes de hipótesis para determinar propiedades interesantes en las muestras que se puedan inferir a la población.

Estos análisis serán importantes para las empresas dedicadas al sector de los vinos al poder dar una ventaja competitiva a las empresas para apoyar las evaluaciones de los enólogos en la cata de vinos, de manera que se pueda analizar cuáles son las propiedades más importantes en las puntuaciones y mejorar la producción de vino.

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

Previamente a comenzar el análisis se carga el dataset y vemos su estructura.

```
#wine = read.csv2("winequality-white.csv")
wine = read.csv("winequality-red.csv")
summary(wine)
```

```
str(wine)
```

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

Vemos como el dataset está compuesto de variables numéricas y una variable de tipo int que resulta ser la variable a predecir, lo que nos permite usar todas las variables para el modelo.

3. Limpieza de los datos.

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

Realizamos un resumen estadístico para analizar los valores vacíos u otras incoherencias o irregularidades en los datos.

En el caso de tener valores vacíos se podrían eliminar si son influyentes en la tendencia de los datos, por ejemplo, visualizando la tendencia de los datos mediante un gráfico de dispersión con una recta de regresión o podemos imputar por el valor medio de la variable u otro método de imputación como kNN en caso de que no quisiéramos perder información. El método de imputación empleado es determinante por lo que se tiene que realizar un análisis de su impacto en el resultado del modelo.

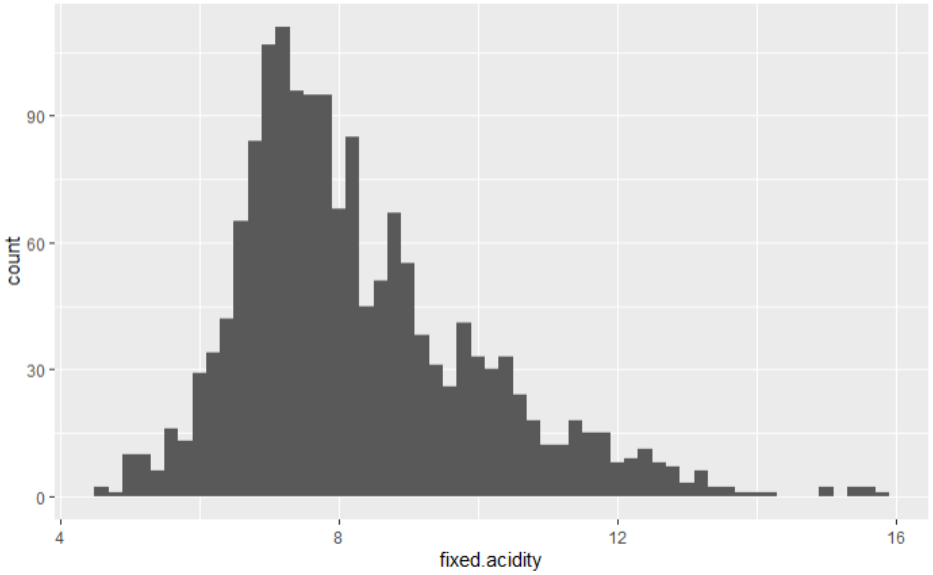
```
wine

 12  Variables      1599  Observations
--------------------------------------------------------------------------------------------------------------------
--------------
fixed.acidity
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0       96    0.999     8.32    1.893      6.1      6.5      7.1      7.9      9.2     10.7     11.8

lowest :  4.6  4.7  4.9  5.0  5.1, highest: 14.3 15.0 15.5 15.6 15.9
--------------------------------------------------------------------------------------------------------------------
--------------
volatile.acidity
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0      143        1   0.5278    0.199    0.270    0.310    0.390    0.520    0.640    0.745    0.840

lowest : 0.120 0.160 0.180 0.190 0.200, highest: 1.180 1.185 1.240 1.330 1.580
--------------------------------------------------------------------------------------------------------------------
--------------
citric.acid
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0       80    0.999    0.271   0.2227    0.000    0.010    0.090    0.260    0.420    0.522    0.600

lowest : 0.00 0.01 0.02 0.03 0.04, highest: 0.75 0.76 0.78 0.79 1.00
--------------------------------------------------------------------------------------------------------------------
--------------
residual.sugar
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0       91    0.996    2.539    1.078     1.59     1.70     1.90     2.20     2.60     3.60     5.10

lowest :  0.9  1.2  1.3  1.4  1.5, highest: 13.4 13.8 13.9 15.4 15.5
--------------------------------------------------------------------------------------------------------------------
--------------
chlorides
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0      153        1  0.08747  0.03217   0.0540   0.0600   0.0700   0.0790   0.0900   0.1090   0.1261

lowest : 0.012 0.034 0.038 0.039 0.041, highest: 0.422 0.464 0.467 0.610 0.611
--------------------------------------------------------------------------------------------------------------------
--------------
free.sulfur.dioxide
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0       60    0.998    15.87    11.24        4        5        7       14       21       31       35

lowest :  1  2  3  4  5, highest: 55 57 66 68 72
--------------------------------------------------------------------------------------------------------------------
--------------
total.sulfur.dioxide
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0      144        1    46.47    34.63     11.0     14.0     22.0     38.0     62.0     93.2    112.1

lowest :   6   7   8   9  10, highest: 155 160 165 278 289
--------------------------------------------------------------------------------------------------------------------
--------------
density
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0      436        1   0.9967 0.002081   0.9936   0.9946   0.9956   0.9968   0.9978   0.9991   1.0000

lowest : 0.99007 0.99020 0.99064 0.99080 0.99084, highest: 1.00260 1.00289 1.00315 1.00320 1.00369
--------------------------------------------------------------------------------------------------------------------
--------------
pH
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0       89        1    3.311   0.1716     3.06     3.12     3.21     3.31     3.40     3.51     3.57

lowest : 2.74 2.86 2.87 2.88 2.89, highest: 3.75 3.78 3.85 3.90 4.01
--------------------------------------------------------------------------------------------------------------------
--------------
sulphates
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0       96    0.999   0.6581   0.1679     0.47     0.50     0.55     0.62     0.73     0.85     0.93

lowest : 0.33 0.37 0.39 0.40 0.42, highest: 1.61 1.62 1.95 1.98 2.00
--------------------------------------------------------------------------------------------------------------------
--------------
alcohol
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1599        0       65    0.998    10.42    1.178      9.2      9.3      9.5     10.2     11.1     12.0     12.5

lowest :  8.40000  8.50000  8.70000  8.80000  9.00000, highest: 13.50000 13.56667 13.60000 14.00000 14.90000
--------------------------------------------------------------------------------------------------------------------
--------------
quality
       n  missing distinct     Info     Mean      Gmd
    1599        0        6    0.857    5.636   0.8431

lowest : 3 4 5 6 7, highest: 4 5 6 7 8

Value          3     4     5     6     7     8
Frequency     10    53   681   638   199    18
Proportion 0.006 0.033 0.426 0.399 0.124 0.011
--------------------------------------------------------------------------------------------------------------------
```

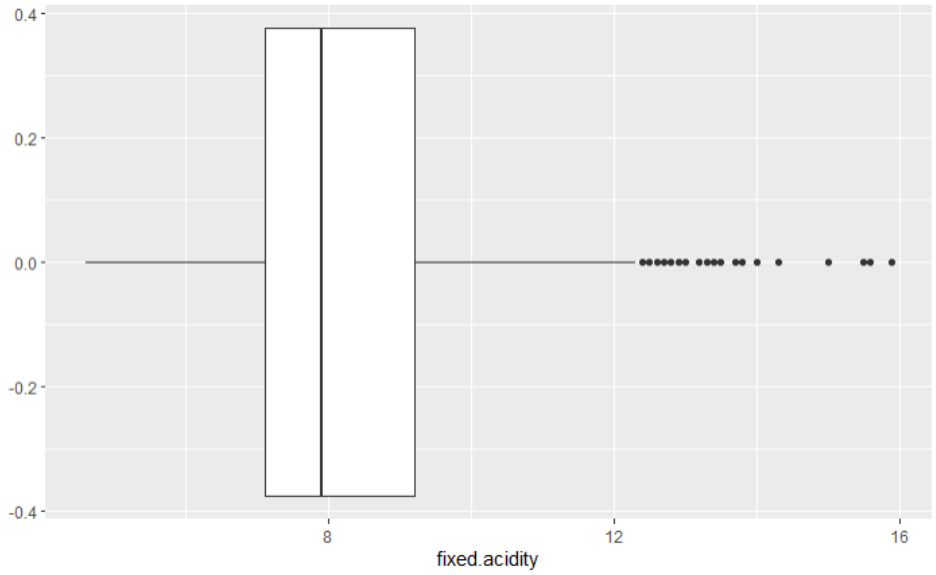
Como podemos observar ninguna variable tiene valores ausentes en el dataset, sin embargo la variable “wine$citric.acid” tiene valores iguales a 0, los cuales caen dentro del rango de valores posibles(0,…,1), estos valores son correctos y no son sustituibles ya que estaríamos sesgando los datos. Con respecto al resto de variables parecen tener rangos de valores factibles por lo que a priori no se observan outliers.

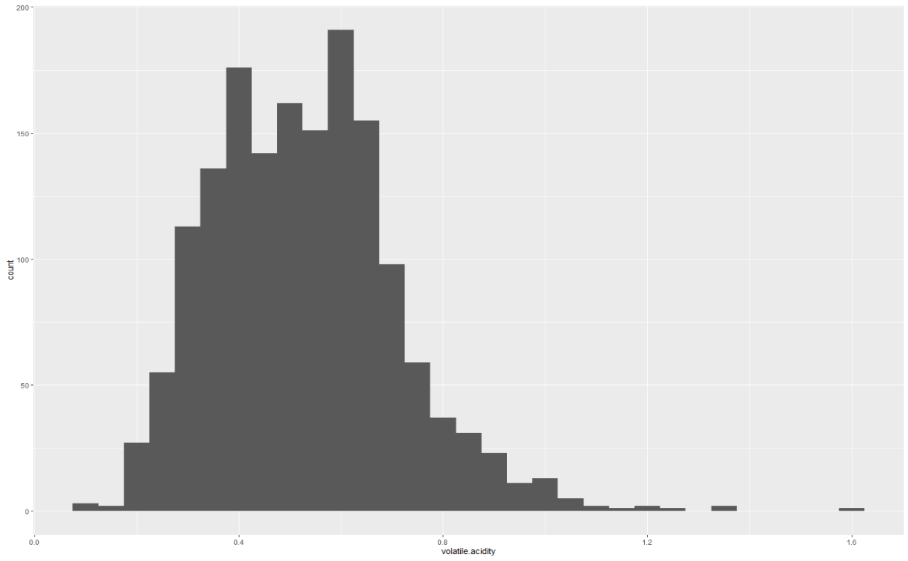
Además, podemos intuir a partir de la media y la mediana que density y pH pueden seguir una distribución normal ya que sus valores coinciden.

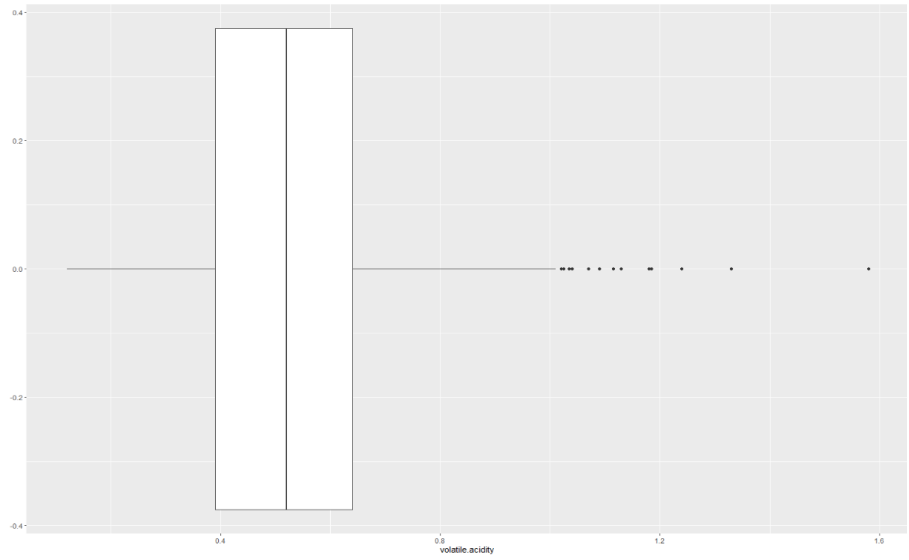
3.2 Identificación y tratamiento de valores extremos.

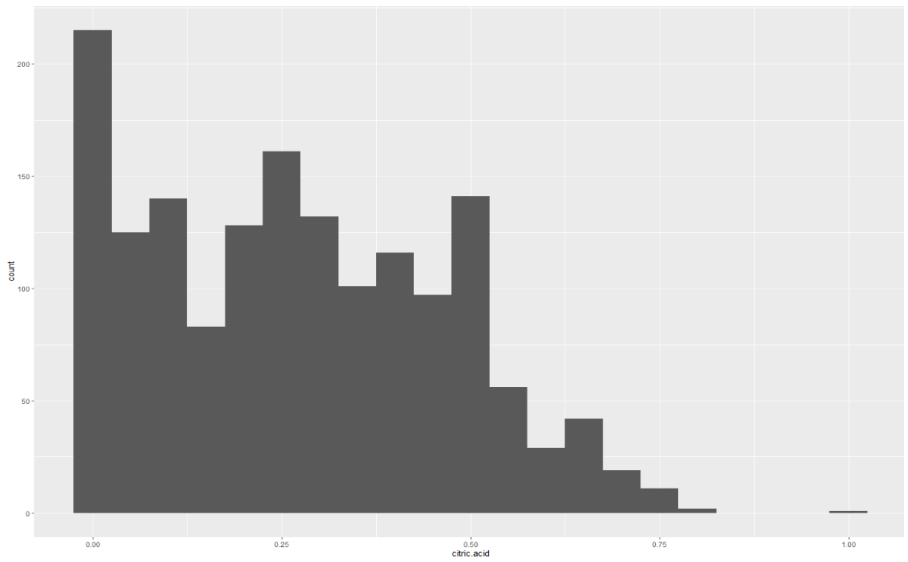
En primer lugar, se lleva a cabo un análisis grafico mediante histogramas y diagramas de cajas para poder visualizar la presencia de valores outliers. Estos valores son aquellas observaciones distantes del resto de los datos, en un diagrama de caja se considera atípico cuando se encuentra a una distancia de 1.5 veces el rango intercuartílico.

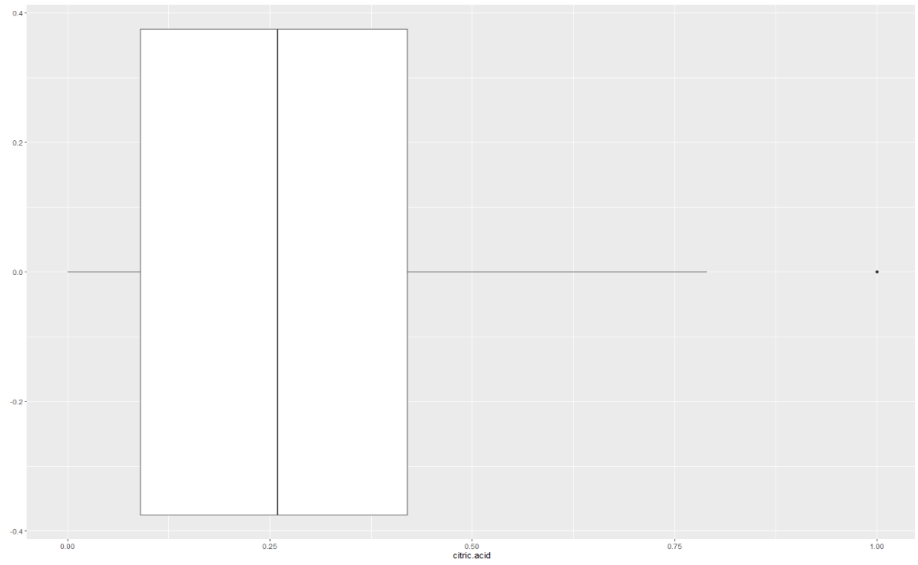


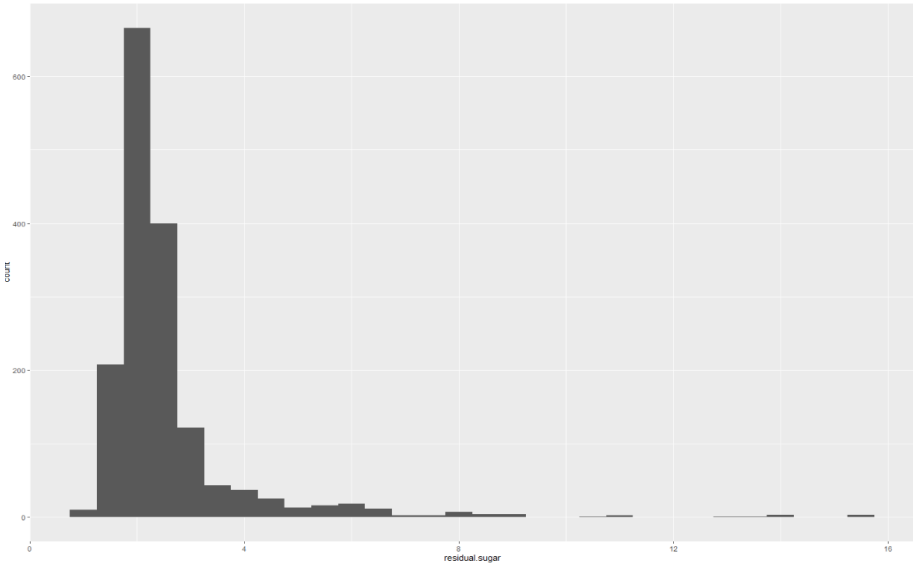


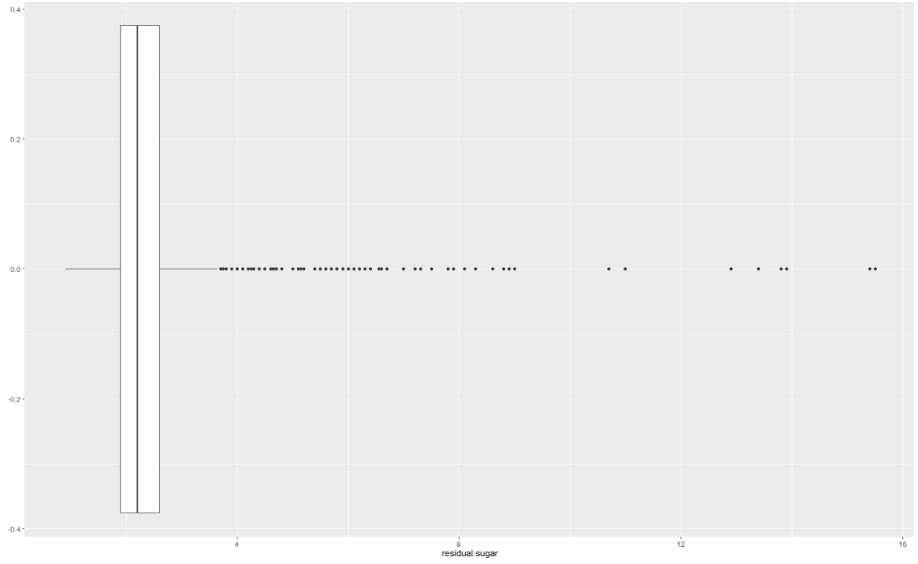

count
volatile.acidity

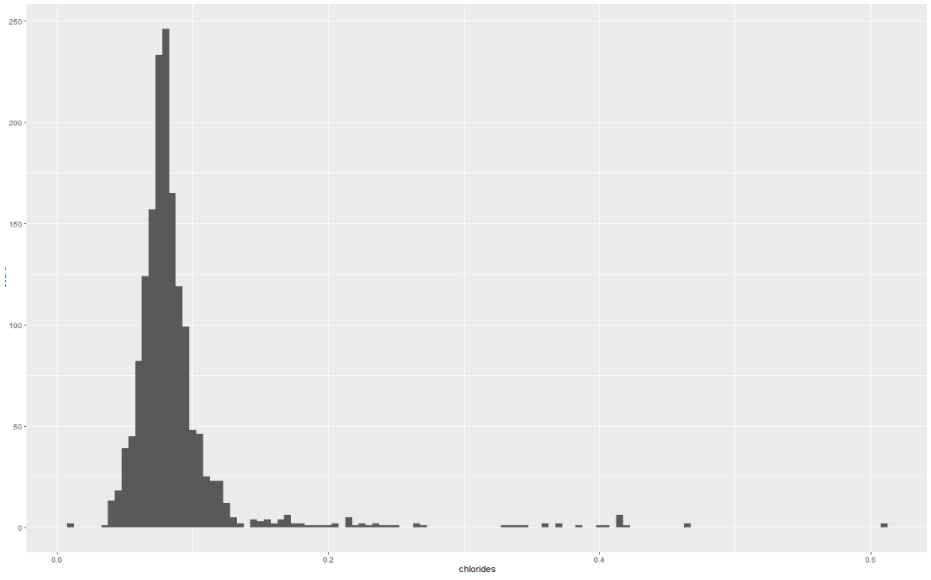

volatile.acidity

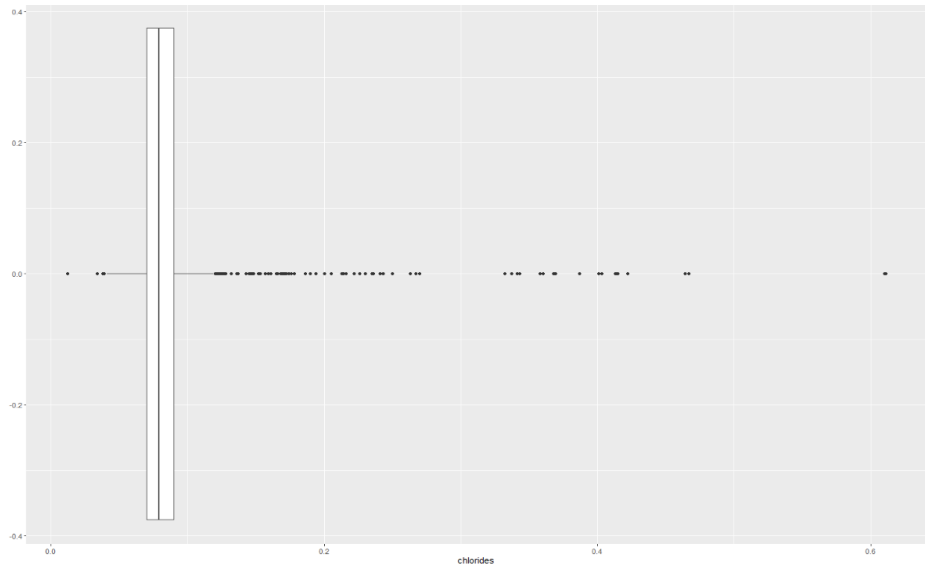

count
citric.acid

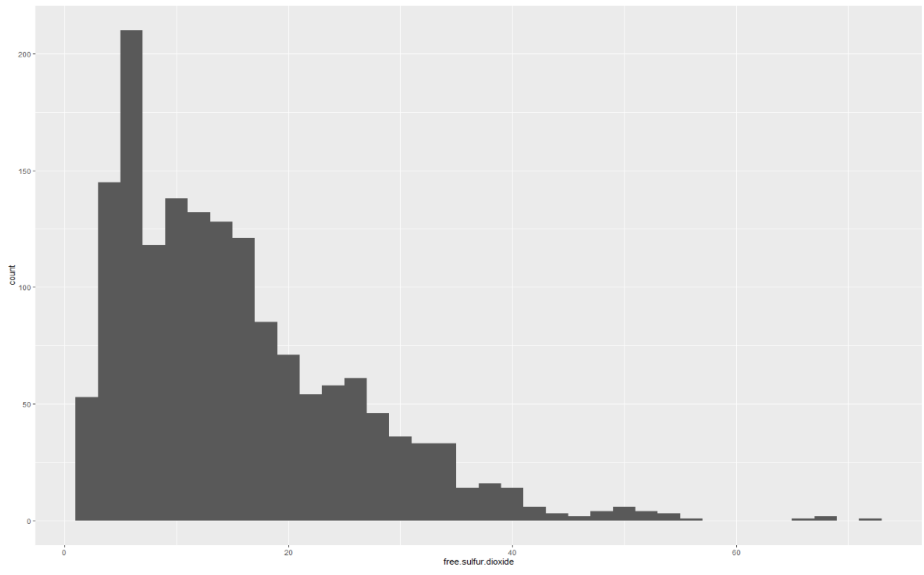

0.4
0.2
0.0
-0.2
-0.4
0.00
0.25
0.50
0.75
1.00
citric.acid

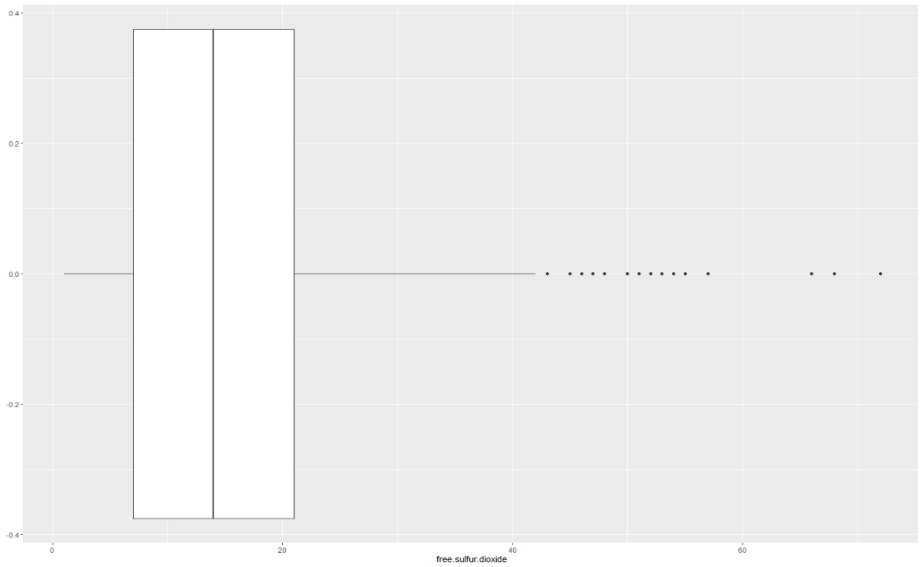

600
400
200
0
count
0
4
8
12
16
residual.sugar

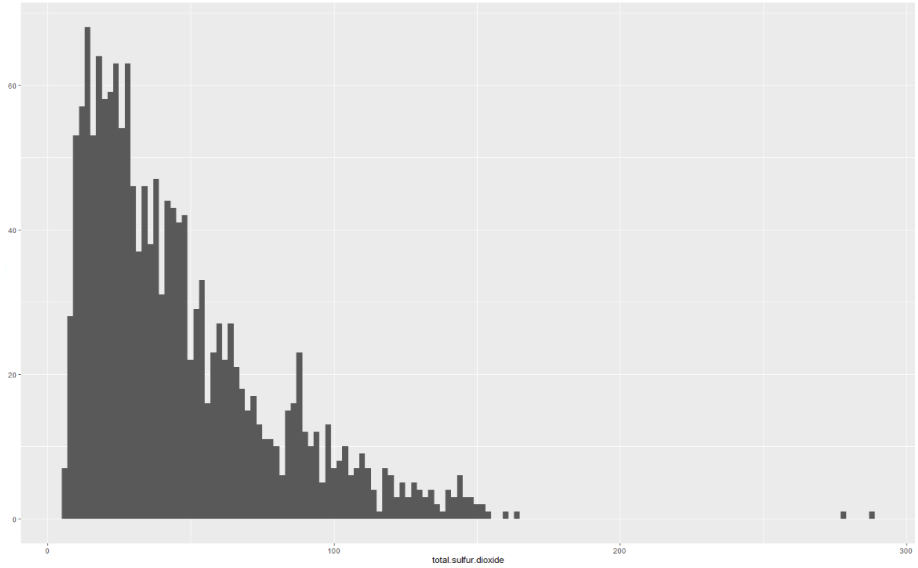

0.4
0.2
0.0
-0.2
-0.4
4
8
12
16
residual.sugar

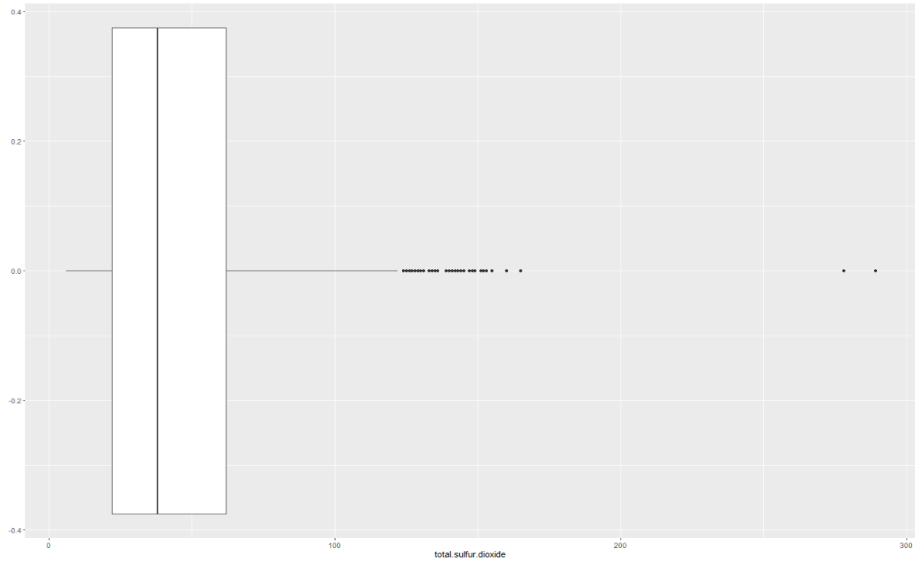

250
200
150
100
50
0
0.0
0.2
0.4
0.6
chlorides

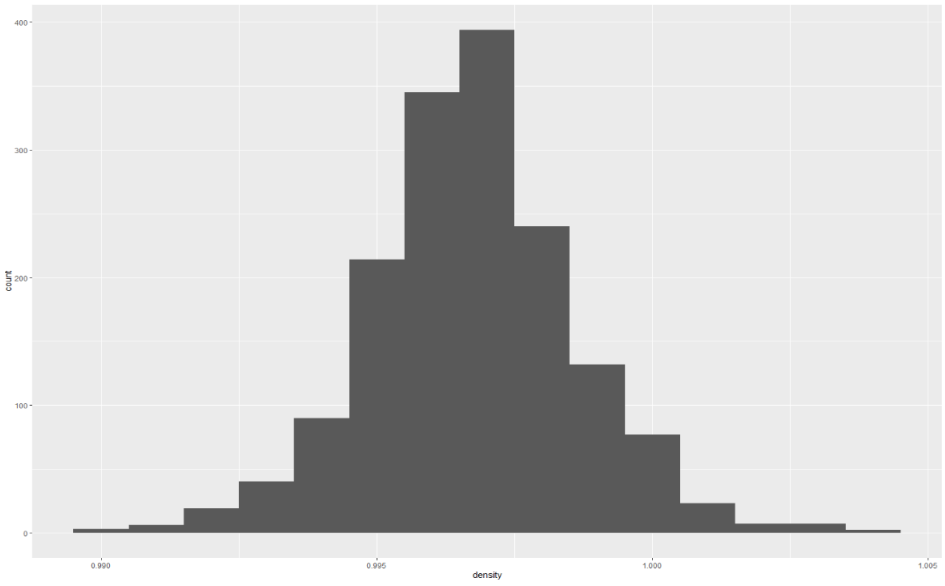

0.4
0.2
0.0
-0.2
-0.4
0.0
0.2
0.4
0.6
chlorides

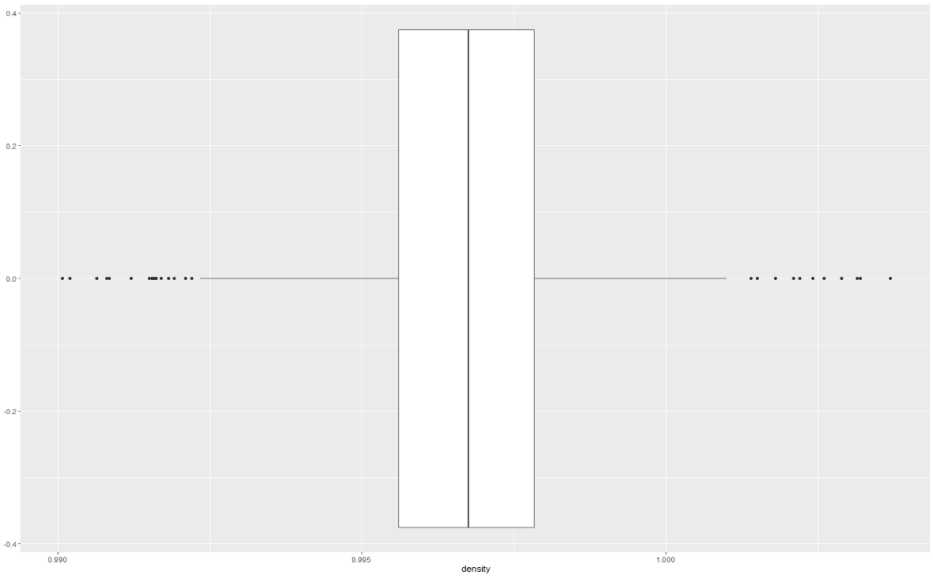

200
150
100
50
0
count
0
20
40
60
free.sulfur.dioxide

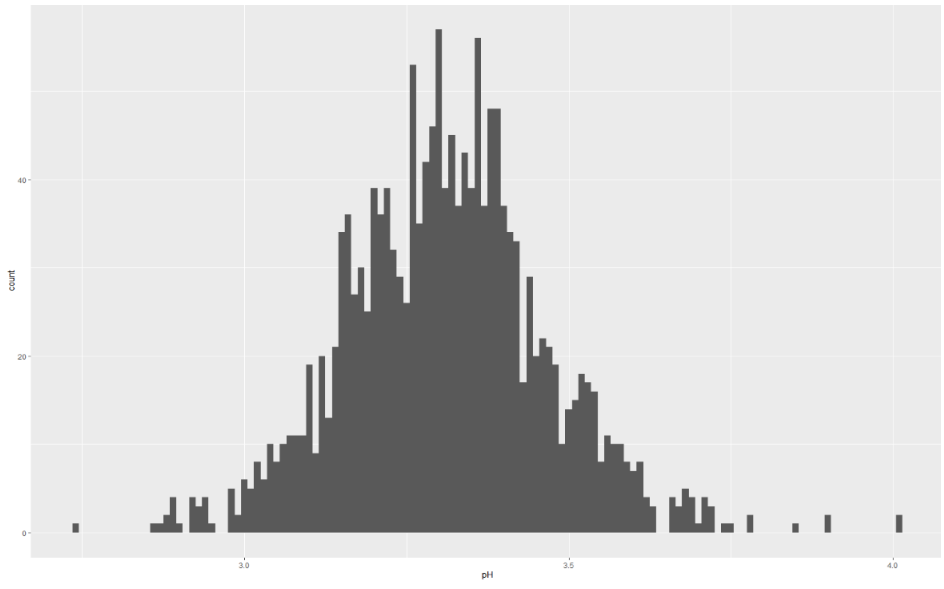

0.4
0.2
0.0
-0.2
-0.4
0
20
40
60
free.sulfur.dioxide

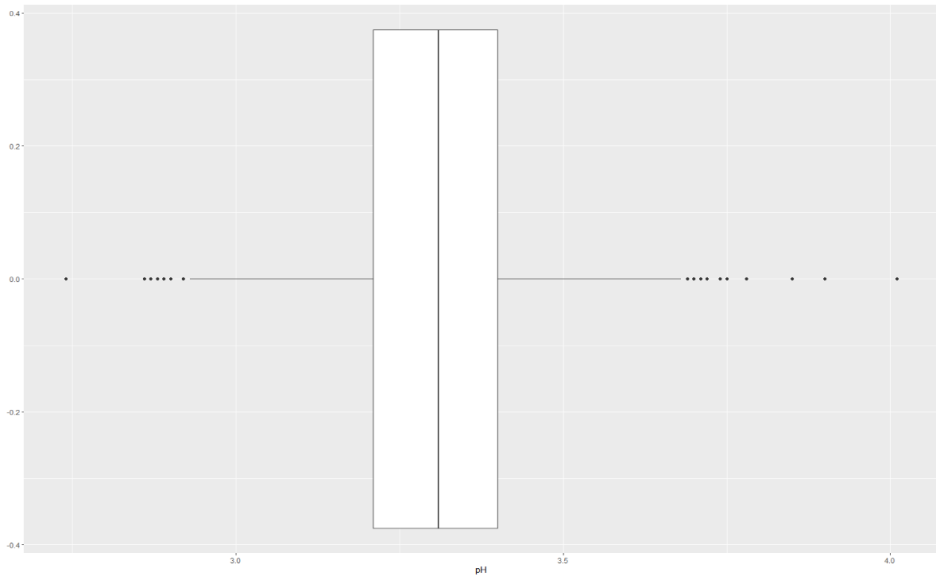

60
40
20
0
0
100
200
300
total.sulfur.dioxide

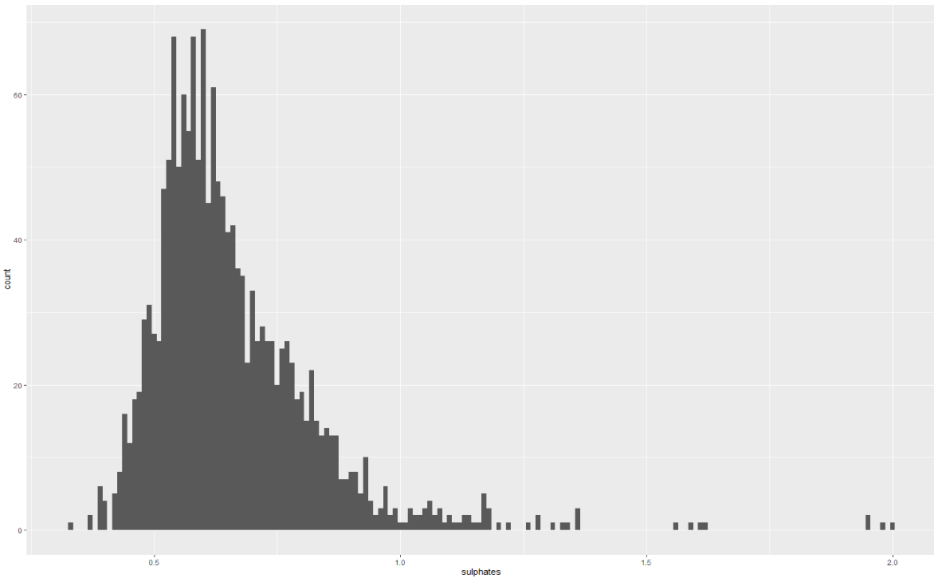

0.4
0.2
0.0
-0.2
-0.4
0
100
200
300
total.sulfur.dioxide

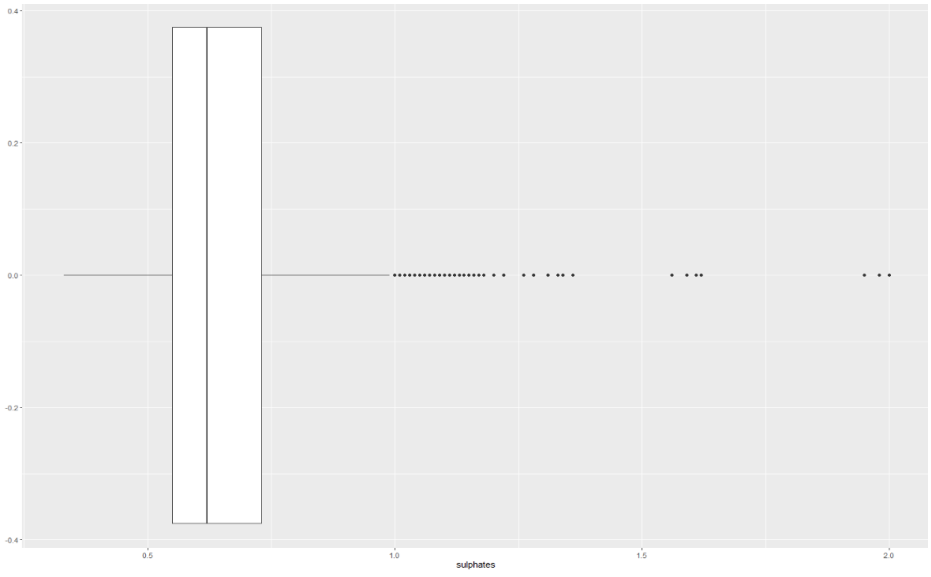

400
300
200
100
0
count
0.990
0.995
1.000
1.005
density

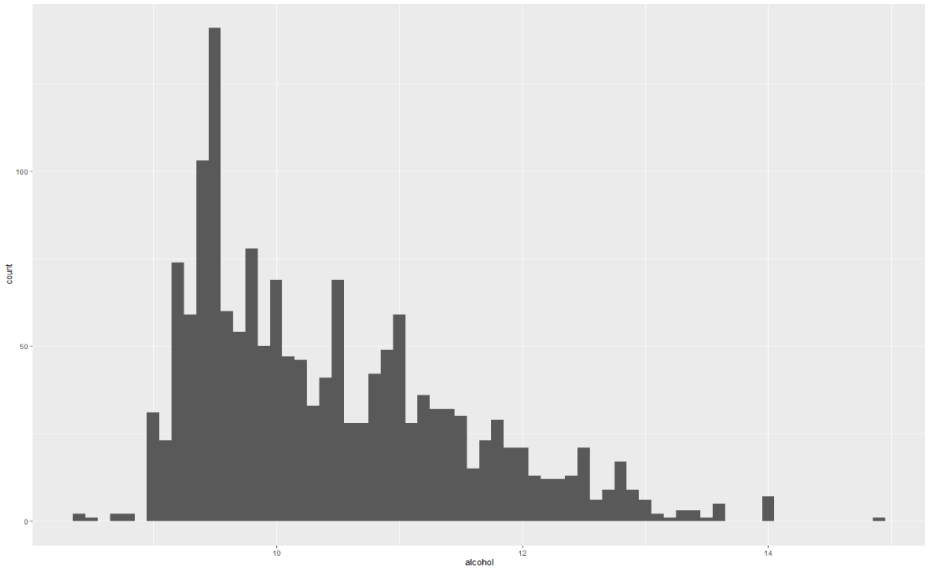

0.4
0.2
0.0
-0.2
-0.4
0.990
0.995
1.000
density

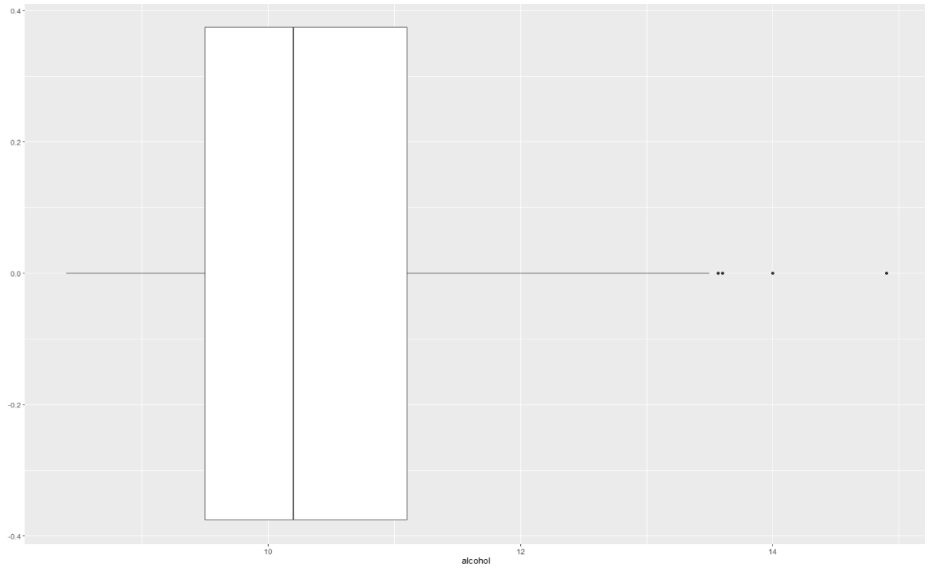

40
20
0
count
3.0
3.5
4.0
pH

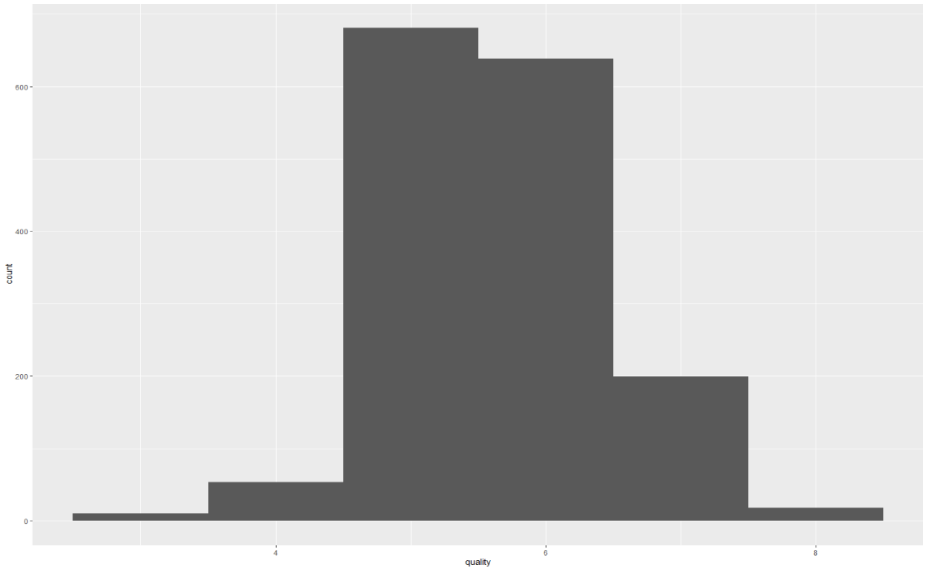

0.4
0.2
0.0
-0.2
-0.4
3.0
3.5
4.0
pH

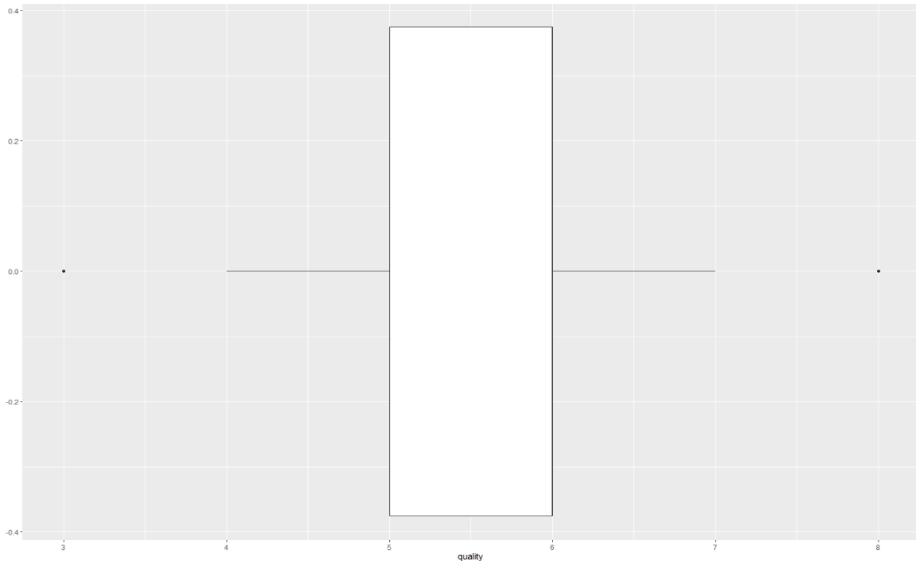

60
40
20
0
count
0.5
1.0
1.5
2.0
sulphates


0.4
0.2
0.0
-0.2
-0.4
0.5
1.0
1.5
2.0
sulphates


count
100
50
0
10
12
14
alcohol


0.4
0.2
0.0
-0.2
-0.4
10
12
14
alcohol


count
600
400
200
0
4
6
8
quality



En todas las variables podemos observar algunos valores atípicos que salen de la distribución normal, pero son valores factibles, por lo tanto, no los eliminaremos ni imputaremos, sino que los dejaremos como vienen en el dataset.

4.Analisis de los datos

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

Con el objetivo de facilitar el análisis una columna dicotómica es añadida al dataset clasificando los vinos en buenos a los que reciben una nota >= 6 y el resto clasificándolos como malos

```
wine <- wine %>% mutate(qualityDct = ifelse(quality >= 6, "Good", "B
ad"))
wine$qualityDct <- factor(wine$qualityDct)
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para la comprobación de que los valores de la muestra provienen de una población con una distribución normal, se realizará la prueba de Anderson-Darling. Así, se comprueba que para cada prueba se obtiene un p-valor superior al nivel de significación prefijado α = 0, 05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```
# seleccionamos las variables cuantitativas #
colnames(wine[sapply(wine, class)=="numeric"]) -> numvars

for (i in numvars) {
  p_val = ad.test(simplify2array(wine[,i]))$p.value
  if(p_val < 0.05){
    cat(i,"-", p_val ,"\n")
  }
}
```

```
## fixed.acidity - 3.7e-24
## volatile.acidity - 5.318894e-14
## citric.acid - 3.7e-24
## residual.sugar - 3.7e-24
## chlorides - 3.7e-24
## free.sulfur.dioxide - 3.7e-24
## total.sulfur.dioxide - 3.7e-24
## density - 1.227494e-09
## pH - 9.245208e-05
## sulphates - 3.7e-24
## alcohol - 3.7e-24
```

Vemos que ningún valor cumple con el nivel de significancia por lo que ninguna variable sigue una distribución normal. Nos centraremos en las variables que más se acercan a una distribución normal que son pH y density las cuales tienen valores más cercanos a cero que el resto y como ya intuíamos a partir de la media y la mediana además de la representación gráfica que parece dibujar una distribución normal, analizaremos más en profundidad la variable pH.

Aplicaremos diferentes métodos de normalidad para ver si alguno de ellos verifica normalidad.

```
# comprobar si los datos siguen una distribucion normal mediante el test de kolmogorov smirnov lilliefors.
lillie.test(wine$pH)

# comprobamos si los datos siguen una distribucion normal mediante el test de anderson-darling
ad.test(wine$pH)

# comprobamos si los datos siguen una distribucion normal mediante el test de shaphiro wilks
shapiro.test(wine$pH)

# comprobamos si los datos siguen una distribucion normal mediante el test de d'agostino pearson
dagoTest(wine$pH)
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  wine$pH
D = 0.040368, p-value = 2.244e-06


        Anderson-Darling normality test

data:  wine$pH
A = 1.8641, p-value = 9.245e-05


        Shapiro-Wilk normality test

data:  wine$pH
W = 0.99349, p-value = 1.712e-06


Title:
 D'Agostino Normality Test

Test Results:
  STATISTIC:
    Chi2 | Omnibus: 33.6847
    Z3   | Skewness: 3.1449
    Z4   | Kurtosis: 4.8779
  P VALUE:
    Omnibus  Test: 4.847e-08
    Skewness Test: 0.001661
    Kurtosis Test: 1.072e-06
```

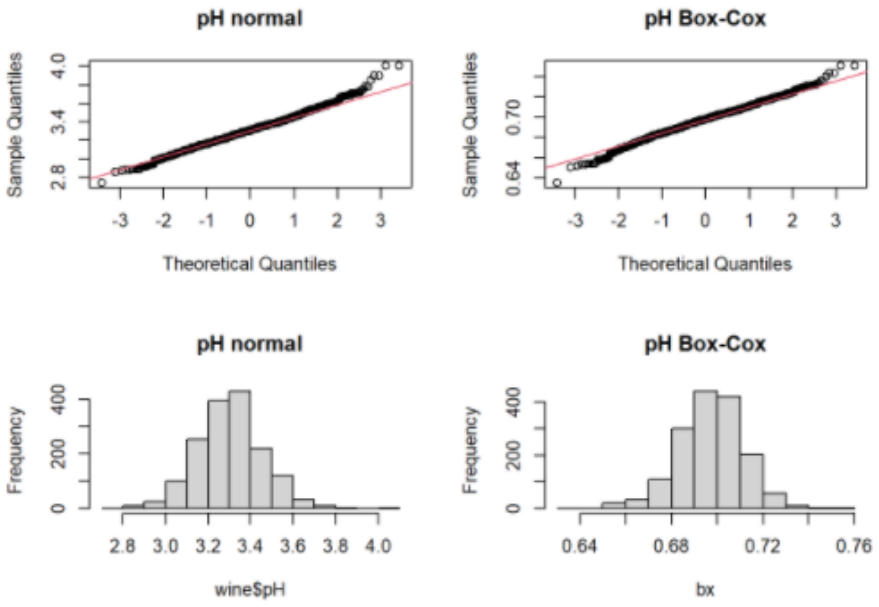
Todos los test devuelven un p-valor muy pequeño, se rechaza la hipótesis nula, los valores de la muestra no siguen una distribución normal.

Realizamos una transformación de box-cox para tratar de mejorar la normalidad y homocedastidad.

```
ad.test(bcPower(wine$pH,lambda = 0.6))

shapiro.test(bcPower(wine$pH,lambda = 0.6))
```

```
        Anderson-Darling normality test

data:  bcPower(wine$pH, lambda = 0.6)
A = 1.6872, p-value = 0.000251


        Shapiro-Wilk normality test

data:  bcPower(wine$pH, lambda = 0.6)
W = 0.99461, p-value = 1.537e-05
```

```
bx = bcPower(wine$pH, lambda = BoxCoxLambda(wine$pH))

par(mfrow=c(2,2))
qqnorm(wine$pH, main="pH normal")
qqline(wine$pH,col=2)
qqnorm(bx, main="pH Box-Cox")
qqline(bx,col=2)
hist(wine$pH,main="pH normal")
hist(bx, main="pH Box-Cox")
```



La transformación boxcox no consigue mejorar la normalidad, por lo que optamos por utilizar una alternativa no paramétrica como las pruebas Wilcoxon o Mann-Whitney. No obstante, como el conjunto de datos se compone de un número de registros grande, por el teorema central del límite, se podría considerar que los datos siguen una distribución normal.

A continuación, aplicamos el test de Levene para analizar la homogeneidad de la varianza, teniendo en cuenta que los datos no cumplen la condición de normalidad también aplicamos el test de fligner-killeen, que se trata de una alternativa no paramétrica.

```
leveneTest(pH~qualityDct, data = wine)

fligner.test(pH~qualityDct, data = wine)
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    1  0.4107 0.5217
      1597

        Fligner-Killeen test of homogeneity of variances

data:  pH by qualityDct
Fligner-Killeen:med chi-squared = 0.62792, df = 1, p-value = 0.4281
```

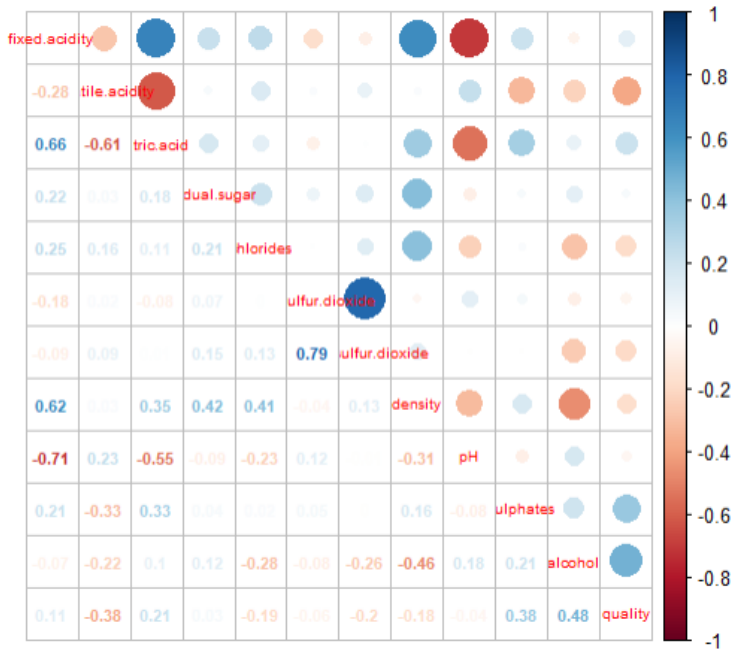
Dado que ambas pruebas resultan en un p-valor superior al nivel de significancia 0.05, se acepta la hipótesis nula de homocedasticidad y se concluye que la variable 'pH' presenta varianzas estadísticamente iguales para los diferentes grupos de 'qualityDct'.

5. Pruebas estadísticas

5.1 ¿Qué variables cuantitativas influyen más en la calidad del vino?

En primer lugar, realizamos un análisis de correlación entre las diferentes variables para determinar cuáles ejercen mayor influencia sobre la calidad del vino.

```
# seleccionamos el df con las variables cuantitativas
wine_num <- wine %>% select(numvars)

# aplicamos la correlacion con el metodo de spearman ya que los datos no siguen una distribucion normal.
cor(wine[,1:12],method = "spearman",use = "complete.obs") -> wine_num.cor
#redondeamos decimales.
round(wine_num.cor, digits = 2)

# visualizamos la matriz de correlacion
#corrplot(wine_num.cor)
corrplot.mixed(wine_num.cor,upper="circle",number.cex=.7,tl.cex=.6)

# Buscamos  las variables con mayor correlación, hemos realizado una selección etre 0.25 y 0.90
zdf <- as.data.frame(as.table(wine_num.cor))
zdf_a <- subset(zdf, abs(Freq) > 0.25 & abs(Freq) <0.9)
# muestro el resultado ordendando la varible Freq
zdf_a <- arrange(zdf_a, desc(Freq))
zdf_a
zdf_a %>% filter(Var1 == 'quality')
```



Var1 <fctr>	Var2 <fctr>	Freq <dbl>
quality	alcohol	0.4785317
quality	sulphates	0.3770602
quality	volatile.acidity	-0.3806465

Las variables que más influyen en la calidad son 'alcohol', 'sulphates' y 'volatile.acidity' la mayor correlación es con el alcohol sin embargo no deja de ser una correlación débil

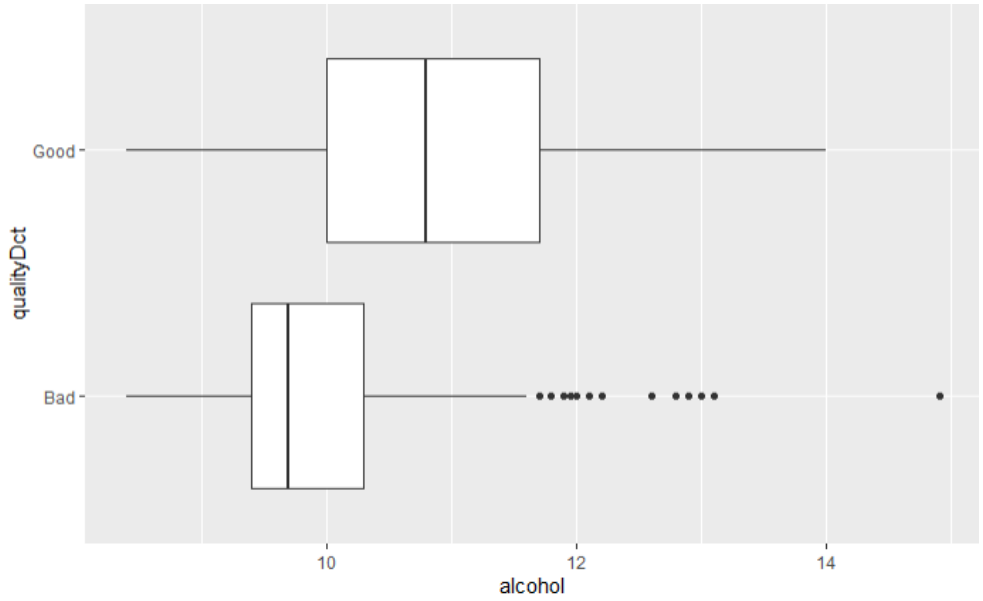
5.3 ¿El nivel de alcohol es igual para los diferentes niveles de la calidad del vino?

Aplicamos pruebas no paramétricas dado que no se cumple la normalidad en los datos. La hipótesis nula asume que las distribuciones de los grupos de datos son las mismas. A continuación, aplicamos el test de Wilcox y Mann-Whitney para realizar la comparación entre los dos grupos ('Good','Bad')

```
wilcox.test(alcohol~qualityDct,data = wine)
```

```
        Wilcoxon rank sum test with continuity correction

data:  alcohol by qualityDct
W = 154807, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Se rechaza la hipótesis nula ya que el p-valor es menor al nivel de significación alfa, se observan diferencias estadísticamente significativas entre el nivel de alcohol y la calidad del vino ('Good','Bad'). Si mostramos gráficamente un diagrama de cajas podemos ver que las distribuciones son estadísticamente diferentes, tanto la media como los rangos de los cuartiles.

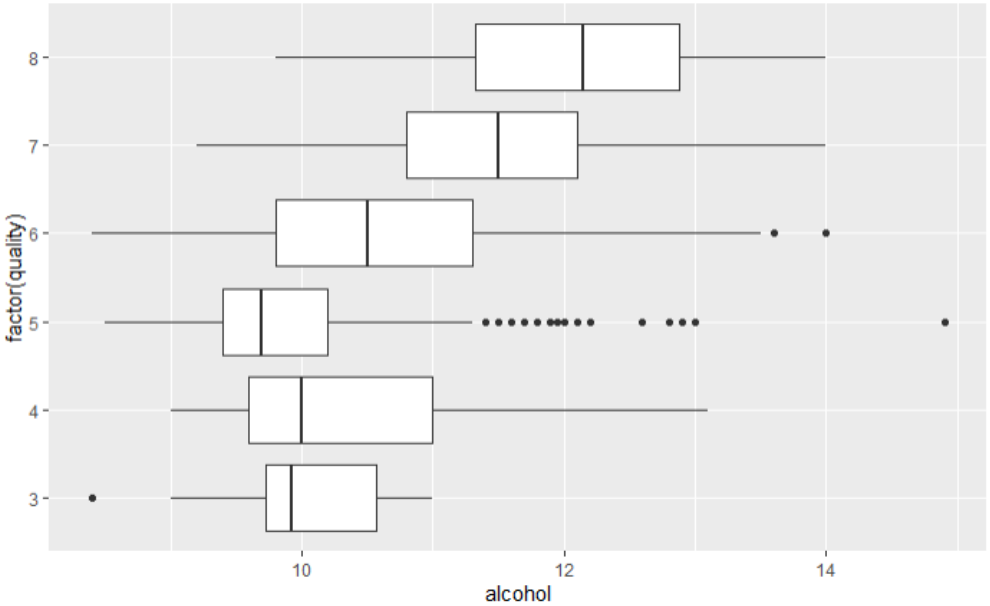


Ahora realizamos el test no paramétrico de Kruskal-Wallis para comparar más de dos grupos (3,4,5,6,7,8)

```
kruskal.test(alcohol~quality, data = wine)
```

```
        Kruskal-Wallis rank sum test

data:  alcohol by quality
Kruskal-Wallis chi-squared = 412.38, df = 5, p-value < 2.2e-16
```

Se rechaza la hipótesis nula y se acepta la hipótesis alternativa, se observan diferencias para los distintos niveles de calidad del vino (3,4,5,6,7,8).



Mirando los diagramas de cajas podemos concluir que los mejores vinos tienen un mayor nivel de alcohol.

5.3 ¿El nivel de acidez es igual para los diferentes niveles de la calidad del vino?

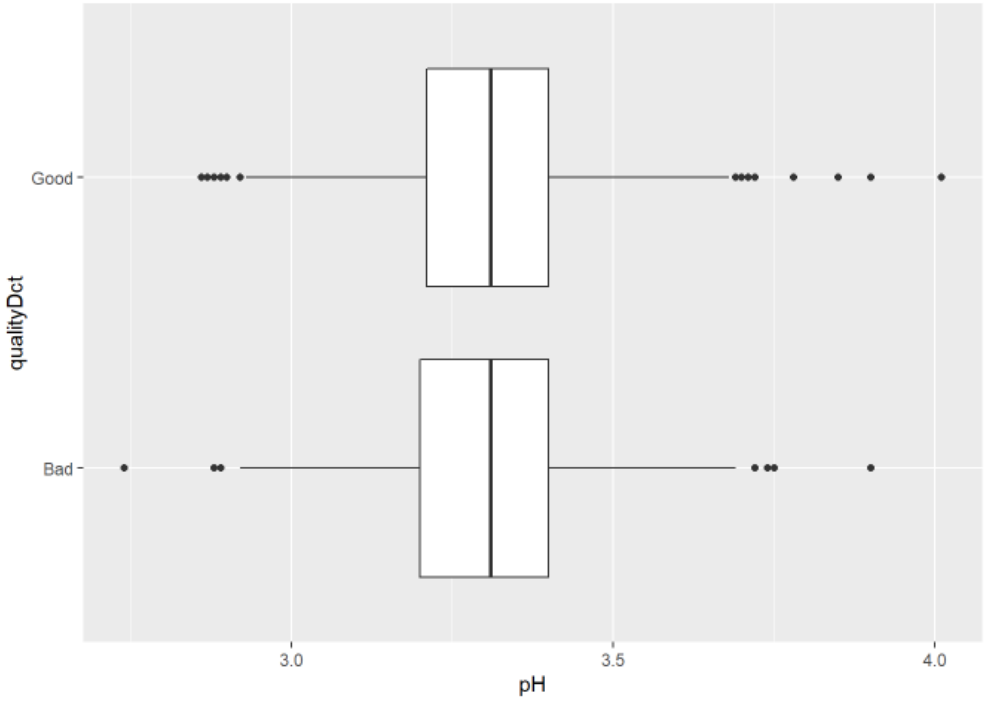
Igual que en el caso anterior realizamos el mismo estudio con la variable pH.

```
# aplicamos pruebas no parametricas dado que no se cumple la normalidad en los datos. Se asume que las distribuci
ones de los grupos de datos son las mismas.

wilcox.test(pH~qualityDct,data = wine)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pH by qualityDct
## W = 319963, p-value = 0.8363
## alternative hypothesis: true location shift is not equal to 0
```

No se rechaza la hipótesis nula ya que el p-valor es mayor al nivel de significación alfa, no se observan diferencias estadísticamente significativas entre el nivel de acidez y la calidad del vino ('Good','Bad'). Si mostramos gráficamente un diagrama de cajas podemos ver que las distribuciones son estadísticamente iguales, tanto la media como los rangos de los cuartiles son similares.

```
# no se observan diferencias estadisticamente significativas entre el nivel de acidez y la calidad del vino(buen
o,malo).
wine %>% ggplot(aes(qualityDct,pH)) + geom_boxplot() + coord_flip()
```

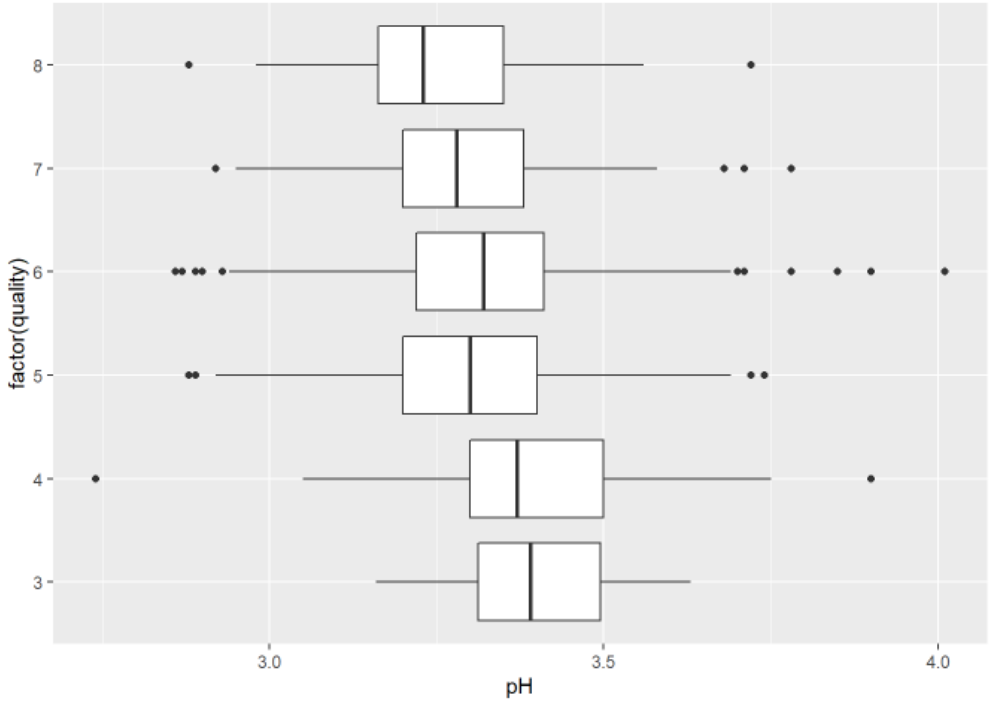


Ahora realizamos el test no paramétrico de Kruskal-Wallis para comparar más de dos grupos (3,4,5,6,7,8) y mostramos el diagrama de cajas.

```
# precio es independiente del año y la ubicacion? mediante test chi cuadrado contraste de independencia . crear u
na tabla de año y ubicacion con el precio.

# estadistico de contraste de hipotesis de mas de 2 grupos, mediante el test de kruskal-wallis. para ver si la di
stribucion de los distintos niveles de calidad del vino con respecto a la acidez son iguales h0 o no h1

kruskal.test(pH~quality, data = wine)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  pH by quality
## Kruskal-Wallis chi-squared = 23.736, df = 5, p-value = 0.000244
```

```
# se rechaza la hipotesis nula y se acepta la hipotesis alternativa, se observan diferencias para los distintos n
iveles de calidad del vino(3,4,5,6,7,8)
wine %>% ggplot(aes(factor(quality),pH)) + geom_boxplot() + coord_flip()
```



Como podemos ver se rechaza la hipótesis nula y se acepta la hipótesis alternativa, se observan diferencias para los distintos niveles de calidad del vino (3,4,5,6,7,8).

Al agrupar los niveles de calidad en dos niveles las distribuciones son muy similares y se pierde detalle, sin embargo, el nivel de acidez es diferente para cada nivel de calidad por separado. Examinando el diagrama de cajas se puede ver una tendencia decreciente del nivel de acidez con respecto a la calidad del vino. Los mejores vinos tienen un menor nivel de acidez, lo cual no se puede apreciar con dos niveles.

5.3 Modelo de regresión logística

Aplicamos un modelo de regresión logística ya que no asume los supuestos de la regresión lineal, particularmente el de normalidad, Linealidad y homoscedasticidad.

Primero dividimos el dataset en conjunto de entrenamiento y test para poder evaluar el mejor modelo. Después creamos cinco modelos diferentes a los cuales iremos añadiendo variables.

```
set.seed(0)
#dividimos los datos en conjunto de entrenamiento y test
split = sample.split(wine$quality,SplitRatio = 0.75)
train_set = subset(wine,split==TRUE)
test_set = subset(wine, split==FALSE)

# aplicamos un modelo de regresion logistica ya que no asume los supuestos de la regresión lineal,
particularmente el de normalidad, Linealidad y homoscedasticidad
modelo1 = glm(formula = qualityDct ~ alcohol+ sulphates + volatile.acidity,family = "binomial",data =
train_set)
modelo1
modelo2 = glm(formula = qualityDct ~ alcohol+ sulphates + volatile.acidity + density,family =
"binomial",data = train_set)
modelo2
modelo3 = glm(formula = qualityDct ~ alcohol+ sulphates + volatile.acidity + density +
total.sulfur.dioxide,family = "binomial",data = train_set)
modelo3
modelo4 = glm(formula = qualityDct ~ alcohol+ sulphates + volatile.acidity + density + total.sulfur.dioxide
+ citric.acid,family = "binomial",data = train_set)
modelo4
modelo5 = glm(formula = qualityDct ~ alcohol+ sulphates + volatile.acidity + density  +
total.sulfur.dioxide + citric.acid + fixed.acidity,family = "binomial",data = train_set)
modelo5
```

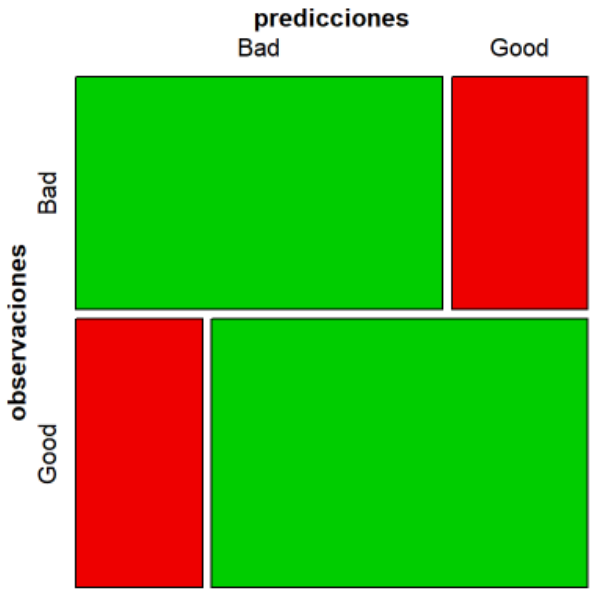
```
Call:  glm(formula = qualityDct ~ alcohol + sulphates + volatile.acidity,
    family = "binomial", data = train_set)

Coefficients:
     (Intercept)           alcohol         sulphates  volatile.acidity
         -9.1599            0.9721            1.5677           -3.3198

Degrees of Freedom: 1199 Total (i.e. Null);  1196 Residual
Null Deviance:      1658
Residual Deviance: 1302         AIC: 1310

Call:  glm(formula = qualityDct ~ alcohol + sulphates + volatile.acidity +
    density, family = "binomial", data = train_set)

Coefficients:
     (Intercept)           alcohol         sulphates  volatile.acidity           density
        -26.2808            0.9855            1.5356           -3.3104           17.0519

Degrees of Freedom: 1199 Total (i.e. Null);  1195 Residual
Null Deviance:      1658
Residual Deviance: 1301         AIC: 1311

Call:  glm(formula = qualityDct ~ alcohol + sulphates + volatile.acidity +
    density + total.sulfur.dioxide, family = "binomial", data = train_set)

Coefficients:
         (Intercept)              alcohol            sulphates      volatile.acidity
           -22.03489              0.88881              1.73294              -3.31038
             density  total.sulfur.dioxide
            14.31762             -0.01429

Degrees of Freedom: 1199 Total (i.e. Null);  1194 Residual
Null Deviance:      1658
Residual Deviance: 1261         AIC: 1273

Call:  glm(formula = qualityDct ~ alcohol + sulphates + volatile.acidity +
    density + total.sulfur.dioxide + citric.acid, family = "binomial",
    data = train_set)

Coefficients:
         (Intercept)              alcohol            sulphates      volatile.acidity
           -81.13383              0.94666              1.88752              -3.97033
             density  total.sulfur.dioxide          citric.acid
            73.51770             -0.01365             -1.12923

Degrees of Freedom: 1199 Total (i.e. Null);  1193 Residual
Null Deviance:      1658
Residual Deviance: 1256         AIC: 1270

Call:  glm(formula = qualityDct ~ alcohol + sulphates + volatile.acidity +
    density + total.sulfur.dioxide + citric.acid + fixed.acidity,
    family = "binomial", data = train_set)

Coefficients:
         (Intercept)              alcohol            sulphates      volatile.acidity
             -4.4851               0.9107               1.9826               -4.0471
             density  total.sulfur.dioxide          citric.acid        fixed.acidity
             -4.0627              -0.0125              -1.7193               0.1377

Degrees of Freedom: 1199 Total (i.e. Null);  1192 Residual
Null Deviance:      1658
Residual Deviance: 1253         AIC: 1269
```

La bondad del modelo se evalua mediante la medida AIC (criterio de Akaike), dado que esta medida tiene en cuenta tanto el error como la complejidad del modelo, elegimos el mejor modelo el que resulte con el menor AIC.

A continuacion realizamos la prediccion en el conjunto de test y creamos la matriz de confusion para evaluar el rendimiento del modelo.

```
# matriz de confusion
cnf = table(test_set[,13],y_pred, dnn = c("observaciones","predicciones"))
cnf

mosaic(cnf, shade = T, colorize = T,
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)),sub = TRUE)

y_pred_num = ifelse(y_pred == "Good", 1, 0)
test_num = ifelse(test_set[,13] == "Good", 1, 0)

pred = ROCR::prediction(y_pred_num,test_num)
perf <- performance(pred, "tpr", "fpr")
plot(perf)

AUCLog1=performance(pred, measure = "auc")@y.values[[1]]
cat("AUC: ",AUCLog1)
```

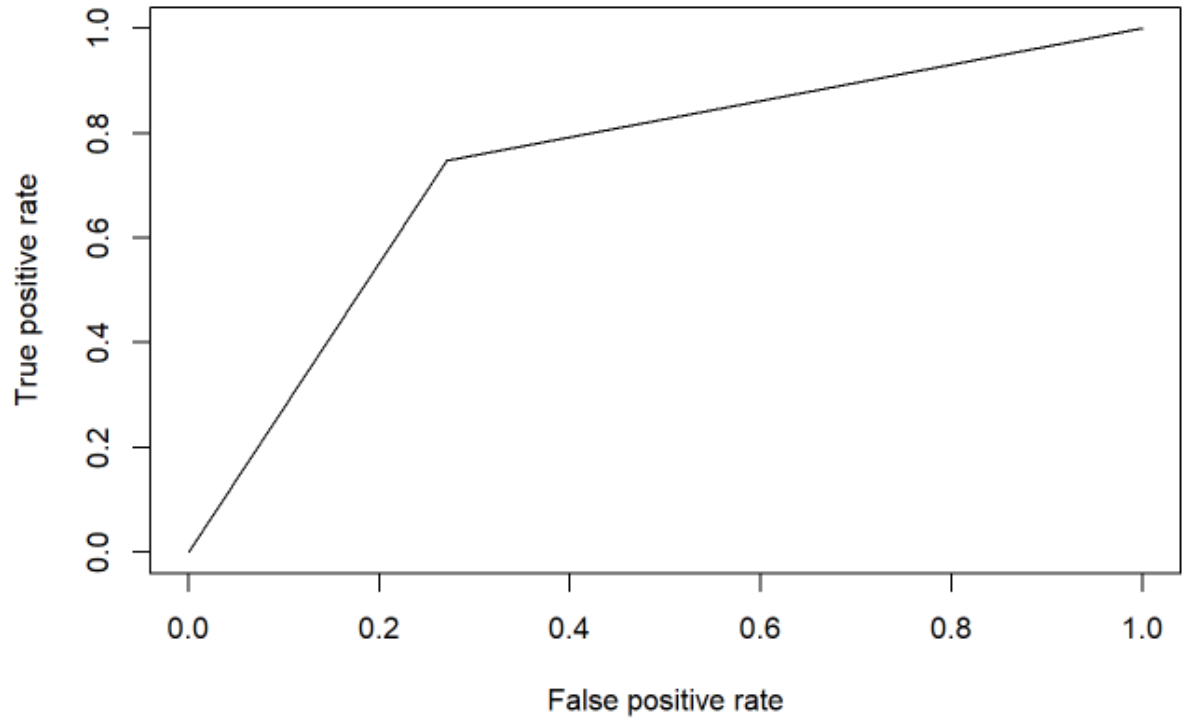
```
             predicciones
observaciones Bad Good
         Bad  134   51
         Good  54  160
AUC:  0.7359939         1
```



cnf

Las predicciones muestran como la mayoria de casos (160 de 210) han sido categorizados correctamente como positivos, y una minoria (50) como falsos positivos y (135 de 189) han sido correctamente categorizados como negativos y 54 como falsos negativos.

El area bajo la curva AUC es de 0.73



Ahora podemos realizar predicciones de la calidad del vino como la siguiente,

```
newdata = data.frame(
  alcohol =  9.3,
  sulphates = 0.34,
  volatile.acidity = 0.75,
  density = 0.9878,
  total.sulfur.dioxide = 45,
  citric.acid = 0.35,
  fixed.acidity = 8.2
)

# mejor modelo modelo5
prob_pred = predict(modelo5,type = "response",newdata = newdata)
prob_pred
y_pred = ifelse(prob_pred > 0.5, "Good", "Bad")
y_pred
```

```
0.08129323
     1
"Bad"
```

6 Resolución del problema

Como se ha visto se han realizado cuatro pruebas estadisticas con el motivo de responder los objetivos que se planteron al inicio en cuanto a la calidad del vino. Cada seccion del analisis ha servido para responder a una pregunta concreta mediante el uso de tablas, graficos o de analisis estadisticos.

Asi en el analisis de correlacion pudimos ver cuales son las propiedades que ejercen mayor influencia sobre la calidad del vino, para posteriormente realizar el analisis de contraste de hipotesis para ver si hay viariabilidad tanto en el nivel de acidez (pH) como de alcohol entre los diferentes niveles de calidad del vino. Lo que nos llevo a sacar conclusiones sobre el efecto del alcohol y el pH en la calidad del vino que puede ser un factor a tener en cuenta para mejorar la produccion de vino.

Por último, el análisis de regresión logística nos permite entrenar un modelo y predecir la calidad de un vino basándonos en sus propiedades básicas además de darnos una idea de la precisión y especificidad del modelo.

Contribuciones	Firma
Investigación previa	Carlos Rea Nogales, Yago Novoa
Redacción de las respuestas	Carlos Rea Nogales, Yago Novoa
Desarrollo código	Carlos Rea Nogales, Yago Novoa