Tipología y Ciclo de Vida de los Dato - Práctica

Autor: Hugo Mourisco Quirós

enero 2022

Descripción del dataset

Antes de comenzar con el análisis del dataset y la limpieza de los datos, procedemos a realizar la lectura de los ficheros en formato CSV, los cuales se encuentran en el siguiente enlace: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

```
## lectura de datos
data_red <- read.csv2("winequality-red.csv")
data_red["type"] <- NA
data_red$type <- "0"

data_white <- read.csv2("winequality-white.csv")
data_white["type"] <- NA
data_white$type <- "1"

data <- rbind(data_red, data_white)

## resumen del data.frame
str(data)
```

```
## 'data.frame':    6497 obs. of  13 variables:
##  $ fixed.acidity       : chr  "7.4" "7.8" "7.8" "11.2" ...
##  $ volatile.acidity    : chr  "0.7" "0.88" "0.76" "0.28" ...
##  $ citric.acid         : chr  "0" "0" "0.04" "0.56" ...
##  $ residual.sugar      : chr  "1.9" "2.6" "2.3" "1.9" ...
##  $ chlorides           : chr  "0.076" "0.098" "0.092" "0.075" ...
##  $ free.sulfur.dioxide : chr  "11" "25" "15" "17" ...
##  $ total.sulfur.dioxide: chr  "34" "67" "54" "60" ...
##  $ density             : chr  "0.9978" "0.9968" "0.997" "0.998" ...
##  $ pH                  : chr  "3.51" "3.2" "3.26" "3.16" ...
##  $ sulphates           : chr  "0.56" "0.68" "0.65" "0.58" ...
##  $ alcohol             : chr  "9.4" "9.8" "9.8" "9.8" ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
##  $ type                : chr  "0" "0" "0" "0" ...
```

El conjunto de datos objeto de análisis está constituido por 13 variables que presentan 6497 observaciones. Las variables de este conjunto de datos son los siguientes:

- **fixed.acidity**: los ácidos fijos predominantes que se encuentran en los vinos son tartárico, málico, cítrico y succínico. Esta variable muestra la cantidad de ácido tartárico por decímetro cúbico.
- **volatile.acidity**: esta variable representa la medida de los ácidos volátiles o gaseosos del vino. El principal ácido volátil del vino es el ácido acético, que también es el principal ácido asociado con el olor y el sabor del vinagre. Esta variable indica la cantidad de ácido acético por decímetro cúbico.
- **citric.acid**: esta variable muestra la cantidad de ácido cítrico por decímetro cúbico.
- **residual.sugar**: el azúcar residual es la cantidad de azúcar que queda en el vino después del proceso de fermentación. Esta variable muestra la cantidad de azúcar residual por decímetro cúbico.
- **chlorides**: la cantidad de cloruro en el vino está influenciada tanto por el terreno como por el tipo de uva, y la importancia de cuantificación radica en el hecho de que el sabor del vino se ve fuertemente afectado por este ión particular, que, en alta concentración, da el vino tiene un sabor salado indeseable y disminuye significativamente su atractivo en el mercado. Esta variable muestra la cantidad de cloruro por decímetro cúbico.
- **free.sulfur.dioxide**: el SO2 evita que el vino reaccione con el oxígeno, lo que puede provocar un pardeamiento y malos olores (oxidación), e inhibe el crecimiento de bacterias y levaduras silvestres indeseables en el mosto y el vino. Esta variable muestra el volumen de sulfitos naturales que se producen durante la fermentación.
- **total.sulfur.dioxide**: esta variable muestra el volumen total de sulfitos.
- **density**: durante la fermentación alcohólica se mide constantemente la densidad. Esta variable muestra el volumen total de sulfitos. Esta variable muestra la densidad de mosto en gramos / centímetros cúbicos.
- **pH**: en la escala de pH cuanto más cerca estamos del 0 más acidez presenta el vino. … Tanto la fermentación alcohólica como la fermentación maloláctica tienden a reducir la Acidez Total del vino, elevando por tanto el pH. La cantidad de acidez es el elemento fundamental, aunque no único, que determina el pH de un vino. Esta variable muestra el nivel de pH.
- **sulphates**: en la elaboración del vino, el sulfato de potasio actúa como antioxidante, eliminando todo el oxígeno suspendido en el vino, lo que ralentiza el envejecimiento. Los tapones de corcho natural permiten la microoxigenación al permitir que pequeñas cantidades de oxígeno regresen al

vino para que los sabores puedan alcanzar su potencial. Esta variable muestra la cantidad de sulfato de potasio por decímetro cúbico.

- **alcohol**: esta variable muestra el volumen de alcohol.
- **quality**: la evaluación de la calidad es a menudo parte del proceso de certificación y se puede utilizar para mejorar la elaboración del vino (identificando los factores más influyentes) y para estratificar vinos como las marcas premium (útil para fijar precios). Está variable establece el nivel de calidad del vino.
- **type**: el tipo de vino (rojo o blanco).

Importancia del dataset

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

Desde hace ya muchos años el interés por el vino es muy importante en el mundo, lo que ha llevado al crecimiento de la industria del vino. La certificación de calidad es un paso crucial para determinar y ajustar el proceso de elaboración del vino y actualmente depende en gran medida de la cata de vinos por parte de expertos humanos. La forma de elaborar el vino en la actualidad, empleando los avances tecnológicos que proporcionan una alta conectividad entre los expertos del mundo ha dado lugar a un nuevo concepto denominado Globalización del vino. Este nuevo concepto hace que viñedos aparentemente separados geográficamente sean tratados de forma similar por un mismo enólogo.

Esta práctica tiene como objetivo analizar dicho dataset con el objetivo de ver que medidas incluídas en el mismo influyen más sobre la certificación de calidad de los vinos. También, se podrá proceder a crear modelos de regresión que sean útiles para apoyar las evaluaciones de cata de vinos, mejorar la producción de vino y desarrollar técnicas que puedan ayudar a modelar campañas de marketing según los gustos de los consumidores en los mercados especializados.

Integración y selección de los datos

Al ser un dataset con un número reducido de variables, creo que no se deben descartar ninguna de las variables antes de la limpieza de los datos. A demás, analizando la descripción de cada variable se comprende que todas ellas son importantes para los objetivos que se plantean. No obstante, a medida que se vaya profundizando en el análisis de cada una de las variables (varianza, valores extremos, etcétera) se seleccionaran grupos de variables.

Limpieza de los datos

En primer lugar vamos a convertir las variables a numeric.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data <- data %>% mutate_if(is.character,as.numeric)
```

Datos vacíos

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

Para conocer si existen datos vacíos en los registros del dataset utilizaremos la función sapply de R junto con la comprobación is.na(variable), la cual se emplea a continuación.

```
# valores desconocidos por variable
sapply(data, function(x) sum(is.na(x)))
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
##                 type
##                    0
```

El resultado es que no existen registros sin valores perdidos, por lo que no realizaremos ninguna operación para resolver este problema. No obstante, para resolver la pregunta de cómo gestionar al encontrarnos valores que contienen ceros o elementos vacíos, se pueden usar diferentes métodos para imputar valores como por ejemplo kNN o métodos de regresion.

Valores extremos

Identificación y tratamiento de valores extremos. En primer lugar observamos los valores mínimos y máximos de cada variable, así como sus cuartiles, la media y la mediana. Para ello utilizaremos la función summary de R, la cual se emplea a continuación por cada variable.

fixed.acidity

```
summary(data$fixed.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.800   6.400   7.000   7.215   7.700  15.900
```

volatile.acidity

```
summary(data$volatile.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0800  0.2300  0.2900  0.3397  0.4000  1.5800
```

citric.acid

```
summary(data$citric.acid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.2500  0.3100  0.3186  0.3900  1.6600
```

residual.sugar

```
summary(data$residual.sugar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.600   1.800   3.000   5.443   8.100  65.800
```

chlorides

```
summary(data$chlorides)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00900 0.03800 0.04700 0.05603 0.06500 0.61100
```

free.sulfur.dioxide

```
summary(data$free.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   17.00   29.00   30.53   41.00  289.00
```

total.sulfur.dioxide

```
summary(data$total.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.0    77.0   118.0   115.7   156.0   440.0
```

density

```
summary(data$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9923  0.9949  0.9947  0.9970  1.0390
```

pH

```
summary(data$pH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.110   3.210   3.219   3.320   4.010
```

sulphates

```
summary(data$sulphates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4300  0.5100  0.5313  0.6000  2.0000
```

alcohol

```
summary(data$alcohol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.30   10.49   11.30   14.90
```

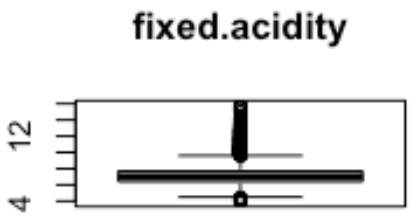
quality

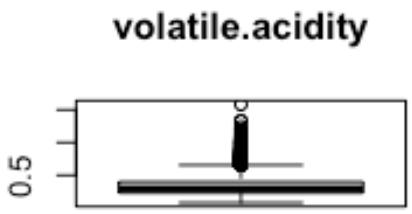
```
summary(data$quality)
```

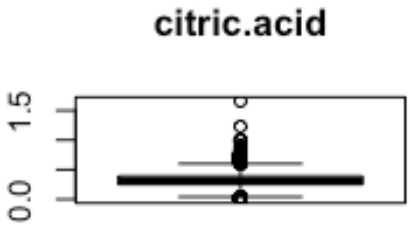
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.818   6.000   9.000
```

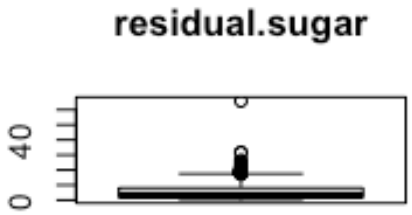
Para observar de una forma más sencilla los valores extremos utilizaremos los diagramas de caja con la función boxplot de R, la cual se emplea a continuación por cada variable.

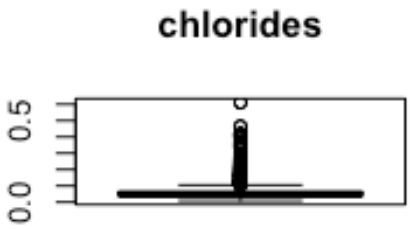
```
par(mfrow=c(2,2))
for (i in 1:ncol(data[1:ncol(data) -1])) {
  boxplot(data[,i], main = colnames(data)[i], width=100)
}
```

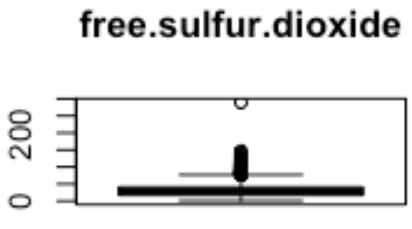

fixed.acidity
4
12

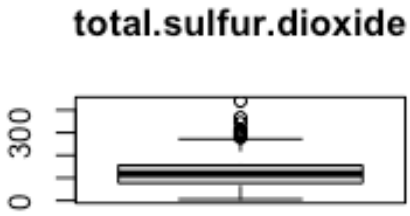

volatile.acidity
0.5

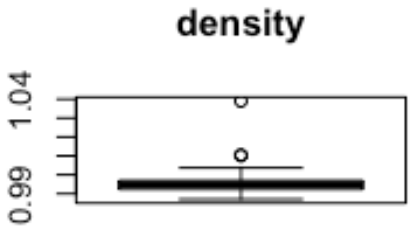

citric.acid
0.0
1.5

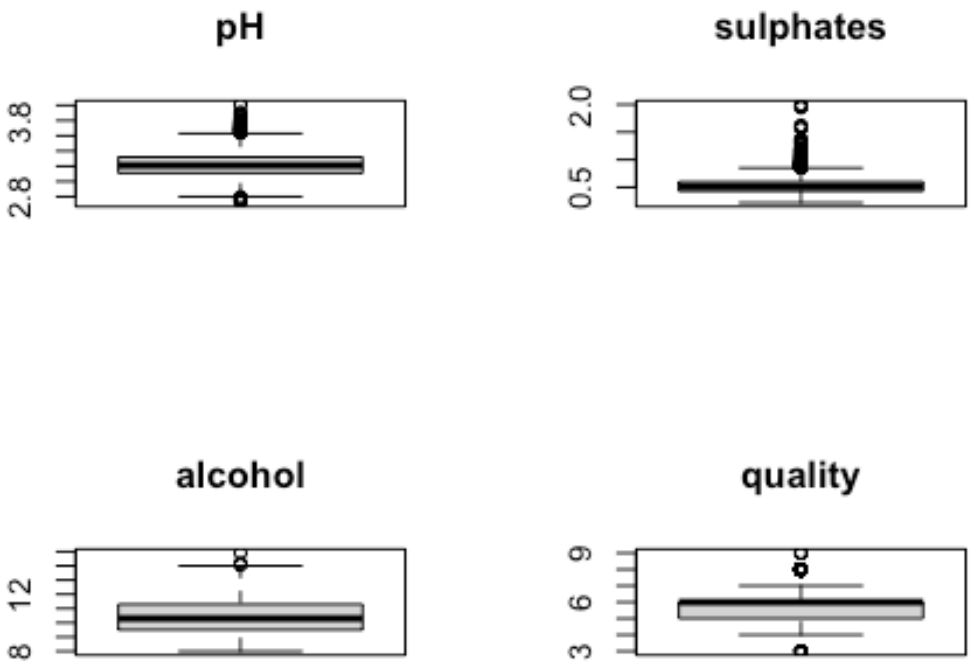

residual.sugar
0
40


chlorides
0.0
0.5


free.sulfur.dioxide
0
200


total.sulfur.dioxide
0
300


density
0.99
1.04



Los valores extremos parecen ser congruentes respecto a los valores que pueden tomar cada una de las variables. Si es cierto que algunas variables tienen valores que distan “mucho” del rango intercuartílico, pero entendemos que no son registros que provengan de datos erróneos.

Análisis de los datos

Selección de los grupos de datos

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

En esta practica se selecciona la variable **type** para determinar la separación en grupos. Recordamos que el valor 0 representa los registros para vino tinto y el valor 1 para los registros para vino blanco.

Normalidad y homogeneidad de la varianza

Comprobación de la normalidad y homogeneidad de la varianza.

Se utiliza la prueba de normalidad de Shapiro-Test para comprobar que los valores que toman las variables cuantitativas estan distribíidos normalmente. Si se obtiene un p-valor superior al nivel de significación prefijado α = 0, 05, entonces se considera que variable en cuestión sigue una distribución normal.

fixed.acidity

```
shapiro.test(data$fixed.acidity[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$fixed.acidity[0:5000]
## W = 0.88397, p-value < 2.2e-16
```

volatile.acidity

```
shapiro.test(data$volatile.acidity[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$volatile.acidity[0:5000]
## W = 0.89185, p-value < 2.2e-16
```

citric.acid

```
shapiro.test(data$citric.acid[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$citric.acid[0:5000]
## W = 0.97068, p-value < 2.2e-16
```

residual.sugar

```
shapiro.test(data$residual.sugar[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$residual.sugar[0:5000]
## W = 0.79355, p-value < 2.2e-16
```

chlorides

```
shapiro.test(data$chlorides[0:5000])

##
##  Shapiro-Wilk normality test
##
```

```
## data:  data$chlorides[0:5000]
## W = 0.62495, p-value < 2.2e-16
```

free.sulfur.dioxide

```
shapiro.test(data$free.sulfur.dioxide[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$free.sulfur.dioxide[0:5000]
## W = 0.95169, p-value < 2.2e-16
```

total.sulfur.dioxide

```
shapiro.test(data$total.sulfur.dioxide[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$total.sulfur.dioxide[0:5000]
## W = 0.97424, p-value < 2.2e-16
```

density

```
shapiro.test(data$density[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$density[0:5000]
## W = 0.96248, p-value < 2.2e-16
```

pH

```
shapiro.test(data$pH[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$pH[0:5000]
## W = 0.99253, p-value = 1.503e-15
```

sulphates

```
shapiro.test(data$sulphates[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$sulphates[0:5000]
## W = 0.89673, p-value < 2.2e-16
```

alcohol

```
shapiro.test(data$alcohol[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$alcohol[0:5000]
## W = 0.94914, p-value < 2.2e-16
```

quality

```
shapiro.test(data$quality[0:5000])

##
##  Shapiro-Wilk normality test
##
## data:  data$quality[0:5000]
## W = 0.88545, p-value < 2.2e-16
```

El test nos indica que ninguna variable está normalizada, ya que el valor p-valor es inferior a $\alpha = 0,05$.

A continuación, se estudia homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. Estudiamos esta homogeneidad en cuanto a los grupos conformados por los vinos tintos con respecto a los vinos blancos. Las varianzas son iguales determinan la hipotesis nula.

```
fligner.test(quality ~ type, data = data)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by type
## Fligner-Killeen:med chi-squared = 0.61775, df = 1, p-value = 0.4319
```

Como resultado obtenemos que el valor p-valor es superiro a $\alpha = 0,05$, entonces aceptamos que las varianzas de ambas muestras son homogéneas.

Aplicación de pruebas estadísticas

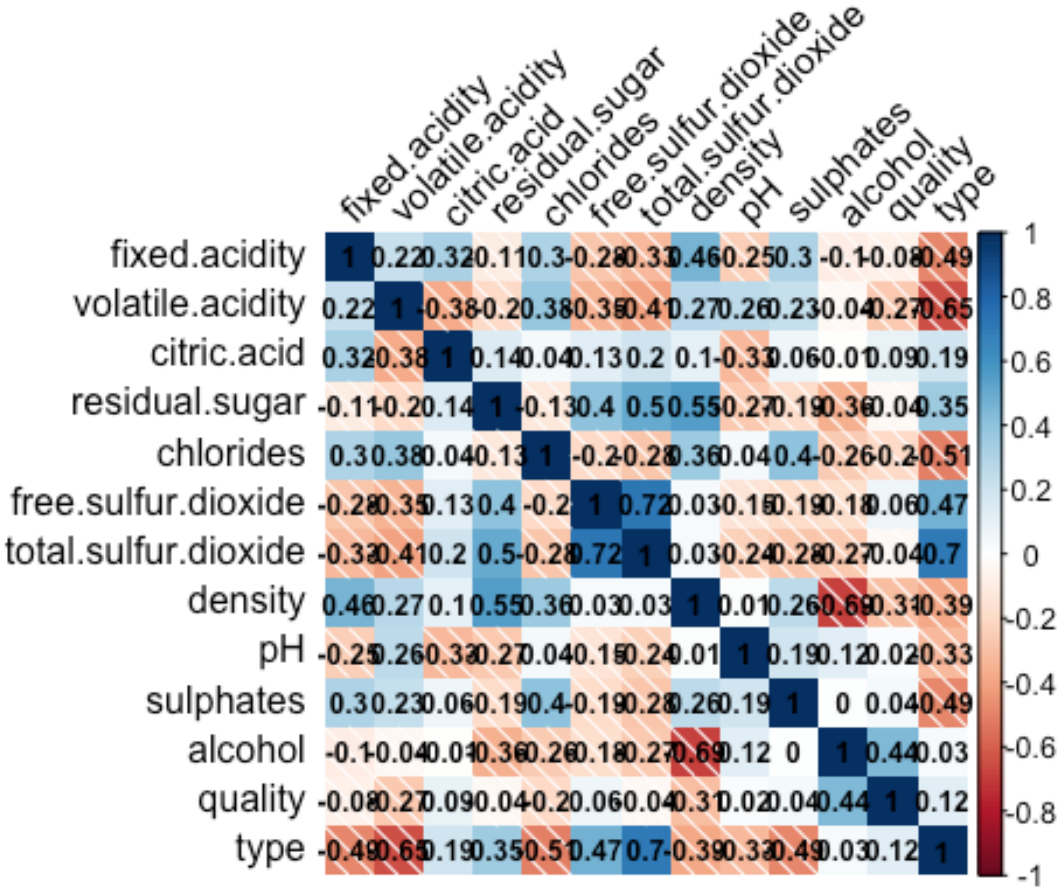
Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

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

El primer método que vamos a aplicar va a ser un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre la calidad del vino tino y del vino blanco. Para ello, se utilizará el coeficiente de

correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.1

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.1.1

## corrplot 0.92 loaded

corData.cor <- round(cor(data, method = "pearson", use = "complete.obs"),
digits=2)
corrplot(corData.cor, method = "shade", addCoef.col = "black", tl.col =
"black", tl.srt = 45, number.cex=0.75)
```



Con el anterior diagrama determinamos que no existe ninguna correlación fuerte entre las variables y la calidad del vino, por lo que entendemos que no se puede determinar que a partir de una variable se pueda determinar la calidad del vino.

¿La calidad del vino es superior si el vino es de tipo vino tinto?

La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si la calidad del vino es superior dependiendo del tipo, vino tinto o vino blanco.

Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa: H0: media de la calidad dela poblacion de vino tinto - media de la calidad dela poblacion de vino blanco = 0 H1: media de la calidad dela poblacion de vino tinto - media de la calidad dela poblacion de vino blanco < 0

```
t.test(data_red$quality, data_white$quality, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  data_red$quality and data_white$quality
## t = -10.149, df = 2950.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2026732
## sample estimates:
## mean of x mean of y
##  5.636023  5.877909
```

Puesto que obtenemos un p-valor menor que el valor de significación fijado a α = 0, 05, rechazamos la hipótesis nula. Aceptamos la hipotesis alternativa, la calidad del vino tinto en las muestras es inferior a la calidad del vino blanco.

Regresión lineal para predecir la calidad del vino

Determinada que no existe una correlación fuerte de las variables respecto a la calidad del vino, vamos a utilizar en primer lugar un modelo de regresión linear con la variable alcohol y a continuación otro modelo con un conjunto de variables para observar que resultados ofrece los modelos de regresión y si podemos precedir la calidad del vino a partir de dichas variables.

```
# modelo de regresión con una variable (alcohol)
quality_lm = lm(quality ~ alcohol, data = data)
summary(quality_lm)
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5042 -0.4957 -0.0488  0.5043  3.2115
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.405269   0.085941   27.99   <2e-16 ***
## alcohol     0.325312   0.008139   39.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7824 on 6495 degrees of freedom
## Multiple R-squared:  0.1974, Adjusted R-squared:  0.1973
## F-statistic:  1598 on 1 and 6495 DF,  p-value: < 2.2e-16
```

```
# modelo de regresión con un conjunto de variables
quality_lm = lm(quality ~ alcohol + volatile.acidity + density, data =
data)
summary(quality_lm)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + density,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4295 -0.4846 -0.0343  0.4706  3.0262
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -35.11411    4.57246  -7.679 1.83e-14 ***
## alcohol            0.38253    0.01095  34.935  < 2e-16 ***
## volatile.acidity  -1.49092    0.05991 -24.887  < 2e-16 ***
## density           37.62502    4.52171   8.321  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7477 on 6493 degrees of freedom
## Multiple R-squared:  0.2673, Adjusted R-squared:  0.267
## F-statistic: 789.6 on 3 and 6493 DF,  p-value: < 2.2e-16
```

Obtenemos que usando varias variables en el modelo, el resultado es más conveniente dado que tiene un mayor coeficiente de determinación (r-squared).

Representación de los resultados

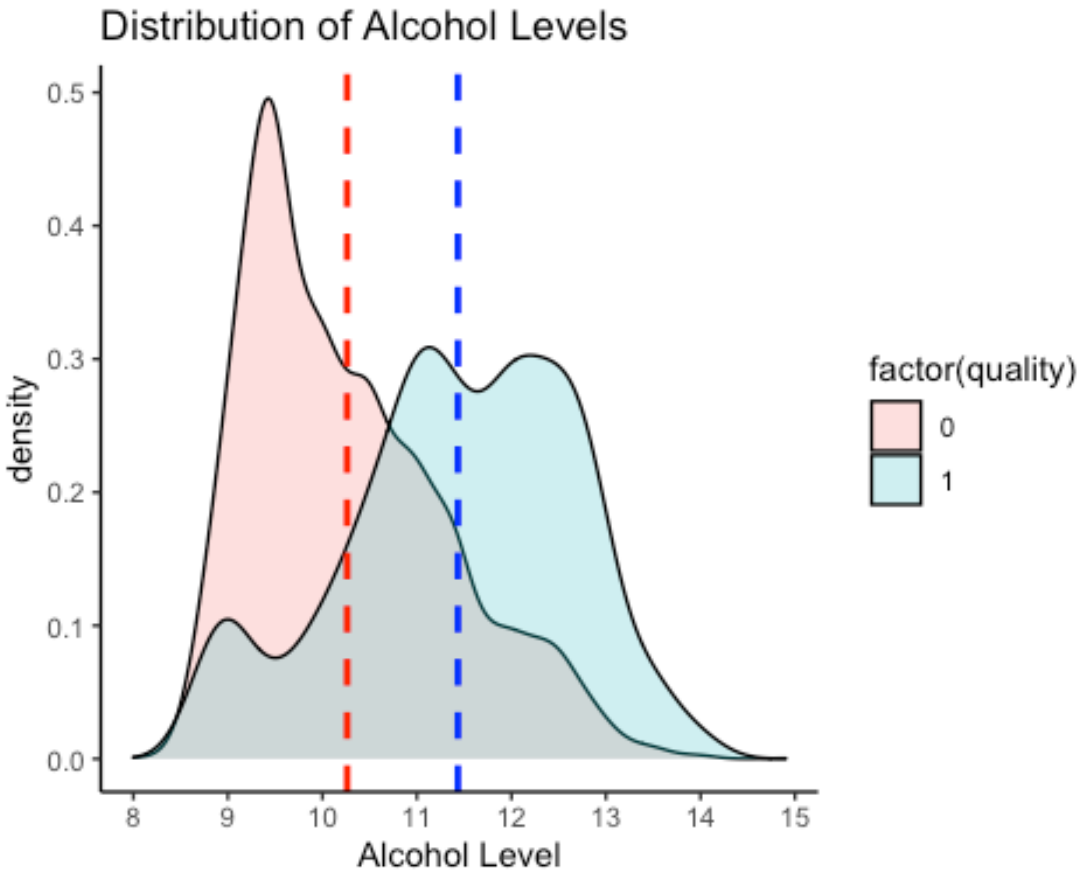
Representación de los resultados a partir de tablas y gráficas.

Vamos a convertir en primer lugar la variable **quality** en binaria, para determinar si un vino es bueno o no.

```
data$quality <- ifelse(data$quality > 6, 1, 0)
```

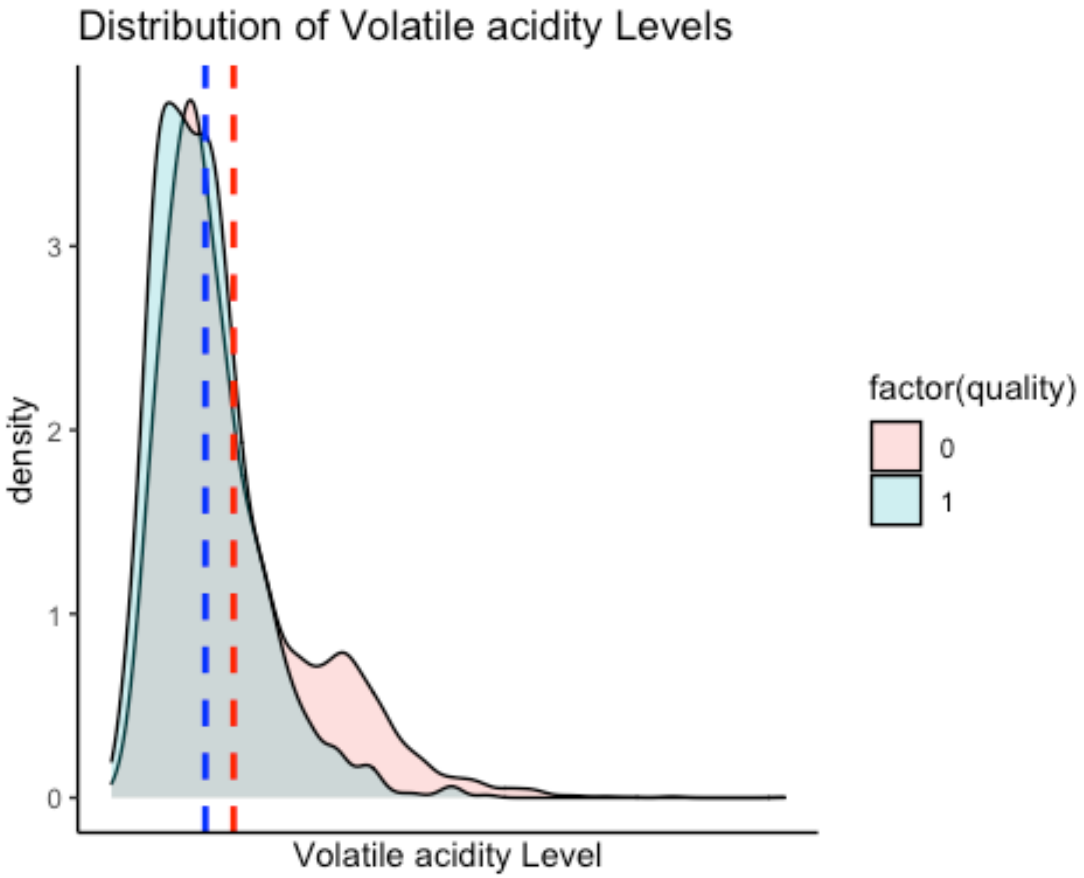
Esta conversión nos permite analizar las variables usadas en el segundo modelo de regresión lineal para determinar si un vino es bueno o malo.

```
ggplot(data,
aes(x=alcohol,fill=factor(quality)))+geom_density(alpha=0.25)+

geom_vline(aes(xintercept=mean(alcohol[quality==0],na.rm=T)),color="red",
linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(alcohol[quality==1],na.rm=T)),color="blue"
,linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(4,16,1))+
  xlab(label = "Alcohol Level")+
  ggtitle("Distribution of Alcohol Levels")+
  theme_classic()
```

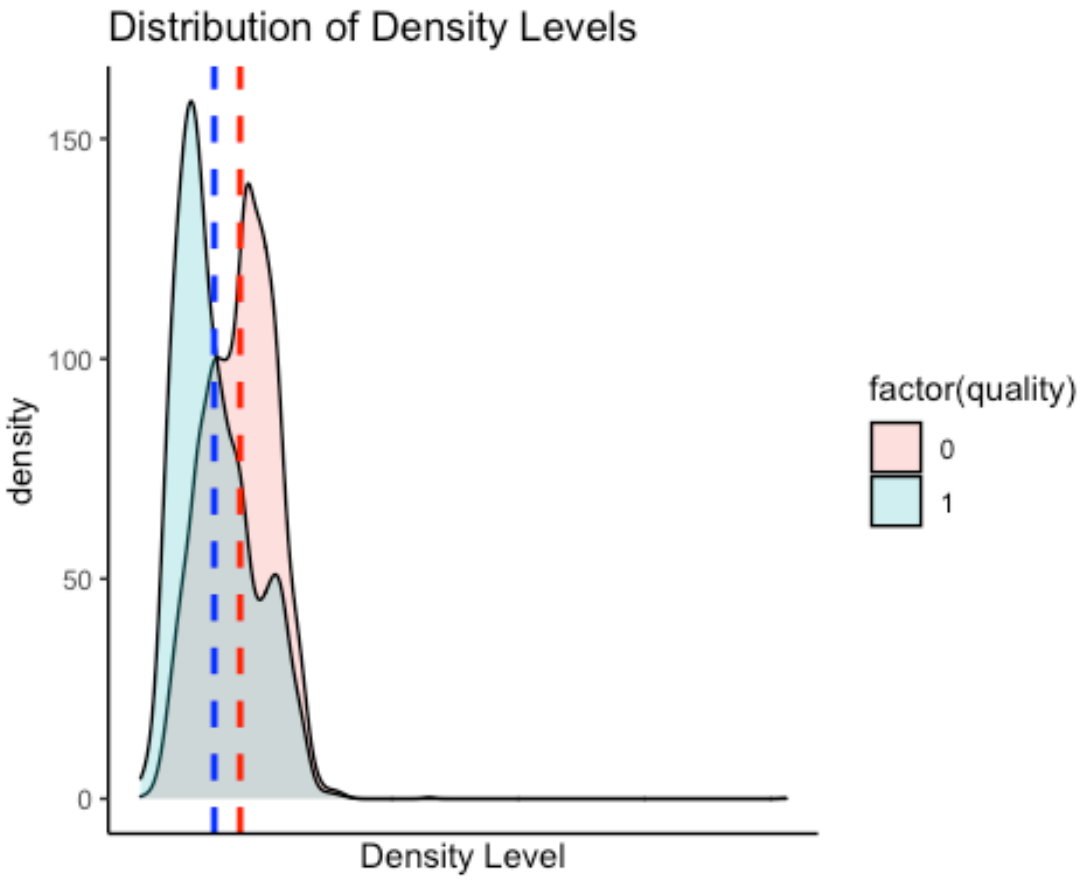



```
ggplot(data,
aes(x=volatile.acidity,fill=factor(quality)))+geom_density(alpha=0.25)+

geom_vline(aes(xintercept=mean(volatile.acidity[quality==0],na.rm=T)),col
or="red",linetype="dashed",lwd=1)+
```

```
geom_vline(aes(xintercept=mean(volatile.acidity[quality==1],na.rm=T)),col
or="blue",linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(4,16,1))+
  xlab(label = "Volatile acidity Level")+
  ggtitle("Distribution of Volatile acidity Levels")+
  theme_classic()
```




```
ggplot(data,
aes(x=density,fill=factor(quality)))+geom_density(alpha=0.25)+

geom_vline(aes(xintercept=mean(density[quality==0],na.rm=T)),color="red",
linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(density[quality==1],na.rm=T)),color="blue"
,linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(4,16,1))+
  xlab(label = "Density Level")+
  ggtitle("Distribution of Density Levels")+
  theme_classic()
```



Conclusiones

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

En primer lugar se ha observado que los dos dataset analizados tenían ya un preprocesamiento bien determinado, posiblemente porque los registros se hayan obtenido de sensores destinados a tal objetivo. Y es que no ha sido necesario realizar ajustes para solucionar posibles casos de valores nulos o vacíos y a valores extremos como resultado de una incorrecta generación del dataset.

En segundo lugar se ha hecho un breve análisis estadístico con el objetivo de determinar la correlación de todos los parámetros con respecto a la calidad del vino, no obteniedo grandes evidencias de que ciertas variables lleguen a influir. También se ha realizado un contraste de hipotesis a partir de las observaciones registradas para el vino tino y el vino blanco, determinando mejores valoraciones en cuanto a calidad para las observaciones de vino blanco. Y por último se ha aplicado dos modelos de

regresión lineal para determinar que para determinar la calidad del vino, cuantas más variables se utilicen mayor determinación tendrá la predicción de la calidad.

Exportación de los datos utilizados

```
write.csv(data, file = "./winequality-data.csv")
```