Práctica 2: Limpieza y Análisis de Datos

Fernando Meza Ibarra y Tania Gualli Culqui

Mayo 2020

Contents

1 Detalles de la actividad

1.1 Descripción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2 Objetivos

Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares. Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico. Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos. Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico. Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación. Desarrollar las habilidades de aprendizaje que les permitan continuar

estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo. Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.3 Competencias

Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo. Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2 Resolución

2.1 Descripción del dataset.

El conjunto de datos objeto de análisis se ha obtenido a partir de: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009 en Kaggle y está constituido por 12 características (columnas) que presentan 1599 muestras de vinos (filas o registros).

Los campos de este conjunto de datos son los siguientes:

Variables de entrada (basadas en pruebas fisicoquímicas):
1 - acidez fija
2 - acidez volátil
3 - ácido cítrico
4 - azúcar residual
5 - cloruros
6 - dióxido de azufre libre
7 - dióxido de azufre total
8 - densidad
9 - pH
10 - sulfatos
11 - alcohol

Variable de salida (basada en datos sensoriales):
12 - calidad (puntuación entre 0 y 10)

2.2 Importancia y objetivos de los análisis

En los últimos años, el interés por el vino ha aumentado, lo que lleva a crecimiento de la industria del vino. Como consecuencia, las empresas deben invertir en nuevas tecnologías para mejorar la producción y venta de vino. La certificación de calidad es un paso crucial para ambos procesos y actualmente depende en gran medida de la cata de vinos por expertos humanos. Sin embargo, las evaluaciones se basan en experiencia y conocimiento de los expertos, que pueden ser propensos a factores subjetivos. Este trabajo es importante para la industria del vino, pues el enfoque propuesto basado en datos, tiene como objetivo la predicción de preferencias de vino a partir de pruebas analíticas objetivas y por lo tanto se puede integrar en un Sistema de Soporte de Decisión (siglas en inglés, DSS), ayudando a la velocidad y calidad del desempeño del enólogo. Por ejemplo, el experto podría repetir la degustación solo si su calificación está lejos de la predicha por el modelo. El modelo también podría usarse para mejorar la capacitación de estudiantes de enología.

Este estudio de caso será abordado por tareas de regresión, donde cada preferencia de tipo de vino se modela en una escala continua, de 0 (muy mal) a 10 (excelente).

Además, la importancia relativa de las variables de entrada trajo ideas interesantes sobre el impacto de la analítica, dado que algunas variables pueden controlarse en el proceso de producción. Esta información se puede utilizar para mejorar la calidad del vino. Por ejemplo, la concentración de alcohol puede aumentarse o disminuirse mediante el monitoreo de la concentración de azúcar de uva antes de la cosecha. (1)

(1) P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

2.3 Análisis descriptivo de los datos

Procedemos a realizar la lectura del fichero en formato CSV en el que se encuentra. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame:

```
#Lectura del fichero
Archivo_Wine_Quality <-"winequality-red.csv"
datos_vinos <-read.csv(Archivo_Wine_Quality, sep=";",na.strings = "NA")
head(datos_vinos) #Para confirmar se muestra las primeras filas
```

```
##   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
```

```
#número total de registros y variables
str(datos_vinos)
```

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

Se confirma que el dataset tiene 1599 registros y 12 variables.

```
#Tipo de dato asignado a cada campo
sapply(datos_vinos,class)
```

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

Se puede observar que los tipos de datos asignados automáticamente por R a las variables, corresponden con el dominio de estas.

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

Todos los atributos presentes en el conjunto de datos corresponden a las características fisicoquímicas que permiten identificar los diferentes tipos de vinos, por lo que será conveniente incluirlos a todos durante la realización de los análisis.

4 Limpieza de los datos

En esta sección se analizarán si los datos contienen ceros o elementos vacíos, así como valores extremos y cómo gestionar cada uno de estos casos.

4.1 Ceros y elementos vacíos

Comúnmente, se utilizan los ceros como centinela para indicar la ausencia de ciertos valores. Así, se procede a conocer a continuación qué campos contienen elementos vacíos:

```
sapply(datos_vinos, 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
```

No se observan registros que contengan valores desconocidos para algún campo. En caso de tenerlos una opción podría ser eliminar los registros que incluyen este tipo de valores, pero ello supondría perder información. Como alternativa, se podría emplear un métodos de imputación de valores que dependerá de las características de los datos.

4.2 Valores extremos

Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso de dos vías: (1) representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja) o (2) utilizar la función boxplots.stats() de R. En este caso se usará la opción 2, que permitirá mostrar los valores atípicos para aquellas variables que los contienen:

```
boxplot.stats(datos_vinos$fixed.acidity)$out
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8
## [16] 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9
```

```
## [31] 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9
## [46] 13.3 12.9 12.6 12.6
```

```
boxplot.stats(datos_vinos$volatile.acidity)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035
## [13] 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

```
boxplot.stats(datos_vinos$citric.acid)$out
```

```
## [1] 1
```

```
boxplot.stats(datos_vinos$residual.sugar)$out
```

```
##   [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65
##  [13]  4.65  5.50  5.50  5.50  5.50  7.30  7.20  3.80  5.60  4.00  4.00  4.00
##  [25]  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80
##  [37]  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70
##  [49]  5.20 15.50  4.10  8.30  6.55  6.55  4.60  6.10  4.30  5.80  5.15  6.30
##  [61]  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90  4.60  5.10  5.60  5.60
##  [73]  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60
##  [85]  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10
##  [97]  6.20  8.90  4.00  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70
## [109]  5.50  5.50  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10 12.90
## [121]  4.30 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80
## [133]  6.10  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75 13.80
## [145] 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90  5.10  7.80
```

```
boxplot.stats(datos_vinos$chlorides)$out
```

```
##   [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146
##  [13] 0.236 0.610 0.360 0.270 0.039 0.337 0.263 0.611 0.358 0.343 0.186 0.213
##  [25] 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121
##  [37] 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148
##  [49] 0.122 0.124 0.124 0.143 0.222 0.039 0.157 0.422 0.034 0.387 0.415 0.157
##  [61] 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.012
##  [73] 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178
##  [85] 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414
##  [97] 0.166 0.168 0.415 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205
## [109] 0.039 0.235 0.230 0.038
```

```
boxplot.stats(datos_vinos$free.sulfur.dioxide)$out
```

```
##  [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52
## [26] 55 55 48 48 66
```

```
boxplot.stats(datos_vinos$total.sulfur.dioxide)$out
```

```
##  [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145
## [20] 144 135 165 124 124 134 124 129 151 133 142 149 147 145 148 155 151 152 125
## [39] 127 139 143 144 130 278 289 135 160 141 141 133 147 147 131 131 131
```

```
boxplot.stats(datos_vinos$density)$out
```

```
##  [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220
## [10] 1.00140 1.00140 1.00140 1.00140 1.00320 1.00260 1.00140 1.00315 1.00315
## [19] 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260 0.99210 0.99154 0.99064
## [28] 0.99064 1.00289 0.99162 0.99007 0.99007 0.99020 0.99220 0.99150 0.99157
## [37] 0.99080 0.99084 0.99191 1.00369 1.00369 1.00242 0.99182 1.00242 0.99182
```

```r
boxplot.stats(datos_vinos$pH)$out
```

```
##  [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89
## [16] 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78 4.01 2.90
## [31] 4.01 3.71 2.88 3.72 3.72
```

```r
boxplot.stats(datos_vinos$sulphates)$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59
## [16] 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06
## [31] 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18
## [46] 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
```

```r
boxplot.stats(datos_vinos$alcohol)$out
```

```
##  [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000
##  [9] 13.60000 14.00000 14.00000 13.56667 13.60000
```

Dado que los valores de las variables de entrada se obtuvieron mediante pruebas físico - químicas, los valores identificados como extremos podrían ser correctos, por lo tanto, el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

5 Análisis de los datos.

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

En esta fase es importante identificar los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. Sin embargo debido al desconocimiento de la industria de vinos se torna complicado definirlos. Pese a ello, de manera general puede resultar interesante analizar y/o comparar el nivel de alcohol con respecto a la calidad del vino. En la sección de pruebas estadísticas se identifican que variables influyen más en la calidad del vino.

```r
#Verificar el valor de la mediana de la variable alcohol
summary(datos_vinos$alcohol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

```r
# Agrupación por nivel de alcohol nivel 1: menor o igual a 10.20, nivel 2:mayor a 10.20
vinos_alcohol1 <- datos_vinos$alcohol[datos_vinos$alcohol<=10.20]
vinos_alcohol2 <- datos_vinos$alcohol[datos_vinos$alcohol > 10.20]
```

5.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Anderson-Darling. Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado alpha = 0,05. Si esto se cumple, entonces se considera que la variable en cuestión sigue una distribución normal.

```r
library(nortest)
alpha = 0.05
col.names = colnames(datos_vinos)
for (i in 1:ncol(datos_vinos)) {
if (i == 1) cat("Variables que no siguen una distribución normal:\n")
if (is.integer(datos_vinos[,i]) | is.numeric(datos_vinos[,i])) {
p_val = ad.test(datos_vinos[,i])$p.value
```

```
if (p_val < alpha) {
cat(col.names[i])
# Format output
if (i < ncol(datos_vinos) - 1) cat(", \n")
if (i %% 3 == 0) cat("\n")
}
}
}
```

```
## Variables que no siguen una distribución normal:
## fixed.acidity,
## volatile.acidity,
## citric.acid,
##
## residual.sugar,
## chlorides,
## free.sulfur.dioxide,
##
## total.sulfur.dioxide,
## density,
## pH,
##
## sulphates,
## alcoholquality
```

Como se puede ver todas las variables de este conjunto de datos no siguen una distribución normal.

Ahora, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por los vinos que presentan nivel de alcohol 1 frente a los vinos que presentan nivel de alcohol 2. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
fligner.test(quality ~ alcohol, data = datos_vinos)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  quality by alcohol
## Fligner-Killeen:med chi-squared = 135.98, df = 64, p-value = 4.157e-07
```

Puesto que obtenemos un p-valor menor a 0.05, rechazamos la hipótesis de que las varianzas de ambas muestras son iguales u homogéneas.

6 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.

6.1 Identificación de variables que influyen más en la calidad del vino

En primer lugar, procedemos a realizar un análisis de correlación entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre en la calidad del vino. Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```r
correlaciones <- cor(datos_vinos, method = 'spearman')
correlaciones
```

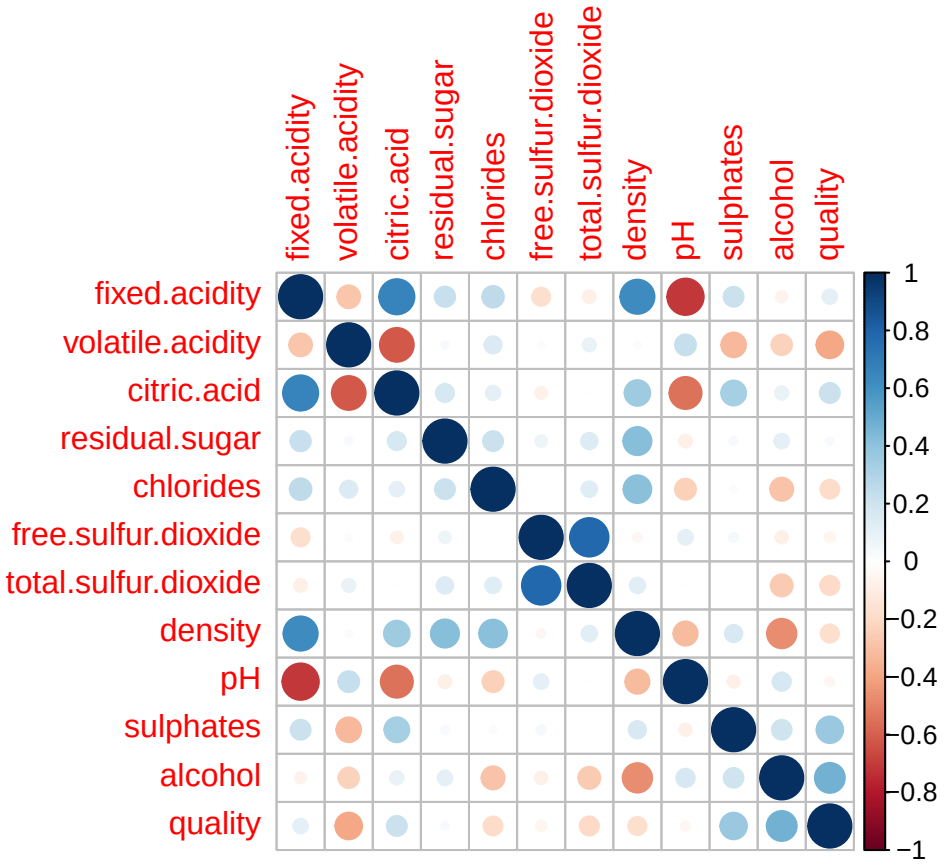
```
##                       fixed.acidity volatile.acidity  citric.acid residual.sugar
## fixed.acidity           1.00000000      -0.27828222  0.661708417     0.22070086
## volatile.acidity       -0.27828222       1.00000000 -0.610259467     0.03238560
## citric.acid             0.66170842      -0.61025947  1.000000000     0.17641731
## residual.sugar          0.22070086       0.03238560  0.176417306     1.00000000
## chlorides               0.25090411       0.15877025  0.112576508     0.21295924
## free.sulfur.dioxide    -0.17513656       0.02116264 -0.076451575     0.07461786
## total.sulfur.dioxide   -0.08841741       0.09411014  0.009399602     0.14537506
## density                 0.62307076       0.02501412  0.352285261     0.42226586
## pH                     -0.70667359       0.23357152 -0.548026276    -0.08997095
## sulphates               0.21265375      -0.32558398  0.331074404     0.03833200
## alcohol                -0.06657566      -0.22493168  0.096455544     0.11654813
## quality                 0.11408367      -0.38064651  0.213480914     0.03204817
##                         chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.2509041064       -0.1751365613        -0.0884174083
## volatile.acidity      0.1587702548        0.0211626414         0.0941101376
## citric.acid           0.1125765077       -0.0764515753         0.0093996024
## residual.sugar        0.2129592419        0.0746178640         0.1453750584
## chlorides             1.0000000000        0.0008051686         0.1300333418
## free.sulfur.dioxide   0.0008051686        1.0000000000         0.7896978767
## total.sulfur.dioxide  0.1300333418        0.7896978767         1.0000000000
## density               0.4113896972       -0.0411776800         0.1293321018
## pH                   -0.2343612736        0.1156791779        -0.0098414382
## sulphates             0.0208254792        0.0458623500        -0.0005038194
## alcohol              -0.2845039422       -0.0813673063        -0.2578060251
## quality              -0.1899223356       -0.0569006455        -0.1967350754
##                          density           pH     sulphates     alcohol
## fixed.acidity         0.62307076 -0.706673595  0.2126537506 -0.06657566
## volatile.acidity      0.02501412  0.233571519 -0.3255839818 -0.22493168
## citric.acid           0.35228526 -0.548026276  0.3310744040  0.09645554
## residual.sugar        0.42226586 -0.089970954  0.0383320002  0.11654813
## chlorides             0.41138970 -0.234361274  0.0208254792 -0.28450394
## free.sulfur.dioxide  -0.04117768  0.115679178  0.0458623500 -0.08136731
## total.sulfur.dioxide  0.12933210 -0.009841438 -0.0005038194 -0.25780603
## density               1.00000000 -0.312055078  0.1614782344 -0.46244458
## pH                   -0.31205508  1.000000000 -0.0803060380  0.17993243
## sulphates             0.16147823 -0.080306038  1.0000000000  0.20732955
## alcohol              -0.46244458  0.179932427  0.2073295535  1.00000000
## quality              -0.17707407 -0.043671935  0.3770601991  0.47853169
##                          quality
## fixed.acidity         0.11408367
## volatile.acidity     -0.38064651
## citric.acid           0.21348091
## residual.sugar        0.03204817
## chlorides            -0.18992234
## free.sulfur.dioxide  -0.05690065
## total.sulfur.dioxide -0.19673508
## density              -0.17707407
## pH                   -0.04367193
## sulphates             0.37706020
## alcohol               0.47853169
```

```
## quality                1.00000000
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(correlaciones)
```



Así, identificamos cuáles son las variables más correlacionadas con la calidad en función de su proximidad con los valores -1 y +1. Con esta consideración se puede observar que las variables más relevantes para la calidad del vino son: volatile.acidity, sulphates y alcohol.

6.2 Prueba de Hipótesis

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 grado de alcohol. Para ello, tendremos dos muestras: la primera corresponde a los vinos con grado de alcohol menor o igual a 10,20 (Mediana) y, la segunda, con aquellos que presentan un grado de alcohol superior a 10,20. Se debe destacar que un test paramétrico como el que a continuación se utiliza necesita que los datos sean normales, si la muestra es de tamaño inferior a 30. Como en nuestro caso, n $>$ 30, el contraste de hipótesis siguiente es válido (aunque podría utilizarse un test no paramétrico como el de Mann-Whitney, que podría resultar ser más eficiente para este caso).

```
vinos_alcohol1 <- datos_vinos$alcohol[datos_vinos$alcohol<=10.20]
vinos_alcohol2 <- datos_vinos$alcohol[datos_vinos$alcohol > 10.20]
```

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 : u1 - u2 = 0, y H1 : u1 - u2 $<$ 0

Donde u1 es la media de la población de la que se extrae la primera muestra y u2 es la media de la población de la que extrae la segunda. El valor de significación se fija en alpha = 0,05.

```
t.test(vinos_alcohol1, vinos_alcohol2, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  vinos_alcohol1 and vinos_alcohol2
## t = -55.004, df = 981.87, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.696035
## sample estimates:
## mean of x mean of y
##  9.595269 11.343637
```

Puesto que obtenemos un p-valor menor que el valor de significación fijado, rechazamos la hipótesis nula. Por tanto, podemos concluir que, la calidad del vino es superior si éste tiene altos niveles de alcohol.

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

Tabla de Estadísticas

Podemos ver una representación estadística muy completa de nuestro conjunto de datos. Con la función **describe()** podemos explorar el contenido de cualquier dataset y obtener una visión general sobre su estructura.
Más allá de ofrecernos una primera mirada a los resultados del estudio, también nos permite identificar cualquier error en la digitación o en la construcción del conjunto de datos.

Ésta función forma parte de la librería "psych" que entrega los descriptivos de cada una de las variables. En cuanto a la calidad de la base de datos lo más importante es revisar el valor de **n**, que nos indica la cantidad de observaciones que han sido representadas en cada una de las variables (un **n** con valor 0, probablemente refleja un error en la creación de la base de datos) y el rango de respuesta **(min y max)** nos permite apreciar la coherencia de rangos para cada variable, es decir que podríamos detectar errores en la creación del conjunto de datos observando los rangos.

Además de esto, describe() entrega una buena cantidad de información descriptiva. El **promedio**, la **mediana**, la **desviación estándar** y el **grado de asimetría (skew)**, Si la distribución es simétrica, ambos índices son iguales a 0; si es asimétrica a la derecha, ambos son positivos; y si es asimétrica a la izquierda, ambos índices son negativos, por otro lado, la **Kurtosis** (también conocida como medida de apuntamiento) es una medida estadística, que determina el grado de concentración que presentan los valores de una variable alrededor de la zona central de la distribución de frecuencias. Si este coeficiente es nulo, la distribución se dice normal (similar a la distribución normal de Gauss) y recibe el nombre de mesocúrtica.
Si el coeficiente es positivo, la distribución se llama leptocúrtica, más puntiaguda que la anterior. Hay una mayor concentración de los datos en torno a la media.
Si el coeficiente es negativo, la distribución se llama platicúrtica y hay una menor concentración de datos en torno a la media. Sería más achatada que la primera.

```
describe(datos_vinos)
```

```
##                      vars    n  mean    sd median trimmed   mad  min    max
## fixed.acidity           1 1599  8.32  1.74   7.90    8.15  1.48 4.60  15.90
## volatile.acidity        2 1599  0.53  0.18   0.52    0.52  0.18 0.12   1.58
## citric.acid             3 1599  0.27  0.19   0.26    0.26  0.25 0.00   1.00
## residual.sugar          4 1599  2.54  1.41   2.20    2.26  0.44 0.90  15.50
## chlorides               5 1599  0.09  0.05   0.08    0.08  0.01 0.01   0.61
```

```
## free.sulfur.dioxide     6 1599 15.87 10.46  14.00   14.58 10.38 1.00  72.00
## total.sulfur.dioxide    7 1599 46.47 32.90  38.00   41.84 26.69 6.00 289.00
## density                 8 1599  1.00  0.00   1.00    1.00  0.00 0.99   1.00
## pH                      9 1599  3.31  0.15   3.31    3.31  0.15 2.74   4.01
## sulphates              10 1599  0.66  0.17   0.62    0.64  0.12 0.33   2.00
## alcohol                11 1599 10.42  1.07  10.20   10.31  1.04 8.40  14.90
## quality                12 1599  5.64  0.81   6.00    5.59  1.48 3.00   8.00
##                       range skew kurtosis   se
## fixed.acidity         11.30 0.98     1.12 0.04
## volatile.acidity       1.46 0.67     1.21 0.00
## citric.acid            1.00 0.32    -0.79 0.00
## residual.sugar        14.60 4.53    28.49 0.04
## chlorides              0.60 5.67    41.53 0.00
## free.sulfur.dioxide   71.00 1.25     2.01 0.26
## total.sulfur.dioxide 283.00 1.51     3.79 0.82
## density                0.01 0.07     0.92 0.00
## pH                     1.27 0.19     0.80 0.00
## sulphates              1.67 2.42    11.66 0.00
## alcohol                6.50 0.86     0.19 0.03
## quality                5.00 0.22     0.29 0.02
```

A continuación veremos como se clasifican los vinos en función de ciertas características específicas de las variables independientes mismas que inciden en la variable dependiente u objetivo **(quality)**.

Clasificación del vino por el contenido de grados de alcohol (alcohol)

En ocasiones es necesario realizar transformaciones u obtener subconjuntos de los datos para poder responder preguntas de nuestro interés.

Por ello, se incluye otra etiqueta, que viene siendo uan variable categórica denominada **(textura)**, que muestra el porcentaje de alcohol en el vino con el fin de agruparlos según su grado de concentración, luego entonces, tenemos 3 grupos:

Suave : Menos de 9
Medio : Entre 9 y 12
Fuerte: Más de 12

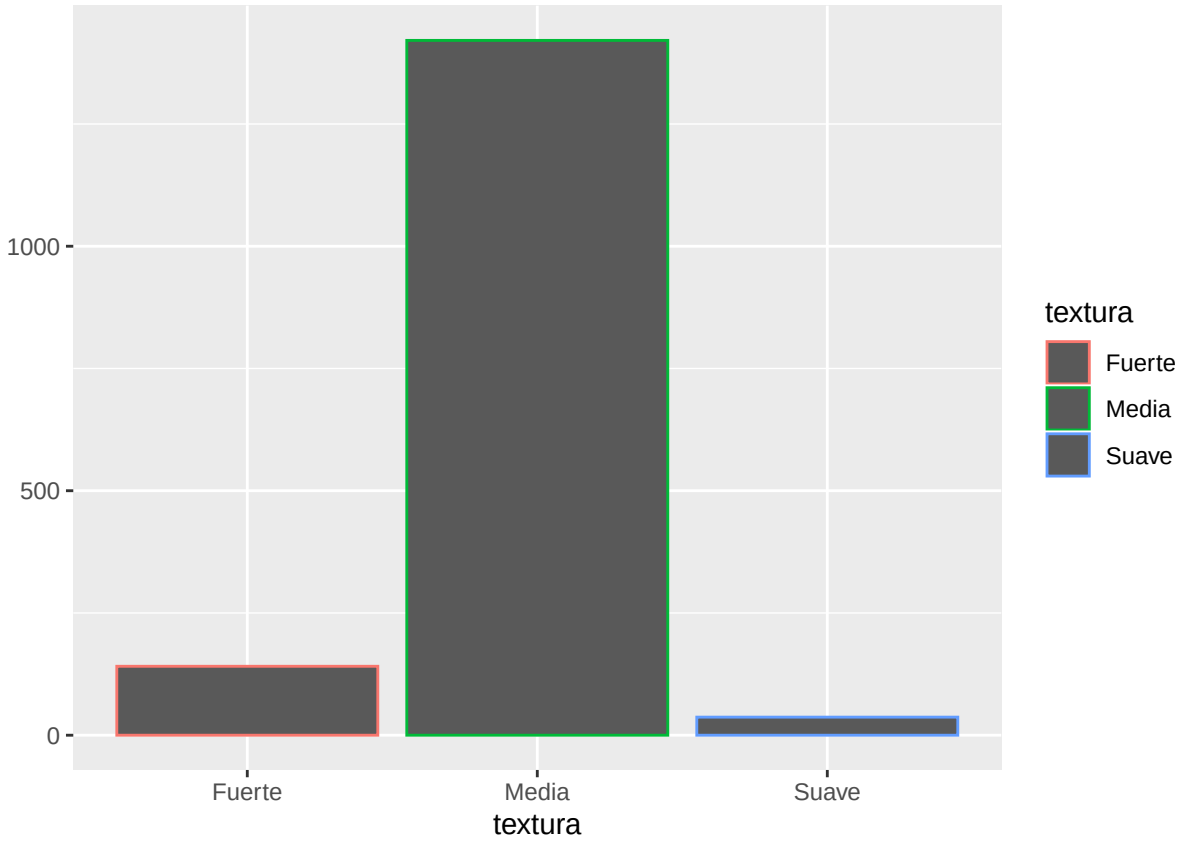
Se aprecia el número de observaciones que cumplen con cada agrupación, además de su representación gráfica.

```
datos_vinos$textura=''
datos_vinos$textura[datos_vinos$alcohol <= 9] ='Suave'
datos_vinos$textura[datos_vinos$alcohol> 9 & datos_vinos$alcohol <= 12]='Media'
datos_vinos$textura[datos_vinos$alcohol > 12]='Fuerte'
datos_vinos$textura=as.factor(datos_vinos$textura)
table(datos_vinos$textura)
```

```
##
## Fuerte  Media  Suave
##    141   1421     37
```

La representación gráfica es:

```r
qplot(x=textura, data = datos_vinos , color = textura)
```



Clasificación del vino por nivel de azúcar residual (residual.sugar)

El vino principalmente obtiene su dulzor del azúcar de la uva (glucosa y fructosa) que queda sin fermentar y que llamamos azúcares residuales. Es el dulzor que podemos detectar en la cata de vino.

Seco: el contenido de acidez total expresado en gramos de ácido tartárico por litro no sea inferior en más de 2 gramos al contenido de azúcar residual, podemos decir que es un vino seco. Es decir, cuando la acidez prevalece sobre el contenido azucaroso en esos estándares.

Semiseco: cuando el contenido de acidez total en gramos de ácido tartárico por litro no es inferior en más de 10 gramos al contenido de azúcar residual, hablamos de un vino semiseco.

Semidulce: cuando el contenido de acidez total en gramos de ácido tartárico por litro es mayor a 10 y menor a 45 gramos.

Dulce: cuando el contenido de acidez total en gramos de ácido tartárico por litro es mayor o igual a 45 gramos.

Para ello calculamos la **(acidez total)**, que es la **media de todos los ácidos contenidos en un vino o de la intensidad ácida del mismo**. En el primer caso, la acidez o acidez total se descompone en acidez fija, acidez volátil y la acidez cítrica, y se suele medir en gramos de ácido tartárico por litro.

Por ello, se incluye otra etiqueta, que viene siendo uan variable categórica denominada **(nivel_azucar)**.

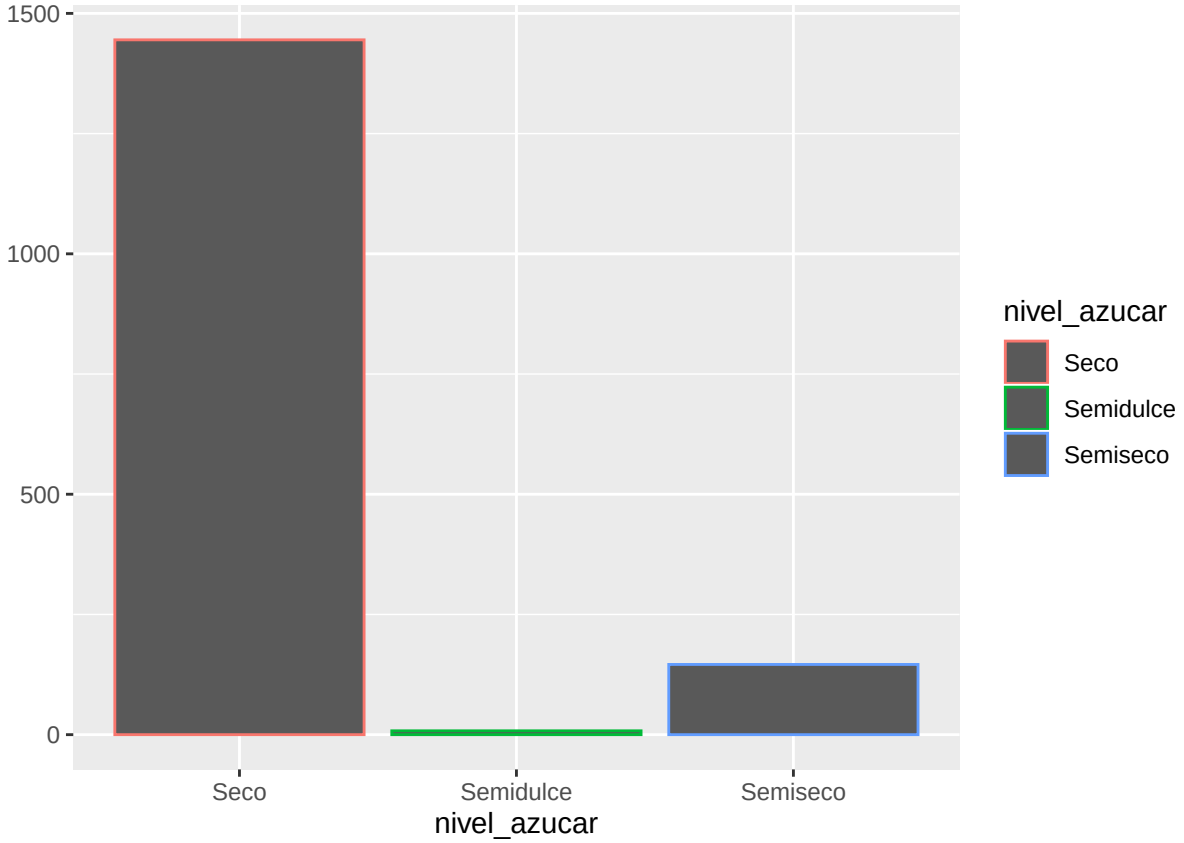
```r
datos_vinos$acidez_total = ((datos_vinos$fixed.acidity + datos_vinos$volatile.acidity + datos_vinos$citi


datos_vinos$nivel_azucar=''
datos_vinos$nivel_azucar[ abs(datos_vinos$residual.sugar - datos_vinos$acidez_total) <= 2] ='Seco'
```

```r
datos_vinos$nivel_azucar[abs(datos_vinos$residual.sugar - datos_vinos$acidez_total) > 2 & abs(datos_vin
datos_vinos$nivel_azucar[abs(datos_vinos$residual.sugar - datos_vinos$acidez_total) > 10 & abs(datos_vi
datos_vinos$nivel_azucar[abs(datos_vinos$residual.sugar - datos_vinos$acidez_total) > 45] ='Dulce'

datos_vinos$nivel_azucar=as.factor(datos_vinos$nivel_azucar)

table(datos_vinos$nivel_azucar)
```

```
##
##      Seco Semidulce  Semiseco
##      1445         8       146
```

Veamos su representación gráfica:

```r
qplot(x=nivel_azucar, data = datos_vinos , color = nivel_azucar)
```



Clasificación del vino por la cantidad de Dióxido de Azufre Libre (SO2) (free.sulfur.dioxide)

Podemos catalogarlos en dos grupos:

Vinos Blancos: El umbral de corrección se sitúa en los **25 miligramos por litro**, siendo necesaria su corrección si su contenido es inferior. En el momento del embotellado se realiza un análisis y se reajusta entre 35 y 40 miligramos por litro. Por supuesto que en zonas donde el pH de los vinos no pasa de 3.5 se reducen un poco estas dosis.

Vinos Tintos: Se suelen mantener con niveles de dióxido de azufre libre entre 25 y 35 miligramos por litro, aumentando la dosis con pH alto.

En general es conveniente mantenerlos en niveles de **30-35 miligramos de dióxido de azufre libre por litro**.

Nuestra data muestra promedios por debajo de los valores referidos, esto se debe a, que en la actualidad se tiende a reducir al máximo la adición de dióxido de azufre (sobre todo en el momento del embotellado), debido a los problemas de alergias e intolerancias y a la presión mediática.

Existen también los Vinos de **Maceración Carbónica**, los cuales pueden llegar a valores de Potencial de Hidrógeno **pH** mayores a 4.0. En nuestra data los valores están por debajo.

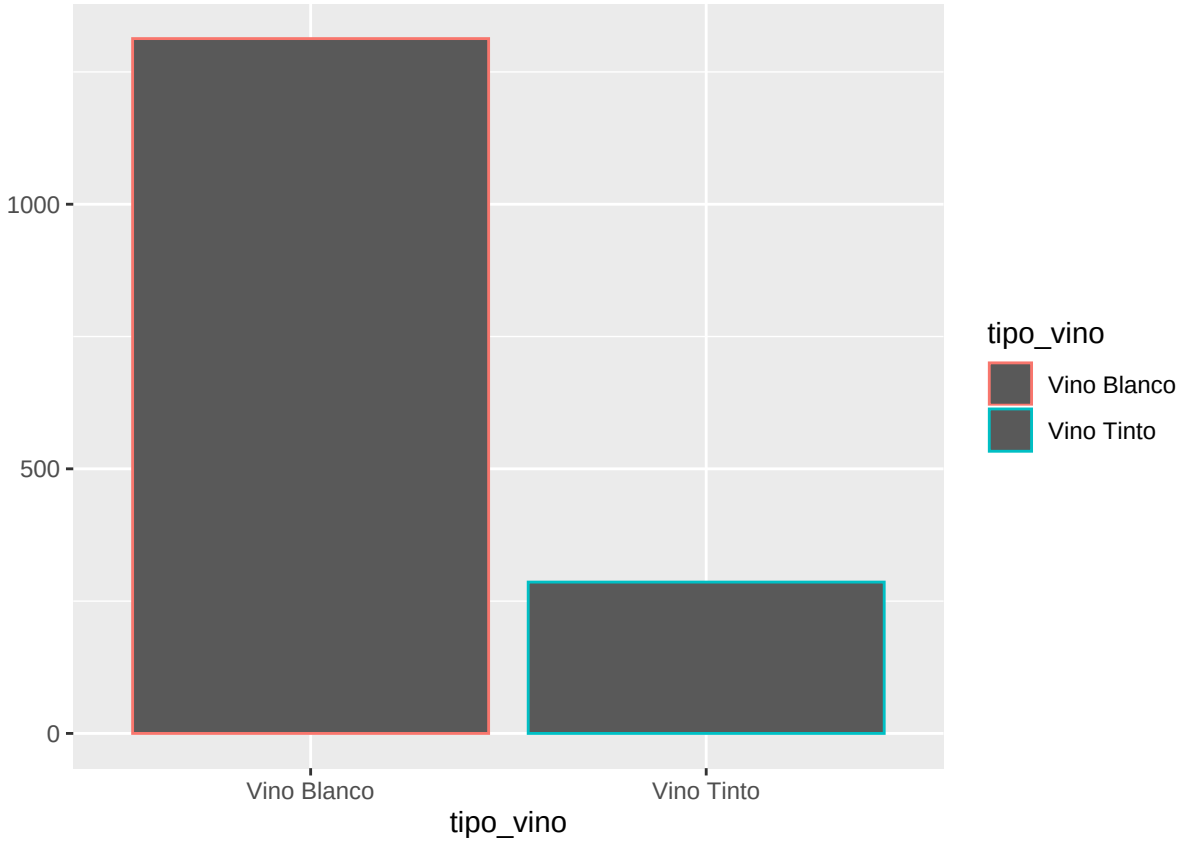
También, se incluye otra etiqueta, que viene siendo uan variable categórica denominada **(tipo_vino)**.

```
datos_vinos$tipo_vino=''
datos_vinos$tipo_vino[datos_vinos$free.sulfur.dioxide <= 25] ='Vino Blanco'
datos_vinos$tipo_vino[datos_vinos$free.sulfur.dioxide > 25]='Vino Tinto'
datos_vinos$tipo_vino=as.factor(datos_vinos$tipo_vino)
table(datos_vinos$tipo_vino)
```

```
##
## Vino Blanco  Vino Tinto
##        1313         286
```

Veamos su representación gráfica:

```
qplot(x=tipo_vino, data = datos_vinos , color = tipo_vino)
```



Gráficos todo en 1

Una manera más sencilla de obtener un gráfico de nuestros datos es con la función **ggpairs()** del paquete GGally que utiliza la librería **ggplot**. Aquí la función reconoce el tipo de variable que le ingresamos y selecciona automáticamente el gráfico adecuado. También le podemos indicar si queremos que le asigne un color en los gráficos a partir de una variable categórica que tenemos. Recordemos que el conjunto ya tiene varias variables categóricas, las cuales son: **textura, nivel_azucar y tipo_vino** que clasifica a los vinos según carcaterísticas específicas.

Cómo es lógico, en éste gráfico se pierde el sentido a medida que aumentamos más variables.

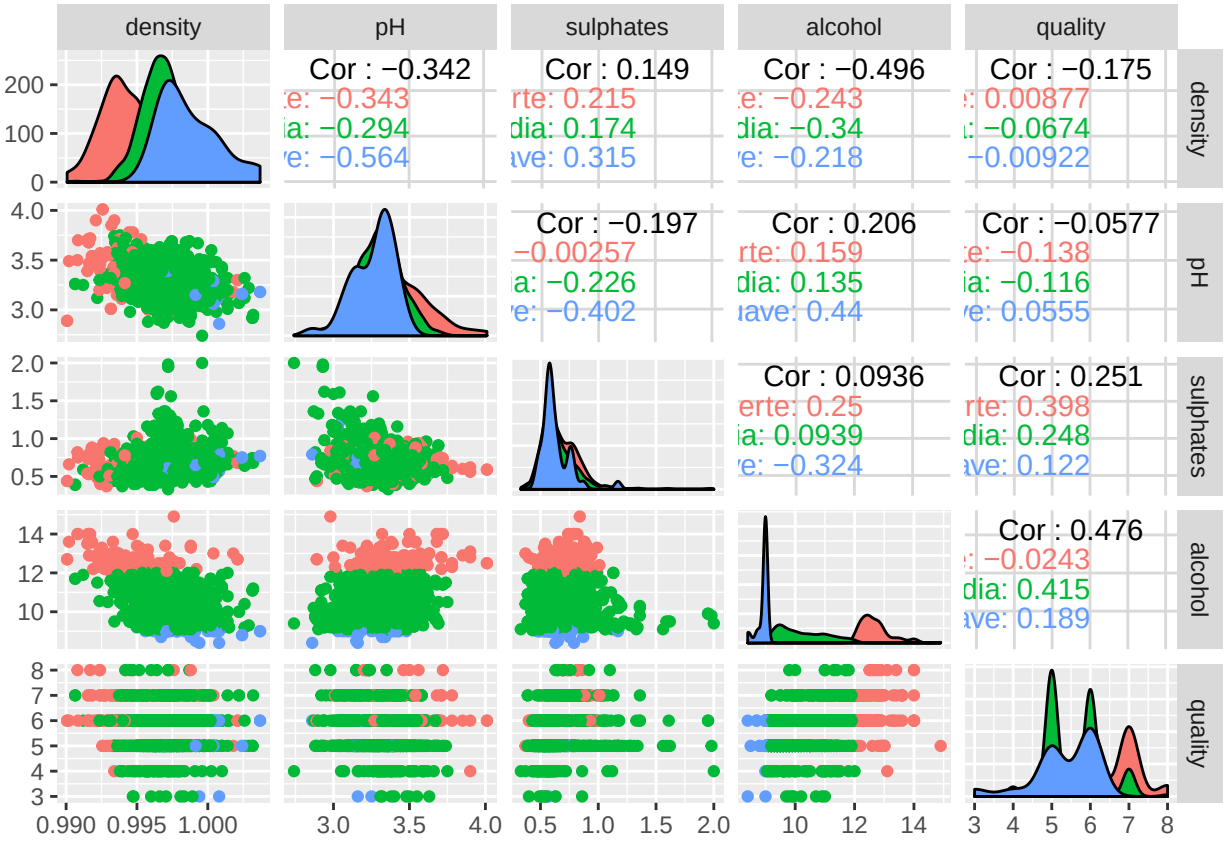
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

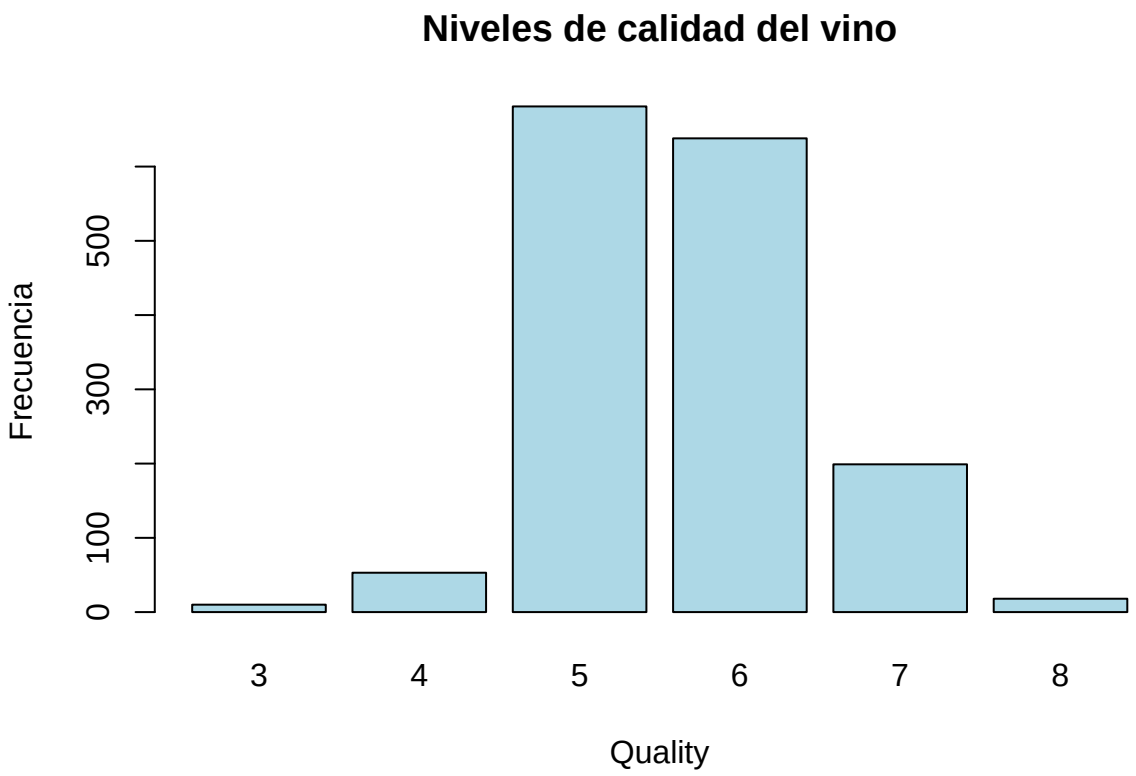
```
ggpairs(datos_vinos, columns = 8:12,aes(colour = textura))
```



Comportamiento de Frecuencias

Abordemos el comportamiento (frecuencias) de la variable objetivo **(quality)**. Se aprecia que los mayores niveles de calidad estan en índices marcados con **5 y 6**, y en menor cantidad los que tienen valores iguales a **7**.

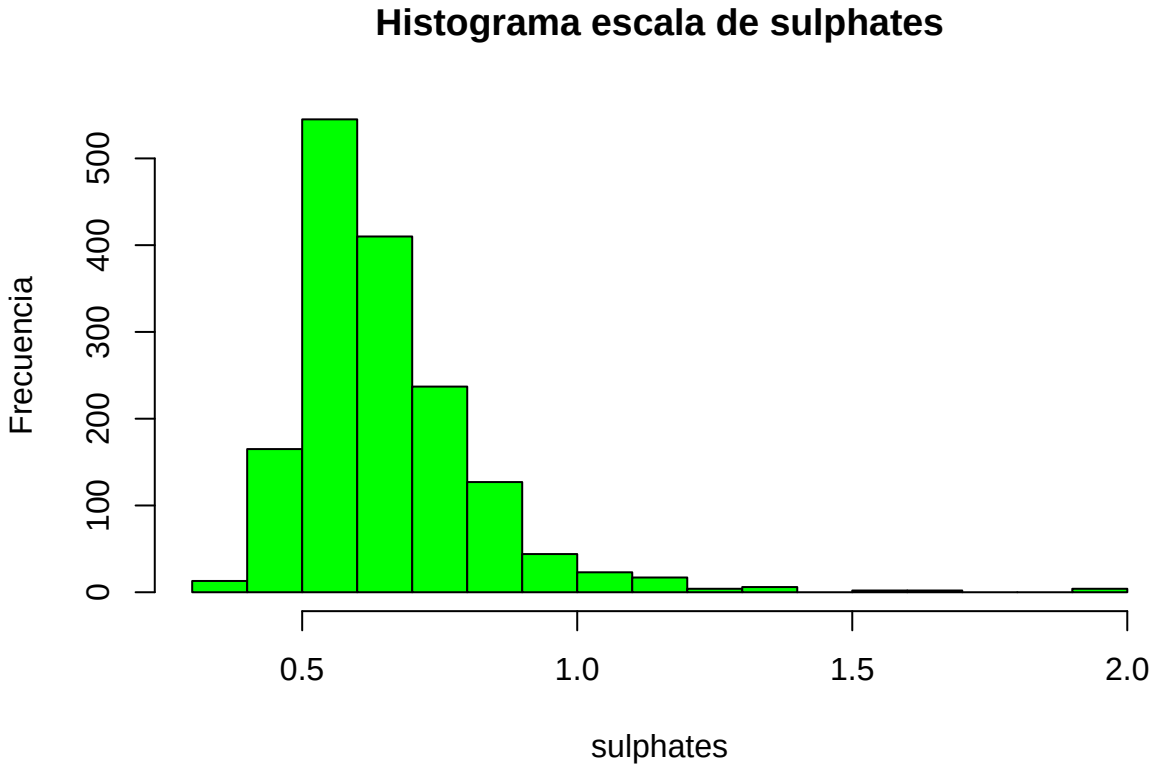
```
barplot(table(datos_vinos$quality), main="Niveles de calidad del vino", xlab="Quality",
        ylab="Frecuencia",
        col="lightblue"
        )
```



Cuando las variables son continuas, hacer una tabla de frecuencia es poco práctico y no entrega información útil. Por ello, la solución es hacer un histograma. Los histogramas grafican una tabla de intervalos, donde queremos saber cuántas observaciones tienen puntajes entre 0 y 2, por ejemplo, veamos el caso de la variable (sulphates).

Se nota qu elas barras están juntas.

```
hist(datos_vinos$sulphates,
     main="Histograma escala de sulphates",
     xlab="sulphates",
     ylab="Frecuencia",
     col="green"
     )
```



Cruzar dos tablas de frecuencia

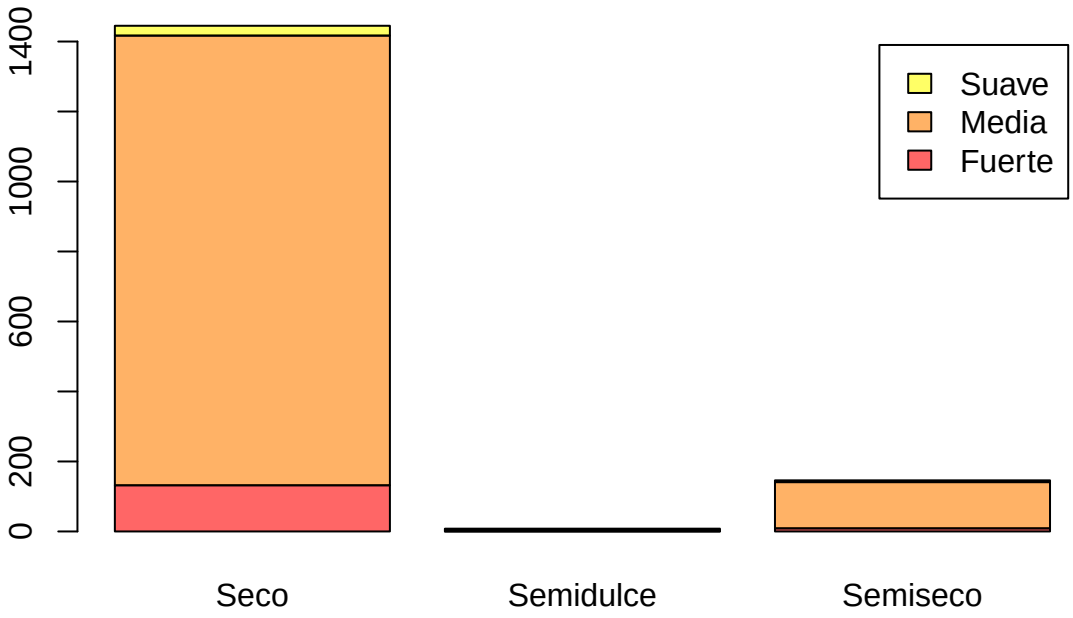
Es común querer cruzar dos tablas de frecuencia (crosstabs), lo que se hace simplemente agregando otra variable al comando **table()**.

```
table(datos_vinos$textura,datos_vinos$nivel_azucar)
```

```
##
##          Seco Semidulce Semiseco
##   Fuerte  132         0        9
##   Media  1285         4      132
##   Suave    28         4        5
```

Veremos un gráfico de barra:

```
barplot(table(datos_vinos$textura, datos_vinos$nivel_azucar), legend=TRUE, col=heat.colors(3, alpha=.6))
```



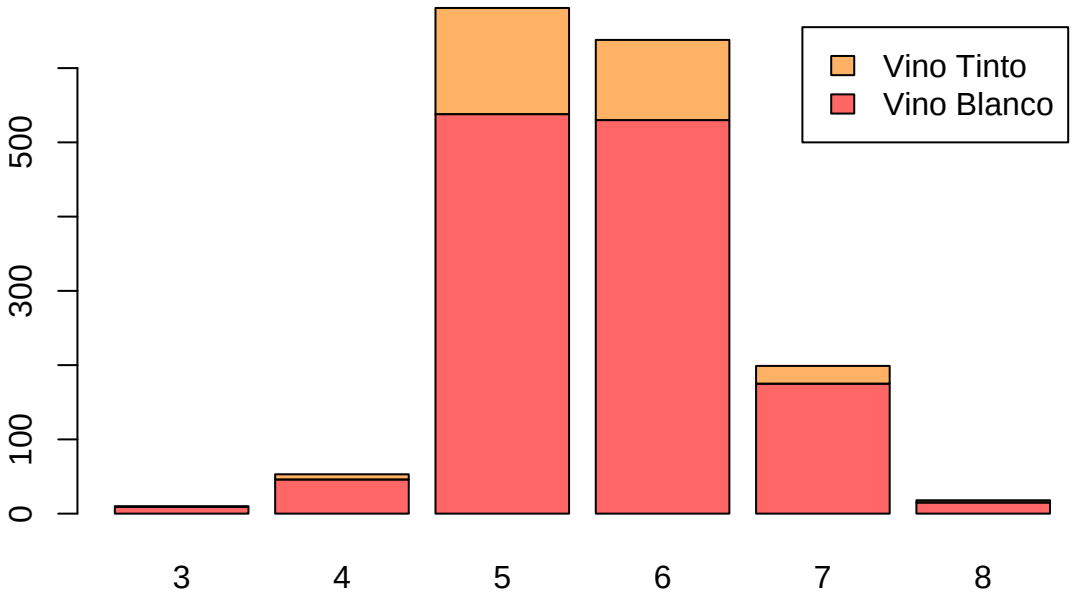
Ahora veamos esta representación para ver el tipo de vino y los niveles de calidad.

```
table(datos_vinos$tipo_vino,datos_vinos$quality)
```

```
##
##                3   4   5   6   7   8
##   Vino Blanco   9  46 538 530 175  15
##   Vino Tinto    1   7 143 108  24   3
```

El gráfico refleja que la mayoría de vinos tienen calidad con textura media, y corresponden al vino blanco.

```
barplot(table(datos_vinos$tipo_vino, datos_vinos$quality), legend=TRUE, col= heat.colors(3, alpha=.6))
```



Comparación de Grupos

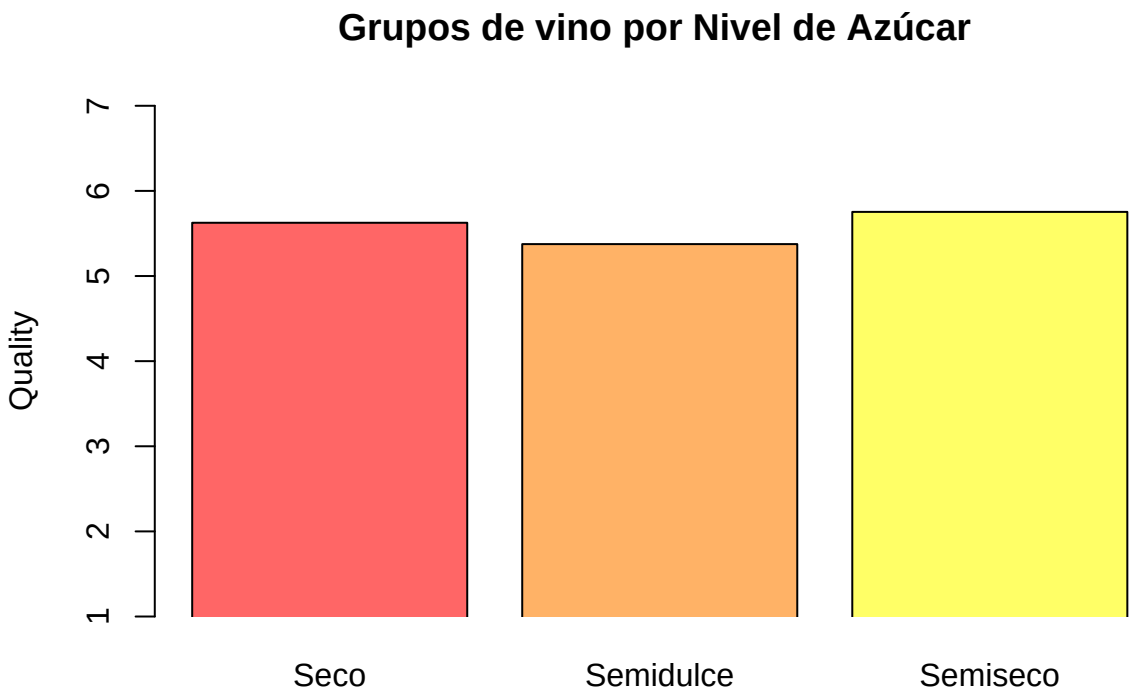
Podemos ver en la siguiente tabla una comparación de los promedios de los grupos de vinos por nivel de azucar con respecto a su calidad.

```
tapply(datos_vinos$quality, datos_vinos$nivel_azucar, mean, na.rm=TRUE)
```

```
##      Seco Semidulce  Semiseco
##  5.625606  5.375000  5.753425
```

La gráfica indica que los tres grupos guardan un equilibrio con respecto a su calidad (5 y 6).

```
barplot(tapply(datos_vinos$quality, datos_vinos$nivel_azucar, mean, na.rm=TRUE),
        main="Grupos de vino por Nivel de Azúcar ",
        ylab="Quality",
        col=heat.colors(3,alpha=.6),
        ylim=c(1,7),
        xpd=F)
```



Finalmente, la función **tapply()** también permite ocupar dos o más factores para definir los grupos. Para ello utilizamos el comando **list()* para hacer una lista con los factores a ocupar.

El comando de abajo hace una tabla con los promedios de (quality) según (tipo_vino) y (textura) grado alcoholico.

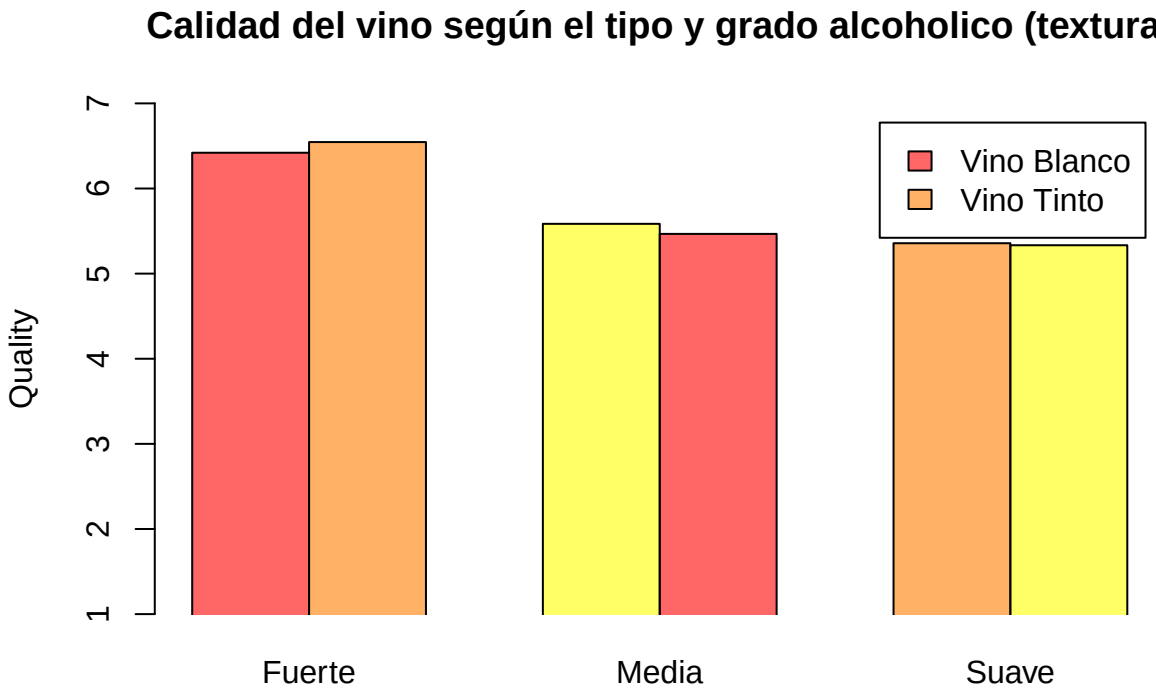
```
tapply(datos_vinos$quality, list(datos_vinos$tipo_vino, datos_vinos$textura), mean, na.rm=TRUE)
```

```
##                Fuerte    Media    Suave
## Vino Blanco 6.420168 5.584906 5.357143
## Vino Tinto  6.545455 5.466667 5.333333
```

la gráfica es la siguiente:

```
x = tapply(datos_vinos$quality, list(datos_vinos$tipo_vino, datos_vinos$textura), mean, na.rm=TRUE)

barplot(x,
        legend=TRUE,
        beside=TRUE,
        main="Calidad del vino según el tipo y grado alcoholico (textura)",
        ylab="Quality",
```

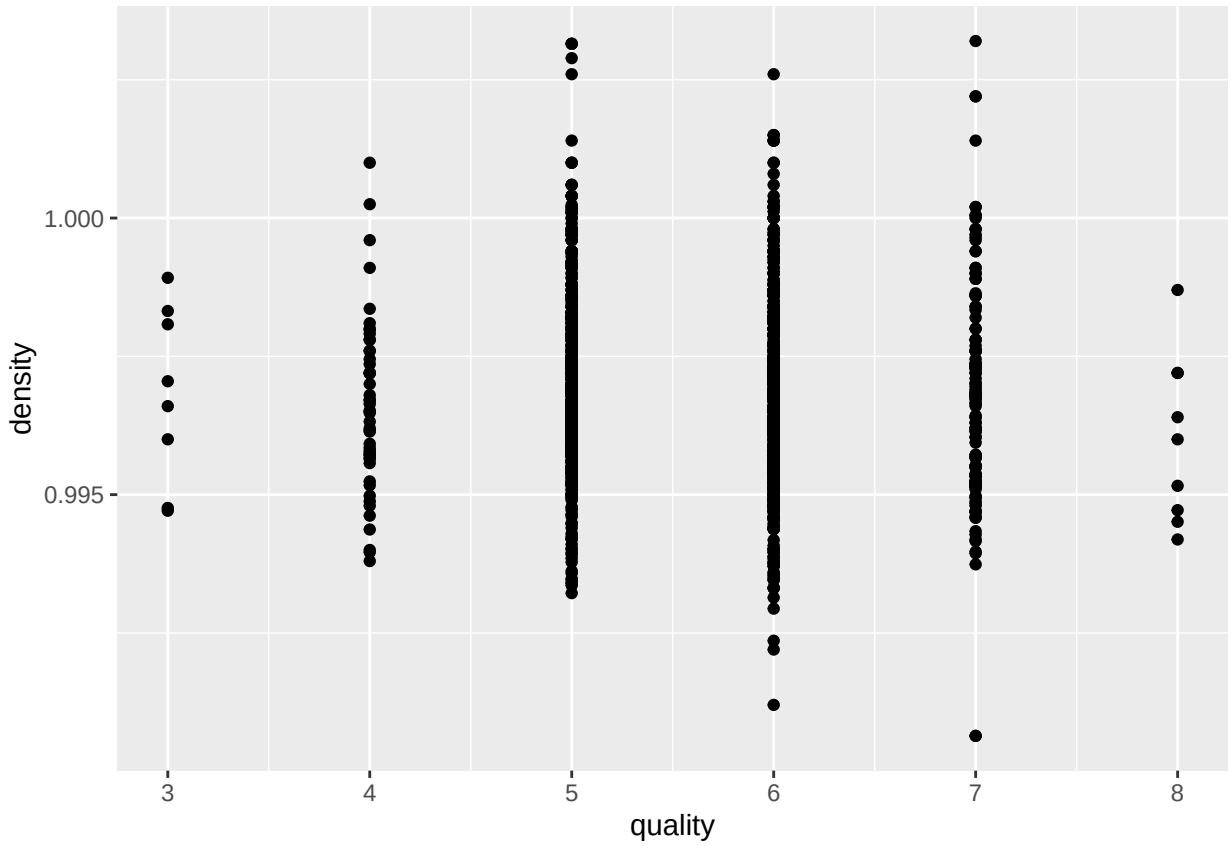
```
        col=heat.colors(3,alpha=.6),
        ylim=c(1,7),
        xpd=F)
```



Tendencia

Supongamos que queremos ver la tendencia del vino con textura "Media", para ello se genera un subconjunto del set de datos.

```
Tendencia_textura_media <- datos_vinos[datos_vinos$textura == "Media", ]

ggplot(Tendencia_textura_media, aes(x = quality, y = density)) +   geom_point()
```



El gráfico muestra que la mayoría de los vinos con textura "Media", que son 1421 que constituyen un 89% del total de observaciones.

Observando el gráfico vemos que la mayoría de los vinos se ratifican con un nivel de calidad entre 5 y 6.

Tablas utilizando Filtrados

A continuació mostramos la manera de aplicar un filtrado al conjunto de datos según los vinos con textura Media.

Podemos obtener una tabla filtrada con los datos que cumplen esta condición, es decir todos los vinos con textura "Media" y cuya densidad está entre 0.90 y 1 y además tienen una calidad comprendida entre 5 y 7.

```
filtra_vinos = filter(datos_vinos, textura == "Media", density > 0.90 & density <= 1.0, quality >= 5 & q
head(filtra_vinos)
```

```
##   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 textura acidez_total nivel_azucar   tipo_vino
## 1       5   Media     2.700000         Seco Vino Blanco
## 2       5   Media     2.893333         Seco Vino Blanco
## 3       5   Media     2.866667         Seco Vino Blanco
## 4       6   Media     4.013333     Semiseco Vino Blanco
## 5       5   Media     2.700000         Seco Vino Blanco
## 6       5   Media     2.686667         Seco Vino Blanco
```

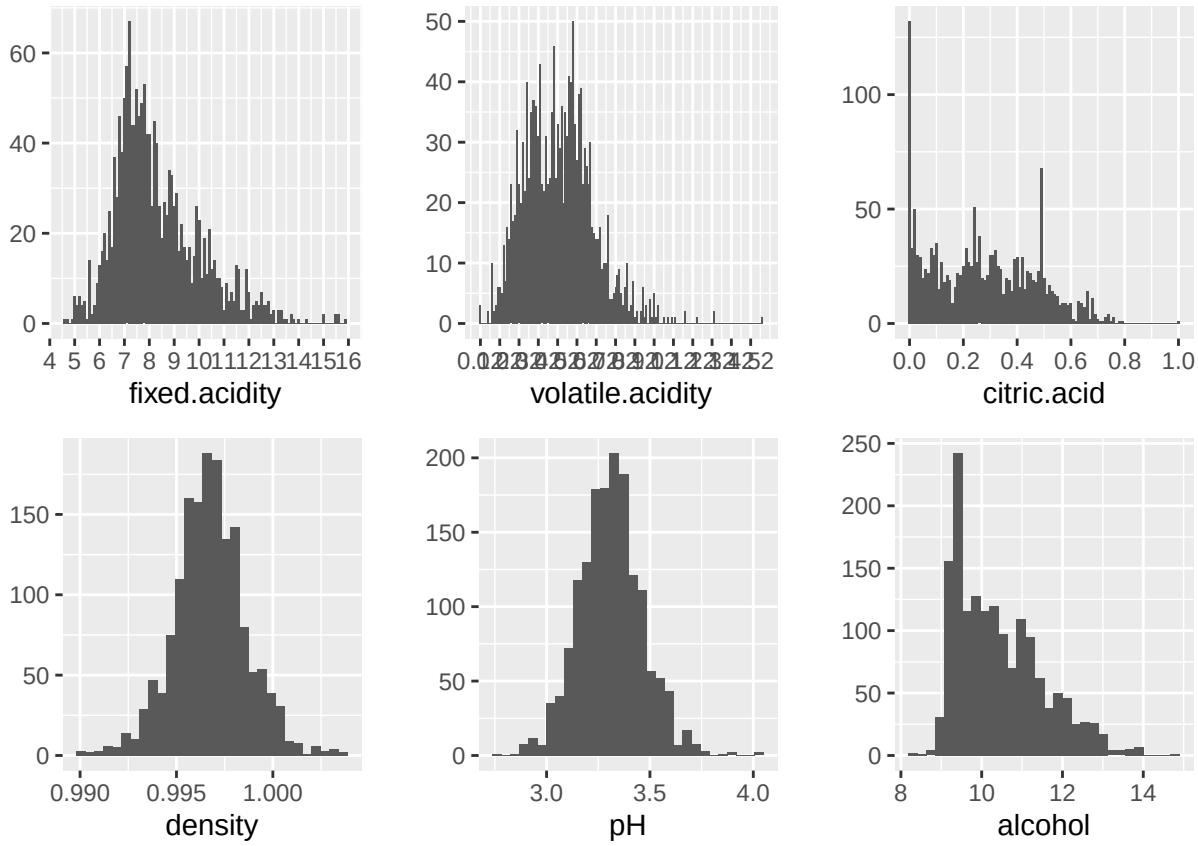
Análisis univariado y plot

Para cada variable en el conjunto de datos, trazo su histograma de frecuencia y diagrama de caja, mostrando el cambio de calidad percibida influenciado por cada característica.

Vemos la Distribución normal de las Frecuencias.

```
n1 = qplot(x = fixed.acidity, data = datos_vinos, binwidth = 0.1) +  scale_x_continuous(breaks = seq(4,

n2 = qplot(x = volatile.acidity, data = datos_vinos, binwidth = 0.01) + scale_x_continuous(breaks = seq

n3 = qplot(x = citric.acid, data = datos_vinos, binwidth = 0.01) + scale_x_continuous(breaks = seq(0, 1

n4 = qplot(x = density, data = datos_vinos)

n5 = qplot(x = pH, data = datos_vinos)

n6 = qplot(x = alcohol, data = datos_vinos)

gridExtra::grid.arrange(n1, n2, n3, n4, n5, n6, ncol = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Las siguientes variables tienen una distribución normal o cercana a la normal: **fixed.acidity, volatile.acidity, density, pH y alcohol**.

La distribución de la frecuencia de la variable **citric.acid* no es normal.
No se transforman estos datos para el propósito del análisis.

concluimos la parte de representación mostrando el conjunto de datos con todas las características (variables con un (*), corresponden a variables categóricas).

```
describe(datos_vinos)
```

```
##                       vars    n  mean    sd median trimmed   mad  min    max
## fixed.acidity            1 1599  8.32  1.74   7.90    8.15  1.48 4.60  15.90
## volatile.acidity         2 1599  0.53  0.18   0.52    0.52  0.18 0.12   1.58
## citric.acid              3 1599  0.27  0.19   0.26    0.26  0.25 0.00   1.00
## residual.sugar           4 1599  2.54  1.41   2.20    2.26  0.44 0.90  15.50
## chlorides                5 1599  0.09  0.05   0.08    0.08  0.01 0.01   0.61
## free.sulfur.dioxide      6 1599 15.87 10.46  14.00   14.58 10.38 1.00  72.00
## total.sulfur.dioxide     7 1599 46.47 32.90  38.00   41.84 26.69 6.00 289.00
## density                  8 1599  1.00  0.00   1.00    1.00  0.00 0.99   1.00
## pH                       9 1599  3.31  0.15   3.31    3.31  0.15 2.74   4.01
## sulphates               10 1599  0.66  0.17   0.62    0.64  0.12 0.33   2.00
## alcohol                 11 1599 10.42  1.07  10.20   10.31  1.04 8.40  14.90
## quality                 12 1599  5.64  0.81   6.00    5.59  1.48 3.00   8.00
## textura*                13 1599  1.93  0.33   2.00    2.00  0.00 1.00   3.00
## acidez_total            14 1599  3.04  0.61   2.91    2.98  0.51 1.76   5.68
## nivel_azucar*           15 1599  1.19  0.58   1.00    1.00  0.00 1.00   3.00
## tipo_vino*              16 1599  1.18  0.38   1.00    1.10  0.00 1.00   2.00
##                        range  skew kurtosis   se
```

```
## fixed.acidity         11.30  0.98     1.12 0.04
## volatile.acidity       1.46  0.67     1.21 0.00
## citric.acid            1.00  0.32    -0.79 0.00
## residual.sugar        14.60  4.53    28.49 0.04
## chlorides              0.60  5.67    41.53 0.00
## free.sulfur.dioxide   71.00  1.25     2.01 0.26
## total.sulfur.dioxide 283.00  1.51     3.79 0.82
## density                0.01  0.07     0.92 0.00
## pH                     1.27  0.19     0.80 0.00
## sulphates              1.67  2.42    11.66 0.00
## alcohol                6.50  0.86     0.19 0.03
## quality                5.00  0.22     0.29 0.02
## textura*               2.00 -1.25     5.46 0.01
## acidez_total           3.93  0.96     1.05 0.02
## nivel_azucar*          2.00  2.78     5.80 0.01
## tipo_vino*             1.00  1.67     0.80 0.01
```

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

La preparación de los datos es un aspecto muy importante del análisis y suele ser la fase que lleva más tiempo, consecuentemente hay que tener una serie de conceptos claros, para poder tenerlos en cuenta a la hora de interpretar los resultados analíticos obtenidos.

Así, debemos considerar los siguientes puntos, para poder abordar el problema:

-Comparar los componentes del mosto (zumo) y del vino.
-Evaluar la importancia analítica de los ácidos y realizar el cálculo de la acidez total.
-Hacer una valoración parecida con el dióxido de azufre (free.sulfur.dioxide), cuyo uso es generalizado, pero que genera numerosas intolerancias.
- Analizar la importancia del control de los azúcares (residual.sugar) y los alcoholes (alcohol), como elementos básicos de los procesos fermentativos y que son mandatorios para obtener el grado de calidad del vino (quality).
-Conocer los principios analíticos que se pueden aplicar al conjunto de datos.

El problema se traduce en analizar si todas las variable predictoras nos inducen a determinar si la variable objetivo **(quality)** tiene el mayor valor.

Según la data analizada, los valores de la calidad del vino fluctúan en un rango cerrado entre 3 y 8 (siendo 3 igual a menor calidad y 8 a mayor calidad).

Con los componentes químicos de 1599 observaciones, se ha podido determinar los niveles de calidad a los que corresponden los vinos.

Conclusiones

1.- Según los análisis, tenemos que la mayoría de los vinos observados tienen nievels de calidad media (entre 5 y 6).

2.- Se han categorizado a 1313 como “vinos blancos” y 286 como “vinos tintos”.

3.- Tenemos 1445 vinos catalogados como vino tipo “Seco”, 8 como “Semidulce y 146 como”Semiseco".

4.- Según el grado de alcohol, tenemos 1421 con textura “Media”, 37 con textura “Suave” y 141 con textura “Fuerte”.

5.- Por la densidad del vino es fuerte (141). Cuanto mayor es el porcentaje de alcohol, menor es la densidad. Es claramente visible que en nuestros datos, los vinos más fuertes tienden a tener una calificación más alta.

6.- El vino de mejor calidad es el que tiene una combinación perfecta de diferentes componentes químicos.

7.- Para realizar una análisis mas exhaustivo en el futuro, debemos recopilar más datos sobre vinos de baja calidad y de alta calidad. Así por ejemplo no existen datos de vinos con pH que puedan llegar a valores de 4.0, que dan lugar a los Vinos de Maceración Carbónica.

8.- Si el conjunto de datos tiene más registros tanto en el extremo bajo como en el extremo superior, se puede mejorar la calidad del análisis.

9.- Podemos estar más seguros acerca de la existencia de una correlación significativa entre algunos componentes químicos y la calidad del vino, así por ejempo: Los componentes de Acidez (fixed.acidity, volatile.acidity, citric.acid) y el azúcar residual (residual.sugar).

10.- El (alcohol) y el ácido cítrico son dos características que aumentan más la calidad percibida del vino. El pH y la acidez volátil, por el contrario, reducen más la calidad percibida.

11.- El alcohol y los sulfatos, junto con otras componentes químicos incrementan la calidad.

12.- Como conclusión final tenemos que en función del grado de (alcohol), el 89% de los vinos tienen calidad **Media**, 2% con calidad **Fuerte** y 9% con calidad **Suave**.

Fin de práctica 2