PRA2

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

MARÍA DEL PILAR FERNÁNDEZ FERNÁNDEZ

AULA 3

Contenido

1. Descripción del dataset. 3

1.1 ¿Por qué es importante y qué pregunta / problema pretende responder? 5

2. Integración y selección de los datos de interés a analizar. 6

3. Limpieza de los datos. 8

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 9

3.2 Identificación y tratamiento de valores extremos. 10

3.3 Exportación fichero después de la limpieza de datos 16

4. Análisis de los datos. 17

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 17

4.2 Comprobación de la normalidad y homogeneidad de la varianza 17

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo de estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes. 20

5. Representación de los resultados a partir de tablas y gráficas. 24

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 27

7. Contribuciones 28

1. Descripción del dataset.

El conjunto de datos que vamos a analizar se ha obtenido descargándolo del siguiente enlace de Kaggle (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)..

Al descargarlo obtenemos un csv separado por ‘,’ que está compuesto por 12 columnas con información relativa a la calidad del vino y contiene una fila de encabezado y 1599 filas de información.

Las columnas que contiene el csv son las siguientes:

- **Fixed acidity:** Acidez fija. Es la marcada por los ácidos que se encuentran en la uva (málico, tartárico, cítrico) y los que surgen en la fermentación (láctico, succínico). Dan al vino el bouquet y vida, pues contribuyen a preservarlo.

- **Volatile acidity:** Acidez volátil. Es debida a las sustancias ácidas volátiles, generalmente ácidos grasos ligeros de la serie acetática. La acidez volátil se encuentra entre 0,20 y 0,70 según el tipo de vino y el proceso de elaboración y en ningún caso debe ser superior a 1,5g/l sobrepasada esta cifra se considera como vinagre.

- **Citric acid:** Ácido cítrico. El ácido cítrico que se encuentra más comúnmente en el vino son los suplementos ácidos producidos comercialmente derivados de la fermentación de soluciones de sacarosa. Los enólogos pueden utilizar estos suplementos económicos en la acidificación para aumentar la acidez total del vino. Se puede agregar ácido cítrico al vino limitado a un contenido total de 1g/l contando el contenido inicial de la uva aunque lo óptimo es entre 150-300mg/l.

- **Residual sugar:** Azúcar residual. Es la cantidad total de azúcar que queda en el vino que no ha sido fermentada por las levaduras. El valor del azúcar residual en el vino puede tomar valores muy diferentes dependiendo del tipo de vino, desde menos de 5 gramos por litro en el caso de los vinos secos hasta más de 45 gramos por litro en el caso de los vinos dulces.

- **Chlorides:** Cloruros. El contenido de cloruros de un vino, fundamentalmente depende de la concentración de este lón en el suelo en el que ha crecido la vid. Su límite se encuentra en 1g/L excpresado en NaCl. El límite máximo fijado por el Instituto Nacional de Vitivinicultura es de 0,80g/l con certificado de análisis para la libre circulación y/o exportación y de 1g/l sin certificado de análisis para la libre circulación y/o exportación.

- **Free sulfur dioxide:** dióxido de azufre libre (sulfitos). Es un conservante ampliamente utilizado en la elaboración del vino y en las muchas industrias alimentarias, debido a sus propiedades antioxidantes y antibacterianas.

- **Total sulfur dioxide:** dióxido de azufre total. La cantidad de sulfitos máximos permitidos en un vino es de 200 mg/l, no pudiendo ser mayor de 20mg/l en vinos naturales y de 150 mg/l en vinos ecológicos.

- **Density:** Densidad. La densidad de los vinos está próxima a 0.994, lo cual significa que el vino contenido en una barrica de 225 litros no llega a pesar 224 Kg. Cuanto más alcohol tenga un vino menor será su densidad.

- **pH:** El pH de la mayoría de los vinos se encuentra en un intervalo de 2,8 a 4, lo que lógicamente recae en el lado ácico de la escala. Un vino con un pH de 2,8 es extremadamente ácido mientras que uno con un pH en torno a 4 es plano, carente de acidez. Los vinos tintos suelen tener un pH entre 2,8 y 3,6 para tener un poco de acidez.

- **sulphates:** Sulfatos. El límite máximo fijado por el Instituto Nacional de Vitivinicultura sigue la siguiente tabla:

	Medida (g/l) con certificado de análisis para la libre circulación y/o exportación	**Medida (g/l) sin certificado de análisis para la libre circulación y/o exportación**
Vino seco	1	1,30
Vinos edulcorados (fermentación superior a 4g/l)	1,50	1,50
Vinos que posean un contenido de azúcares reductores remanentes naturales de fermentación superior a 4g/l	1,50	1,50
Con añejamiento mínimo de 2 años en barrica	2	2,00
Provenientes de procedimientos especiales de elaboración con denuncia previa para controles oficiales pertinentes	2	2,00

De lo que podemos deducir que el valor máximo permitido en el vino es de 2 g/l.

- **alcohol:** Es uno de los elementos esenciales que podemos encontrar en el vino, junto con la acidez, los taninos, la fruta y el cuerpo. Los vinos verdes portugueses se encuentra en torno a 12-13º.

- **quality:** Calidad. La calidad del vino se basa en determinados parámetros, siendo los más importantes:

    - Densidad
    - Alcohol
    - pH
    - Acidez volátil
    - Color
    - Hierro

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

Partiendo de este conjunto de datos queremos comprobar que parámetros de los ofrecidos en el dataset son los más relevantes para poder conocer la calidad de un vino. Como hemos detallado anteriormente la calidad de un vino se basa sobre todo en la densidad, alcohol, pH, acidez volátil, color y hierro. De los dos últimos parámetros, color y hierro, no tenemos información por lo que no podremos analizar el impacto en la calidad, del resto de parámetros si tenemos información para poder comprobar cómo afectan realmente a la calidad. También podemos analizar si el resto de parámetros ofrecidos en el dataset tiene relevancia en alguno de los parámetros que si se tienen en cuenta para calcular la calidad.

Este dataset es importante porque conociendo de antemano los valores que alcanzan determinados parámetros podremos evaluar si nuestro vino tendrá mejor o peor calidad, lo cual será importante para poder conocer las futuras ventas de nuestras cosechas y nos ofrecerá información para saber cómo mejorar la calidad de las futuras.

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

En nuestro caso sólo disponemos de un único dataset en un único fichero .csv por lo que no será necesario realizar el proceso de integración.

Antes de comenzar con el proceso de selección de los datos realizaremos la lectura del fichero csv en el que se encuentran: winequality-red.csv:

```
#Lectura fichero winequality-red.csv

#Por defecto los campos vienen separados por coma (,) y los decimales
por punto(.) como en nuestro fichero.

#Indicamos con el valor TRUE en header que el fichero viene con
encabezado

datos_vino<-read.csv('C:\\PRA2\\winequality-red.csv', header=TRUE)
```

Mostramos las primeras filas del fichero leído para ver que se ha hecho correctamente:

```
head(datos_vino)
```

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

1           7.4             0.70        0.00            1.9     0.076
11                  34  0.9978 3.51      0.56     9.4       5

2           7.8             0.88        0.00            2.6     0.098
25                  67  0.9968 3.20      0.68     9.8       5

3           7.8             0.76        0.04            2.3     0.092
15                  54  0.9970 3.26      0.65     9.8       5

4          11.2             0.28        0.56            1.9     0.075
17                  60  0.9980 3.16      0.58     9.8       6

5           7.4             0.70        0.00            1.9     0.076
11                  34  0.9978 3.51      0.56     9.4       5

6           7.4             0.66        0.00            1.8     0.075
13                  40  0.9978 3.51      0.56     9.4       5
```

Una vez que hemos comprobado que los valores que se han leído corresponden con lo esperado vamos a comprobar el tipo de dato que se ha asignado a cada campo para constatar que sea correcto.

```
sapply(datos_vino, function(x) class(x))
```

Con esta función comprobamos que todos los valores obtenidos son de tipo numérico excepto el último (quality) que es de tipo integer. Vemos que todos los valores se han leído con su tipo correcto.

```
> sapply(datos_vino, function(x) class(x))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide total.sulfur.dioxide
           "numeric"           "numeric"             "numeric"           "numeric"            "numeric"             "numeric"             "numeric"
             density                  pH             sulphates             alcohol              quality
           "numeric"           "numeric"             "numeric"           "numeric"            "integer"
> |
```

Una vez que tenemos el data frame datos_vino con los datos leídos del csv procedemos a realizar la selección de datos de interés a analizar.

SELECCIÓN DE LOS DATOS

Analizando el dataset hemos visto que todos los parámetros de nuestro dataset son de tipo numérico o integer por lo que son susceptibles de estudio. Si nos encontráramos con algún tipo de variable categórica o cualitativa sería el momento de eliminarla.

3. Limpieza de los datos.

Para realizar la limpieza de los datos, vamos a comprobar si tenemos algún registro duplicado para eliminarlo antes de realizar cualquier otro proceso, esto nos permitirá reducir la muestra. Realizando un screening del fichero vemos que existen filas duplicadas.

Vamos a comprobar el número de éstas:

```
#Comprobamos si tenemos registros duplicados

datos_duplicados<-datos_vino[duplicated(datos_vino),]

#Contamos las filas duplicadas

nrow(datos_duplicados)
```

El resultado obtenido es el siguiente:

```
[1] 240
```

Podemos comprobar que tenemos 240 filas duplicadas que no van a aportarnos nada de información y que por tanto podremos eliminar del dataset. Procedemos al borrado:

```
#Contamos las filas que tenemos en el dataframe antes del borrado

nrow(datos_vino)

#Procedemos al borrado de las filas duplicadas

datos_vino<-datos_vino[!duplicated(datos_vino),]

#Contabilizamos las filas que nos quedan en el dataframe

nrow(datos_vino)
```

El resultado de la ejecución del código mostrado es el siguiente:

```
> #Contamos las filas que tenemos en el dataframe antes del borrado

> nrow(datos_vino)

[1] 1599

> #Procedemos al borrado de las filas duplicadas

> datos_vino<-datos_vino[!duplicated(datos_vino),]

> #Contabilizamos las filas que nos quedan en el dataframe

> nrow(datos_vino)

[1] 1359
```

Donde podemos ver que antes de eliminar los duplicados el dataset constaba de 1599 filas y después de eliminarlos está compuesto por 1359 filas, la diferencia entre ambos valores se corresponde con los 240 valores duplicados que habíamos detectado previamente.

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

En primer lugar vamos a comprobar si los datos contienen elementos vacíos para analizar cómo tratarlos:

```
#Vamos a comprobar si existen elementos vacíos

sapply(datos_vino, function(x) sum(is.na(x)))
```

```
> sapply(datos_vino, 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
```

Podemos ver que la función nos devuelve 0 para todos los parámetros existentes en el dataset por lo que no tenemos ningún elemento vacío.

En caso de encontrarnos elementos vacíos deberíamos rellenarlos para reducir los problemas que nos pueden acarrear. Existen determinadas técnicas para rellenarlos y evitar encontrarnos con estos casos. Algunas de dichas técnicas son las siguientes:

- Completar manualmente los registros. No aplicaría en el caso del dataset con el que estamos trabajando ya que no conocemos la información con la que rellenarlos.
- Podríamos reemplazar los valores perdidos por una misma constante. Esto sería más útil cuando estos valores tienen un significado común como “No procede”. En nuestro caso no aplicaría ya que no se trata de valores no aplicables, si no de datos perdidos.
- Reemplazar los valores perdidos por una misma medida de tendencia central como la medida o la mediana. En el caso de nuestro dataset no sería lo más recomendable ya que podría alterar los resultados y encontrar incoherencias al calcular la calidad.
- Implementación de métodos probabilistas como regresiones, inferencias basadas en modelos bayesianos o los árboles de decisión. Este caso sería el más apropiado para nuestro dataset teniendo en cuenta de utilizar los métodos adecuados para no introducir errores y no falsear los resultados.

Vamos a comprobar si hay algún parámetro con valor 0 en el conjunto de datos:

```
#Vamos a comprobar si existen valores a cero

sapply(datos_vino, function(x) sum(x == 0))
```

```
> sapply(datos_vino, function(x) sum(x == 0))
       fixed.acidity     volatile.acidity          citric.acid
                   0                    0                  118
      residual.sugar            chlorides  free.sulfur.dioxide
                   0                    0                    0
total.sulfur.dioxide              density                   pH
                   0                    0                    0
           sulphates              alcohol              quality
                   0                    0                    0
> |
```

Como podemos ver en el resultado la única variable que toma valor 0 es citric.acid, (ácido cítrico), que tiene 118 casos con valor 0. Habría que analizar si sería posible que existan valores 0 de ácido cítrico o si pueden ser valores perdidos que se han rellenado con 0. El ácido cítrico se puede encontrar en mayor o menor medida en los vinos de manera natural o añadido artificialmente por lo que cabría la posibilidad real de tomar valor 0. Además si comprobamos el resto de valores de la variable vemos que el valor 0.1 lo encontramos 29 veces y el valor 0.2 lo encontramos 21 veces por lo que parece perfectamente plausible poder encontrar valores a 0 y asumiremos que son valores correctos.

3.2 Identificación y tratamiento de valores extremos.

Los valores extremos, o outliers, son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población. Se desvían tanto del resto que levantan sospechas sobre si son valores correctos o posibles errores. Estos valores hay que encontrarlos y analizarlos ya que podrían afectar de forma adversa al resultado de los análisis de datos.

Vamos a comprobar la existencia de outliers con la función boxplot() que nos mostrarás los valores atípicos para cada una de las variables. Vamos a ir analizando cada una de las variables y como proceder para los valores atípicos encontrados en cada una de ellas:

FIXED.ACIDITY

```
#Outliers de fixed.acidity

boxplot(datos_vino$fixed.acidity)$out
```

```
[1] 12.8 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 14.0
13.7 12.7

[16] 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9 14.3 12.4
15.5 15.6

[31] 13.0 12.7 12.4 12.7 13.2 13.2 15.9 13.3 12.9 12.6 12.6
```

El valor de los ácidos fijos que se encuentran en los vinos puede ser muy variable, en el conjunto de los outliers que se nos muestra los valores parecen muy homogéneos y observando el conjunto total de valores vemos que no parecen outliers si no valores perfectamente válidos por lo que vamos a considerar los outliers como valores correctos.

VOLATILE.ACID

```
#Outliers de volatile.acid

boxplot(datos_vino$volatile.acid)$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
```

El valor máximo que se permite de ácidos volátiles en un vino no debe superar los 1,5 g/l, como vemos ninguno de los outliers obtenidos supera dicha cifra por lo que damos la información como buena.

CITRIC.ACID

```
#Outliers de citric.acid

boxplot(datos_vino$citric.acid)$out

[1] 1
```

Vemos que únicamente tenemos un outlier con valor 1, como hemos indicado al definir los parámetros el valor del ácido cítrico es óptimo que sea menor pero se puede llegar a añadir hasta el valor de 1g/l por lo que este valor se considerará como válido.

RESIDUAL.SUGAR

```
#Outliers de residual.sugar

boxplot(datos_vino$residual.sugar)$out

[1]  6.10  3.80  3.90  4.40 10.70  5.50  5.90  3.80  5.10  4.65  5.50
5.50

 [13]  7.30  7.20  3.80  5.60  4.00  4.00  4.00  7.00  6.40  5.60
11.00  4.50

 [25]  4.80  5.80  3.80  4.40  6.20  4.20  7.90  3.70  4.50  6.70
6.60  3.70

 [37]  5.20 15.50  4.10  8.30  6.55  4.60  6.10  4.30  5.80  5.15
6.30  4.20
```

```
 [49]  4.60  4.20  4.30  7.90  4.60  5.10  5.60  6.00  8.60  7.50
4.40  4.25

 [61]  6.00  3.90  4.20  4.00  4.00  6.60  6.00  3.80  9.00  4.60
8.80  5.00

 [73]  3.80  4.10  5.90  4.10  6.20  8.90  4.00  3.90  8.10  6.40
8.30  8.30

 [85]  4.70  5.50  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80
4.10 12.90

 [97]  4.30 13.40  4.80  6.30  4.50  4.30  3.90  3.80  5.40  3.80
6.10  3.90

[109]  5.10  3.90 15.40  4.80  5.20  5.20  3.75 13.80  5.70  4.30
4.10  4.10

[121]  4.40  3.70  6.70 13.90  5.10  7.80
```

El azúcar residual en el vino puede tomar valores muy dispersos dependiendo del tipo de vino del que estemos hablando. En el caso que nos aplica vemos que toma valores entre 0,9 y 15,5 gramos por litro, valores totalmente válidos para este parámetro por lo que daremos por buenos estos valores.

CHLORIDES

```
#Outliers de chlorides

boxplot(datos_vino$chlorides)$out

[1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.178 0.146
0.236

[13] 0.610 0.360 0.270 0.337 0.263 0.611 0.358 0.343 0.186 0.213 0.214
0.128

[25] 0.159 0.124 0.174 0.127 0.413 0.152 0.152 0.125 0.200 0.171 0.226
0.250

[37] 0.148 0.124 0.143 0.222 0.157 0.422 0.034 0.387 0.415 0.157 0.157
0.243

[49] 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.194 0.132
0.161

[61] 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136 0.132 0.123
0.123

[73] 0.403 0.137 0.414 0.166 0.168 0.415 0.153 0.267 0.123 0.214 0.169
0.205

[85] 0.235 0.230 0.038
```

Los valores de cloruros están permitidos hasta un máximo de 1g/l por lo que los valores que obtenemos como outlier no son tales ya que no alcanzan ni superan la cantidad de 1.

FREE.SULFUR.DIOXIDE

```
#Outliers de free.sulfur.dioxide

boxplot(datos_vino$free.sulfur.dioxide)$out

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

[26] 66
```

Analizando los valores obtenidos como outliers vemos son valores entre los cuales no hay ninguno que destaque sobremanera y que parece ser que se encuentran dentro de un rango válido por lo que aceptaremos estos valores como válidos.

TOTAL.SULFUR.DIOXIDE

```
#Outliers de total.sulfur.dioxide

boxplot(datos_vino$total.sulfur.dioxide)$out

[1] 145 148 136 125 140 133 153 134 141 129 128 143 144 127 126 145
144 135 165

[20] 134 129 151 133 142 149 147 145 148 155 151 152 125 127 139 143
144 130 278

[39] 289 135 160 141 133 147 131
```

Los valores máximos de sulfitos permitidos en el vino, como se ha indicado anteriormente varían entre los permitidos para vinos naturales que no pueden ser mayores de 20mg /l y los permitidos como máximo en vinos convencionales que no pueden superar los 200 mg/l. Revisando los outliers devueltos vemos que hay dos valores que exceden las cantidades máximas y que se encuentran muy alejados del resto de valores que son 278 y 289.

Revisando las filas en las que encontramos los valores 278 y 289 vemos que son exactamente iguales a excepción del este valor. Eliminaremos una de las dos filas y a continuación trataremos la fila que tiene el outlier.

En este caso se ha optado por no tratarlo si no directamente eliminarlo, se ha barajado la posibilidad de sustituirlo por la mediana pero esta operación desvirtuaría el cálculo de la calidad en función de los parámetros ya que este valor no sería correcto y por esta razón eliminamos la fila del outlier completamente:

```
#Eliminamos las dos filas que tiene outlieres: 289 y 278

datos_vino<-datos_vino[!(datos_vino$total.sulfur.dioxide == 289),]

datos_vino<-datos_vino[!(datos_vino$total.sulfur.dioxide == 278),]
```

Comprobamos el resultado para ver que se han eliminado los dos valores:

```
#Comprobamos que se han eliminado las filas de los outliers

boxplot(datos_vino$total.sulfur.dioxide)$out

[1] 145 148 136 125 140 133 153 134 141 129 128 143 144 127 126 145
144 135 165

[20] 134 129 151 133 142 149 147 145 148 155 151 152 125 127 139 143
144 130 135

[39] 160 141 133 147 131
```

DENSITY

```
#Outliers de density

boxplot(datos_vino$density)$out

[1] 0.99160 1.00140 1.00150 1.00180 0.99120 1.00220 1.00140 1.00140
1.00140

[10] 1.00320 1.00260 1.00140 1.00315 1.00315 1.00210 0.99170 0.99220
1.00260

[19] 0.99210 0.99154 0.99064 1.00289 0.99162 0.99007 0.99020 0.99220
0.99150

[28] 0.99157 0.99080 0.99084 0.99191 1.00369 1.00242 0.99182 0.99182
```

Los valores de la densidad del vino están normalmente próximos a 0,994 comparando con los outliers que hemos obtenido vemos que podrían considerarse perfectamente válidos y por tanto los vamos a considerar como tales.

pH

```
#Outliers de pH

boxplot(datos_vino$pH)$out

[1] 3.90 3.75 3.85 2.74 3.69 2.88 2.86 3.74 2.92 2.92 3.72 2.87 2.89
2.92 3.90

[16] 3.71 3.69 3.71 2.89 3.78 3.70 3.78 4.01 2.90 4.01 3.71 2.88 3.72
```

Los outliers que observamos para el pH tienen valores perfectamente válidos, lo óptimo para un vino es que tenga valores entre 2,8 y 3,6 pero los valores que se muestran como outliers vemos que están muy cercanos a estos valores y por tanto no podemos considerarlos como valores extremos si no como valores válidos.

SULPHATES

```
#Outliers de sulphates

boxplot(datos_vino$sulphates)$out

[1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.98 1.31 2.00 1.08
1.59 1.02

[16] 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.07 1.06
1.06 1.05

[31] 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
1.07 1.34

[46] 1.16 1.10 1.15 1.17 1.33 1.18 1.17 1.03 1.10 1.01
```

Como hemos indicado anteriormente los valores de sulfatos máximos son 2 g/l por lo que los valores que se muestran como outliers vemos que se encuentran dentro de los rangos válidos para esta variable.

ALCOHOL

```
#Outliers de alcohol

boxplot(datos_vino$alcohol)$out

[1]  14.00000  14.00000  14.00000  14.90000  14.00000  13.60000  13.60000
13.60000

[9] 14.00000 14.00000 13.56667 13.60000
```

El alcohol del vino verde se encuentra en torno a 12-13 grados por lo que los valores arrojados como outliers vemos que son valores perfectamente válidos.

QUALITY

```
#Outliers de quality

boxplot(datos_vino$quality)$out

[1] 8 8 8 8 8 3 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8
```

En el caso de la calidad vemos que los valores atípicos toman valores de 3 y 8. Realmente estos valores no pueden considerarse atípicos ya que la calidad puede variar entre 0 y 10 por lo que son valores totalmente válidos.

3.3 Exportación fichero después de la limpieza de datos

Una vez terminada la limpieza de datos grabamos el fichero nuevamente:

```
write.csv(datos_vino, 'C:\\PRA2\\winequality-red_clean.csv')
```

4. Análisis de los datos.

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

A continuación seleccionamos los grupos de datos que son interesantes para analizar / comparar. En nuestro caso utilizaremos los valores que más influyen a la hora de calcular la calidad de un vino, además de la propia calidad misma. Los valores seleccionados para analizar / comparar son:

- Densidad
- Alcohol
- pH
- acidez volátil
- calidad

Esta es la selección de datos con la que vamos a trabajar para realizar los análisis siguientes:

- ¿a mayor acidez volátil mayor calidad del vino?
- ¿está la densidad relacionada con la calidad?
- ¿cuáles son las variables que más influyen en la calidad del vino?

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

NORMALIDAD

Para verificar la normalidad vamos a utilizar el test de Shapiro-Wilk, el cual se considera uno de los métodos más potentes para contratar la normalidad. Asumiendo como hipótesis nula que la población está distribuida normalmente, si el p-valor es menor al nivel de significancia, generalmente $\alpha=0.05$, entonces la hipótesis nula es rechazada y se concluye que los datos no cuentan con una distribución normal. Si, por el contrario p-valor es mayor que α, se concluye que no se puede rechazar dicha hipótesis y se asume que los datos siguen una distribución normal.

```
#Calculamos la normalidad
alpha=0.05
col.names =colnames(datos_vino)
for (i in 1:ncol(datos_vino))
{
      if(i==1) cat ("Variables que no siguen una distribución normal:\n")
      if(is.integer(datos_vino[,i]) | is.numeric(datos_vino[,i]))
      {
            p_valor = shapiro.test(datos_vino[,i])$p.value
            if (p_valor<alpha)
            {
                  cat(col.names[i])
                  #Formato de salida
                  if (i <= ncol(datos_vino) - 1) cat(", \n")
            }
      }
}
```

El resultado que hemos obtenido es el siguiente:

```
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,
alcohol,
quality>
```

HOMOGENEIDAD DE LA VARIANZA

Seguidamente vamos a estudiar la homocedasticidad (homogeneidad de la varianza).

Cuando los datos no cumplen la condición de normalidad, como es nuestro caso, aplicaremos el test de Fligner-Killeen. La hipótesis nula asume igualdad de varianzas en los diferentes grupos e datos, por lo que p-valores inferiores al nivel de significancia indicarán heterocedasticidad, esto es que la varianza de los errores no es igual en todas las observaciones realizadas.

```
#Cálculo de la homogeneidad de la varianza

#Recuperamos los valores que queremos ver la relación

volatile.acid<-datos_vino$volatile.acid

density<-datos_vino$density

pH<-datos_vino$pH
```

```
alcohol<-datos_vino$alcohol

quality<-datos_vino$volatile.acid


#Test de Fligner-Killen

#volatile-acid y quality

print(fligner.test(volatile.acid ~ quality, data = datos_vino))

#density y quality

print(fligner.test(density ~ quality, data = datos_vino))

#pH y quality

print(fligner.test(pH ~ quality, data = datos_vino))

#alcohol y quality

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

El resultado que obtenemos al comparar la varianza entre los diferentes parámetros y la calidad es el siguiente:

```
> #Test de Fligner-Killen
> #volatile-acid y quality
> print(fligner.test(volatile.acid ~ quality, data = datos_vino))

        Fligner-Killeen test of homogeneity of variances

data:  volatile.acid by quality
Fligner-Killeen:med chi-squared = 30.336, df = 5, p-value = 1.266e-05

> #density y quality
> print(fligner.test(density ~ quality, data = datos_vino))

        Fligner-Killeen test of homogeneity of variances

data:  density by quality
Fligner-Killeen:med chi-squared = 38.34, df = 5, p-value = 3.224e-07

> #pH y quality
> print(fligner.test(pH ~ quality, data = datos_vino))

        Fligner-Killeen test of homogeneity of variances

data:  pH by quality
Fligner-Killeen:med chi-squared = 1.8735, df = 5, p-value = 0.8664

> #alcohol y quality
> print(fligner.test(alcohol ~ quality, data = datos_vino))

        Fligner-Killeen test of homogeneity of variances

data:  alcohol by quality
Fligner-Killeen:med chi-squared = 107.7, df = 5, p-value < 2.2e-16
```

Únicamente obtenemos un p-valor>0,05 en el caso del ph y la calidad por lo que aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas. Para el resto de casos al tener valores menores de 0,05 rechazamos la hipótesis nula de homocedasticidad y se concluye que las variables presentan varianzas estadísticamente diferentes para los diferentes grupos.

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

CORRELACIÓN

Lo primero que queremos responder es cuales de las variables que disponemos tienen mayor peso en el cálculo de la calidad del vino. Como anteriormente hemos visto que no cumplen con la normalidad el test más adecuado en este caso será el test de Spearman.

```
#Correlación con el test de Spearman entre cada variable y la calidad
matriz_correlacion<-matrix(nc=2, nr=0)
colnames(matriz_correlacion)<-c("estimate", "p-value")
#Eliminamos la calidad ya que haremos el test con ella
for (i in 1:(ncol(datos_vino) -1))
{
      if(is.integer(datos_vino[,i]) | is.numeric(datos_vino[,i]))
      {
            test_Spearman =  cor.test(datos_vino[,i],
                                      datos_vino$quality,
                                      method= "spearman",
                                      exact = FALSE)

            coeficiente_correlacion = test_Spearman$estimate
            p_valor = test_Spearman$p.value

            #Añadir fila a la matriz a mostrar
            pair = matrix(ncol=2, nrow=1)
            pair[1][1] = coeficiente_correlacion
            pair[2][1] = p_valor
            matriz_correlacion<-rbind(matriz_correlacion, pair)
            rownames(matriz_correlacion)[nrow(matriz_correlacion)]<-colnames(datos_vino)[i]
      }
}
#Imprimimos la matriz con los resultados obtenidos
print(matriz_correlacion)
```

El resultado obtenido es la siguiente matriz:

```
> print(matriz_correlacion)
                       estimate      p-value
fixed.acidity         0.11187186 3.623674e-05
volatile.acidity     -0.38530525 2.898306e-49
citric.acid           0.21629078 7.895694e-16
residual.sugar        0.02181203 4.220585e-01
chlorides            -0.20128996 7.159047e-14
free.sulfur.dioxide  -0.06268736 2.092206e-02
total.sulfur.dioxide -0.20144519 6.844797e-14
density              -0.18104700 1.832706e-11
pH                   -0.03898340 1.512095e-01
sulphates             0.38483506 3.870494e-49
alcohol               0.48604707 2.242926e-81
>
```

El coeficiente de Spearman puede tomar un valor entre +1 y -1 donde: +1 significa una perfecta asociación de rango, 0 que no hay asociación de rango y -1 significa una perfecta asociación negativa entre los rangos. Aplicando esta información a los valores obtenidos vemos que cuanto más cercano esté a 1 más correlacionadas estarán las variables. En el caso anterior vemos que la variable que más influye en la calidad del vino es el alcohol, seguida de los sulfatos. Según este conjunto de datos vemos que la que menos influencia tiene en la calidad del vino sería la acidez volátil.

CONTRASTE DE HIPÓTESIS:

El contrate de hipótesis permite la comparación entre dos grupos. En nuestro caso deberemos aplicar pruebas como Wilcoxon o Mann-Whitney ya que en nuestras variables no se cumple la normalidad y en la mayoría de los casos tampoco la homocedasticidad. Esta alternativa generalmente implicará la pérdida de potencia estadística por lo que, aunque en el caso de estudio no vayamos a realizar, sería recomendable realizar previamente una transformación de Box-Cox para tratar de mejorar la normalidad y homocedasticidad.

Queremos comprobar si la densidad influye en la calidad del vino, para ello tenemos en cuenta sólo aquellos vinos que tengan calidad entre 5 y 8:

```
#Contraste de hipótesis
wilcox.test(density ~ quality, data=datos_vino, subset = quality %in% c(5,8))
```

El resultado que obtenemos es el siguiente:

```
> #Contraste de hipótesis
> wilcox.test(density ~ quality, data=datos_vino, subset = quality %in% c(5,8))

        Wilcoxon rank sum test with continuity correction

data:  density by quality
W = 7256, p-value = 0.0007478
alternative hypothesis: true location shift is not equal to 0
```

Observando el resultado vemos que se observan diferencias estadísticas entre la densidad y la calidad cuando nos encontramos en valores de calidad superiores a 5.

REGRESIÓN

La regresión lineal es un modelo matemático que tiene como objetivo aproximar la relación de dependencia entre variable dependiente y una variable o una serie de variables independientes. En el caso de estudio que estamos tratando vamos a aplicar un modelo de regresión lineal para intentar predecir la calidad del vino en función de los valores que tomen el resto de parámetros.

Compararemos varios modelos de regresión para ver cuáles son las variables que más afectan a la calidad del vino.

Modelo A

En este modelo utilizaremos las variables que inicialmente leímos (apartado 1), que son las que más influyen en la calidad del vino. Estas variables son: densidad, acidez_volátil, pH y alcohol.

```
modelo_A <- lm(calidad ~ acidez_volatil + alcohol + pH + densidad, data = datos_vino)
```

Modelo B

Vamos a realizar el modelo de regresión simple con las variables más correlacionadas que hemos visto anteriormente aplicando la correlación. Estas variables son: acidez_fija, acido_citrico, azucar_residual, sulfatos y alcohol.

```
modelo_B <- lm(calidad ~ acidez_fija + acido_citrico + azucar_residual + sulfatos + alcohol, data= datos_vino)
```

Modelo_C

En el modelo C vamos a comprobar cómo afectan a la calidad del vino aquellas variables menos correlacionadas: acidez_volatil, cloruros, sulfitos_libres, sulfitos_totales, densidad y pH.

```
modelo_C <- lm(calidad ~ acidez_volatil + cloruros + sulfitos_libres + sulfitos_totales + densidad + pH,
               data = datos_vino)
```

Los resultados obtenidos para los tres modelos son los siguientes:

```
> print(summary(modelo_A)$r.squared)
[1] 0.3277009
> print(summary(modelo_B)$r.squared)
[1] 0.2886324
> print(summary(modelo_C)$r.squared)
[1] 0.2187698
>
```

De entre los tres modelos elegidos vemos que el modelo que mejor predice la calidad del vino es el que se detalló en el apartado 1. Hay que tener en cuenta que entre la variables que se detallaban se encontraba también el hierro y el color de las cuáles no disponemos información pero que suponemos que si se incorporan mejorarán todavía más el coeficiente de determinación.

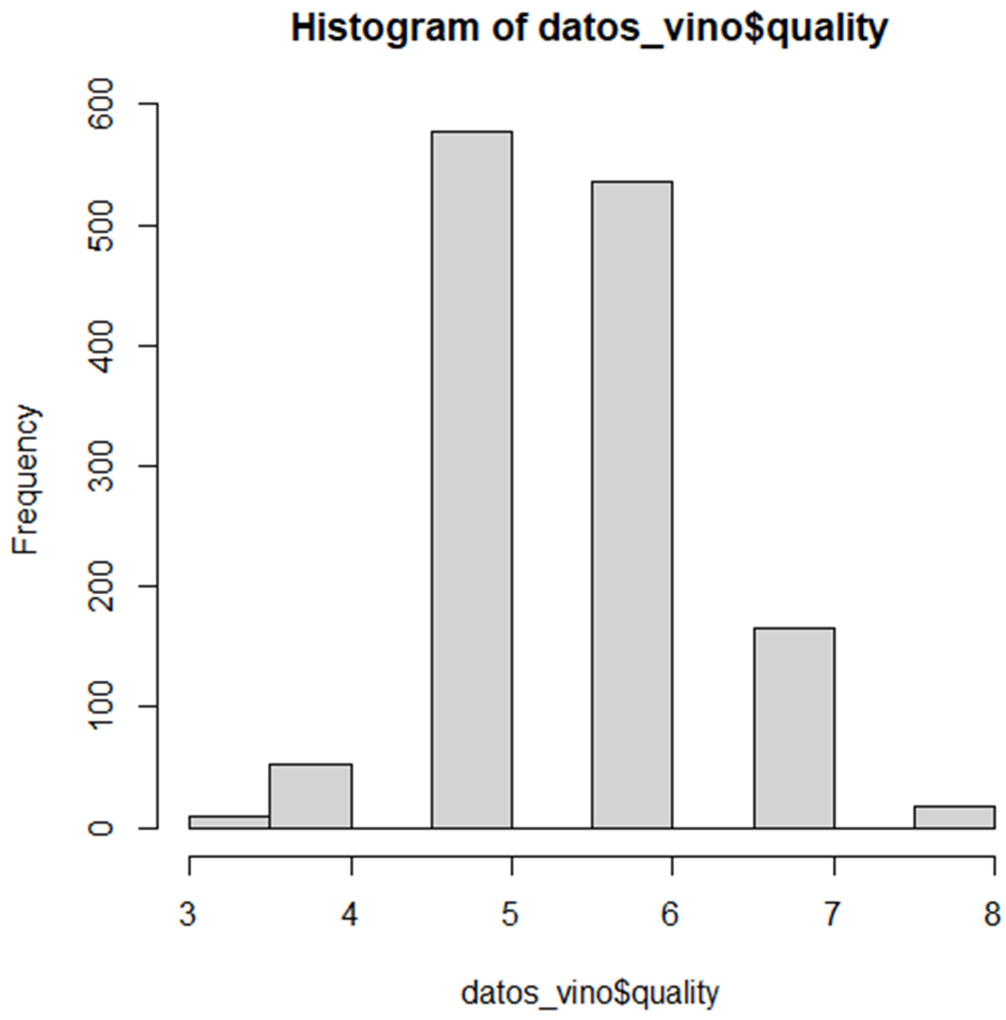
5. Representación de los resultados a partir de tablas y gráficas.

En este apartado vamos a representar mediante gráficas y tablas algunos de los pasos realizados anteriormente, para ayudar a realizar el análisis de los datos de una manera más visual.

Antes hemos analizado la normalidad con el test de Shapiro-Wilk, con la que hemos visto que ninguna de las variables seguía una distribución normal, ahora vamos a representar con histogramas los valores de alguna variable para confirmarlo visualmente.

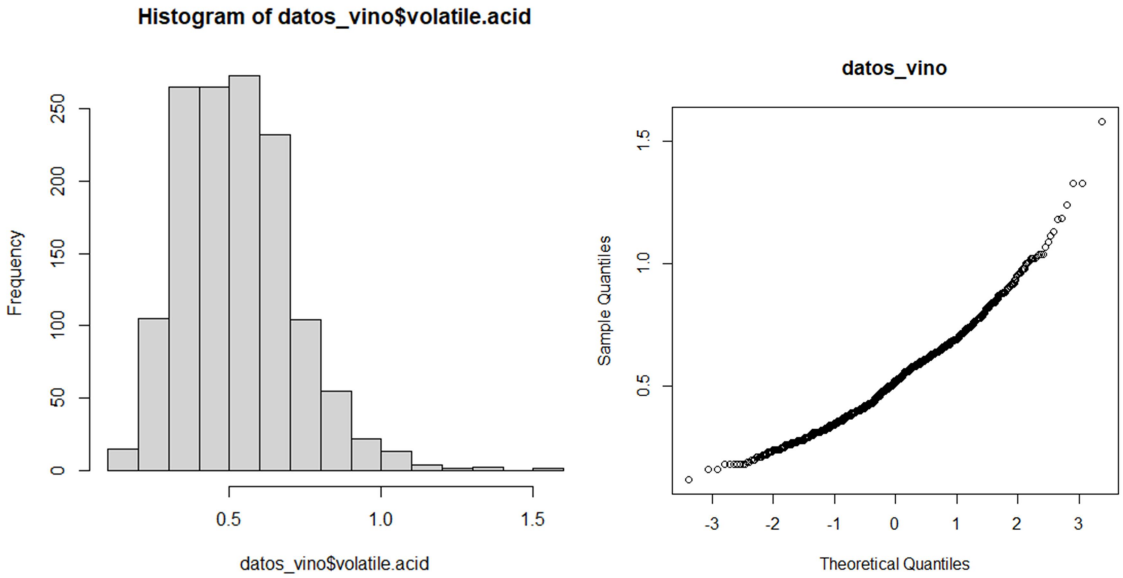
Aquí vemos que la calidad no sigue una distribución normal:

```
#Normalidad mediantes histogramas
hist(datos_vino$quality)
```

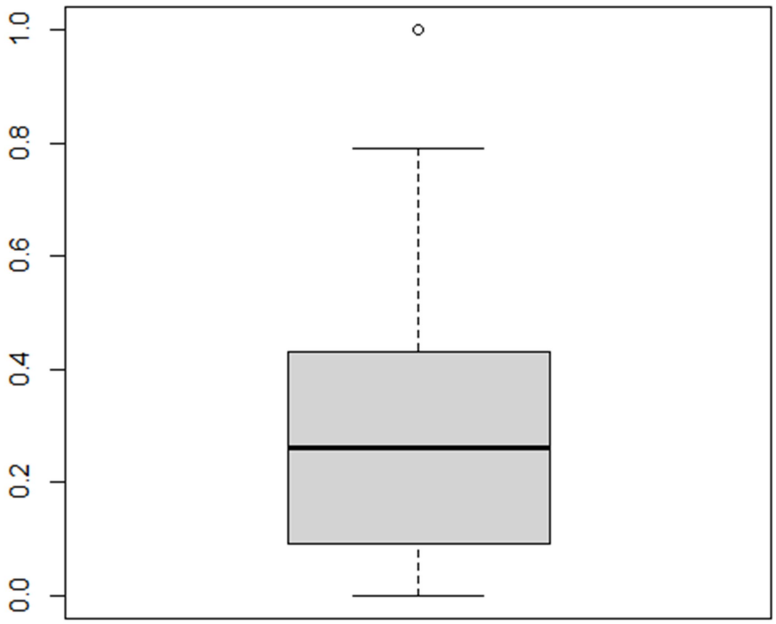


Comprobando los ácidos volátiles, comprobamos que tampoco cumple la normalidad:

```
hist(datos_vino$volatile.acid)
qqnorm(datos_vino$volatile.acid, main="datos_vino")
```



Con la representación de box-plot podremos ver los outliers gráficamente, vamos a mostrar como ejemplo el resultado de citric.acid donde sólo se nos devolvió un outlier con valor 1 que confirmamos viendo la gráfica:



En cuanto a la correlación podemos mostrar la siguiente tabla con los valores obtenidos:

```
> print(matriz_correlacion)
                      estimate      p-value
fixed.acidity        0.11187186 3.623674e-05
volatile.acidity    -0.38530525 2.898306e-49
citric.acid          0.21629078 7.895694e-16
residual.sugar       0.02181203 4.220585e-01
chlorides           -0.20128996 7.159047e-14
free.sulfur.dioxide -0.06268736 2.092206e-02
total.sulfur.dioxide -0.20144519 6.844797e-14
density             -0.18104700 1.832706e-11
pH                  -0.03898340 1.512095e-01
sulphates            0.38483506 3.870494e-49
alcohol              0.48604707 2.242926e-81
>
```

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

Después de analizar el conjunto de datos disponible y haber realizado sobre él una limpieza de datos previa, comprobando que no hubiera registros duplicados, valores vacíos, valores a cero no válidos y de comprobar si existen outliers que no tuvieran una explicación plausible hemos procedido a realizar un análisis de la información sobre el nuevo conjunto de datos ya limpio.

Primeramente hemos realizado un análisis de la normalidad y de la homocedasticidada para poder valorar que pruebas estadísiticas encajaban mejor con la distribución de datos de la que disponemos. Posteriormente hemos realizado un análisis de correlación para analizar que variables influyen más en la calidad del vino, que es el objetivo que perseguíamos analizando este dataset. Después de tener esta información hemos realizado un análisis de regresión sobre varios modelos para poder saber en base a qué variables podemos predecir mejor la calidad del vino. Para poder realizar todos estos pasos nos hemos ayudado de diferentes diagramas y tablas que nos han permitido analizar mejor los datos.

Después de todos los análisis realizados hemos confirmado la hipótesis inicial que nos indicaba que los valores que mejor predecían la calidad del vino, de entre los disponibles en este dataset son los ácidos volátiles, el alcohol, el pH y la densidad.

7. Contribuciones

Contribuciones	Firma
Investigación previa	María del Pilar Fernández Fernández
Redacción de las respuestas	María del Pilar Fernández Fernández
Desarrollo código	María del Pilar Fernández Fernández