Pra2 TCVD

Javier Díaz Domínguez

6/6/2022

#Práctica 2: Limpieza y validación de los datos #### Javier Díaz Domínguez #### 6 de junio de 20223

Índice

Detalles de la actividad

Descripción

La práctica 3 consiste en la realización de un caso práctico de análisis y tratamiento de un conjunto de datos, con el objetivo principal de identificar los datos relevantes y el tratamiento necesario para llevar a cabo un proyecto analítico.

Objetivos

Los objetivos concretos de esta práctica son:

- 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.

Competencias

Así, las competencias del Máster en Data Science que se desarrollan son:

- 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.

Resolución

Carga del dataset

Se procede con la carga del conjunto de datos

```
data <- read.csv("songs_normalize.csv", header = TRUE)
```

Descripción del dataset

El dataset empleado para la elaboración de esta prática ha sido extraído de la página web Kaggle, y puede ser obtenido en el siguiente enlance (https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019?resource=download).

```
dim(data)
```

```
## [1] 2000   18
```

Se trata de un conjunto de datos de 2000 filas por 18 columnas, que recoge información sobre los *hits* de Spotify desde el año 2000 hasta el año 2019. El título original es “Top Hits Spotify from 2000-2019”. Está compuesto por los siguientes campos:

- artist: Nombre del artista.
- song: Nombre de la canción
- duration_ms: Duración de la canción en milisegundos
- explicit: Si la canción contiene palabras explícitas o no.
- year: Año de lanzamiento de la canción.
- popularity: Popularidad de la canción (Higher better).
- danceability: Cómo de “bailable” es la canción en base a una serie de parámetros (intensidad, estabilidad del ritmo, fuerza del beat…)
- energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- key: Tono en el que está la canción
- loudness: Volumen medio de la canción en decibelios.
- mode: Mayor 1 o menor 0
- speechiness: La medida en la que aparecen palabras en la canción. Rango entre 0 y 1.
- acousticness: Medida en la que la canción es acústica (1 acústica, 0 no acústica)
- instrumentalness: Medida en la que la canción contiene sonidos vocales. 0 más vocal, 1 más instrumental.
- liveness: La medida en la que se detecta que la canción fue grabada en directo
- valence: Medida en la que se considera la cancion positiva, vívida.
- tempo: El ritmo medio estimado de la canción (en BPM).
- genre: Género de la canción.

Importancia y objetivos de los análisis

Las conclusiones que se pueden sacar en base a estos datos son realmente interesantes. Se pueden llegar a conclusiones sobre cuáles son las características de una canción que más gustan o son más propensas a gustar, así como elaborar un modelo predictivo para estimar la popularidad de una canción que se pretende lanzar.

Limpieza de los datos

```
summary(data)
```

```
##     artist              song            duration_ms       explicit
##  Length:2000        Length:2000        Min.   :113000   Length:2000
##  Class :character   Class :character   1st Qu.:203580   Class :character
##  Mode  :character   Mode  :character   Median :223280   Mode  :character
##                                        Mean   :228748
##                                        3rd Qu.:248133
##                                        Max.   :484146
##       year        popularity     danceability        energy
##  Min.   :1998   Min.   : 0.00   Min.   :0.1290   Min.   :0.0549
##  1st Qu.:2004   1st Qu.:56.00   1st Qu.:0.5810   1st Qu.:0.6220
##  Median :2010   Median :65.50   Median :0.6760   Median :0.7360
##  Mean   :2009   Mean   :59.87   Mean   :0.6674   Mean   :0.7204
##  3rd Qu.:2015   3rd Qu.:73.00   3rd Qu.:0.7640   3rd Qu.:0.8390
##  Max.   :2020   Max.   :89.00   Max.   :0.9750   Max.   :0.9990
##       key            loudness            mode         speechiness
##  Min.   : 0.000   Min.   :-20.514   Min.   :0.0000   Min.   :0.02320
##  1st Qu.: 2.000   1st Qu.: -6.490   1st Qu.:0.0000   1st Qu.:0.03960
##  Median : 6.000   Median : -5.285   Median :1.0000   Median :0.05985
##  Mean   : 5.378   Mean   : -5.512   Mean   :0.5535   Mean   :0.10357
##  3rd Qu.: 8.000   3rd Qu.: -4.168   3rd Qu.:1.0000   3rd Qu.:0.12900
##  Max.   :11.000   Max.   : -0.276   Max.   :1.0000   Max.   :0.57600
##   acousticness       instrumentalness       liveness         valence
##  Min.   :0.0000192   Min.   :0.0000000   Min.   :0.0215   Min.   :0.0381
##  1st Qu.:0.0140000   1st Qu.:0.0000000   1st Qu.:0.0881   1st Qu.:0.3867
##  Median :0.0557000   Median :0.0000000   Median :0.1240   Median :0.5575
##  Mean   :0.1289549   Mean   :0.0152260   Mean   :0.1812   Mean   :0.5517
##  3rd Qu.:0.1762500   3rd Qu.:0.0000683   3rd Qu.:0.2410   3rd Qu.:0.7300
##  Max.   :0.9760000   Max.   :0.9850000   Max.   :0.8530   Max.   :0.9730
##      tempo           genre
##  Min.   : 60.02   Length:2000
##  1st Qu.: 98.99   Class :character
##  Median :120.02   Mode  :character
##  Mean   :120.12
##  3rd Qu.:134.27
##  Max.   :210.85
```

De la ejecución de summary, podemos ver cómo los tipos de todas las columnas han sido identificados correctamente. Además, no existen valores nulos.

La columna explicit debería de tener clase boolean: se procede con su conversión.

```
data$explicit = as.logical(data$explicit)
```

Selección de datos de interés

Todos los campos de la tablas son relevantes de cara a un análisis completo del tema en cuestión, y pueden ser útiles para la elaboración de un modelo predictivo en el que la variable dependiente sea la popularidad y las demás sean variables independientes.

Ceros y elementos vacíos

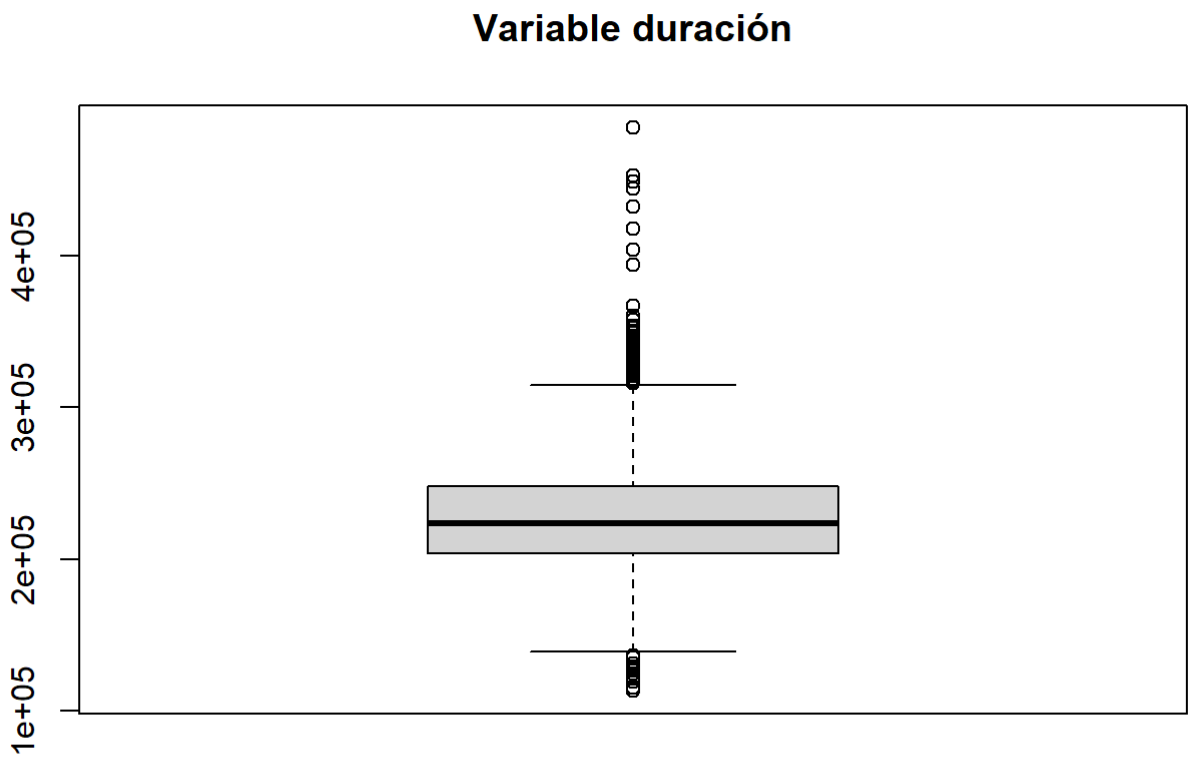
```
colSums(is.na(data))
```

```
##           artist             song      duration_ms         explicit
##                0                0                0                0
##             year       popularity     danceability           energy
##                0                0                0                0
##              key         loudness             mode      speechiness
##                0                0                0                0
##     acousticness instrumentalness         liveness          valence
##                0                0                0                0
##            tempo            genre
##                0                0
```

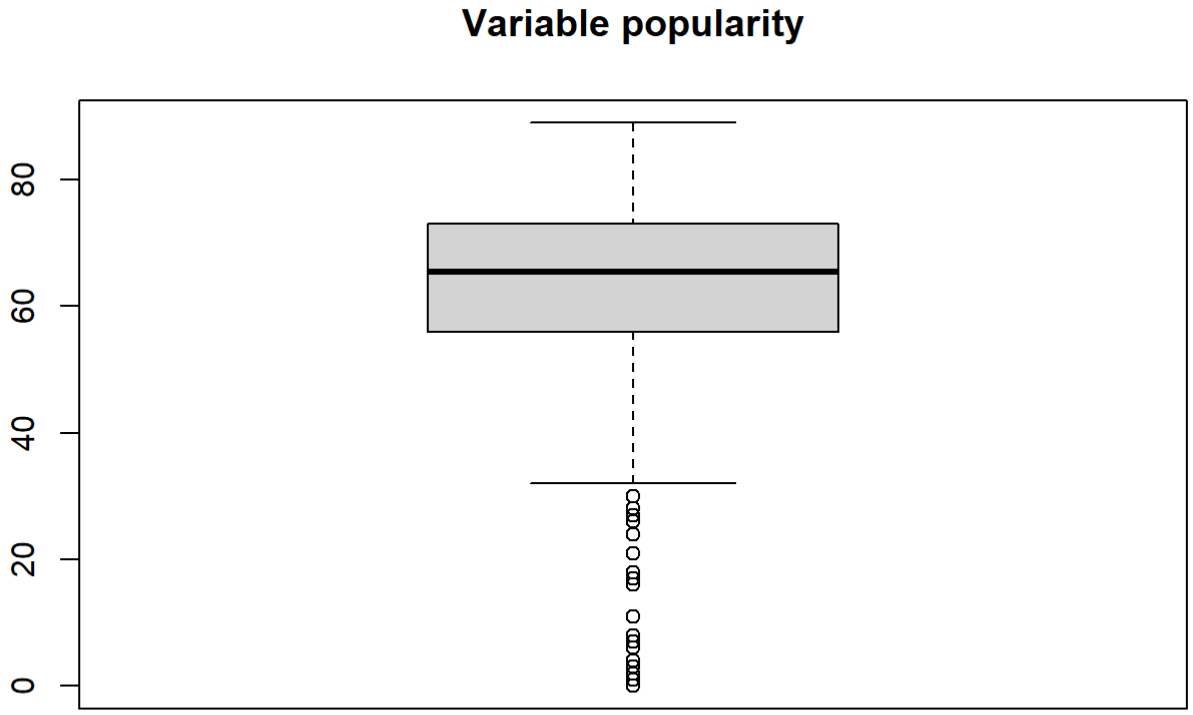
Se comprueba la inexistencia de valores nulos.

Valores extremos

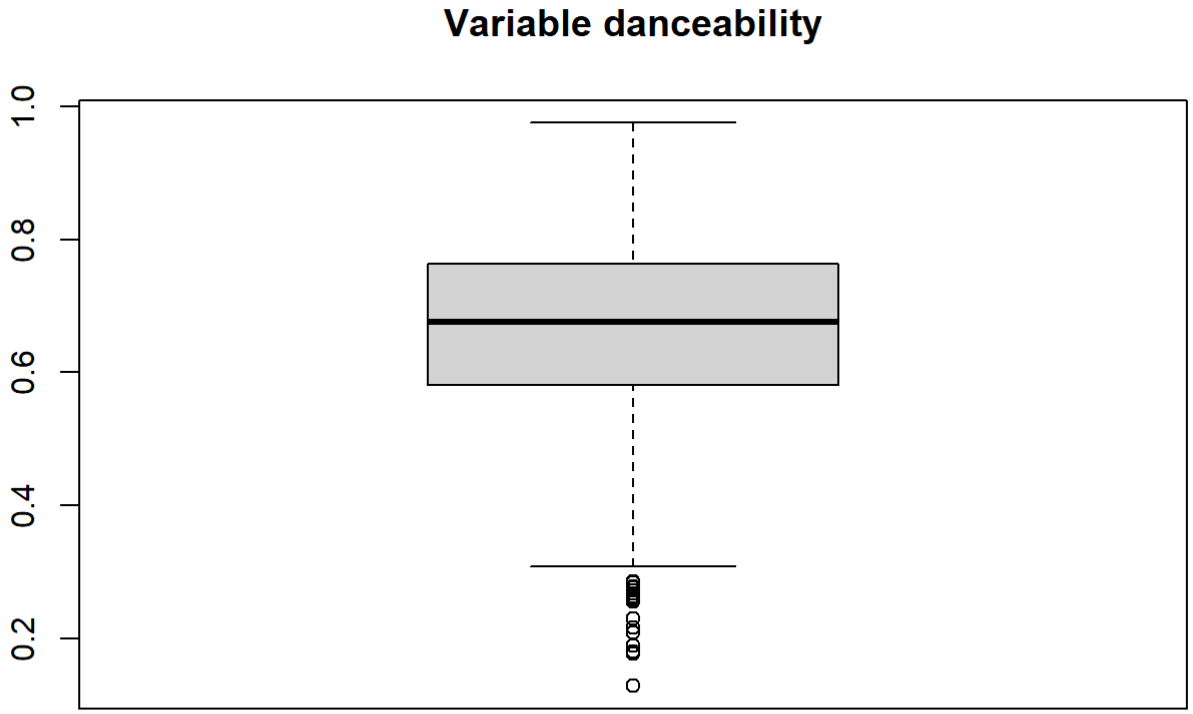
```
boxplot(data$duration_ms, main="Variable duración")
```



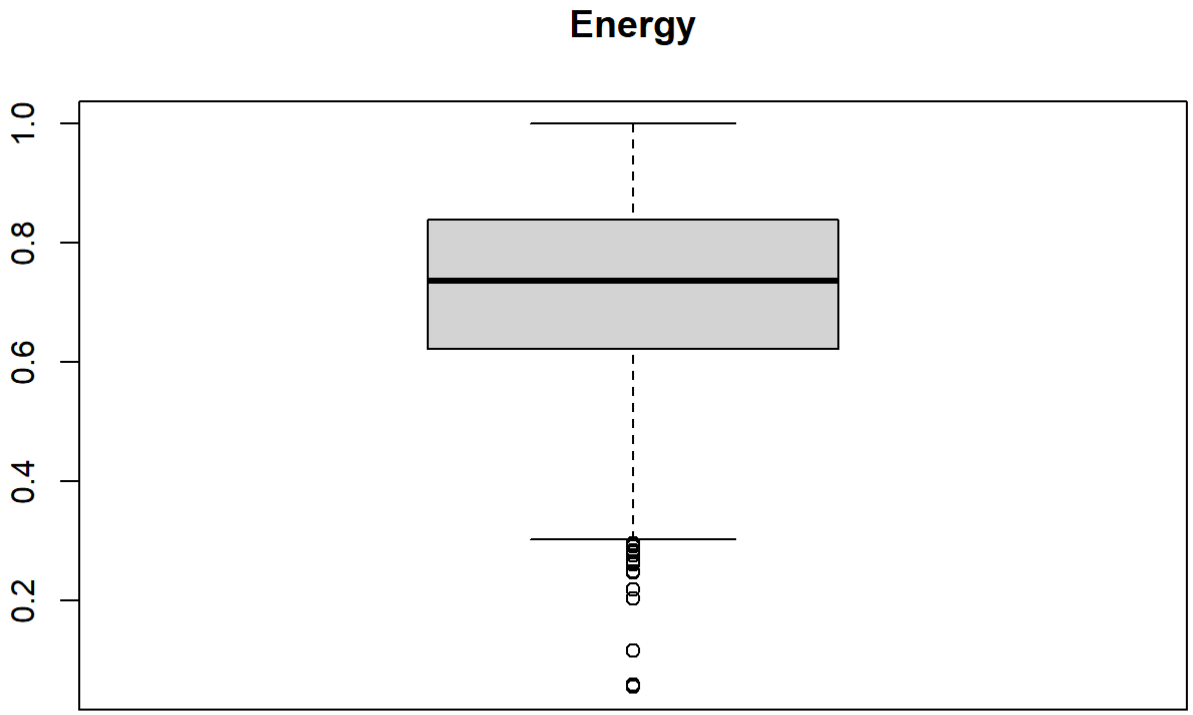

```
boxplot(data$popularity,  main="Variable popularity")
```



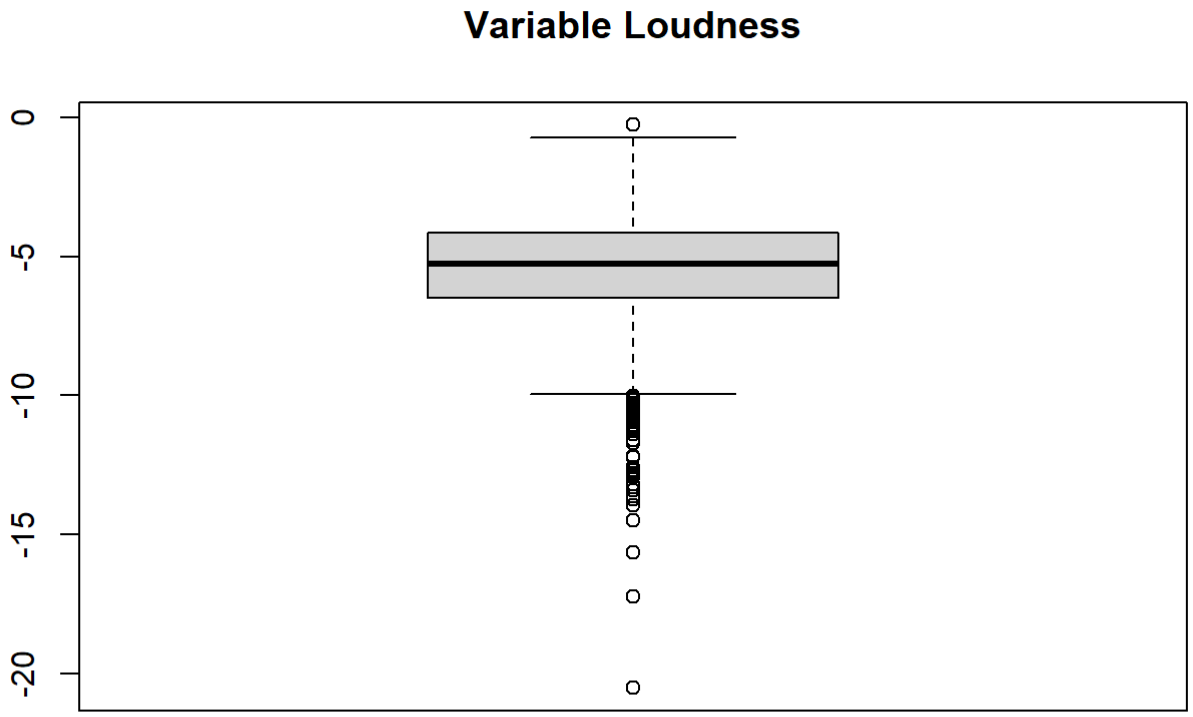

```
boxplot(data$danceability,  main="Variable danceability")
```



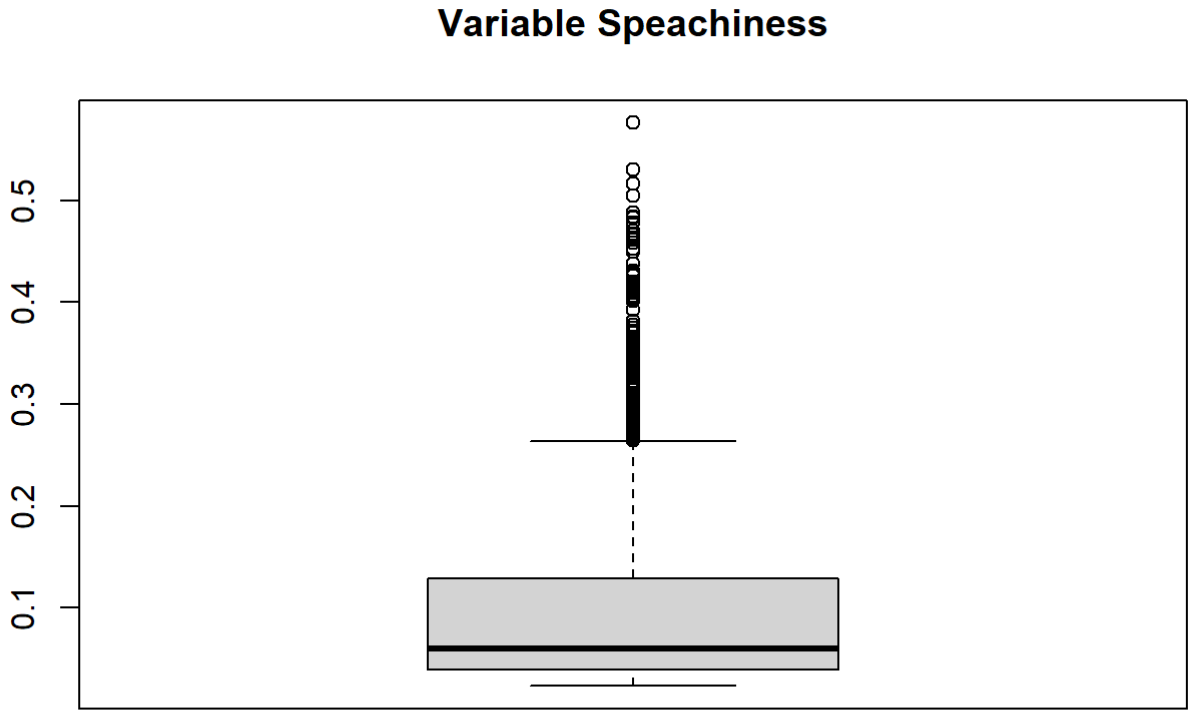
```
boxplot(data$energy,  main="Energy")
```



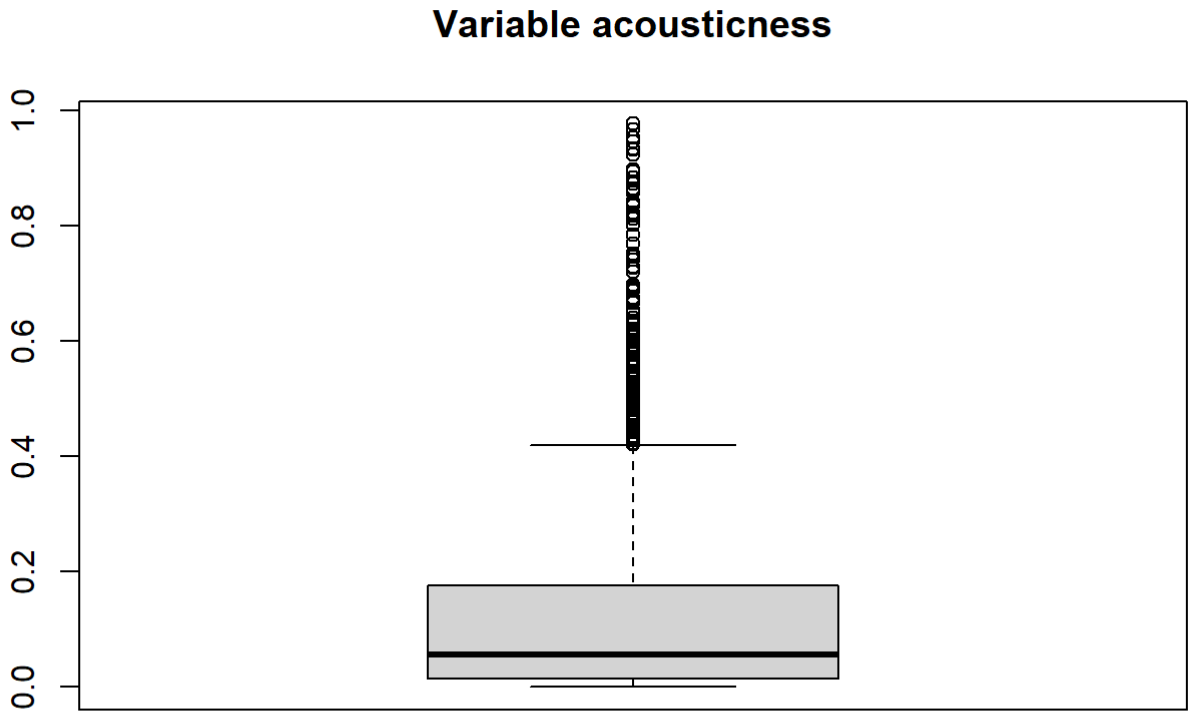

```
boxplot(data$loudness, main ="Variable Loudness")
```



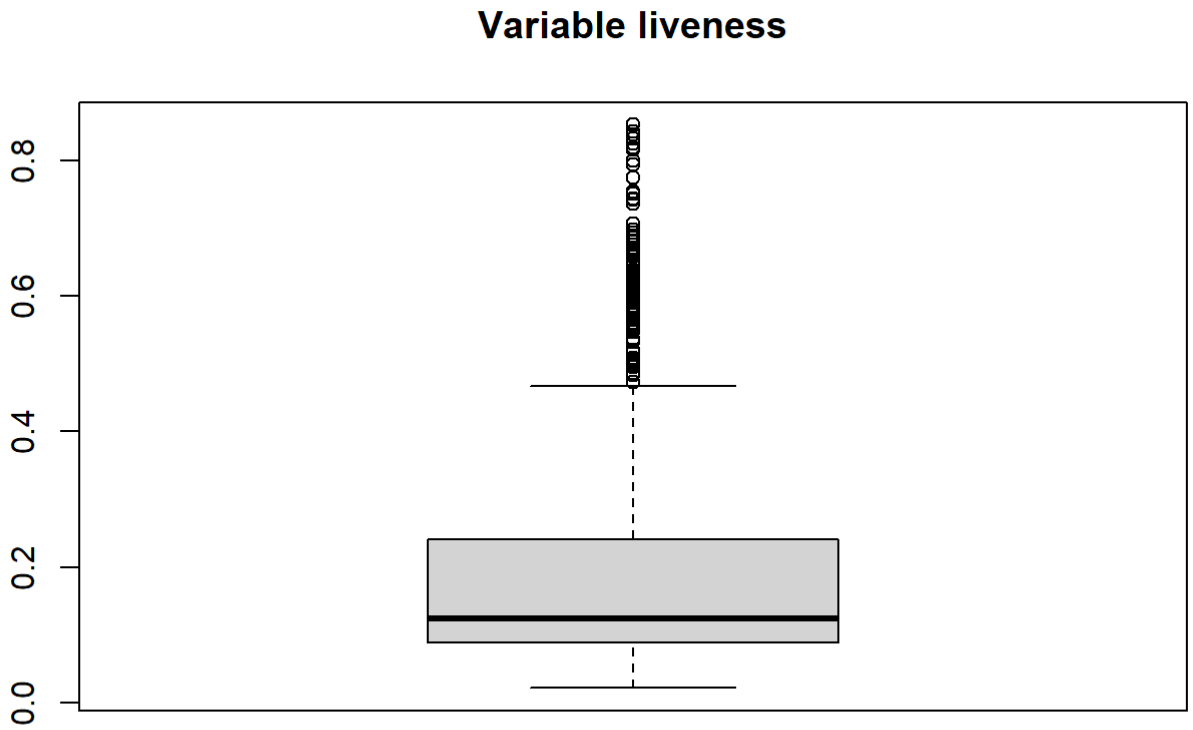

```
boxplot(data$speechiness, main ="Variable Speachiness")
```



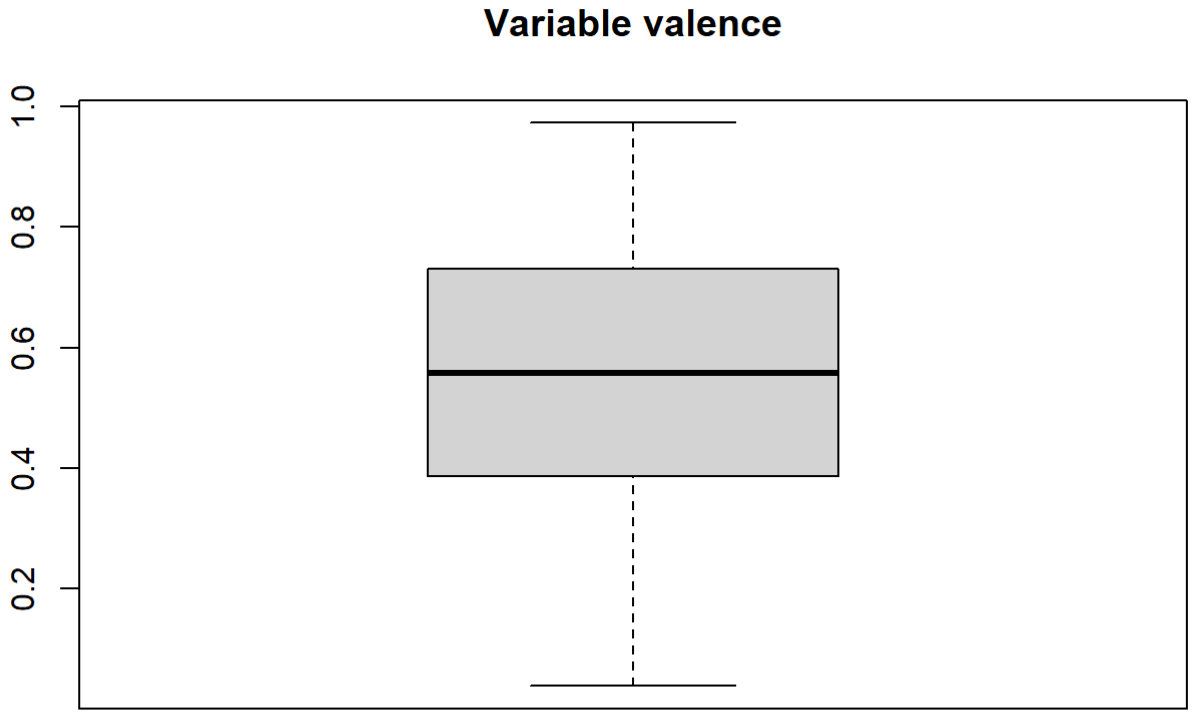
```
boxplot(data$acousticness, main ="Variable acousticness")
```



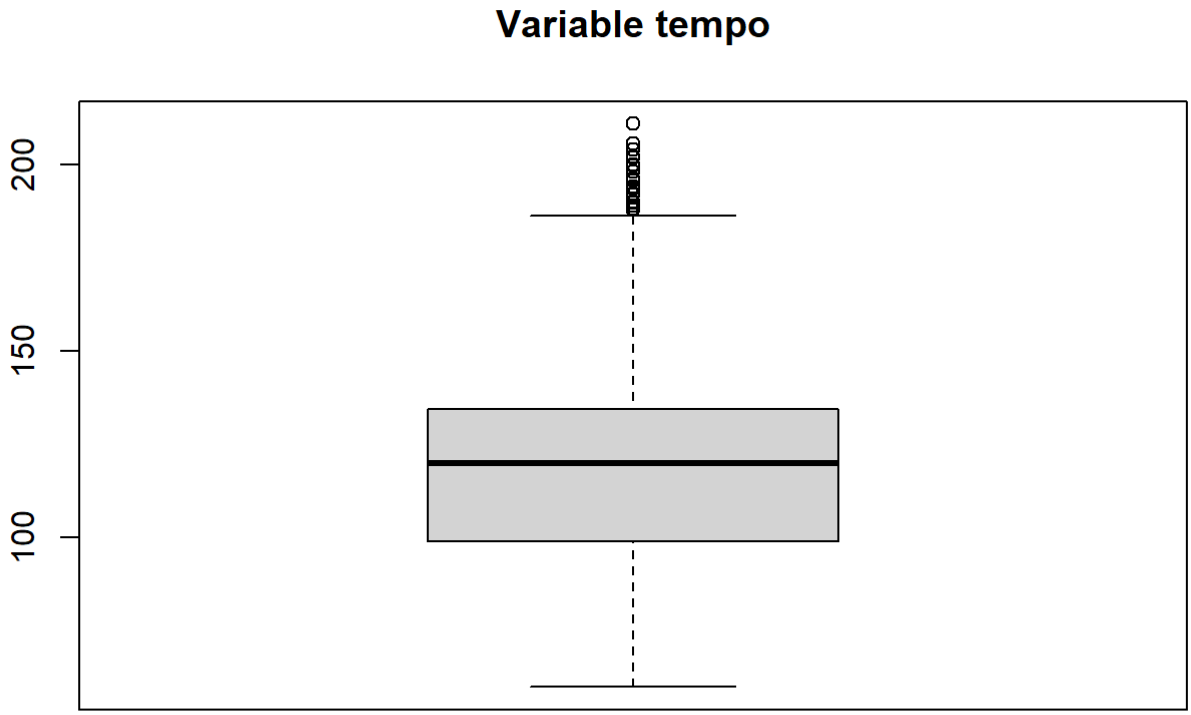

```
boxplot(data$liveness, main ="Variable liveness")
```




```
boxplot(data$valence, main ="Variable valence")
```



```
boxplot(data$tempo, main ="Variable tempo")
```



Todos los valores *outliers* están dentro de los límites descritos por la semántica del conjunto de datos. En la duración, es aceptable que existan canciones con duraciones muy por encima de la media, así como canciones que estén muy por debajo de la media de la popularidad.

Exportación de datos preprocesados

Pese a que no se han tenido que realizar a penas acciones sobre el dataset para adecuarlo al procesamiento, se procede a almacenar el resultado de la fase de preparación en un fichero csv.

```
write.csv(data, "SpotifyTopHits.csv")
```

Análisis de los datos

Selección de los grupos de datos a analizar

Para analizar algunas cuestiones interesantes, se procede a agrupar el conjunto de datos según diferentes valores de las variables. Durante esta división, haremos una asunción (detallada por el creador del conjunto de datos): consideraremos a las canciones con un valor de “acusticidad” por encima de 0.5 como acústicas, mientras que las que tengan este parámetro por debajo de 0.5 serán consideradas como no acústicas. Además, para la división por géneros se han considerado los 4 géneros con más ocurrencias. Para cononcer el número de ocurrencias, se ha ejecutado el siguiente bloque.

```
sort(table(data$genre))
```

```
##
##                       country, latin                       easy listening
##                                    1                                    1
##                  Folk/Acoustic, rock             Folk/Acoustic, rock, pop
##                                    1                                    1
##                     hip hop, country      hip hop, latin, Dance/Electronic
##                                    1                                    1
##                hip hop, pop, country pop, easy listening, Dance/Electronic
##                                    1                                    1
##             pop, R&B, easy listening                      rock, classical
##                                    1                                    1
##               rock, Dance/Electronic                  rock, easy listening
##                                    1                                    1
##   rock, Folk/Acoustic, easy listening             rock, Folk/Acoustic, pop
##                                    1                                    1
##    rock, pop, metal, Dance/Electronic         rock, R&B, Folk/Acoustic, pop
##                                    1                                    1
##      World/Traditional, Folk/Acoustic               World/Traditional, pop
##                                    1                                    1
##                   Folk/Acoustic, pop                   hip hop, rock, pop
##                                    2                                    2
##            pop, easy listening, jazz             pop, rock, Folk/Acoustic
##                                    2                                    2
##                          rock, blues                   rock, blues, latin
##                                    2                                    2
##           World/Traditional, hip hop World/Traditional, pop, Folk/Acoustic
##                                    2                                    2
##              World/Traditional, rock          World/Traditional, rock, pop
##                                    2                                    2
##   hip hop, pop, R&B, Dance/Electronic             hip hop, pop, R&B, latin
##                                    3                                    3
##                          hip hop, R&B                      rock, pop, metal
##                                    3                                    4
##           pop, R&B, Dance/Electronic                         pop, country
##                                    6                                    8
##                    pop, Folk/Acoustic           rock, pop, Dance/Electronic
##                                    8                                    8
##                    hip hop, pop, rock                                metal
##                                    9                                    9
##                               country           pop, rock, Dance/Electronic
##                                   10                                   13
##                                  R&B                  hip hop, pop, latin
##                                   13                                   14
##                      pop, rock, metal                                latin
##                                   14                                   15
##             hip hop, Dance/Electronic                               set()
##                                   16                                   22
##                             pop, rock                            pop, latin
##                                   26                                   28
##                           rock, metal                     Dance/Electronic
##                                   38                                   41
##                             rock, pop                                rock
##                                   43                                   58
##        hip hop, pop, Dance/Electronic                             hip hop
##                                   78                                  124
```

```
##                              pop, R&B                 pop, Dance/Electronic
##                                   178                                  221
##                     hip hop, pop, R&B                      hip hop, pop
##                                   244                                  277
##                                   pop
##                                   428
```

```
### Agrupación por modo (mayor o menor)

songs.mayor <- data[data$mode == 1,]
songs.menor <- data[data$mode == 0,]

### Agrupación explícitas o no

songs.explicit <- data[data$explicit,]
songs.nexplicit <- data[!data$explicit,]

### Agrupación por "acousticness"

songs.acoustic <- data[data$acousticness >= 0.5,]
songs.nacoustic <- data[data$acousticness < 0.5,]

### Agrupación por géneros: "pop" ,"rock, pop", "hip hop, pop, R&B", "pop, Dance/Elec
tronic"
songs.pop <- data[data$genre == "pop",]
songs.rock <- data[data$genre == "rock, pop",]
songs.hiphop <- data[data$genre == "hip hop, pop, R&B",]
songs <- data[data$genre == "pop, Dance/Electronic",]
```

Comprobación de la normalidad y homogeneidad de la varianza

El siguiente paso es la comprobación de la normalidad de las distribuciones de las diferentes variables cuantitativas que componen el conjunto de datos. Para ello, se implementa una función que indicará cuales de las variables siguen una distribución normal según el test de Saphiro-Wilk con un nivel de significación del 0.05.

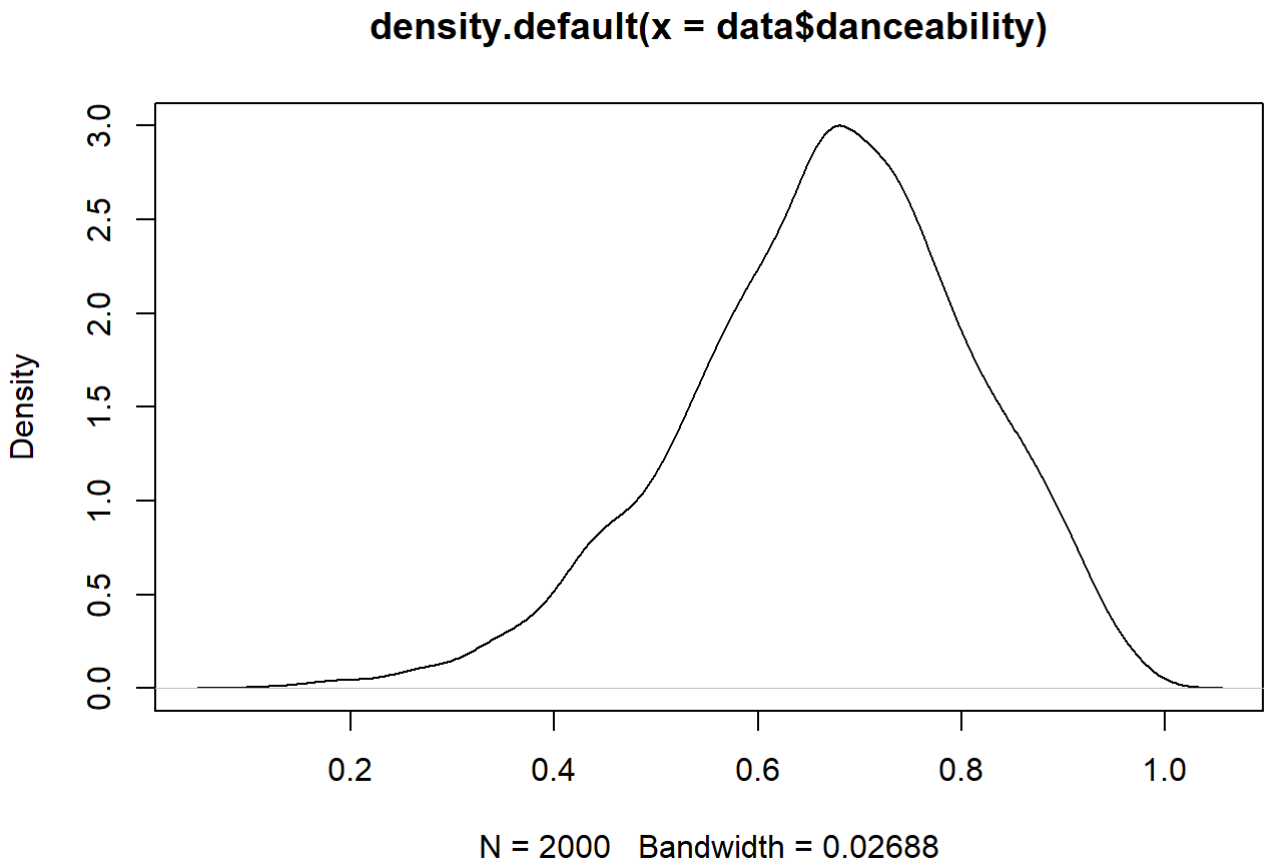
```
library("stats")

alpha = 0.05
col.names = colnames(data)
for (i in 1:ncol(data)) {
  if(is.integer(data[,i]) | is.numeric(data[,i])) {
    if(shapiro.test(sample(data[,i], 500))$p.value < alpha) {
      cat(col.names[i])
      cat(" tiene distribución normal")
      cat("\n")
    } else {
      cat(col.names[i])
      cat(" no tiene distribución normal")
      cat("\n")
    }
  }
}
```

```
## duration_ms tiene distribución normal
## year tiene distribución normal
## popularity tiene distribución normal
## danceability tiene distribución normal
## energy tiene distribución normal
## key tiene distribución normal
## loudness tiene distribución normal
## mode tiene distribución normal
## speechiness tiene distribución normal
## acousticness tiene distribución normal
## instrumentalness tiene distribución normal
## liveness tiene distribución normal
## valence tiene distribución normal
## tempo tiene distribución normal
```

Del análisis de normalidad de las diferentes variables, se extrae la conclusión de que todas ellas siguen una distribución normal.

Se comprueba esta conclusión mediante la visualización de la distrbución de la variable `danceability` :

```
plot(density(data$danceability))
```



El siguiente paso consistirá en el estudio de la homogeneidad de varianzas, mediante la aplicación del test de Fligner-Killeen. Se estudiará la homogeneidad en cuanto a los casos activos agrupados por países. La hipótesis nula del siguiente test es que las varianzas son iguales.

```
fligner.test(popularity ~ genre, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  popularity by genre
## Fligner-Killeen:med chi-squared = 130.03, df = 58, p-value = 1.91e-07
```

Dado que el p-value es menor a α=0.05, se rechaza la hipótesis nula, es decir, no podemos afirmar que las varianzas son homogéneas para los casos activos por país.

Pruebas estadísticas

¿Influye la explicitud de una canción en su popularidad? ¿Son más populares las canciones no explícitas que las explícitas?

Ya contamos con la división de las canciones por explicitud, por lo que se puede proceder con el contraste de medias para comprobar la hipótesis nula de que las canciones explítitas son, en promedio, menos populares que las no explícitas.

$$H_0 : \mu_0 - \mu_1 = 0$$
$$H_1 : \mu_0 - \mu_1 > 0$$

Siendo μ_0 la popularidad media de las canciones no explícitas y μ_1 la popularidad media de las canciones explícitas.

```
t.test(songs.explicit$popularity, songs.nexplicit$popularity, alternative = "greater"
)
```

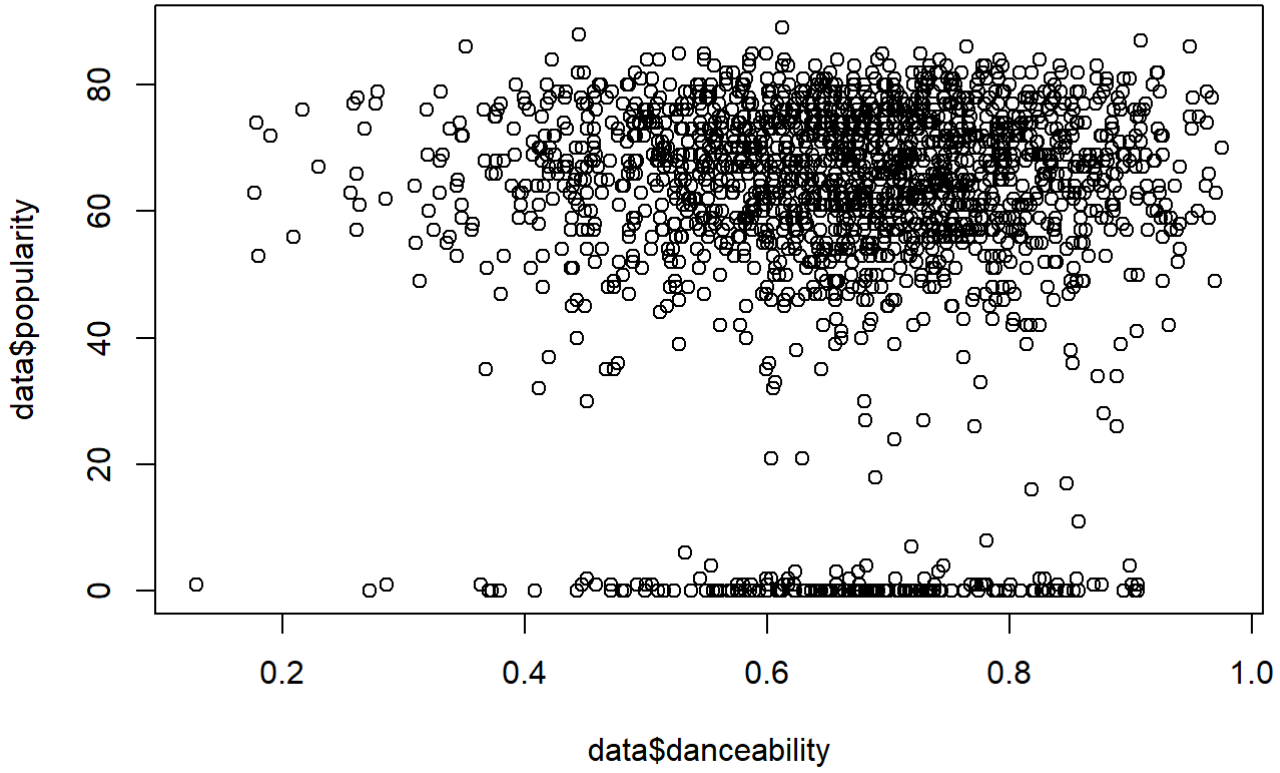
```
##
##  Welch Two Sample t-test
##
## data:  songs.explicit$popularity and songs.nexplicit$popularity
## t = 2.127, df = 1033.8, p-value = 0.01683
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.502834      Inf
## sample estimates:
## mean of x mean of y
##  61.48457  59.25949
```

La ejecución de este test indica que no es posible aceptar la hipótesis nula, por lo que es posible concluir que las canciones no explícitas tienen mayor popularidad que las canciones no explícitas.

¿Son las canciones más “bailables” automáticamente más populares?

Para responder a esta pregunta debemos conocer si existe correlación entre las variables `popularity` y `danceability` . Para ello, procedemos primeramente con un análisis visual de la relación entre las variables

```
plot(data$danceability, data$popularity)
```



Como se puede apreciar claramente en la visualización, de primeras no parece existir corelación alguna entre las dos variables. Para comprobar esto, se procede con un estudio de correlación

```
cor.test(data$popularity, data$danceability, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$popularity and data$danceability
## t = -0.15849, df = 1998, p-value = 0.8741
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04736930  0.04029146
## sample estimates:
##         cor
## -0.00354573
```

Se comprueba que efectivamente no existe relación alguna entre la “bailabilidad” de las canciones y su popularidad.

Modelo de regresión lineal

Para este último apartado, se propone la elaboración de un modelo de regresión lineal que permita obtener una predicción de la popularidad de una canción en base a los parámetros incluidos en el dataset. En primer lugar se realiza la selección de las mejores variables mediante la visualización de la matriz de correlación.

```
library("corrplot")
```

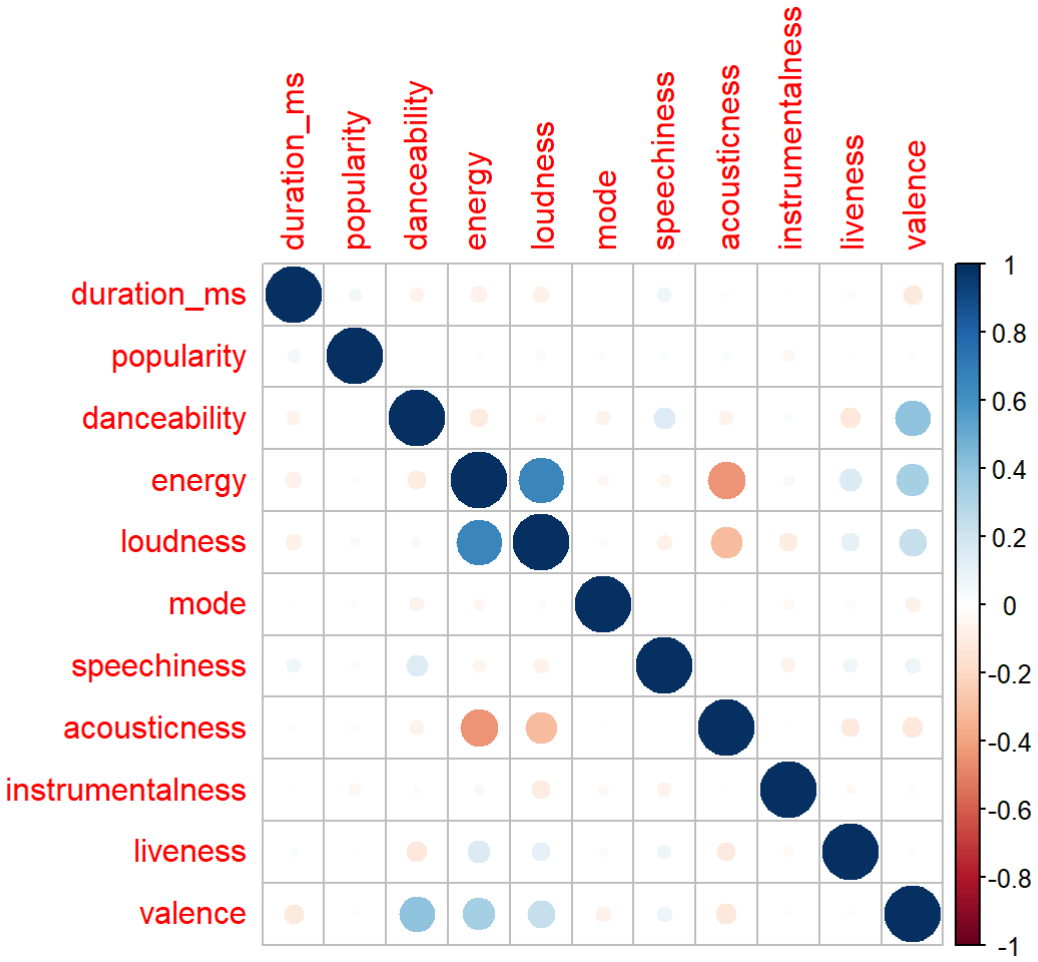
```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
colnames(data)
```

```
##  [1] "artist"           "song"             "duration_ms"      "explicit"
##  [5] "year"             "popularity"       "danceability"     "energy"
##  [9] "key"              "loudness"         "mode"             "speechiness"
## [13] "acousticness"     "instrumentalness" "liveness"         "valence"
## [17] "tempo"            "genre"
```

```
corrplot(cor(data[,c(3, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16)]))
```



Popularity no tiene relación lineal con ningua de las variables. Emplearemos duration, energy y valence. Para comparar, utilizaremos otro modelo elaborado con las variables speachiness, valence y danceability.

```
model <- lm(popularity ~ duration_ms + energy + valence, data = data)
model1 <- lm(popularity ~ speechiness + valence + danceability, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = popularity ~ duration_ms + energy + valence, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.419  -3.629   5.577  13.425  28.789
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.493e+01  3.834e+00  14.325   <2e-16 ***
## duration_ms  2.677e-05  1.228e-05   2.180   0.0294 *
## energy      -1.050e+00  3.316e+00  -0.317   0.7516
## valence     -7.622e-01  2.302e+00  -0.331   0.7406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.32 on 1996 degrees of freedom
## Multiple R-squared:  0.002718,   Adjusted R-squared:  0.001219
## F-statistic: 1.813 on 3 and 1996 DF,  p-value: 0.1426
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = popularity ~ speechiness + valence + danceability,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.596  -3.492   5.757  13.425  29.212
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.27080    2.34547  25.697   <2e-16 ***
## speechiness   4.97489    5.01858   0.991    0.322
## valence      -1.73565    2.36218  -0.735    0.463
## danceability  0.06594    3.74536   0.018    0.986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.34 on 1996 degrees of freedom
## Multiple R-squared:  0.000763,   Adjusted R-squared:  -0.0007389
## F-statistic: 0.508 on 3 and 1996 DF,  p-value: 0.6768
```

Parece ser que ambos modelos han fracasado a la hora de intentar establecer una relación lineal entre las variables introducidas. De cara a la finalización del análisis, realizaremos predicciones con el primer modelo, que tiene una puntuación de R-squared más alta.

```
pred.data <- data.frame(
  duration_ms= 180000,
  energy= 0.7,
  valence = 0.85
)
predict(model, pred.data)
```

```
##        1
## 58.36158
```

Según el modelo realizado, para una canción de 3 minutos con puntuaciones de 0.7 en energía y 0.85 en *vivance*, la popularidad deberá de estar al rededor de 58.361 puntos.

Conclusiones

El conjunto de datos empleado para el análisis contiene una información muy interesante sobre los gustos musicales globales y sus raíces: permite el análisis de patrones en las canciones más populares, para averiguar cuáles son las características que hacen un buen éxito. Durante las pruebas estadísticas, se han planteado y resuelto cuestiones en esta línea, y se ha averiguado que no existe relación entre la “bailabilidad” de una canción y su popularidad (aunque a priori parezca que tiene sentido), y que las canciones no tienen que ser menos populares por ser explícitas.

Finalmente, con la elaboración de un modelo lineal en base a este conjunto de variables que permita realizar predicciones sobre la variable `popularity` , se ha averiguado que un modelo predictivo basado en relaciones lineales no es suficiente para modelar la relación entre estas variables y su variable dependiente. Se necesitará un modelo de regresión más potente, como un árbol de decisión o una red neuronal para correctamente modelar la relación.

Recursos

- Mireia, C., Diego, P., Laia, S.(2019). *Introducción a la limpieza y análisis de los datos.* Material UOC.
- Mireia, C., Diego, P., Laia, S.(2019). *Introducción al ciclo de vida de los datos*. Material UOC.