Emmanuel Cardozo
Lucas Didone

Trabajo práctico integrador.
Análisis de datos

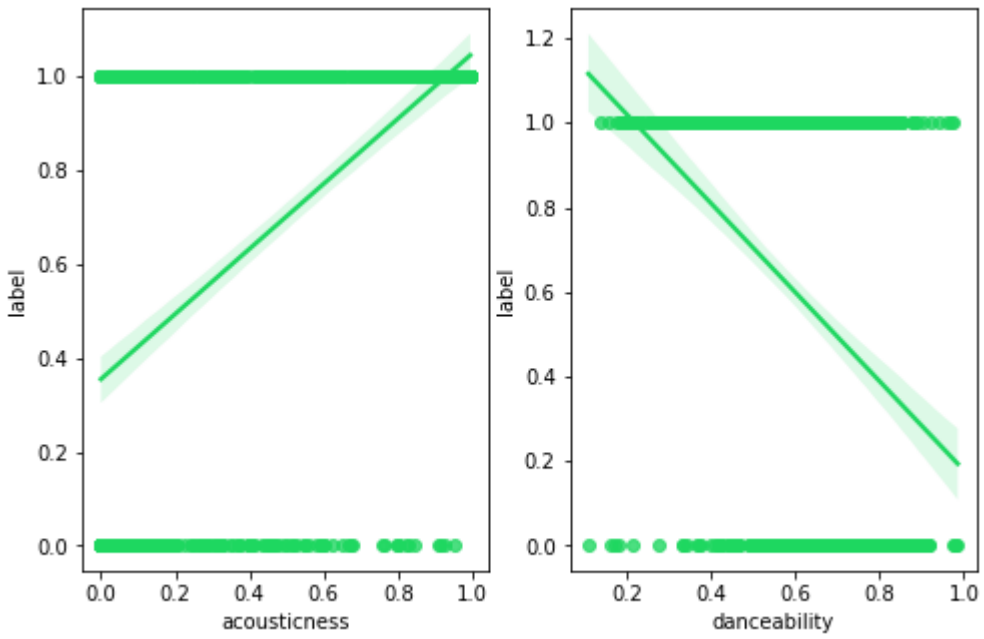
3. Limpieza y preparación de datos / Ingeniería de features.

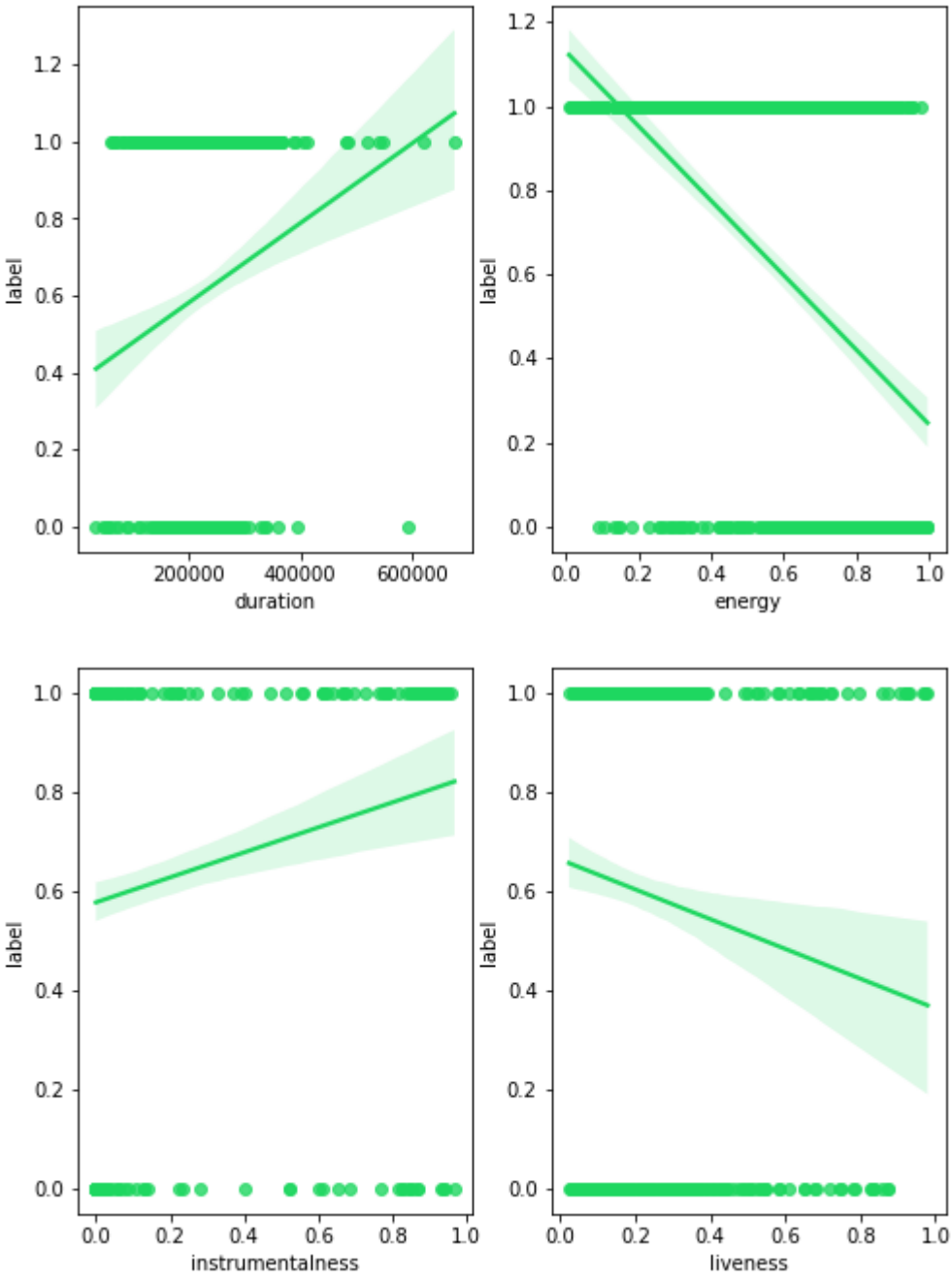
Dado que no se evidencia la presencia de valores faltantes o nulo, se pasa responder la pregunta:

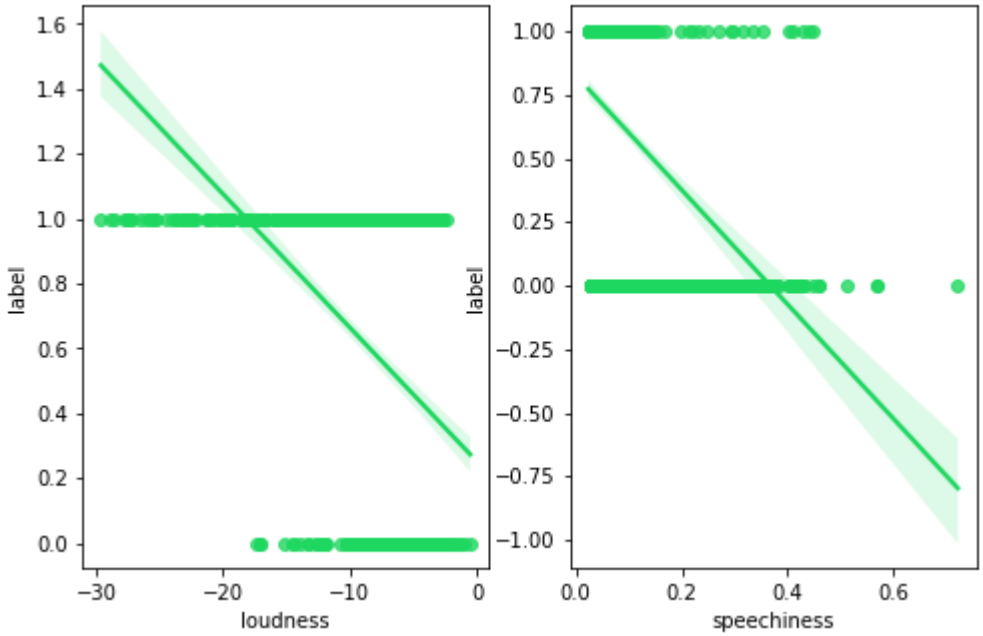
¿Qué puede decir acerca de las relaciones entre las variables de entrada?

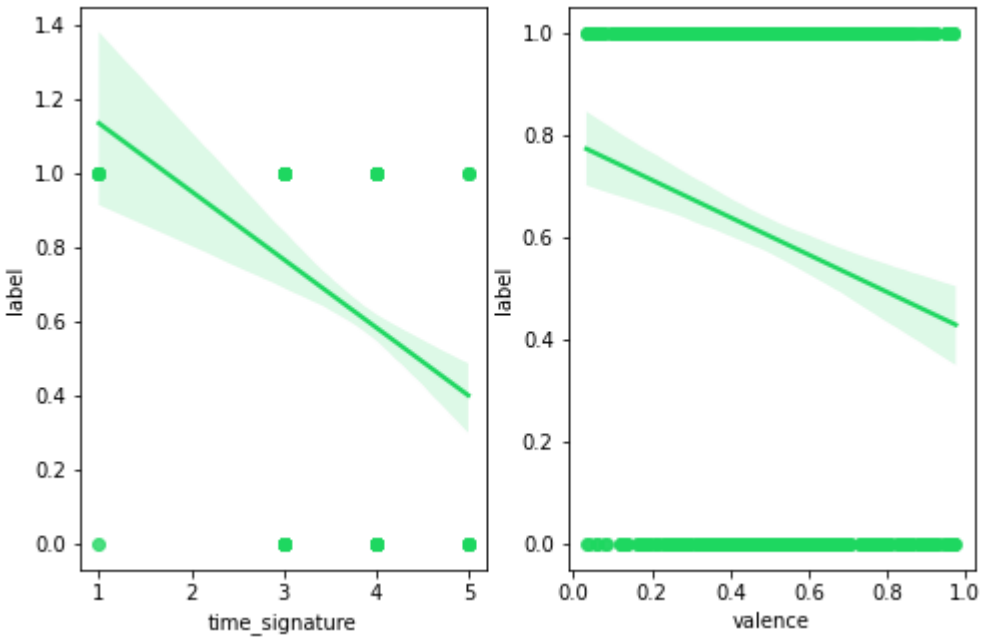
Spotify

En primer término, se grafican los cada una de las features numéricas vs la variable de salida “label”, de manera de intentar conocer si existen relaciones lineales entre features, por la similitud entre las pendientes de las gráficas:




label
duration
energy
instrumentalness
liveness





A simple vista, no se encuentran gráficas similares. No se considera que exista de momento, relación lineal entre features de entrada.

Se procede a graficar la matriz de Pearson para encontrar relaciones entre “label” y las demás features:

Se observa para “label” una fuerte relación lineal con “acousticness” y una menos fuerte con “duration” e “instrumentalness”.

Se grafica a su vez, la relación entre las features y la variable de salida “label” utilizando información mutua.

Si bien se observan más relaciones entre la variable de salida y otras features (acousticness nuevamente, danceability, energy, loudness y speechiness), ninguna de ellas es fuerte.

Missing values:

Si bien no se evidenciaron valores nulos en nuestro dataset, los generamos con el scrytp enviado en clase (data besmircher.ipynb). Con un valor de un 20% (y no con filas esparsas), decidimos ver si las distribuciones, relaciones entre variables, entre features y variables de salida, entre otras, se modifican al efectuar una imputación por medianas.

Las columnas afectadas por el script, fueron las siguientes:

-Acousticness

-Danceability

-Energy

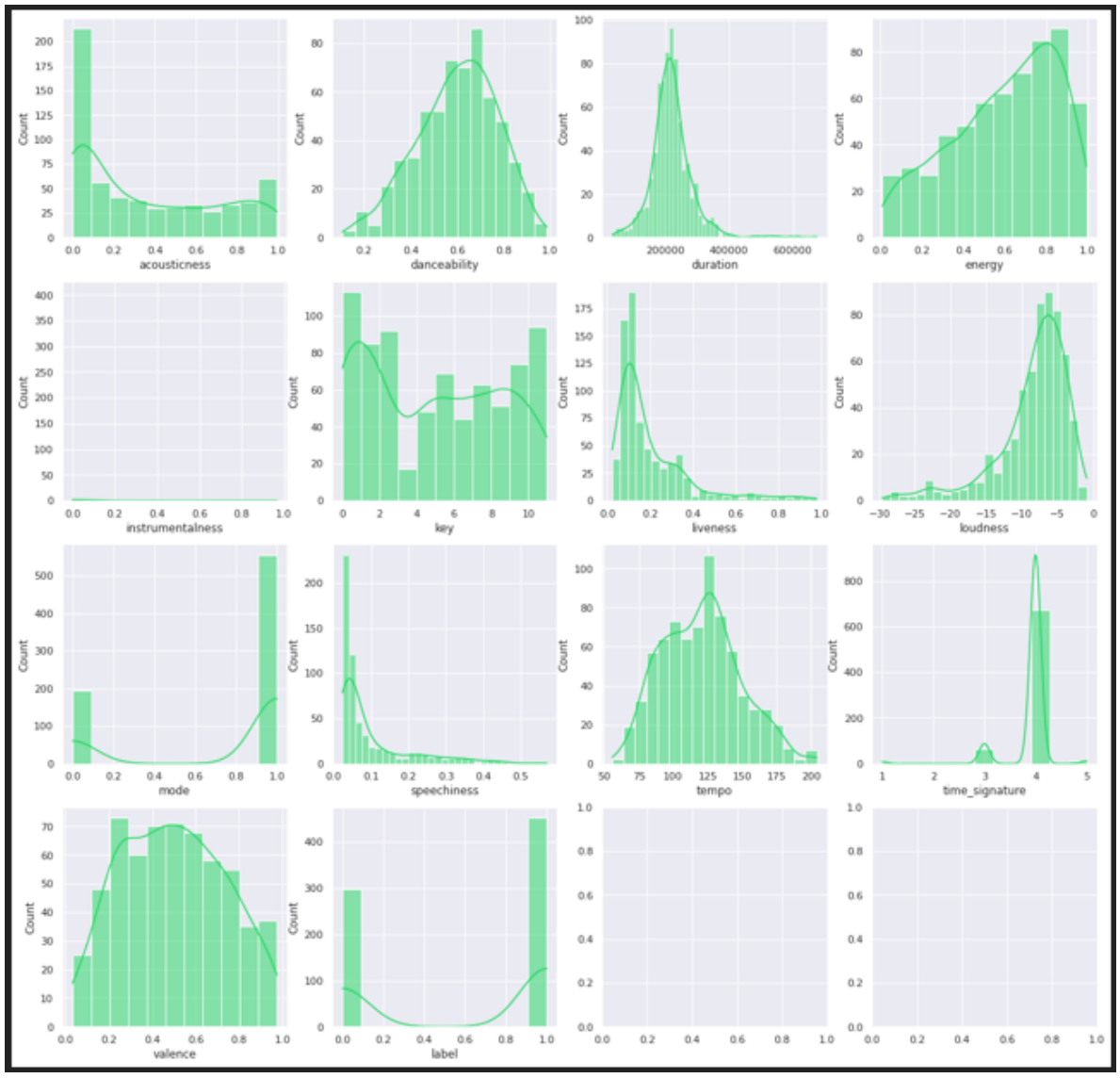
-Instrumentalness

-Loudness

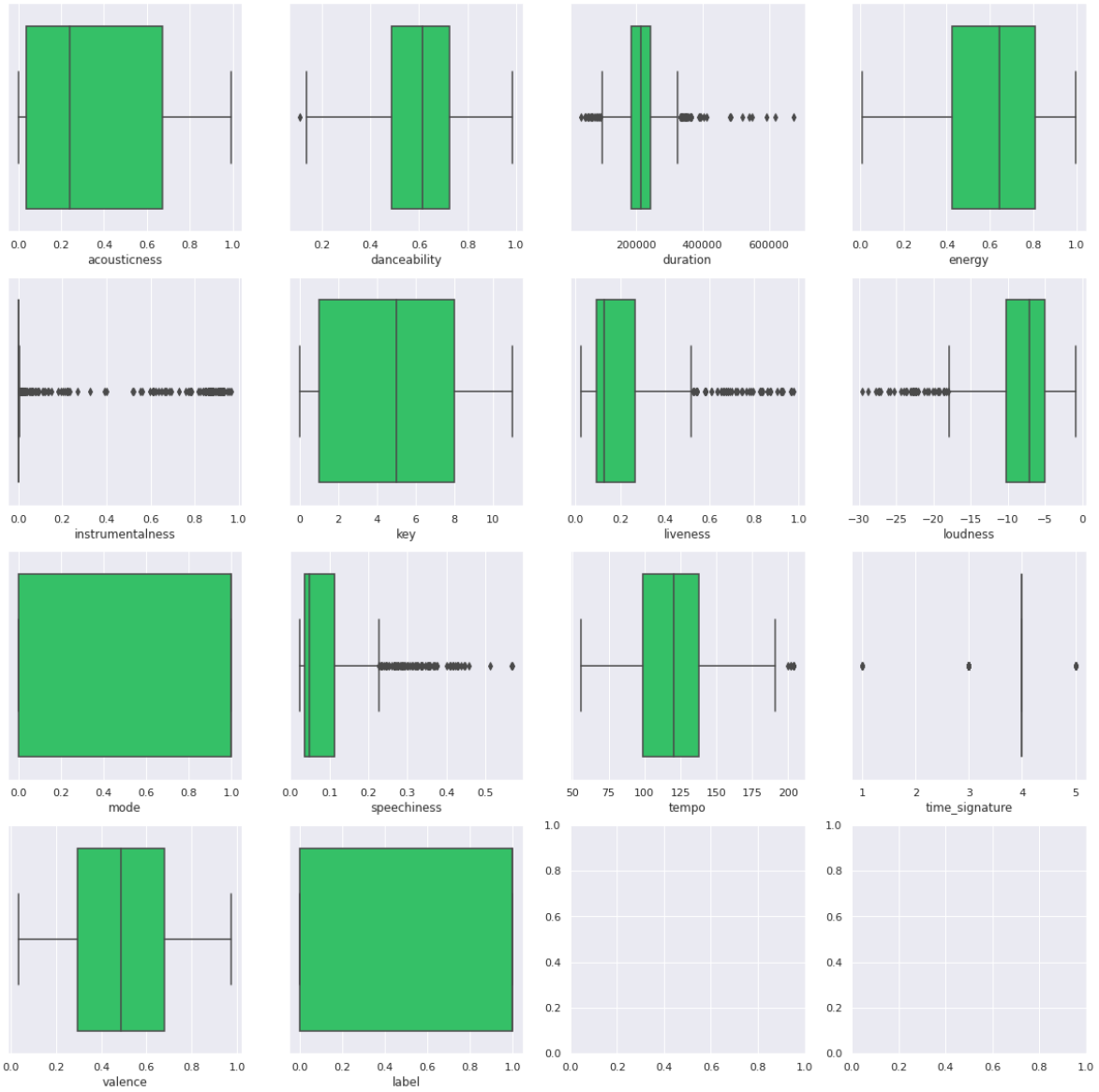
-Speechiness

-Valence

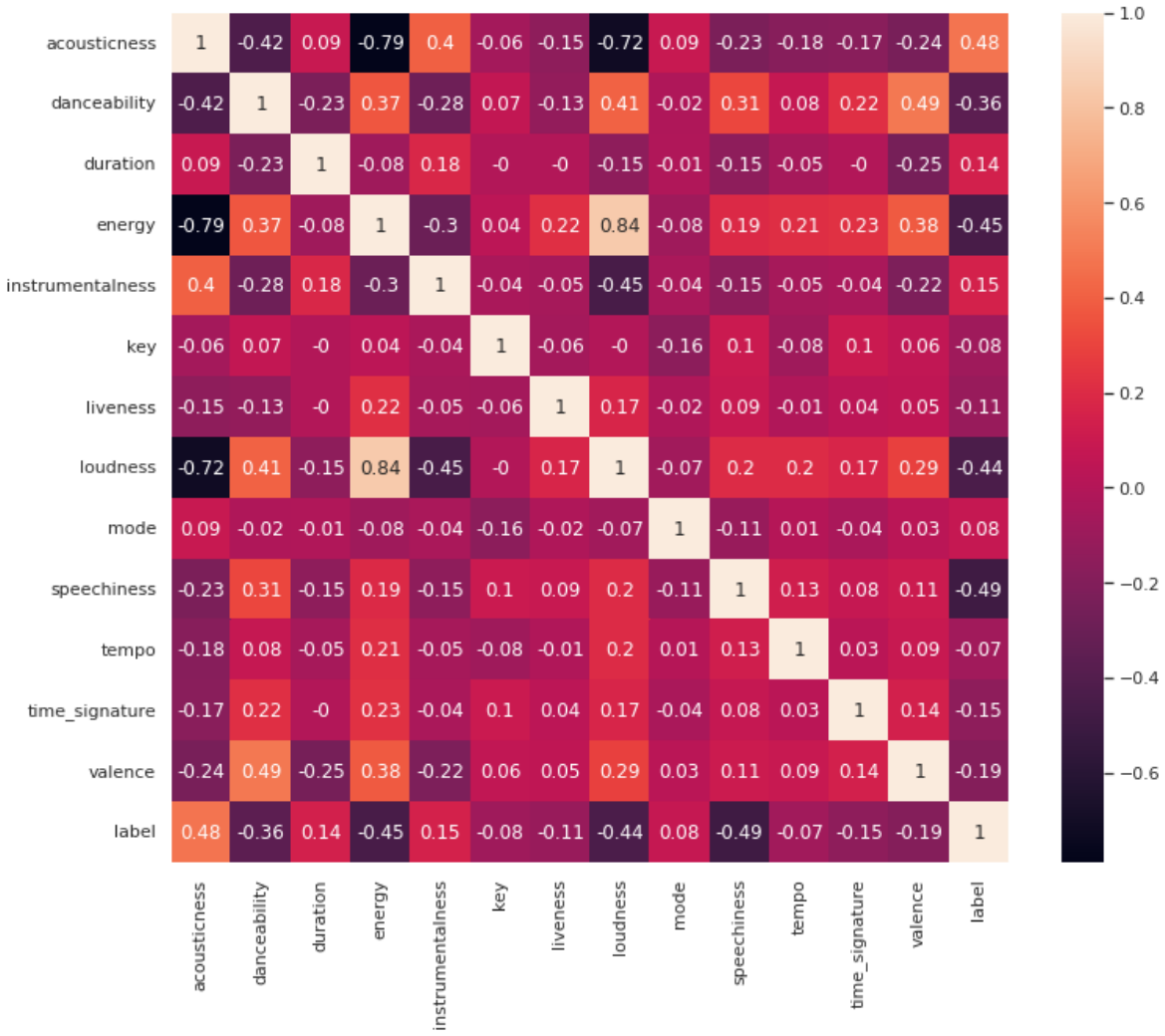
Histogramas del Dataframe afectado:



Box-Plots:


acousticness
danceability
duration
energy
instrumentalness
key
liveness
loudness
mode
speechiness
tempo
time_signature
valence
label

Matriz de correlación de Pearson con un mapa de calor:



Conclusiones: Si bien no se experimentan modelos de clasificación en el dataset, como si se hará con el original, con un 20% de valores nulos generados con el script, y una posterior imputación de valores faltantes por medianas, si sigue manteniendo una fuerte relación lineal entre la variable de salida “label” y “acousticness”, como también una menor relación lineal entre “duration” e “instrumentalness”.