Trabajo Práctico
Análisis Predictivo

Makk - Gonzalez - Virgili

Índice

01 **Objetivo**
Caso de negocios.

02 **Análisis preliminar**
Información exploratoria.

03 **Análisis de variables**
Correlaciones.

04 **Conclusión**
Reflexiones finales. Links de Interés.

01

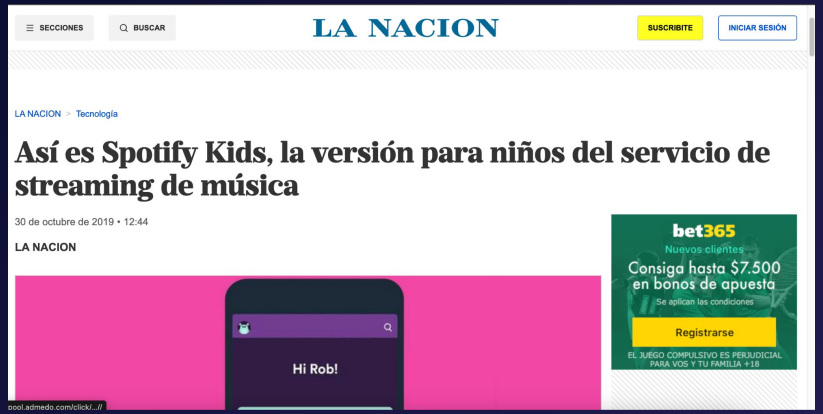
Objetivo

Caso de negocio. Motivación.

Caso de negocio

Predecir qué contenidos son explícitos en función a sus características para el desarrollo del segmento Spotify Kids.
El mismo implica:

- Entorno seguro y adecuado
- Contenido curado y apropiado para la edad


SECCIONES
BUSCAR
LA NACION
SUSCRIBITE
INICIAR SESIÓN
LA NACION > Tecnología
Así es Spotify Kids, la versión para niños del servicio de streaming de música
30 de octubre de 2019 • 12:44
LA NACION
Hi Rob!
bet365
Nuevos clientes
Consiga hasta $7.500 en bonos de apuesta
Se aplican las condiciones
Registrarse
EL JUEGO COMPULSIVO ES PERJUDICIAL PARA VOS Y TU FAMILIA +18


ámbito
Dólar
Economía
Finanzas
ámbito.com • Negocios • Spotify
17 de marzo 2020 - 10:46
Spotify Kids ya está disponible en Argentina
Se trata de una nueva aplicación, independiente de la tradicional de Spotify. Es libre de avisos publicitarios y con contenidos cuidadosamente seleccionados.
• Devaluación: a cuánto subirán los abonos de Netflix, Apple TV, y Spotify
• Tener un Smart Home a bajo costo, ya no es una misión imposible


30 Ago, 2023
Argentina
Colombia
España
México
Perú
Mundo
Últimas Noticias
infobae
Últimas Noticias
Rusia invade Ucrania
Deportes
Venezuela
Tecnología
ESPN
Qué Puedo Ver
Entretenimiento
Leamos
EEUU
TECNO ›
Spotify Kids: la apuesta de la plataforma vía streaming para atraer al público infantil
Las playlist incluirán bandas sonoras de películas de Disney e incluso algunos éxitos pop de artistas que son muy seguidos por este sector de la población. Por el momento, sólo está disponible en Irlanda

02

Análisis preliminar

Información exploratoria

Base de datos

kaggle

El dataset analizado contiene canciones de Spotify en una variedad de 114 géneros musicales. Cada género incluye 1000 registros e incluye las características sonoras del mismo.

Considerando la divergencia que pueden presentar individualmente cada uno de los géneros, hemos decidido hacer una selección de 10 a fines de especializar el análisis. Los elegidos fueron:

- Clasico
- Metal
- Jazz
- Punk-Rock
- Techno
- Reggae
- Sleep
- Trance
- Hip-Hop
- Study

Variables

Extras

- Track_id
- Artista
- Album Name
- Track Name

Categóricas

- Explicit T/F
- Key
- Genre
- Mode 0/1
- Time Signature

Numéricas

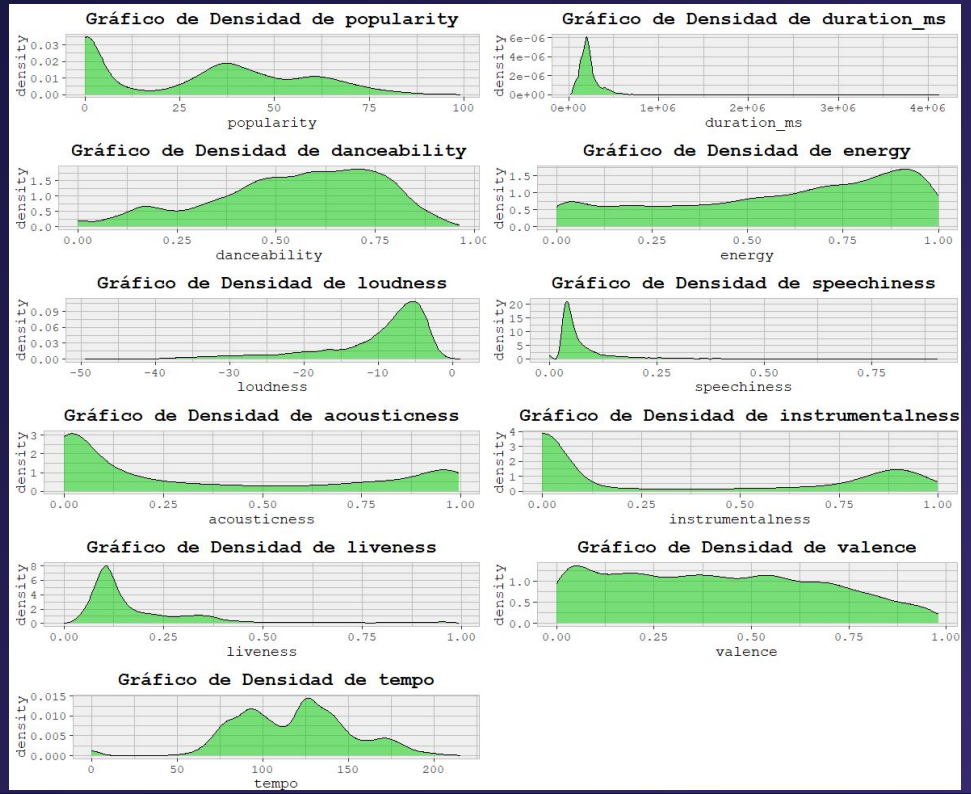
- Speechiness [0-1]
- Acousticness [0-1]
- Instrumentalness [0-1]
- Liveness [0-1]
- Valence [0-1]
- Tempo
- Popularidad [0-100]
- Duración en milisegundos
- Danceability [0-1]
- Energy [0-1]
- Loudness (dB)

Missings

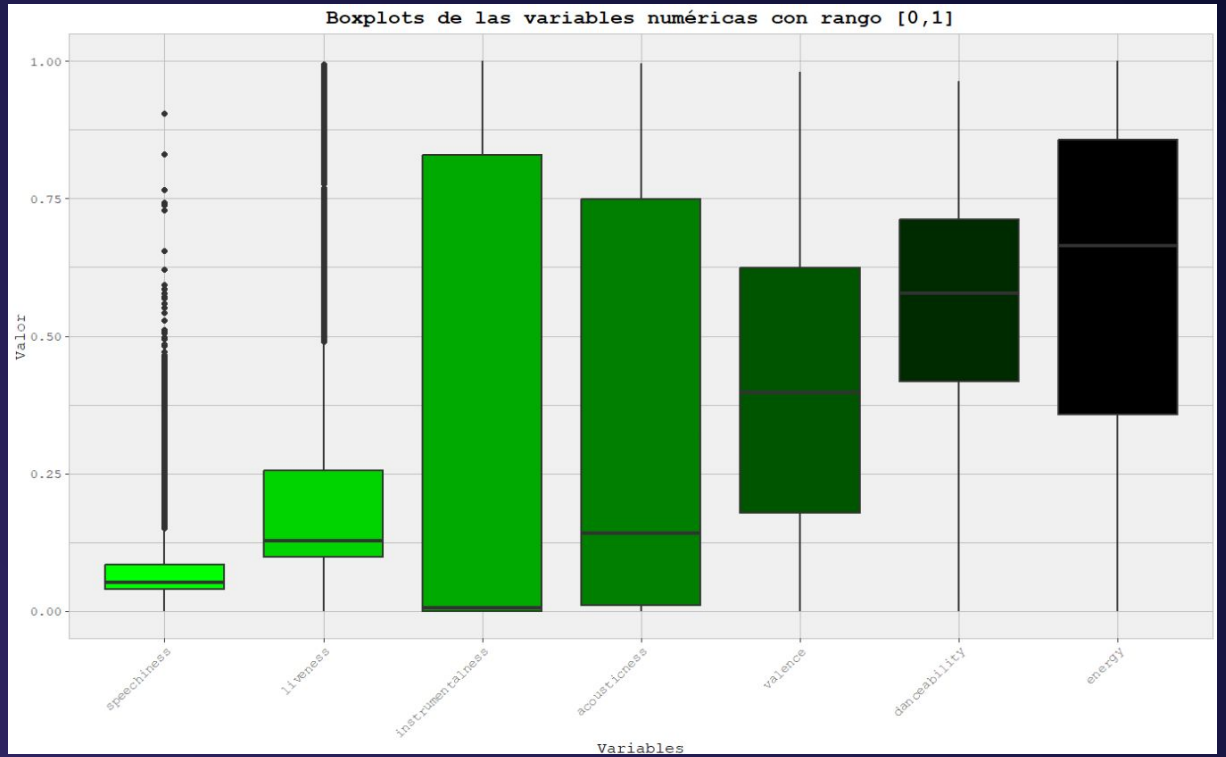
Ninguna de las variables presenta valores nulos (NAs)

Distribución de las variables numéricas

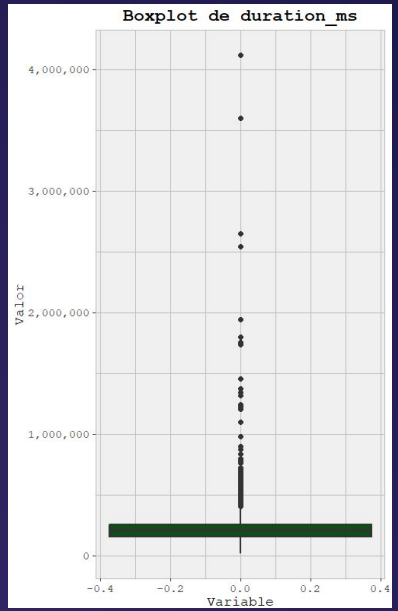
Se observa que hay pocas canciones con 60 minutos. La distribución más simétrica es loudness, valence o danceability. El tempo parece tener 2 modas.

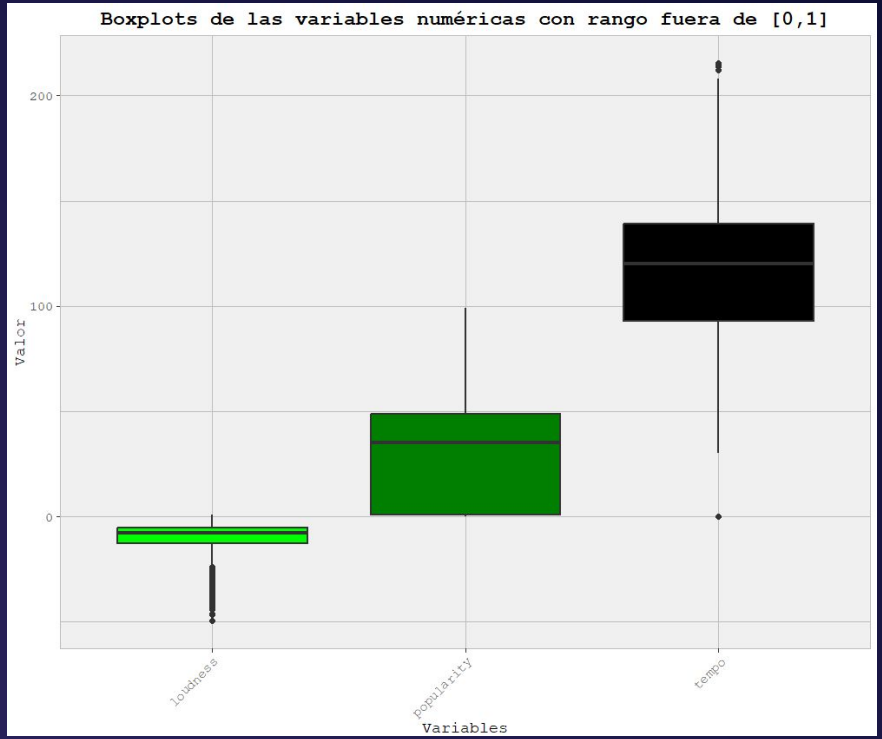


Outliers


Boxplots de las variables numéricas con rango [0,1]
Valor
1.00
0.75
0.50
0.25
0.00
speechiness
liveness
instrumentalness
acousticness
valence
danceability
energy
Variables

Outliers





4,000,000ms = 66.6667 minutos

03

Análisis de variables

Correlaciones. Variable Objetivo.

Transformación de los datos

Para la fase de análisis de variables se ha realizado una transformación de los datos. Esta decisión se ha optado basado en el hecho de que cada variable posee su propia escala, siendo que algunas toman valores negativos, otras entre [0-1] y otras en escalas positivas mayores. Para ello se ha optado por el método **min-max**

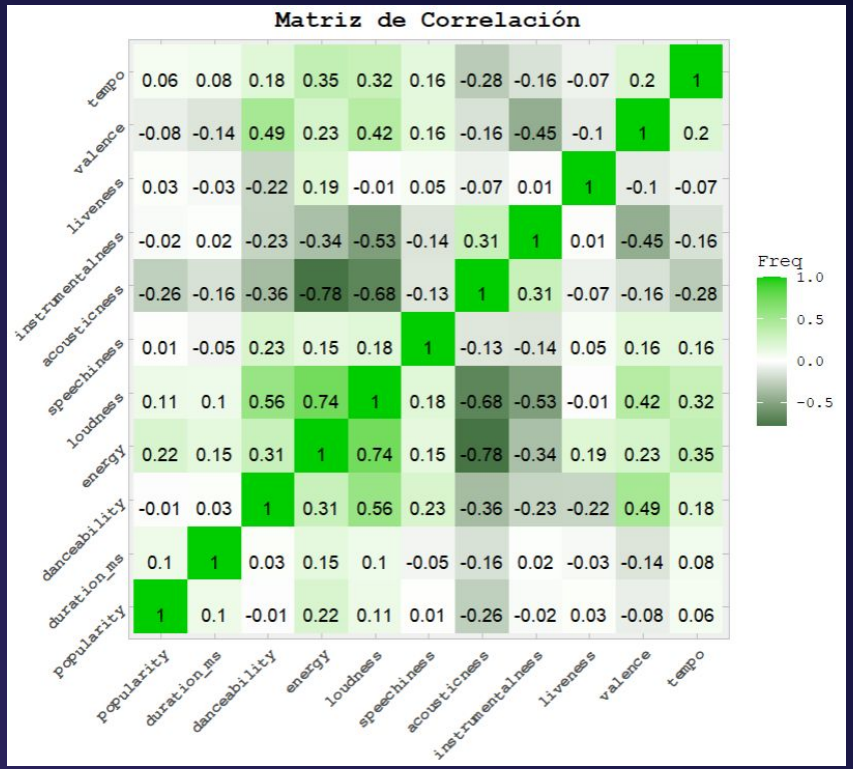
$$\frac{x - min(x)}{max(x) - min(x)}$$

Todas las variables quedan expresadas en el rango 0-1. Las medias y los desvíos de las variables no son iguales, a diferencia de la estandarización convencional.

Matríz de correlación

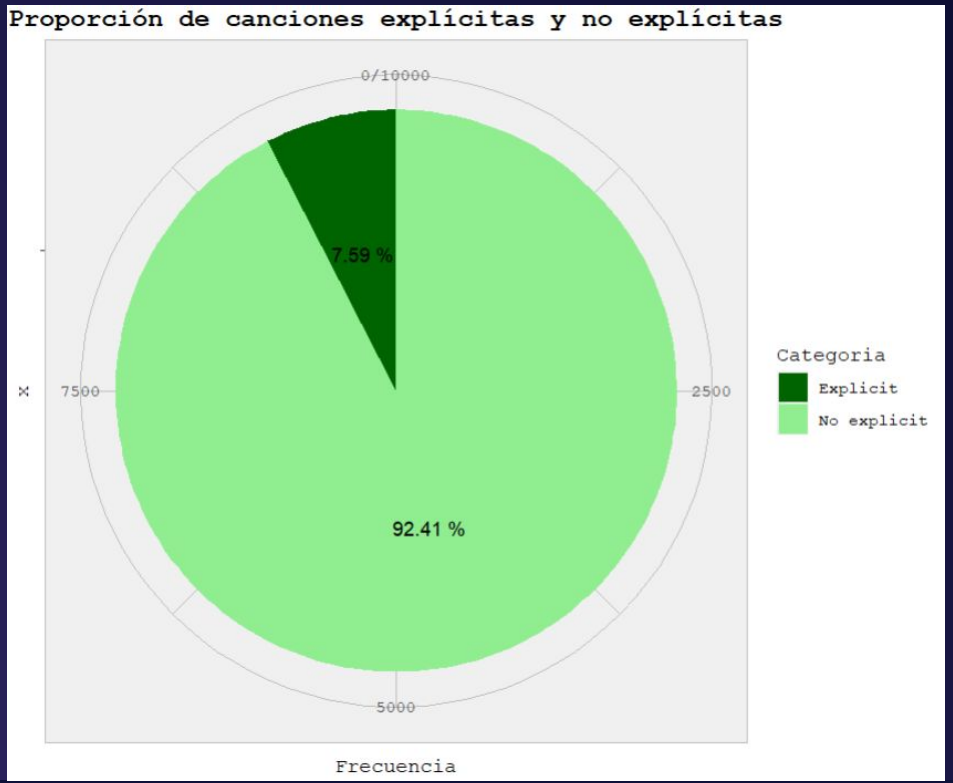
Realizamos un gráfico de correlación de Pearson con las variables numéricas.

Popularity, duration, speechiness y liveness parecen no tener ninguna correlación lineal.



Analizando Explicit

Realizamos un gráfico de tortas para ver la proporción de canciones explícitas y no explícitas en nuestros datos.



Analizando Explicit

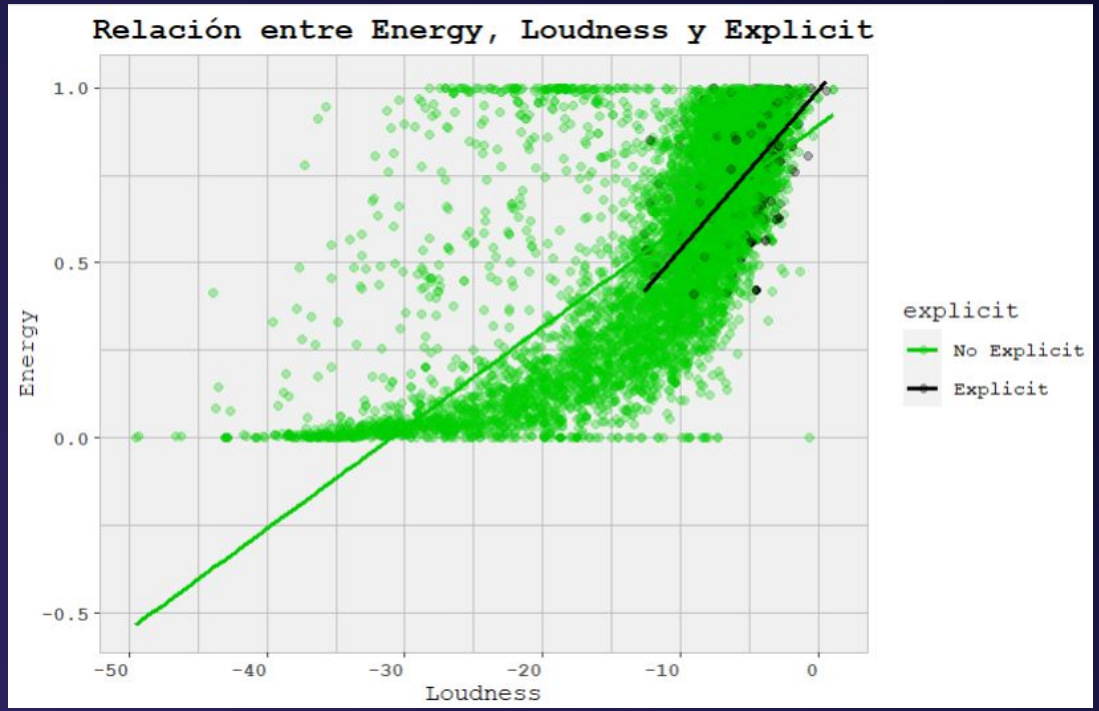
Se realizó un test de comparación de medias de las variables numéricas en los casos de canciones explícitas y no explícitas.

Las únicas variables en las que explicit no tiene inferencia es en popularity y en liveness. En algunas variables, como energy, acousticness y loudness, el impacto que tiene es considerablemente alto.

Variable1	Variable2	p-valor	Límite inferior	Límite superior	Media explicit	Media no explicit
explicit	energy	≈ 0	0.144	0.170	0.740	0.583
	tempo	≈ 0	0.021	0.042	0.574	0.543
	loudness	≈ 0	0.099	0.018	0.087	0.067
	popularity	0.81	-0.021	0.028	0.311	0.308
	speechiness	≈ 0	0.051	0.066	0.141	0.082
	duration_ms	≈ 0	-0.004	0.002	0.048	0.050
	liveness	0.86	-0.011	0.014	0.021	0.208
	valence	≈ 0	0.078	0.112	0.057	0.413
	acousticness	≈ 0	-0.2731	-0.246	0.105	0.364
	danceability	≈ 0	0.140	0.167	0.709	0.555
	instrumentalness	≈ 0	-0.337	-0.313	0.028	0.353

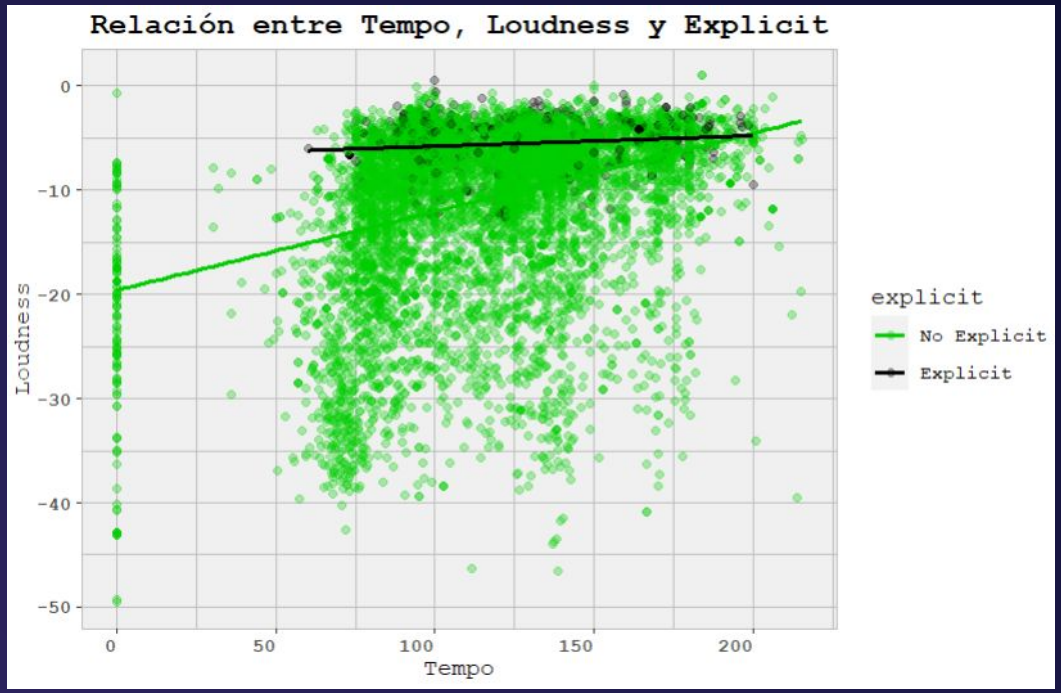
Explicit en relación a otras variables

Las canciones Explícitas son las que tienen mayor Loudness y Energy.



Explicit en relación a otras variables

Las canciones Explícitas son las que tienen mayor Loudness y Tempo mayor a 55.



Asociación con las categóricas y explicit

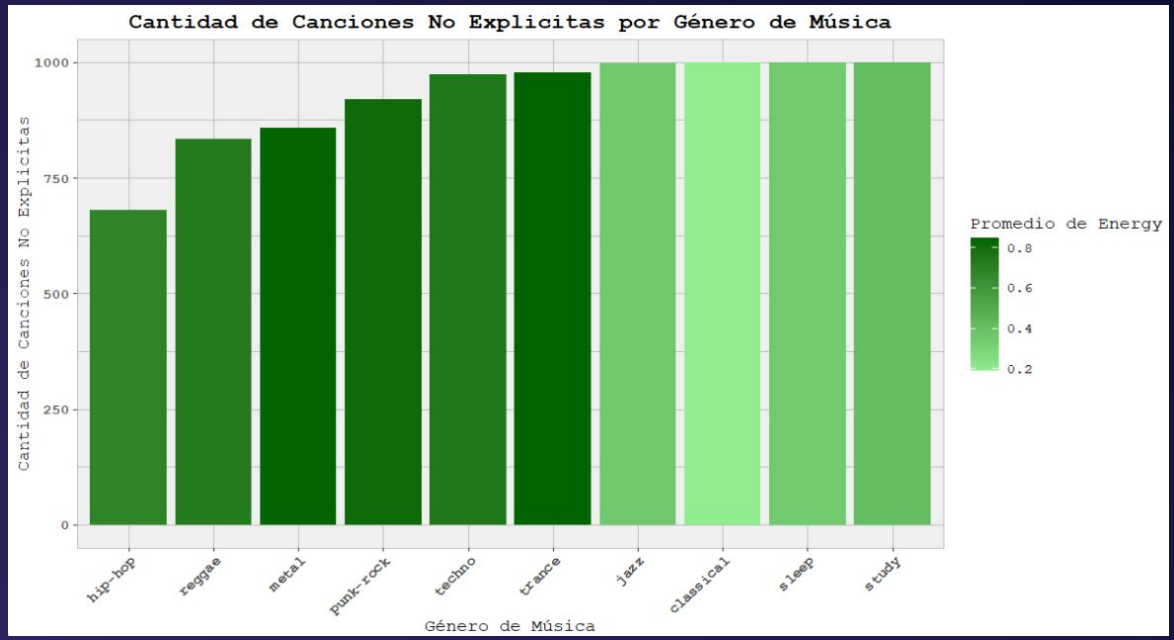
Para medir el nivel de asociación entre las variables categóricas con explicit se utilizó el valor de la V de Cramer.

Las variables que dan cercanas a cero indican una falta de asociación. El género es la única variable que si se asocia con explicit.

Variable1	Variable2	V-Cramer
explicit	key	0.08
	mode	0.02
	time_signature	0.07
	genre	0.38

Explicit en relación a otras variables

Los géneros que menos canciones explícitas tienen son jazz, classical, sleep y study. A su vez, son los que tienen un menor promedio de energy.

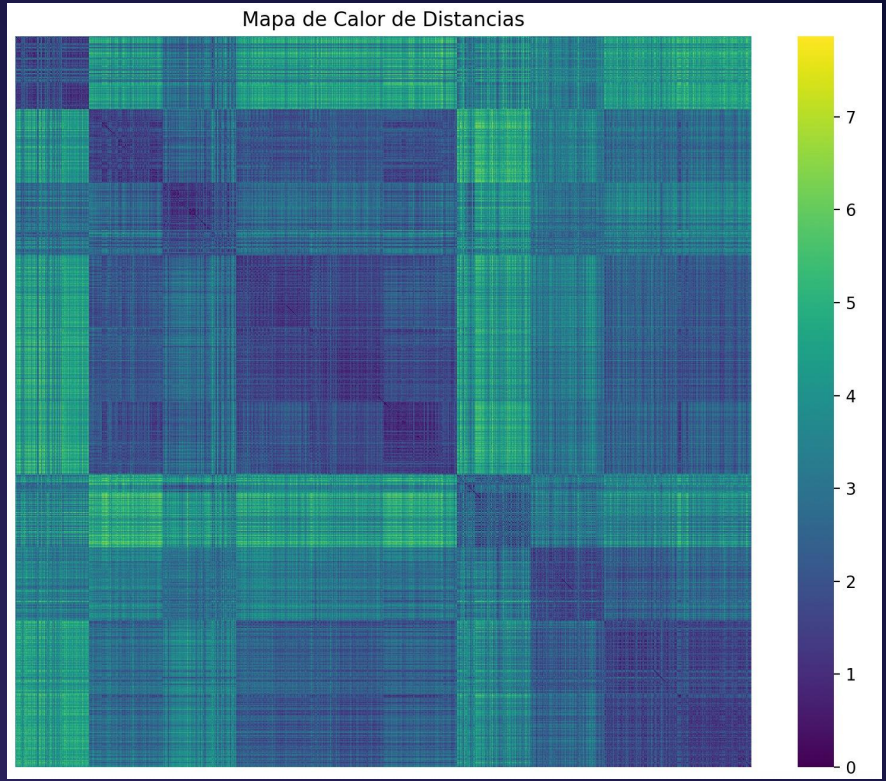


Matriz de distancias

La disposición de la matriz conserva un formato de cuadrícula al estar los datos ordenados en función a su género.

Clustering heatmap muestra la posibilidad de la existencia de diversos grupos de clusters, ya que las regiones más amarillas indican una mayor distancia entre los datos.

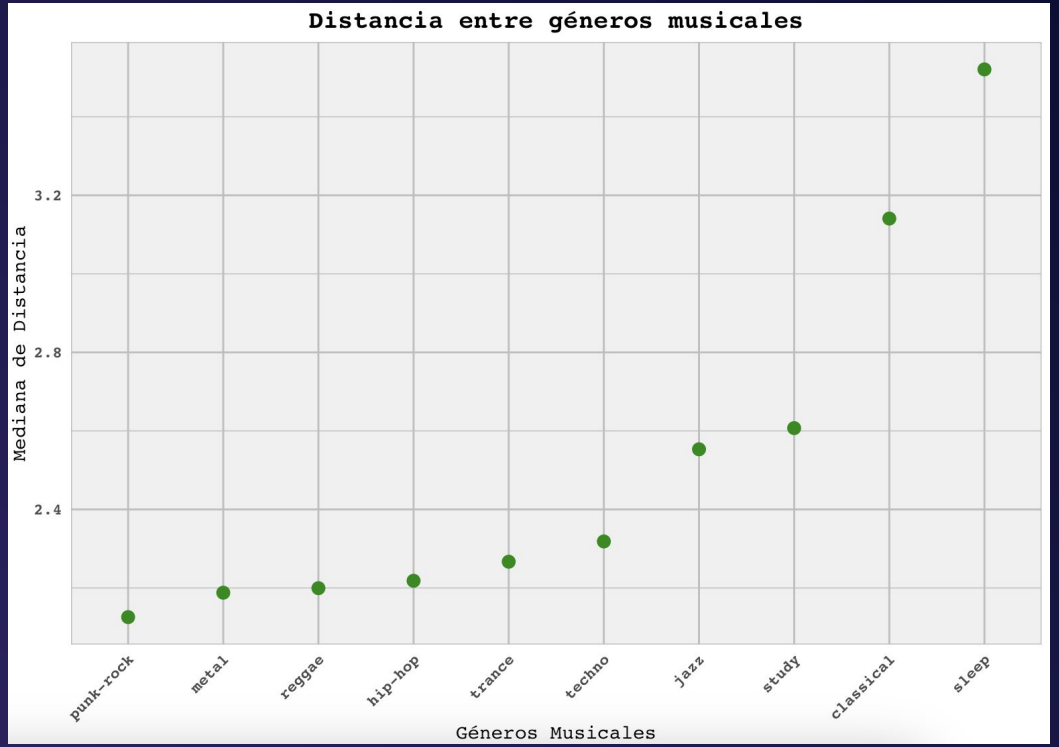
Nota: Dada la dimensión de nuestro dataset filtrado, RStudio ha presentado dificultad de realizar esta visualización. Alternativamente, se ha realizado con Python.



Distancias medianas entre géneros

Podemos identificar que puntualmente los generos classical y sleep se comportan de manera notablemente distinta al resto de los géneros, lo que señala que podrían ser outliers.

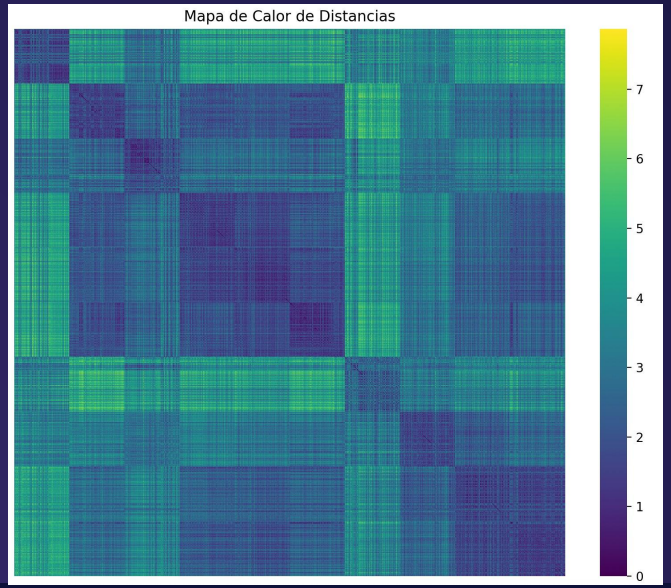
Por eso, decidimos tratar la muestra con y sin ellos en el heat map y de manera diferenciada en el gráfico de coordenadas paralelas a continuación.



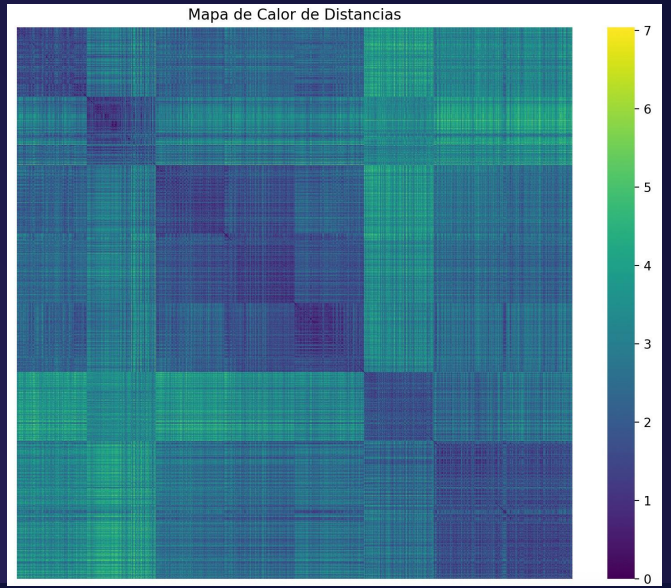
Matriz de distancias

Se observa que, efectivamente las distancias entre las canciones decrecen dado a que hay menor presencia de distancias cuyos valores se asemejan a 7.

Con outliers



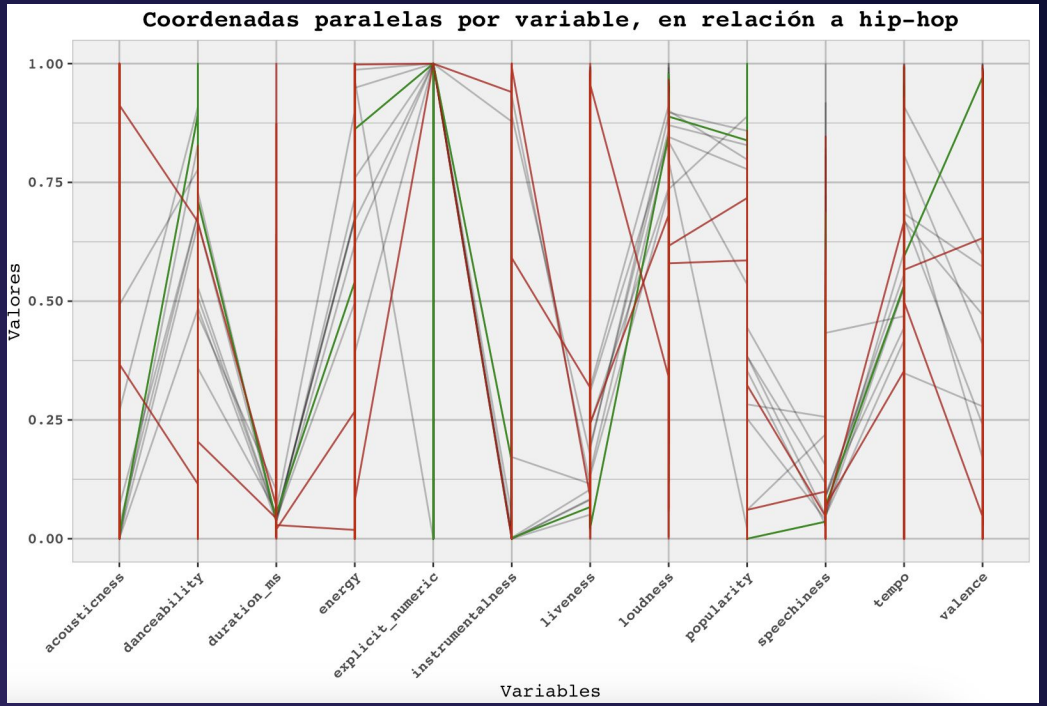
Sin outliers



Coordenadas paralelas por variable

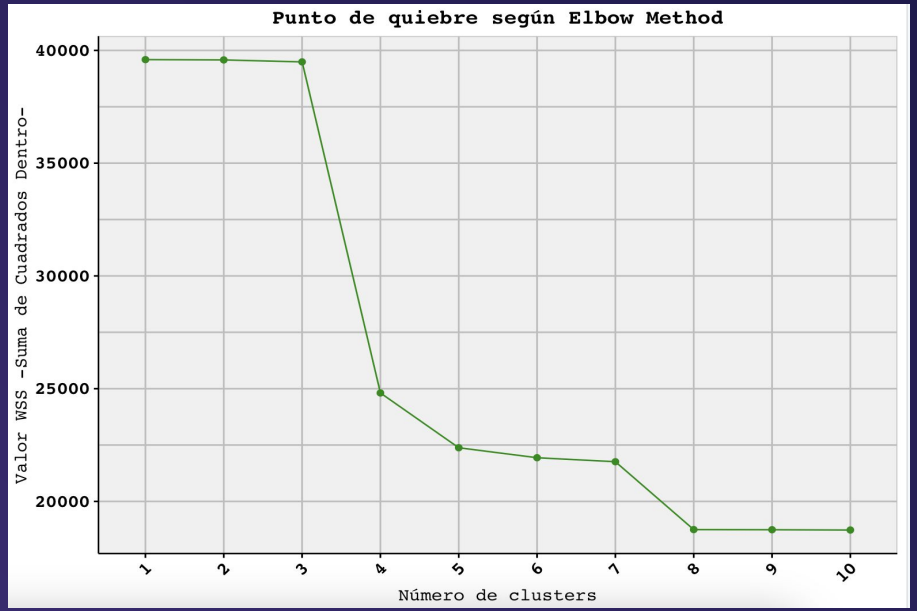
En el siguiente gráfico, a fines de poder hacer un seguimiento de los géneros y sus diferenciaciones, marcamos en color rojo los generos classical y sleep -que conservan mayores diferencias-.

Por otro lado, se marco en color verde al género musical hip-hop, ya que es el que posee mayor proporción de explicit -variable de interés- con un 31,9%.

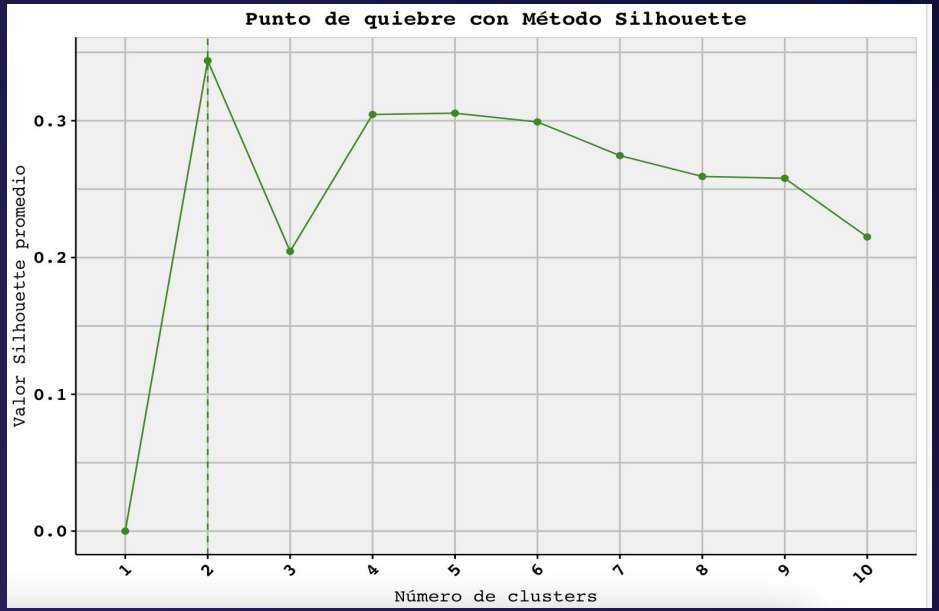


Clusters óptimos

Elbow Method



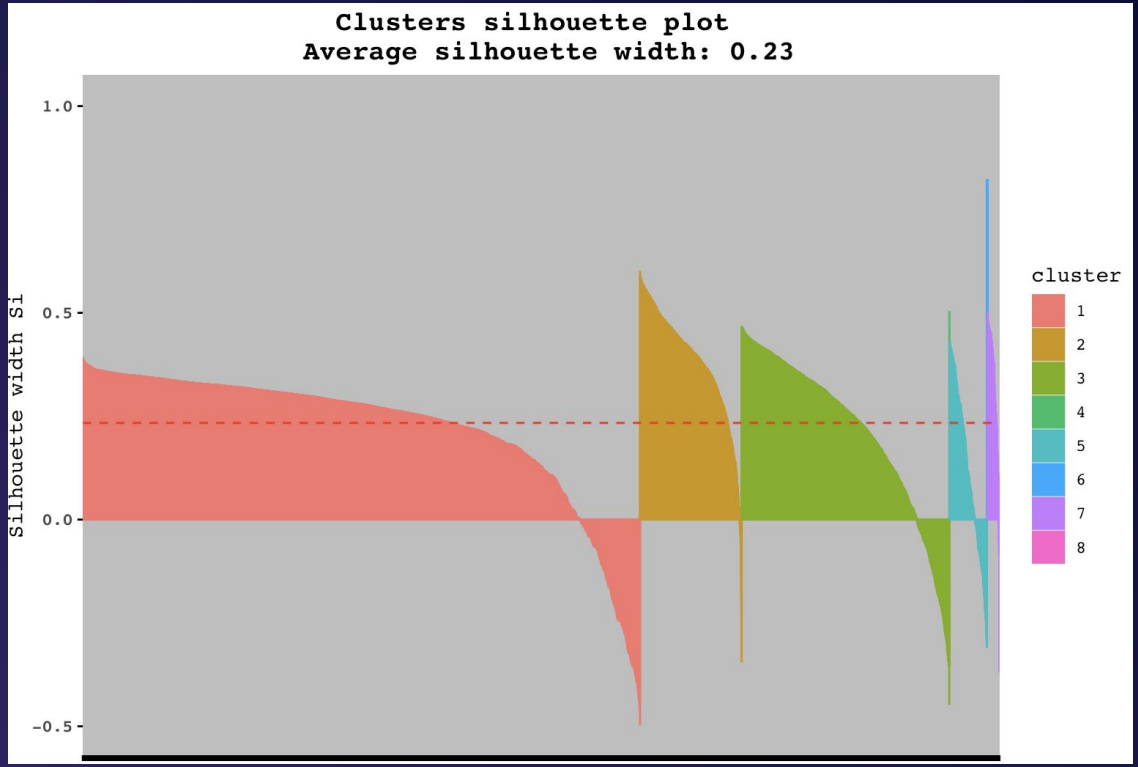
Silhouette

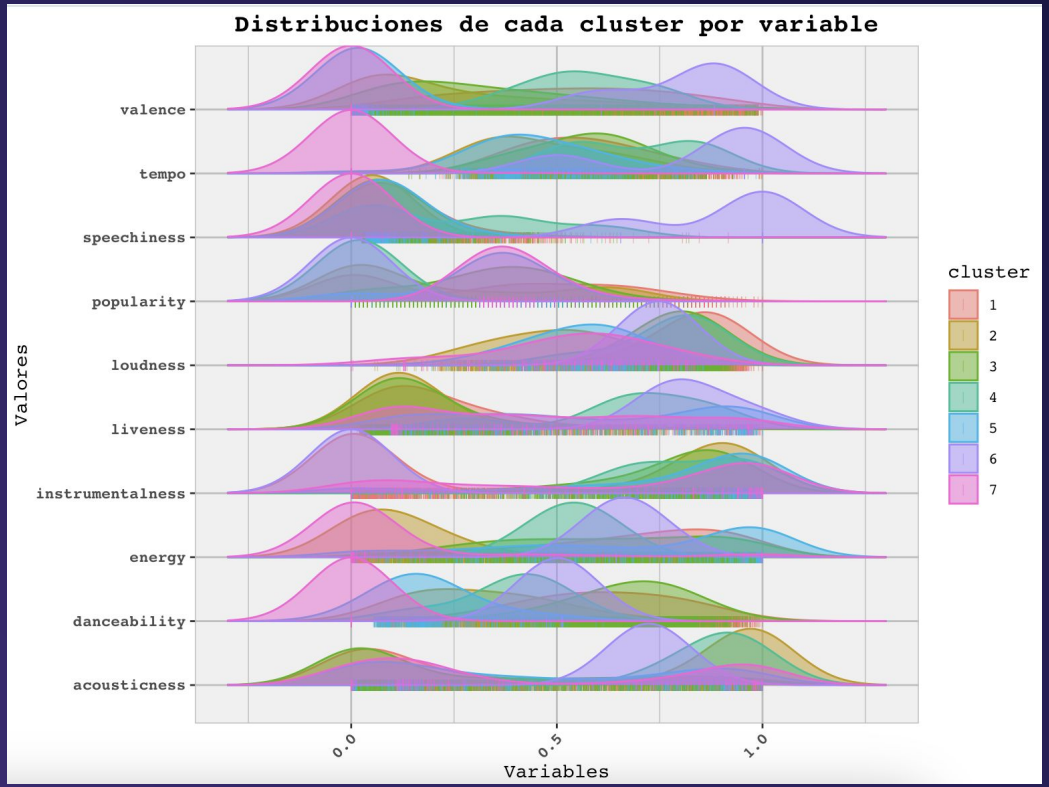


Optamos por utilizar 8 clusters en los cuales se identifican 3 principales. En estos casos, no tenemos una gran confianza en que las canciones que exhiben un valor negativo en el índice Silhouette sean realmente similares al resto del conjunto.

Los clusters 6 y 7 presentan un alto grado de similitud entre las canciones que los componen.

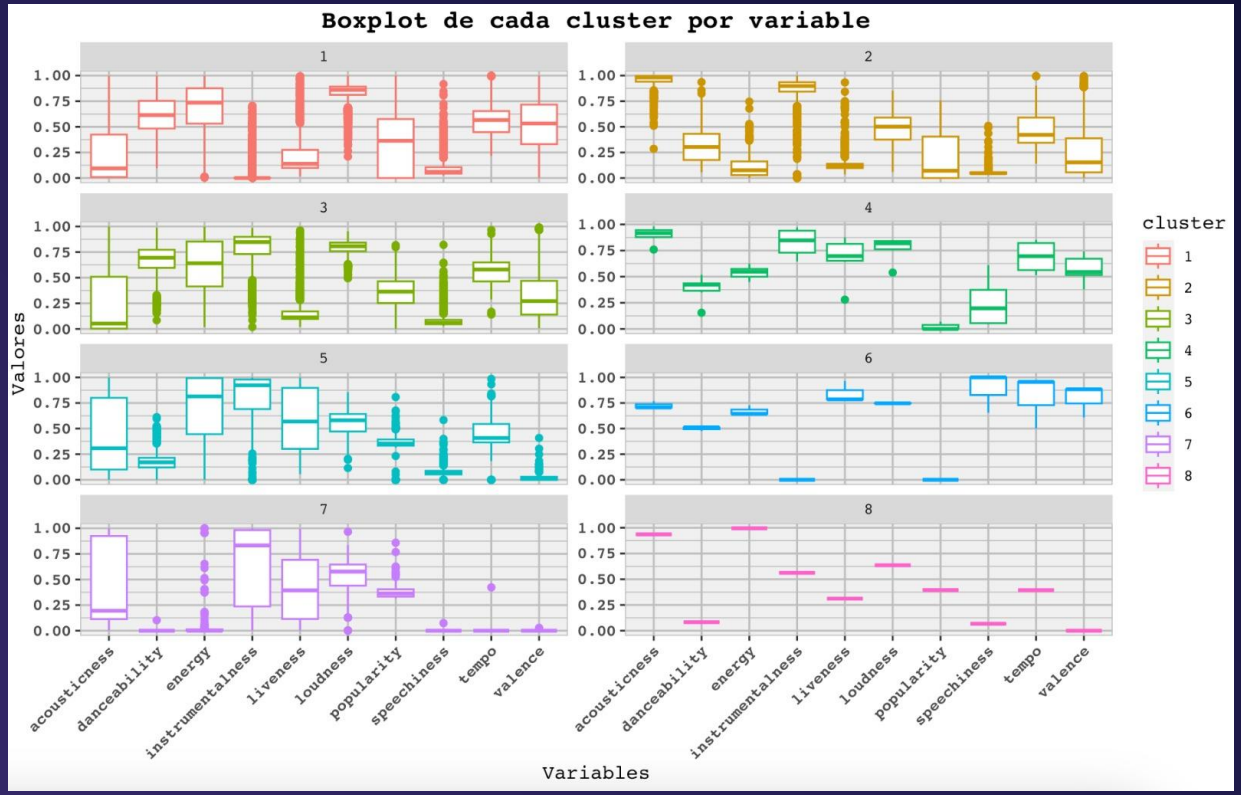
El cluster 8 no se ve porque contiene una sola canción (Ocean Waves Sounds to Relax and Sleep - Ocean Sounds, dura 1,14 horas, es el mayor outlier dentro de la variable duration_ms)





Se observa que las variables musicales en todos los clusters exhiben distribuciones bastante similares en términos generales. Esto sugiere que las características musicales que hemos evaluado tienden a mantener una coherencia relativa y un comportamiento consistente en todos los grupos.

Sin embargo, es interesante notar que existe una excepción importante en este patrón general: la variable 'popularity', lo cual corresponde a su correlación previamente mencionada.


Boxplot de cada cluster por variable
1
2
3
4
5
6
7
8
1.00
0.75
0.50
0.25
0.00
Valores
acousticness
danceability
energy
instrumentalness
liveness
loudness
popularity
speechiness
tempo
valence
Variables
cluster
1
2
3
4
5
6
7
8

Porcentaje de explícitas por cluster

Los clusters 1 y 3 son los que mayor porcentaje de canciones explícitas tienen.

Cluster	Tamaño	Porcentaje Explícito
1	6079	12%
3	2265	1%
2	1109	0%
5	405	0%
7	128	0%
6	7	0%
4	6	0%
8	1	0%

04

Conclusión

Reflexiones finales. Links de Interés.

Reflexiones Finales

- Sólo el 7,59% de las canciones son explícitas, por lo que no habría que filtrar tantas canciones para el armado del Spotify Kids.
- Si bien el género de la canción tiene cierta inferencia en sí la canción es explícita o no, hay otras variables que tienen mayor peso, asi como su loudness o scheechiness.
- Una canción movida, enérgica y bailable tienden más a ser explícitas que una canción tranquila y acústica.
- De los 8 clusters identificados, solo 2 contienen canciones explícitas, lo que sugiere que sería preferible considerar los otros clusters al crear Spotify Kids.

Links de Interés

Repositorio de GitHub

https://github.com/azulamakk/analisisPredictivo/tree/master/Entrega%201

Base de datos -Google Drive-

https://drive.google.com/file/d/112y_xKu2a1d4qpejeFHKB8UrvzLDowjQ/view?usp=sharing

Documentación en Kaggle

https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

Muchas gracias!