Fundamentos de la Ciencia de Datos

Integrantes:

- Kiara Aranda
- Emanuel Esperon
- Luciano Losardo

Fecha de entrega:14/11/2024

Introducción:

En el presente trabajo práctico, se solicitó que se analizará de forma detallada un conjunto de datos sobre covers de la década de los 90 sacados de You Tube Music.
El cual cuenta con 980 canciones de 536 artistas diferentes y que está compuesto de los siguientes valores:

- **Track:** el título de la canción.
- **Artist:** el intérprete o grupo que grabó la canción.
- **Duration:** la duración de la canción, medida en minutos y segundos.
- **Time_Signature:** la métrica musical de la canción, indica el número de pulsaciones por compás.
- **Danceability:** una medida de qué tan adecuada es una pista para bailar, basada en el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general.
- **Energy:** una medida de intensidad y actividad en la canción, donde los valores más altos indican una pista más enérgica.
- **Key:** la tonalidad musical en la que está compuesta la canción, representada por un número entero.
- **Loudness:** el volumen promedio de la canción, medido en decibelios (dB).
- **Mode:** la modalidad de la pista, indica si la canción está en tono mayor o menor.
- **Speechiness:** una medida de la presencia de palabras habladas en una pista, valores más altos indican cualidades más parecidas al habla.
- **Acousticness:** una medida de la calidad acústica de la pista, valores más altos indican una mayor probabilidad de ser acústica.
- **Instrumentalness:** una medida que indica la presencia de voces, valores más altos representan pistas más instrumentales.
- **Liveness:** una medida de la probabilidad de que la pista se haya interpretado en vivo, valores más altos indican más ruido de audiencia.
- **Valence:** Una medida de la positividad musical de la pista, valores más altos indican música más positiva o alegre.

- **Tempo:** la velocidad o ritmo de la pista, medida en pulsaciones por minuto (BPM).
- **Popularity:** una puntuación que refleja la popularidad de la pista, generalmente basada en los recuentos de transmisiones y otras métricas.
- **Year:** el año en que se lanzó la canción.

Sobre el dataset dado se realizó un análisis exploratorio y una limpieza de los datos (link a git o no). Además se plantearon seis hipótesis, de las cuales se analizaron cuatro.

Pasos

Análisis Exploratorio

1- Clasificación de datos: Empezamos viendo los tipos de datos que teníamos en el dataset para hacer una primera toma de contacto con el dominio de análisis.

- Track: cualitativa nominal.
- Artist: cualitativa nominal.
- Duration: cuantitativa continua.
- Time_signature: cuantitativa discreta.
- Danceability: cuantitativa continua.
- Energy: cuantitativa continua.
- Key: cuantitativa discreta.
- Loudness: cuantitativa continua.
- Mode: cualitativa nominal.
- Speechiness: cuantitativa continua.
- Acousticness cuantitativa continua.
- Instrumentalness: cuantitativa continua.
- Liviness: cuantitativa continua.
- Valence: cuantitativa continua.
- Tempo: cuantitativa continua.
- Popularity: cuantitativa discreta.
- Year: cuantitativa discreta.

2 - Limpieza de datos

- Se analizó el dataset para ver si se encontraban valores atípicos (nulos, NaNs, valores fuera de rango, variables tipo object, Strings que sean Unknown, etc). De los posibles valores atípicos que se mencionaron, solo se encontró un valor fuera de rango para la variable Loudness (loudness positivo), pero al ser único se consideró que no era necesario eliminar la observación por no perder el resto de variables de la misma.
- Se buscaron líneas repetidas en el dataset. Si bien no se encontraron líneas repetidas en sí, observamos que algunas eran casi idénticas, difiriendo sólo en algunos valores. De estas, eliminamos aquellas que variaban en el año de lanzamiento.

	Track	Artist	Duration	Time_Signature	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Popularity	Year
397	Again	Janet Jackson	3:13	4	0.664	0.738	9	-4.095	0	0.0338	0.188000	0.000058	0.1350	0.2030	119.973	76	1994
300	Again	Janet Jackson	3:13	4	0.664	0.738	9	-4.095	0	0.0338	0.188000	0.000058	0.1350	0.2030	119.973	76	1993
5	Alright	Janet Jackson	3:39	4	0.796	0.766	7	-5.974	1	0.2380	0.074200	0.000000	0.0827	0.5580	110.034	80	1990
296	Alright	Kris Kross	3:39	4	0.796	0.766	7	-5.974	1	0.2380	0.074200	0.000000	0.0827	0.5580	110.034	80	1993

  De los dos casos que vemos en este ejemplo, eliminamos pares como el primero.
- Para un mejor manejo de los datos, convertimos la columna de Duration a segundos y cambiamos su tipo de dato de object a entero.

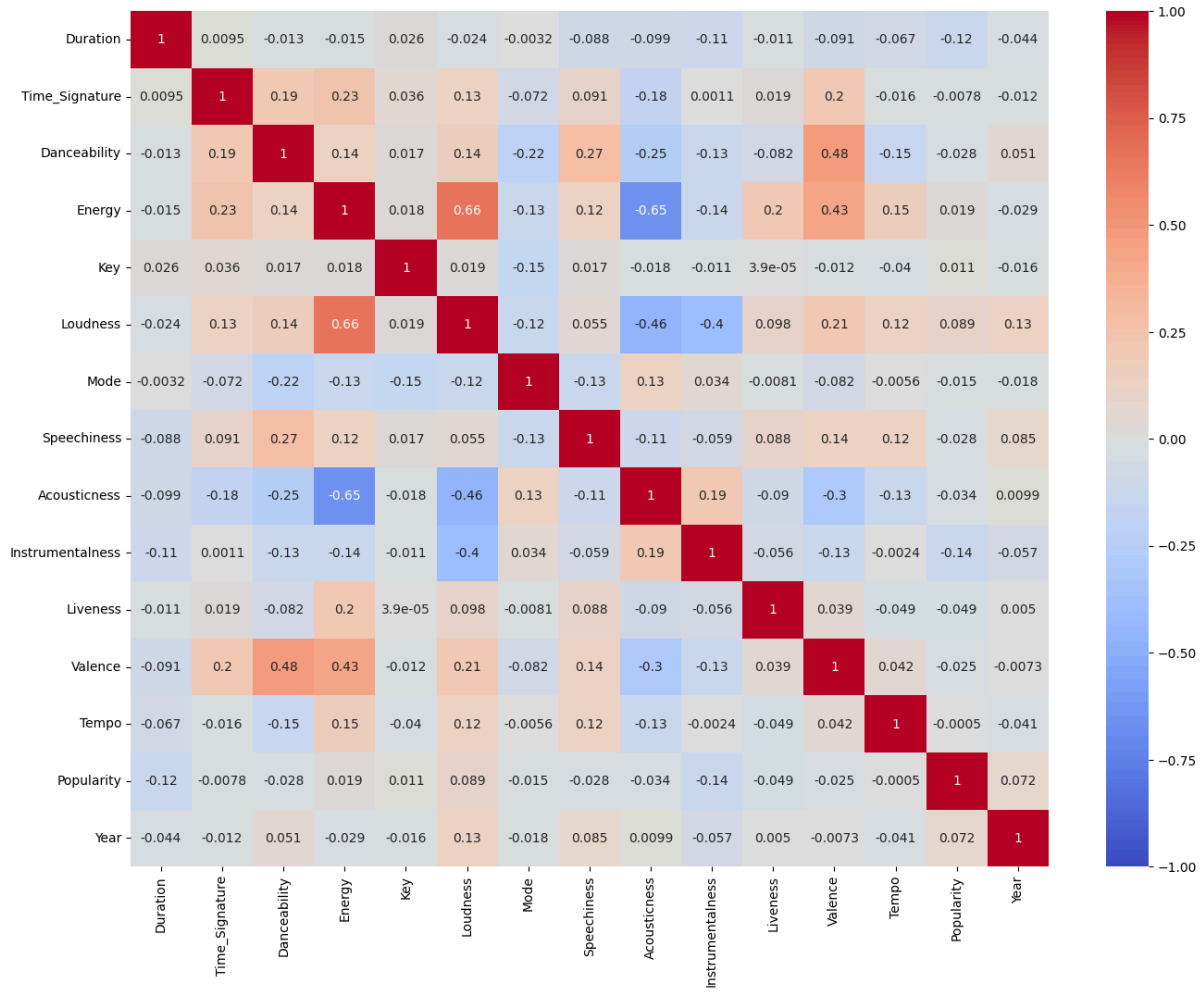
Aclaración: si bien no se eliminó el año de lanzamiento, no se usó en análisis subsecuentes.

3 - Análisis univariado

Se realizaron diversos gráficos para analizar la distribución de las variables y su comportamiento individual. Principalmente, se buscó hallar distribuciones normales.
Se le dio especial importancia a la variable Popularity por considerarla una variable de interés.

4 - Análisis bivariado

Se realizó un heatmap con la matriz de correlaciones. Si bien no se encontró ningún valor de correlación superior a 0.7 o menor a -0.7, posteriormente se prestó especial atención a aquellas variables con un coeficiente de correlación relativamente alto. También se hicieron scatterplots para ver la tendencia de la correlación entre las variables con un coeficiente más alto. Viendo estos gráficos notamos que la variable Energy tiene una correlación alta con varias variables, lo cual se tuvo en cuenta para análisis futuros.



5 - Análisis multivariado

En base a lo que se vio en el punto anterior, se decidió hacer un análisis para estudiar si Energy tiene una relación no lineal con el resto de variables. Este análisis se hizo también con la variable Popularity debido a que se consideró relevante dada la temática del dataset, **COVERS DE UNA DÉCADA ICÓNICA DEL SIGLO PASADO (YouTube Music).**

En el caso de Energy, se vio una tendencia creciente de derecha a izquierda. Esto permitió corroborar que estaba relacionada con otras variables. Para Popularity, no se pudo observar ningún comportamiento específico en los diagramas realizados.

Los resultados de estos análisis se tomaron en cuenta para armar las hipótesis.

<u>Clusters:</u>

Primero, se realizó un elbow plot para decidir la cantidad de clusters que mejor se ajusta a nuestros datos. 8, 5 y 3 se consideraron opciones viables, pero nos decantamos por 3 ya que se veía una mejor separación de los clusters. Una vez hecho esto, notamos que los dos grupos más grandes se

diferenciaban principalmente por su modalidad. Esto también fue utilizado a la hora de plantear las hipótesis.

6- Hipótesis planteadas y justificación de porqué se planteó:

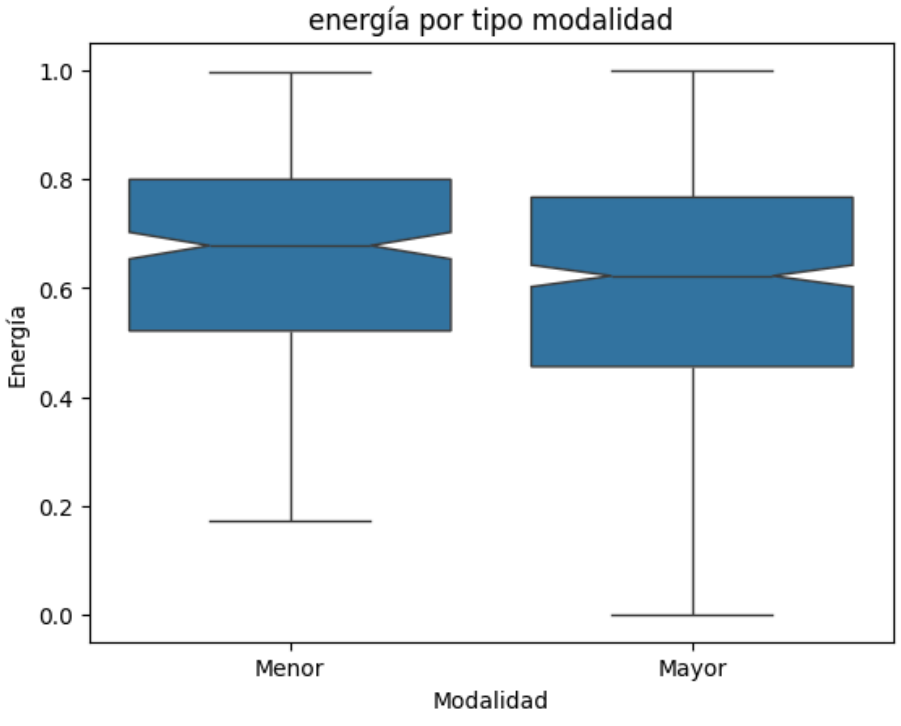
h1. ¿Las canciones que son de modalidad mayor tienen una energía más fuerte?

Justificación: dado que la modalidad mayor se caracteriza por tener un sonido alegre y luminoso, mientras que el modo menor transmite tristeza y melancolía, y que por lo general las pistas enérgicas se sienten rápidas, fuertes y ruidosas, se pensó en la posibilidad de que entre las modalidades la energía sea diferente.

Análisis de hipótesis:

Ho = no existen diferencias entre los dos grupos
H1 = existen diferencias entre los dos grupos
nivel de significancia: 0.05
normal (Shapiro): no
    modalidad baja p = 0
    modalidad alta p = 0.001
homocedasticidad: no (Levene)
    p=0.001

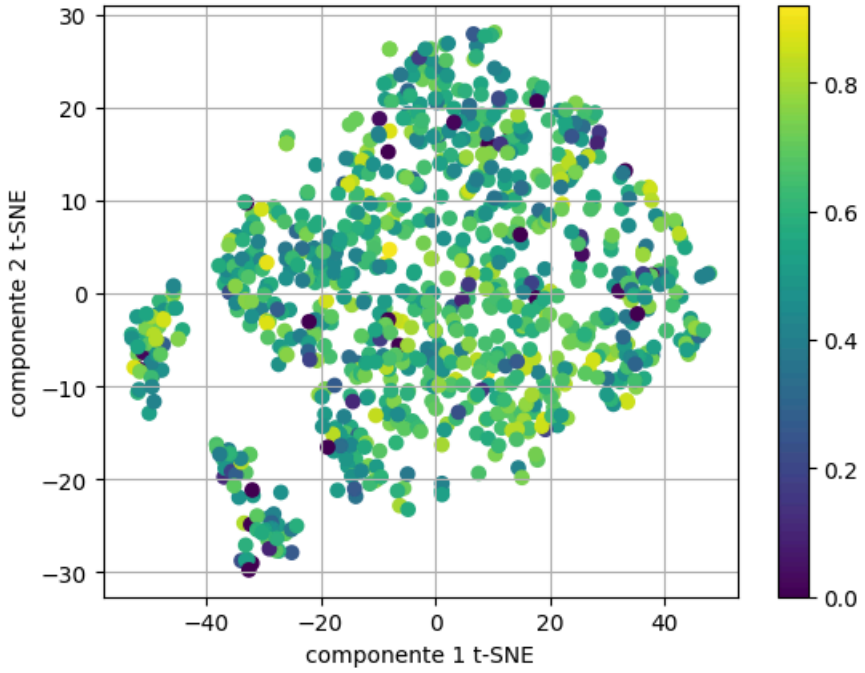
Dado que los datos no son normales y no cumplen con la homocedasticidad, se aplicó un test Kruskal-Wallis. El resultado fue un valor de p = 0.001, rechazando la hipótesis nula.
Entonces existe una diferencia entre ambas modalidades pero si vemos el siguiente gráfico:



La hipótesis de que el valor de energía es mayor si el modo es mayor es falsa.

h2. ¿La popularidad entre las dos modalidades son iguales ?

Justificación:

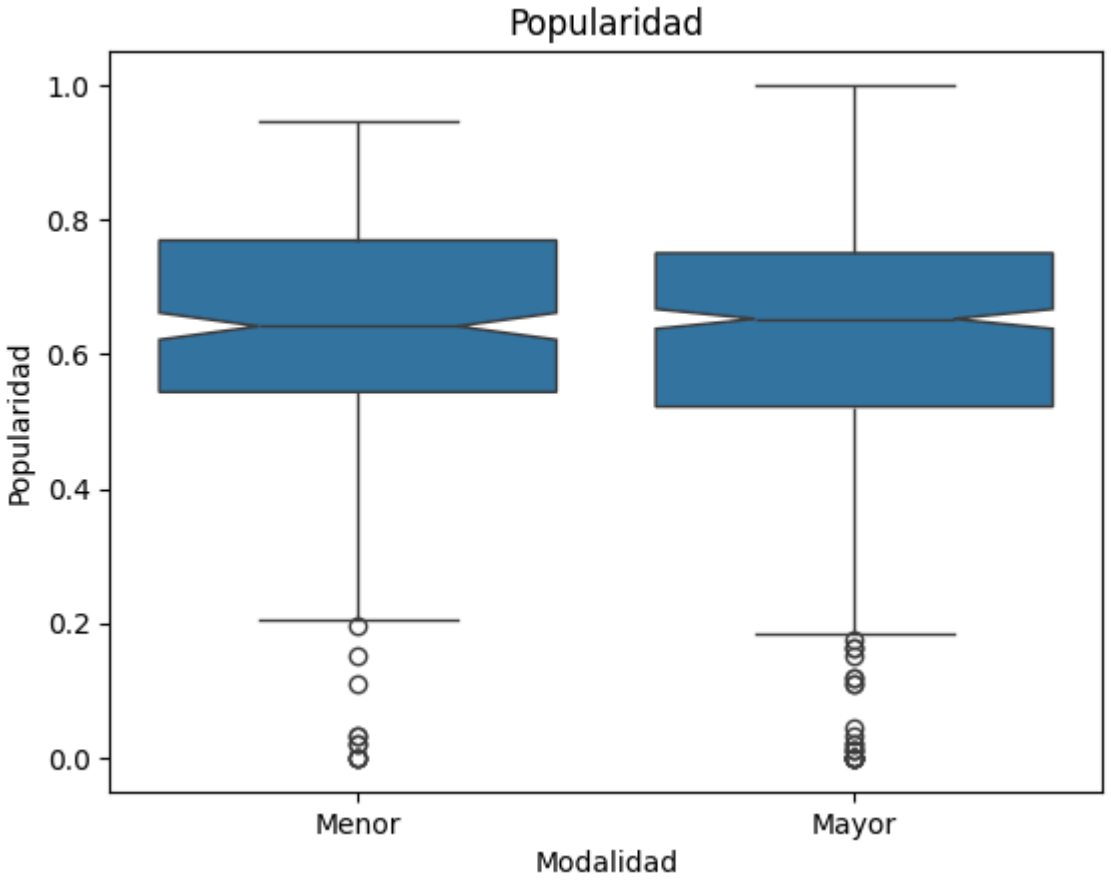


Dado que a la hora de hacer el análisis sobre la popularidad en el dataset no se observó ningún comportamiento que nos indicara algún agrupamiento en torno a la popularidad.
Debido a que no encontramos información relevante sobre la popularidad, ya sea comportamiento con respecto a la forma en la que se calcula, solo nos basamos en las observaciones dichas anteriormente.

Ho = no existen diferencias entre los dos grupos
H1 = existen diferencias entre los dos grupos
nivel de significancia: 0.05
normal (Shapiro): no
    modalidad baja p = 0
    modalidad alta p = 0
homocedasticidad: si (Levene)
    p=0.572

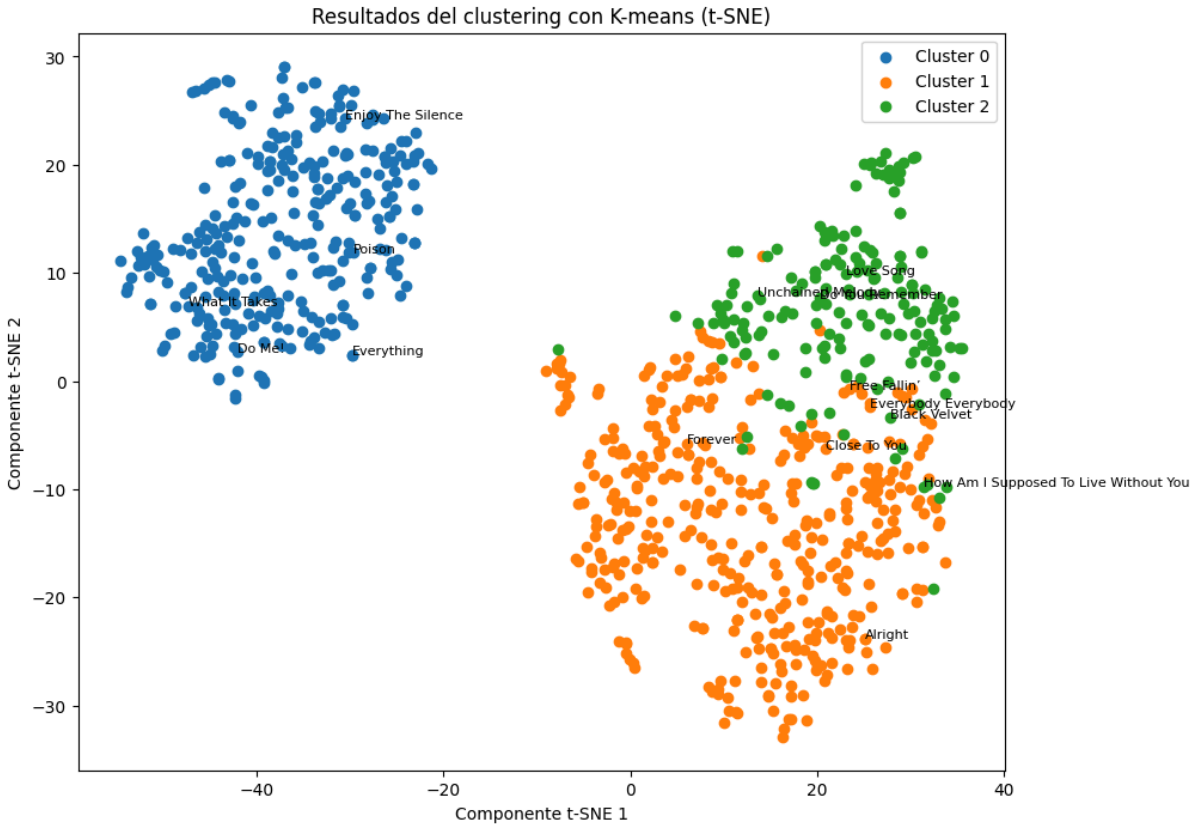
Dado que los datos no son normales y si son homocedásticos, se aplicó un test de Mann-Whitney. El resultado fue de p =0.827, lo cual no permite rechazar la hipótesis nula.

Entonces no existe una diferencia significativa en la popularidad de las canciones con modalidad alta y baja.



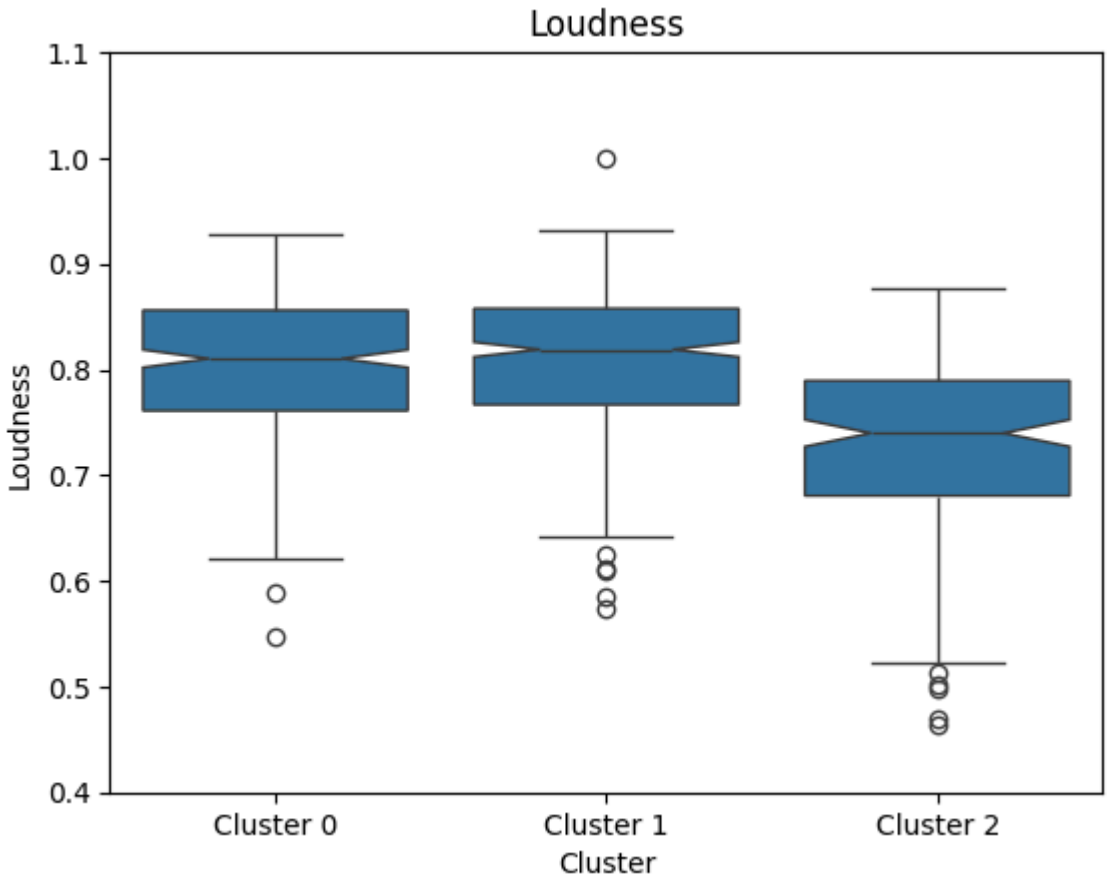
h3. una tonalidad mayor puede determinar una energía mayor

Justificación: dado que la tonalidad y el modo están estrechamente relacionados a la hora de describir la armonía de una canción, es posible intuir que los valores de energía entre tonalidades distintas va a ser distintas.



Ho = no existen diferencias entre los dos grupos
H1 = existen diferencias entre los dos grupos
nivel de significancia: 0.05
normal (Shapiro): no
    modalidad baja p = 0
    modalidad alta p = 0
homocedasticidad: no (Levene)
    p = 0

Dado que los datos no son normales y no son homocedásticos, se aplicó un test de Kruskal-wallis. El resultado fue de p =0 , lo cual permite rechazar la hipótesis nula.
Entonces existe una diferencia significativa en el loudness de las canciones pertenecientes al cluster 1 y 2.
Además podemos decir que las canciones del cluster 1 tienen un loudness mayor al cluster 2.

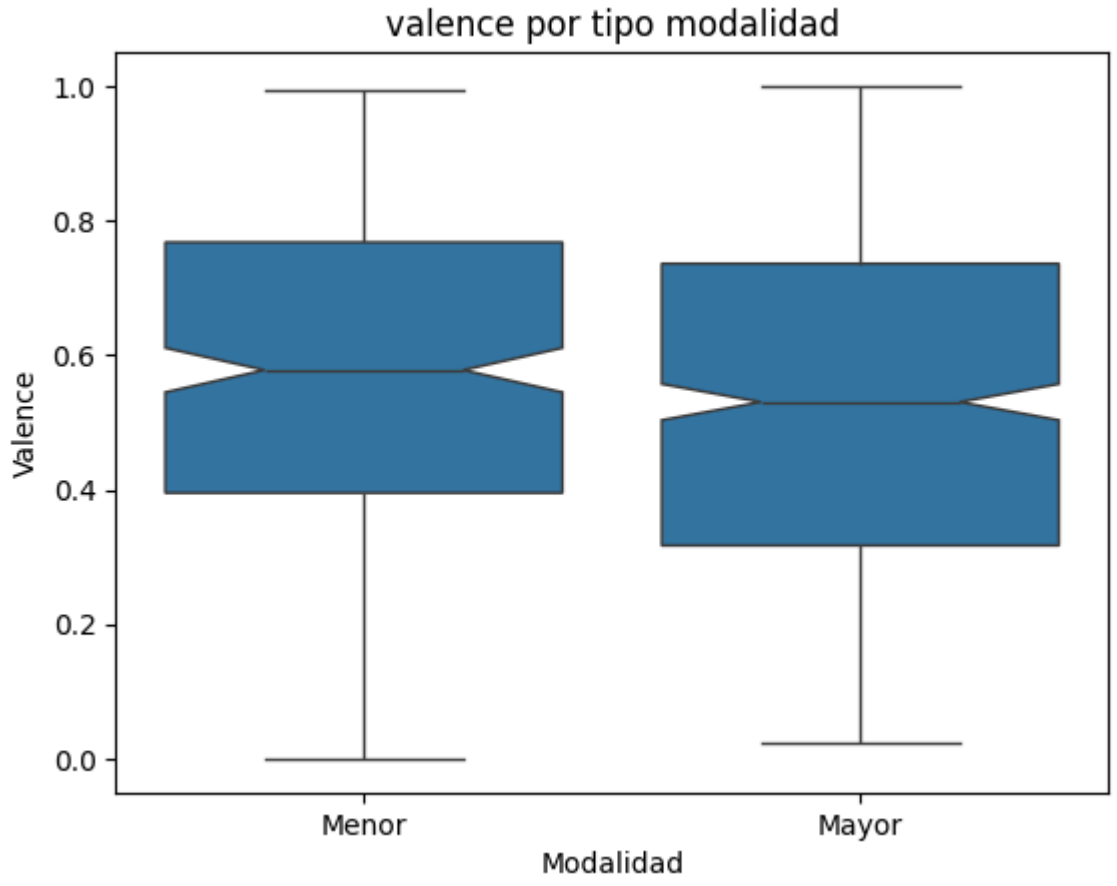

Loudness
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Loudness
Cluster 0
Cluster 1
Cluster 2
Cluster

h4. Las canciones de modalidad mayor se asocian a Valence mayor

Justificación: dado que para las canciones con modalidad mayor tienden a asociarse con emociones más alegres y que valence mide la positividad de una canción.

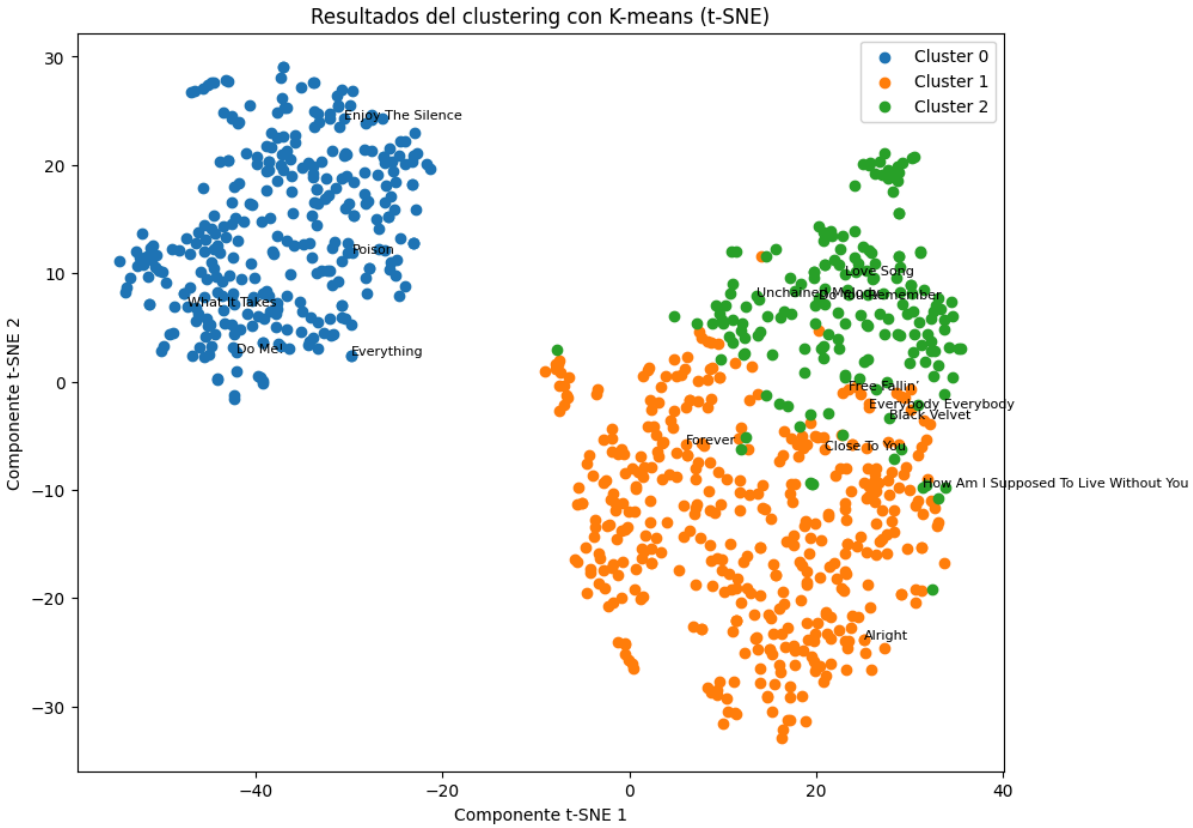
Ho = no existen diferencias entre los dos grupos
H1 = existen diferencias entre los dos grupos
nivel de significancia: 0.05
normal (Shapiro): no
    modalidad baja p = 0
    modalidad alta p = 0
homocedasticidad: ni (Levene)
    p=0.078

Dado que los datos no son normales y no son homocedásticos, se aplicó un test de Kruskal-wallis. El resultado fue de p =0.016, lo cual permite rechazar la hipótesis nula.
Entonces existe una diferencia significativa en el Valence de las canciones con modalidad alta y baja.



h5. las canciones del cluster 1 tienen un loudness mayor que las del cluster 2

Justificación: dentro de los temas con un modo mayor se visualizaba una separación entre las mismas de los datos que vimos, la variación en loudness era la más notable, lo cual llevó a plantear hipótesis.



h6. ¿las canciones con energy y valence altas son más bailables?

Justificación: sabiendo que la valencia describe la positividad y la energía la intensidad de una canción, supusimos que entre más altos estos valores, es más probable que las canciones sean más bailables.

bibliografía:

https://www.undiaunacancion.es/como-funciona-el-modo-en-la-musica/

Referencia de API web | Spotify para desarrolladores

https://es.linkedin.com/pulse/qu%C3%A9-es-el-porcentaje-de-popularidad-micaela-blue-bo2xf

Music extractor — Essentia 2.1-beta6-dev documentation

https://composicionmusical.es/la-tonalidad-musical/