87.17 Análisis Multivariado

Comisión A

Segundo Trabajo Práctico

Azul de los Ángeles Makk

Paula Ariana González

Cosatto Ammann, Pedro Camilo

Índice

1. Introducción

El presente trabajo se centra en el análisis de una base de datos obtenida de Spotify, la cual abarca una amplia gama de géneros musicales y describe las características asociadas a diversos géneros. El objetivo principal de este trabajo radica en la aplicación del Escalamiento Multidimensional Métrico (MDS) para visualizar en baja dimensión la similitud entre las canciones de Spotify, permitiendo la identificación de patrones y estructuras subyacentes en estos datos complejos.

Este enfoque se dirige específicamente a la comparación de similitudes entre las observaciones, apuntando a proporcionar una visualización precisa que permita identificar qué canciones tienen características musicales similares. La aplicación del MDS se vuelve crucial para esta tarea, ya que este método transforma las similitudes originales en un espacio bidimensional o tridimensional, reflejando las distancias entre las observaciones de manera fiel.

2. Base de datos

La base de datos analizada fue tomada del sitio Kaggle, el cual refleja registros que varían en 125 géneros y posee las características musicales que se le asocian. Tal originalmente posee 114.000 registros y 21 columnas. A fines de realizar un análisis sobre un mix de variables numéricas y categóricas, se conservaron las siguientes variables:

- Popularity: La popularidad de un registro es un valor entre 0 y 100, siendo 100 el más popular. La popularidad es calculada por un algoritmo que está basado en mayor parte en la cantidad de reproducciones y que tan recientes son.
- Danceability: Describe qué tan acorde es una canción para bailar acorde a una combinación de elementos musicales que incluyen tempo, estabilidad del ritmo, estabilidad del ritmo y regularidad. Un valor de 0 indica que no es bailable y de 1 que es totalmente bailable.
- Energy: Medida que varía entre 0 y 1 y representa la medida perceptual de la intensidad y actividad. Típicamente, una canción energética se siente rápida y fuerte, como podría ser un registro de death metal mientras que la música clásica se ubica bajo en esa escala.
- Mode: Indica la modalidad (mayor o menor) de un registro. El tipo de escala de la cual el contenido melódico deriva. Mayor es representado por 1 y menor por 0.
- Time_signature: Compás estimado. El compás es una conversión notarial que especifica cuántos pulsos hay en una barra. Un compás varía desde 3 a 7 indicando los compases ¾ y 7/4.

3. Alcance y objetivo

El objetivo del presente informe es producir una visualización de la selección de las canciones de Spotify en baja dimensión, mediante un escalado multidimensional. El escalado multidimensional es una técnica que transforma datos de similitud en un espacio bidimensional o tridimensional de manera que las distancias en ese espacio reflejan las similitudes originales entre los datos. Se utiliza para visualizar datos complejos y encontrar patrones o estructuras ocultas en ellos.

Se opta por analizar las observaciones en lugar de las características que las definen, ya que es valioso identificar qué

canciones son más parecidas entre sí. Esto puede sugerir cuáles canciones podrían ser incluidas en una lista de reproducción, como la función de *radio* de Spotify, que agrupa canciones similares en función de su estilo. Esta característica es ampliamente apreciada por los usuarios, ya que les permite descubrir y disfrutar de música afín a sus gustos musicales.

Asimismo, reconocer cuáles canciones de Spotify son similares entre sí permite brindar a los usuarios recomendaciones precisas de canciones y artistas basadas en sus preferencias musicales, lo que mejora la experiencia de escucha. También, permite que los usuarios puedan descubrir canciones y artistas nuevos que se ajustan a sus gustos al explorar canciones similares a las que ya les gustan. Por último, ayuda a la industria musical y a los profesionales a identificar tendencias y similitudes en la música popular, lo que puede influir en las decisiones de marketing y producción.

4. Definición de una medida de proximidad

Como el objetivo del presente trabajo es realizar un mapeo de las similitudes entre los registros del dataset de Spotify, y las variables son de dos tipos, se definieron dos medidas de proximidad para cada tipo de variable. La métrica utilizada para medir estas distancias combina la distancia euclidiana para variables continuas y la distancia de Gower para variables categóricas, ponderadas por un vector de pesos.

La distancia de Gower es una medida de distancia para datos mixtos que combina las distancias euclidianas para variables numéricas y distancias de igualdad (0 si dos categorías son iguales, 1 en caso contrario) para variables categóricas. Para ilustrar en un formato académico, la fórmula para calcular la distancia de Gower entre dos observaciones i y j en una variable k se define de la siguiente manera:

Si $x^{(k)}$ *es numérica*

$$S_{ij}^{(k)} = \frac{\left|x_i^{(k)} - x_j^{(k)}\right|}{R^{(k)}}$$

Si $x^{(k)}$ *es categórica*

$$S_{ij}^{(k)} = 1\left(x_i^{(k)} = x_j^{(k)}\right)$$

Donde:

- $S_{ij}^{(k)}$ es la distancia de las observaciones i y j en la variable k
- $x_i^{(k)}$ y $x_j^{(k)}$ son las variables de observaciones i y j respectivamente en la variable k
- $1(.)$ es una función indicadora que devuelve 1 si las categorías son iguales y 0 en caso contrario.
- $R^{(k)}$ es el rango (diferencia entre el valor máximo y mínimo) de la variable k

En cada iteración, calcula la distancia entre un par de observaciones (representadas por índices i y j) utilizando la fórmula de distancia de Gower ponderada. Este cálculo se realiza para cada variable en el conjunto de datos, considerando si es una variable numérica o categórica.

Las distancias euclidianas entre valores continuos pueden ser interpretadas de manera análoga a un PCA en términos de encontrar la distancia o similitud entre observaciones en un espacio multidimensional. En esencia, el cálculo de distancias euclidianas entre variables continuas refleja la magnitud de las diferencias en múltiples dimensiones, de manera similar a cómo el PCA busca reducir estas dimensiones a componentes principales que maximizan la varianza.

5. Escalado Multidimensional

5.1. MDS vs. PCA

El método del escalamiento multidimensional (MDS) parte de una matriz de proximidades, que pueden ser similaridades, disimilaridades o distancias, de un conjunto de observaciones. Como el objetivo del MDS es reducir el espacio de dimensiones, se busca transformar la matriz de proximidades en una matriz de disimilaridades para poder reescalar los datos conservando las distancias de los puntos originales. Es especialmente útil en la visualización de datos de similitud, como matrices de distancia o similitud.

Esta técnica se diferencia con el PCA, ya que este análisis de componentes principales parte de los datos originales y transforma las variables en un nuevo conjunto de variables no correlacionadas llamadas "componentes principales" usando las varianzas de los datos. Busca

identificar la combinación lineal de las variables originales que explican la mayor parte de la variabilidad.

Por estas razones, hace que el MDS sea especialmente útil para visualizar datos de similitud, como matrices de distancia o similitud, mientras que PCA es más adecuado para reducir la complejidad de los datos al conservar la información en términos de varianza.

5.2. Graficando MCA en dos dimensiones

Se realiza un escalado multidimensional con los resultados arrojados por la función *delta_nuestra* y, producto de ello, se obtiene el gráfico de la *Figura 5.2.1*.

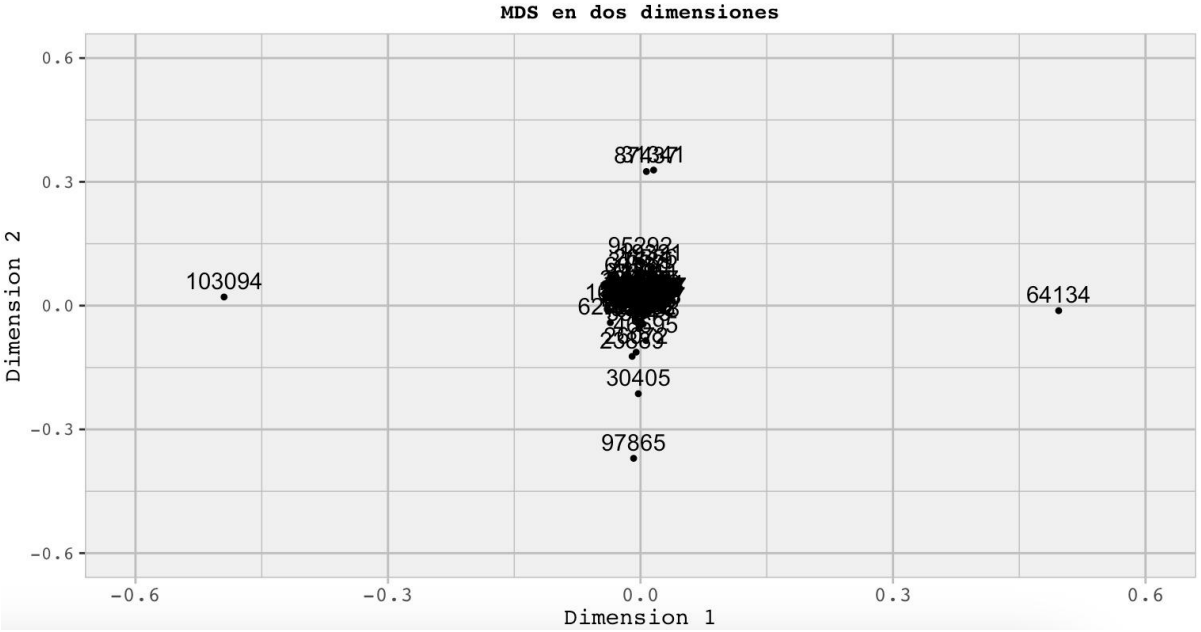

Figura 5.2.1

Para empezar, es evidente la presencia de cinco valores atípicos significativos identificados como "103094", "97865", "64134", “87437” y “31341”. La existencia de estos outliers causa una ampliación de la escala del gráfico, lo que resulta en una visualización más distante del gráfico en su conjunto. Esto, a su vez, hace que los puntos en el centro del gráfico sean difíciles de identificar o de distinguir claramente.

Para un análisis más detallado de las características de los registros atípicos, se crearon dos tablas: *Tabla 5.2.1* para las variables numéricas contínuas y *Tabla 5.2.2* para las variables numéricas categóricas. En la fila "Media" de la *Tabla 5.2.1*, se proporciona el

promedio de las variables. Este valor de promedio sirve como punto de referencia para las variables y facilita la evaluación de si los registros atípicos están significativamente alejados de estas medias. Esto, a su vez, ayuda a entender su posición en el gráfico MDS. En la fila de "Porcentaje del total" de la *Tabla 5.2.2*, se indica el porcentaje que representa la frecuencia de ocurrencia en su respectiva categoría. Este valor tiene el mismo propósito que el promedio y ayuda a contextualizar la frecuencia relativa de los outliers en sus categorías respectivas.

	popularity	**energy**	**danceability**
Media	34.01	0.668	0.562
97865	45	0.564	0.788
64134	0	0.368	0.401
87437	45	0.555	0.766
31341	0	0.945	0.788
103094	0	0.377	0.406

Tabla 5.2.1

	mode	**time_signature**
Porcentaje del total	68%	91%
97865	1	4
64134	1	4
87437	1	4
31341	1	4
103094	1	4

Tabla 5.2.2

Echando un vistazo en la *Tabla 5.2.2*, es posible identificar que los cinco valores atípicos en la variable "mode" tienen un valor de 1 en el 68% de las instancias en el conjunto de datos. Además, estos valores atípicos tienen un valor de 4 en la variable "time_signature" en el 91%

de los casos. Esto sugiere que las variables categóricas no son las responsables de que estos valores sean considerados atípicos.

Al examinar los datos en la *Tabla 5.2.1*, se destaca que en la variable "popularity," que tiene un promedio de 34.1, hay tres registros que tienen un valor de 0 en esa característica. Esto parece ser la razón por la cual se consideran valores atípicos. Al analizar la variable "energy," que tiene un promedio de 0.668, se observa que el registro 31341 tiene un valor de 0.277 por encima de la media, mientras que los registros 64134 y 103094 tienen valores de 0.3 y 0.291 por debajo de la media. Esto sugiere que estas diferencias con respecto a la media son la causa de que se clasifiquen como valores atípicos. En el caso de la variable "danceability", no se encuentran valores que estén significativamente alejados de la media, lo que nos lleva a suponer que estos puntos no influyen en el comportamiento de los valores atípicos.

Luego, se procedió a realizar ajustes en los pesos asignados a las variables con el propósito de explorar cómo estos ajustes impactan en la representación gráfica de los puntos en cuestión. En este contexto, se optó por asignar un peso de 100 a la variable numérica "popularity" y un peso de 50 a la variable numérica "energy". Estos ajustes en los pesos se encuentran reflejados en la *Figura 5.2.2* que se presenta a continuación.

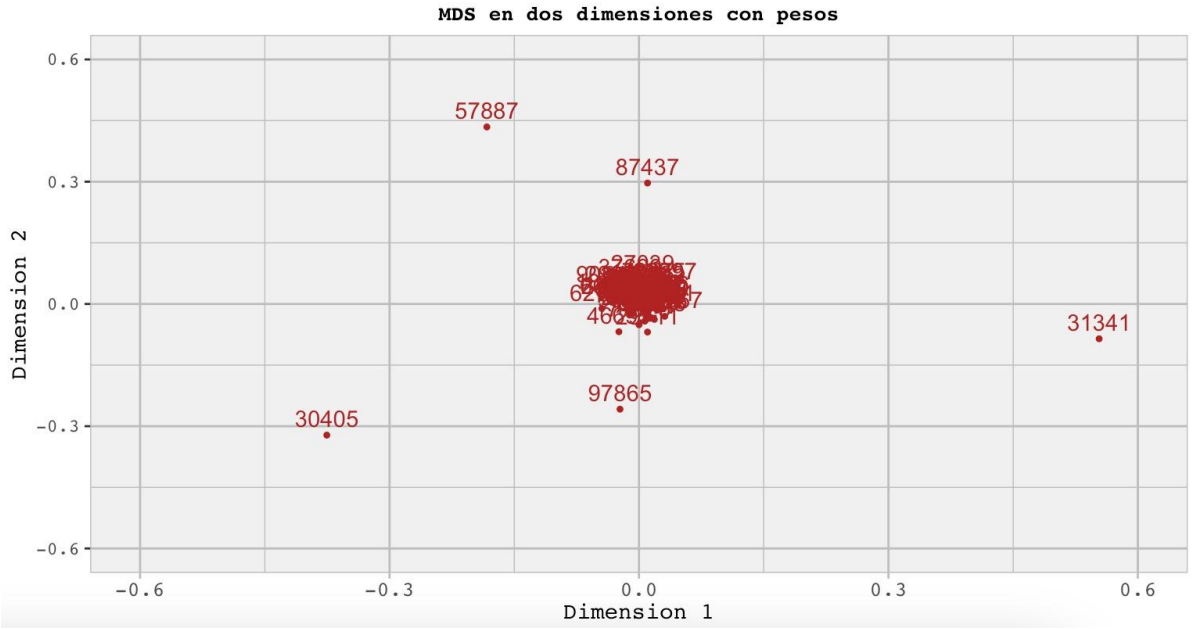

Figura 5.2.2

En este gráfico, se pueden apreciar una serie de valores atípicos diferentes en comparación con el gráfico anterior. Algunos de los valores atípicos son los mismos, mientras que otros han cambiado. Otra diferencia con respecto al gráfico anterior es que los valores atípicos están claramente destacados en el gráfico y están distribuidos de manera más dispersa. Además, se nota que los datos en la región central se agrupan más estrechamente debido a la mayor ponderación de "popularity" y "energy". Esto dificulta aún más la identificación de grupos de datos similares en el centro del gráfico.

Los únicos nuevos valores atípicos que aparecen en la Figura 5.2.2 son analizados en las siguientes tablas: *Tabla 5.2.3* para las variables numéricas categóricas y *Tabla 5.2.4* para las variables numéricas categóricas.

	popularity	**energy**	**danceability**
Media	34.01	0.668	0.562
57877	0	0.949	0.651
30405	0	0.941	0.73

Tabla 5.2.3

	mode	**time_signature**
Porcentaje del total	68%	91%
97865	1	4
64134	1	4

Tabla 5.2.4

En la *Tabla 5.2.4*, no se encuentra ningún valor atípico en las variables categóricas, lo que sugiere que estas variables no desempeñan un papel en la clasificación de los valores como atípicos. Por otro lado, al observar la *Tabla 5.2.3*, se nota que los nuevos valores atípicos tienen un valor de 0 en la característica "popularidad" y valores muy cercanos a uno en "energía". Esto tiene sentido, ya que al asignarle pesos a las variables, aquellos registros que tengan un valor atípico en esa variable puntual se van a acentuar o destacar más con respecto a la

nube de puntos mientras que si tienen un valor alejado de la media en una variable que no tiene peso, si bien antes ese registro podía ser un valor atípico, ahora en la visualización con peso no será tan perceptible.

5.3. Graficando MCA en tres dimensiones

A continuación, se decidió realizar un MCA en tres dimensiones, que se puede apreciar en la *Figura 5.3.1*. Se ha proporcionado un mosaico con cuatro capturas de pantalla desde distintos ángulos para que puedas tener una mejor percepción de la tridimensionalidad del MCA.

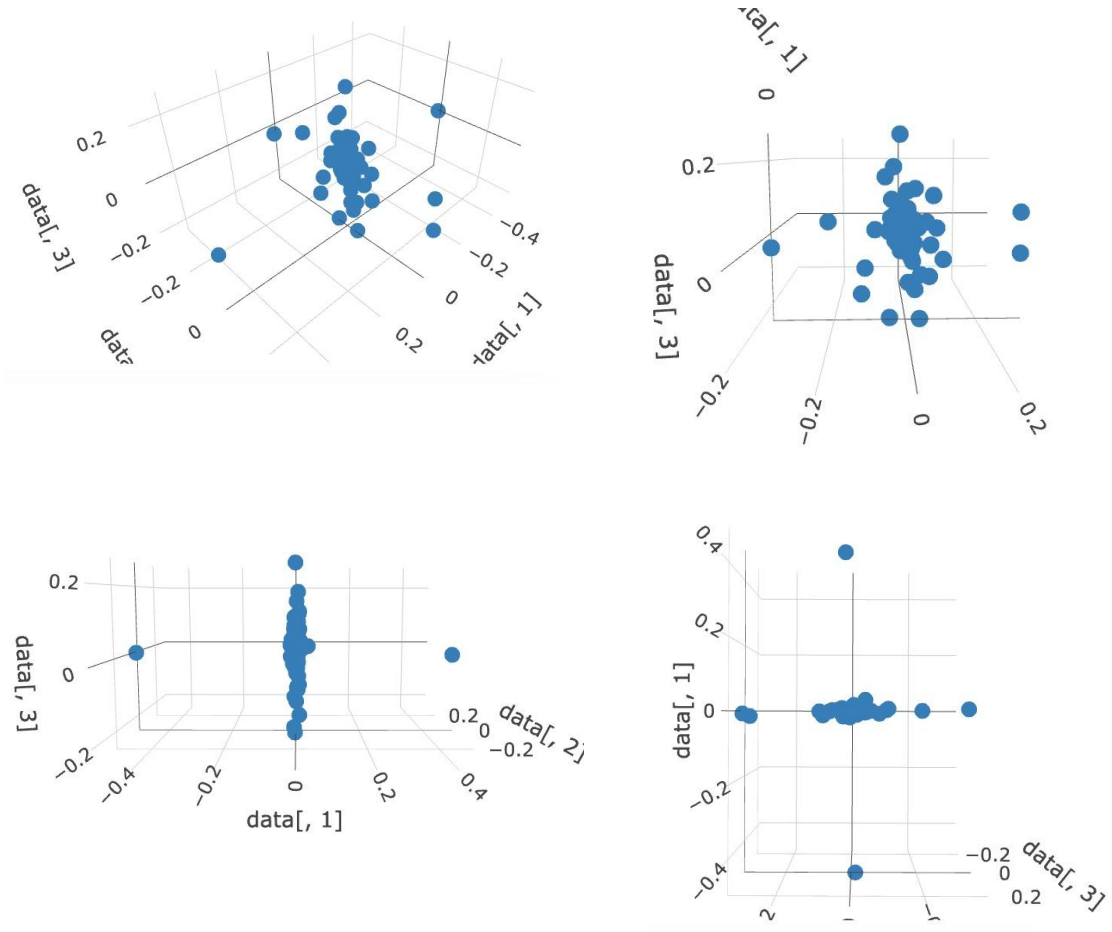

Figura 5.3.1

Como se puede ver en la Figura 5.3.1, hay una representación más clara de la distribución de los puntos a diferencia de verlo en dos dimensiones.

Si bien las representaciones en dos dimensiones proyectan la información en un plano bidimensional, pueden a su vez limitar la representación exacta de la relación entre los puntos. La representación 2D es óptima para observar relaciones lineales y estructuras simples, ya que preserva la simplicidad y facilita la interpretación al presentar datos de manera clara y fácilmente comprensible. Sin embargo, puede ocultar relaciones más complejas y no lineales entre los datos al reducir una dimensión.

Por otro lado, la visualización en tres dimensiones ofrece una representación más fiel a la distribución espacial de los datos. Añadir una dimensión adicional puede permitir la identificación de patrones más complejos y relaciones intrínsecas entre los puntos, ya que se incorpora una dimensión extra que ofrece una representación más cercana a la realidad tridimensional. No obstante, la visualización en 3D puede generar desafíos para la interpretación, dado que visualizar datos en tres dimensiones puede generar obstrucciones visuales y dificultar la comprensión general de la distribución de los datos. Por ejemplo, si se compara la Figura 5.2.1 comparado con la imagen en el cuadrante derecho superior de la Figura 5.3.1, es posible observar que desde una perspectiva en dos dimensiones, cuando la Dimensión 1 es 0 la gran mayoría de los datos se acomodan en una nube de puntos indiferenciada -a excepción de outliers previamente mencionados-, siendo que desde la otra perspectiva hay datos que se alejan del centro claramente y hay una mayor dispersión en líneas generales.

Por otro lado, a su vez se ha decidido analizar una perspectiva en tercera dimensión -véase Figura 5.3.2-, utilizando los datos a los que se les ha asignado pesos a las variables. A diferencia de la figura 5.3.1 se visualiza una estructura de datos más clara y concentrada, donde independientemente de la perspectiva se conserva una distribución similar. A su vez, si se toma la comparación de la Figura 5.2.2. y 5.3.2., los outliers remarcados en dos dimensiones son claramente delimitables en tres dimensiones, lo que tiene sentido al no haber mayor dispersión de los datos en la dimensión adicional.

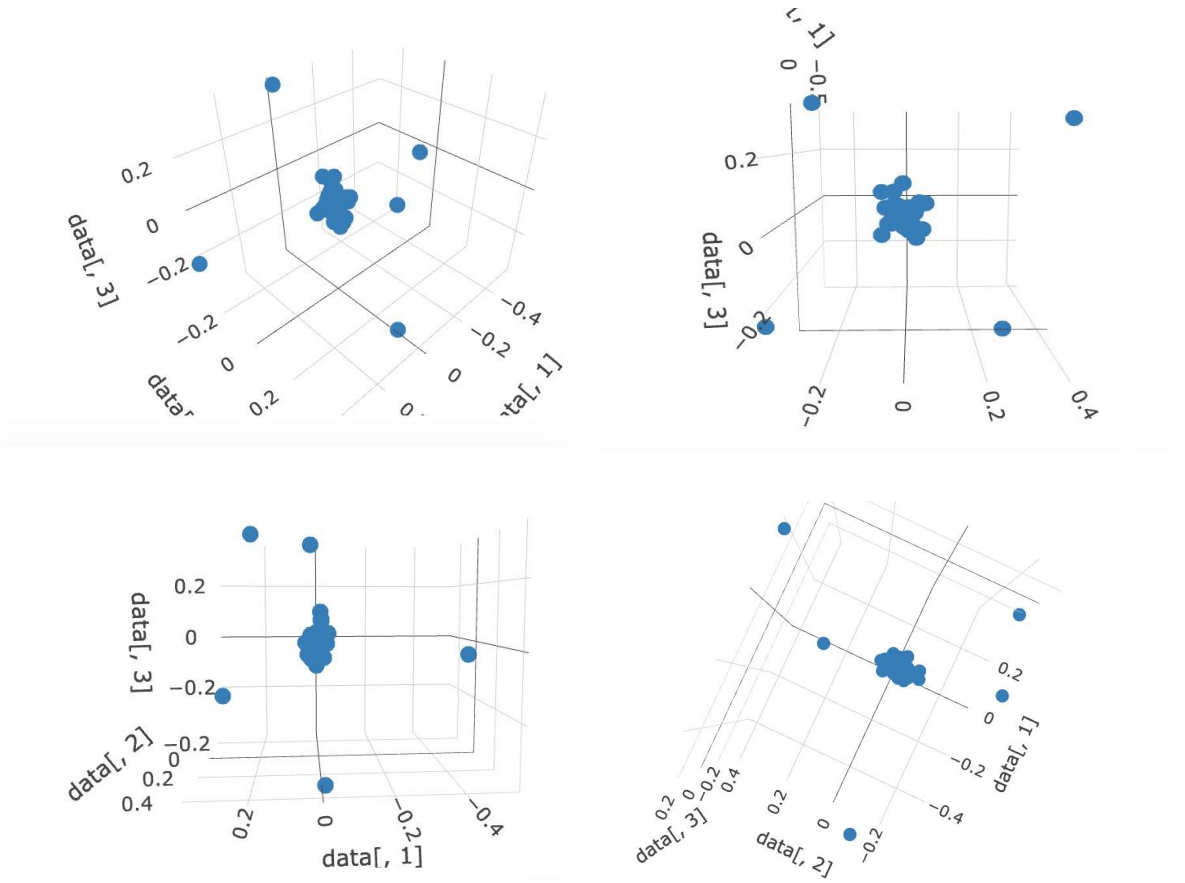

Figura 5.3.2

5.4. Calidad de ajuste de los MDS

Para analizar si conviene representar los puntos en dos dimensiones o tres hay que evaluar la calidad de ajuste de los modelos MDS a los datos originales. Para eso, se decide utilizar la medida de stress de Kruskal.

La medida de stress de Kruskal se calcula haciendo la raíz cuadrada de la suma de diferencias ponderadas cuadradas sobre la suma de diferencias ponderadas cuadradas totales. Esto se puede ver en la siguiente fórmula.

$$\sigma_1 = \sqrt{\frac{\sum_{j \geq i} [d_{ij} - f(\delta_{ij})]^2}{\sum_{j \geq i} d_{ij}^2}}$$

Un stress de Kruskal cercano a 0 indica un buen ajuste del modelo MDS a los datos originales. Esto significa que las distancias en el espacio MDS son muy similares a las distancias originales, lo que sugiere que el modelo MDS representa eficazmente la estructura de los datos. Cuanto más se acerque a 1, mayor será la discrepancia entre las distancias originales y las distancias en el espacio MDS. En este caso, el modelo MDS no representa bien la estructura de los datos y se considera que el ajuste es deficiente.

En la siguiente *Tabla 5.4.1* se observan los resultados al stress de Kruskal del escalamiento multidimensional en dos y tres dimensiones.

	Stress de Kruskal
Dos dimensiones	0.406
Tres dimensiones	0.329

Tabla 5.4.1

Como se puede observar en la *Tabla 5.4.1*, el stress de Kruskal de un MDS de tres dimensiones es menor que el stress de Kruskal de un MDS de dos dimensiones, esto sugiere que el modelo de MDS tridimensional se ajusta mejor a los datos y preserva de manera más fiel las relaciones entre los datos originales que el modelo bidimensional.

5.5. Detección de grupos

Es posible afirmar que la detección de grupos en los MDS realizados en dos y tres dimensiones no es evidente a simple vista. En el MDS en dos dimensiones original, como se muestra en la *Figura 5.2.1*, los datos parecen más dispersos y no muestran agrupamientos claramente identificables. En el MDS en dos dimensiones con pesos modificados, los datos se acercan más entre sí en el centro, lo que dificulta la identificación de grupos, en caso de que existan. Lo mismo se aplica a los MDS realizados en tres dimensiones; al agregar pesos a las variables "popularity" y "energy", las observaciones se concentran en el centro de los ejes, y no se logra discernir patrones claros entre ellos. Sin embargo, en la imagen en la parte superior derecha de la *Figura 5.3.1*, se pueden observar las posiciones relativas de las observaciones con mayor claridad, pero aún no se revelan patrones distintivos entre los datos.

5.6. Coordenadas de la visualización

El procedimiento que se llevó a cabo implica una serie de pasos para comprender la relación entre las coordenadas resultantes del MDS y las coordenadas originales de un conjunto de datos.

En primer lugar, se calculó la medida de distancia modificada Gower. Esta medida es utilizada para evaluar la distancia entre diferentes elementos en un conjunto de datos que puede contener tanto variables numéricas como categóricas. Posteriormente, se aplicó el método MDS que reduce la dimensionalidad de los datos, transformando la información original en un espacio de menor número de dimensiones. Las coordenadas resultantes del MDS representan una versión reducida de los datos, conservando en la medida de lo posible las relaciones de distancia entre los puntos originales.

A partir de las coordenadas resultantes del MDS en dos y tres dimensiones, se calculó una matriz de correlaciones entre las coordenadas MDS y las coordenadas originales para cada caso. Los resultados de las matrices de correlaciones reveló valores mayormente próximos a cero que indican una correlación débil entre las coordenadas generadas por el MDS y las distancias originales. Esto implica que no existe una correlación lineal apreciable entre las dimensiones del MDS y los registros originales.

6. Conclusión

El análisis a través del Escalamiento Multidimensional (MDS) nos proporciona una visión de las similitudes entre las canciones de Spotify. Es relevante destacar que el peso asignado a las variables "popularity" y "energy" tiene un impacto significativo en la agrupación y destacado de las canciones en el espacio MDS. Esto resalta la importancia crítica de estas características en la identificación de similitudes y patrones musicales. Sin embargo, la presencia de valores atípicos en estas variables dificulta el análisis de las observaciones, ya que tiende a agrupar todos los datos en el centro de los gráficos, dificultando la identificación de patrones.

Además, se observa una diferencia notable entre las representaciones en dos y tres dimensiones. A simple vista, se aprecia

que analizar las coordenadas en tres dimensiones facilita la comprensión de los datos, lo que llevó a considerar que la visualización tridimensional es más adecuada en este caso. Además, con la ayuda de la medida de stress de Kruskal, se confirmó que el modelo de MDS tridimensional se ajusta mejor a los datos y preserva de manera más precisa las relaciones entre los datos originales en comparación con el modelo bidimensional. A pesar de ello, el modelo tridimensional tampoco permitió detectar grupos.

En resumen, estos resultados enfatizan la utilidad del MDS para comprender las similitudes entre registros con múltiples variables de distintos tipos. Sin embargo, la presencia de valores atípicos dificultó el análisis al no poder visualizar claramente las relaciones entre las canciones ni identificar grupos similares de canciones de Spotify.

7. Anexo

7.1. Bootstrap

En relación con MDS, el proceso de bootstrap implica tomar múltiples muestras (con reemplazo) del conjunto de datos original y aplicar el procedimiento de MDS a cada muestra generada. Luego, se obtienen múltiples representaciones espaciales (coordenadas MDS) a partir de cada una de estas muestras.

Al ejecutar MDS en cada muestra, se generan múltiples conjuntos de coordenadas que representan diferentes perspectivas o variaciones en la disposición de los datos en el espacio reducido. Esta variabilidad en las coordenadas MDS revela cómo cambia la disposición de los puntos al re-muestrear el conjunto de datos.

Los resultados se visualizan en la *Figura 6.1* Cada círculo corresponde a una variable o punto de datos, y su ubicación en el gráfico refleja su posición en el espacio reducido, generado por el análisis de MDS. Por otro lado, la superposición de círculos -como time_signature5 y energy o mode y danceability- indica que esos elementos (variables) tienen posiciones similares o idénticas en el espacio MDS, lo que sugiere una mayor estabilidad o consistencia en la representación de esas variables a través de las diferentes muestras bootstrap. En contraste, los círculos que no se superponen -así como popularity- representan

elementos que varían más en su posición en el espacio MDS a lo largo de las muestras bootstrap, lo que sugiere una mayor variabilidad o inestabilidad en la representación de esos elementos.

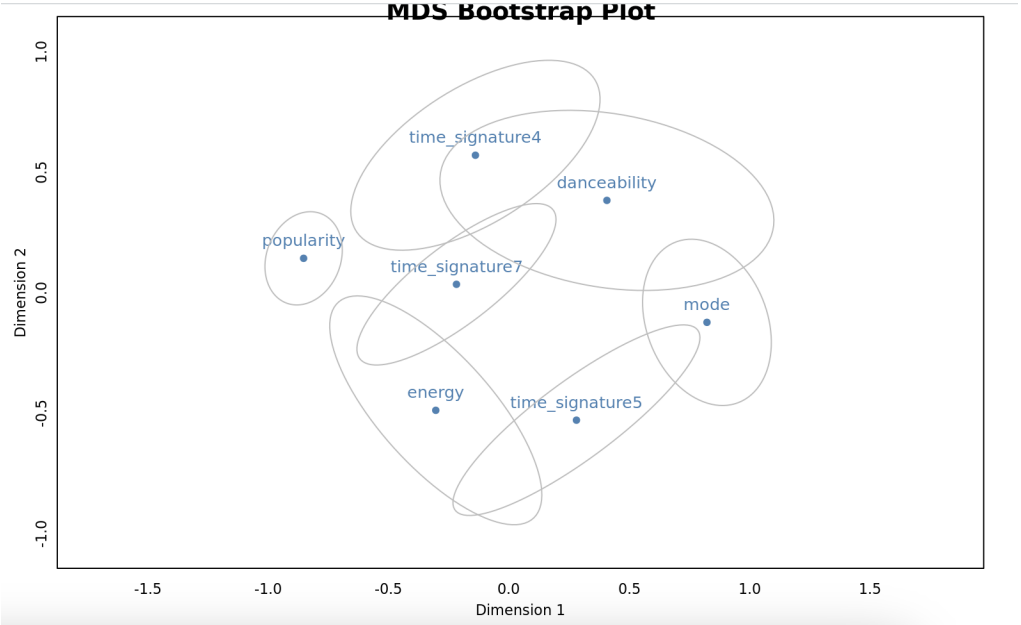

Figura 6.1