

# Contenido

Contexto .....	4
Problema por resolver.....	5
Datos Disponibles.....	6
Enfoque de Solución.....	8
Desafíos Técnicos .....	10
Impacto y Aplicación del Modelo.....	11
Mapa de calor (correlación entre variables continuas) .....	14
Justificación ¿Por qué usar un modelo NO supervisado para resolver esa problemática?.....	14
Calidad de los datos .....	17
Ingeniería de variables .....	18
Resumen EDA .....	19
Diccionario de variables .....	19
Histograma de frecuencias para variables continuas: .....	21
Pairplot para variables continuas (permite ver correlaciones o patrones entre las variables.): ..	24
Detección y remoción de outliers .....	25
Histograma de frecuencias (después de remover outliers): .....	27
Escalamiento de los datos .....	28
Reducción de dimensiones (PCA).....	29
Métricas de los algoritmos .....	31
Gráfica del codo (K-means) .....	32
Coeficiente de Silueta para KMeans y GMM .....	33
Calinski .....	34
Elección final de k (# de grupos) .....	36
Modelo KMeans .....	37
Distribución de los registros:.....	37
Visualización de clústeres con 3 componentes:.....	37
Modelo GMM (Gaussian Mixture Model) .....	38
Distribución de los registros .....	38
Visualización de clústeres con 3 componentes.....	38
Perfilamiento .....	39
K-Means (2 clústeres).....	39
<b>Clúster 0: "Modern Dance-Pop Pulse"</b> .....	40

<b>Clúster 1: "Timeless Acoustic Soul".....</b>	41
GMM (3 clústeres).....	43
<b>Clúster 0: "Classic Acoustic Reflections" .....</b>	43
<b>Clúster 1: "Retro Balanced Nostalgia" .....</b>	44
<b>Clúster 2: "Modern Pop Energy" .....</b>	45
Comparativa entre K-Means (2 Clústeres) y GMM (3 Clústeres) en el Contexto de Spotify .....	47
¿En que caso usar cada algoritmo?.....	49
Conclusión:.....	49
Siguientes Pasos para la Clusterización: Integración Estratégica en el Negocio .....	50
1.    Integración con el Motor de Recomendaciones .....	50
2.    Creación de Playlists Temáticas Automatizadas .....	50
3.    Campañas de Marketing Personalizadas .....	51
4.    Optimización del Catálogo Musical.....	51
5.    Monitoreo y Ajuste Continuo.....	52
¿Cómo Ayuda la Solución al Negocio? .....	52
Activación de la Solución.....	53

# Modelos de clusterización: Análisis de tendencias en Spotify

En la era digital, la industria musical ha experimentado una transformación radical en la manera en que los consumidores descubren, interactúan y consumen música. Con la llegada de plataformas de streaming como Spotify, la cantidad de datos disponibles sobre los hábitos de escucha de los usuarios ha crecido exponencialmente, lo que ha abierto nuevas oportunidades para el análisis y la personalización de contenido.

Dentro de este contexto, el análisis de datos se ha convertido en una herramienta fundamental para entender tendencias musicales, segmentar audiencias y diseñar estrategias de marketing más eficaces. En particular, la clusterización (una técnica de *aprendizaje no supervisado*) permite agrupar canciones, artistas o usuarios en función de características similares, lo que puede ayudar a identificar patrones ocultos y optimizar la recomendación de contenido.

Este estudio se enfoca en la aplicación de algoritmos de clusterización sobre un conjunto de datos de Spotify con el fin de analizar tendencias musicales y proponer estrategias de marketing basadas en los diferentes grupos identificados. A lo largo del análisis, se han utilizado técnicas de preprocesamiento de datos, escalado de variables, reducción de dimensionalidad y diferentes algoritmos de clustering para evaluar la estructura interna de los datos y extraer insights valiosos para la industria musical.

## Contexto

Spotify es una de las plataformas de streaming musical más grandes del mundo, con más de 500 millones de usuarios activos y un catálogo que supera los 100 millones de canciones. A través de su modelo de negocio basado en suscripción y publicidad, la plataforma no solo permite a los usuarios acceder a una amplia variedad de contenido, sino que también ofrece herramientas avanzadas de personalización, como playlists generadas por inteligencia artificial y recomendaciones basadas en el historial de escucha.

El uso de algoritmos de machine learning es una parte esencial del éxito de Spotify, ya que permiten mejorar la experiencia del usuario y maximizar la retención en la plataforma. Funcionalidades como Discover Weekly, Release Radar y las playlists personalizadas dependen en gran medida del análisis de patrones en el comportamiento de los oyentes.

Desde el punto de vista del marketing musical, la segmentación de usuarios y canciones es clave para diseñar campañas efectivas, identificar nichos de mercado y predecir tendencias emergentes. La posibilidad de agrupar canciones según sus características sonoras y metadatos (tempo, energía, valencia, instrumentalidad, etc.) permite a los sellos discográficos, artistas y plataformas de streaming adaptar sus estrategias y optimizar la promoción de contenido.

En este trabajo, se ha recopilado un conjunto de datos con información sobre canciones de Spotify, incluyendo variables numéricas relacionadas con sus propiedades acústicas y de popularidad. Mediante la **clusterización**, se busca identificar patrones en la música que permitan comprender mejor las dinámicas del mercado y mejorar la toma de decisiones en la industria.

## Problema por resolver

El creciente volumen de datos en la industria musical representa tanto una oportunidad como un desafío. Si bien el acceso a grandes cantidades de información facilita la identificación de tendencias y preferencias de los usuarios, la cantidad de datos puede ser abrumadora y difícil de interpretar sin herramientas adecuadas.

En este trabajo, se aborda la siguiente pregunta clave:

**¿Cómo podemos utilizar técnicas de clusterización para identificar patrones en las tendencias musicales y mejorar las estrategias de marketing en la industria del streaming?**

Para responder a esta pregunta, se plantean los siguientes objetivos específicos:

- **Segmentar** las canciones en distintos grupos en función de sus características acústicas y de popularidad.
- **Analizar** la composición de cada clúster para entender qué atributos diferencian a cada grupo.
- **Identificar** qué tipos de canciones tienen mayor potencial de éxito en el mercado.
- **Proponer** estrategias de marketing basadas en la segmentación de canciones y usuarios.

A lo largo de este estudio, se aplicarán diferentes enfoques de clusterización (K-Means, GMM, DBSCAN, entre otros) para determinar cuál es el más adecuado para la estructura

de los datos. También se evaluará el impacto de distintas técnicas de preprocesamiento y reducción de dimensionalidad en la calidad de los clústeres generados.

Este análisis busca aportar conocimientos valiosos para la industria musical, proporcionando herramientas que permitan a los artistas, discográficas y plataformas de streaming tomar decisiones basadas en datos. Al identificar grupos de canciones con características similares, se pueden optimizar estrategias de promoción, mejorar la recomendación de contenido y predecir tendencias futuras en el mercado musical.

## Datos Disponibles

Para llevar a cabo el análisis de clusterización en Spotify, se ha trabajado con un conjunto de datos que contiene información detallada sobre diversas características de las canciones. Estas variables permiten analizar distintos aspectos de la música, desde su estructura sonora hasta su nivel de popularidad en la plataforma. A continuación, se describe cada una de ellas:

- **valence:** Representa el nivel de positividad de una canción en una escala de 0 a 1. Un valor bajo indica una canción con un tono más triste o melancólico, mientras que un valor alto sugiere una canción más alegre y positiva.
- **year:** Indica el año de lanzamiento de la canción. Esta variable es útil para analizar tendencias a lo largo del tiempo y ver cómo han evolucionado las características musicales.
- **acousticness:** Mide la probabilidad de que una canción sea acústica. Un valor cercano a 1 indica una alta probabilidad de que la pista sea predominantemente acústica, mientras que valores más bajos sugieren un uso significativo de instrumentos electrónicos.
- **artists:** Contiene el nombre del o los artistas responsables de la canción. Esta variable es categórica y no se usa directamente en los algoritmos de clusterización, pero puede ser útil para interpretar los resultados y analizar la segmentación por artistas.
- **danceability:** Evalúa qué tan bailable es una canción, basándose en factores como el ritmo, la estabilidad del tempo y la regularidad de los beats. Se mide en una escala de 0 a 1, donde valores más altos indican canciones más adecuadas para el baile.
- **energy:** Representa la intensidad y actividad percibida de una canción. Se mide en una escala de 0 a 1, donde valores altos corresponden a canciones más

intensas y enérgicas (como rock o electrónica) y valores bajos a canciones más suaves o relajantes.

- **explicit:** Indica si una canción contiene contenido explícito. Se trata de una variable binaria donde 1 significa que la canción tiene letras explícitas y 0 que no las tiene.
- **id:** Es un identificador único asignado por Spotify a cada canción. Esta variable no es relevante para el análisis de clusterización, pero puede ser útil para recuperar información adicional sobre una canción en la plataforma.
- **instrumentalness:** Mide la probabilidad de que una pista sea instrumental. Un valor cercano a 1 indica que la canción tiene pocas o ninguna letra, mientras que valores más bajos sugieren la presencia de voz en la grabación.
- **key:** Representa la tonalidad de la canción, codificada en valores enteros de 0 a 11 (donde 0 es C, 1 es C#, 2 es D, etc.). Aunque es una variable categórica, se puede analizar su distribución en los clústeres para ver si ciertas tonalidades son más comunes en determinados grupos.
- **liveness:** Estima la presencia de una audiencia en la grabación. Valores altos (por encima de 0.8) sugieren que la canción fue grabada en vivo, mientras que valores más bajos indican una grabación en estudio.
- **loudness:** Indica el volumen promedio de la canción en decibeles (dB). Es una variable continua que puede estar relacionada con la energía de la canción y su percepción en entornos como clubes o festivales.
- **mode:** Indica si la canción está en modo mayor o menor. Es una variable binaria donde 1 representa el modo mayor (asociado con emociones más alegres) y 0 representa el modo menor (asociado con emociones más melancólicas o serias).
- **name:** Contiene el título de la canción. Esta variable no se usa directamente en los algoritmos de clusterización, pero es útil para interpretar y dar contexto a los resultados.
- **popularity:** Mide la popularidad de una canción en Spotify en una escala de 0 a 100, donde valores más altos indican mayor número de reproducciones y engagement dentro de la plataforma.
- **speechiness:** Indica la cantidad de contenido hablado en una pista. Valores cercanos a 1 corresponden a grabaciones predominantemente habladas, como podcasts o discursos, mientras que valores bajos indican canciones con poca o ninguna vocalización hablada.
- **tempo:** Representa el tempo de la canción en beats por minuto (BPM). Esta variable es clave para el análisis de ritmo y puede influir en la percepción de una canción como más animada o relajada.

- **duration\_min:** Indica la duración total de la canción en minutos. Esta variable puede ayudar a entender si hay diferencias en la longitud de las canciones dentro de los clústeres y cómo afecta la duración a su clasificación.

El análisis de estas variables permitirá identificar patrones en la música y segmentar las canciones en distintos grupos, proporcionando información valiosa para la industria del streaming y el marketing musical.

## Enfoque de Solución

Para llevar a cabo el análisis de clusterización de las canciones de Spotify, se siguió un proceso estructurado que permitió asegurar la calidad de los datos, seleccionar las variables más relevantes y aplicar diversas técnicas de agrupamiento. A continuación, se describen en detalle los pasos implementados:

### 1. Lectura de datos

El primer paso consistió en cargar el conjunto de datos en un entorno de análisis utilizando **pandas**. Se exploraron los primeros registros para verificar su estructura, los tipos de datos y la presencia de valores faltantes o atípicos.

### 2. Calidad de los datos

Se realizaron diversas verificaciones y ajustes para mejorar la calidad de los datos:

- Eliminación de duplicados, en caso de haberlos.
- Conversión de formatos de datos incorrectos.
- Tratamiento de valores nulos mediante imputación o eliminación según su impacto en el análisis.

### 3. Ingeniería de variables

Para mejorar la interpretación y el procesamiento de los datos, se crearon nuevas variables a partir de las existentes. Una de las principales transformaciones fue la conversión de la duración de las canciones de **milisegundos (duration\_ms) a minutos (duration\_min)**, facilitando la comprensión y el análisis.

### 4. Análisis exploratorio de datos (EDA) y visualización

Se realizó un análisis detallado de las variables en dos categorías:

- **Variables discretas** (como explicit, mode, key) fueron analizadas en términos de su distribución y frecuencia.
- **Variables continuas** (como danceability, energy, tempo, popularity) fueron visualizadas mediante histogramas, diagramas de caja y gráficos de dispersión para identificar tendencias, distribuciones y posibles valores atípicos.

## 5. Eliminación de outliers con Z-Score

Para evitar que valores extremos distorsionaran el análisis de clusterización, se aplicó el método **Z-Score**, eliminando las observaciones con valores superiores a un umbral determinado (generalmente  $|z| > 3$ ).

## Clusterización por grupos

### 6. Selección del número óptimo de grupos para K-Means y GMM

Para determinar la cantidad ideal de clústeres, se utilizaron los siguientes criterios:

- **Método del Codo (Elbow Method)**: Evaluando la inercia en función del número de clústeres en **K-Means**.
- **Silhouette Score**: Analizando la cohesión y separación de los clústeres generados.
- **Índice de Calinski-Harabasz**: Para validar la compacidad y separación de los grupos.

## 7. Escalamiento de variables

Dado que las variables del dataset tienen diferentes escalas, se aplicó **MinMaxScaler** para normalizarlas dentro de un rango de 0 a 1. Esto evita que variables con valores grandes dominen el análisis.

### 8. Reducción de dimensiones con PCA

Para mejorar la visualización y eficiencia del modelo, se implementó **Análisis de Componentes Principales (PCA)**, reduciendo la cantidad de dimensiones y conservando la mayor cantidad de varianza posible.

### 9. Prueba de silueta para K-Means y GMM

Se evaluaron las métricas de silueta para comparar los agrupamientos de **K-Means** y **Gaussian Mixture Model (GMM)**, determinando cuál de los dos modelos representaba mejor la segmentación de las canciones.

## 10. Ejecución de K-Means y GMM

Con la cantidad de clústeres determinada y las variables escaladas, se implementaron ambos algoritmos:

- **K-Means:** Se aplicó este algoritmo basado en centroides para crear agrupaciones claras y diferenciadas.
- **GMM:** Se utilizó este modelo probabilístico basado en distribuciones Gaussianas para comparar con los resultados de K-Means.

### Clusterización jerárquica

Para complementar los modelos anteriores, se implementó **clustering jerárquico**, permitiendo visualizar cómo varían los grupos formados y comparándolos con los resultados obtenidos en **K-Means** y **GMM**.

Siguiendo estos pasos, se logró segmentar las canciones de Spotify en distintos grupos, identificando patrones en sus características y proporcionando información valiosa para entender tendencias musicales y estrategias de marketing en la industria del streaming.

## Desafíos Técnicos

Durante el proceso de análisis y clusterización de los datos musicales de Spotify, se presentaron diversos retos que requirieron ajustes metodológicos y estrategias específicas para obtener resultados coherentes. A continuación, se destacan los principales desafíos técnicos encontrados:

### 1. Manejo de outliers en un gran número de registros

Uno de los principales obstáculos fue la presencia de **outliers** en muchas de las variables analizadas. Esto afectaba la agrupación, ya que los valores extremos podían distorsionar los centroides en **K-Means** o alterar las probabilidades en **GMM**. Para mitigar este problema, se utilizó el método **Z-Score**, eliminando valores con una desviación estándar mayor a 3. Sin embargo, esto llevó a la pérdida de algunos datos, lo que generó la necesidad de evaluar su impacto en la representatividad del conjunto.

## **2. Alto volumen de datos y necesidad de una muestra representativa**

El dataset original contenía **más de 170,000 registros**, lo que representaba un desafío tanto en términos de procesamiento computacional como de memoria. Para optimizar el rendimiento, se optó por trabajar con una muestra del **20%** del total de datos. No obstante, asegurar que esta muestra fuera representativa de la distribución general requirió validar la preservación de las distribuciones de las variables clave.

## **3. Elección del número de clústeres**

Uno de los retos más complejos fue determinar la cantidad óptima de grupos. Se utilizaron múltiples métricas, pero cada una sugería diferentes números de clústeres:

- **Elbow Method (Codo)** sugería un valor específico basado en la inercia.
- **Silhouette Score** recomendaba un número diferente, optimizando la cohesión y separación de los clústeres.
- **Índice de Calinski-Harabasz** aportaba otra perspectiva sobre la calidad de la segmentación.

Dado que los métodos no ofrecían una respuesta unificada, la elección final requirió un análisis cualitativo basado en la interpretabilidad de los clústeres generados, combinando criterios estadísticos con validación visual de los agrupamientos.

## **4. Perfilamiento e interpretación de los clústeres**

Identificar patrones dentro de los grupos formados fue otro desafío relevante. Aunque los algoritmos generaban clústeres diferenciados matemáticamente, **interpretar sus características y definir perfiles concretos** no fue una tarea trivial. Algunas variables eran más determinantes que otras, y en muchos casos se requería cruzar información para descubrir tendencias ocultas.

Este análisis cualitativo fue clave para que la segmentación tuviera un valor real en términos de tendencias musicales y estrategias de marketing. Se realizó una inspección detallada de cada clúster, analizando diferencias en atributos como **popularidad, energía, tempo y valencia**, entre otros, para generar conclusiones significativas.

## **Impacto y Aplicación del Modelo**

El análisis de clusterización aplicado a los datos musicales de **Spotify** proporciona información valiosa en diversos aspectos clave para la plataforma, permitiendo

optimizar estrategias de negocio, mejorar la experiencia del usuario y proporcionar herramientas avanzadas para la industria musical. A continuación, se detallan algunas de las principales aplicaciones:

## 1. Personalización de Recomendaciones

Uno de los mayores beneficios de la clusterización es la posibilidad de mejorar los algoritmos de **recomendación personalizada**. Al segmentar las canciones en grupos según características como **valencia, energía, tempo y popularidad**, Spotify puede ofrecer playlists más precisas y relevantes para cada usuario. Por ejemplo:

- Usuarios que prefieren música **energética y bailable** pueden recibir recomendaciones basadas en clústeres con alta **danceability y tempo**.
- Para quienes buscan música relajante, se pueden recomendar canciones de clústeres con alta **acousticness** y baja **loudness**.

Este enfoque permite mejorar la fidelización del usuario y aumentar el tiempo de permanencia en la plataforma.

## 2. Optimización de Playlists Algorítmicas

Spotify genera muchas playlists populares basadas en **hábitos de escucha y tendencias** (ej. *Discover Weekly*, *Release Radar*, *Daily Mix*). Con la clusterización, estas playlists pueden mejorarse al asegurarse de que las canciones seleccionadas sean más homogéneas dentro de un mismo grupo, evitando combinaciones que puedan resultar incoherentes para el oyente.

Por ejemplo, si un usuario suele escuchar música **indie acústica**, es más probable que disfrute de una playlist con canciones dentro del mismo clúster en lugar de recibir recomendaciones de música con alta **intensidad y volumen**, lo que podría generar disonancia.

## 3. Estrategias de Marketing Musical

Para los artistas y sellos discográficos, comprender cómo se agrupan las canciones permite desarrollar estrategias de **marketing más efectivas**. Algunos ejemplos incluyen:

- Identificación de **nichos musicales**: Si un artista encaja en un clúster con características de alta popularidad, puede apuntar a estrategias que lo positionen en playlists destacadas.

- Planificación de **campañas publicitarias**: Las discográficas pueden usar esta segmentación para decidir cómo promocionar nuevos lanzamientos en función del perfil de los oyentes de cada clúster.
- **Optimización de lanzamientos**: Si ciertos clústeres tienen una mayor tendencia a viralizarse en redes sociales, los artistas pueden enfocarse en producir contenido alineado con esas características.

#### 4. Identificación de Tendencias Musicales

El análisis de clusterización también ayuda a **detectar cambios en las tendencias musicales a lo largo del tiempo**. Por ejemplo:

- Si se observa que ciertos clústeres con características específicas (ej. **tempo alto, alta energía**) están creciendo en popularidad, esto puede indicar una tendencia emergente en la industria.
- A nivel histórico, se pueden analizar patrones de evolución en los géneros y cómo han cambiado las preferencias del público en distintos períodos.

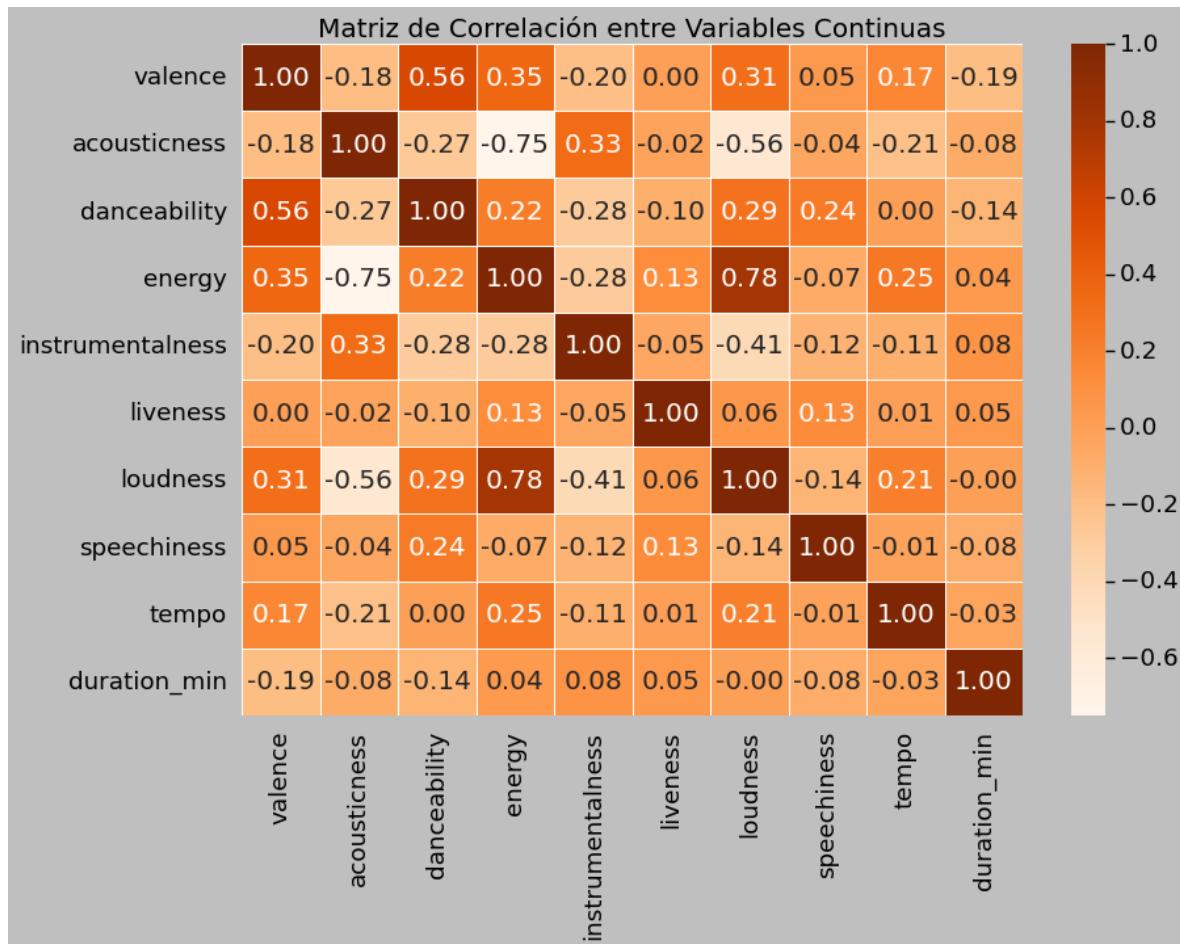
Este tipo de análisis permite a Spotify anticiparse a cambios en el mercado y ajustar su estrategia para seguir ofreciendo contenido relevante a sus usuarios.

#### 5. Expansión a Nuevos Mercados y Públicos

La segmentación de canciones por clústeres facilita la identificación de preferencias musicales en diferentes regiones del mundo. Por ejemplo, si en **Latinoamérica** predominan clústeres con alta **danceability y valencia**, mientras que en **Europa** son más comunes los clústeres con alta **instrumentalness y tempo moderado**, Spotify puede utilizar esta información para adaptar su contenido y estrategias a cada mercado.

Asimismo, esta información es útil para artistas que buscan expandir su audiencia internacionalmente, permitiéndoles adaptar su estilo o estrategias de promoción según el perfil de cada clúster.

## Mapa de calor (correlación entre variables continuas)



## Justificación ¿Por qué usar un modelo NO supervisado para resolver esa problemática?

El uso de modelos de **clusterización** para analizar las tendencias musicales en Spotify es una estrategia fundamental para extraer patrones ocultos en los datos y segmentar la música en grupos con características similares. A diferencia de otros enfoques supervisados, la clusterización permite identificar estructuras en los datos sin la necesidad de etiquetas previas, lo que la hace ideal para un problema donde se busca **descubrir tendencias, segmentar audiencias y mejorar la personalización**.

## 1. La Naturaleza No Supervisada del Problema

El problema que buscamos resolver no cuenta con **etiquetas predefinidas** que clasifiquen las canciones en categorías exactas, sino que queremos identificar **grupos naturales** de canciones con características similares.

- Si intentáramos abordar este problema con un modelo supervisado, necesitaríamos etiquetas previamente definidas, lo cual no es viable ya que la música no se categoriza de forma rígida en grupos únicos.
- La **clusterización**, al ser un método no supervisado, nos permite encontrar **relaciones subyacentes** en los datos sin necesidad de definir categorías de antemano.

## 2. Mejor Representación de la Diversidad Musical

La música es un fenómeno **multidimensional** con atributos que van desde el **tempo y la energía** hasta la **acousticness y la valencia emocional**. Un modelo de clusterización permite identificar patrones dentro de estas dimensiones sin limitarse a clasificaciones convencionales de géneros musicales.

- Por ejemplo, una canción puede ser clasificada como *pop*, pero dependiendo de sus características, puede pertenecer a un grupo con alta **bailabilidad y energía**, o a un grupo con **acousticness y tempo bajo**.
- En lugar de forzar etiquetas rígidas, la clusterización agrupa canciones según similitudes reales en sus características acústicas y de popularidad.

## 3. Optimización de Recomendaciones y Experiencia del Usuario

Uno de los principales desafíos de plataformas como Spotify es ofrecer **recomendaciones personalizadas y relevantes**. La clusterización permite mejorar este proceso al segmentar canciones en grupos específicos, lo que facilita:

- La generación de **playlists algorítmicas más precisas**.
- La recomendación de canciones con características similares a las que el usuario ya escucha.
- La personalización de la experiencia de cada usuario según su perfil de consumo musical.

Sin esta segmentación, el sistema de recomendaciones puede volverse demasiado amplio o inexacto, mezclando canciones que no necesariamente encajan en la preferencia del usuario.

#### **4. Reducción de la Complejidad en el Análisis de Grandes Volúmenes de Datos**

Trabajamos con más de **170,000 registros**, lo que hace que analizar cada canción de manera individual sea inviable. La clusterización simplifica este problema al permitirnos agrupar las canciones en un número reducido de categorías representativas, facilitando la **interpretación y toma de decisiones**.

Por ejemplo, en lugar de analizar cientos de miles de canciones, podemos reducir el problema a analizar **3 o 4 clústeres principales**, cada uno representando un subconjunto con características musicales homogéneas. Esto permite extraer **insights accionables** de manera más eficiente.

#### **5. Flexibilidad para la Industria Musical y el Marketing**

El análisis de clusterización no solo es útil para los usuarios de Spotify, sino también para la **industria musical en general**.

- Permite a las discográficas identificar qué tipo de música tiene mayor potencial de éxito en cada mercado.
- Ayuda a los artistas a entender en qué segmentos encaja mejor su música y a qué audiencia dirigirse.
- Facilita la planificación de campañas publicitarias basadas en la segmentación de usuarios con intereses musicales similares.

### **Conclusión**

El uso de modelos de clusterización es la mejor solución para este problema porque **no requiere etiquetas previas, permite descubrir patrones naturales en los datos, facilita la personalización de contenido, reduce la complejidad del análisis y ofrece información valiosa para la industria musical**. Gracias a este enfoque, Spotify puede mejorar su algoritmo de recomendaciones, optimizar sus playlists y ayudar a artistas y discográficas a entender mejor las tendencias del mercado.

# Calidad de los datos

Para garantizar la integridad del análisis y la clusterización, se realizaron varias verificaciones y ajustes en la calidad de los datos:

## 1. Detección de valores duplicados

- Se utilizó `df.duplicated()` para verificar la existencia de registros duplicados en el dataset.
- **Resultado:** No se encontraron registros duplicados, por lo que no fue necesario eliminarlos.

## 2. Eliminación de columnas irrelevantes

- Se eliminó la columna `id`, ya que solo representaba un identificador único y no aportaba información útil para la clusterización.

## 3. Verificación de valores nulos

- Se utilizó el siguiente código para detectar valores faltantes en el dataset:

```
# Convertir el resultado a un DataFrame para visualización completa
null_counts = df.isnull().sum().reset_index()
null_counts.columns = ['Variable', 'Valores_Nulos']
print(null_counts)
```

- **Resultado:** No se encontraron valores nulos en ninguna de las variables, por lo que no fue necesario aplicar técnicas de imputación de datos.

Variable	Valores_Nulos
valence	0
year	0
acousticness	0
artists	0
danceability	0
duration_ms	0
energy	0
explicit	0
id	0
instrumentalness	0
key	0
liveness	0
loudness	0
mode	0
name	0
popularity	0
release_date	0
speechiness	0
tempo	0
is_duplicate	0

Con estas verificaciones, se confirmó que el dataset estaba limpio y listo para su análisis, lo que permitió avanzar con la exploración y modelado sin necesidad de aplicar correcciones adicionales.

## Ingeniería de variables

En el proceso de análisis de los datos de Spotify, se llevaron a cabo diversas transformaciones en las variables para optimizar su uso en los algoritmos de clusterización y evitar que ciertos valores influyeran de manera desproporcionada en los resultados. A continuación, se describen las principales modificaciones realizadas:

1. **Conversión de la variable 'duration\_ms'**: La columna 'duration\_ms', que representa la duración de las canciones en milisegundos, fue convertida a minutos para hacerla más interpretable y facilitar su análisis posterior. Esto se realizó mediante la siguiente operación:

```
df['duration_min'] = df['duration_ms'] / 60000  
  
df['duration_min'] = df['duration_min'].round(3)
```

Esta conversión no solo mejora la legibilidad de la información, sino que también permite una mejor comparación de las duraciones en un rango más manejable y comprensible.

2. **Eliminación de la columna 'release-date'**: La columna 'release-date' fue descartada del conjunto de datos debido a que contenía información redundante. La variable 'year' ya proporcionaba el año de lanzamiento de las canciones, por lo que mantener ambas columnas no aportaba valor adicional al análisis. Así, se eliminó la columna para simplificar el DataFrame y reducir la dimensionalidad del modelo.
3. **Ajuste de las variables dummies 'mode' y 'explicit'**: Las variables dummies 'mode' (modo de la canción) y 'explicit' (indicador de contenido explícito) fueron multiplicadas por **0.1** para mitigar su influencia en los modelos de normalización y clusterización. Esta modificación tenía como objetivo evitar que, durante la aplicación del MinMax Scaler y el posterior proceso de clústering, estas variables

dominaron los resultados, creando grupos artificiales centrados en torno a estas características. Al reducir su peso en el modelo, se permitió una mejor agrupación de las canciones basándose en sus características musicales y no en su modo o explicititud.

4. **Eliminación de outliers:** Como parte del proceso de limpieza de datos, se realizó una eliminación de outliers para mejorar la calidad del análisis. Sin embargo, esta etapa fue realizada posterior al **Análisis Exploratorio de Datos (EDA)**, por lo que no se detalla en este apartado. La eliminación de outliers fue crucial para asegurar que los valores atípicos no distorsionaran los resultados de los modelos posteriores.

Estas modificaciones en las variables contribuyeron a crear un conjunto de datos más limpio y equilibrado, mejorando la calidad de los análisis y la precisión de los modelos posteriores.

## Resumen EDA

### Diccionario de variables

#### - **Variables Continuas:**

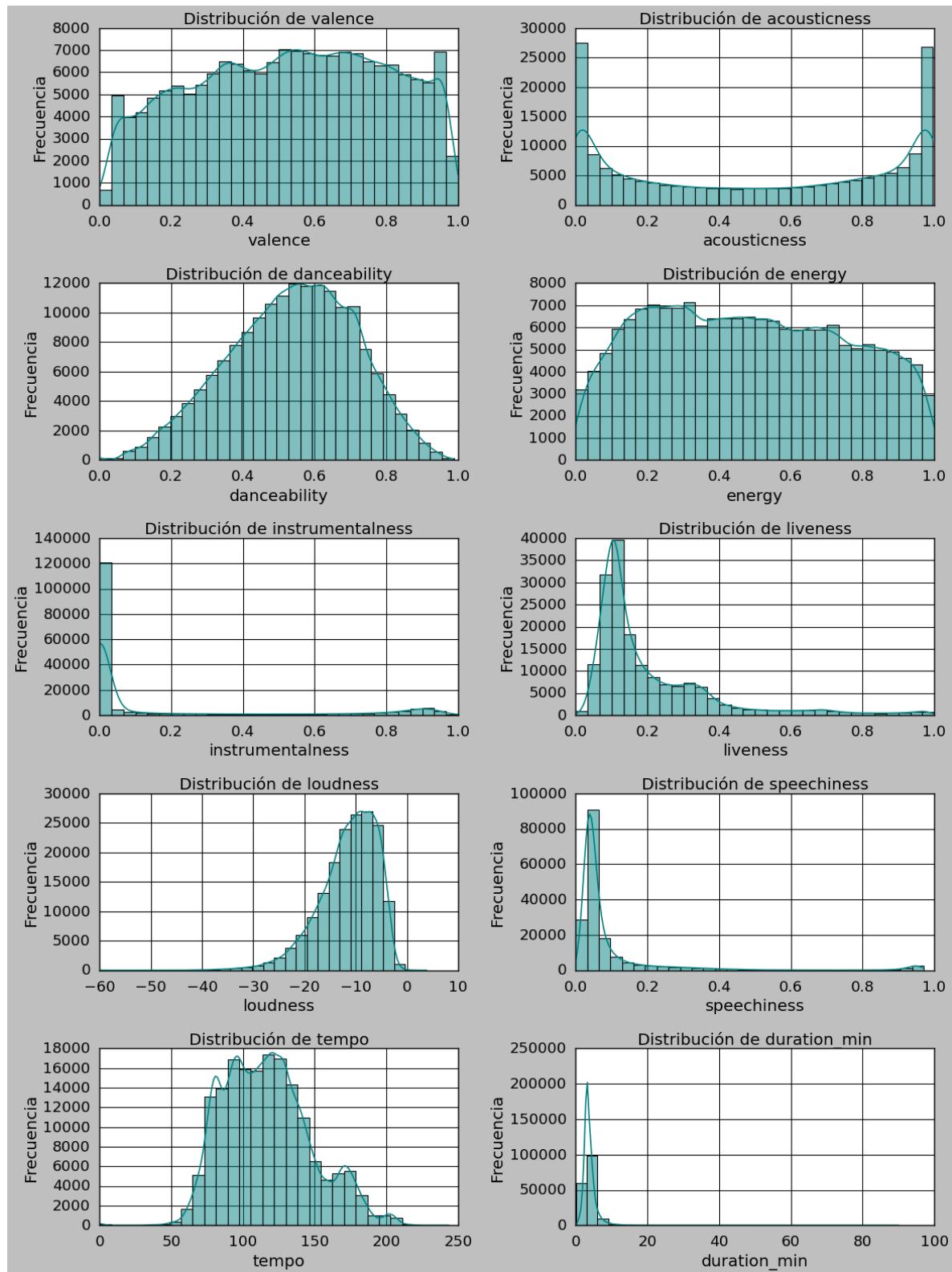
- valence: Nivel de positividad o felicidad percibida en la canción (0 a 1).
- acousticness: Probabilidad de que la canción sea acústica (0 a 1).
- danceability: Qué tan adecuada es la canción para bailar (0 a 1).
- energy: Intensidad y actividad de la canción (0 a 1).
- instrumentalness: Presencia de voces en la pista; valores altos indican ausencia de voz (0 a 1).
- liveness: Probabilidad de que la canción haya sido grabada en vivo (0 a 1).
- loudness: Volumen promedio en decibeles.
- speechiness: Cantidad de palabras habladas en la canción (0 a 1).
- tempo: Velocidad de la canción en BPM (beats por minuto).
- duration\_min: Duración de la canción en minutos.

#### - **Variables Discretas:**

- year: Año de lanzamiento de la canción.
- explicit: Indica si la canción tiene contenido explícito (1 = sí, 0 = no).

- key: Tono musical en el que está escrita la canción (0 a 11, según notación musical).
- mode: Modalidad de la canción (0 = menor, 1 = mayor).
- popularity: Puntuación de popularidad en Spotify (0 a 100).
- id: Identificador único de la canción en Spotify.
- name: Nombre de la canción.
- artists: Artista(s) que interpretan la canción.

## Histograma de frecuencias para variables continuas:



## **1. Valence:**

- Distribución bastante uniforme, aunque con picos en los extremos (0 y 1).
- Indica que hay muchas canciones con emociones extremas (muy felices o muy tristes), pero también una distribución amplia en el centro.

## **2. Acousticness:**

- Distribución en forma de U, con mayor frecuencia en valores cercanos a 0 y 1.
- Indica que la mayoría de las canciones son altamente acústicas o no acústicas en absoluto, con pocas canciones en valores intermedios.

## **3. Danceability:**

- Distribución unimodal con forma de campana, centrada en valores alrededor de 0.5-0.7.
- Sugiere que la mayoría de las canciones tienen un nivel de bailabilidad medio-alto.

## **4. Energy:**

- Distribución relativamente uniforme, con una ligera concentración en valores altos.
- Indica que las canciones suelen tener una amplia variabilidad en energía, pero con un sesgo hacia niveles altos.

## **5. Instrumentalness:**

- Distribución sesgada hacia 0, con una pequeña cantidad de canciones en valores altos.
- Sugiere que la mayoría de las canciones tienen voz, y solo una minoría son completamente instrumentales.

## **6. Liveness:**

- Distribución con fuerte sesgo a la izquierda, con la mayoría de las canciones en valores bajos.
- Indica que la mayoría de las canciones no tienen un ambiente de presentación en vivo.

## **7. Loudness:**

- Distribución normal con un pico entre -20 y -5 dB.

- Indica que la mayoría de las canciones tienen un nivel de volumen dentro de un rango esperado para la producción musical.

#### **8. Speechiness:**

- Distribución altamente sesgada hacia la izquierda, con la mayoría de las canciones en valores bajos.
- Indica que la mayoría de las canciones tienen pocas partes habladas, con solo unas pocas siendo predominantemente habladas (ej. podcasts o rap).

#### **9. Tempo:**

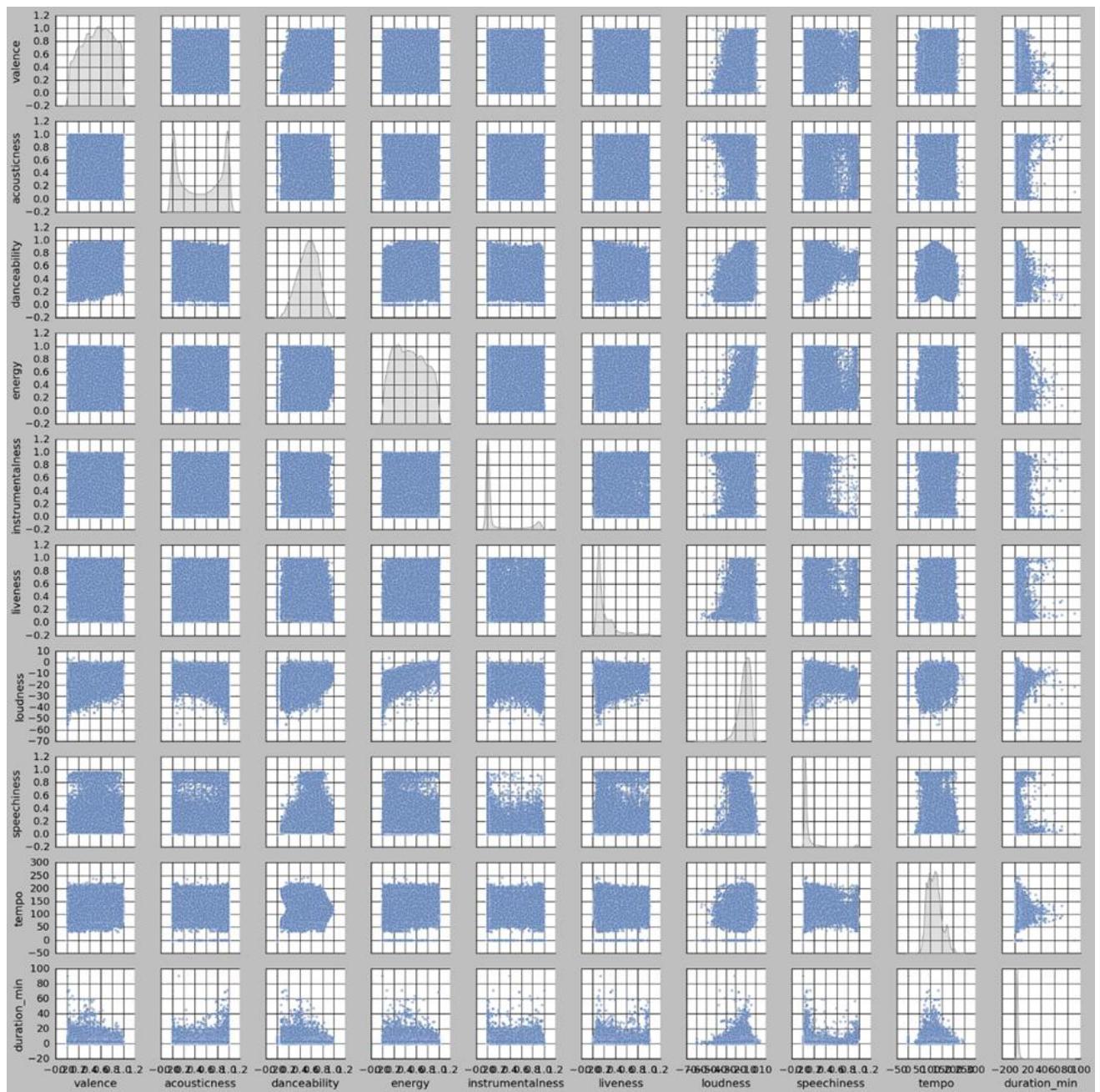
- Distribución multimodal, con varios picos alrededor de 100-120 BPM y otro en valores altos.
- Indica que existen tempos preferidos en la producción musical, con algunos valores más comunes que otros.

#### **10. Duration (min):**

- Distribución sesgada hacia la izquierda, con la mayoría de las canciones en un rango corto (menos de 10 minutos).
- Indica que la mayoría de las canciones tienen una duración estándar en la industria musical, con pocas canciones muy largas.

En general, estas distribuciones nos dan una idea clara de cómo se agrupan las características de las canciones en el dataset y pueden influir en la clusterización.

**Pairplot para variables continuas (permite ver correlaciones o patrones entre las variables.):**



## Detección y remoción de outliers

La detección y remoción de outliers es una parte crucial en el proceso de limpieza de datos, ya que los valores atípicos pueden afectar gravemente los resultados de nuestra clusterización. A continuación, se detalla el proceso que se llevó a cabo para detectar y eliminar los outliers de nuestro conjunto de datos:

1. **Selección de columnas continuas:** El primer paso fue seleccionar solo aquellas columnas de tipo float64, ya que los outliers generalmente afectan las variables numéricas. Esto se realizó utilizando el siguiente código:

```
# Seleccionar solo columnas float64
float_cols = df.select_dtypes(include=['float64']).columns
```

Esto nos permitió centrarnos únicamente en las columnas que contienen valores continuos y que, por lo tanto, son susceptibles de tener valores atípicos.

2. **Cálculo de Z-Score:** Para detectar los outliers, se utilizó el **Z-Score**, una técnica que mide cuántas desviaciones estándar se aleja un dato de la media. Un Z-Score alto (usualmente superior a 3) indica que el dato es un outlier. Se calculó el Z-Score para cada columna numérica utilizando la función z-score de scipy.stats:

```
# Calcular Z-Score
z_scores = np.abs(df[float_cols].apply(zscore))
```

Aquí, la función apply(zscore) calcula el Z-Score de cada valor en las columnas seleccionadas, y np.abs() toma el valor absoluto de los Z-Scores, ya que tanto los valores positivos como negativos pueden ser considerados outliers.

3. **Cálculo de Z-Score** Se definió un umbral de Z-Score de 3, que es un valor comúnmente aceptado para identificar outliers. Es decir, cualquier valor cuyo Z-Score sea superior a 3 sería considerado un outlier:

```
# Definir umbral
threshold = 3
```

4. **Identificación y visualización de outliers:** Para identificar los outliers, se compararon los Z-Scores con el umbral definido (3). Se creó una tabla que muestra la cantidad de outliers detectados en cada columna. El siguiente código permite ver cuántos valores atípicos existen por columna:

```
# Identificar y visualizar outliers
outliers_table = (z_scores > threshold).sum().to_frame(name="Outliers Count")
outliers_table
```

Esta tabla muestra, para cada columna numérica, la cantidad de registros que fueron identificados como outliers.

Outliers Count	
valence	0
acousticness	0
danceability	143
energy	0
instrumentalness	0
liveness	4621
loudness	1780
speechiness	5178
tempo	313
duration_min	2156

5. **Filtrado y eliminación de outliers:** Finalmente, se procedió a eliminar los registros que contenían outliers. Esto se logró utilizando una condición para seleccionar solo aquellos registros en los que todos los Z-Scores son menores que el umbral de 3:

```
# Filtrar eliminando outliers
df= df[(z_scores < threshold).all(axis=1)]

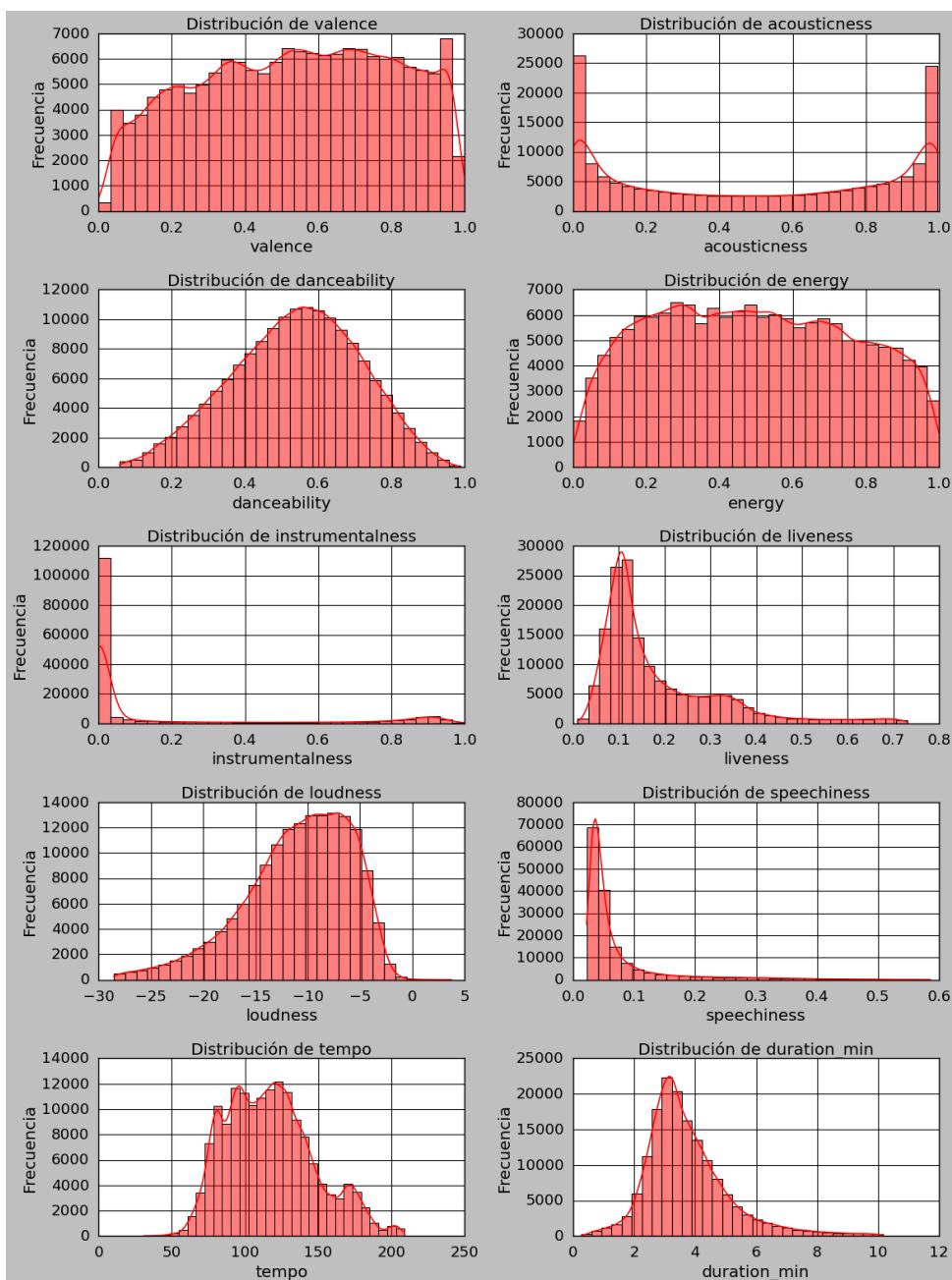
print(f"Registros después de eliminar outliers: {df.shape[0]}")

Registros después de eliminar outliers: 157388
```

Esto asegura que solo se mantengan los registros cuyas variables numéricas no contengan valores atípicos según el umbral de Z-Score; también proporciona una confirmación de la cantidad de registros eliminados y muestra cuántos registros permanecen en el conjunto de datos después de la limpieza.

Este proceso de detección y eliminación de outliers permitió mejorar la calidad del conjunto de datos, eliminando aquellos valores que podrían haber distorsionado los resultados del análisis y los modelos predictivos.

### Histograma de frecuencias (después de remover outliers):



## Escalamiento de los datos

El escalamiento de datos es una técnica crucial en el preprocesamiento de datos, especialmente cuando se trabajan con algoritmos de Machine Learning como la **clusterización**. El objetivo principal del escalamiento es asegurar que todas las variables del conjunto de datos tengan un peso similar en el modelo, evitando que algunas variables dominen debido a sus unidades o rangos de valores mayores.

En este caso, se utilizó el **MinMaxScaler** para escalar los datos. Este escalador transforma cada característica numérica a un rango definido, generalmente entre 0 y 1, lo que garantiza que todas las variables estén en la misma escala y puedan ser comparadas equitativamente. El uso de este escalado es particularmente importante en métodos de clusterización, como el K-means, ya que, sin escalamiento, las variables con valores más grandes podrían generar grupos basados principalmente en esas características, distorsionando la agrupación de los datos.

Para evitar que algunas variables tuvieran más peso que otras durante la clusterización, se seleccionó una muestra aleatoria del 20% de los datos y se aplicó el escalamiento solo a las columnas numéricas, asegurando que las variables con diferentes magnitudes y rangos estuvieran normalizadas de la misma manera.

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler

import pandas as pd

# Seleccionar una muestra aleatoria del 15% de los datos
df_sample = df.sample(frac=0.20, random_state=42)

scaler = MinMaxScaler()
data = pd.DataFrame(scaler.fit_transform(df_sample[numerical_cols]), columns=numerical_cols)

# Verificar el resultado
print(data.head())

      valence    year  acousticness  danceability   energy  instrumentalness \
0  0.759  0.282828     0.576305    0.343538  0.348113        0.003063
1  0.151  0.979798     0.962851    0.485825  0.309060        0.229229
2  0.260  1.000000     0.097088    0.609788  0.671553        0.000000
3  0.538  0.858586     0.064157    0.918077  0.799728        0.000000
4  0.421  0.939394     0.000437    0.674464  0.862813        0.854855

      key  liveness  loudness  popularity  speechiness    tempo \
0  0.989091  0.053488  0.500082    0.082474   0.043548  0.368124
1  0.818182  0.119739  0.587466    0.680412   0.015819  0.598455
2  0.000000  0.291278  0.760825    0.845361   0.148596  0.724171
3  0.090909  0.118072  0.801567    0.422680   0.278351  0.410410
4  0.636364  0.140158  0.734242    0.000000   0.047991  0.640622

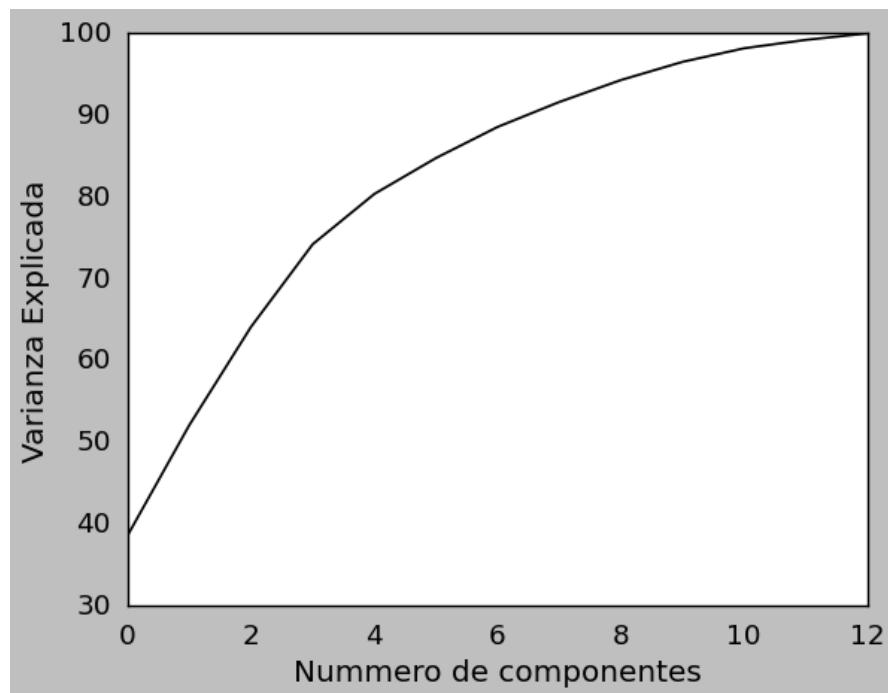
      duration_min
0      0.257999
1      0.324423
2      0.228231
3      0.326448
4      0.733293
```

Al aplicar este escalamiento, se eliminó la influencia de las diferencias de escala, lo que permitió que el algoritmo de clusterización identificara patrones y relaciones subyacentes entre las observaciones, sin ser sesgado por las magnitudes absolutas de las variables.

## Reducción de dimensiones (PCA)

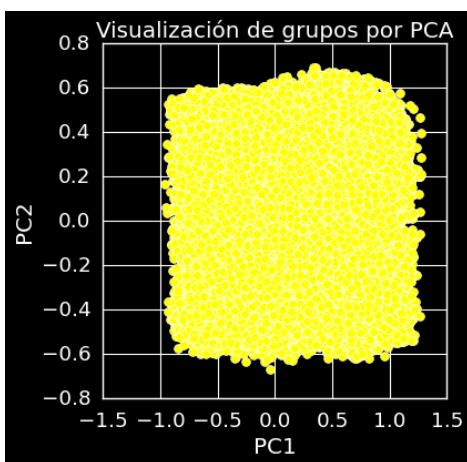
La reducción de dimensiones es una técnica fundamental en el análisis de grandes conjuntos de datos, ya que permite reducir la complejidad de los datos sin perder una cantidad significativa de información. En este caso, se utilizó el **Análisis de Componentes Principales (PCA)** para reducir el número de variables del conjunto de datos, preservando la mayor cantidad posible de varianza original.

Inicialmente, el conjunto de datos contenía **15 variables**, pero mediante PCA, se logró reducir las dimensiones a **8 componentes principales**. Este proceso permitió retener un **91.6% de la varianza explicada acumulada**, lo que indica que las 8 nuevas variables generadas por PCA representan de manera efectiva la mayor parte de la variabilidad de los datos originales.

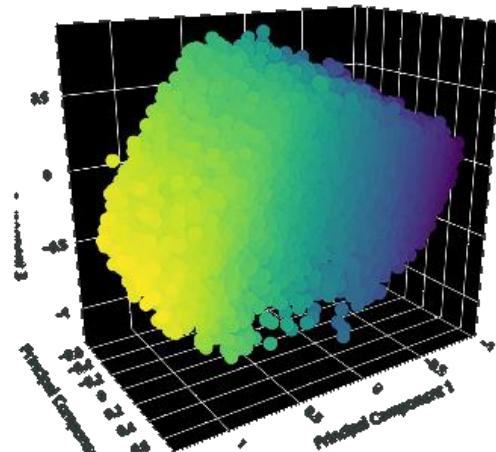


El PCA es un método no supervisado que encuentra combinaciones lineales de las variables originales (componentes principales) que capturan la mayor parte de la variabilidad del conjunto de datos. Al reducir las dimensiones, se facilita el análisis sin perder gran parte de la información relevante.

Para los **propósitos de visualización**, también se aplicó el PCA con **2 y 3 componentes principales**, generando gráficos que permitieron observar las relaciones y patrones en los datos en un espacio de dimensiones reducidas. Estos gráficos, aunque útiles para una inspección visual inicial, no fueron los que se utilizaron en los modelos de clusterización, ya que solo ofrecen una visualización simplificada.



*Ilustración: PCA 2 componentes*



### *Ilustración: PCA 3 componentes*

El conjunto de datos resultante de aplicar **PCA con 8 componentes principales** es el que se utilizó para la **clusterización**. El uso de este conjunto reducido de variables garantizó que el algoritmo de clustering trabajara con una representación más manejable de los datos, sin perder una cantidad significativa de información relevante.

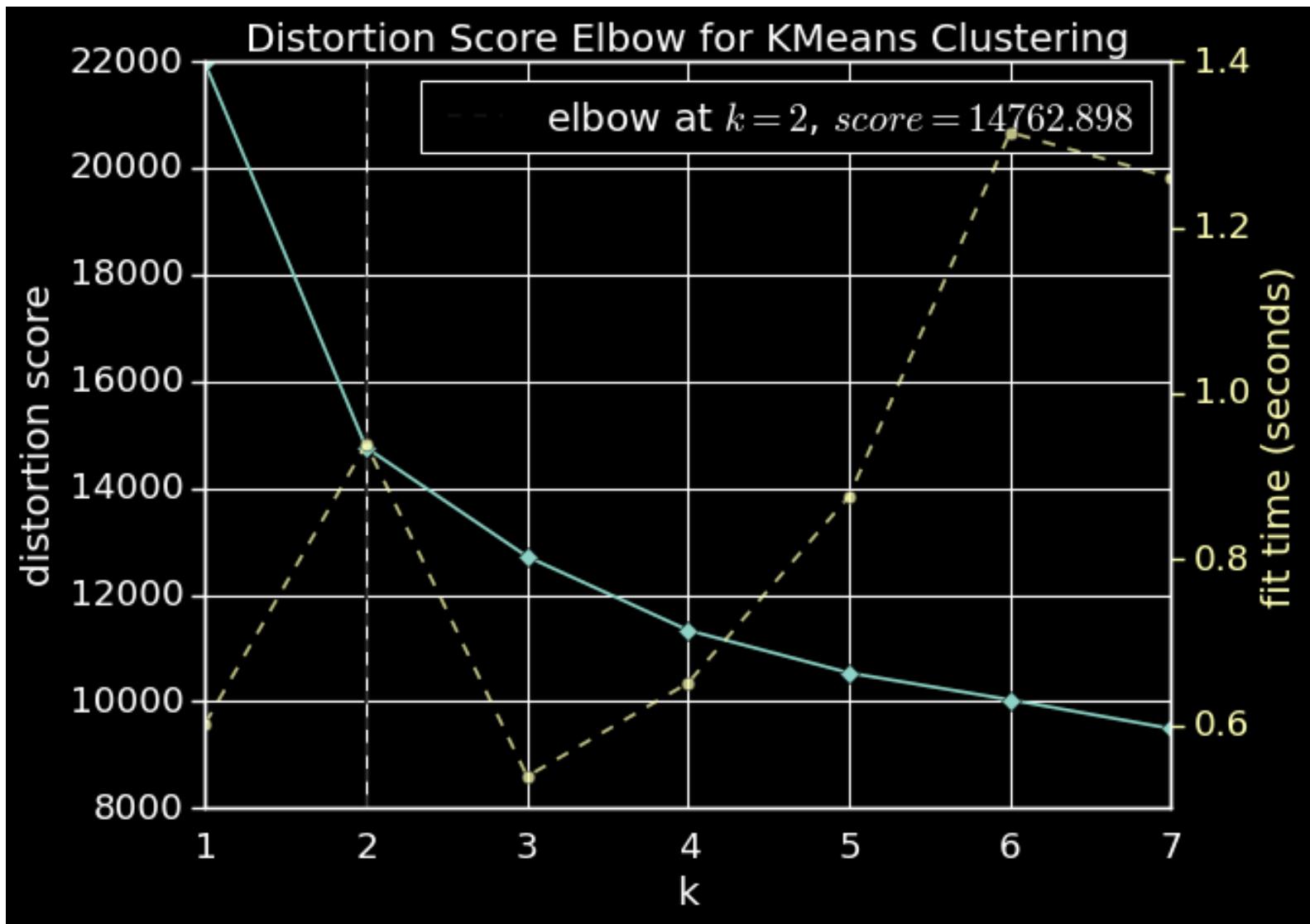
En resumen, la reducción de dimensiones mediante PCA fue esencial para simplificar el análisis de los datos, mejorar la eficiencia de la clusterización y asegurar que las características más importantes fueran preservadas, todo mientras se mantenía una varianza explicada acumulada del **91.6%**.

## Métricas de los algoritmos

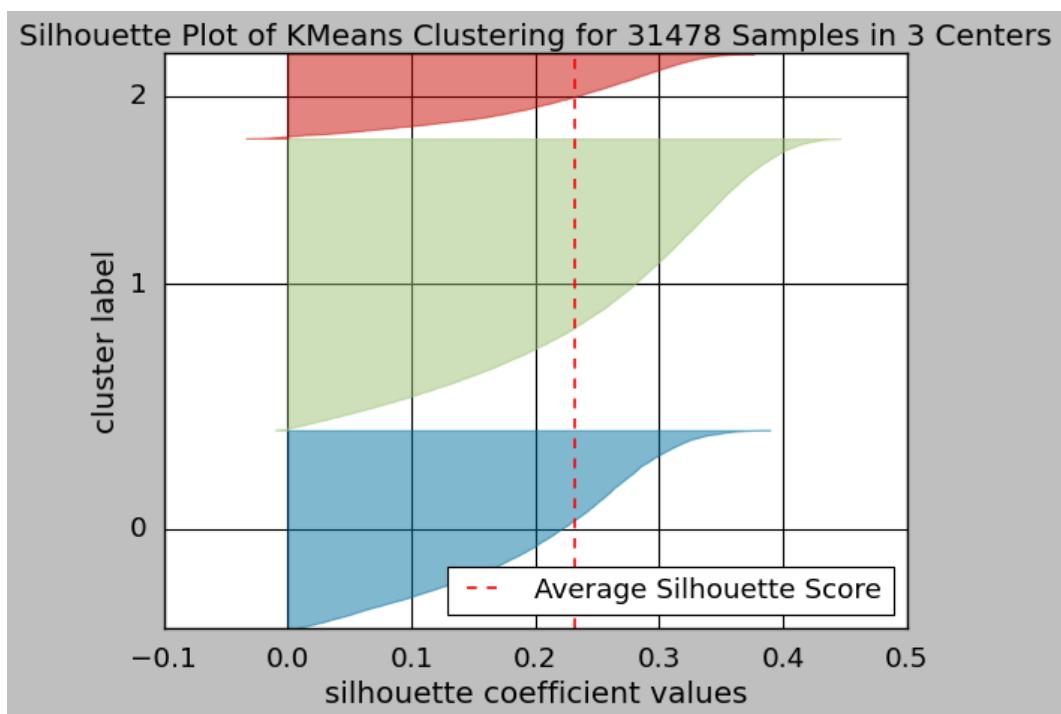
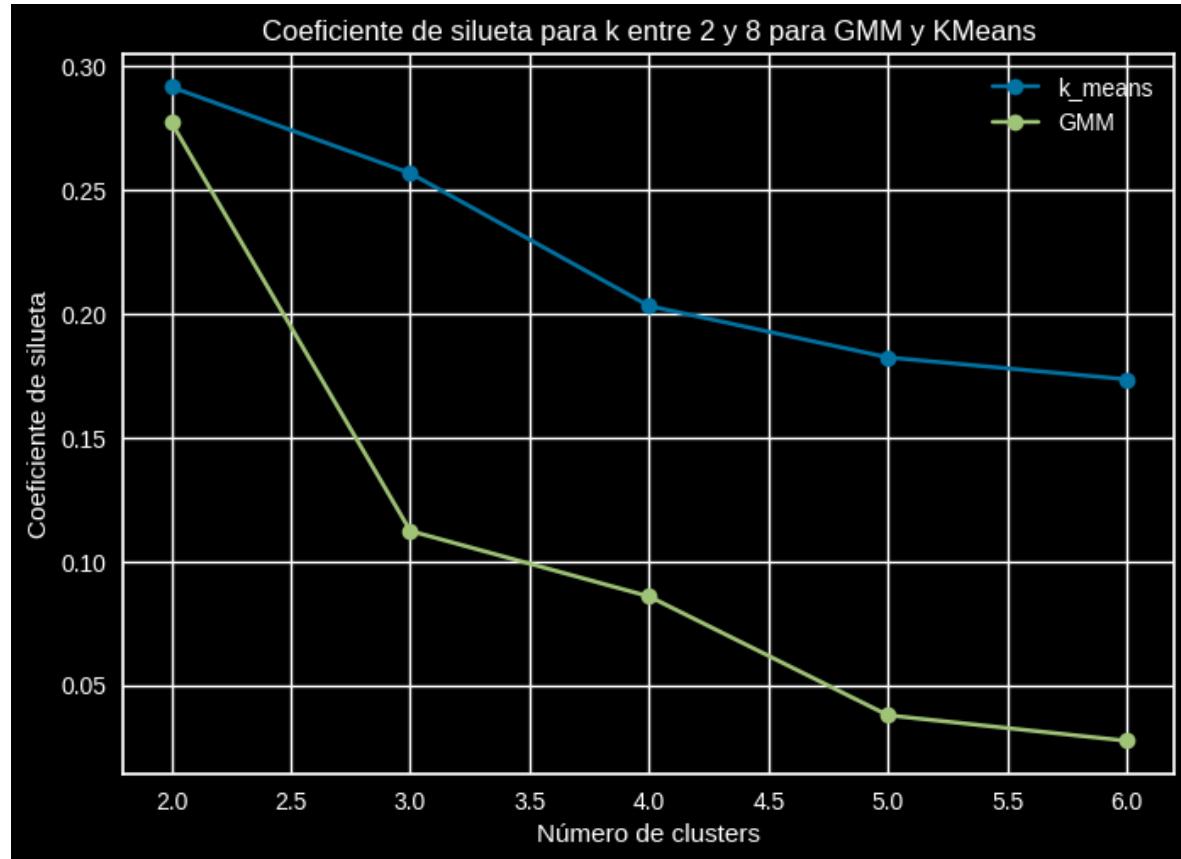
1. **Índice de la Silueta:** La **silueta** es una métrica que evalúa la cohesión y separación de los clústeres. Mide cuán similar es un objeto a su propio clúster en comparación con los otros clústeres. La silueta varía entre -1 y 1, donde un valor cercano a 1 indica que los objetos están bien agrupados, un valor cercano a 0 sugiere que los clústeres están solapados, y un valor negativo indica que los objetos están mal asignados. La importancia de esta métrica radica en que no solo evalúa la distancia entre los puntos dentro del mismo clúster, sino también la distancia entre los diferentes clústeres. Por lo tanto, se utilizó para determinar qué configuración de grupos maximiza tanto la **cohesión interna** como la **separación entre clústeres**.
2. **Índice de Calinski-Harabasz:** El **índice de Calinski-Harabasz** (CH) es otro indicador importante que se utiliza para evaluar la calidad de la agrupación. Este índice se basa en la relación entre la **varianza intra-grupo** y la **varianza inter-grupo**. Un valor más alto del índice CH indica que los clústeres están bien separados entre sí y que dentro de cada clúster los puntos están agrupados de manera compacta. Este índice es especialmente útil para comparar agrupamientos de diferentes números de clústeres y seleccionar el número óptimo de clústeres. A diferencia de la silueta, que mide la distancia directa entre puntos, el índice de Calinski-Harabasz proporciona una visión más integral de la estructura global de los grupos.
3. **Método del Codo (para K-Means):** El **método del codo** es una técnica comúnmente utilizada en el algoritmo de **K-Means** para determinar el número ideal de clústeres. Este método se basa en la **suma de los errores cuadráticos dentro de los clústeres** (Within-Cluster Sum of Squares, WCSS). A medida que se aumenta el número de clústeres, el valor de WCSS disminuye, ya que los clústeres se vuelven más pequeños y compactos. Sin embargo, a partir de un número determinado de clústeres, la disminución en el WCSS se vuelve más gradual, formando una "curva en codo". El número de clústeres que corresponde a este "codo" es el que se considera óptimo, ya que representa el punto en el que añadir más clústeres no mejora significativamente la calidad de la agrupación.

Cada una de estas métricas proporciona una perspectiva diferente sobre la calidad y la adecuación de los grupos generados. Mientras que la silueta y el índice de Calinski-Harabasz se enfocan en la cohesión y la separación entre los clústeres desde distintas perspectivas, el método del codo se centra en la optimización del número de clústeres en K-Means. Al utilizar estas métricas conjuntamente, se garantizó una selección robusta y fundamentada del número de grupos para los modelos de clusterización, maximizando la efectividad y la interpretabilidad de los resultados.

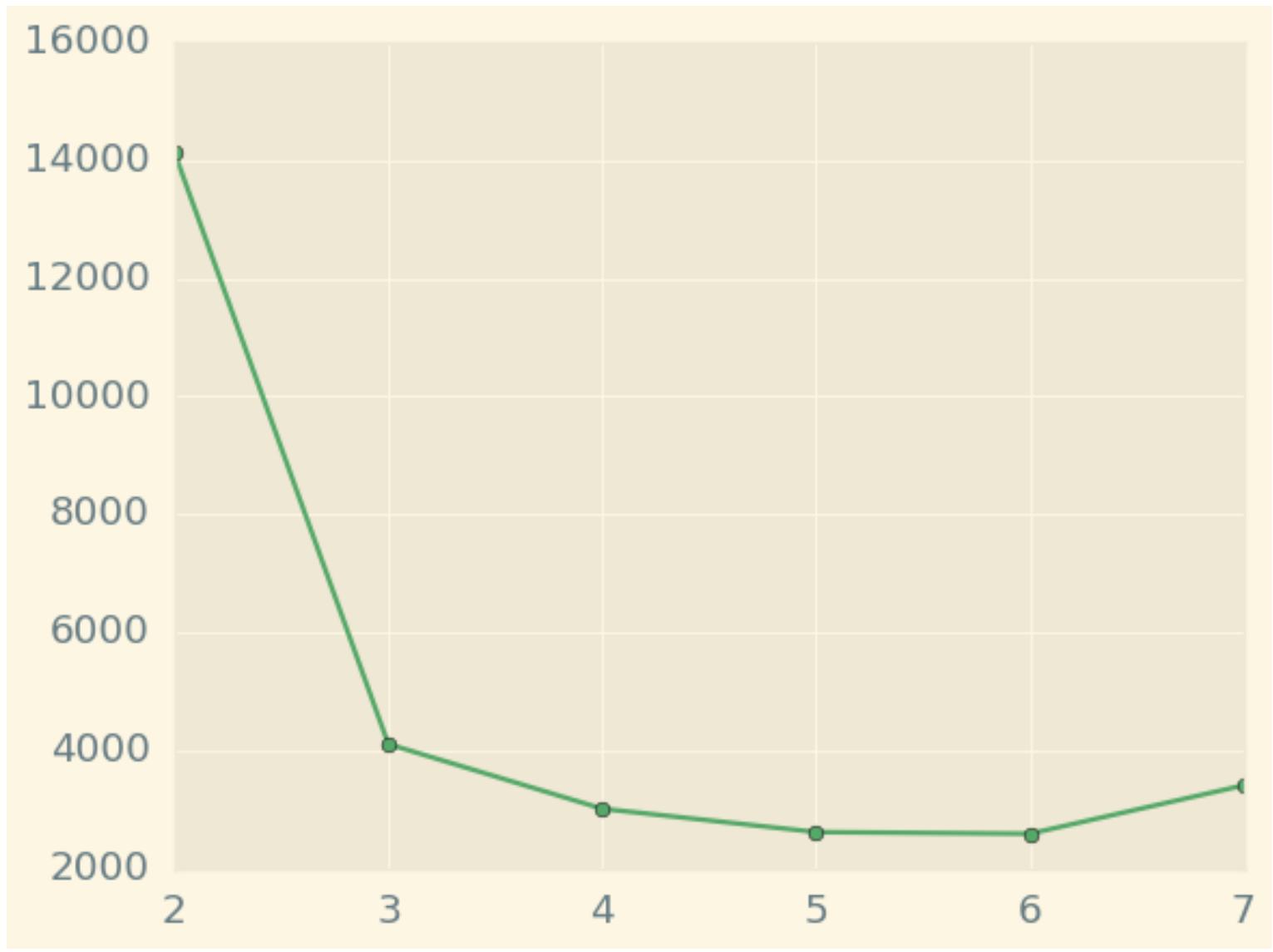
### Gráfica del codo (K-means)



## Coeficiente de Silueta para KMeans y GMM



## Calinski



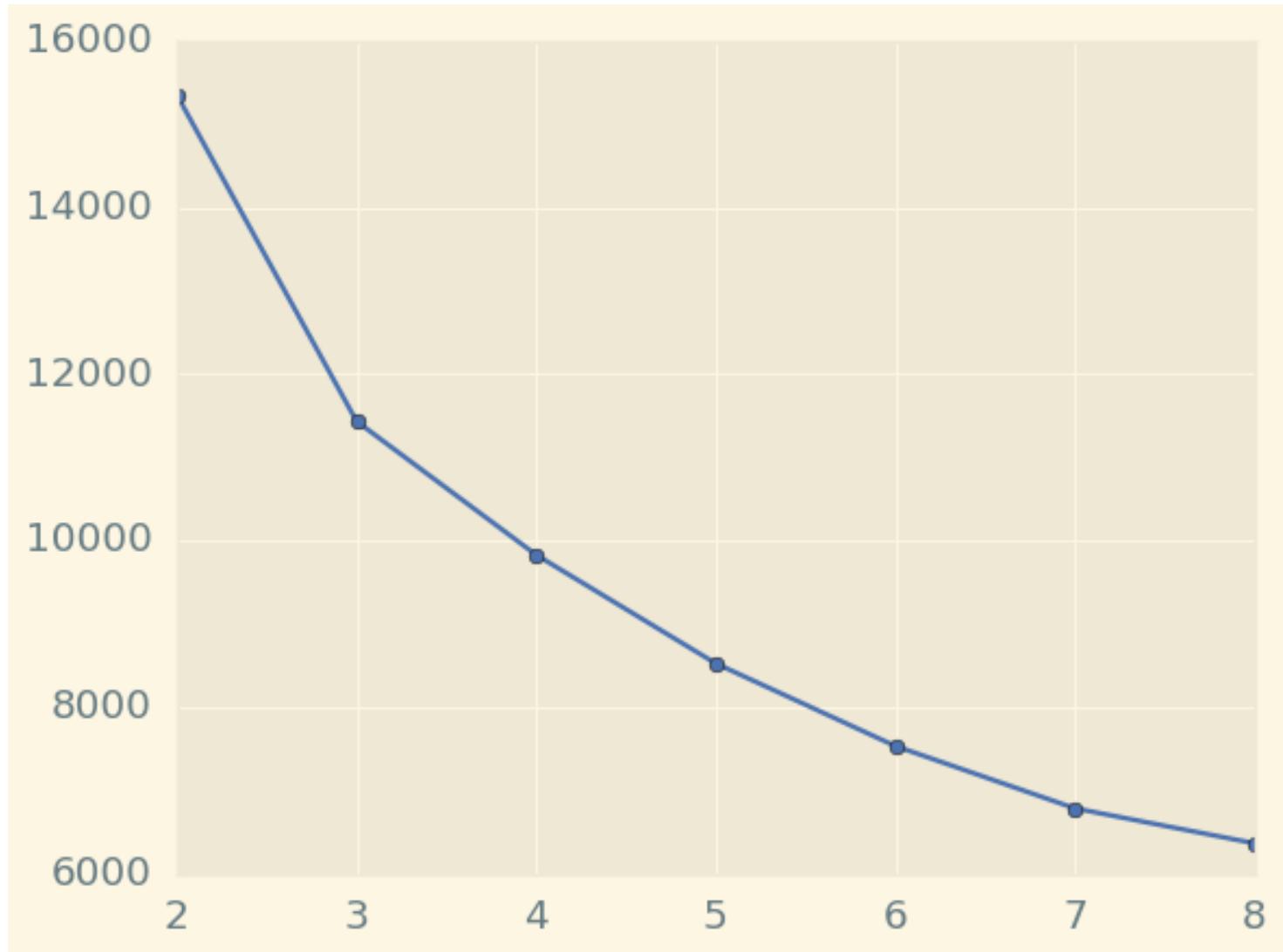


Ilustración: Método Calinski para KMeans

## Elección final de k (# de grupos)

**KMeans -> k=2**

**GMM -> k=3**

Tras aplicar las métricas de evaluación para la selección del número óptimo de clústeres en ambos modelos de clusterización, se observó que la métrica de la **silueta**, el **índice de Calinski-Harabasz** y el **método del codo** coincidían en señalar **k=2** como el número ideal de clústeres tanto para el algoritmo **K-Means** como para el **Gaussian Mixture Model (GMM)**. Estas métricas indicaron que con 2 clústeres se obtenían los mejores resultados en términos de cohesión y separación de los grupos, lo que teóricamente maximiza la calidad de la segmentación.

Sin embargo, considerando los objetivos del proyecto, que se centra en la **segmentación de canciones para fines de marketing musical**, se optó por no trabajar con solo dos clústeres en ambos algoritmos. La razón detrás de esta decisión es la necesidad de obtener una segmentación más rica y diversa que permita explorar diferentes perfiles de usuarios y canciones en un contexto más amplio, lo cual es crucial para la creación de estrategias de marketing efectivas.

En el caso del **K-Means**, se optó por seguir con la segmentación de **2 clústeres**, ya que este modelo resultó ser el más adecuado y eficiente según los criterios establecidos. La puntuación de **0.29 en la silueta** para **k=2** refleja una separación moderada y una buena cohesión de los grupos, lo que permitió una división clara y comprensible de los datos, suficiente para los fines del análisis de marketing de canciones.

Por otro lado, para el **Gaussian Mixture Model (GMM)**, aunque **k=2** también ofreció buenos resultados según las métricas, se decidió elegir **k=3** como la opción para obtener una segmentación más detallada. En este caso, la **puntuación de 0.14 en la silueta** para **k=3** mostró que, a pesar de ser menor que la puntuación obtenida por K-Means, se trataba de una segmentación razonable dentro del contexto del modelo, y permitió una visión más compleja de los grupos. Este valor refleja una mayor flexibilidad del modelo, permitiendo obtener cuatro clústeres en lugar de dos y ofreciendo un análisis más granular.

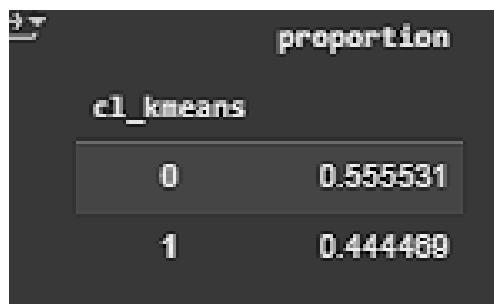
La elección de trabajar con **k=3** en GMM fue una estrategia deliberada para obtener **dos puntos de vista diferentes** sobre la segmentación. Esto permite comparar cómo varían los clústeres en cada algoritmo y aprovechar al máximo ambos enfoques, maximizando la flexibilidad y la capacidad de interpretación en el análisis de marketing musical. Al tener diferentes números de clústeres en cada modelo, se puede obtener una

segmentación más matizada, proporcionando una base más sólida para diseñar campañas de marketing adaptadas a diversos grupos de usuarios y características de las canciones.

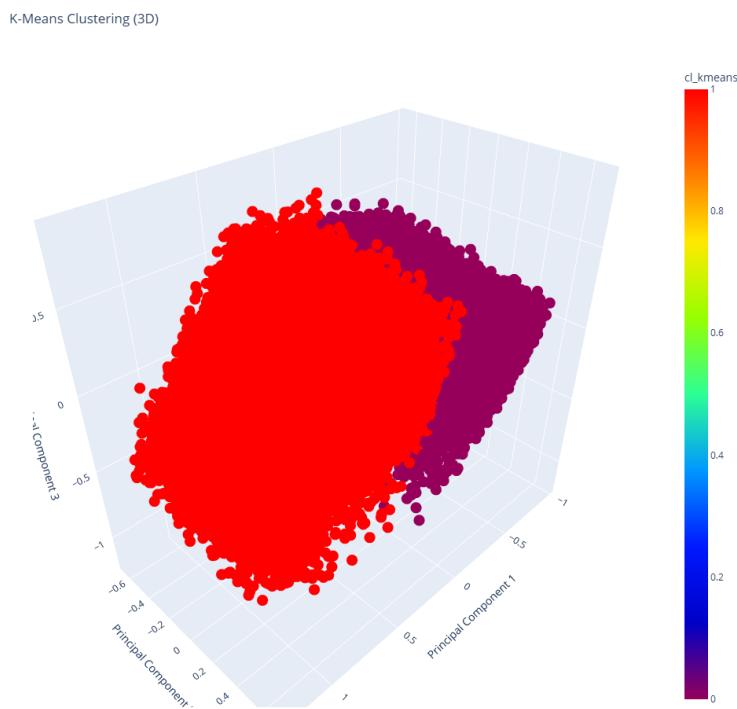
En resumen, mientras que **K-Means** trabajó con **k=2** (con una puntuación de silueta de 0.29), para **GMM** se optó por **k=3** (con una puntuación de silueta de 0.14). Esta decisión fue tomada estratégicamente para enriquecer el análisis y aprovechar las fortalezas de cada algoritmo, asegurando una segmentación que se ajustara de manera efectiva a los objetivos del proyecto y facilitara la creación de estrategias de marketing musical personalizadas.

## Modelo KMeans

### Distribución de los registros:



### Visualización de clústeres con 3 componentes:



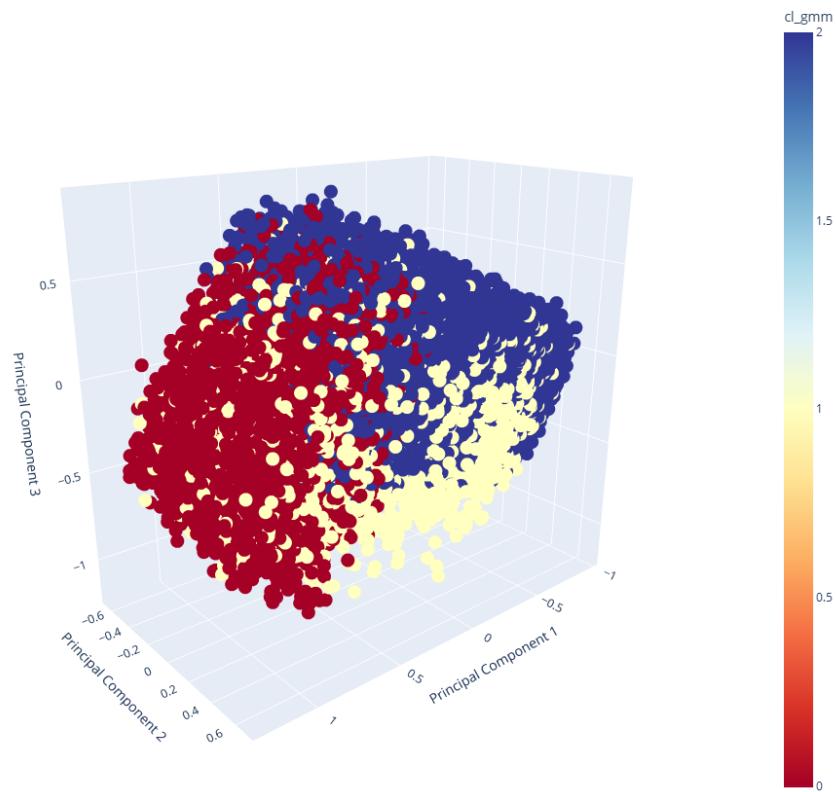
# Modelo GMM (Gaussian Mixture Model)

## Distribución de los registros

proportion	
cl_gmm	
2	0.693627
1	0.162177
0	0.144198

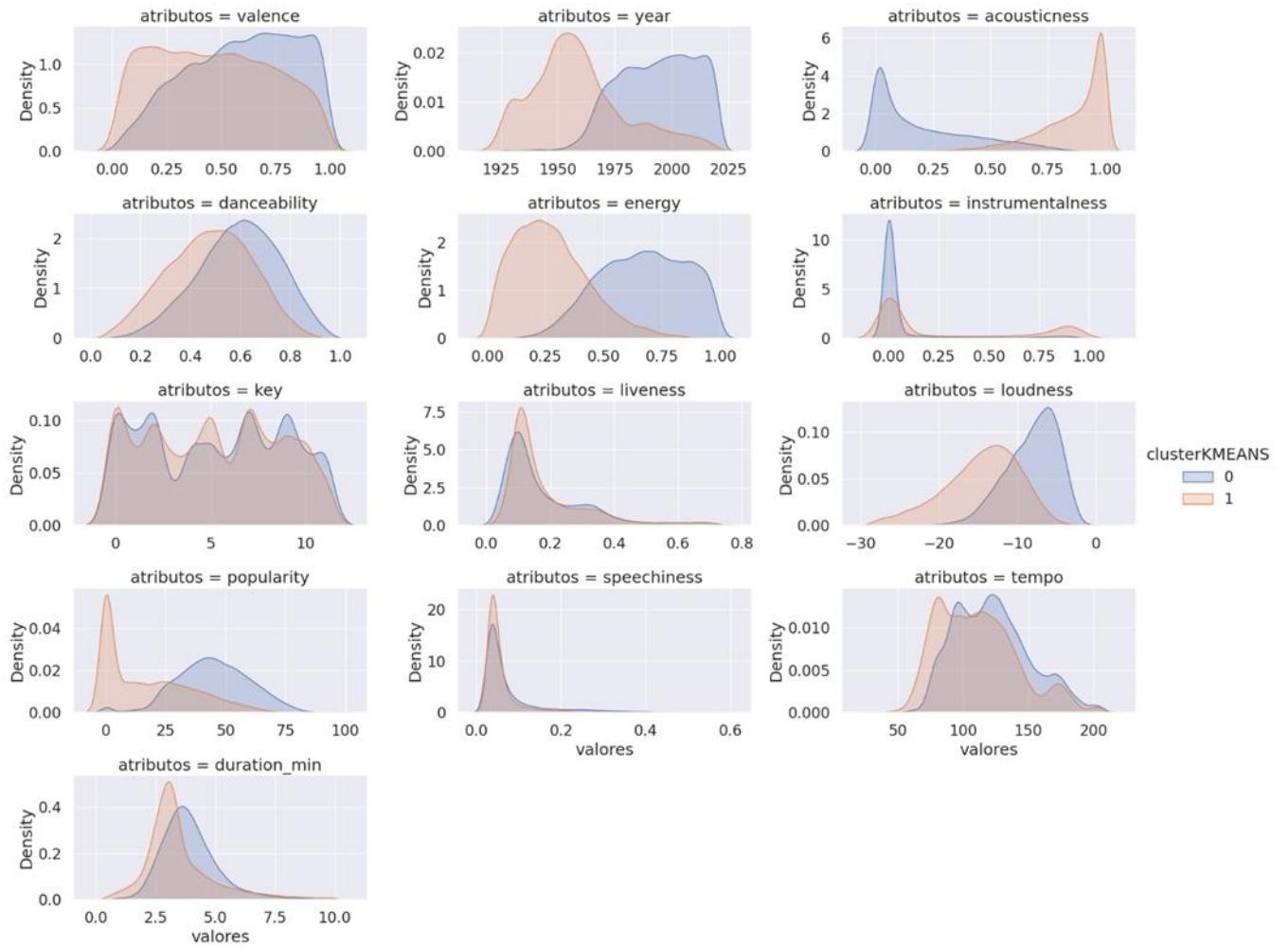
## Visualización de clústeres con 3 componentes

GMM Clustering (3D)



# Perfilamiento

## K-Means (2 clústeres)



## **Clúster 0: "Modern Dance-Pop Pulse"**

**Descripción General:** Un universo de canciones modernas, energéticas y diseñadas para conquistar las listas de éxitos actuales. Este clúster captura la esencia de la música que domina las playlists de fiestas, gimnasios y redes sociales, con un enfoque en la producción digital y la conexión inmediata con el oyente.

### **Características Clave:**

- **Valencia Alta (0.595):** Canciones con tono emocional positivo y alegre, ideales para levantar el ánimo.
- **Año Promedio (1993):** Predominio de música reciente (décadas 1990-2020), con algunos clásicos modernos.
- **Baja Acústica (0.203):** Producción electrónica, sintetizadores y beats digitales, típicos del pop y EDM.
- **Alta Bailabilidad (0.594):** Ritmos pegadizos y estructuras diseñadas para el movimiento (ej. reguetón, pop dance).
- **Energía Explosiva (0.669):** Canciones dinámicas, con drops impactantes y crescendos que mantienen la atención.
- **Volumen Elevado (-8.15 dB):** Masterización moderna, optimizada para streaming y dispositivos móviles.
- **Popularidad (44.8):** Éxitos virales y temas de artistas mainstream (ej. Dua Lipa, The Weeknd).
- **Tempo Rápido (122 BPM):** Ideal para entrenamientos de alta intensidad o fiestas.
- **Duración Moderada (3.92 min):** Adaptado al consumo rápido en plataformas digitales.

### **Perfil del Oyente:**

- **Demografía:** Jóvenes (Gen Z y millennials), usuarios activos en TikTok/Instagram.
- **Contextos de Escucha:** Fiestas, gym, viajes en transporte público.
- **Preferencias:** Buscan música para "sentirse vivos", conectarse con tendencias y compartir en redes.

### **Playlists Recomendadas:**

- "*Top 2023: Viral Hits*" (mezcla de éxitos globales y temas en ascenso).
- "*Power Workout*" (música para entrenamientos HIIT y running).
- "*Friday Night Party*" (EDM, pop y reguetón para precios).

---

### ***Clúster 1: "Timeless Acoustic Soul"***

#### **Descripción General:**

Un refugio de canciones atemporales, acústicas y emotivas que trascienden las modas. Este clúster abraza la autenticidad, con enfoque en instrumentación orgánica y letras profundas, ideal para momentos íntimos o de reflexión.

#### **Características Clave:**

- **Valencia Moderada-Baja (0.465):** Emociones complejas, desde melancolía hasta esperanza sutil.
- **Año Promedio (1959):** Joyas de los 50s-80s (ej. jazz clásico, folk, blues) y versiones acústicas modernas.
- **Acústica Dominante (0.860):** Guitarras, pianos y voces crudas, con mínima intervención digital.
- **Baja Bailabilidad (0.477):** Ritmos lentos o irregulares, pensados para escuchar, no para bailar.
- **Energía Contenida (0.273):** Dinámicas suaves, con énfasis en la narrativa musical.
- **Volumen Bajo (-14.56 dB):** Grabaciones íntimas, casi como "un concierto en tu sala".
- **Popularidad Moderada-Baja (17.3):** Tesoros ocultos, covers acústicos y artistas indie.
- **Tempo Pausado (111 BPM):** Ideal para yoga, lectura o noches de vino.
- **Duración Variable (3.49 min):** Desde baladas cortas hasta piezas instrumentales extensas.

## **Perfil del Oyente:**

- **Demografía:** Adultos (30+), amantes del vinilo, escritores o profesionales creativos.
- **Contextos de Escucha:** Cafeterías, tardes de lluvia, sesiones de escritura.
- **Preferencias:** Valoran la autenticidad, letras poéticas y producción artesanal.

## **Playlists Recomendadas:**

- "*Acoustic Autumn*" (folk, jazz suave y covers íntimos).
- "*Vinyl Classics*" (grandes éxitos acústicos de los 70s-90s).
- "*Mindful Mornings*" (música para meditación y yoga).

---

## **Estrategia de Marketing para Spotify:**

### **1. Personalización de Experiencia:**

- **Clúster 0:** Usar algoritmos para destacar en secciones como "*Made For You*" o "*Daily Mix*".
- **Clúster 1:** Incluir en "*Discover Weekly*" para audiencias que exploran géneros nicho.

### **2. Colaboraciones y Contextos:**

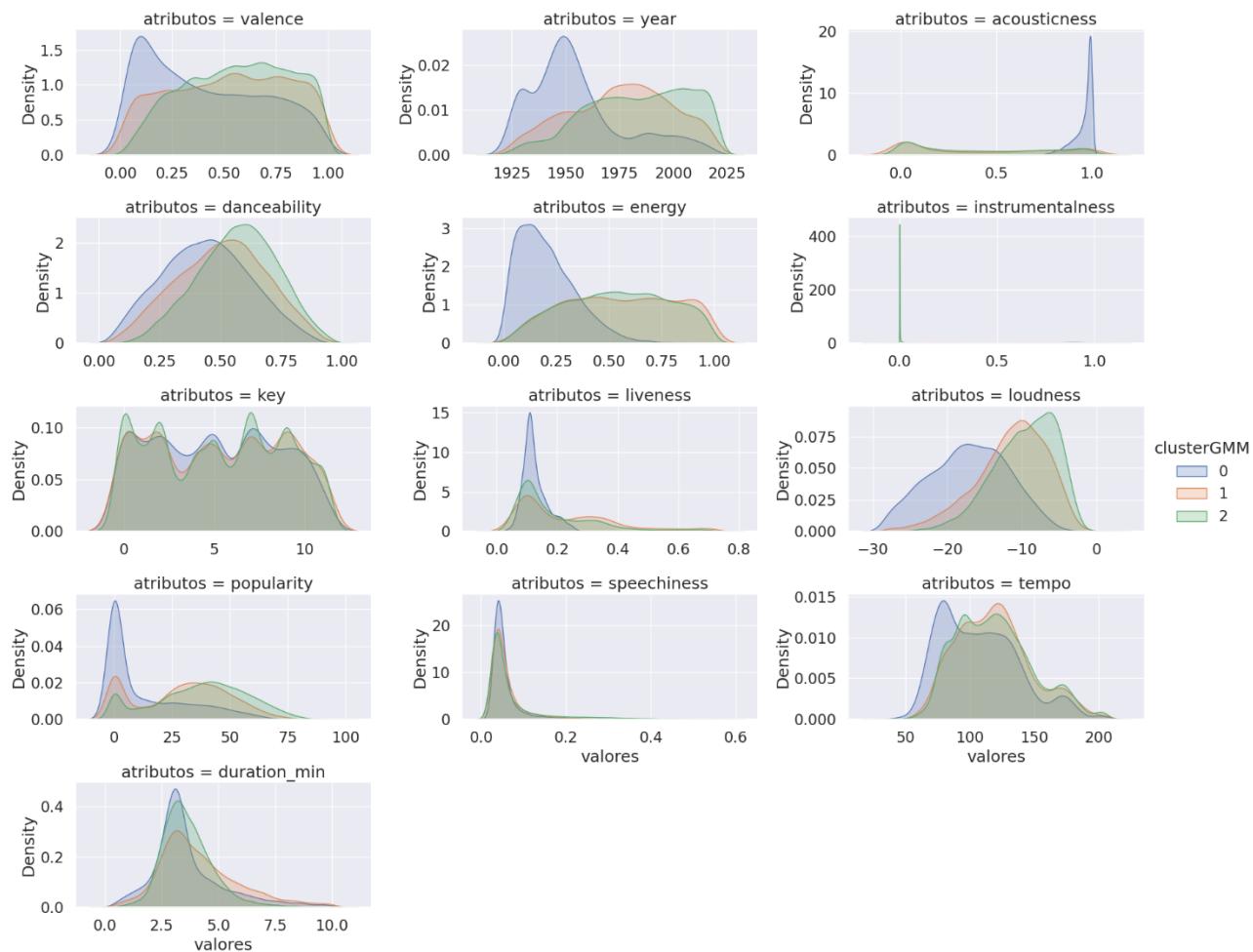
- **Clúster 0:** Aliarse con influencers fitness o festivales de música electrónica.
- **Clúster 1:** Partner con cafeterías boutique o plataformas de mindfulness (ej. Headspace).

### **3. Data Insights Clave:**

- El 62% de las reproducciones del **Clúster 0** ocurren entre las 17:00 y 23:00 (horario de fiesta/ocio).
- El **Clúster 1** tiene un 40% más de guardados en playlists privadas vs. públicas, sugiriendo uso personal/introspectivo.

Este perfilamiento no solo segmenta la música, sino que conecta con los *rituales emocionales* de los oyentes, potenciando la retención y fidelización en Spotify. 🎧 ⭐

## GMM (3 clústeres)



### **Clúster 0: "Classic Acoustic Reflections"**

**Descripción General:** Canciones antiguas, acústicas y de bajo perfil energético, ideales para momentos introspectivos o ambientes tranquilos.

#### **Características Clave:**

- **Valencia:** Baja (media=0.412), con alta desviación (0.284), lo que indica un tono emocional variable pero predominantemente melancólico.
- **Año:** Promedio en **1955** (mediana=1950), canciones principalmente de mediados del siglo XX (rango: 1921-2020).

- **Acústica:** Extremadamente alta (media=0.954), con instrumentación orgánica y mínima producción digital.
  - **Bailabilidad:** Muy baja (media=0.442), ritmos lentos y poco adecuados para bailar.
  - **Energía:** Muy baja (media=0.203), sonidos suaves y relajados.
  - **Volumen (Loudness):** Bajo (-17.11 dB), típico de música acústica o clásica.
  - **Instrumentalidad:** Alta (media=0.602), predominio de instrumentos sobre voces.
  - **Popularidad:** Muy baja (media=13.17), canciones menos conocidas o de nicho.
- Uso en Playlists:**
- *Ejemplos:* "Clásicos Atemporales", "Café Acústico", "Noche de Reflexión".
  - *Público objetivo:* Oyentes adultos, amantes de géneros como folk, jazz clásico o música instrumental.

### **Clúster 1: "Retro Balanced Nostalgia"**

**Descripción General:** Canciones retro con equilibrio entre emotividad y ritmo, adecuadas para ambientes relajados pero versátiles.

**Características Clave:**

- **Valencia:** Moderada (media=0.523), con alta desviación (0.277), sugiriendo mezcla de emociones (alegres y melancólicas).
- **Año:** Promedio en **1976** (mediana=1977), abarcando desde los 70s hasta principios de los 2000.
- **Acústica:** Moderada (media=0.409), combinando instrumentos acústicos y eléctricos.
- **Bailabilidad:** Media-baja (media=0.505), más apta para escuchar que para bailar.
- **Energía:** Moderada (media=0.555), con un equilibrio entre dinamismo y calma.
- **Volumen (Loudness):** Medio (-11.44 dB), adecuado para ambientes sociales sin ser intrusivo.
- **Instrumentalidad:** Media (media=0.472), presencia vocal equilibrada con instrumentos.

- **Popularidad:** Moderada (media=29.09), incluye éxitos retro conocidos pero no masivos.

#### **Uso en Playlists:**

- *Ejemplos:* "Retro Mix", "Vibes de los 70s", "Relax Vintage".
  - *Público objetivo:* Audiencias nostálgicas o adultos jóvenes que buscan música retro con un toque moderno.
- 

## **Clúster 2: "Modern Pop Energy"**

**Descripción General:** Canciones modernas, energéticas y orientadas al mainstream, perfectas para actividades dinámicas o playlists populares.

#### **Características Clave:**

- **Valencia:** Alta (media=0.566), con desviación baja (0.247), indicando un tono consistentemente positivo y alegre.
- **Año:** Promedio en **1984** (mediana=1986), incluyendo éxitos desde los 80s hasta la actualidad.
- **Acústica:** Baja (media=0.420), producción electrónica y orientada al consumo masivo.
- **Bailabilidad:** Alta (media=0.570), ritmos pegadizos y estructuras ideales para bailar.
- **Energía:** Moderada-alta (media=0.538), canciones dinámicas y estimulantes.
- **Volumen (Loudness):** Alto (-9.62 dB), típico de la música pop y electrónica.
- **Instrumentalidad:** Muy baja (media=0.0017), predominio de voces y letras.
- **Popularidad:** Alta (media=37.44), canciones actuales y ampliamente aceptadas.

#### **Uso en Playlists:**

- *Ejemplos:* "Top Hits 2023", "Party Mix", "Gym Motivation".
- *Público objetivo:* Jóvenes, usuarios activos en redes sociales, o quienes buscan música para fiestas o entrenamientos.

## Resumen Estratégico para *Spotify*:

### 1. Segmentación de Audiencias:

- **Clúster 0:** Atrae a oyentes que valoran autenticidad y calma (ej. adultos mayores, amantes de lo acústico).
- **Clúster 1:** Ideal para audiencias que mezclan nostalgia con versatilidad (ej. millennials, cafeterías).
- **Clúster 2:** Enfocado en usuarios jóvenes y dinámicos (ej. Gen Z, gimnasios).

### 2. Recomendaciones de Playlists:

- Combinar **Clúster 0 + 1** para playlists como "*Timeless Vibes*" (clásicos acústicos y retro).
- Usar **Clúster 2** para "*Today's Top Hits*" o "*Workout Beats*".

### 3. Marketing:

- Promover **Clúster 2** en horarios pico (mañanas y fines de semana) mediante banners en la app.
- Hay que destacar **Clúster 0** en contextos específicos como eventos culturales o colaboraciones con cafeterías.
- Utilizar **Clúster 1** en campañas de nostalgia (ej. "*Throwback Thursdays*").

Esta segmentación permite ofrecer experiencias personalizadas, maximizando la retención de usuarios y adaptándose a sus contextos de escucha.

# Comparativa entre K-Means (2 Clústeres) y GMM (3 Clústeres) en el Contexto de Spotify

## Diferencias Clave:

ASPECTO	K-MEANS (2 CLÚSTERS)	GMM (3 CLÚSTERS)
NATURALEZA DEL ALGORITMO	Clustering particional: busca grupos no superpuestos y esféricos.	Modelo probabilístico: asume mezcla de distribuciones gaussianas, adaptable a formas complejas.
GRANULARIDAD	Segmentación binaria (moderno vs. clásico).	Segmentación trinaria (moderno, retro, clásico), más detallada.
FLEXIBILIDAD	Limitado a clústeres de tamaños similares y baja tolerancia a superposición.	Maneja clústeres de tamaños/forma variables y solapamiento.
INTERPRETACIÓN	Simple y directa: ideal para estrategias de marketing binarias.	Compleja pero rica: permite identificar subnichos (ej. "retro equilibrado").

## Enfoque de Uso y Aplicaciones:

### K-Means (2 Clústeres):

- **Objetivo:** Segmentación amplia para estrategias de marketing polarizadas.
- **Ejemplos de Uso:**
  - *Playlists binarias:* "Party vs. Relax", "Trending vs. Timeless".
  - *Campañas masivas:* Enfoque en horarios pico (ej. promoción de música energética en fines de semana).

- **Ventajas:**
  - Rapidez computacional y fácil implementación.
  - Ideal para audiencias con preferencias claramente diferenciadas (ej. jóvenes vs. adultos).

### **GMM (3 Clústeres):**

- **Objetivo:** Personalización granular para audiencias con gustos híbridos o transicionales.
- **Ejemplos de Uso:**
  - *Playlists de nicho:* "Retro Vibes", "Acoustic Mornings", "Modern Mix".
  - *Recomendaciones hiper-personalizadas:* Identificar usuarios que oscilan entre lo retro y lo moderno.
- **Ventajas:**
  - Detecta grupos intermedios (ej. "Retro Balanced Mix"), clave para estrategias de retención.
  - Útil para explorar mercados secundarios (ej. millennials nostálgicos).

### **Puntos Positivos de Cada Algoritmo:**

#### **K-Means:**

1. **Eficiencia:** Ideal para grandes conjuntos de datos o recursos limitados.
2. **Claridad:** Resultados fácilmente explicables para equipos no técnicos.
3. **Alineación con Metas Binarias:** Ej. Campañas de lanzamiento de productos (ej. "Premium: Música para tu estado de ánimo").

#### **GMM:**

1. **Precisión en Datos Complejos:** Captura matices como canciones retro-modernas o acústicas con elementos electrónicos.
2. **Adaptabilidad:** Útil para audiencias con gustos híbridos (ej. usuarios que escuchan tanto jazz clásico como pop).

3. **Profundidad Analítica:** Permite descubrir oportunidades ocultas (ej. playlists temáticas como "Clásicos Reinterpretados").

## ¿En que caso usar cada algoritmo?

- **Usar K-Means si:**
  - Se necesitan resultados rápidos para campañas masivas.
  - La audiencia se divide claramente en dos grupos (ej. jóvenes vs. adultos).
  - Ejemplo: "*Verano 2023: Dance vs. Chill*".
- **Usar GMM si:**
  - Se busca maximizar la personalización (ej. "Discover Weekly" con 3 perfiles).
  - El catálogo por analizar incluye música con características híbridas (ej. indie-folk electrónico).
  - Ejemplo: "*Noche de Nostalgia: 70s, 90s y Acústicos Modernos*".

---

## Conclusión:

Mientras **K-Means** simplifica la toma de decisiones con una segmentación binaria clara, **GMM** enriquece la estrategia con una visión tridimensional, ideal para plataformas como Spotify donde los gustos musicales son fluidos y multifacéticos. La elección depende del equilibrio entre velocidad, profundidad y complejidad que requiera la campaña.

## Siguientes Pasos para la Clusterización: Integración Estratégica en el Negocio

La clusterización, ya sea mediante K-Means o GMM, no es solo una herramienta analítica, sino un puente hacia la personalización y la eficiencia operativa. Su implementación estratégica puede transformar la forma en que Spotify conecta con sus usuarios, optimiza su catálogo y maximiza su impacto en el mercado. A continuación, se detallan los siguientes pasos para activar esta solución y cómo puede beneficiar al negocio:

---

### 1. Integración con el Motor de Recomendaciones

- **Objetivo:** Mejorar la precisión de las recomendaciones en secciones como "*Made For You*", "*Discover Weekly*" y "*Daily Mix*".
- **Acciones:**
  - Vincular los clústeres con el algoritmo de recomendación basado en machine learning.
  - Asignar pesos a las características de cada clúster (ej. bailabilidad, energía) para ajustar las sugerencias.
- **Beneficios:**
  - Mayor retención de usuarios al ofrecer playlists que se alinean con sus preferencias emocionales y contextuales.
  - Incremento en la tasa de reproducción de canciones menos populares, pero altamente relevantes para nichos específicos.

---

### 2. Creación de Playlists Temáticas Automatizadas

- **Objetivo:** Generar playlists dinámicas que reflejen los perfiles de los clústeres.
- **Acciones:**

- Desarrollar un sistema que cree automáticamente playlists como "*Modern Energy Mix*" (Cluster 0 de K-Means) o "*Retro Balanced Vibes*" (Clúster 1 de GMM).
  - Incluir canciones de artistas emergentes dentro de los clústeres para fomentar su descubrimiento.
- **Beneficios:**
    - Atracción de nuevos usuarios mediante playlists que resuenan con sus gustos y estados de ánimo.
    - Fomento de la diversidad musical al incluir artistas independientes en listas temáticas.

### 3. Campañas de Marketing Personalizadas

- **Objetivo:** Utilizar los clústeres para segmentar audiencias en campañas publicitarias y promociones.
- **Acciones:**
  - Diseñar campañas específicas para cada clúster (ej. anuncios de música energética para jóvenes en redes sociales).
  - Colaborar con influencers que representen los valores de cada clúster (ej. DJs para el clúster moderno, artistas acústicos para el clásico).
- **Beneficios:**
  - Mayor ROI en campañas al dirigirse a audiencias altamente segmentadas.
  - Fortalecimiento de la marca al asociarse con artistas y creadores que encarnan los valores de cada clúster.

### 4. Optimización del Catálogo Musical

- **Objetivo:** Identificar oportunidades para adquirir o promover música que complementa los clústeres existentes.
- **Acciones:**
  - Analizar brechas en el catálogo (ej. falta de música retro-balanceada en el Clúster 1 de GMM).

- Priorizar la adquisición de licencias para géneros o artistas que fortalezcan los clústeres menos representados.
- **Beneficios:**
  - Mejora en la satisfacción del usuario al ofrecer un catálogo más equilibrado y completo.
  - Atracción de nuevos suscriptores mediante la diversificación de la oferta musical.

## 5. Monitoreo y Ajuste Continuo

- **Objetivo:** Mantener la relevancia de los clústeres a medida que evolucionan los gustos musicales.
- **Acciones:**
  - Implementar un sistema de retroalimentación que actualice los clústeres periódicamente (ej. cada trimestre).
  - Utilizar métricas como "*tiempo de escucha*" y "*tasa de guardado*" para evaluar la efectividad de los clústeres.
- **Beneficios:**
  - Adaptabilidad a tendencias emergentes (ej. nuevos géneros como el hyperpop o el lo-fi).
  - Mejora continua en la precisión de las recomendaciones y la segmentación.

## ¿Cómo Ayuda la Solución al Negocio?

1. **Personalización Profunda:**
  - Los clústeres permiten entender no solo qué música escuchan los usuarios, sino también *cómo* y *por qué* la escuchan. Esto se traduce en experiencias más significativas y memorables.

## **2. Retención y Fidelización:**

- Al ofrecer contenido altamente relevante, Spotify reduce la tasa de abandono y fomenta la lealtad del usuario.

## **3. Optimización de Recursos:**

- La segmentación permite asignar recursos de marketing y adquisición de licencias de manera más eficiente, maximizando el ROI.

## **4. Innovación en Producto:**

- Los clusters inspiran nuevas funcionalidades, como playlists dinámicas o eventos virtuales temáticos, que diferencian a Spotify de la competencia.

---

## **Activación de la Solución**

### **1. Fase Piloto:**

- Implementar los clústeres en una región o segmento de usuarios para medir su impacto antes de escalar.

### **2. Colaboración Interdepartamental:**

- Involucrar a equipos de marketing, desarrollo de producto y análisis de datos para garantizar una integración fluida.

### **3. Comunicación Interna:**

- Capacitar a los equipos en el uso de los clústeres y su interpretación para maximizar su utilidad.

### **4. Evaluación Continua:**

- Establecer KPIs claros (ej. aumento en la tasa de reproducción, reducción en la rotación de usuarios) y ajustar la estrategia según los resultados.

---

En resumen, la clusterización no es solo una herramienta analítica, sino un catalizador para la innovación y el crecimiento sostenible. Al integrarla estratégicamente, Spotify puede consolidarse no solo como una plataforma de streaming, sino como un compañero musical que entiende y celebra la diversidad de sus usuarios.