240

Proyecto final

recolección, análisis y visualización de datos

STEPHEN LUNA RAMÍREZ – 2203066748

EDUARDO GONZALEZ GONZALEZ – 2213064505

JOSÉ ALBERTO POSADAS GUDIÑO - 2213026541

Fundamentos y Preparación de Datos: Una Introducción

Antes de que nuestras palabras sigan su curso y alcancen la espumosa cresta de las olas del mar, presentamos una introducción necesaria para comprender nuestros objetivos y esfuerzos, los cuales otorgarán valor a nuestro trabajo. En esta introducción, abordaremos conceptos y características fundamentales que marcan el inicio de la ***preparación de los datos.*** *Esta fase incluye la limpieza y transformación de los datos obtenidos, los cuales provienen de diversas fuentes, como nuestro entorno API. Posteriormente, exploraremos las técnicas matemáticas que nos han permitido mantener eficiencia y estabilidad en nuestro conjunto de datos. A través de un riguroso proceso de preprocesamiento, construiremos el modelo correspondiente para obtener un análisis preciso de esta colección de datos.*

La introducción inicial se ajusta a las entregas previas del proyecto (fases uno y dos). Por lo tanto, se reducirá cualquier información redundante para lograr claridad y brevedad. No obstante, se mantendrá el enfoque en los puntos críticos del documento asociado, especialmente en lo que respecta a la evaluación.

Con distinguida consideración, los integrantes del equipo del proyecto se despiden atentamente.

México, 20 de enero de 2025.

Contenido

Capítulo I

Introducción

Una introducción concisa y técnica a la técnica empleada para extraer datos de contenidos dinámicos, específicamente en el contexto de la aplicación Spotify, junto con sus características relevantes para su posterior análisis.

1.1. Análisis y Relevancia del Tema seleccionado

Para nuestro proyecto, hemos seleccionado el tema de **Spotify** debido a su relevancia en la industria musical y su papel en el **análisis de datos**. Como una de las principales plataformas de *streaming* a nivel mundial, Spotify proporciona información valiosa sobre tendencias, preferencias de usuarios y patrones de consumo. Utilizaremos la **API de Spotify Developer** para un análisis exhaustivo de datos musicales, explorando desde detalles técnicos de canciones hasta factores de éxito comercial. Esta API nos permite acceder a datos en tiempo real (datos totalmente dinámicos), lo que nos brinda una visión dinámica del consumo musical global.

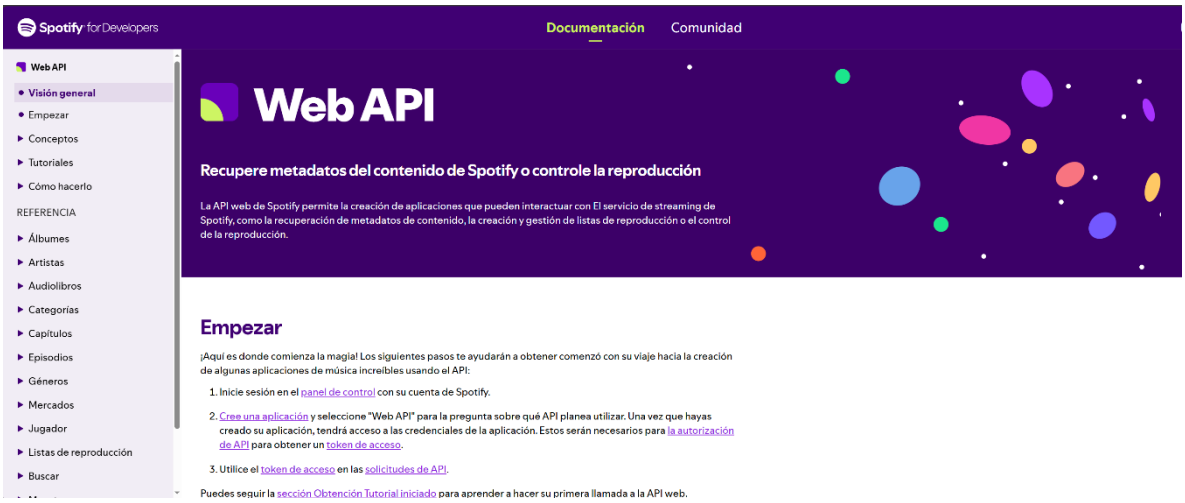


Figura 1.1: *Visualización del sitio original y disponible de la API de Spotify según se documenta.*

La **música**, a lo largo de la historia, ha experimentado una transformación significativa. Aunque su origen sigue siendo misterioso, sabemos que va más allá de

la simple combinación de sonidos. Además, de los elementos básicos como el *tono* y el *ritmo*, la música se ha enriquecido al incorporar **poesía** y **danza**, evocando emociones como *alegría* y *tristeza*. En sus inicios, estaba estrechamente ligada al canto y la danza, pero con el tiempo se emancipó para convertirse en el arte que conocemos hoy. A pesar de su diversidad, la música y sus componentes están unidos por leyes naturales. Las notas musicales, en esencia, se traducen en números según sus vibraciones, otorgándole un carácter científico. Los matemáticos griegos estudiaron minuciosamente las escalas y descubrieron que estas leyes aplicables generaban fenómenos capaces de evocar emociones profundas en nosotros.

Con la evolución del **Software**, acceder a melodías se ha vuelto más sencillo a través de aplicaciones y sitios web que permiten a los usuarios crear sus propias bibliotecas musicales basadas en datos personales. Esto ahorra tiempo y, en ocasiones, dinero al evitar visitas a tiendas de disco y decisiones de compra. En la era moderna, la música es esencial a nivel global, y la amplia base de usuarios genera datos que nos ayudan a anticipar tendencias y medir la popularidad de artistas y grupos.

Por otro lado, las compañías discográficas, conscientes de la importancia de los datos, utilizan análisis estratégicos para tomar decisiones informadas. Estudian tendencias y géneros musicales emergentes para establecer una producción adecuada y lograr crecimiento económico. En el contexto de aplicaciones como Spotify, la experiencia del usuario se ha optimizado gracias a un equipo multidisciplinario de expertos en programación y diseño. Además de disfrutar de música, los usuarios pueden explorar datos interesantes y aplicar técnicas avanzadas.

El procesamiento de señales y el aprendizaje automático permiten analizar composiciones musicales en profundidad. Identificar patrones en ritmo, armonía y melodía nos ayuda a desentrañar estructuras subyacentes y tendencias en diferentes géneros. Por ejemplo, podemos descubrir qué elementos hacen que una canción sea

pegajosa o por qué ciertos géneros son más populares en ciertas regiones. Las **letras de las canciones** también son valiosas. Al evaluar las emociones expresadas en ellas, comprendemos mejor la conexión emocional que los oyentes tienen con la música. Esto no sólo enriquece nuestra experiencia, sino también proporciona datos útiles para mejorar la calidad del servicio.

1.2. Razonamiento y Selección del Conjunto de Datos Utilizado

En el ámbito del entretenimiento, los datos relacionados con la música son fundamentales. Los sistemas informáticos pueden inferir con precisión las preferencias de los usuarios incluso a partir de referencias vagas, como fragmentos de letras o sonidos. Estos datos no se limitan a una sola canción; más bien, se recopilan detalles detallados sobre artistas, géneros y otros atributos relevantes. Al analizar estos datos, podemos identificar tendencias en diferentes grupos demográficos y obtener *insights* valiosos para la industria musical.

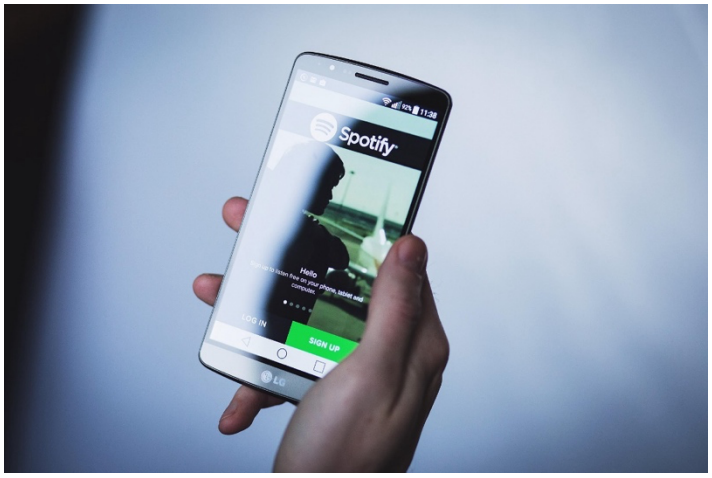
Por tanto, el análisis de datos relacionados con la música nos permite descubrir patrones clave en la industria del entretenimiento. Al examinar los metadatos incrustados en archivos de audio, como **título**, **género**, **energía**, **duración**, **sonoridad**, posibilidad de ser **bailable** o no, entre otros, podemos identificar artistas populares, géneros buscados y canciones reproducidas con mayor frecuencia. Estos datos son esenciales para funciones como recomendaciones personalizadas en plataformas de *streaming*, la creación de listas de reproducción y la distribución justa de regalías a los creadores musicales.

Observación: Si surge interés en comprender su valor o su potencial impacto en el mundo moderno, se recomienda consultar la **primera entrega** del mismo proyecto.

En las siguientes secciones, se detallan las características relevantes de las columnas correspondientes en los datos musicales (*consulte la **Figura 3.3, Capítulo III***).

1. **Nombre:** Titulo de la canción.
2. **Bailabilidad:** Una métrica que indica qué tan apta es una canción para bailar, basada en características como tempo, estabilidad del ritmo y regularidad. Los valores van de **0** (menos bailable) a **1** (más bailable).
3. **Energía**: Representa la intensidad y actividad percibida en la canción. Las canciones con mayor energía tienden a sentirse rápidas y ruidosas. Valores de **0** a **1**.
4. **Acústica**: Mide que tan acústica es una canción, **1** indica una alta presencia de elementos acústicos y **0** indica la producción completamente digital o electrónica.
5. **Duración (Ms)**: Duración total de la canción en milisegundos.

6. **Instrumentalidad**: Indica la probabilidad de que la canción sea instrumental (sin voces). Valores cercanos a 1 indican mayor probabilidad de que no haya voces.
7. **Sonoridad:** El nivel general de volumen percibido en la canción, medido en decibeles (dB). Valores negativos más bajos indican canciones más suaves, mientras que valores más altos indican canciones más fuertes.
8. **Tempo:** El ritmo o velocidad de la canción, medido en pulsaciones por minuto (BPM).
9. **Discursividad:** Mide qué tan hablada es la canción. Valores altos indican mayor presencia de palabras habladas, como en un podcast o rap.
10. **ID:** Identificador único de la canción generado por Spotify.
11. **Modo**: Representa si la canción está en modo mayor o menor:
    - 1: *Modo mayor* (asociado con canciones alegres).
    - 0: *Modo menor* (asociado con canciones más melancólicas).
12. **Valencia:** Mide el grado de positividad o felicidad percibida en la canción. Valores cercanos a **1** indican canciones felices, y valores cercanos a **0** indican canciones más tristes o sombrías.
13. **Género:** La categoría musical o estilo al que pertenece la canción.



Capítulo II

Metodología

El procedimiento esencial para el análisis exploratorio de datos consiste en aplicar métodos y herramientas que permitan comprender la naturaleza y distribución de los datos antes de construir la estructura del modelo seleccionado de manera apropiada.

2.1. Descripción del Proceso de Recolección de Datos

Utilizando la API oficial de Spotify (*véase* ***Figura 1.1, Capítulo I***), hemos recopilado datos (guardados a través de un archivo XLSX) de alta calidad que nos permitirán analizar tendencias en la industria musical. El conjunto de datos original (*ver* ***Figura 3.2, Capítulo III***), con aproximadamente 500 registros, servirá como base para nuestro análisis. Además de cuantificar los datos sobre las canciones y artistas más populares, también podemos explorar aspectos cualitativos relacionados con las *emociones* de los usuarios. Las múltiples variables en las columnas nos brindan información esencial para identificar patrones, como la tonalidad de las melodías o la calidad rítmica de los géneros musicales. En suma, la **calidad** de estos datos se sustenta en la intersección entre sensaciones emocionales y relaciones aritméticas.

La **cantidad** de datos con la que contamos es un aspecto crucial para cualquier análisis. En nuestro caso, disponemos de entre 500 y 530 registros recolectados a través de la API oficial de Spotify (visite, para mayor comprensión, el siguiente sitio: https://developer.spotify.com/). Sin embargo, al comparar esto con conjuntos de datos del mundo real, que a menudo contienen miles o incluso millones de registros, podemos notar una diferencia significativa en términos de profundidad y representatividad.

En un escenario ideal, podríamos aumentar la cantidad de datos a un rango de 1000 a 2500 registros. Esto nos permitirá obtener una visión más completa y robusta de las tendencias musicales y los patrones de consumo; pero aquí entra en juego una limitación importante: la API de Spotify impone restricciones en el número de consultas que podemos realizar, incluso en sus planes de pago. Esto significa que debemos trabajar dentro de esos límites y encontrar soluciones creativas para maximizar la utilidad de los datos disponibles. A pesar de esta limitación, la cantidad actual de registros es suficiente para comenzar con la **transformación** y el **análisis de datos**. Es importante estar conscientes de estas restricciones desde el

principio y adaptarnos en consecuencia. A lo largo del proyecto, enfrentaremos este desafío y buscaremos optimizar nuestros enfoques para obtener resultados valiosos dentro de las limitaciones impuestas por la API.

2.2. Procedimientos de Limpieza y Transformación de Datos

En este momento, deseo dirigir mi atención hacia el águila que domina los rápidos vientos de los cuatro puntos cardinales. Entre sus garras, prevalece sobre el voraz enemigo. A continuación, presento las siguientes novedades:

2.2.1. Limpieza

La obtención de esta colección de datos proviene de un archivo en formato XLSX (Excel), cuya información se extrae directamente de la API de Spotify. Es imperativo preparar estos datos de manera adecuada para su posterior transformación y estructuración en el análisis correspondiente. Por lo tanto, iniciamos el **proceso de limpieza de datos**: Al observar el conjunto de datos representado en la ***Figura 3.3*** del ***capítulo III***, se evidencia que cada elemento se asocia con un valor numérico. Esto sugiere que estamos tratando con datos **cuantitativos**, es decir, valores que pueden medirse, evaluarse o clasificarse. En consecuencia, la preparación y limpieza de estos datos se llevará a cabo mediante **medidas de localización**, como la *media*, la *mediana*, entre otras. No obstante, esta situación presenta tanto ventajas como desventajas para nuestro análisis.

En primer lugar, debido a que la colección de datos está completa y prácticamente libre de errores, no se han identificado valores nulos o vacíos. Por lo tanto, no es necesario aplicar medidas estadísticas como la **media** (que se utiliza cuando no hay *valores extremos*) o la **mediana** (útil cuando la *distribución es sesgada* o presenta valores extremos). Estas medidas suelen emplearse para reemplazar o llenar **valores faltantes**, un problema común en diversos conjuntos de datos. No obstante, para evitar cualquier duda, se realizó una exploración visual utilizando herramientas como los **Mapas de Calor** (*ver* ***Figura 3.1, capítulo III***). El objetivo era verificar rápidamente si existían posibles valores faltantes en nuestro conjunto de datos, ya sea en las filas o columnas extraídas previamente de la aplicación de Spotify. Afortunadamente, se confirmó que no había ningún valor faltante.

En segundo lugar, debido a las circunstancias particulares, se presentaron obstáculos que dificultaron el inicio del proceso de limpieza de datos. Afortunadamente, no se identificaron otros problemas, como valores fuera de rango o cambios en el formato, que requirieran la aplicación de métodos o medidas estadísticas específicas; sin embargo, al examinar detalladamente cada fila (aunque no todas), se observó que algunas canciones se repetían hasta cuatro ocasiones. Esta redundancia o **duplicación** innecesaria de datos (en este caso, la lista de canciones) constituía un error inicial en nuestro proceso de recolección de datos. La presencia

de duplicados podría afectar significativamente nuestro análisis por dos razones fundamentales:

i) **Impacto en la calidad de los resultados:** Si los resultados no son precisos debido a la duplicación, podríamos generar informes incorrectos o tomar decisiones basadas en datos distorsionados.

ii) **Desorganización de la información relevante:** La falta de un orden apropiado podría llevar a conclusiones erróneas o a una interpretación ambigua de los datos.

Además de los aspectos mencionados previamente, la presencia de duplicados en nuestro conjunto de datos puede afectar positivamente la **eficiencia operativa**, pero también puede generar confusión al aplicar modelos específicos, como el SVM (***Support Vector Machine***, por sus siglas en inglés). Por lo tanto, es imperativo eliminar esta duplicación para garantizar la integridad de nuestros datos. Para abordar este problema, se implementó un *script* en **Python** diseñado específicamente para procesar el archivo Excel original. El proceso fue el siguiente:

1. **Identificación de canciones por género:** El *script* analizó los datos y utilizó los identificadores únicos (ID) de las canciones para agruparlas por género. Esto permitió una clasificación más precisa y facilitó la detección de duplicados.

2. **Búsqueda y eliminación de duplicados:** A continuación, el *script* iteró sobre los registros y detectó las entradas duplicadas. Si se encontraban dos o más registros con el mismo ID de canción, se eliminaba la duplicación. Este paso aseguró que obtuviéramos una versión depurada del conjunto de datos, completamente libre de redundancias.

El resultado final fue un conjunto de datos limpio y listo para su posterior análisis (*véase* ***Figura 3.2*** *y* ***Figura 3.3, capítulo III***). Este proceso de **de duplicación** es fundamental para garantizar resultados confiables en cualquier análisis o modelo que se aplique posteriormente.

2.2.2. Transformación

Al comenzar un análisis de datos, es crucial aplicar los primeros pasos de limpieza al conjunto de datos inicial. Cualquier error o conflicto, por pequeño que sea, podría tener consecuencias significativas, no sólo durante la transformación de los datos (como se detallará más adelante), sino también en la construcción inicial del modelo. Por lo tanto, una vez que los datos están libres de ambigüedades arbitrarias, es fundamental aplicar técnicas matemáticas como la **normalización** o la **estandarización**. Estas técnicas, aunque diferentes, permiten que las variables

sean comparables y, al mismo tiempo, maximizan el rendimiento del modelo que se aplicará en el análisis final. Además, al reducir el impacto de valores extremadamente altos o bajos, evitamos descontrol y cálculos sin sentido en el conjunto de datos. Con el objetivo de facilitar una comprensión más profunda, procederé a exponer de manera concisa ambos métodos:

i) **Normalización:** Consiste en escalar los valores de las variables para que estén dentro de un rango específico, generalmente entre 0.0 y 1.0. Esto es especialmente útil cuando las variables tienen diferentes unidades o escalas. La normalización garantiza que todas las características contribuyan de manera **equitativa** al modelo.

ii) **Estandarización:** En este caso, transformamos las variables para que tengan una media de 0 y una desviación de 1. La estandarización es útil cuando queremos eliminar el efecto de las diferencias de escala entre las variables. Al estandarizar, hacemos que todas las variables tengan la misma *escala relativa*.

En una palabra, la **transformación de datos** es esencial para escalar diferentes rangos, facilitar la visualización y el análisis, y convertir valores abstractos en valores comparables. Tanto la normalización como la estandarización nos permiten *ajustar* y *organizar* adecuadamente nuestro conjunto de datos, centralizando todos los valores. Este paso es crucial en el **preprocesamiento** de datos de entrenamiento para modelos, ya que no basta con eliminar duplicados (*véase* ***sección 2.2.1***); también debemos establecer datos eficientes y controlados.

En el contexto de la transformación de datos, se plantea la necesidad de normalizar o estandarizar los valores; sin embargo, antes de aplicar cualquiera de estas técnicas, es crucial determinar cuál de ellas es más adecuada para nuestro conjunto de datos. Aunque ambas opciones pueden escalar valores numéricos, cada una tiene sus propias características y objetivos específicos, a los cuales fueron explicados anteriormente.

Existen dos enfoques principales que ofrecen cierta certidumbre en este proceso. El primero, ampliamente utilizado, se basa en **métricas descriptivas**, específicamente *medidas de localización* y *variabilidad*. Estas métricas nos permiten comprender la distribución de un conjunto de datos y evaluar si existen niveles bajos o altos de dispersión. Las dos métricas clave en este contexto son el **rango** y la **desviación estándar**. El rango (***ecuación 1)*** se calcula como la diferencia entre el *valor máximo* y el *valor mínimo* de una serie de observaciones o características. Por otro lado, la desviación estándar (***ecuación 2***) se obtiene mediante la raíz cuadrada de la suma de los cuadrados de las desviaciones de los datos con respecto a su *media*, dividida por el número total de observaciones. Una mayor desviación estándar indica una mayor dispersión de los datos, mientras que una desviación baja sugiere que los datos se agrupan alrededor de la media.

2.2.2.1. ***Expresiones Matemáticas***

$$R = (Max_x) - (Min_x) \quad [\textbf{\textit{ecuación 1}}].$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{x})^2}{N}} \qquad [\textbf{\textit{ecuación 2}}].$$

Para tomar decisiones informadas, es fundamental establecer umbrales que nos permitan determinar si los valores calculados superan o no ciertos límites. Esto nos proporciona la certeza necesaria para elegir entre *normalizar* o *estandarizar* nuestros datos, en función de los objetivos específicos de nuestro análisis (*consulte la* ***Figura 3.4, capítulo III***).

Observación: En el contexto de este documento, los cálculos se realizan exclusivamente a través de programas informáticos (acceder directamente al directorio de código, incorporado en el mismo archivo con formato .zip). Aunque es posible resolverlos manualmente, la eficiencia y el ahorro de tiempo inherentes a la automatización justifican su implementación en un entorno de programación. En nuestro caso, utilizaremos el lenguaje de programación **Python**. Por lo tanto, se prescindirá de capturas de pantalla o ilustraciones de código para evitar redundancias o confusiones. Únicamente se presentarán los resultados relevantes (*véase* ***Capítulo III***), que se citarán y analizarán en este segundo capítulo (Metodología).

La segunda y última opción para determinar qué técnica matemática es recomendable o favorable consiste en utilizar la **visualización de datos**. Mediante herramientas de observación sencilla y práctica, podemos identificar qué datos se encuentran fuera del rango definido, es decir, aquellos que presentan desequilibrio o valores sesgados dentro del conjunto. Esta aproximación nos permite evaluar la sensibilidad del dominio ante **valores atípicos** y determinar en qué casos nuestros datos podrían enfrentar circunstancias desfavorables o problemas futuros. En última instancia, esta comprensión más profunda nos ayuda a alcanzar nuestros objetivos de análisis de manera más precisa. En otras palabras, la visualización de datos nos permite evaluar la distancia entre nuestro conjunto de datos y su comportamiento ideal.

En función de las medidas de localización y variabilidad, como la media, la varianza y los cuartiles, se han empleado dos tipos de visualización o gráficas: **el Diagrama de Cajas y Bigotes** (***BoxPlot***, en inglés) y los **Histogramas**. Ambas gráficas se encuentran en el capítulo de resultados (*ver* ***Figura 3.5*** *y* ***Figura 3.6****,* ***capítulo III***). Estas representaciones gráficas nos permiten comprender cómo los datos se distribuyen en función de sus valores; sin embargo, para evitar confusiones o

malentendidos, es relevante destacar sus diferencias y características correspondientes:

A. **Diagrama de Cajas y Bigotes:** Conocido como **boxplot**, es una representación gráfica que permite visualizar la distribución de un conjunto de datos estadísticos. Se basa en los cuartiles y muestra información relevante sobre la dispersión de los datos. En un *boxplot*, se representan los siguientes elementos:

   - **Caja (*Box*):** La caja abarca desde el primer cuartil (Q_1) hasta el tercer cuartil (Q_3). La mediana (valor o segmento central) se encuentra en el interior de la caja. La longitud (entre Q_1y Q_3) de la caja indica la variabilidad o valor intercuartílica.
   - **Bigotes (*Whiskers*):** Los bigotes se extienden desde la caja hacia los valores extremos (dado por límites admisibles, uno inferior y otro superior). Indican, asimismo, la presencia de valores atípicos.
   - **Valores Atípicos:** Representados como puntos individuales, a los cuales indican que están fuera dentro de la caja o bigotes.

   El *boxplot* es útil para analizar la simetría de la muestra estadística. Si la mediana no está en el centro de la caja, sugiere que la distribución no es simétrica.

B. **Histogramas:** Gráfico estadístico que representa la distribución de frecuencias de un conjunto de datos. En él, se utilizan barras rectangulares para visualizar la frecuencia con la que aparecen los valores en diferentes intervalos. Algunas características del histograma son:

   - **Barras:** Cada barra representa un intervalo de valores. La altura de la barra es proporcional a la frecuencia de los datos en ese intervalo.
   - **Anchura de las Barras:** La anchura de cada barra refleja la amplitud del intervalo.
   - **Forma de la Distribución:** La forma del histograma puede indicar si la distribución es simétrica, sesgada o multimodal.

En suma, ambas gráficas son herramientas valiosas para *explorar* y *comprender* la distribución de datos en un conjunto estadístico. Cada uno ofrece una perspectiva diferente y complementaria.

En la ***Figura 3.6*** del ***capítulo III***, se observa una característica clave de nuestro conjunto de datos. Esta observación nos permite tomar decisiones precisas sobre cómo escalar los datos mediante técnicas de transformación matemática. En particular, notamos que la mayoría de los datos requieren **estandarización**. Por lo

tanto, como primer paso en el preprocesamiento, sería recomendable aplicar el **Análisis de Componentes Principales** (**PCA**, por sus siglas en inglés).

2.3. Propósitos Generales

Mediante la limpieza y transformación de los datos, hemos logrado un avance significativo; sin embargo, aún debemos determinar cómo convertir esta colección de datos (una lista de canciones) en un formato más eficaz mediante técnicas más avanzadas. Antes de proceder, es crucial establecer nuestros objetivos. Una vez que hemos escalado adecuadamente nuestros datos aplicando la técnica matemática de estandarización, podemos identificar las columnas o datos más relevantes para nuestro análisis. Esto nos permitirá seleccionar apropiadamente y avanzar hacia la etapa final: el modelado.

Dada las diversas columnas o datos extraídos de la API de Spotify (para una comprensión más detallada, consulta el ***Capítulo I*** y la ***sección 2.1*** del mismo capítulo), presentamos de manera general algunos posibles objetivos:

i. Utilizando la lista de canciones arbitrarias, podemos aplicar técnicas de agrupamiento (como el **Clustering**) basadas en características musicales. Esto nos permitirá descubrir patrones y tendencias en diferentes géneros musicales e incluso identificar subgéneros emergentes. Por ejemplo, podríamos agrupar canciones con alta energía y tempo rápido en un grupo, mientras que las canciones acústicas y lentas podrían formar otro grupo.

ii. Al clasificar cada canción de la lista según sus características (como energía, bailabilidad y valencia), podemos predecir su popularidad. ¿Qué características están más relacionadas con una canción exitosa en términos de reproducciones o inclusiones en listas de reproducción?

iii. Aunque los datos son numéricos (cuantitativos), también pueden tener implicaciones cualitativas. Por ejemplo, podemos examinar cómo la valencia y la energía se correlacionan con la percepción emocional de las canciones. ¿Las canciones más felices tienden a tener una mayor valencia (**relación positiva**, representada como 1)? ¿O las canciones más tristes son menos bailables (**relación negativa**, con un valor representado como -1)?

iv. Si deseamos profundizar aún más, podemos identificar qué canciones se desvían significativamente de las tendencias generales. ¿Existen canciones extremadamente largas o cortas? ¿Hay algún género musical con características inusuales?

2.4. Método de Preprocesamiento

La embarcación ha alcanzado una distancia considerable, al punto de que, al colisionar con las olas cuya espuma blanca se asemeja al verso entonado por las voces de las Musas, ha iluminado la tarde sombría con un sonido semejante al rayo dorado del Sol.

2.4.1. Estandarización

Una vez que hemos establecido las ideas iniciales y el primer esbozo, es fundamental determinar la técnica más adecuada para equilibrar nuestro conjunto de datos y eliminar distorsiones, como la influencia de **valores atípicos**. Ahora debemos perfeccionarlo para llevarlo al nivel de un modelo general para nuestro análisis; sin embargo, antes de utilizarlo de manera efectiva, debemos **transformarlo**, similar a pulir el oro crudo, convirtiéndolo en datos sólidos y estructurados.

Dado que la mayoría de nuestros datos se han sometido a la técnica de estandarización (*consultar* ***Figura 3.4***), es imperativo seguir las mejores prácticas para garantizar el éxito de nuestra empresa. Esto implica considerar técnicas más avanzadas que nos permitan estructurar los datos de manera más eficiente antes de construir o aplicar el modelo final de análisis.

Una vez que hemos identificado la técnica de transformación adecuada, procedemos a aplicarla para **estandarizar** todo el conjunto de datos. Esto implica crear un segundo archivo de datos, similar al formato de Excel, que contendrá únicamente los datos limpios y centralizados. En otras palabras, mediante la técnica de estandarización, mitigamos el impacto de valores atípicos (aquellos que se encuentran fuera del rango definido) y logramos una distribución completamente normal. En esta distribución, cada dato tiene una media de cero y una desviación estándar de uno, lo que convierte la estructura de datos en una dimensión simétrica.

Como resultado, obtenemos un nuevo conjunto de datos con valores ordenados de manera uniforme, donde todas las características tiene igual peso e importancia en su formación. Este proceso se ilustra claramente en la ***Figura 3.7*** del próximo capítulo (***Capítulo III, Resultados***), donde, gracias a los pasos anteriores, aplicamos la técnica apropiada para obtener datos sólidos y libres de ambigüedad.

2.4.2. Optimización del uso del PCA (Análisis de Componentes Principales)

Al igual que un valiente guerrero que estudia a sus adversarios antes de lanzarse al combate, nosotros hemos preparado meticulosamente nuestro análisis. Utilizando diversas técnicas y herramientas matemáticas, hemos obtenido datos limpios y evitado errores, como la **multicolinealidad**. Nuestro objetivo es construir un modelo eficiente y audaz que pueda procesar correctamente los datos detallados en nuestra ***sección 1.2*** (*ver* ***Capítulo I***). Además, buscamos determinar aspectos como el interés de un usuario en una canción específica o la calidad del sonido según su *energía* y *valencia*. Para lograrlo, evaluaremos una técnica muy diferente a las

anteriores, considerando tanto los resultados previos como alternativas más amplias. Este paso es crucial para diseñar y construir un modelo efectivo.

La técnica que abordaremos a continuación es el **Análisis de Componentes Principales** (**PCA**, por sus siglas en inglés), un método fundamental en el preprocesamiento de datos. PCA se utiliza para reducir la dimensionalidad de un conjunto de datos, lo que nos permite obtener una representación más concisa y significativa de la información contenida en las variables originales. En primer lugar, es importante destacar que PCA se basa en técnicas previas (vistas desde la ***sección 2.2.2***), como la estandarización (véase la ***sección 2.4.1***) de los datos. La estandarización es crucial, ya que afecta directamente la eficacia de PCA. Una vez que hemos estandarizado nuestros datos (es decir, los hemos escalado para que tengan media cero y varianza unitaria), estamos listos para aplicar PCA.

¿Cómo funciona PCA? En esencia, PCA reorganiza nuestros datos en nuevas variables llamadas **componentes principales**. Estos componentes son combinaciones lineales de las variables originales y están diseñados para capturar la mayor variabilidad posible en los datos. En otras palabras, representan las direcciones en las que los datos varían más. La **ventaja** clave de PCA es que nos permite reducir la cantidad de información redundante o innecesaria en nuestro conjunto de datos. Al seleccionar sólo los componentes principales más relevantes, podemos simplificar la representación sin perder la estructura esencial de los datos. Esto es especialmente útil cuando trabajamos con conjuntos de datos de alta dimensionalidad.

Para fundamentar inicialmente esta afirmación, y considerando si la estrella aún brilla en nuestro pecho, permíteme explicar cómo se determina la eficiencia representativa en nuestra colección de datos y mediante qué medios:

1. **Centralización de Datos:**
    - Calculamos la media de cada variable en nuestro conjunto de datos.
    - Luego, restamos cada valor individual de la media correspondiente. Esto nos da los datos centralizados.
2. **Matriz de Covarianza:**
    - Se determina una matriz de covarianza numérica.
    - Esta misma nos muestra cómo las variables se combinan entre sí. En otras palabras, revela cómo se comportan en conjunto.
    - No obstante, la covarianza mide la relación lineal entre dos variables. Un **valor positivo** indica que aumentan juntas, mientras que un **valor negativo** sugiere que una disminuye cuando la otra aumenta.

3. **Autovalores y Autovectores:**
    - Se determina, desde luego, los componentes principales.
    - Los **autovalores** representan la importancia o peso de una dirección en nuestros datos.
    - Los **autovectores** indican la dirección misma. En conjunto, nos dice cuánta variabilidad hay en esa dirección.
    - Un autovalor grande significa que esa dirección captura mucha información relevante.
4. **Varianza Explicada:**
    - Como final, se evalúa la varianza explicada.
    - Este valor, expresado como un porcentaje, nos dice cuánta información captura nuestro componente principal.
    - Si es suficientemente alto, podemos estar seguros de que estamos reteniendo la mayor cantidad posible de información relevante.

En relación a estos pasos o cálculos, no necesariamente se han llevado a cabo explícitamente en el código. Esto se debe a que el lenguaje de programación utilizado (en este caso, **Python**, que es de tipo dinámico) ya los incorpora internamente. De hecho, en una única instrucción (mediante un método proporcionado por la biblioteca PCA), se obtiene como resultado final el componente o el número de componentes principales creados.

Dado un conjunto de componentes principales, cada uno de ellos contiene su propia información con **alta variabilidad**. Estos componentes indican, a través de la **varianza explicada**, cuánta información o datos han sido capturados y considerados de mayor importancia. En otras palabras, mediante la varianza explicada, se seleccionan aquellos datos (columnas) que tienen más peso o valor, expresado como un porcentaje entre el 0.0% y el 100%. Este porcentaje determina si es suficiente para capturar la mayor cantidad posible de información dentro de nuestro conjunto de datos. Basándonos en este porcentaje, construimos un número específico de componente principales.

En nuestro caso, hemos utilizado tres de estos componentes principales, los cuales representan entre el 50% y el 60% de la información total. Estos componentes se utilizarán en nuestro análisis para ejecutar el experimento de manera más precisa y formal, de acuerdo al modelo seleccionado. Posteriormente, evaluaremos si se han cumplido los objetivos planteados previamente. Los resultados detallados se encuentran en la ***Figura 3.8*** del ***capítulo III***.

2.5. Análisis de Datos (Selección del Modelo)

Como conclusión final y resultado de un proceso meticuloso, alejado de las limitaciones impuestas por los objetivos iniciales, hemos alcanzado una resolución al problema real. Para ello, aplicamos las excelentes funcionalidades proporcionadas por nuestro modelo. En la sección 2.3, previamente describimos los objetivos, y ahora es pertinente enfatizar que cada uno de ellos era factible de ejecución. En una palabra, nuestro análisis inicial, basado en datos sólidos y bien estructurados, se presta a aplicarse a dos algoritmos distintos: el SVM (Máquina de Vectores de Soporte) y el Clustering . Ambos son herramientas idóneas para el análisis de datos, pero, dada nuestra intención principal, hemos seleccionado la segunda opción. El Clustering, perteneciente al campo del Aprendizaje no Supervisado (donde los datos no necesariamente están etiquetados), se materializa mediante el algoritmo de K-means .

Una vez que hemos identificado el algoritmo central de nuestro análisis de datos, es fundamental justificar su elección mediante una explicación detallada de sus funciones y características clave. En este contexto, profundizaremos en las siguientes áreas:

1. El objetivo principal del Clustering es agrupar datos similares en clústeres o grupos. Cada clúster debe contener datos con características afines, mientras que los datos de diferentes clústeres deben ser distintos entre sí.

2. Existen varios algoritmos de Clustering, y cada uno tiene su enfoque particular para formar clústeres. Algunos ejemplos, además de K-means, son el DBSCAN , el Agglomerative Hierarchical Clustering y el Mean Shift . En nuestro caso, hemos optado por el algoritmo K-means debido a su simplicidad y eficiencia. Veamos, a detalle, sobre éste mismo:

   K-means es un algoritmo de particionamiento que busca dividir un conjunto de datos en K clústeres, donde K es un valor predefinido. Funciona de la siguiente manera:

   I. Inicialización: Se selecciona K centroides iniciales (puntos representativos) de manera aleatoria o mediante algún método específico.
   II. Asignación Iterativa: Para cada punto en el conjunto de datos, se calcula la distancia a cada centroide. El punto se asigna al clúster cuyo centroide está más cerca.
   III. Actualización de Centroides: Una vez que todos los puntos han sido asignados, se recalculan los centroides de cada clúster como el promedio de los puntos asignados a ese clúster.

Los pasos de asignación y actualización se repiten hasta que los centroides convergen (es decir, ya no cambian significativamente).

El objetivo final es minimizar la suma de las distancias cuadradas entre los puntos y sus centroides asignados.

Para concluir adecuadamente este segundo capítulo, presentaremos los resultados generales obtenidos por nuestra empresa. Estos resultados se encuentran representados en la ***figura 3.9*** del ***capítulo III***.

Como se puede observar, a través del modelo o algoritmo de Clustering, hemos procesado y analizado nuestro conjunto de datos de la siguiente manera:

La ***figura 3.9*** muestra los resultados del clustering K-Means aplicado a datos de Spotify después de una reducción de dimensionalidad con PCA. Cada punto representa una canción, su color indica el clúster al que pertenece y su posición está determinada por los **3 componentes principales** (ejes X, Y, Z). La interpretación se basa en observar la separación de los clústeres, la ubicación de cada clúster en relación a las características originales (usando las cargas de PCA), identificar valores atípicos y analizar el tamaño de cada clúster. Combinando estas observaciones con otros análisis, se puede obtener una comprensión profunda de los patrones y relaciones entre las canciones de Spotify en función de sus características de audio.

La ***figura 3.9*** también muestra la cantidad de canciones que pertenecen a cada cluster identificado por el algoritmo K-Means. Cada barra representa un cluster diferente, su altura indica la cantidad de canciones dentro de ese cluster y el color ayuda a distinguir visualmente entre los clusters. Esta visualización permite comprender la distribución de las canciones en los diferentes grupos, identificando rápidamente **clusters populares** (con más canciones) y **clusters menos populares**, y ofreciendo una visión general de cómo se agrupan las canciones según sus características.

La***figura 3.10*** del ***capítulo III*** un heatmap, muestra las características promedio de las canciones dentro de cada cluster generado por el algoritmo K-Means. Cada fila representa un cluster y cada columna una característica musical como bailabilidad o energía. Los colores indican el valor promedio de cada característica en cada cluster: colores intensos (rojos o azules) indican valores altos o bajos respectivamente, mientras que colores claros indican valores promedio. De esta manera, el heatmap permite visualizar rápidamente las características que definen a cada grupo de canciones y entender las diferencias entre los clusters, por ejemplo, identificando clusters de canciones bailables y energéticas o clusters de canciones acústicas y tranquilas.

La ***figura 3.10*** también contiene otro "heatmap" que muestra la distribución de **géneros musicales** dentro de cada cluster o grupo de canciones. Cada fila del mapa representa un cluster diferente, mientras que cada columna representa un género musical. La intensidad del color azul en cada celda indica la cantidad de canciones de ese género que pertenecen a ese cluster. Un azul más oscuro significa que hay más canciones de ese género en ese cluster. De esta manera, el gráfico permite visualizar rápidamente qué géneros son más comunes en cada cluster, y por lo tanto, entender mejor las características musicales que definen a cada grupo de canciones. Por ejemplo, se podría observar que un cluster (0) tiene una alta concentración de canciones de k-pop y reggae, mientras que black-metal se concentra en dos clusters (3 y 9).

CAPÍTULO III

RESULTADOS

Las ilustraciones desempeñan un papel crucial en la comprensión de los desafíos inherentes a la recolección de datos y al análisis exploratorio. Asimismo, aplicación a técnicas avanzadas en la construcción del modelado.

En este tercer capítulo, continuando con la metodología presentada en el capítulo anterior, cada figura incorporada en este documento recibirá su propia etiqueta. Estas figuras no se presentan meramente como ilustraciones para el entretenimiento del lector; más bien, se les asignarán números correspondientes en un orden apropiado, y se citarán en relación con la explicación proporcionada en el segundo capítulo. Esta estrategia tiene un propósito específico: facilitar la visualización y comprensión para el lector. Una vez que se ha descrito una solución o proceso, se hace referencia a la figura relevante como ejemplo y justificación. Esto enriquece la comprensión de los procedimientos y pasos utilizados para resolver el problema en cuestión. Además, es fundamental aplicar el modelo apropiado según nuestros objetivos, basándonos en la información recopilada.

En una palabra, los capítulos segundo y tercero están estrechamente relacionados y son esenciales para alcanzar conclusiones significativas.

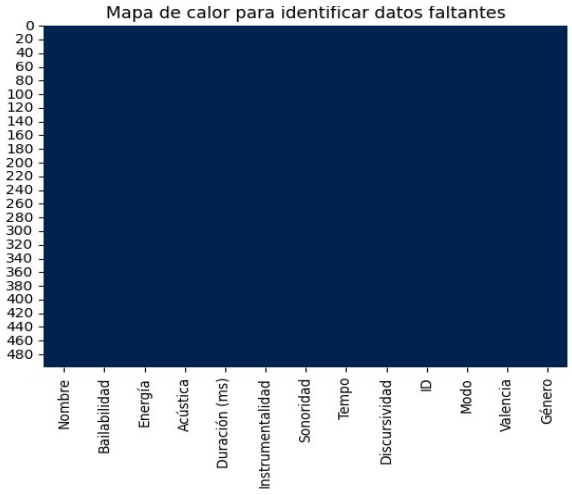


Figura 3.1. *Se realiza una verificación exhaustiva utilizando la herramienta de* ***Mapa de Calor****, confirmando que no se presentan valores faltantes en la colección de datos originales.*

310	Mana God	0.363	0.96	0.000165	220560	0.00151	-4.364	88.468	0.177	5dXZHtpvc	1	0.216 black-metal
311	Sun//Eater	0.245	0.889	0.000093	370493	0.835	-3.401	139.775	0.0884	1PZ1po0vz	1	0.052 black-metal
312	Pain Remains III: In a Sea of Fire	0.181	0.839	0.000202	552613	0.659	-4.733	135.13	0.083	6wgKN7m	0	0.0433 black-metal
313	A Miserable Life...	0.382	0.649	0.0864	255660	0.825	-10.079	130.13	0.0394	19F3MeGH	1	0.547 black-metal
314	Doomswitch	0.249	0.951	2.49E-05	275205	0.0134	-4.062	143.018	0.211	6pUUtKVR	0	0.117 black-metal
315	Weaponized	0.381	0.951	3.27E-05	187293	0.0712	-4.185	109.901	0.124	3EMS8JKV	0	0.299 black-metal
316	My Meds Aren't Working	0.288	0.792	0.000435	241625	0.587	-6.818	137.077	0.0486	3iubkenxO	1	0.16 black-metal
317	No Hard Feelings	0.401	0.95	4.17E-06	243987	0.397	-4.339	135.018	0.118	1Rx4mE1y	1	0.427 black-metal
318	Oscillator	0.454	0.951	1.57E-05	240787	0.000281	-4.54	100.011	0.128	6eVAkkl0U	0	0.243 black-metal
319	Venusian Blues	0.375	0.947	4.62E-06	168800	0.701	-4.779	160.029	0.062	7faxrkubPj	1	0.436 black-metal
320	Mana God	0.363	0.96	0.000165	220560	0.00151	-4.364	88.468	0.177	7efiu7ceW	1	0.216 black-metal

Figura 3.2. *Datos originales recopilados, segmentados por problemas de duplicación.*

Nombre	Bailabilidad	Energía	Acústica	Duración (ms)	Instrumentalidad	Sonoridad	Tempo	Discursividad	ID	Modo	Valencia	Género
In Death	0.551	0.965	0.000124	134560	0.71	-4.852	115.003	0.0424	6hPoIXMBQuX7Af4XyqBcSX	1	0.151	black-metal
Soulless Existence	0.414	0.93	0.0000077	432507	0.0311	-2.411	110.018	0.0571	2DGnHjfMxtSPINMAWFKcvq	0	0.214	black-metal
Nymphetamine Fix	0.462	0.905	0.00066	302360	0.0402	-3.825	122.925	0.0438	6WuqJLVZcyJklg7IIozAO8	0	0.253	black-metal
...And I Return to Nothingness	0.392	0.922	0.0000306	370154	0.068	-3.161	130.096	0.0692	6UTNQ8YN8da3ExqI5AOJPE	1	0.1	black-metal
Freezing Moon	0.168	0.974	0.0000305	383053	0.641	-9.116	93.424	0.112	4AP3a7eEOlz5sTJiWnv2C6	1	0.0943	black-metal
Take Me to Church	0.566	0.664	0.634	241693	0	-5.303	128.945	0.0464	1CS7Sd1u5tWkstBhpssyjP	0	0.437	pop
Apocalypse	0.369	0.468	0.0205	290617	0.566	-9.013	94.434	0.0273	0yc6Gst2xkRu0eMLeRMGCX	1	0.18	pop
It's ok I'm ok	0.754	0.514	0.0257	156522	0.0000398	-7.721	114.997	0.0471	24XihnoVPWXlKJ4BgXqjVM	0	0.363	pop
Saturn	0.411	0.619	0.62	186192	0	-6.834	177.937	0.0368	1bjeWoagtHmUKputLVyDxQ	1	0.337	pop
Night Changes	0.672	0.52	0.859	226600	0	-7.747	120.001	0.0353	5O2P9iiztwhomNh8xkR9IJ	1	0.37	pop
NIGHTS LIKE THIS	0.482	0.757	0.0137	86984	0.0016	-4.586	142.579	0.0355	1XBYiRV30ykHw5f4wm6qEn	1	0.12	hip-hop
No Scrubs	0.743	0.675	0.0251	214400	0.000717	-4.267	92.909	0.0953	1KGi9sZVMeszgZOWivFpxs	0	0.59	hip-hop
See You Again (feat. Kali Uchis)	0.558	0.559	0.371	180387	0.00000749	-9.222	78.558	0.0959	7KA4W4McWYRpgf0fWsJZWB	1	0.62	hip-hop
Pink + White	0.545	0.545	0.667	184516	0.0000548	-7.362	159.94	0.107	3xKsf9qdS1CyvXSMEid6g8	1	0.549	hip-hop
Not Like Us	0.898	0.472	0.0107	274192	0	-7.001	101.061	0.0776	6Al3ezQ4o3HUoP6Dhudph3	1	0.214	hip-hop
Middle	0.583	0.695	0.0138	220573	0	-5.336	104.879	0.0423	0g5EKLgdKvNlln7TNqBByK	1	0.224	electronic
Playground Love	0.33	0.423	0.265	211500	0.731	-10.504	143.086	0.0279	052z2UsE2wPrHsBJ9tly0g	0	0.117	electronic
La leçon particulière - Bande or	0.185	0.469	0.00652	105560	0.909	-7.729	150.319	0.0314	4OWa2dOlmtvMDhFrFL0QA1	0	0.422	electronic
Awake	0.552	0.877	0.0585	283636	0.899	-5.765	176.05	0.0353	5MhMXTuVODDF234VDvSxQx	1	0.597	electronic
Ghosts 'n' Stuff - Extended Mix	0.622	0.63	0.000047	328733	0.0961	-6.941	127.995	0.127	0KR5i46RXvDy8YzOZRvTAI	1	0.433	electronic

Figura 3.3. *Representación de datos depurados, completamente libres de duplicaciones, obtenidos durante la fase de* ***deduplicación****.*

```
Columna: Bailabilidad
Rango: 0.7829999999999999, Desviación estándar: 0.1811593480823759
-> Recomendación: Bailabilidad podría beneficiarse de la estandarización.

Columna: Energía
Rango: 0.9677, Desviación estándar: 0.18790984837404906
-> Recomendación: Energía podría beneficiarse de la estandarización.

Columna: Acústica
Rango: 0.9959985, Desviación estándar: 0.2581654884822389
-> Recomendación: Acústica podría beneficiarse de la estandarización.

Columna: Duración (ms)
Rango: 499040, Desviación estándar: 77618.61334634684
-> Recomendación: Duración (ms) podría beneficiarse de la normalización.

Columna: Instrumentalidad
Rango: 0.965, Desviación estándar: 0.23231185626625173
-> Recomendación: Instrumentalidad podría beneficiarse de la estandarización.

Columna: Sonoridad
Rango: 28.514, Desviación estándar: 3.230855308303301
-> Recomendación: Sonoridad podría beneficiarse de la estandarización.

Columna: Tempo
Rango: 145.223, Desviación estándar: 26.345050035089955
-> Recomendación: Tempo podría beneficiarse de la estandarización.

Columna: Discursividad
Rango: 0.4974, Desviación estándar: 0.07034694523038036
-> Recomendación: Discursividad podría beneficiarse de la estandarización.
```

Figura 3.4. *Se ilustra cómo las* ***métricas descriptivas*** *pueden proporcionar información sobre si los datos deben normalizarse o estandarizarse.*

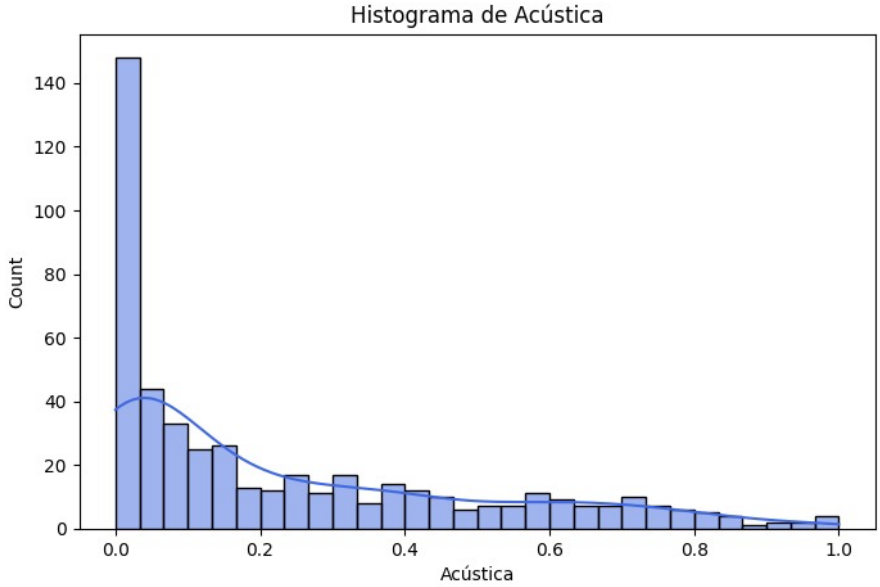


Figura 3.5. *Se muestra la representación de algunas columnas (datos) del conjunto mediante un histograma. Observamos que la distribución no sigue una forma normal, lo que sugiere la presencia de valores extremos significativos.*

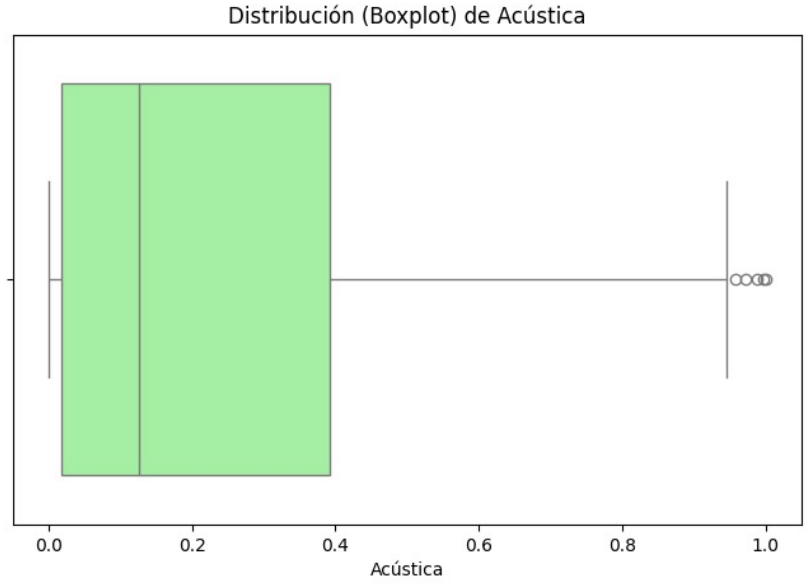


Figura 3.6. *Se presenta la misma columna que fue representada en la* ***Figura 3.5,*** *pero esta vez mediante una visualización de datos diferente. Se observa de manera inmediata que los valores atípicos se encuentran en el extremo superior del bigote derecho, y además, se aprecia un sesgo en la línea central (mediana) de la caja.*

	Nombre	Bailabilidad	Energía	Acústica	Duración (ms)	Instrumentalidad	Sonoridad	Tempo	Discursividad	ID	Modo	Valencia	Género
2	In Death	-0.30164635	0.974166	0.000123	0.129875361	0.735751295	0.882513853	0.415458	0.035786088	6hPoIXMB	0.837708	-1.2802	black-metal
3	Soulless E›	-1.05866749	0.937997	6.22E-06	0.726915678	0.032227979	0.968120923	0.381131	0.065339767	2DGnHjfM	-1.19373	-1.05012	black-metal
4	Nympheta	-0.7934338	0.912163	0.000661	0.466120952	0.041658031	0.918531248	0.470008	0.038600724	6WuqJLVZ	-1.19373	-0.90769	black-metal
5	...And I Re	-1.18023293	0.92973	2.92E-05	0.601969782	0.070466321	0.941818054	0.519387	0.089666265	6UTNQ8Yf	0.837708	-1.46645	black-metal
6	Freezing M	-2.41799012	0.983466	2.91E-05	0.627817409	0.664248705	0.732973276	0.266865	0.175713711	4AP3a7eE(	0.837708	-1.48727	black-metal
7	Bones	-1.7438545	0.874961	5.96E-05	0.224729481	0	0.871887494	0.381441	0.099718536	0FP5lpRlJK	0.837708	-0.24668	black-metal
8	Hollowed	-1.67754608	0.994833	5.54E-05	0.324927861	0.000515026	0.952058638	0.248893	0.280257338	3ffljpjKfPIS	0.837708	-0.50232	black-metal
9	Cursed to	-1.16918152	0.956598	8.84E-06	0.421314925	0.01388601	0.975275303	0.518637	0.095295537	1nqXQo7Y	0.837708	-0.94421	black-metal
10	Leech	-1.94277977	0.989666	0.000139	0.431654777	0.068290155	0.900189381	0.484455	0.227985525	5XaueS6W	0.837708	-1.68594	black-metal

Figura 3.7. *Se muestra el resultado de aplicar la técnica de* ***estandarización*** *a un conjunto de datos. Mediante este proceso, se logra una escala uniforme en todas las variables, lo que da como resultado un nuevo conjunto de datos completamente escalado.*

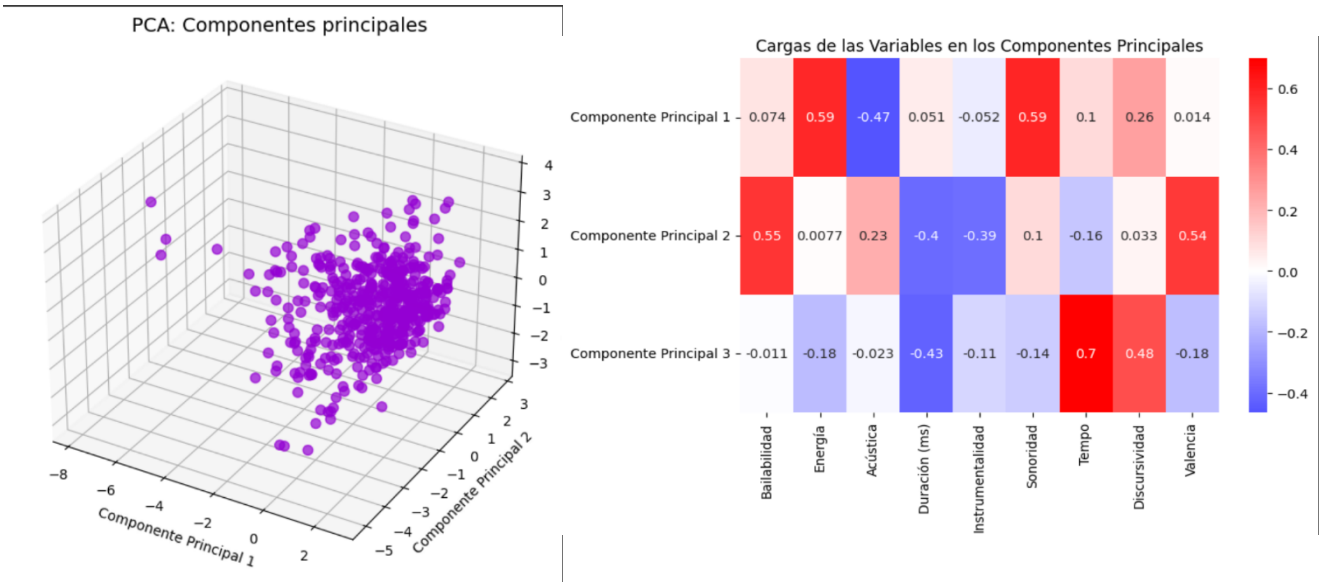


Figura 3.8. *Se presentan diversos diagramas que representan los tres componentes principales generados mediante el método de* ***Análisis de Componentes Principales*** *(**PCA**). Estos componentes capturan aproximadamente el 50% o 60% de la información contenida en nuestro conjunto de datos.*

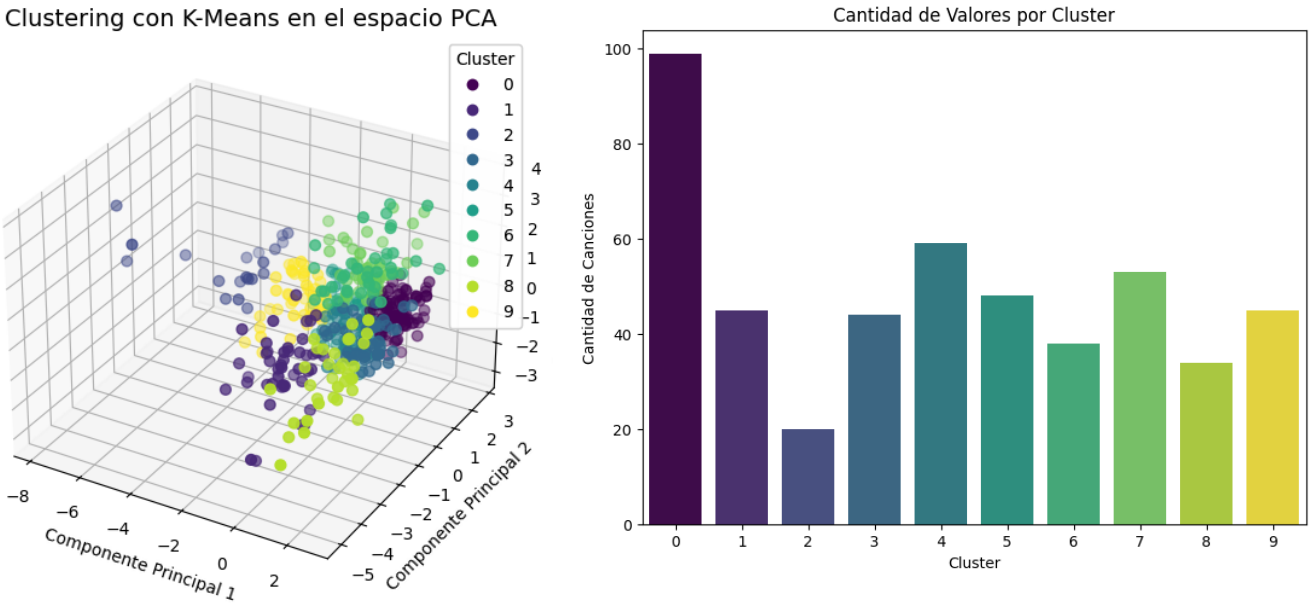


Figura 3.9. *Se muestran diversos diagramas que representan cómo los datos se agruparon en distintos clústeres, los cuales se formaron en función de sus características similares.*

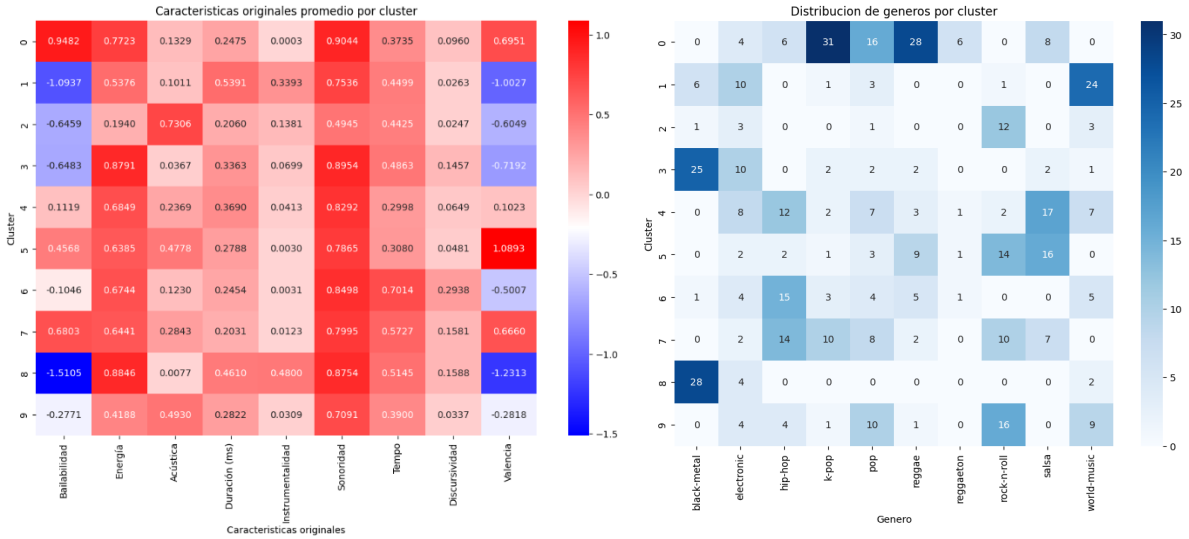


Figura 3.10. *Se presentan diagramas de calor en del lado izquierdo indica los valores promedio de las características que componen cada clúster, mientras que el del lado derecho refleja la frecuencia con la que se repiten los géneros dentro de un clúster.*

CAPÍTULO IV
Conclusión

Se presenta una justificación explícita que evalúa
si se logró con éxito el propósito establecido
durante el análisis del conjunto de datos musicales,
identificando tanto los logros como los fallos.

En conclusión, el uso del Clustering ha proporcionado una base sólida para explorar y segmentar datos musicales complejos, permitiendo una categorización basada en las características extraídas de las canciones. Este análisis ha demostrado que las características musicales como la energía, la bailabilidad, la instrumentalidad y otros atributos cuantificables pueden utilizarse como indicadores clave para agrupar canciones con perfiles similares.

El modelo ha cumplido con los objetivos establecidos, generando clusters que reflejan los distintos estilos musicales, lo que resulta valioso tanto para el análisis descriptivo como para la implementación de aplicaciones prácticas. Por ejemplo, los resultados obtenidos pueden ser utilizados para mejorar sistemas de recomendación de música personalizada, que podrían sugerir canciones basadas en los intereses del usuario y su similitud con otros grupos musicales previamente definidos. Asimismo, la identificación de tendencias dentro de los clusters permite explorar nuevos estilos o subgéneros que podrían no haber sido evidentes previamente, lo que beneficia tanto a oyentes como a creadores de contenido musical.

El análisis también destacó cómo ciertos géneros musicales tienden a agruparse en clusters específicos, lo que demuestra que las características técnicas de las canciones reflejan, las categorizaciones subjetivas de los géneros. Sin embargo, el modelo también identificó casos en los que un mismo género se distribuye entre diferentes clústeres, lo que podría interpretarse como una muestra de la diversidad interna de ciertos estilos musicales o una fusión entre diferentes influencias. Esto abre la puerta a un análisis más detallado sobre cómo los géneros y subgéneros evolucionan con el tiempo.

Además, la capacidad de detectar canciones con características atípicas o que se desvían de los patrones comunes dentro de un clúster destaca el potencial del modelo para descubrir música innovadora. Estas canciones, que podrían ser valores atípicos o innovaciones, representan una oportunidad para identificar artistas emergentes o estilos únicos que desafían las tendencias establecidas. Este tipo de análisis tiene implicaciones prácticas significativas para la industria musical, ya que podría utilizarse para listas de reproducción más diversas o incluso para el análisis competitivo entre artistas y géneros.