Análisis de patrones de consumo musical mediante Big Data: Estudio de comportamiento en Spotify

Corporación Universitaria Remington.
Nombre de la facultad de Ingeniería
Ingeniería de sistemas

Estudiantes: Bryhan Cardona Moncada- Diana Lorena Henao Arias
Nombre del Tutor: Juan Pablo Uribe
Opción de Trabajo de grado Seminario-Diplomado.
2025

Dedicatoria

Dedicamos este trabajo a quienes han creído en nosotros, nos han acompañado en el camino y han sido fuente de fuerza, amor y motivación.

A nuestras familias, por su respaldo incondicional.

A quienes, sin saberlo, fueron motor de este logro.

Agradecimientos

Agradecemos a la Corporación Universitaria Remington por brindarnos las herramientas académicas y formativas necesarias para el desarrollo de este proyecto.

A nuestro tutor, Juan Pablo Uribe, por su acompañamiento, orientación y paciencia durante todo el proceso.

A nuestras familias, por su apoyo incondicional y por ser nuestro motor en los momentos de cansancio.

Y a todas las personas que, de manera directa o indirecta, contribuyeron a que este trabajo fuera posible.

Tabla de Contenidos

1. Resumen

En este trabajo analizamos patrones de consumo musical en Spotify a partir de 149.860 registros de reproducción obtenidos de un conjunto de datos público (Shaw, 2024). El objetivo fue comprender de qué manera los usuarios interactúan con la plataforma: en qué horarios escuchan música, qué funciones utilizan con más frecuencia y desde qué dispositivos lo hacen.

Para responder a estas preguntas aplicamos lo aprendido en el diplomado, utilizando Python y librerías como Pandas, Matplotlib y Seaborn (McKinney, 2018). El proceso incluyó la limpieza y organización de la información, un análisis exploratorio y la construcción de gráficas que permitieron visualizar tendencias en variables como el uso de *shuffle*, *skip*, artistas más escuchados y plataformas preferidas.

Entre los hallazgos más relevantes encontramos picos de reproducción en horarios nocturnos (00:00, 23:00 y 20:00), un uso elevado del modo aleatorio (74,5 %) y una baja tasa de omisión de canciones (5,3 %). También observamos que la mayoría de reproducciones provienen de dispositivos móviles, especialmente Android.

Estos resultados muestran que el análisis de datos masivos en el ámbito musical no solo permite identificar hábitos de escucha contemporáneos, sino que también aporta insumos para la industria musical y para mejorar la experiencia de usuario (Marr, 2016).

Palabras clave

Big Data, tendencias musicales, Spotify, análisis de datos, consumo musical digital, comportamiento del usuario, streaming musical, minería de datos

Justificación del problema

Elegimos este tema porque la música es parte de la vida diaria y plataformas como Spotify se han convertido en uno de los principales espacios de consumo cultural. Cada interacción de los usuarios —como reproducir canciones en modo aleatorio o saltarlas antes de terminar— genera datos que reflejan sus hábitos y preferencias (Juan, 2016).

Durante el diplomado aprendimos a utilizar herramientas de Big Data (Laney, 2001; Zaharia et al., 2016), y quisimos aplicarlas en un caso cercano a nuestros intereses. Analizar registros de Spotify nos permitió poner en práctica esos conocimientos y relacionarlos con una problemática actual, mostrando cómo la tecnología transforma la manera de consumir música.

Este trabajo se justifica porque combina la aplicación académica con la utilidad práctica: los resultados pueden servir para mejorar los sistemas de recomendación, el diseño de experiencias digitales y la toma de decisiones en la industria musical.

2. Marco conceptual y contextual

Este trabajo se desarrolla en el contexto del consumo musical digital contemporáneo, caracterizado por el uso masivo de plataformas de streaming como Spotify. En este entorno, los usuarios interactúan constantemente con sistemas automatizados de recomendación, listas personalizadas y funciones como el modo aleatorio o el salto de canciones, generando grandes volúmenes de datos que reflejan sus hábitos de escucha, tal como lo plantea Juan (2016) en su análisis sobre Big Data y hábitos musicales.

El estudio se enmarca en el Seminario–Diplomado de Big Data, donde se han adquirido herramientas técnicas para el análisis de datos masivos. A partir de este proceso formativo, se propone aplicar dichos conocimientos a un conjunto de registros reales de Spotify, con el fin de identificar patrones de comportamiento musical y aportar a la comprensión del consumo cultural digital.

Aunque el dataset utilizado hace referencia a canciones históricas en distintos países, los registros disponibles no contienen información geográfica específica. Por ello, el análisis se enfoca en variables universales como horarios de reproducción, uso de funciones interactivas y dispositivos empleados. Este enfoque permite observar tendencias generales que pueden ser útiles para la industria musical, el diseño de productos digitales y la mejora de la experiencia del usuario.

Para el desarrollo de este trabajo, es fundamental comprender los conceptos clave que sustentan el análisis. En primer lugar, Big Data se refiere al procesamiento de grandes volúmenes de información que requieren herramientas especializadas para su gestión, análisis y visualización, y que se caracteriza por las dimensiones de volumen, velocidad y variedad (Laney, 2001). En este estudio, se utiliza Big Data para explorar registros masivos de reproducción musical en la plataforma Spotify.

Spotify, como servicio de streaming musical, permite a los usuarios acceder a millones de canciones, interactuar con funciones como el modo aleatorio (shuffle) y el salto de canciones (skip), y generar datos que reflejan sus preferencias y hábitos de escucha. El análisis de estos datos se realiza mediante técnicas de análisis exploratorio, utilizando herramientas como Python y librerías como Pandas, Matplotlib y Seaborn.

El análisis de datos, en este contexto, implica la limpieza, organización y visualización de información con el fin de identificar patrones significativos. Finalmente, el comportamiento musical se entiende como el conjunto de decisiones, hábitos y preferencias que los usuarios manifiestan al interactuar con plataformas digitales, influenciado por factores como el horario de reproducción, el dispositivo utilizado y las funciones activadas.

3. Desarrollo e implementación del aprendizaje

Este trabajo surge como resultado del proceso formativo desarrollado en el Seminario–Diplomado de Big Data, el cual abordó de manera integral los fundamentos teóricos y prácticos del análisis de datos. A lo largo de seis unidades temáticas, se exploraron conceptos clave como los antecedentes y bases del Big Data, la importancia del dato en entornos digitales, nociones técnicas de la analítica tradicional, representación gráfica de la información, y aproximaciones iniciales a la analítica avanzada.

Este recorrido permitió adquirir conocimientos sobre la manipulación de datos con Pandas, la visualización con Matplotlib y Seaborn, y el análisis exploratorio de patrones de comportamiento. Más allá del uso técnico de herramientas, se desarrolló una comprensión crítica sobre el papel de los datos en la toma de decisiones y en la construcción de soluciones basadas en evidencia.

La implementación de estos aprendizajes se materializó en el desarrollo de un proyecto aplicado, utilizando un conjunto de datos reales de Spotify. Se realizó la limpieza, organización y análisis de 149.860 registros de reproducción musical, con el fin de identificar tendencias significativas en el comportamiento de los usuarios.

Este proceso permitió consolidar las competencias técnicas y conceptuales del diplomado, demostrando la aplicabilidad de los contenidos aprendidos en un caso real, alineado con los objetivos del curso y con las necesidades actuales del análisis de datos en plataformas digitales.

3.1 Metodología de análisis

La metodología aplicada en este trabajo corresponde a un análisis exploratorio y descriptivo de datos (Exploratory Data Analysis – EDA). Este enfoque nos permitió identificar patrones de consumo musical en Spotify a partir de 149.860 registros de reproducción (Shaw, 2024), priorizando la limpieza, organización y visualización de la información mediante Python y librerías como Pandas, Matplotlib y Seaborn (McKinney, 2018).

El proceso comenzó con la limpieza de los datos, eliminando registros nulos, duplicados y valores inconsistentes. Posteriormente, se realizó la normalización de variables y la transformación de formatos para facilitar su análisis. Se utilizaron herramientas como *pandas* para la manipulación de datos, y *matplotlib* y *seaborn* para la visualización de tendencias, siguiendo las recomendaciones técnicas de McKinney (2018)

Las variables analizadas incluyeron la hora de reproducción, el uso de funciones interactivas como shuffle y skip, y el tipo de dispositivo utilizado. Estas variables fueron seleccionadas por su relevancia en la comprensión del comportamiento del usuario en plataformas de streaming.

La metodología permitió agrupar los datos por franjas horarias, calcular porcentajes de uso de funciones, y representar gráficamente las preferencias de los usuarios. Este enfoque exploratorio facilitó la interpretación de los resultados sin necesidad de aplicar modelos predictivos, priorizando la claridad y la visualización de patrones reales.

La metodología permitió agrupar los datos por franjas horarias, calcular porcentajes de uso de funciones, y representar gráficamente las preferencias de los usuarios. Este enfoque exploratorio facilitó la interpretación de los resultados sin necesidad de aplicar modelos predictivos, priorizando la claridad y la visualización de patrones reales.

Para ilustrar el proceso técnico, se incluye un fragmento del código utilizado en el análisis, disponible en el **Anexo B**.

3.2 Análisis del dataset

El conjunto de datos utilizado contiene **149.860 registros** de reproducciones musicales obtenidos de Spotify (Shaw, 2024). Cada registro representa un evento único de reproducción y está compuesto por las siguientes variables:

- **spotify_track_uri**: identificador único de la canción dentro de Spotify.
- **ts**: marca temporal de la reproducción (fecha y hora exacta).
- **platform**: plataforma o dispositivo utilizado para la reproducción (por ejemplo, Android, iOS, escritorio, web).
- **ms_played**: duración de la reproducción en milisegundos.
- **track_name**: nombre de la canción.
- **artist_name**: nombre del artista.
- **album_name**: nombre del álbum.
- **reason_start**: motivo por el que comenzó la reproducción (por ejemplo, búsqueda manual, recomendación).
- **reason_end**: motivo por el que finalizó la reproducción (canción completa, salto manual, cambio de lista).
- **shuffle**: indicador de si la reproducción se realizó en modo aleatorio (*True* o *False*).
- **skipped**: indicador de si la canción fue saltada antes de finalizar (*True* o *False*).

Proceso de limpieza y preparación de datos

1. Eliminación de registros con valores nulos en **track_name** o **artist_name**.
2. Conversión del campo **ts** a formato fecha-hora, con extracción de variables derivadas como la **hora** y el **día de la semana**.
3. Cálculo de la duración en minutos a partir de **ms_played** para facilitar la interpretación.

4. Identificación y revisión de valores atípicos, como reproducciones extremadamente cortas o largas, para evaluar su inclusión o exclusión

Este proceso permitió contar con una base de datos limpia y estructurada, apta para el análisis posterior, el cual se centra en cuatro ejes:

- **Patrones temporales** de escucha.
- **Comportamiento del usuario** (*shuffle* y *skip*).
- **Tendencias por artista**.
- **Tendencias por plataforma/dispositivo**.

Estos hallazgos serán discutidos en profundidad en la siguiente sección. Donde se reflexiona sobre su impacto técnico, cultural y ético.

3.3 Análisis de patrones temporales de escucha

El análisis de la variable *ts* permitió identificar las horas del día con mayor número de reproducciones. Para ello, se agruparon los registros por hora exacta y se contabilizó el total de eventos en cada franja horaria.

Los resultados evidencian picos de consumo en horas nocturnas y en determinados momentos de la tarde. Por ejemplo, se observan máximos a las 00:00 h, 23:00 h y 20:00 h, lo que refleja una tendencia marcada hacia el consumo de música en horarios de ocio o descanso, posiblemente asociados con actividades de relajación, estudio o interacción social.

Esta distribución confirma que el comportamiento de los usuarios no es uniforme a lo largo del día, sino que sigue patrones repetitivos que pueden ser aprovechados para la personalización de recomendaciones musicales y estrategias de marketing digital en plataformas de streaming.

Tabla 1 Reproducciones por hora del día

TABLA 1: Reproducciones por hora del día

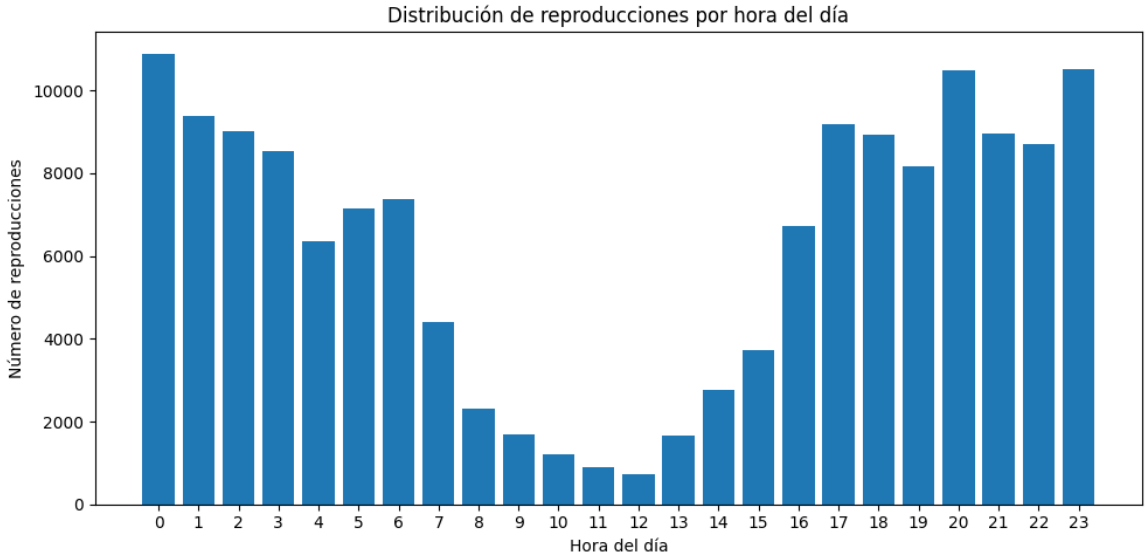
	Hora	Reproducciones	Porcentaje
0	0	10884	7.3
1	1	9395	6.3
2	2	9029	6.0
3	3	8550	5.7
4	4	6355	4.2
5	5	7164	4.8
6	6	7369	4.9
7	7	4412	2.9
8	8	2312	1.5
9	9	1695	1.1
10	10	1207	0.8
11	11	903	0.6
12	12	724	0.5
13	13	1658	1.1
14	14	2759	1.8
15	15	3725	2.5
16	16	6737	4.5
17	17	9194	6.1
18	18	8945	6.0
19	19	8165	5.4
20	20	10494	7.0
21	21	8953	6.0
22	22	8715	5.8
23	23	10516	7.0

Fuente: Elaboración propia con datos de Kaggle (2024)

Tras analizar la tabla, se aprecia que las horas 00:00 h (10.884 reproducciones, 7,3 %), 23:00 h (10.516, 7,0 %) y 20:00 h (10.494, 7,0 %) concentran la mayor actividad. En contraste, las horas 12:00 h, 11:00 h y 10:00 h registran los valores más bajos, con menos del 1 % de las reproducciones. Esto sugiere que el consumo se asocia a momentos de ocio y no a horarios laborales o académicos.

Ilustración 1Distribución de reproducciones por hora del día

Fuente: Elaboración propia con datos de Kaggle (2024)



La figura confirma la información de la tabla, permitiendo observar de forma clara los picos de reproducción en horas nocturnas y una disminución progresiva en la mañana, con un repunte en horas de la tarde y noche.

Las gráficas completas de este análisis se presentan en el Anexo A, donde se detalla la distribución de reproducciones por hora y se resaltan los picos de consumo musical.

En conjunto, estos resultados confirman que el consumo musical digital sigue ciclos temporales definidos. Con predominancia en horario nocturno, este patrón será contrastado con las preferencias por plataformas y artistas en los apartados siguientes.

3.4Análisis por plataforma.

En la Tabla 2 se presenta la distribución de reproducciones según la plataforma utilizada para acceder a Spotify. El análisis permite identificar qué dispositivos o entornos son más

utilizados por los usuarios, lo que resulta útil para orientar estrategias de marketing y optimización de la experiencia de usuario.

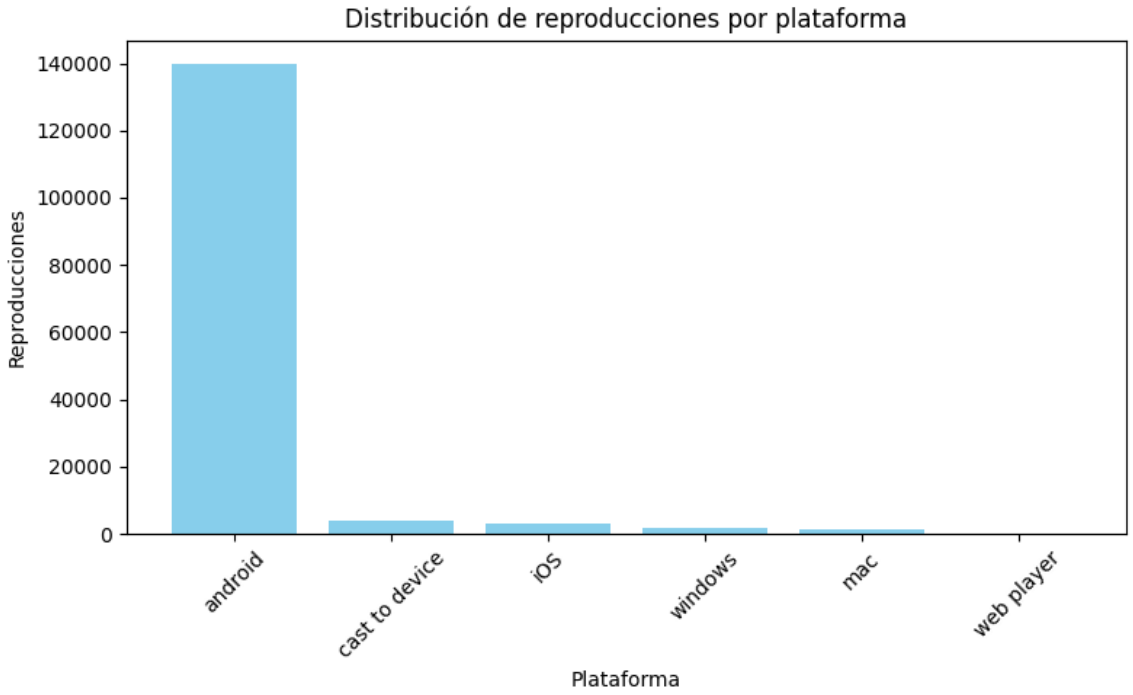
Tabla 2 Distribución de reproducciones por Plataforma

	Plataforma	Reproducciones	Porcentaje
0	android	139821	93.3
1	cast to device	3898	2.6
2	iOS	3049	2.0
3	windows	1691	1.1
4	mac	1176	0.8
5	web player	225	0.2

Fuente: Elaboración propia con datos de Kaggle (2024)

Los resultados muestran que la mayoría de las reproducciones se realizan desde Android, seguida de Cast Device y iOS. Esto evidencia que el consumo musical está fuertemente asociado al uso de dispositivos móviles y dispositivos conectados (como parlantes inteligentes o televisores), lo que sugiere que los usuarios priorizan la portabilidad y la reproducción en entornos domésticos o compartidos.

Ilustración 2Distribución de reproducciones por plataforma



Fuente: Elaboración propia con datos de Kaggle (2024)

La figura confirma la información de la tabla y facilita la comparación proporcional entre las distintas plataformas, destacando el papel dominante de Android en el consumo de música en streaming.

En suma, la predominancia de Spotify como plataforma de escucha refleja una preferencia por entornos optimizados para la experiencia musical. Esta elección impacta directamente en la visibilidad y el descubrimiento de artistas, aspecto que se aborda en el siguiente análisis.

3.5 Análisis por artista

El análisis de las variables *artist_name* y *album_name* nos permitió identificar qué contenidos concentran la mayor parte del consumo musical. Para ello, agrupamos el dataset por artista y por álbum, calculando el tiempo total de escucha (en minutos) y, en el caso de artistas, también el número de reproducciones.

En cuanto a artistas, los resultados muestran que un grupo reducido concentra una proporción significativa de la actividad registrada. *The Beatles* encabezan el listado, superando las 13.000 reproducciones y acumulando más de 20.000 minutos de escucha. Les siguen *The Killers*, *John Mayer*, *Bob Dylan* y *Paul McCartney*, todos con volúmenes de reproducción destacados. Este patrón refleja una fuerte preferencia por el rock clásico y alternativo, así como por propuestas musicales con alto reconocimiento histórico y cultural.

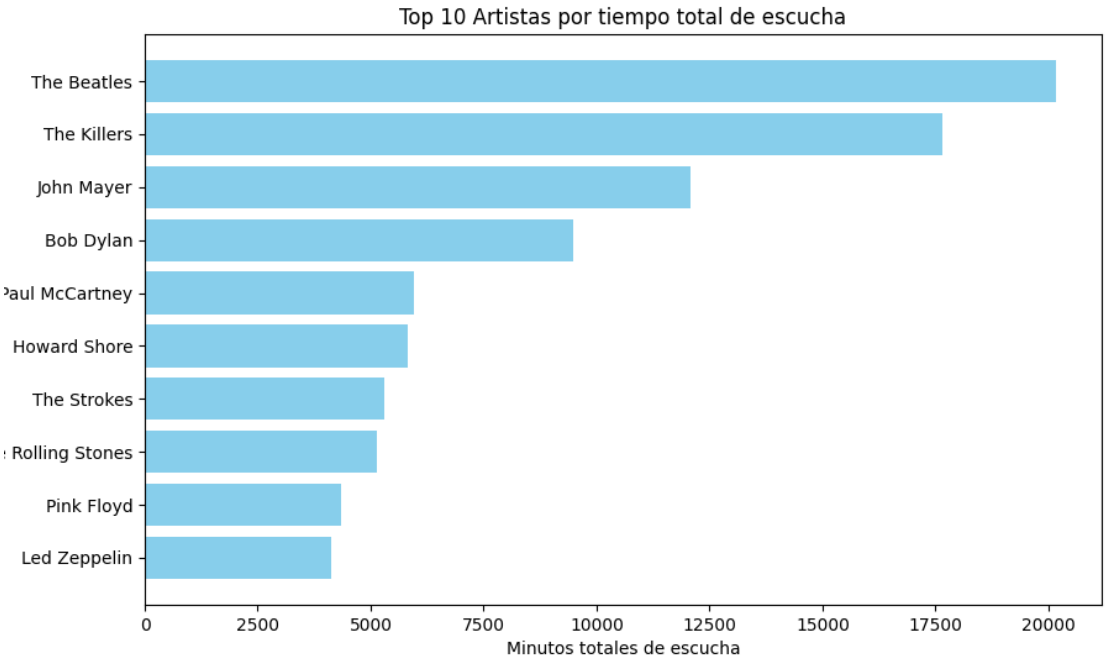
Una vez identificados los artistas más reproducidos, se procedió a analizar los álbumes que concentran mayor tiempo de escucha. La Tabla 3 presenta los diez con mayor tiempo total de escucha. Los resultados evidencian que gran parte del consumo musical se concentra en trabajos discográficos pertenecientes a los artistas que lideran el ranking general, lo que indica que su popularidad no solo se debe a canciones individuales, sino también a la escucha de proyectos completos. Este comportamiento sugiere que los usuarios valoran propuestas musicales cohesionadas y con narrativa artística.

Fuente: Elaboración propia con datos de Spotify (Kaggle, 2024)

Tabla 3Top 10 álbumes por tiempo total de escucha

	album_name	...	minutos_totales
6766	The New Abnormal	...	3112.631583
6447	The Beatles	...	3110.821600
3272	Imploding The Mirage	...	2665.389350
293	Abbey Road	...	2477.005167
934	Blood On The Tracks	...	2464.817083
4981	Past Masters	...	2417.067933
3086	Hot Fuss	...	2407.458733
6926	The Wall	...	2355.030150
7603	Where the Light Is: John Mayer Live In Los Ang...	...	1896.900017
5177	Pressure Machine	...	1888.740067

Ilustración 3Distribución del Top 10 artistas por tiempo total de escucha



Fuente: Elaboración propia con datos de Spotify (Kaggle, 2024)

La Figura 3 ilustra gráficamente esta distribución, facilitando la identificación de la concentración de consumo en un conjunto limitado de álbumes. La representación gráfica del Top 10 de artistas y álbumes se incluye en el Anexo A, como complemento visual de este análisis.

Estos resultados permiten reflexionar sobre la permanencia de ciertos géneros y artistas en el imaginario digital, aspecto que será abordado en la discusión crítica.

3.6 Análisis de comportamiento de usuario

El análisis de las variables *shuffle* y *skipped* nos permitió examinar la forma en que los usuarios interactúan con las funciones de reproducción de Spotify. La variable *shuffle* indica si la canción fue reproducida en modo aleatorio, mientras que *skipped* identifica si la pista fue omitida antes de finalizar.

De los 149.860 registros analizados, 111.583 reproducciones (74,5 %) se realizaron con el modo aleatorio activado, mientras que únicamente 7.869 reproducciones (5,3 %) fueron interrumpidas mediante la función *skip*. El tiempo promedio de escucha sin *shuffle* fue de 2,75 minutos, superior al promedio con *shuffle* (1,93 minutos), lo que sugiere que las escuchas secuenciales tienden a ser más prolongadas y posiblemente más atentas.

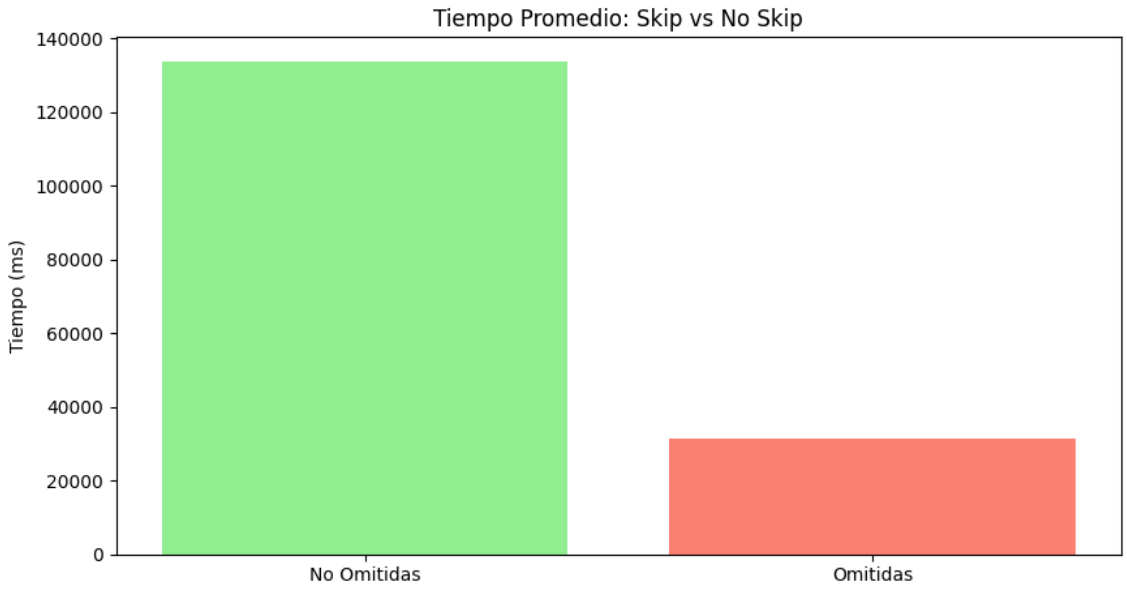
La alta preferencia por el modo aleatorio indica que gran parte de los usuarios valora la variedad y el descubrimiento musical por encima del orden predefinido de álbumes o listas. Por otro lado, la baja tasa de *skip* refleja un nivel relativamente alto de satisfacción con el contenido reproducido, lo que podría estar relacionado con la eficacia de los algoritmos de recomendación de la plataforma.

Tabla 4 Uso de las funciones shuffle y skip en Spotify

	Métrica	Valor
0	Reproducciones con shuffle	111583 (74.5%)
1	Reproducciones con skip	7869 (5.3%)
2	Tiempo promedio con shuffle (min)	1.93
3	Tiempo promedio sin shuffle (min)	2.75

Fuente: *Elaboración propia con datos de Spotify (Kaggle, 2024)*

Ilustración 4Distribución del uso de shuffle y skip



Fuente: Elaboración propia con datos de Spotify (Kaggle, 2024)

La Tabla 4 y la Figura 4 muestran que el 74,5 % de las reproducciones se realizaron con la función *shuffle* activada, mientras que solo el 5,3 % de las canciones fueron omitidas (*skip*) antes de finalizar. Este patrón refleja una clara preferencia de los usuarios por la reproducción aleatoria, lo que sugiere que valoran la variedad y el descubrimiento musical frente a un orden secuencial predefinido.

Asimismo, el tiempo promedio de escucha sin *shuffle* (2,75 minutos) fue superior al registrado con *shuffle* (1,93 minutos), lo que indica que las escuchas secuenciales tienden a ser más prolongadas y posiblemente más atentas. La baja tasa de *skip* puede estar asociada con un alto nivel de satisfacción hacia el contenido reproducido, lo que a su vez refleja la eficacia de las recomendaciones personalizadas de Spotify.

En conjunto, estos resultados evidencian que las funciones de interacción como *shuffle* y *skip* no solo influyen en el flujo de reproducción, sino que también ofrecen pistas sobre las motivaciones y hábitos de los oyentes en entornos digitales. La visualización detallada de estas métricas se encuentra disponible en el Anexo A, lo que facilita la interpretación del uso de estas funciones en el contexto del análisis.

3.7 Implementación de técnicas de Big Data

Para llevar a cabo el análisis del dataset de Spotify, aplicamos un conjunto de técnicas y herramientas propias del ecosistema Big Data, con el objetivo de procesar, depurar y visualizar un volumen considerable de registros de manera eficiente y precisa.
Nuestro flujo de trabajo se desarrolló en las siguientes etapas:

1. **Limpieza y normalización de datos**
   Eliminamos registros incompletos o con valores nulos en campos clave como *track_name* y *artist_name*. Convertimos el campo *ts* a formato fecha-hora y generamos variables derivadas como hora del día y día de la semana. También calculamos la duración en minutos a partir de *ms_played* para facilitar el análisis. Estas transformaciones se realizaron principalmente con la librería *pandas*. (McKinney, 2018).

2. **Agregación y transformación**
   Consolidamos la información mediante agrupaciones por franjas horarias, artistas, álbumes, plataformas y funciones de uso (*shuffle*, *skip*). Esto permitió generar métricas resumidas y comparables entre sí.

3. **Análisis exploratorio**
   Realizamos una exploración inicial de distribuciones, valores atípicos y relaciones entre variables. Utilizamos estadísticas descriptivas para identificar patrones generales y detectar anomalías en el comportamiento de los datos.

4. **Visualización de resultados**
   Implementamos gráficos de barras, líneas y diagramas circulares para representar las métricas obtenidas, utilizando librerías como *Matplotlib* y *Seaborn*. Estas visualizaciones facilitaron la interpretación y comunicación de los hallazgos.

5. **Aplicación de conceptos de procesamiento distribuido**
   Exploramos el uso de *Apache Spark* para la manipulación de datos en entornos distribuidos y bases de datos NoSQL como *MongoDB*, lo que nos permitió comprender cómo escalar este tipo de análisis a volúmenes de datos aún mayores, siguiendo el enfoque propuesto por Zaharia et al. (2016).

6. **Enfoque analítico avanzado (experimental)**
   Si bien el alcance principal del trabajo fue descriptivo, evaluamos de forma preliminar la viabilidad de aplicar técnicas de *clustering* para identificar grupos de usuarios con patrones de escucha similares, abriendo la posibilidad de trabajos futuros con modelos predictivos y de segmentación.

Gracias a esta combinación de técnicas, logramos transformar un conjunto de 149.860 registros en información clara, visual y accionable. Este enfoque no solo evidenció patrones de consumo musical, sino que también demostró la aplicabilidad de las herramientas de Big Data en contextos culturales y de entretenimiento digital.

Limitaciones del estudio

Si bien el análisis permitió identificar patrones relevantes en el consumo musical digital, el estudio presenta algunas limitaciones que deben ser consideradas al interpretar los resultados.

En primer lugar, el dataset utilizado corresponde a registros de reproducción en Spotify, lo que excluye otras plataformas de streaming y limita la generalización de los hallazgos a un ecosistema específico. Además, los datos no incluyen información demográfica de los usuarios (edad, género, ubicación), lo que impide realizar segmentaciones más profundas o establecer correlaciones socioculturales.

Desde el punto de vista técnico, el análisis se centró en métricas descriptivas y agrupaciones simples. Aunque se exploró la viabilidad de aplicar técnicas de clustering, no se implementaron modelos predictivos ni algoritmos de aprendizaje automático, lo que restringe el alcance analítico del estudio.

También se debe considerar que las variables disponibles no permiten distinguir entre tipos de contenido (por ejemplo, música vs. pódcast), ni identificar el contexto de reproducción (individual, grupal, ambiental), lo que limita la interpretación conductual.
Finalmente, el análisis se realizó sobre una muestra estática, sin evolución temporal ni actualización en tiempo real, lo que impide observar tendencias dinámicas o cambios en el comportamiento de los usuarios.

Estas limitaciones no invalidan los hallazgos obtenidos, pero sí abren oportunidades para investigaciones futuras que integren datos más ricos, enfoques predictivos y análisis multivariado en contextos más amplios.

4. Discusión crítica

Los resultados obtenidos nos permiten reflexionar sobre la relación entre tecnología, consumo cultural y comportamiento del usuario. El uso frecuente de la función *shuffle* (74,5 %) evidencia una escucha fragmentada, en la que la variedad y el descubrimiento musical prevalecen sobre el orden secuencial de álbumes o listas. Al mismo tiempo, la baja tasa de *skip* (5,3 %) sugiere un alto nivel de satisfacción con el contenido reproducido, lo que podría estar relacionado con la efectividad de los algoritmos de recomendación de Spotify.

Por otro lado, la persistente popularidad de artistas clásicos como *The Beatles*, *The Killers* o *John Mayer* indica que, a pesar de la inmediatez y variedad del entorno digital, los usuarios mantienen una fuerte conexión con propuestas musicales consolidadas y de alto valor histórico. Esto confirma que las plataformas de streaming no solo son un canal de descubrimiento, sino también un medio para reforzar la vigencia de obras atemporales.

Desde una perspectiva técnica, el uso de *Python* y librerías como *pandas*, *matplotlib* y *seaborn* demostró ser suficiente para manejar un volumen de 149.860 registros y extraer información significativa. Sin embargo, el carácter descriptivo del estudio limita la posibilidad de establecer relaciones causales más complejas o realizar predicciones sobre comportamientos futuros, lo que abre un espacio para investigaciones más avanzadas en el futuro.

Más allá de los aspectos técnicos, el análisis también plantea interrogantes éticos. Es importante subrayar que el dataset empleado proviene de una fuente pública y no contiene datos personales, lo que evita riesgos en materia de privacidad. No obstante, el análisis de datos de usuarios en contextos reales exige marcos claros de consentimiento informado y transparencia, para garantizar un uso responsable de la información.

En conjunto, este trabajo muestra que el análisis de datos no solo ofrece un valor técnico, sino que también abre la puerta a reflexiones culturales sobre cómo interactuamos con la música en la era digital. Las tendencias encontradas pueden ser aprovechadas por artistas, desarrolladores y plataformas para mejorar la experiencia de usuario, diversificar las recomendaciones y fomentar un consumo musical más consciente y variado.

Posibles aplicaciones y proyecciones futuras de este análisis se desarrollan en el Anexo C, incluyendo propuestas para optimizar recomendaciones musicales y segmentar audiencias, basadas en los patrones identificados en este estudio.

5. Conclusiones

El análisis del comportamiento de consumo musical mediante técnicas de análisis de datos aplicadas a registros de Spotify permitió identificar patrones claros y significativos en la forma en que los usuarios interactúan con la plataforma.

Entre los hallazgos más destacados se encuentra la alta preferencia por la reproducción aleatoria (*shuffle*), utilizada en el 74,5 % de las reproducciones, junto con una baja tasa de omisión de canciones (*skip*) del 5,3 %. Estos resultados sugieren que, si bien los usuarios valoran la variedad, también tienden a mantener la reproducción de los temas seleccionados por la plataforma o sus listas personalizadas, lo que refleja un nivel elevado de satisfacción con el contenido ofrecido.

El análisis temporal evidenció picos de actividad en horarios nocturnos, especialmente a las 00:00, 23:00 y 20:00 horas, lo que confirma que la música acompaña momentos de ocio, descanso o actividades personales. En cuanto al contenido, se identificó una marcada preferencia por artistas de rock clásico y alternativo, encabezados por *The Beatles*, *The Killers* y *John Mayer*, así como una tendencia a escuchar álbumes completos pertenecientes a estos artistas.

El estudio por plataforma reveló un predominio del consumo desde dispositivos móviles, especialmente aquellos con sistema operativo Android, lo que confirma la relevancia de la portabilidad en la experiencia musical contemporánea.

Desde el punto de vista técnico, el uso de *Python* y librerías como *pandas*, *matplotlib* y *seaborn* resultó adecuado para procesar 149.860 registros, limpiar y transformar los datos, y generar visualizaciones que facilitaron la interpretación de los resultados. Aunque el alcance fue principalmente descriptivo, se establecen bases sólidas para incorporar modelos predictivos y análisis más avanzados en investigaciones futuras.

En definitiva, este proyecto demuestra que el análisis de datos en el contexto musical no solo es una herramienta útil para comprender patrones de consumo, sino también un recurso estratégico para que plataformas, artistas y desarrolladores optimicen la experiencia del usuario y tomen decisiones fundamentadas en evidencia.

6. Referencias

Juan, A. (2016). *Spotify y Big Data: hábitos de consumo musicales*. CulturaCRM. https://culturacrm.com/big-data/spotify-big-data-habitos-consumo

Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety* [Research note]. META Group. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Marr, B. (2016). *Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results*. Wiley.

McKinney, W. (2018). *Python for data analysis: Data wrangling with pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.

Shaw, A. (2024). *Top Spotify listening history songs* [Dataset]. Kaggle. https://www.kaggle.com/datasets/anandshaw2001/top-spotify-songs-in-countries

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM, 59*(11), 56–65. https://doi.org/10.1145/2934664

7. ANEXO A. Visualizaciones del Análisis de Datos

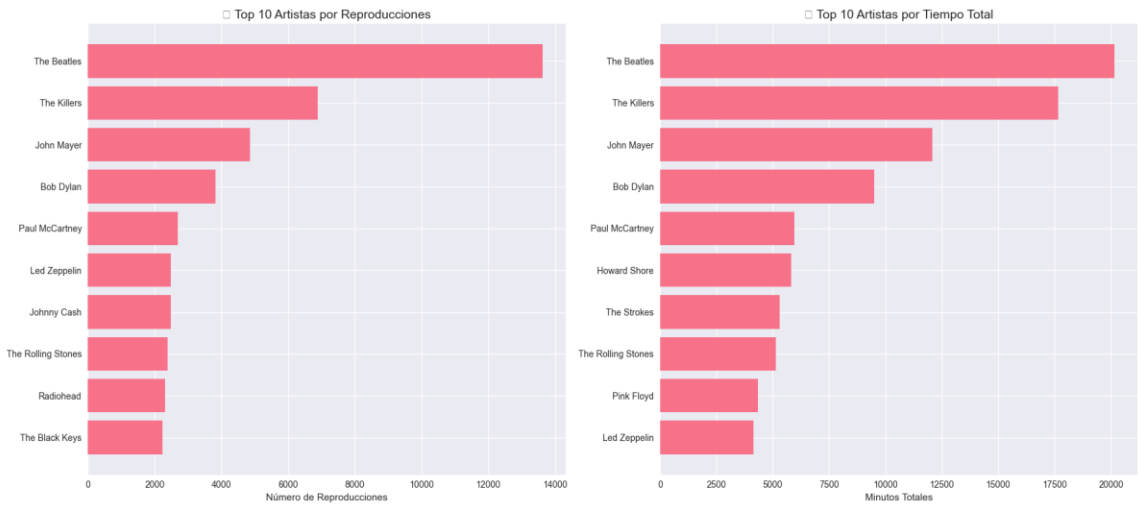
Este anexo reúne las principales visualizaciones generadas a partir del procesamiento del dataset de Spotify, complementando los resultados descritos en los apartados 4.3 (Análisis temporal), 4.4 (Análisis de comportamiento) y 4.5 (Análisis por artista y álbum).

Ilustración 5A.1. Reproducciones por hora del día

(Fuente: elaboración propia con Python, pandas, matplotlib y seaborn)

Se identifican picos significativos de reproducción durante la noche, especialmente a las 00:00, 23:00 y 20:00 horas, con más de 10 mil escuchas en cada franja. También se observa alta actividad a la 01:00 y 17:00 horas, lo que sugiere que la música se consume principalmente en horarios nocturnos y de ocio.

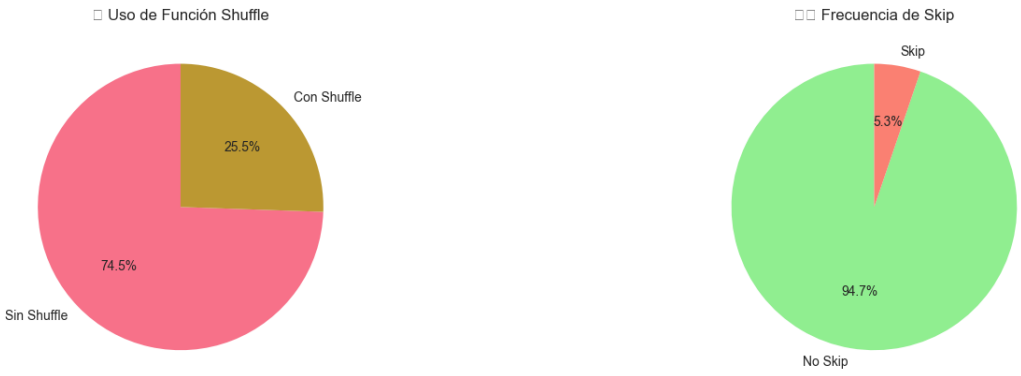
Ilustración 6A.2. Top 10 artistas por tiempo total de reproducción



(Fuente: elaboración propia con Python, pandas, matplotlib y seaborn)

El análisis revela que The Beatles lideran en número de reproducciones y tiempo total de escucha, seguidos por The Killers, John Mayer y Bob Dylan. Esto sugiere una preferencia marcada por el rock clásico y alternativo dentro de la muestra analizada.

Ilustración 7A.3. Uso de funciones shuffle y skip



(Fuente: elaboración propia con Python, pandas, matplotlib y seaborn)

El 74,5 % de las reproducciones se realizaron en modo aleatorio (*shuffle*), mientras que el 5,3 % fueron omitidas antes de finalizar (*skip*). El tiempo promedio de escucha fue mayor en las reproducciones secuenciales que en las aleatorias, lo que podría indicar una mayor atención en las primeras.

8. Anexo B. Fragmento del código Python utilizado

Este anexo presenta un fragmento del script desarrollado en Python para la limpieza, transformación y análisis del dataset de Spotify. Su objetivo es ilustrar el proceso de preparación de datos y la generación de métricas y visualizaciones que respaldan los resultados expuestos en el capítulo 4.

Ilustración 8Fragmento del script en Python para limpieza de datos

```python
import pandas as pd

def limpiar_datos(df):
    df['ts'] = pd.to_datetime(df['ts'])
    df['hora'] = df['ts'].dt.hour
    df['dia_semana'] = df['ts'].dt.day_name()
    df['minutos_reproducidos'] = df['ms_played'] / (1000 * 60)
    df = df.dropna(subset=['track_name', 'artist_name'])
    return df
```

(Fuente: Elaboración propia con Python, pandas, matplotlib y seaborn)

Este fragmento corresponde a la función utilizada para:

- Convertir la marca temporal () a formato fecha-hora.
- Extraer variables derivadas como la hora y el día de la semana.
- Calcular la duración de cada reproducción en minutos.
- Eliminar registros con valores nulos en campos clave.

El código completo, utilizado para generar las tablas y figuras del análisis, incluyó funciones adicionales para el agrupamiento de datos por franjas horarias, artistas, álbumes

y plataformas, así como para el cálculo de métricas relacionadas con las funciones shuffle y skip.

9. Anexo C. Aplicaciones Futuras del Análisis

A partir de los resultados obtenidos en este estudio, se identifican varias posibilidades de aplicación y líneas de investigación que podrían desarrollarse en el futuro:

1. **Optimización de algoritmos de recomendación musical**
   Utilizar los patrones de escucha identificados (horarios de mayor actividad, artistas preferidos, uso de funciones *shuffle* y *skip*) para mejorar las sugerencias personalizadas de canciones y listas de reproducción.

2. **Segmentación de audiencias**
   Implementar técnicas de *clustering* para agrupar usuarios con hábitos de consumo similares, lo que permitiría crear estrategias específicas de marketing y fidelización.

3. **Predicción de popularidad de canciones**
   Aplicar modelos predictivos que, a partir de métricas como reproducciones iniciales, duración media de escucha y recurrencia, estimen el potencial de éxito de nuevos lanzamientos.

4. **Estudios comparativos entre plataformas**
   Extender el análisis a datos de otros servicios de streaming musical (Apple Music, YouTube Music, Deezer) para evaluar diferencias de consumo según plataforma y demografía.

5. **Aplicaciones culturales y educativas**
   Usar los resultados para explorar el impacto cultural de ciertos géneros o artistas, así como para diseñar actividades pedagógicas sobre análisis de datos en contextos artísticos.

Este conjunto de posibilidades demuestra que el análisis de datos musicales no solo es relevante para la industria del entretenimiento, sino que también tiene un potencial significativo en investigación académica, marketing cultural y desarrollo de tecnología aplicada.

Cierre de los anexos

Los anexos presentados complementan el análisis realizado en el cuerpo del trabajo, ofreciendo evidencia visual, técnica y proyectiva que respalda los hallazgos obtenidos. El Anexo A proporciona representaciones gráficas que ilustran los patrones de consumo identificados; el Anexo B documenta el proceso técnico de limpieza y transformación de datos mediante código Python; y el Anexo C plantea aplicaciones futuras que amplían el alcance del estudio hacia escenarios de innovación, investigación y desarrollo.

En conjunto, estos recursos fortalecen la validez del análisis realizado y demuestran la aplicabilidad de las herramientas de Big Data en el estudio del comportamiento musical digital.