Universidad del Valle de Guatemala
Facultad de ingeniería

Data Science
Análisis Exploratorio - Proyecto
Sistema de recomendación de Spotify

Pablo Gonzalez
Javier Mombiela
Jose Hernandez
Jun Woo Lee
Andrés de la Roca

Guatemala, 2023

Índice

Índice 1
Situación Problemática 2
Problema Científico 2
Objetivos 2
Objetivo General 2
Objetivos Específicos 2
Descripción de los datos 3
Análisis Exploratorio 3
Hallazgos 16
Conclusiones 16
Referencias Bibliográficas 16

Situación Problemática

Las circunstancias que dieron lugar a la necesidad de desarrollar un sistema de recomendación preciso en Spotify se relaciona directamente con la creciente cantidad de música disponible en la plataforma y la demanda por diversidad de géneros que buscan los usuarios, por esto, los usuarios en repetidas ocasiones se encuentran con una abrumadora cantidad de opciones para escuchar música, por lo que se les dificulta el poder descubrir música nueva que realmente se acoplara a sus gustos musicales.

Este problema se complica aún más debido a que los gustos musicales, son muy subjetivos y varían significativamente de un individuo a otro, además, los usuarios a menudo no tienen tiempo para dedicarse a buscar manualmente entre canciones y playlists que canciones les gusta más.

Ante esta situación, una de las soluciones más viables es el utilizar técnicas de Machine Learning y Ciencia de los datos para poder crear un sistema de recomendación efectivo y que pueda recomendarle a un usuario cierta playlist o canciones que se adapten a sus gustos musicales particulares.

Se quiere desarrollar un sistema de recomendación que analice los gustos del usuario a través de sus playlists y ofrezca canciones recomendadas basadas ya sea en el título de la playlist o en las canciones que ya contenga.

Problema Científico

¿Cómo se pueden identificar patrones en los gustos musicales de usuarios en plataformas de streaming de música para ofrecerles mejores recomendaciones mediante Machine Learning?

Objetivos

Objetivo General

- Proponer un sistema de recomendación efectivo y preciso que le ofrezca recomendaciones valiosas a los usuarios de Spotify por medio de Machine Learning.

Objetivos Específicos

- Reconocer patrones en los gustos musicales que ayuden a predecir la relevancia de las recomendaciones que se le hagan a los usuarios
- Identificar qué variables representan mejor los gustos musicales de los usuarios.
- Relacionar de manera precisa las canciones con las características que buscan los usuarios en su música preferida.
- Desarrollar un modelo de manera iterativa para encontrar la mejor combinación de parámetros para el sistema de recomendación

Descripción de los datos

El conjunto de datos consiste en aproximadamente 10,000 playlists públicas de Spotify que han sido creadas entre enero 2010 y noviembre 2017. Cada playlist contiene datos como; el nombre de la playlist y la lista de canciones, la lista de canciones contiene información específica referente a cada una de las canciones, a continuación se profundizará más acerca de qué significa cada uno de los campos del conjunto de datos.

Análisis Exploratorio

Tareas de limpieza y procesamiento de datos:

En el código presentado lleva a cabo una serie de tareas críticas de limpieza y preprocesamiento de datos en un archivo JSON llamado 'challenge_set.json'. En primer lugar, se inicia la limpieza al cargar los datos del archivo JSON en una estructura de datos manejable mediante la función 'json.load()'.Luego a partir de esto , se crea un DataFrame de Pandas ('df') a partir de las listas de playlists contenidas en el diccionario 'data'. Esta acción transforma los datos en un formato tabular que simplifica su manipulación y análisis.

La etapa de preprocesamiento se intensifica al iterar a través de cada fila del DataFrame 'df' utilizando un bucle 'for'. Durante esta iteración, se extraen datos relevantes de cada track, incluyendo su posición, nombre del artista, URI del track, URI del artista, nombre del track, URI del álbum, duración en milisegundos y nombre del álbum. Estos detalles se almacenan cuidadosamente en listas de Python, lo que facilita su posterior utilización en análisis posteriores.

Uno de los aspectos cruciales del proceso de limpieza es la identificación y gestión de listas de tracks vacías. Para lograrlo, se crea una lista llamada 'empty_track_indices' donde se registran los índices de las playlists que carecen de tracks. Esta información se utiliza para calcular y presentar el número de listas de tracks vacías en el conjunto de datos.

Finalmente, se lleva a cabo una acción crítica de limpieza al eliminar las filas correspondientes a las playlists con listas de tracks vacías del DataFrame 'df'. Este paso asegura que los datos estén libres de inconsistencias y se mantengan únicamente las playlists con información válida. En conjunto, estas tareas de limpieza y preprocesamiento son esenciales para garantizar que los datos estén en un estado óptimo para análisis posteriores, permitiendo un estudio más eficiente y preciso de la información contenida en 'challenge_set.json'.

Descripción de variables y sus tipos:

El conjunto de datos analizado contiene un total de 10000 observaciones de las cuales se reducen a 9000 debido a que se eliminan 1000 por la limpieza de listas vacías se cuenta con un total de 6 variables globales y dentro de la variable de track se cuenta con 8 variables mas lo que daría un total de 14 variables . Cada observación representa un caso único en el

conjunto de datos, mientras que las variables representan distintos atributos o características asociados a cada observación.

Tabla de variables

Nombre	Descripción	Tipo
Pid	Identificador de la playlist	Categórica nominal
Name	Nombre de la playlist, para algunas playlists este campo es omitido.	Categórica nominal
num_holdouts	Número de canciones que han sido omitidas para esta playlist.	Cuantitativa discreta
tracks	Una colección que contiene las canciones de cada playlist	Cuantitativa discreta
pos	Posición de la canción en la playlist.	Cuantitativa discreta
track_name	Nombre de la canción.	Categórica nominal
track_uri	El URI de la canción en Spotify.	Categórica nominal
artist_name	Nombre del álbum de la canción.	Categórica nominal
artist_uri	El URI del artista en Spotify.	Categórica nominal
album_name	Nombre del álbum en donde esta la música.	Categórica nominal
album_uri	El URI del álbum en Spotify.	Categórica nominal
duration_ms	La duración de la canción en milisegundos.	Cuantitativa continua
num_samples	Número de canciones en total incluidas en la playlist	Cuantitativa discreta

num_tracks	Número de canciones en total de la playlist, incluyendo num_holdouts.	Cuantitativa discreta

Esta descripción inicial de las variables y observaciones proporciona una visión general de la estructura del conjunto de datos y sienta las bases para un análisis más detallado y específico de sus contenidos.

Resumen de variables categóricas y numéricas:

Resumen de variable numéricas:

	count	mean	std	min	25%	50%	75%	max
num_holdouts	9000.0	74.506889	42.088348	5.0	43.0	67.0	94.0	225.0
num_tracks	9000.0	105.729111	65.049312	10.0	51.0	87.0	163.0	250.0
num_samples	9000.0	31.222222	37.607579	1.0	5.0	10.0	25.0	100.0
duration_ms	281000.0	232601.692505	63151.588982	0.0	200026.0	224574.0	256013.0	9158194.0

Se pueden obtener interesantes detalles sobre los patrones de uso y las preferencias de los usuarios analizando las listas de reproducción. Con un total de 9,000 listas de reproducción en el conjunto de datos, la cantidad promedio de "holdouts" por lista es de 74.51, aunque hay una gran variación debido a que algunas listas contienen hasta 225 "holdouts". Las listas de reproducción son muy personalizadas, como lo demuestra su amplia variedad, y los usuarios pueden preferir mantener ciertas canciones en espera durante más tiempo que otras.

Además, cada lista de reproducción generalmente contiene aproximadamente 105.73 pistas, pero hay algunas que tienen hasta 250 pistas. Esto muestra la diversidad de gustos y el nivel de compromiso de los usuarios con su música; algunos pueden preferir listas más cortas y específicas, mientras que otros pueden preferir listas más extensas que probablemente escucharán durante más tiempo.

La cantidad de "samples" que hay en cada lista también indica una gran dispersión. Aunque el promedio es de 31.22 "samples", la mediana es solo de 10, lo que indica que la mayoría de las listas suelen tener un número menor de "samples". Sin embargo, algunas listas, probablemente las más populares o destacadas, acumulan un número significativamente más alto.

Finalmente, la duración promedio de las pistas es de alrededor de 3.88 minutos, lo que es una duración típica de muchas canciones populares. Sin embargo, es notable que haya pistas que duren hasta 152.64 minutos, que podrían ser compilaciones o mezclas. Sin embargo, la presencia de pistas con una duración de 0 milisegundos puede indicar errores de datos o pistas especiales sin sonido.

Tablas de frecuencias de variables categóricas

Playlist Name	Frequency	Artist Name	Frequency	Track Name	Frequency	Album Name	Frequency
country	148	drake	4877	closer	335	views	1126
rap	97	kanye west	2592	roses	241	coloring book	806
oldies	75	kendrick lamar	1902	ride	229	stoney	762
workout	74	rihanna	1734	broccoli (feat. lil yachty)	226	more life	743
rock	72	the weeknd	1644	ignition - remix	222	the life of pablo	713
throwback	59	eminem	1494	gold digger	215	beauty behind the madness	701
throwbacks	54	luke bryan	1486	forever	213	greatest hits	660
party	54	j. cole	1472	no role modelz	211	2014 forest hills drive	657
jams	43	chris brown	1418	home	203	original album classics	656
classic rock	40	future	1308	let me love you	203	good kid, m.a.a.d city	632
classics	37	ed sheeran	1265	humble.	202	damn.	609
road trip	35	florida georgia line	1261	jumpman	199	take care	604
disney	32	beyoncé	1243	one dance	199	blurryface	598
music	31	justin bieber	1126	no problem (feat. lil wayne & 2 chainz)	196	montevallo	596
lit	30	kenny chesney	1115	alright	193	culture	595
pop	29	big sean	1095	t-shirt	193	nothing was the same	555
christmas	29	twenty one pilots	1086	caroline	188	starboy	541
car	29	the chainsmokers	1082	bad and boujee (feat. lil uzi vert)	186	if you're reading this it's too late	540
chill	29	zac brown band	1069	congratulations	186	crash my party	536
tbt	28	jay z	1041	stay	185	american teen	531

Con 148 menciones, el género "country" se destaca en las listas de reproducción más populares. Los géneros "rap" y "oldies" también son notables, ocupando el segundo y tercer lugar respectivamente con 97 y 75 menciones. Según estos datos, algunos géneros musicales tienen una resonancia más fuerte entre los usuarios. Además, es interesante notar que nombres como "throwback", "party" y "workout" aparecen en la lista, lo que indica que las listas de reproducción no solo se basan en géneros musicales, sino también en momentos o actividades específicas.

Con una impresionante cantidad de 4,877 menciones, Drake es claramente el artista más mencionado de la lista, superando por mucho an artistas más conocidos como Kanye West y Kendrick Lamar. Esta éxito de Drake refleja su gran popularidad y su influencia en la industria musical en ese momento.

Además, las canciones más populares brindan información útil. "Closer" está en la cima, seguida de temas como "rosas" y "bailar". A pesar de la diversidad de género, estas elecciones parecen haber captado la atención de los oyentes.

Finalmente, en términos de álbumes, las "vistas" de Drake son las más destacadas, lo que aumenta aún más la popularidad y el impacto del artista. "Coloring book" y "stoney" le siguen en la lista, lo que demuestra la diversidad en las preferencias de los usuarios por los álbumes.

Cruzar Variables:
Proceso de Creación y Análisis de la Matriz de Correlación:

Para comenzar, se realizó una copia del DataFrame original llamada 'df2'. Luego, se abordó la necesidad de calcular la duración total de las canciones representada en milisegundos

('duracion_cancion_ms') contenidas en la columna 'tracks'. Esto se logró mediante el uso de una función lambda aplicada a cada lista de canciones en la columna 'tracks', extrayendo y sumando las duraciones individuales de las canciones. A continuación, se seleccionaron las variables numéricas de interés: 'num_holdouts', 'num_tracks', 'num_samples' y 'duracion_cancion_ms'. Finalmente, se calculó la matriz de correlación entre estas variables utilizando Pandas y se visualizó mediante un mapa de calor con Seaborn.

Resultados de la Matriz de Correlación:

La matriz de correlación revela relaciones significativas entre las variables seleccionadas. En primer lugar, se observa una fuerte correlación positiva (0.84) entre 'num_holdouts' y 'num_tracks', indicando que las listas de reproducción con más 'holdouts' suelen tener un mayor número de pistas. Esto sugiere una posible conexión entre la diversidad de listas y la cantidad de pistas incluidas.

Además, se encuentra una correlación positiva (0.79) entre 'num_tracks' y 'num_samples', lo que implica que las listas con más pistas tienden a contener más muestras. Esto podría estar relacionado con la riqueza musical de las listas y la cantidad de datos de audio disponibles.

El hallazgo más destacado es la correlación casi perfecta (0.99) entre 'num_samples' y 'duracion_cancion_ms', lo que sugiere que la duración de las canciones está altamente relacionada con la cantidad de muestras utilizadas para representarlas. Esto podría tener implicaciones en la calidad y el detalle de las representaciones de audio utilizadas en el conjunto de datos.

En resumen, la matriz de correlación proporciona una visión valiosa de las relaciones entre estas variables, lo que puede ser fundamental para comprender la estructura de nuestro conjunto de datos y guiar futuros análisis relacionados con la duración de las canciones, la composición de listas de reproducción y la representación de datos de audio. Estos hallazgos pueden ser útiles para investigaciones más profundas y toma de decisiones informadas en el contexto de la música y el análisis de datos de audio.

Test de anova:
Los resultados del análisis de varianza (ANOVA) para las variables numéricas en relación con la variable categórica 'name' arrojan hallazgos interesantes. Para la variable 'num_holdouts', se observa una diferencia significativa entre los grupos definidos por 'name', con un valor p extremadamente bajo ($p < 0.001$), lo que sugiere que la variable 'name' tiene un efecto significativo en 'num_holdouts'.

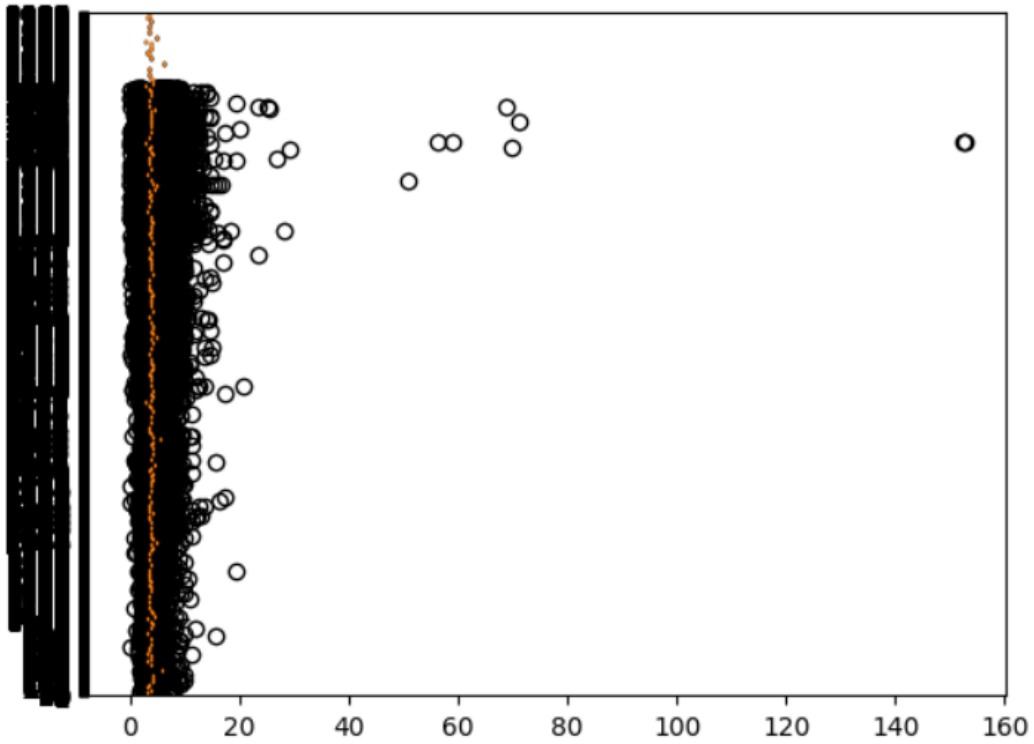
Lo mismo se aplica a 'num_tracks' y 'num_samples', donde los valores p son igualmente muy bajos ($p < 0.001$). Esto indica que la variable 'name' también tiene un impacto significativo en estas dos variables numéricas, lo que podría reflejar diferencias importantes en la cantidad de pistas y muestras en función de las categorías de 'name'.

El resultado más destacado es para 'duracion_cancion_ms', donde nuevamente se observa un valor p extremadamente bajo ($p < 0.001$). Esto sugiere que la variable 'name' tiene un impacto altamente significativo en la duración de las canciones en milisegundos. En otras palabras, diferentes categorías de 'name' están asociadas con diferencias significativas en la duración de las canciones.

En resumen, estos resultados del ANOVA indican claramente que la variable categórica 'name' tiene un efecto significativo en todas las variables numéricas evaluadas ('num_holdouts', 'num_tracks', 'num_samples' y 'duracion_cancion_ms'). Esto puede ser fundamental para comprender cómo las categorías de 'name' influyen en las características numéricas de los datos y puede guiar análisis posteriores y la toma de decisiones relacionadas con estas variables.

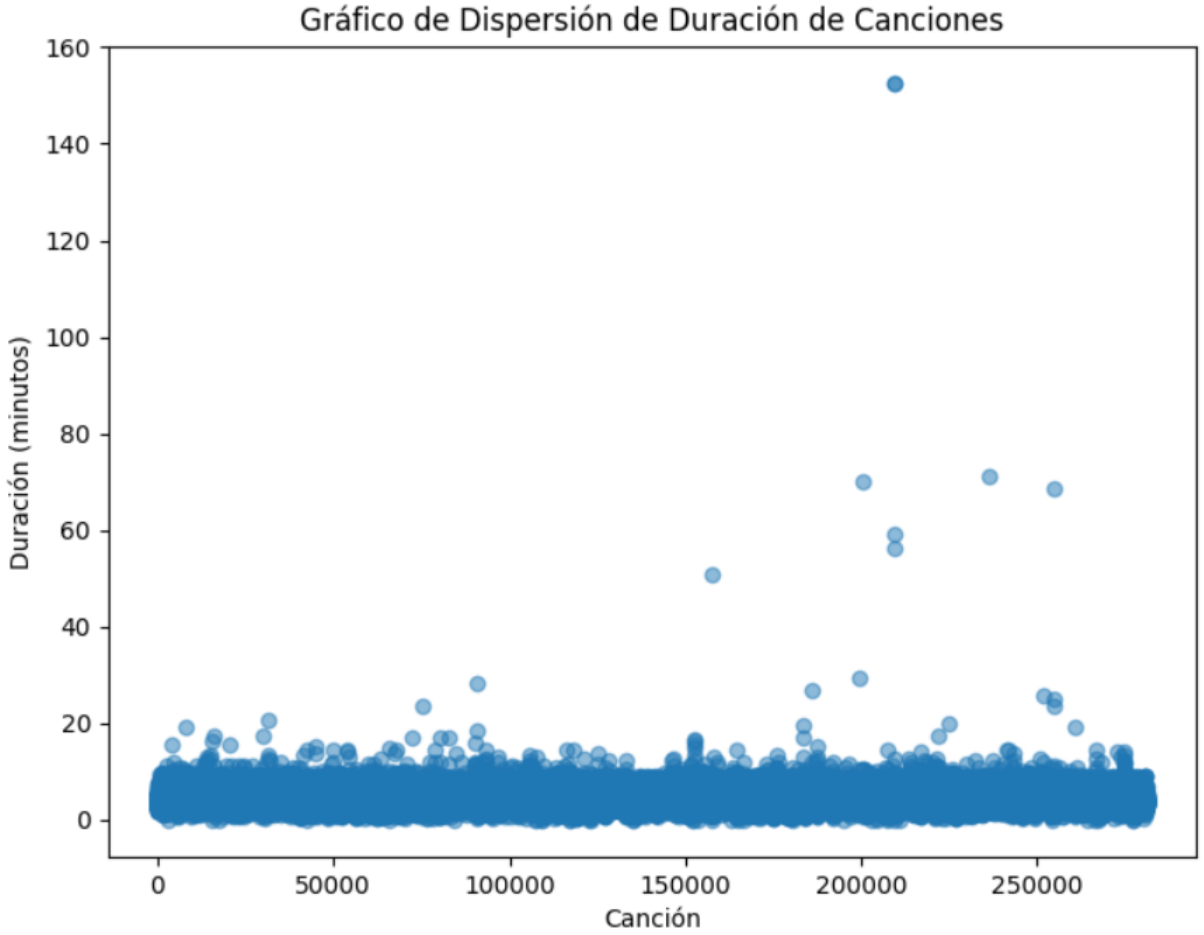
Gráficos exploratorios de la data:

- Gráfica 1: Diagrama de caja y bigotes para la variable duration_ms



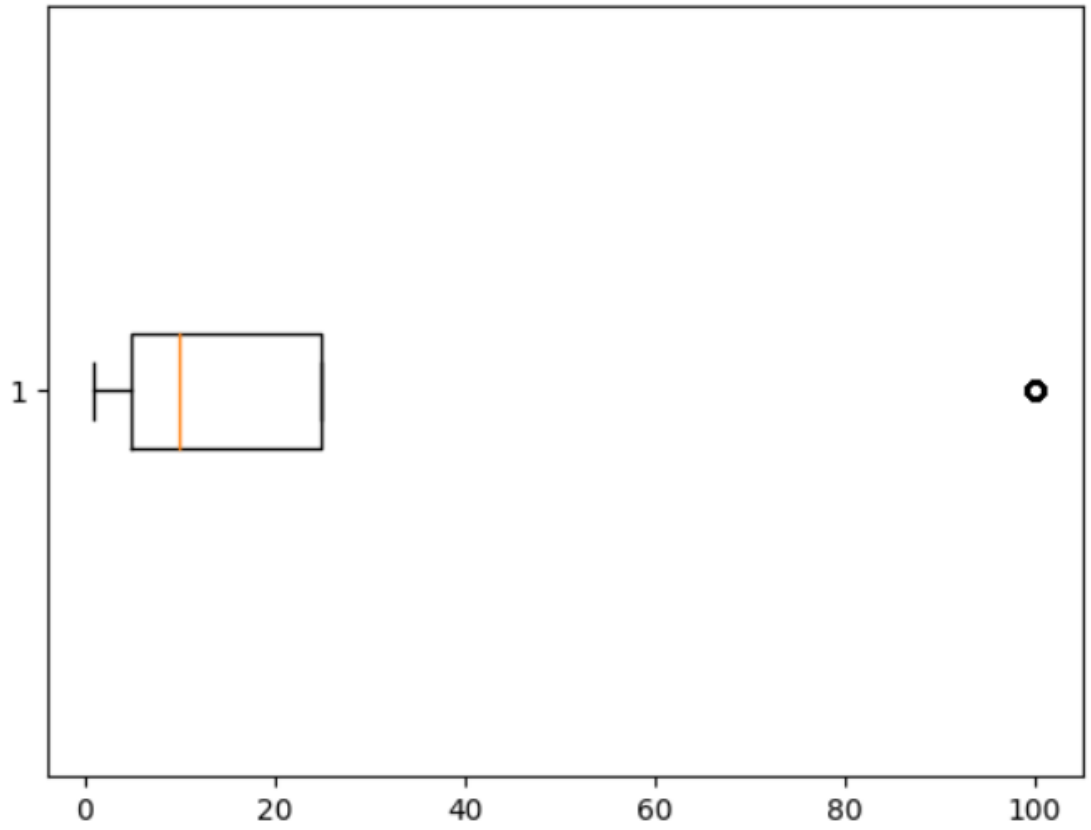
En el diagrama de caja y bigotes de la variable 'duration_ms', se observa que la mediana (Q2) se encuentra en el valor de 0.75. Esto significa que el 50% de las observaciones en esta variable tienen una duración de canción igual o menor a 0.75, Por otro lado, el Cuartil 3 (Q3) se encuentra en el valor de 1.25. Esto indica que el tercer cuartil de los datos (75% de las observaciones) tiene una duración de canción igual o menor a 1.25 unidades. También se puede observar que en la gráfica existen puntos atípicos que están entre los 40,60, 80 y 150 minutos lo cual no es un tiempo lógico que dure una canción.

- Gráfica 2: Dispersión de duración de canciones



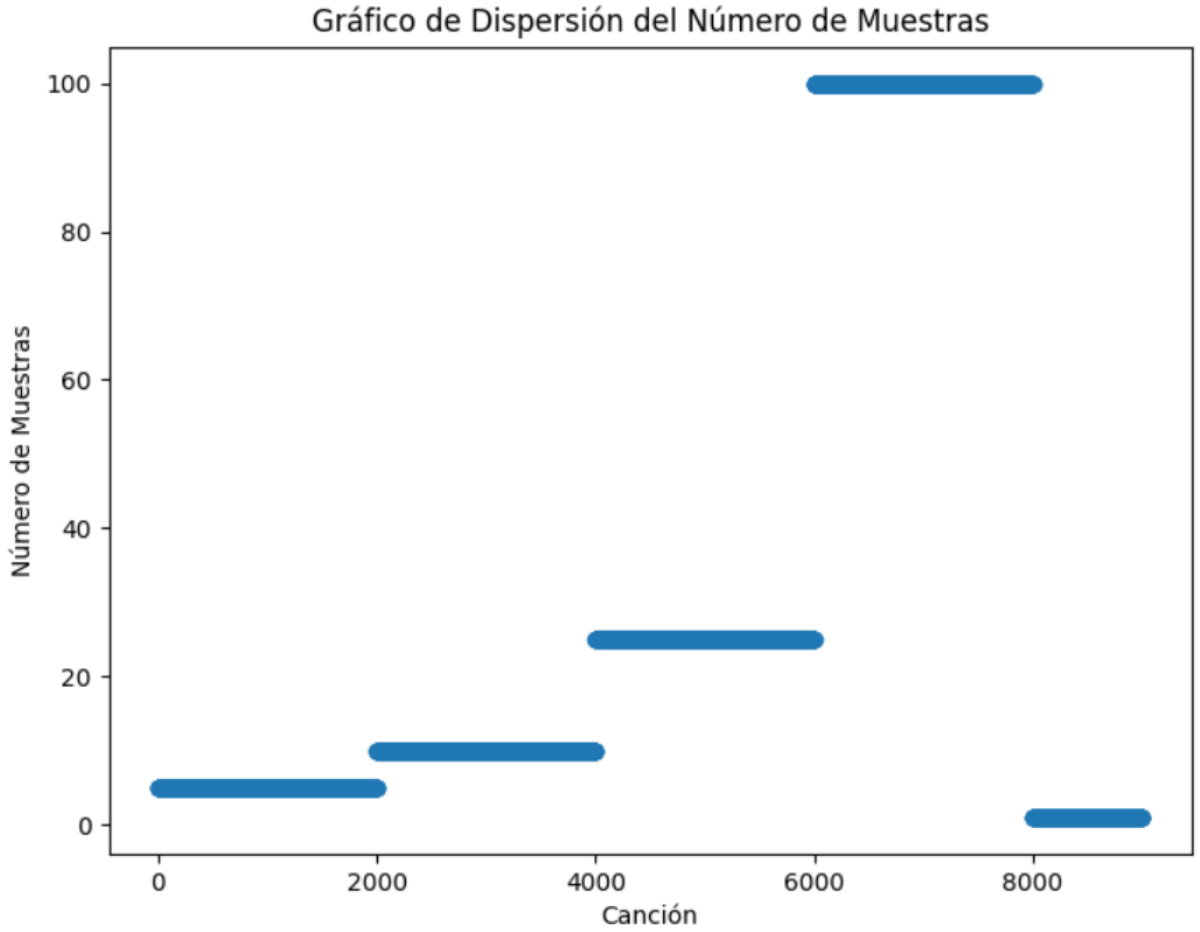
En el diagrama de dispersión para la variable ‘duration_ms’ se puede visualizar que la gran mayoría de cancione se concentran en un área de 0 a 15 minutos mientras que también existen puntos que salen de esta área en donde se pueden visualizar los puntos ya antes descritos en el gráfico de caja y bigotes los cuales puntos 40,60, 80 y 150.

- Gráfica 3: Diagrama de caja y bigotes para num_samples



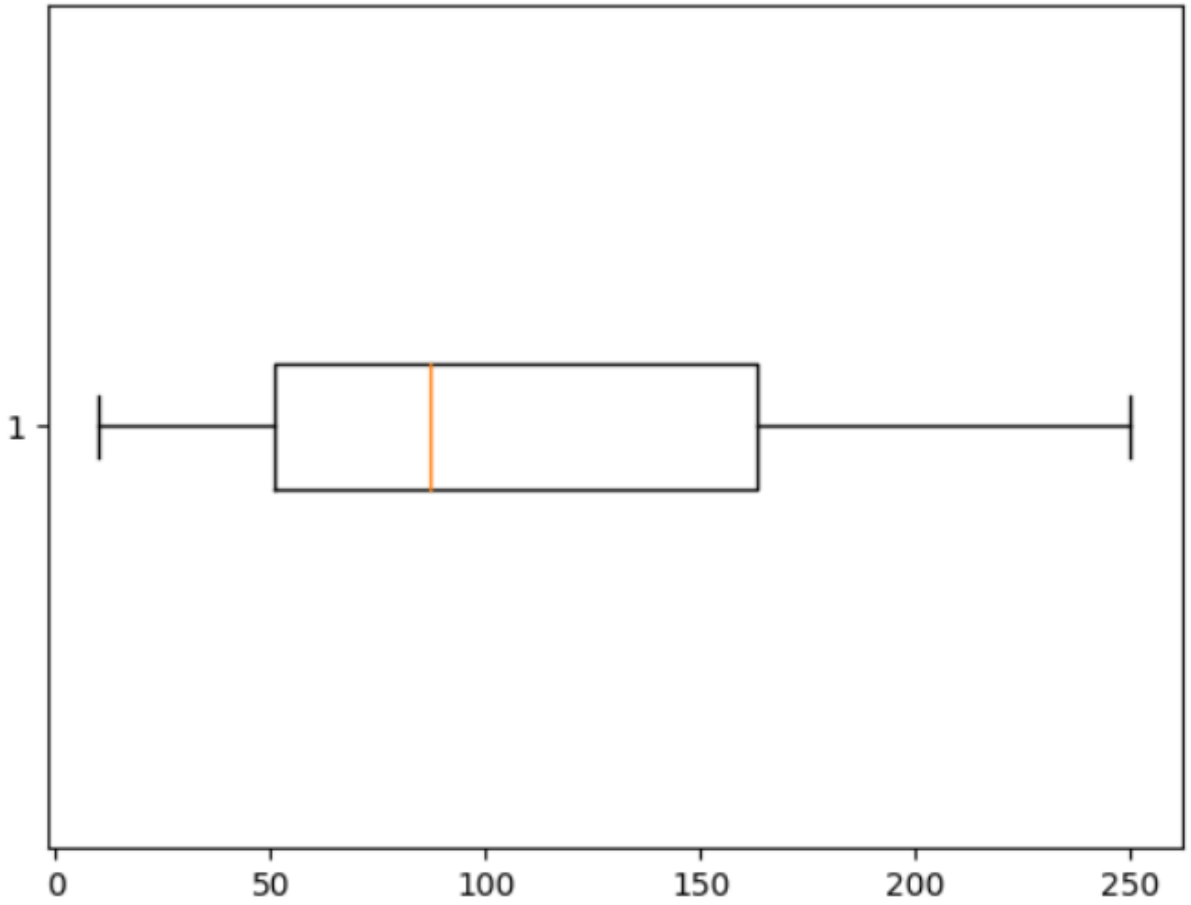
En el diagrama de caja y bigotes de la variable 'num_samples', se observa que la mediana (Q2) se encuentra en el valor de 0.925. Esto significa que el 50% de las observaciones en esta variable tienen una duración de canción igual o menor a 0.925. Por otro lado, el Cuartil 3 (Q3) se encuentra en el valor de 1.075. Esto indica que el tercer cuartil de los datos (75% de las observaciones) tiene una duración de canción igual o menor a 1.075 unidades. También se puede observar que en la gráfica existen puntos atípicos que están en el valor de 100, lo cual no es un tiempo lógico que dure una canción.

- Gráfica 4: Diagrama de dispersión para num_samples



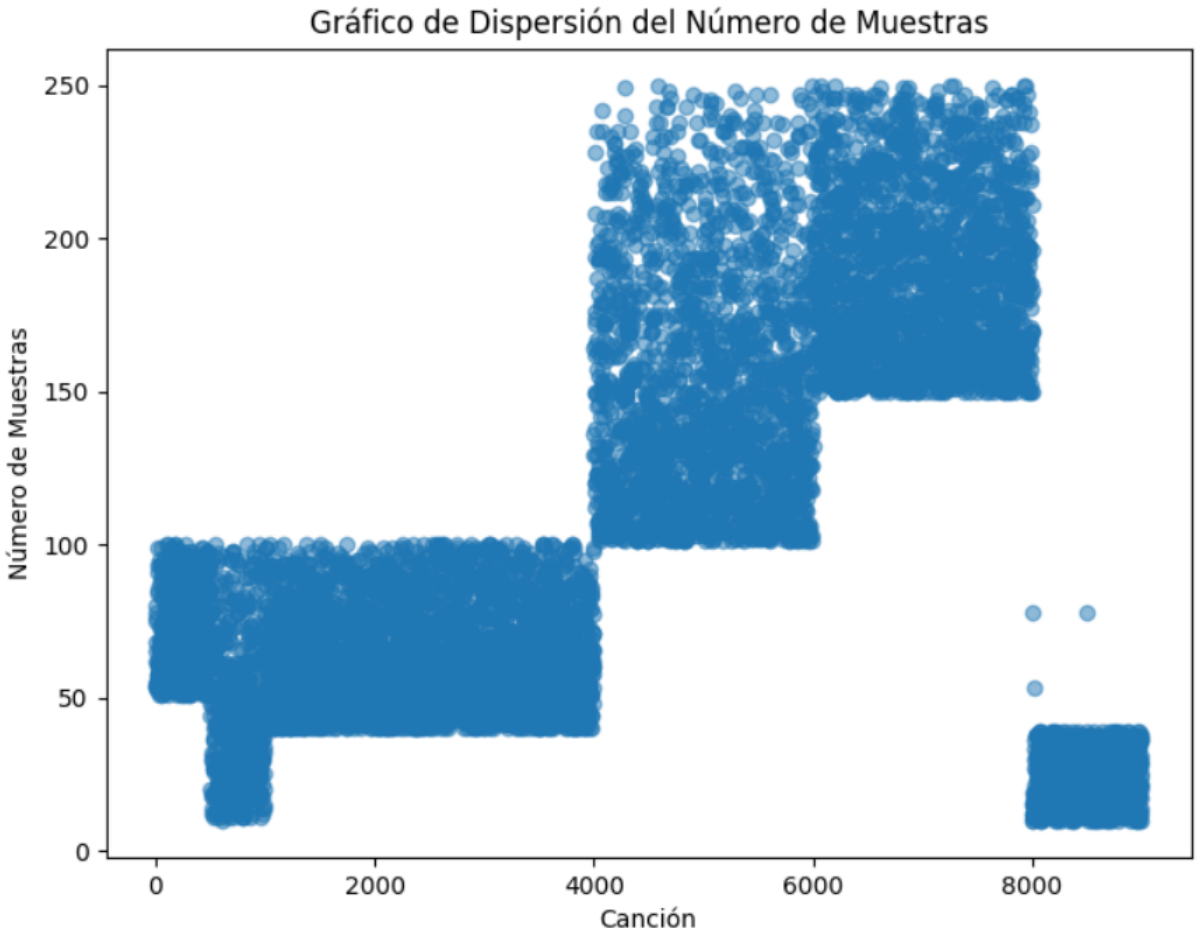
En el diagrama de dispersión para la variable “Num_samples” se puede observar que de la playlist 0 a la 2000 se cuentan con alrededor de 8 canciones, de la playlist 2001 a la 4000 se puede observar se cuenta con alrededor de de 12 canciones mientras que para la playlist 4001 a la 6000 se puede observar que las playlist contienen alrededor de 28 canciones, con la cantidad de más alta de canciones tenemos al grupo de 6001 a 8000 la cual contiene alrededor de de 100 canciones mientras que por último de la 8001 a la 9000 se cuenta con la cantidad más baja de canciones de alrededor de 3 canciones.

- Gráfica 5: Diagrama de caja y bigotes para num_tracks



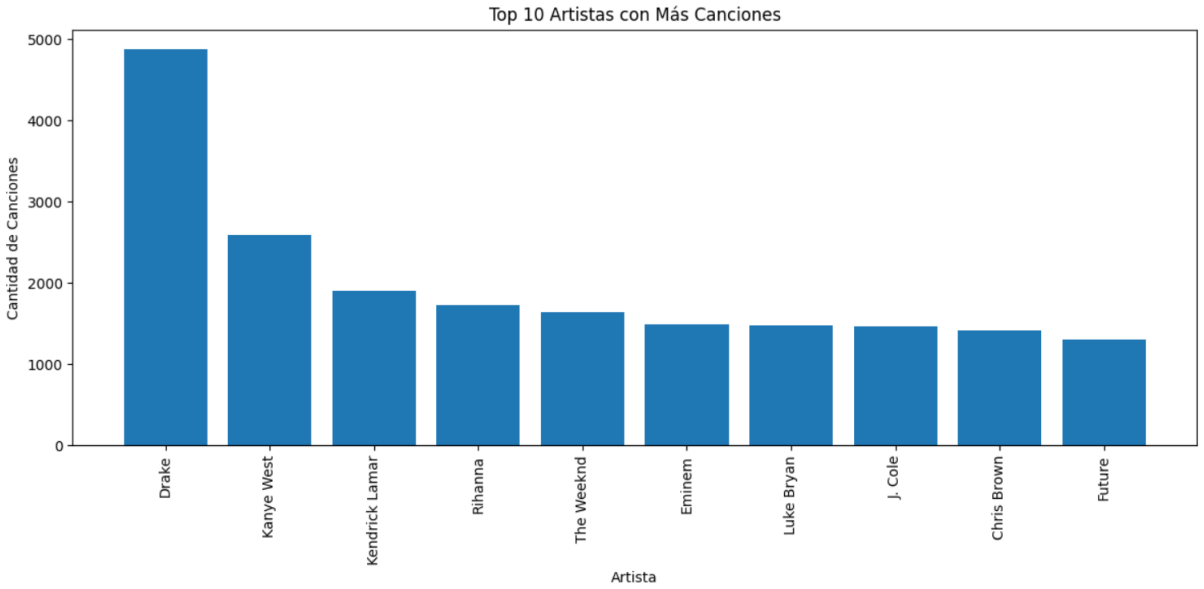
En el gráfico de caja y bigotes para la variable 'num_tracks', se destaca que la mediana (Q2) se sitúa en 0.925. Esto señala que el 50% de las observaciones en esta variable presentan un número de pistas igual o menor a 0.925. Por otra parte, el Cuartil 3 (Q3) se encuentra en 1.075, indicando que el tercer cuartil de los datos (el 75% de las observaciones) posee un número de pistas igual o menor a 1.075. Sin embargo, es importante notar que en el gráfico no se logró identificar ningún punto atípico.

- Gráfica 6: Diagrama de dispersión num_tracks



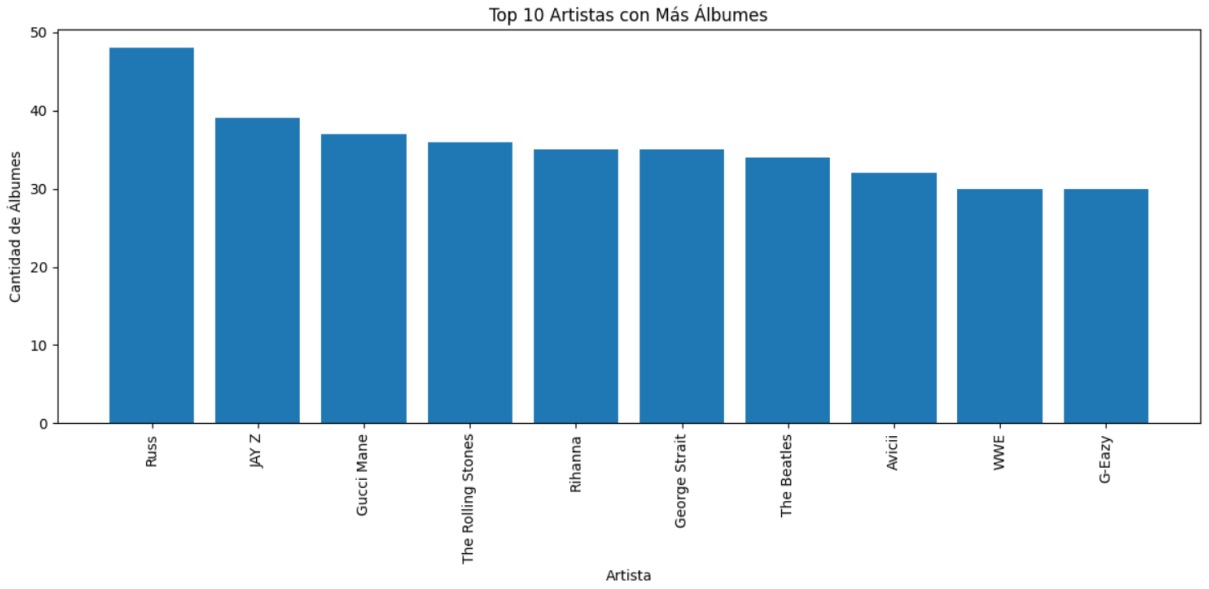
En el diagrama de dispersión para la variable ‘num_tracks’ se puede visualizar que a partir de la playlist 1 hasta la 400 se puede visualizar que la cantidad de las canciones en las playlist está dentro de un rango de 40 canciones a 100 canciones mientras que de la playlist 4000 a la 6000 estas cuentan con alrededor de 100 a 150 canciones, con respecto a la playlist 6001 a la 8000 se puede ver que las playlist cuentan con 150 a 250 canciones y por último de la 8000 a la 9000 cuenta con alrededor de de 10 canciones a 50 canciones.

- Gráfica 7: Top 10 artistas con más canciones



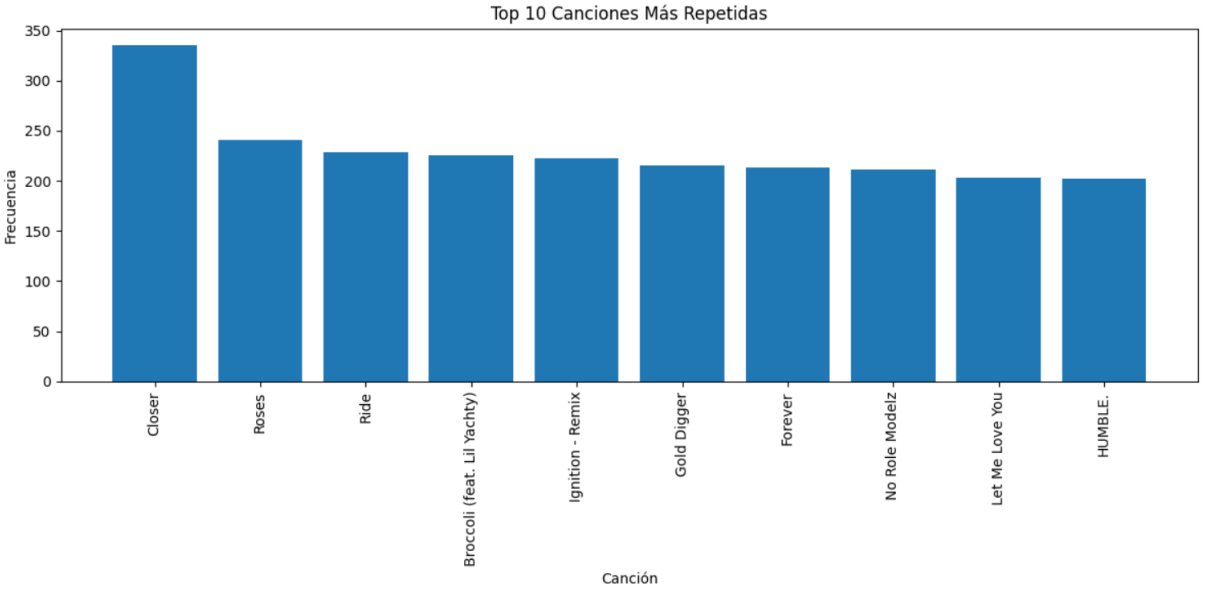
Esta gráfica de barras presenta a los 10 artistas con la mayor cantidad de canciones en el conjunto de datos. Se puede observar que "Drake" es el artista con más canciones con un total de 4,877 canciones. Le siguen "Kanye West", "Kendrick Lamar", "Rihanna" y "The Weeknd" en el top 5. Se puede observar también que el promedio de canciones entre el top 10 artistas con más canciones es de 1,892.

- Gráfica 8: Top 10 artistas con más álbumes



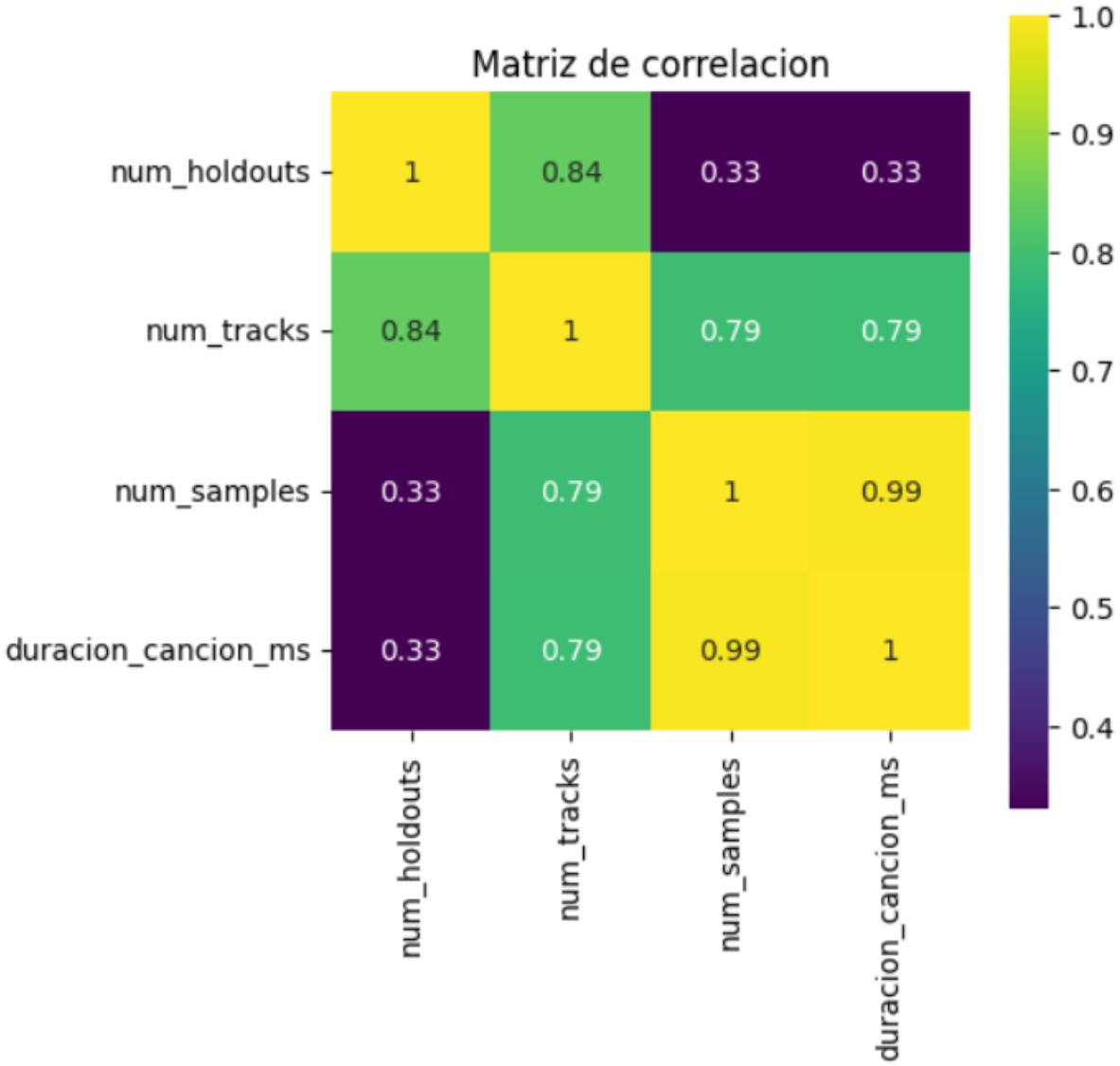
En esta gráfica se pueden observar a los 10 artistas con el mayor número lanzamientos de álbumes, liderados por 'Russ' con 48 álbumes. Le sigue 'JAY Z' con 39 álbumes y 'Gucci Mane' con 37. En promedio, estos artistas principales han lanzado aproximadamente 36 álbumes cada uno, lo que significa que la diferencia entre cada artista no es mucha.

- Gráfica 9: Top 10 canciones más repetidas



Esta gráfica de barras muestra las 10 canciones más repetidas. Se puede observar que la canción 'Closer' se destaca como la más repetida en el conjunto de datos, con un total de 335 repeticiones. A esta le siguen ‘Roses’ con 241 reproducciones y ‘Ride’ con 229 reproducciones. En promedio, estas canciones tienen alrededor de 222 menciones, lo que destaca la popularidad significativa de las canciones en el análisis.

- Gráfica 10: Matriz de correlación



Para la gráfica de correlación en primer lugar, se destaca una fuerte correlación positiva entre 'num_holdouts' y 'num_tracks', con un valor de correlación de aproximadamente 0.84. Esto sugiere que, en general, a medida que aumenta el número de 'holdouts' en las listas de reproducción, también tiende a aumentar el número de pistas en esas listas, lo que puede indicar una relación entre la diversidad de las listas y su longitud. Además, se encuentra una correlación positiva robusta entre 'num_tracks' y 'num_samples' con un valor cercano a 0.79. Esto sugiere que las listas de reproducción con un mayor número de pistas también tienden a contener más muestras, lo que podría estar relacionado con la riqueza musical de las listas y la cantidad de datos de audio recopilados. El resultado más llamativo es la correlación casi perfecta (cercana a 1) entre 'num_samples' y 'duracion_cancion_ms'. Esto sugiere que la duración de las canciones en milisegundos está altamente relacionada con la cantidad de muestras utilizadas para representarlas.

Hallazgos

Este análisis proporcionó información valioso sobre las relaciones entre variables numéricas, la influencia de la variable categórica “name” sobre algunas de las variables numéricas y estadísticas descriptivas de las variables clave, lo que va a ser fundamental para el desarrollo del modelo y a tomar decisiones importantes sobre la arquitectura a utilizar.

Uno de los grandes hallazgos de este análisis fue la correlación entre variables, se logró identificar una correlación positivas entre ‘num_holdouts’ y ‘num_tracks’, lo que sugiere que las que las listas de reproducción con mayor cantidad de ‘holdouts’ tienden a tener un mayor número de canciones, lo cual indica una relación entre la diversidad de las listas y su longitud.
Adicionalmente, se identificaron correlaciones entre la duración de las canciones y la cantidad de muestras para cada lista, lo que demuestra un factor importante a tomar en cuenta al momento de optimizar el modelo.

En el test ANOVA, se reveló que la variable ‘name’ tenía un efecto significativo sobre las variables numéricas que representan el número de muestras y la duración de cada canción, lo cual la hace una variable a tomar en consideración para realizar las futuras predicciones con el modelo.

Se observó la distribución de las diferentes variables numéricas para poder tener una mejor visión de cómo se distribuye el dataset, además, se identificaron los 10 artistas con mayor cantidad de canciones, los 10 artistas con mayor número de álbumes y las 10 canciones más repetidas dentro de las listas del conjunto, lo cual nos permite conocer la influencia que tienen ciertos artistas y canciones individuales en los gustos musicales de cada personas y da un vistazo a cómo están distribuidas las listas dentro del conjunto de datos.

Conclusiones

- La popularidad de los artistas puede llegar a tener una correlación importante al momento de realizar predicciones de acorde a los gustos de cada persona dados ciertos parámetros.
- La cantidad de canciones en cada playlist son un factor importante al momento de identificar patrones para poder realizar recomendaciones.
- La variable ‘name’ influye de manera significativamente en todas las variables numéricas evaluadas en el test de ANOVA.
- Las relaciones encontradas para la cantidad de canciones y duración de las canciones pueden llegar tener una correlación con el género o gusto del usuario.

Referencias Bibliográficas

Spotify, AICrowd (2020) Spotify Million Playlist Dataset Challenge. Extraído de AICrowd.

C.W, Chen, P. Lamere, M. Schedl, H. Zamani. (2018) Recsys Challenge 2018: Automatic Music Playlist Continuation, in Proceedings of the 12th ACM Conference on Recommender Systems.

C.W, Chen, P. Lamere, M. Schedl, H. Zamani. (2018) An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation.