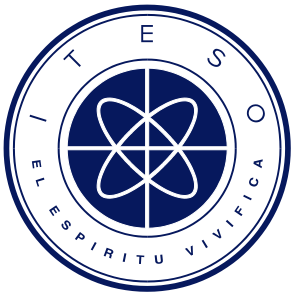
Instituto Tecnológico y de Estudios Superiores de Occidente

PROGRAMACIÓN PARA ANÁLISIS DE DATOS



ANÁLISIS MUSICAL (DE BASE DE DATOS DE SPOTIFY)

Presentan

Ángel Ramírez Carrillo 739611

Alejandro Samuel Romero Mora 741745

Luis Raúl Acosta Mendoza 739199

José Juan Díaz Campos Díaz Campos 740313

Profesor: Mtra. Gisel Hernández Chávez

Fecha 29/11/2023

Elaborado por: Mtra. Gisel Hernández Chávez

Revisado por: Mtra. Gisel Hernández Chávez

Aprobado por: Mtra. Gisel Hernández Chávez

Contenido

Índice de Tablas

Lista de Figuras

Introducción

1.1 Enunciado del problema y preguntas de investigación

Gracias a los avances tecnológicos, existen enormes bases de datos que calculan (ya sea con algoritmos, modelos de inteligencia artificial) metadatos sobre la música que los usuarios de servicios de streaming consumen. El ejemplo en cuestión corresponde a Spotify, quien tiene disponible esta información para todos público.

Asimismo, esta información se ordena, clasifica y submuestrea en páginas web especializadas en compartir muestras de datos, como lo es Kaggle. Teniendo una considerable cantidad de información por canciones, clasificadas por género, hace que la investigación y análisis de la muestra se preste para revisar el comportamiento y características de las canciones de forma general o específica, dependiendo de las categorías en las que se desee filtrar.

El dataset que tomamos nosotros, fue de uno considerablemente grande en comparación a otros proyectos (alrededor de 250,000 observaciones) y al alcance de este proyecto, donde cada observación compone a una canción. Una canción puede estar en hasta dos géneros, pero no dos veces en el mismo género, por lo tanto, la **unidad de observación** es la combinación del track_id y de género.

Cada canción cuenta con información para identificarla:

- Su ID de spotify (track ID)
- El nombre del track (track name)
- Artist (artist)
- Género (genre)

Información objetiva de los archivos:

- Duración (duration_ms)
- Rango dinámico (loudness)

Información objetiva de las canciones, aproximada por un algoritmo:

- Modo (Mode)
- Tonalidad (Key)
- Pulsos por Minuto (BPM)
- Compás (Time signature)

E información calculada por algoritmos de Spotify, en base a los archivos o su comportamiento en la plataforma:

- Speechiness
- Danceability
- Valence
- Popularity
- Liveness
- Acousticness

- Instrumentalness
- Liveness

Si bien, nuestras preguntas de investigación fueron variando mucho conforme el análisis se fue desarrollando, al final nos quedamos con las siguientes incógnitas:

1. ¿Cuál es la distribución de frecuencias de los géneros musicales en la muestra de datos?
    a. Unidad de análisis: genre
2. ¿Hay una relación entre la duración de una canción y su capacidad de baile?
    a. Unidad de análisis: track, es decir, cada fila.
3. ¿Cómo varía la característica “speechiness” de las canciones según su género?
    a. Unidad de análisis: track
4. ¿Existe alguna relación entre la popularidad de una canción y el género pop?
    a. Unidad de análisis: track
5. ¿Existe una relación lineal entre la valencia y la danceability de las canciones del género rock?
    a. Unidad de análisis: tracks del género rock.

Dado que el dataset sólo contiene información sobre características específicas de canciones (género, nombre del artista, características de audio, etc.) y no existe una dimensión temporal en la que se siga el cambio a lo largo del tiempo para las mismas canciones, podemos concluir que se trata de un estudio transversal.

1.2 Justificación

El estudio nace de la necesidad de querer conocer qué tan acertadas son los algoritmos de Spotify para obtener información de las canciones, básicamente, haciendo hipótesis sobre cómo se comporta la información en base a estereotipos (la música de rap es más hablada que las otras, la música de pop es popular).

Además, este es un tema para el cual independientemente de nuestras preferencias, todos podremos estar medianamente interesados en descubrir más acerca de, porque al final, somos muchos los individuos que consumimos música de esta manera, en streaming, y todavía más específicamente desde spotify.

1.3 Objetivos

1. Aplicar conocimientos de programación para análisis de datos en Python empleando bibliotecas como numpy, pandas, matplotlib, seaborn, statsmodels y scikit learn.
2. Recopilar datos y estructurar el data set que permita responder a las preguntas de investigación, limpiando las observaciones incorrectas y transformándolos para facilitar su interpretación gráfica.
3. Aplicar conocimientos de Probabilidades y Estadísticas durante la etapa de exploración y preparación de datos, fundamentalmente las relativas a estadística descriptiva, formulación y pruebas de hipótesis, así como análisis de regresión con fines explicativos, en el contexto de nuestra información.
4. Visualizar datos durante la exploración de estos, y hacer posibles hipótesis para responderlas, o proponerlas al siguiente equipo que decida retomar este proyecto.
5. Documentar los resultados del proceso de análisis de datos exploratorio, con énfasis en la redacción de hallazgos y conclusiones
6. Seguir procesos y etapas de una metodología de Minería de Datos, fundamentalmente en sus fases iniciales de comprensión del negocio, comprensión de los datos, preparación de los datos para el modelado y modelado con fines explicativos

1.4 Vista general del documento

1. Introducción:
    a. Breve introducción del documento, objetivos, justificación del análisis, entre otras cosas.
2. Plan de análisis de datos:
    a. Comprensión sobre todas las columnas de nuestro dataset, los archivos con los que trabajamos inicialmente, de dónde proviene el dataset
3. Adquisición y comprensión de los datos
    a. La información del dataset bajo un enfoque analítico, su importación en los notebooks, tipo de dato, entre otras cosas.
4. Análisis de Datos exploratorio
    a. Análisis de las columnas del dataset. Cómo se comportan, de qué forma se distribuyen en las observaciones, la presencia de outliers y sus transformaciones, tanto para un enfoque univariado, como bivariado, y en un caso, multivariado.
5. Implementación

    a. Información sobre los notebooks y archivos generados o necesarios para su ejecución, la cual dependerá de lo que el cliente requiera o no ver
6. Conclusiones
    a. Conclusiones del proyecto y recomendaciones para quien le interese retomarlo.
7. Referencias
    a. Referencias bibliográficas
8. Anexo
    a. Anexo de información extra

2 Plan de Análisis de Datos

a) Comprensión del negocio

Spotify es una compañía que ofrece un servicio de música, podcast y otros medios audiovisuales que permite a usuarios acceder a millones de canciones y contenido relacionado con la música. Kaggle es una plataforma web que junta a una comunidad de científicos de datos globalmente. Hay medio millón de usuarios activos en esta plataforma y cuenta con una inmensa cantidad de recursos que ayudan a muchos especialistas en la materia a buscar y publicar bases de datos, construir modelos, etc.

En la plataforma anteriormente mencionada tuvimos la posibilidad de encontrar un set de datos de Spotify. Este dataset cuenta con canciones distribuidas en un rango de 26 géneros. Cada canción cuenta con diferentes características de audio, unas dadas por medio de algoritmos y otras por medio de otras mediciones tangibles.

Según la descripción obtenida de Kaggle, este dataset puede ser utilizado para crear un sistema de recomendaciones basado en un input de usuario o una preferencia en específico. También, puede ser utilizado para la clasificación de las diferentes características de audio o género. Finalmente, se comenta que puede ser utilizada para aplicar cualquier otra aplicación que se nos pueda llegar a ocurrir.

Nosotros contamos con una serie de preguntas de investigación que nos gustaría responder al final de nuestro análisis del csv.

Antes de contestar nuestras preguntas de investigación, decidimos hacer un análisis univariado para lograr identificar outliers y llegar a una conclusión más completa e interesante.

b) Comprensión de los datos

Tenemos que tomar en cuenta que nuestra unidad de observación para este proyecto, son cada uno de los archivos de audio encontrados en el dataset identificados a través de la columna 'track_id' y 'genre' que puede identificarse como llave primaria compuesta.

En la figura 1 podemos encontrar una clasificación de los datos que vamos a analizar.

Tipo	Datos
Nominales	Genre, Artist_name, Track_name, Track_id, Mode, Key
---	---
Razón	Popularity, Acousticness, Danceability, Duration_ms, Energy, Instrumentalness, Liveness, Speechiness, Tempo, Valence
---	---
Ordinal:	Time_signature

Figura 2.1. Clasificación de las columnas del dataset.

A continuación, para lograr comprender los datos de una mejor manera, explicaremos a detalle cada uno de estos junto con su clasificación como variables.

Acousticness

Una medida de 0,0 a 1,0 para saber si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica. La escala continua (razón) se utiliza porque puede variar ampliamente y permitir mediciones precisas.

Artist_name

Es el nombre del artista de la canción. Aunque no tiene un orden inherente, se clasifica como una variable nominal porque se utiliza para la identificación y categorización.

Danceability

Danceability describe qué tan adecuada es una pista para bailar en función de una combinación de elementos musicales que incluyen el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general. Un valor de 0,0 es el menos bailable y 1,0 es el más bailable. Se mide en una escala continua (razón) de 0,0 a 1,0 para reflejar la diversidad en la capacidad de baile.

Duration_ms

La duración de la pista en milisegundos. Se considera una variable continua (razón) para permitir mediciones precisas y cálculos matemáticos.

Energy

La energía es una medida de 0,0 a 1,0 y representa una medida perceptiva de intensidad y actividad. Normalmente, las pistas enérgicas se sienten rápidas, ruidosas y ruidosas. Por ejemplo, el death metal tiene mucha energía, mientras que un preludio de Bach obtiene una puntuación baja en la escala. Las características de percepción que contribuyen a este atributo incluyen rango dinámico, volumen percibido, timbre, velocidad de inicio y entropía general. La energía de una

pista se mide entre 0,0 y 1,0 y se clasifica como una variable continua (razón) para representar la diversidad en la intensidad percibida de la música.

Genre

El género de la pista. Una pista puede tener más de un género, repitiendo observaciones en la lista. Debido a que se utiliza para clasificar las pistas en diferentes grupos, el género de la pista es una variable nominal que no implica un orden o una jerarquía específicos.

Id

Es un identificador de Spotify que se le asigna a una pista de audio. Es una variable nominal que identifica de manera única cada pista en la base de datos de Spotify.

Instrumentalness

Predice si una pista no contiene voces. Los sonidos "Ooh" y "aah" se tratan como instrumentales en este contexto. Las pistas de rap o de palabra hablada son claramente "vocales". Cuanto más cerca esté el valor de instrumentalidad de 1,0, mayor será la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0,5 pretenden representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0. Se mide en una escala continua (razón) de 0,0 a 1,0 para predecir la probabilidad gradual de que una pista tenga contenido vocal.

Key

La tonalidad en la que se encuentra la pista. Debido a que representa varios tonos musicales, pero no tiene un orden específico, la tonalidad de la pista es una variable nominal.

Liveness

Detecta la presencia de una audiencia en la grabación. Los valores de liveness más altos representan una mayor probabilidad de que la pista se haya interpretado en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista es una grabación en vivo. Se mide la probabilidad de que una pista se haya interpretado en vivo y se mide en una escala continua (razón) de 0,0 a 1,0 para reflejar la probabilidad gradual.

Loudness

El volumen general de una pista en decibeles (dB). Los valores de sonoridad se promedian en toda la pista y son útiles para comparar el volumen relativo de las pistas. El volumen es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db. Debido a que las variaciones en decibeles tienen un valor cuantitativo, pero no son un punto de partida absoluto, se considera una variable de intervalo.

Mode

Modo indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. Se clasifica como una variable nominal porque representa una característica categórica de la música.

Speechiness

Speechiness detecta la presencia de palabras habladas en una pista. Cuanto más exclusivamente hablada sea la grabación (por ejemplo, un programa de entrevistas, un audiolibro, poesía), más cercano a 1,0 será el valor del atributo. Los valores superiores a 0,66 describen pistas que probablemente estén compuestas exclusivamente de palabras habladas. Los valores entre 0,33 y 0,66 describen pistas que pueden contener música y voz, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores inferiores a 0,33 probablemente representen música y otras pistas que no sean de voz. Se mide de 0,0 a 1,0 en una escala continua (razón) para reflejar la proporción de contenido hablado.

Tempo

El tempo general estimado de una pista en pulsaciones por minuto (BPM). En terminología musical, el tempo es la velocidad o el ritmo de una pieza determinada y se deriva directamente de la duración promedio del tiempo. Se utiliza como una variable continua (razón) para mostrar la variabilidad en la velocidad de la música.

Time_signature

Un compás estimado. El tipo de compás es una convención de notación para especificar cuántos tiempos hay en cada compás. El tipo de compás oscila entre 3/4 y 7/4. es una variable ordinal porque representa una jerarquía de varios tipos de compás; sin embargo, no se refiere a cantidades numéricas absolutas.

Track_name

Es el nombre de la pista. Debido a que se utiliza para identificar una canción, se considera una variable nominal.

Valence

Una medida de 0,0 a 1,0 que describe la positividad musical que transmite una pista. Las pistas con valencia alta suenan más positivas (por ejemplo, felices, alegres, eufóricas), mientras que las pistas con valencia baja suenan más negativas (por ejemplo, tristes, deprimidas, enojadas). Se mide en una escala continua (razón) de 0,0 a 1,0

c) Preparación de los datos para el modelado

Para que hubiera una lógica correcta al momento de analizar las variables, lo que tuvimos que hacer es comprender cada una de ellas profundamente. Por ejemplo, en la variable de duración, el tiempo de los audios está en milisegundos, lo que provoca que exista una variación mínima que no resulta para nada significativa si queremos ver patrones en común entre las canciones. De igual manera, ver la duración de un audio en milisegundos no es lo habitual, por lo que para hacer el modelado tuvimos que agregar una columna de duración en minutos con segundos, los segundos fueron redondeados a 2 decimales. Lo que le hicimos a la duración fue una discretización debido a que expresar la duración de las pistas en milisegundos no es una unidad de medida habitual para el análisis musical.

Otros datos que tuvimos que preparar para el modelado fueron Liveness, Instrumentalness, Loudness, Speechiness. Convertimos estas variables numéricas en categóricas para poder simplificar nuestros resultados y hacerlos más comprensibles. Fue utilizada la binarización, antes de hacer el modelado de los datos. De igual manera, para Speechiness tuvimos que realizar una transformación logarítmica para reducir el sesgo de los valores extremos y poder entender perfectamente nuestros modelos.

d) Modelado de datos.

Para el modelado de datos, lo que tuvimos que utilizar fueron las librerías numpy(fórmulas probabilísticas), pandas (procesamiento de los datos en formato TIDY), Seaborn y MatPlotLib (gráficas). Con ellas pudimos graficar los datos que posteriormente analizamos y presentamos.

Dependiendo del tipo de variable, le corresponden distintos gráficos:

1. Nominales
    a. Barra
    b. Pastel
2. Ordinales, Intérvalo y Razón
    a. Histoplot
    b. Boxplot
    c. Violín

Distribución de Trabajo

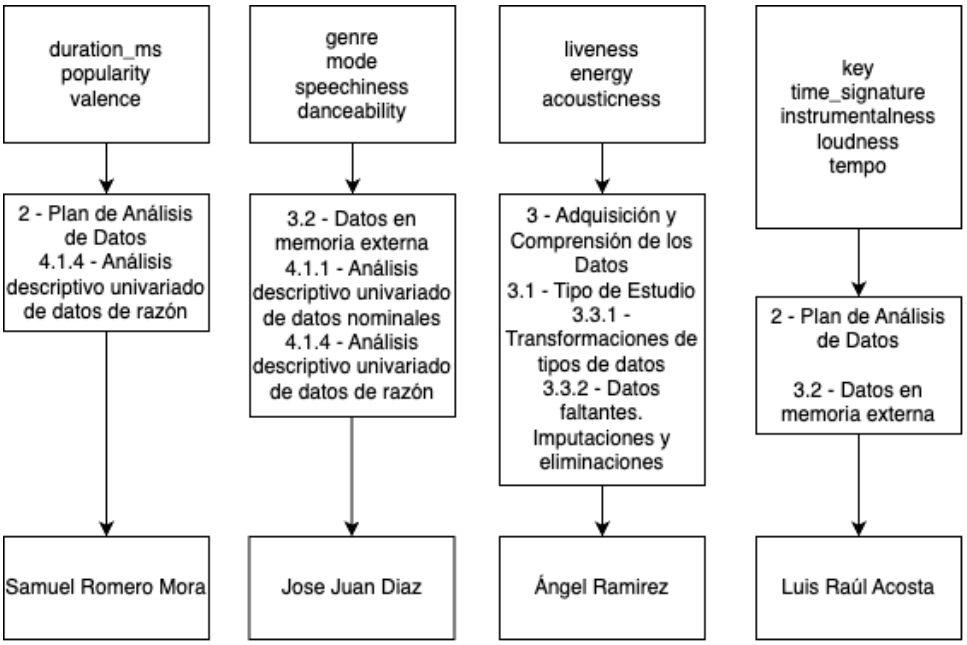


Figura 2.2. Diagrama de distribución de trabajo

3 Adquisición y comprensión de los datos

El dataset utilizado fue obtenida de una fuente confiable y reconocida en el ámbito de la investigación: el conjunto de datos de pistas de Spotify, Este recurso proporcionado a través de Kaggle, representa una muestra de datos invaluable para analizar y comprender una amplia gama de características relacionadas con las pistas musicales dentro de spotify, La disponibildiad de este dataset en formato accesible y organizado facilita en gran medida el proceso de interpretación y extracción de información relevante

Spotify Tracks DB. (2019, Julio 23). Kaggle.

https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db?resource=download

Además, para una comprensión aún más detallada de los datos, se utilizó esta página que nos da una referencia del WEB API de spotify, es actualizada pues es del presente año del 2023, por lo cual ofrece una información más reciente, directa y detallada sobre las características acústicas, técnicas de cada pista musical en el dataset, Al aprovechar esta herramienta se logra una mayor claridad en cuanto como se describen y que significan los valores específicos en cada una de las celdas del Dataset

Get Track's Audio Features. (2023). Spotify.

https://developer.spotify.com/documentation/web-api/reference/get-audio-features

3.1 Tipo de estudio

Al analizar detenidamente los datos proporcionados, hemos llegado a la conclusión de que, debido a la ausencia de una extensión temporal en la base de datos y al hecho de que se trata únicamente de un lapso de tiempo específico, la naturaleza intrínseca del estudio se revela como eminentemente transversal. Esta característica es crucial en el contexto de la investigación, ya que nos permite observar y analizar un conjunto de variables en un momento determinado, brindando una visión precisa y detallada de la situación.

3.2 Datos en memoria externa (archivos)

Tabla 3-1 Descripción de archivos de datos originales

Nombre del archivo	*Breve descripción*	*Formato (txt, csv, xlsx, json, mp4, jpg, etc.)*
Spotify_features.csv	*csv que contiene toda la información del dataset*	*csv*

Fuente: elaboración propia

Tabla 3-2 Especificación de columnas de archivo registros.csv

Nombre del campo o columna	*Tipo csv*	*Tipo de dato según escala de medición estadística*	*Tipo sugerido en Python pandas*	*Valores válidos*
genre	*Cadena de caracteres*	nominal	category	texto
artist_name	*Cadena de caracteres*	nominal	category	texto
track_name	*Cadena de caracteres*	nominal	object	texto
track_id	*Cadena de caracteres*	nominal	object	texto
popularity	*Cadena de caracteres*	razón	float32	int del 0 al 100
acousticness	*Cadena de caracteres*	razón	float32	float del 0 al 1
danceability	*Cadena de caracteres*	razón	float32	float del 0 al 1
duration_ms	*Cadena de caracteres*	razón	float32	float apartir del 0
energy	*Cadena de caracteres*	razón	float32	float del 0 al 1
instrumentalness	*Cadena de caracteres*	razón	float32	float del 0 al 1
key	*Cadena de caracteres*	nominal	category	12 valores posibles (C,C#,D,D#,E,F,F#,G,G#,A,A#,B)
liveness	*Cadena de caracteres*	razón	float32	int
loudness	*Cadena de caracteres*	razón	float32	int
mode	*Cadena de caracteres*	nominal	category	Major o Minor
speechiness	*Cadena de caracteres*	razón	float32	float del 0 al 1
tempo	*Cadena de caracteres*	razón	float32	Entero a partir del 0
time_signature	*Cadena de caracteres*	ordinal	category	7 valores posibles. 2/4, 3/4, 4/4, 5/4, 6/4, 7/4, 6/8
valence	*Cadena de caracteres*	razón	float32	float del 0 al 1

Fuente: Elaboración propia

3.3 Preprocesamiento. Primera parte.

3.3.1 Transformaciones de tipos de datos

Hasta este momento, las transformaciones de los datos han sido las siguientes:

- Binarización (división por cuartiles)
  - Liveness
  - Instrumentalness
  - Loudness
  - Speechiness
- Logaritmo:
  - Speechiness
- Discretización:

- Duration(in ms)

3.3.2 Datos faltantes. Imputaciones y eliminaciones

Para asegurar una representación más digerible y fiable de los datos, se implementó una estrategia para abordar los outliers dentro de las columnas de razón del dataset, siendo:

- Liveness
- Instrumentalness
- Loudness
- Speechiness
- Duration_ms
- Tempo

Las columnas a las que se optó por utilizar división por cuartiles con la función qcut, esto debido a que había una gran cantidad de outliers en las anteriores columnas mencionadas cuando se les hacia un gráfico de cajas y bigotes. Además, algunos de nosotros, también probamos con utilizar transformaciones de columnas de razón, y almacenarlas en otras columnas, siguiendo un formato como [*nombre_de_la_columna*2]. Al momento de gráficar, esto ayuda a procesar la información de una forma más interpretable.

Además, se procedió a llevar a cabo una depuración del Dataset donde se eliminaron un total de 9353 filas, de un total de alrededor de 250,000. Esta decisión se tomó en función de la detección de datos que resultaban incoherentes con el contexto de las variables. Esta situación se presentó en la presencia de canciones con contenido explícito clasificadas con la categoría (genre) de Children´s music, a sabiendas de la existencia de otro Children's music que si contenía información fiable.

Esto se asume que fue un error de las personas que publicaron este dataset, en el sitio web de kaggle. Se sabe que esta información es almacenada por Spotify de una forma que priorice el uso de almacenamiento, por lo tanto, se utilizan claves numéricas para almacenar elementos como las canciones. Es por eso, que el dataset fue manejado por las personas que lo proporcionaron, para que información como el tempo, género, modo, entre otras cosas, fuera fácil de interpretar para aquellos que hagan el análisis con esta muestra.

Como existía un error tipográfico en Children's Music (donde el incorrecto, en vez de escribir el género con un apóstrofe, utilizaba una coma "Children´s Music", fue fácil encontrar el rango con información errónea y hacer un drop de ese rango de filas).

Esta meticulosa revisión y limpieza de la base de datos es crucial para garantizar la integridad y fiabilidad de los resultados obtenidos en el estudio. Al eliminar datos inconsistentes o erróneos, se asegura que el análisis se base en información precisa y confiable, lo que a su vez fortalece la validez y robustez de las conclusiones que se derivan del estudio.

3.3.3 Eliminación de duplicados

En este caso, no hubo eliminación de duplicados. Realizamos una prueba como unas 50 canciones, posicionadas en distintos rangos del dataset, para ver si aparecían más de una vez. Nos dimos cuenta de que, en efecto, una canción puede aparecer más de una vez en la muestra. Sin embargo, esto funciona de esta manera para indicar que una canción puede tener más de un género, habiendo una relación N:M. Por eso, es que la llave aquí es compuesta (Género y TrackID).

3.4 Conclusiones del capítulo

Conclusivamente, el dataset:

- Necesitó ser transformado en algunas de sus columnas de razón
- No hubieron duplicados, no al menos de forma muy evidente.
- Tuvo que ser limpiado por datos de naturaleza incorrecta, ya que existían dos géneros musicales con la etiqueta children's music, uno con información errónea y otro correcto.

Esto nos da la seguridad que podremos obtener información relativamente verídica del análisis que realicemos.

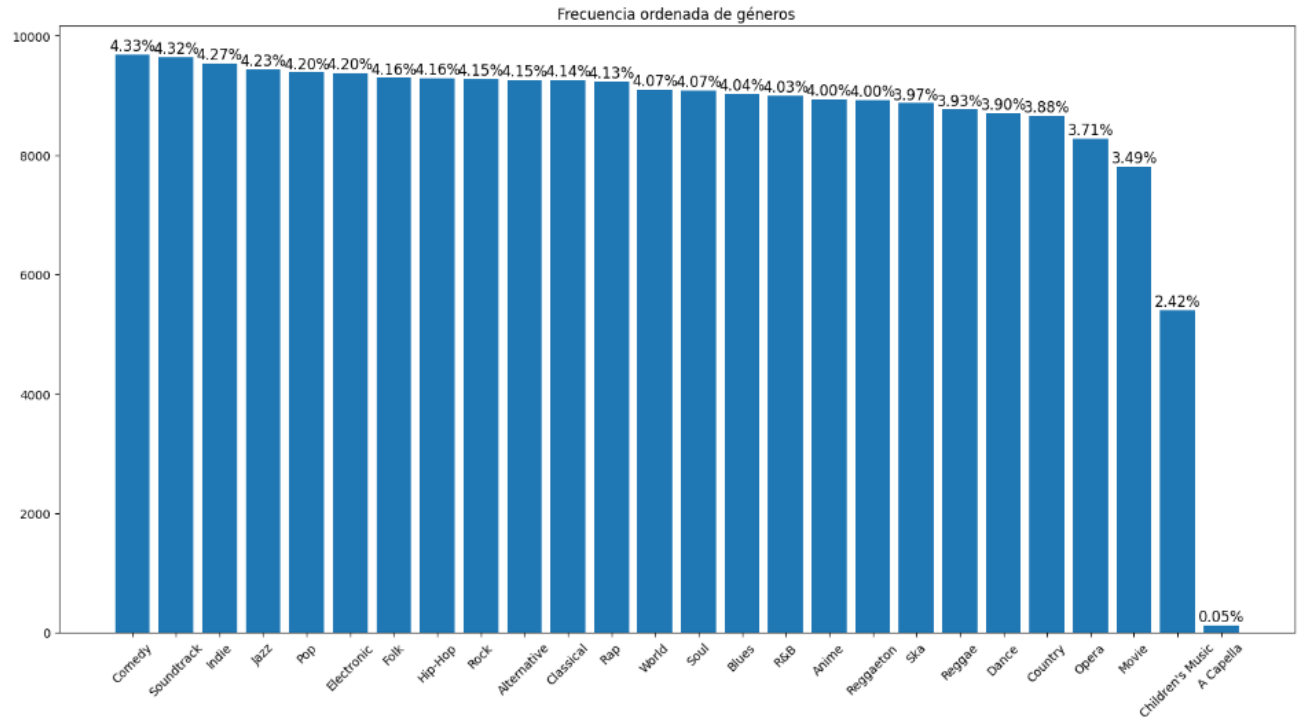
4 Análisis de Datos Exploratorio

4.1 Análisis descriptivo univariado

4.1.1 Análisis descriptivo univariado de datos nominales

Genre

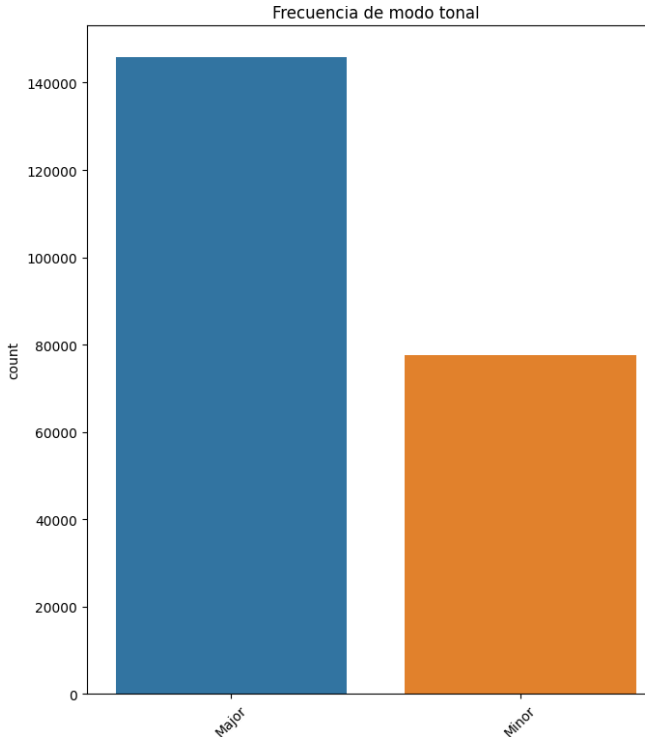
Figura 4.1. Gráfica de barras con las frecuencias y porcentaje de la columna genre



- El valor de la moda es Comedy, con una frecuencia de 9681 repeticiones.
- El rango tiene un valor de 26, es decir, hay 26 géneros

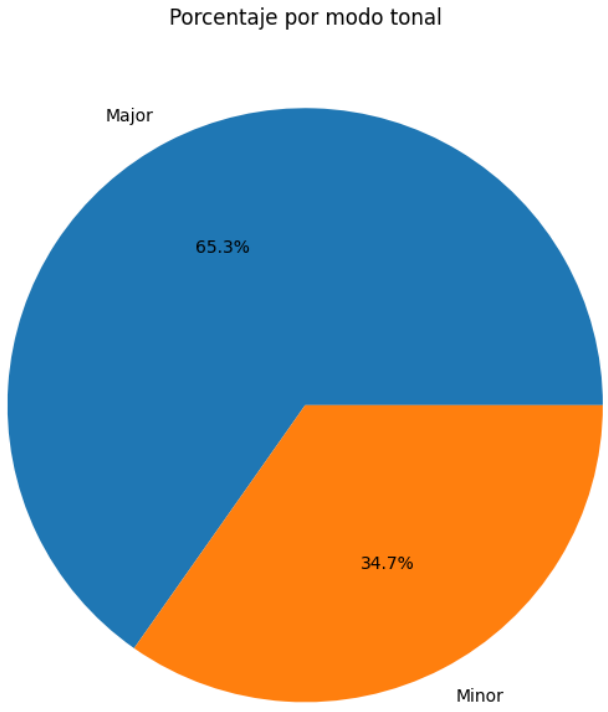
Mode

Figura 4.2. Gráfica de barras con las frecuencias de la columna mode.



- El valor de la moda es Major, con una frecuencia de 145,767 repeticiones.
- El rango tiene un valor de 2, Major o Mi

Figura 4.3. Gráfica de pastel con el porcentaje de la columna mode.



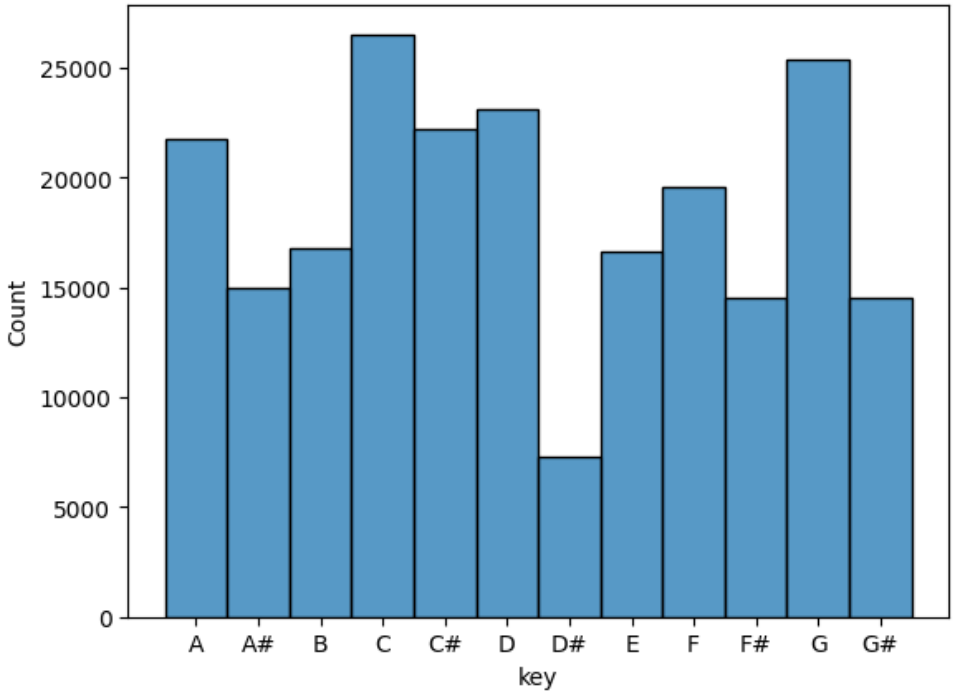
Key

El valor de la moda es C, con una frecuencia de 26,518 repeticiones.

El rango de esta columna comprende un total de 12 categorías: A, A#, B, C, C#, D, D#, E, F, F#, G y G#.

Figura __. Gráfica de barras de densidad de categorías de la columna key.

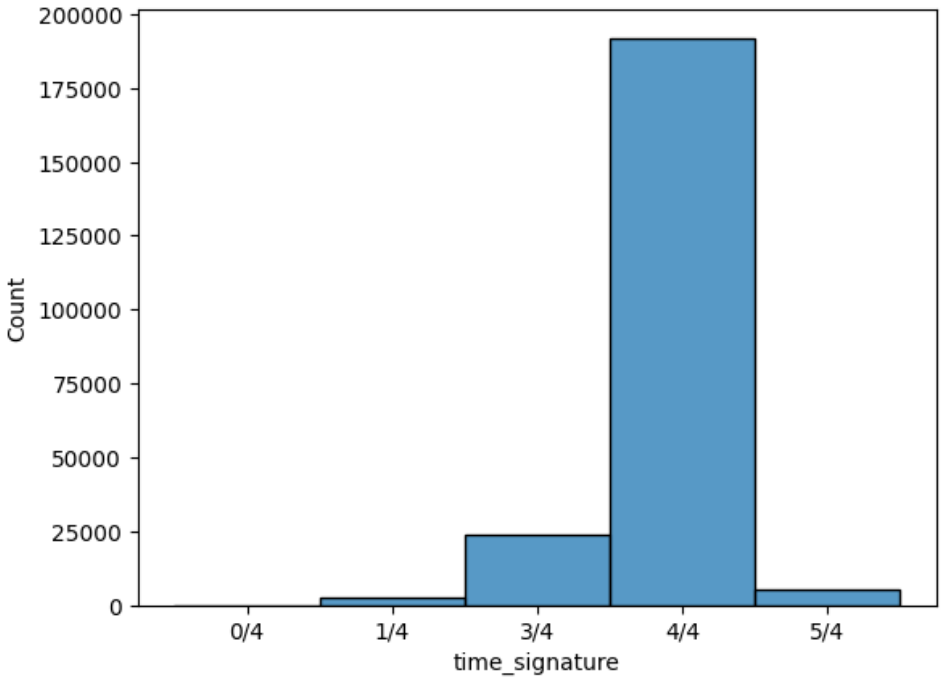
Figura 4.4. Gráfica de barras con frecuencia de cada clave.



4.1.2 Análisis descriptivo univariado de datos ordinales

Time_signature

Figura 4.5. Gráfica de barras con la frecuencia de las categorías de time_signature

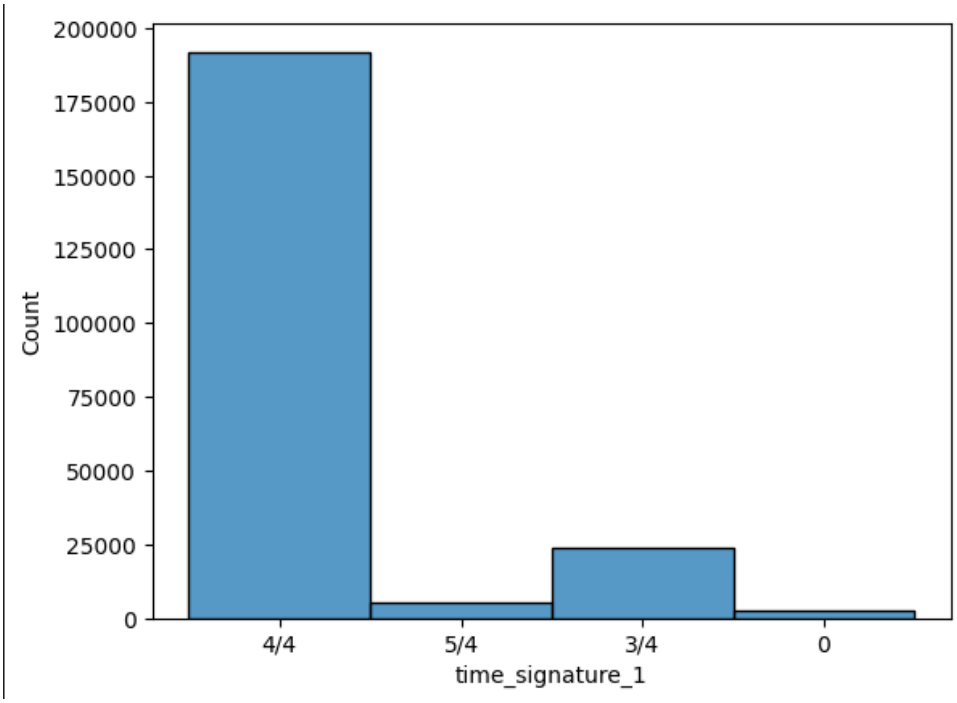


Datos estadísticos:

En esta columna, la moda es 4/4. Es el valor que se repite con mucha más frecuencia. La mediana en este caso también es 4/4, por lo que la mitad de los datos son 4/4 o menor, mientras que la otra mitad tiene valores de 4/4 o mayor.

El rango, obtenido de restar el valor mayor (5/4) menos el valor menor (0/4), es de 5/4. Según la documentación de la API de Spotify, el rango de valores admitidos comprende de 3/4 a 7/4; por lo que creamos una nueva columna "time_signature_1" que trate los valores 0/4 y 1/4 como desconocidos al referirse a ellos con "0".

Figura 4.6. Gráfica de barras con la frecuencia de las categorías de time_signature_1



4.1.3 Análisis descriptivo univariado de datos de intervalo

Loudness

Figura 4.7. Histograma de la columna loudness

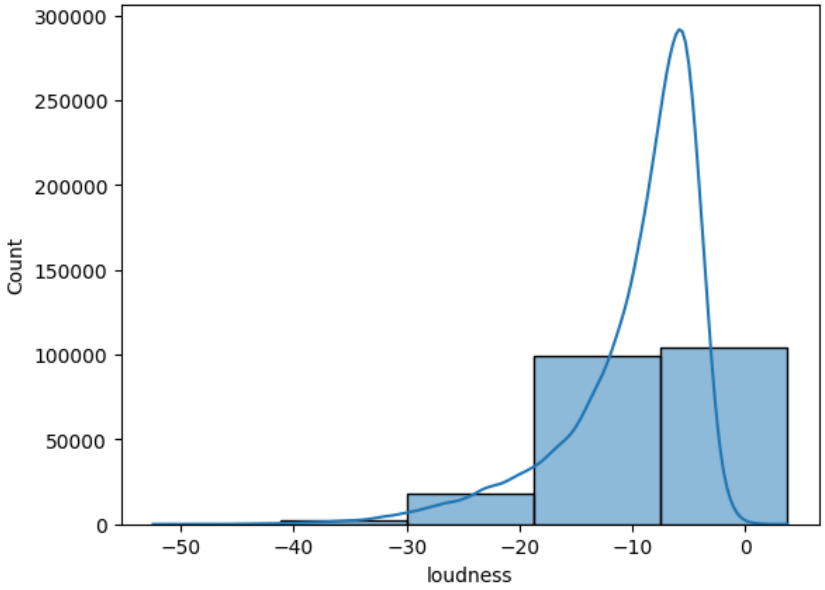
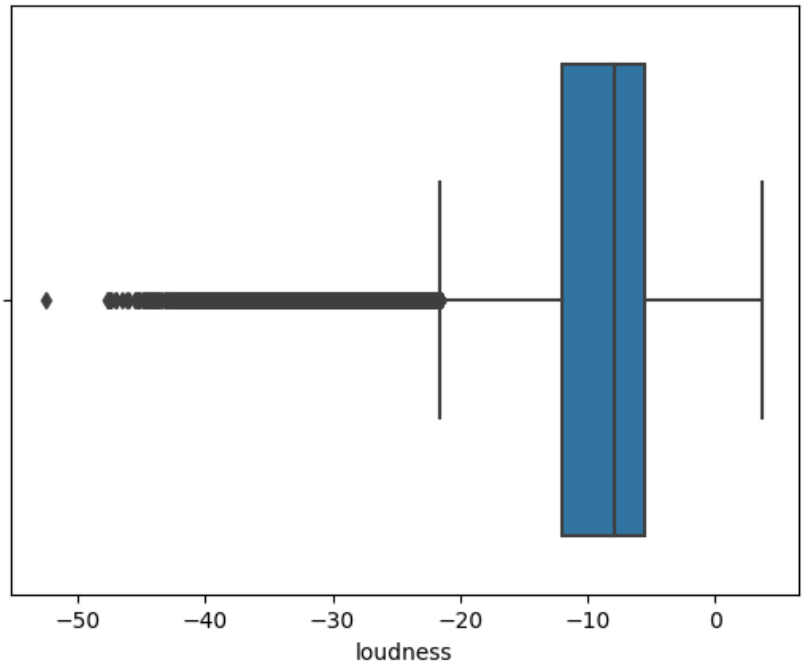


Figura 4.8. Diagrama de cajas y bigotes de la columna loudness



Estadísticas descriptivas

Promedio: -9.697311

Desviación estándar: 6.063335

Mínimo: -52.457

25%: -11.984

50%: -7.87

75%: -5.562

Máximo: 3.744

Los valores de esta columna típicamente se encuentran entre -60 y 0. De nuestra muestra encontramos que, de hecho, la mayoría se encuentra dentro de este rango esperado ya que, el 75% inferior es menor a 0 y el mínimo valor encontrado es -52.457.

4.1.3.1 Valores atípicos

<Comentar los hallazgos y explicar, si fuera el caso, la eliminación o transformación de datos derivada del análisis de datos atípicos.

Usar histogramas, boxplots o gráficas especializadas para outliers>

Loudness

Los valores de esta columna presentan una cantidad significativa de datos atípicos, posicionados por debajo de la mediana. Con herramientas de Pandas, realizamos un corte por cuantiles para manejar estos valores a través de rangos.

Figura 4.9. Histograma de la columna loudness sin valores atípicos

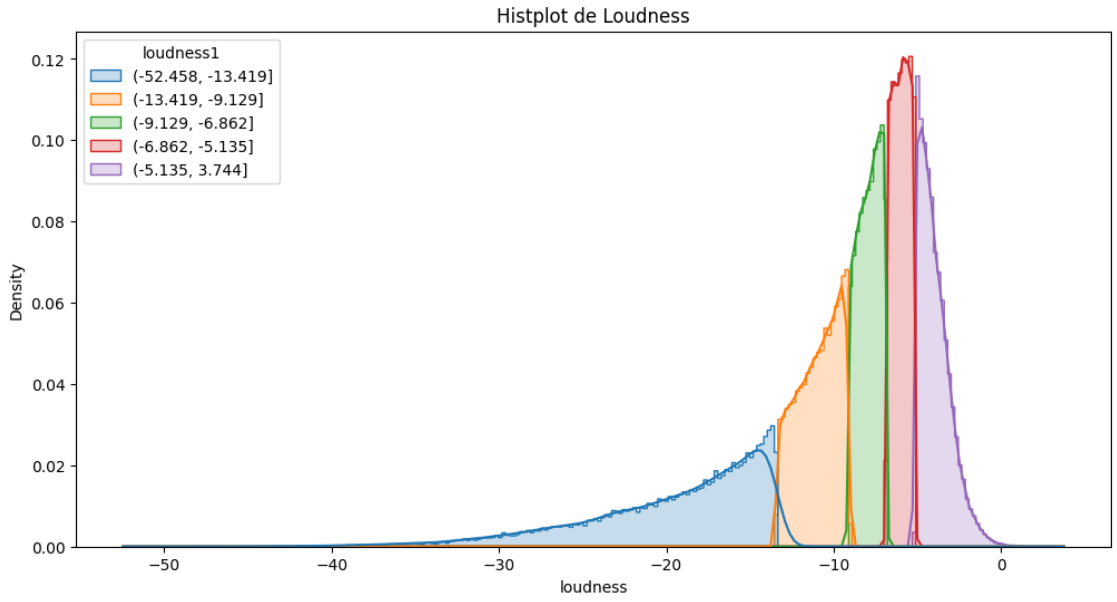
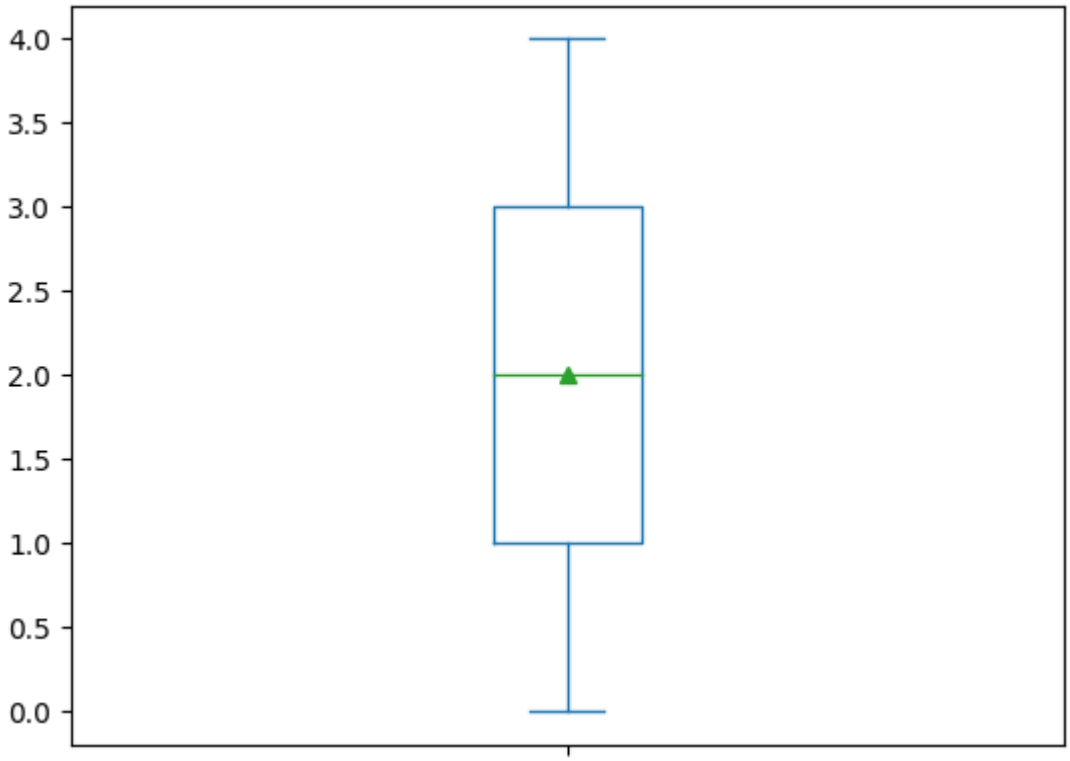


Figura 4.10. Diagrama de cajas y bigotes de la columna loudness sin valores atípicos



Los valores de loudness presentan una distribución con asimetría hacia la derecha, lo que quiere decir que la mayoría de las observaciones tiene una cantidad de decibeles mayor al valor de la media.

4.1.4 Análisis descriptivo univariado de datos de razón

Danceability

Figura 4.11. Histograma de la columna danceability

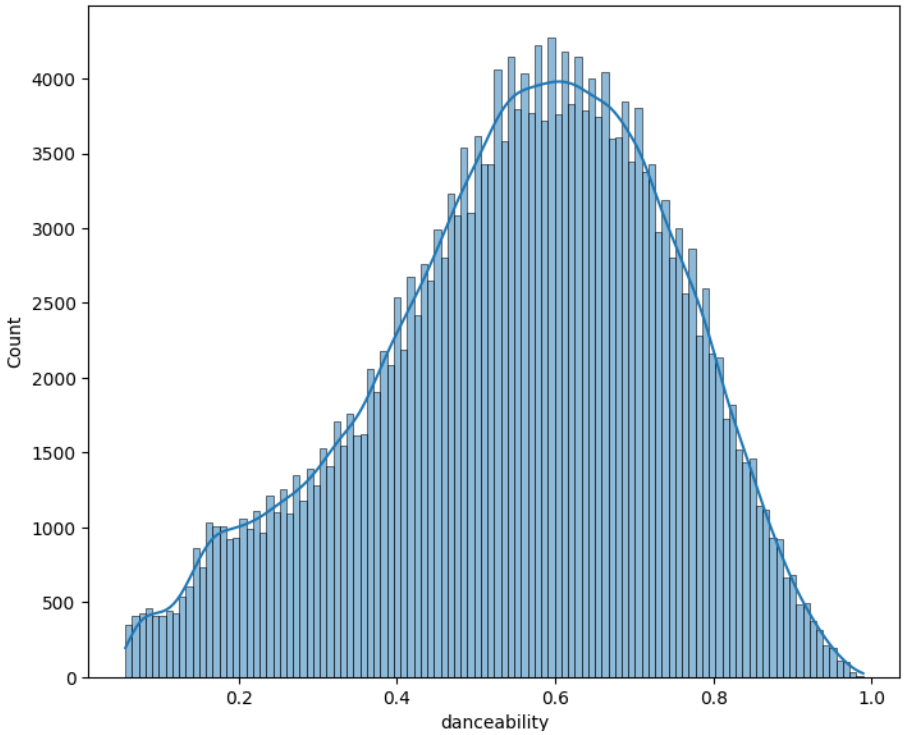
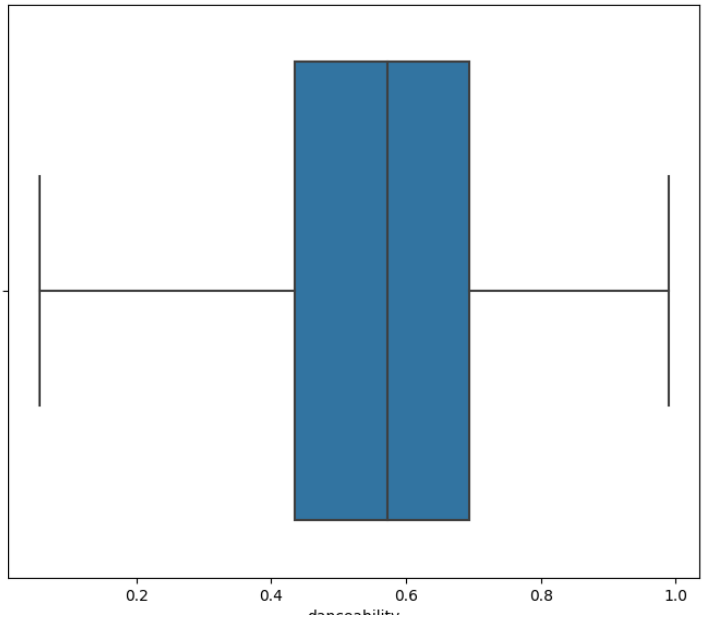


Figura 4.12. Diagrama de cajas y bigotes de la columna danceability



- La densidad de kernel nos permite ver este acampanamiento de una manera más evidente. No parece haber una distribución anormal de los datos ni outliers.
- El de caja confirma, la ausencia de outliers tan evidentes. Además, es posible presenciar que la media y la mediana están muy cerca, lo que asegura de nuevo que la distribución de los datos es relativamente simétrica y no está sesgada de manera significativa.

```
• #danceability
• ##count    223372.000000
```

```
##mean          0.554889
##std           0.187030
##min           0.056900
##25%           0.435000
##50%           0.572000
##75%           0.694000
##max           0.989000
```

Interpretación:

- La mayoría de las observaciones tienen una danzabilidad media (0.5548).
- En el estudio, predominan las canciones que son bailables, por poco.
- La canción menos bailable tiene 0.05 y la más, 0.98.
- La desviación estándar es de 0.18.

Speechiness

Figura 4.13. Histograma de la columna speechiness

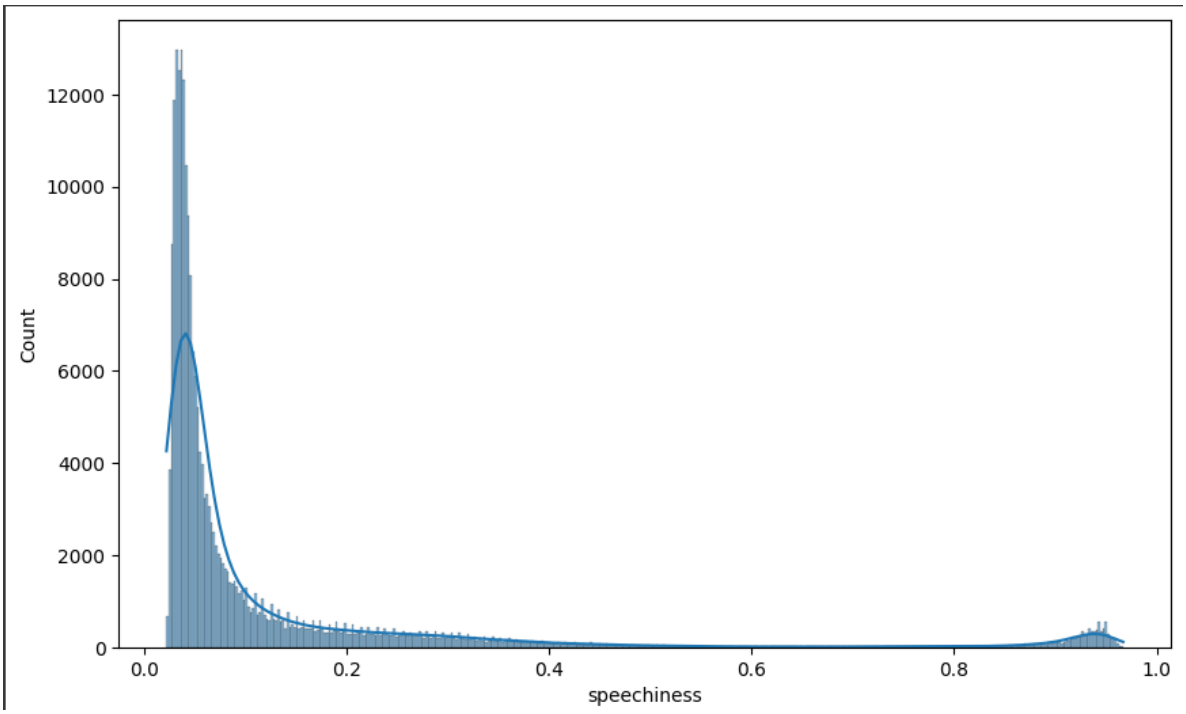


Figura 4.14. Histograma con transformación log10 de la columna speechiness

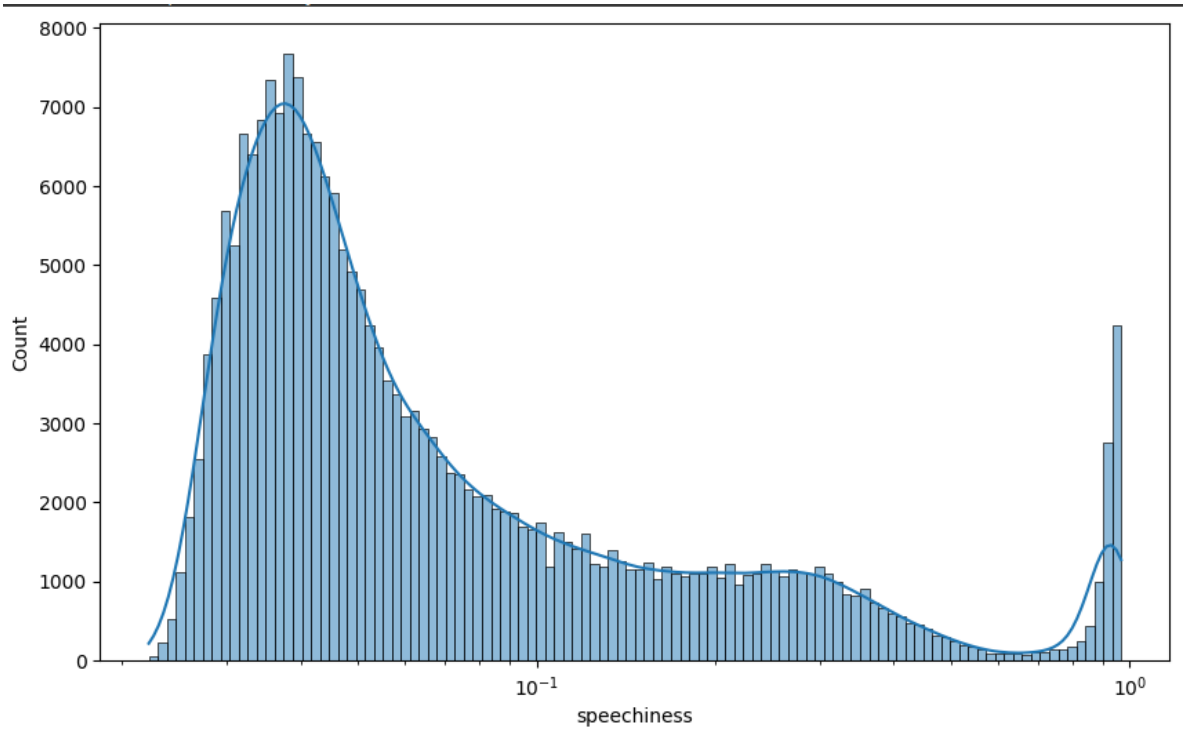


Figura 4.15. Diagrama de cajas y bigotes de la columna speechiness

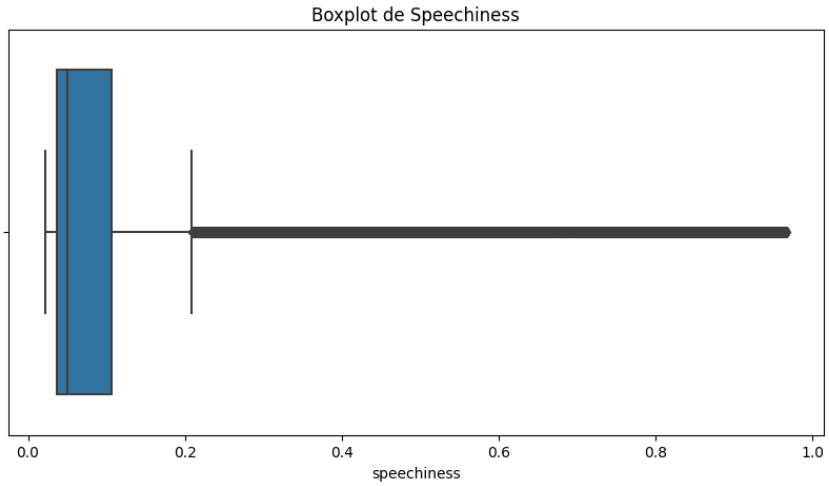
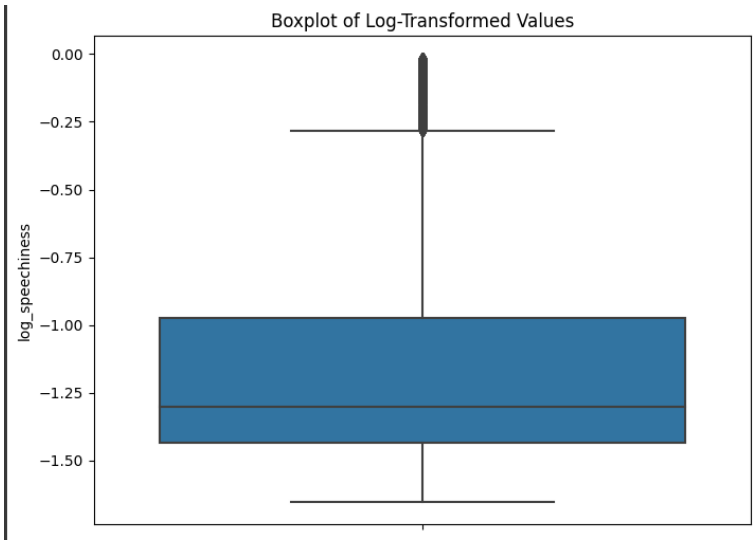


Figura 4.16. Diagrama de cajas y bigotes con transformación de log 10 de la columna speechiness



- Utilizando una escala logaritmica en el histoplot, es mucho más fácil interpretar la información, ya que, como mencionamos anteriormente, hay un desbalance de outliers importante.
- Asimismo, existe un rango muy amplio en esta columna, ya que los valores oscilnan los 6 decimales de presición, de una muestra de más de 200,000 observaciones

```
• #speechiness
    ○ ##count    223372.000000
    ○ ##mean          0.122200
    ○ ##std           0.188338
    ○ ##min           0.022200
    ○ ##25%           0.036700
    ○ ##50%           0.050000
    ○ ##75%           0.106000
    ○ ##max           0.967000
```

Interpretación:

- La basta mayoría de las observaciones no tienen voces habladas, y si las tienen, es en calidad melódica (menor a 0.3) o rapeada (menor a 0.5).
- La música instrumental y poco lírica mantienen valores cercanos a 0.
- Los picos de alta speechiness están probablemente relacionados a los géneros de Comedia, Children's Music y Acapella.

Duration_ms

```
[ ]  df['duration_ms'].describe()

count    2.233720e+05
mean     2.352299e+05
std      1.207831e+05
min      1.538700e+04
25%      1.820500e+05
50%      2.202270e+05
75%      2.660670e+05
max      5.552917e+06
Name: duration_ms, dtype: float64
```

Se realizó un análisis de la duración en milisegundos de 223,372 canciones.

Lo que podemos concluir con este análisis, es que el tiempo promedio por canción (Mean) es de 2.233720e+05 milisegundos, lo que a minutos equivale a 3.9204983333. Esto quiere decir que en promedio las canciones duran aproximadamente 3 minutos con 55 segundos.

La desviación estándar de esta variable es de alrededor de 2 minutos o 120783.1 milisegundos. Con esto nos damos cuenta de que, del conjunto de datos, los valores individuales tienden a dispersarse alrededor del promedio (3 minutos con 55 segundos) en aproximadamente 2 minutos. Podemos decir que la mayoría de las canciones no tienden a pasarse de los 5 minutos con 55 segundos ni ser menores a 1 minuto con 55 segundos.

El valor mínimo que encontramos de duración de una canción es de 0.25645 minutos, lo que equivale a 15 segundos, mientras que el valor máximo es de 5.552917e+06 milisegundos, equivalente alrededor de 92 minutos y 33 segundos.

De esta información también pudimos obtener los cuartiles 1, 2 y 3. El primer cuartil nos dice que el 25% de los datos tiene un valor menor o igual a 1.820500e+05 milisegundos o 3 minutos con 2 segundos. El segundo cuartil nos indica que el 50% de las canciones observadas tiene una duración igual o menor a 2.202270e+05 milisegundos o 3.67 minutos o 3 minutos con 40 segundos. El tercer percentil nos da a entender que el 75% de los valores analizados son iguales o menores a 2.660670e+05 milisegundos o 4.43445 minutos o 4 minutos con 46 segundos.

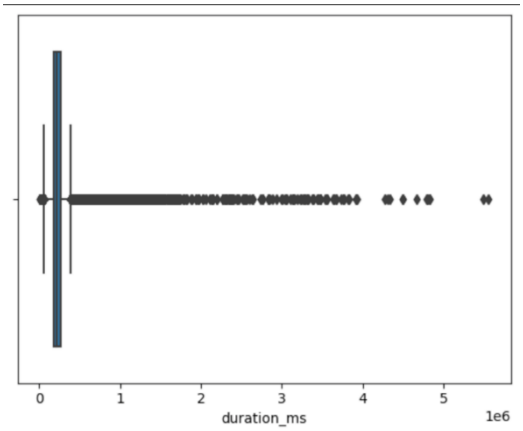


Figura 4.17. Diagrama de cajas y bigotes de duration_ms

En este diagrama podemos observar que hay muchos valores que son outliers, varían bastante por la naturaleza de los datos. Como son canciones no todas duran lo mismo y esto toma más sentido cuando tenemos en cuenta que puede haber variaciones de milisegundos que hagan parecer que hay muchos outliers. Para este análisis podemos crear una nueva columna que incluya el valor en minutos redondeado a dos decimales y posteriormente verificar si sigue habiendo la misma cantidad de outliers.

La columna fue creada, y se ve de la siguiente manera:

```
   duration_minutes
0              1.66
1              2.29
2              2.84
3              2.54
4              1.38

[5 rows x 53 columns]
```

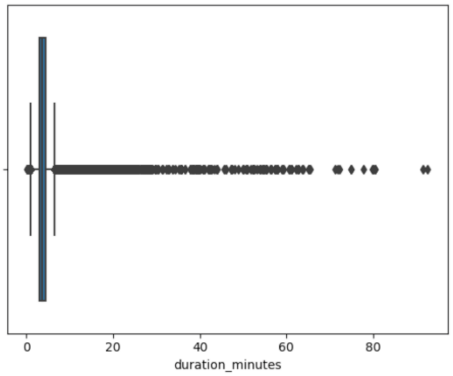


Figura 4.18. Diagrama de cajas y bigotes de duration_minutes

Si hacemos el boxplot de nueva cuenta podemos notar que hay outliers pero no a la misma escala que la anterior gráfica, ya que la anterior gráfica estaba representada con 1e6 (Un millón). Ahora está a otra escala que no está tan descabellada. Aún debemos de tener en cuenta que, debido a la naturaleza de los datos, las canciones siempre tendrán ligeras variaciones de tiempo.

Finalmente pudimos obtener los siguientes datos:

```
La moda es: 0    240000
Name: duration_ms, dtype: int64
La varianza es: 14588549549.888058
La curtosis es: 245.51052532192674
```

La moda es de 240000 milisegundos o 4 minutos, lo que nos dice que el tiempo que más se repite son 4 minutos exactamente. La varianza nos indica que los valores individuales varían con respecto a la media de manera bastante dispersa. La curtosis, por otra parte, es de 245.51052532192674 lo que nos hace ver que la distribución de las duraciones es puntiaguda, lo que indica que hay valores extremadamente altos y también extremadamente bajos.

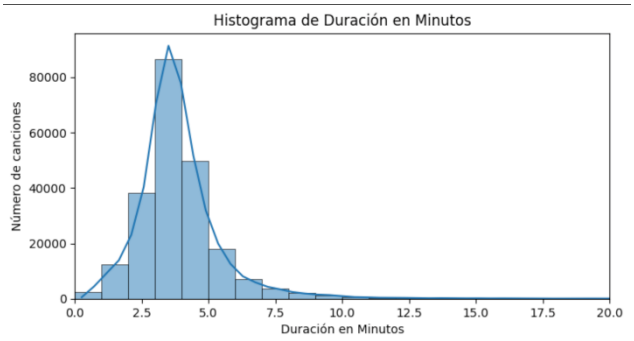


Figura 4.19. Histograma de duración en minutos

Este histograma nos da información importante acerca del tiempo de las canciones. Primero que nada, vemos que hay un claro rango de 0 a 10 minutos aproximadamente, claro que puede haber valores extremos, pero la mayoría se encuentra en ese rango. También podemos ver que la punta del histograma es muy puntiaguda, lo que significa que la mayoría de los valores se acerca

bastante a la media. Podemos apreciar que hay una distribución normal, no hay brincos muy pronunciados, los datos no varían extremadamente y se distribuyen simétricamente

Popularity

```
[ ]  df['popularity'].describe()

count    223372.000000
mean         40.560912
std          18.280096
min           0.000000
25%          28.000000
50%          42.000000
75%          54.000000
max         100.000000
Name: popularity, dtype: float64
```

Se realizó un análisis de la popularidad de 223,372 canciones.

Lo que podemos concluir con este análisis, es que el promedio de popularidad del set de canciones (Mean) es de 40.560912. Esto quiere decir que en promedio las canciones de este set tienen una popularidad de 40.560912/100, lo que indica que tomando en cuenta todo el dataset, no son tan populares en promedio.

La desviación estándar de esta variable es de alrededor de 18.280096. Con esto nos damos cuenta de que, del conjunto de datos, los valores individuales tienden a dispersarse alrededor del promedio (40.560912) en aproximadamente 18.280096. Podemos decir que la mayoría de las canciones no tienden a pasarse de los 58.84 puntos de popularidad ni ser menores a 22.28081 puntos.

El valor mínimo que encontramos de popularidad de una canción es de 0, Mientras que el valor máximo es de 100.

De esta información también pudimos obtener los cuartiles 1, 2 y 3. El primer cuartil nos dice que el 25% de los datos tiene un valor menor o igual a 28 puntos sobre 100. El segundo cuartil nos indica que el 50% de las canciones observadas tiene una duración igual o menor a 42 puntos sobre

100. El tercer percentil nos da a entender que el 75% de los valores analizados son iguales o menores a 54 puntos sobre 100.

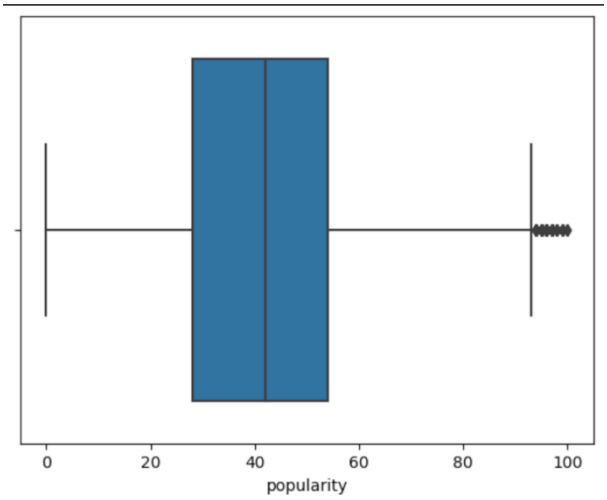


Figura 4.20. Diagrama de cajas y bigotes sobre la distribución de popularity

Gracias a este diagrama podemos ver un ligero sesgo a la derecha. La distribución de datos es un poco asimétrica. Podemos concluir que hay ciertos valores atípicos a la derecha, esto representa a las canciones que son extremadamente populares, se consideran valores atípicos porque se salen completamente de la caja.

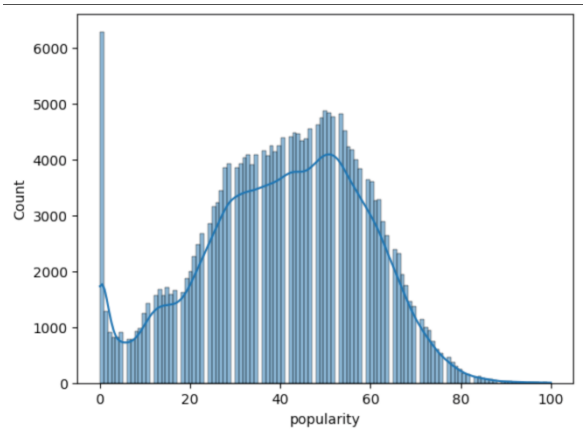


Figura 4.21. Histograma de la distribución de popularity

Una cantidad considerable de datos obtuvieron un grado de popularidad de 0 debido a que una considerable parte de las canciones no despegan en cuanto a esta métrica o se quedan atascadas en puntajes bajos. La forma que tiene el histograma es semi acampanado, hay algunas irregularidades, pero notablemente podemos ver que son muy pocas las canciones que obtienen puntajes de 80 para arriba, lo que quiere decir que es muy poco probable que una canción sea popular. En otras palabras, son pocas las canciones que tienen una popularidad alta. Gracias a la densidad de la línea podemos observar cómo se distribuyen los puntajes de popularidad de las canciones.

Valence

```
[ ]  df['valence'].describe()

count    223372.000000
mean          0.455155
std           0.261695
min           0.000000
25%           0.235000
50%           0.444000
75%           0.663000
max           1.000000
Name: valence, dtype: float64
```

Se realizó un análisis de la valencia entre valores de 0.0 a 1.0 en el dataset

Lo que podemos concluir con este análisis, es que la valencia promedio de las canciones del dataset es de 0.455155, esto nos dice que hay canciones con sentimientos más negativos que positivos.

La desviación estándar de esta variable es de alrededor de 0.261695. Con esto nos damos cuenta de que, del conjunto de datos, los valores individuales tienden a dispersarse alrededor del promedio (0.455155) en aproximadamente 0.261695. Podemos decir que la mayoría de las canciones no tienden a pasarse del valor 0.71685, ni ser menores a 0.19346 de valencia.

El valor mínimo que encontramos de valencia de una canción es de 0.000000. El valor máximo que encontramos de valencia en una canción es de 1.0

De esta información también pudimos obtener los cuartiles 1, 2 y 3. El primer cuartil nos dice que el 25% de los datos tiene un valor menor o igual a 0.235000. El segundo cuartil nos indica que el 50% de las canciones observadas tiene una duración igual o menor a 0.444000. El tercer percentil nos da a entender que el 75% de los valores analizados son iguales o menores a0.663000

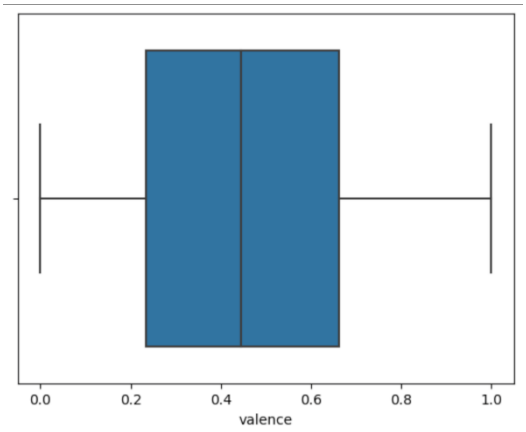


Figura 4.22. Diagrama de cajas y bigotes de la distribución de valence

En este diagrama podemos ver que la caja está ligeramente inclinada a la izquierda, lo que hace sentido cuando tenemos en cuenta que el promedio de valencia de nuestra data frame es de 0.455155 Podemos ver que dentro de la caja hay una simetría muy marcada, podemos decir que los datos a la derecha de la línea que separa a la caja son datos arriba de la media y los datos a la izquierda son datos menores. También podemos darnos cuenta de que no hay outliers, lo que nos da a entender que hay una distribución uniforme, no hay valores externos que extremadamente alejados de la mayoría de los datos.

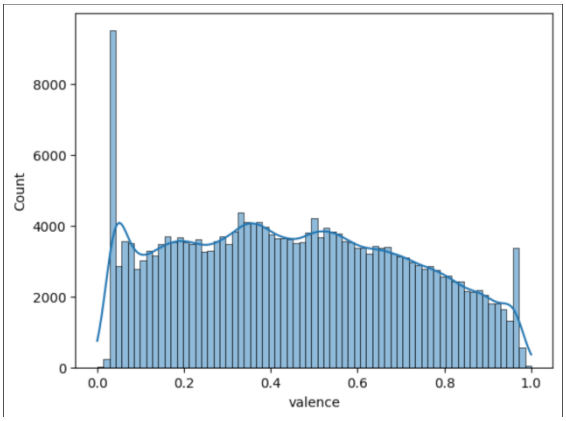


Figura 4.23. Histograma de la distribución de valence

En este histograma podemos ver cómo hay bastantes canciones en el valor de valencia de aproximadamente 0.1, no se compara con la cantidad de canciones que se encuentran en los valores extremos de valencia cercana a 1. Con esto podemos entender que, hablando de extremos, hay más canciones con emociones más negativas que positivas. Podemos ver una gráfica un poco más inclinada a la izquierda, lo que hace sentido con el análisis que habíamos realizado previamente sobre las valencias.

COLUMNAS DE ANGEL

Figura 4.24. ***Diagrama de la distribución de "energy"***

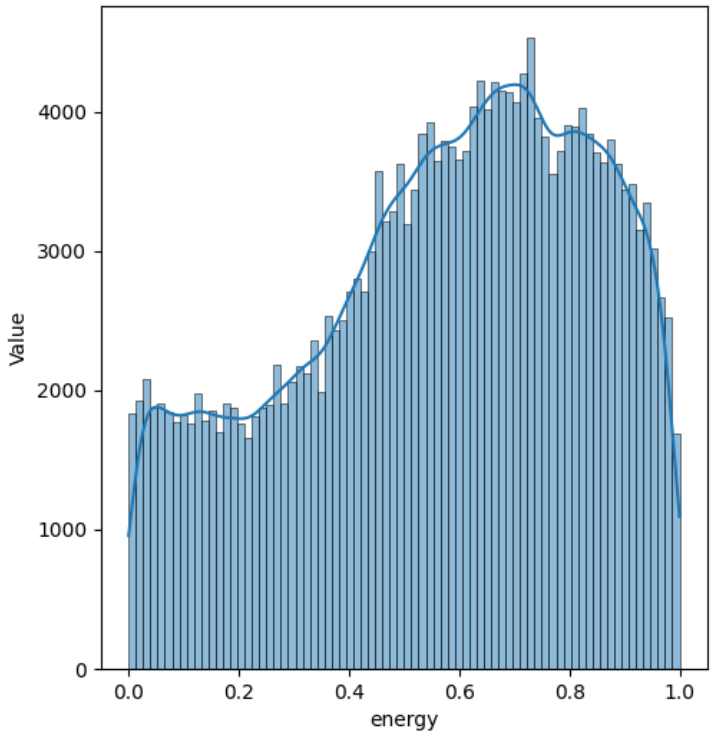
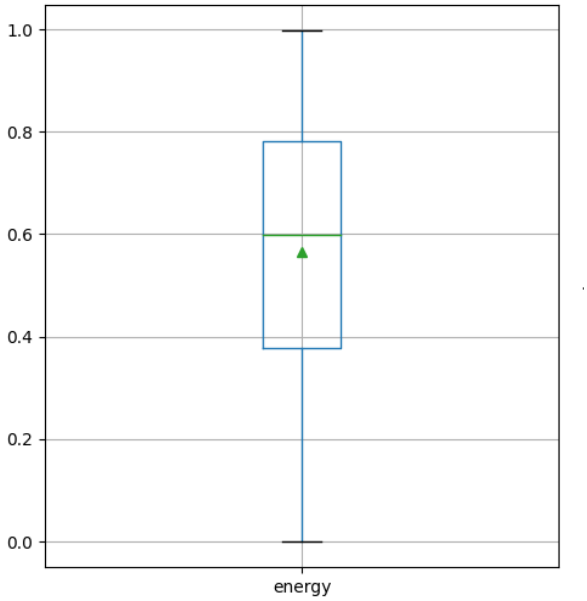


Figura 4.25. Diagrama de caja y bigotes de "energy"



Datos estadísticos

Media 0.565277

Desviación estándar 0.264097

mínimo 0.000020

25% 0.377000

50% 0.599000

75% 0.781000

Maximo 0.999000

La columna energy mide valores del 0 al 1 de que tanta “energía” hay en una canción, implicando ruido, rapidez, por ejemplo, una canción de black metal tendrá una presencia de energía alta, a base de los estadísticos podemos determinar que en promedio las canciones tienden a inclinarse a tener una cantidad de energía considerable con un 0.5652, en este diagrama no obtuvimos presencia de outliers.

Figura 4.26. ***Diagrama de la distribución de “Liveness””***

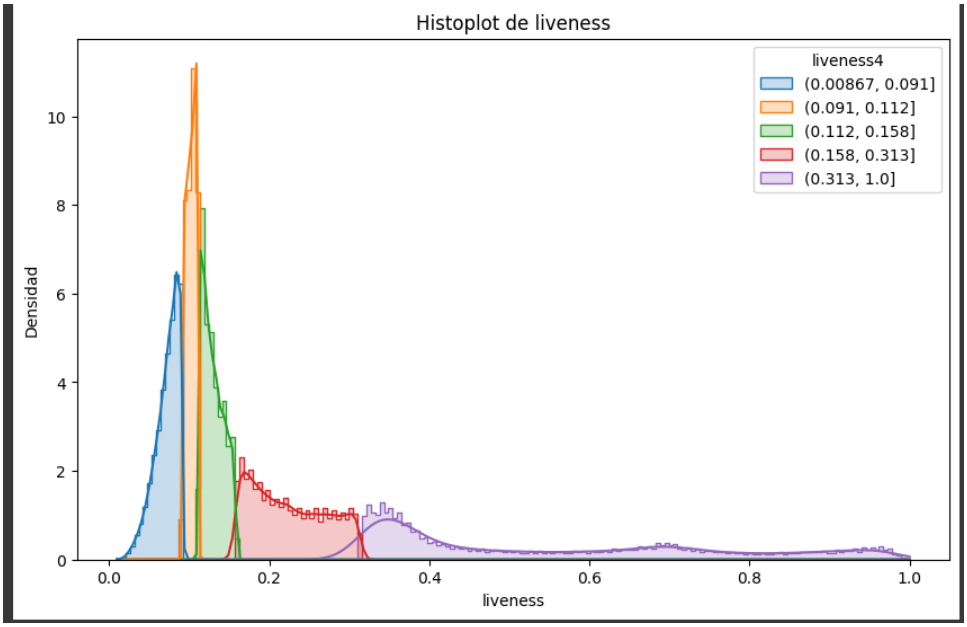
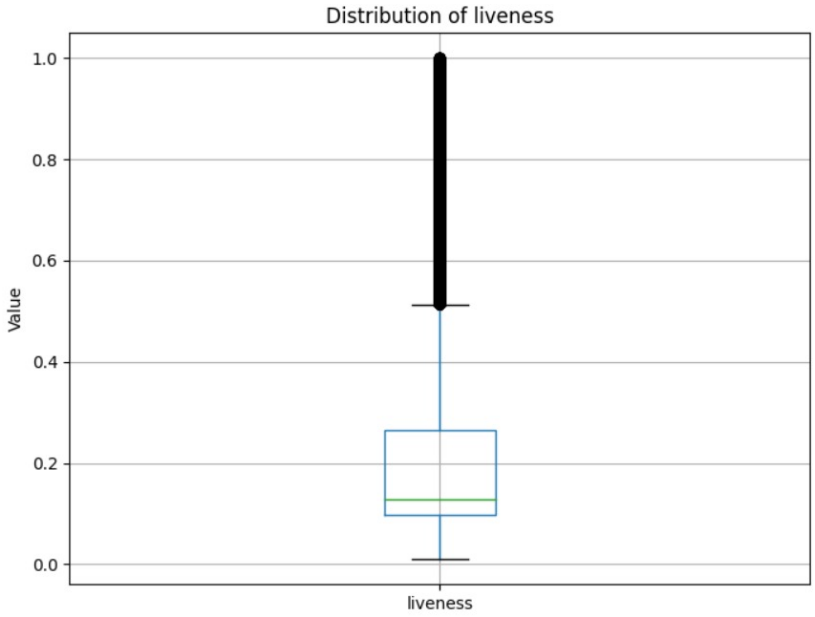
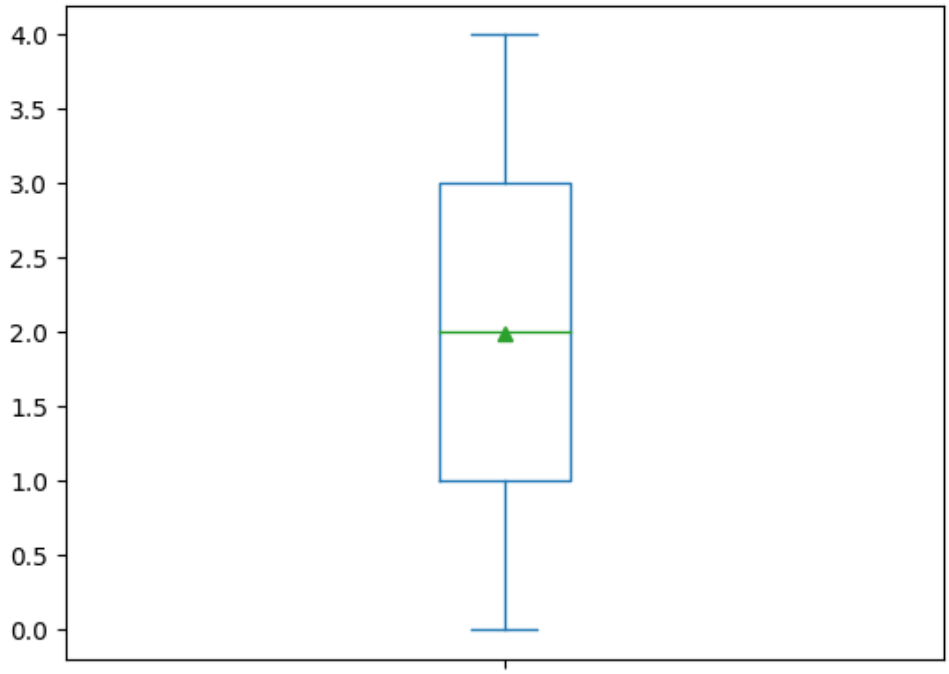


Figura 4.27. Diagrama de caja y bigotes de "***Liveness***" con valores atipicos



Como se puede observar en el histograma implica que hay muchos valores pero no los suficientes como para ser considerados parte del histograma que rayan arriba de la media valores que estn entre 0.5 y 1 pero son muy pocos pero son demasiados valores unicos que son considerados outliers y no se puede realizar una eliminacion entonces se hace una binarizacion y apartir de eso obtenemos la siguiente figura.

Figura 4.28. Diagrama de caja y bigotes de "***Liveness***" sin valores atipicos



Datos estadísticos

Media 0.215846

Desviación estándar 0.199904

mínimo 0.009670

25% 0.097300

50% 0.128000

75% 0.264000

Maximo 1.000000

Liveness es una medición del 0 al 1 que determina que tan posible es que la canción haya sido tocada en vivo, como podemos ver y es algo lógico la mayoría de las canciones no tienen una gran posibilidad de ser tocadas en vivo, y tienen una presencia de 0.21 muy improbable que sean tocadas en vivo, sin embargo, como valor máximo tenemos un 1 asegurando que si hay canciones que fueron tocadas en vivo.

Figura 4.29. ***Diagrama de la distribución de "acousticness"***

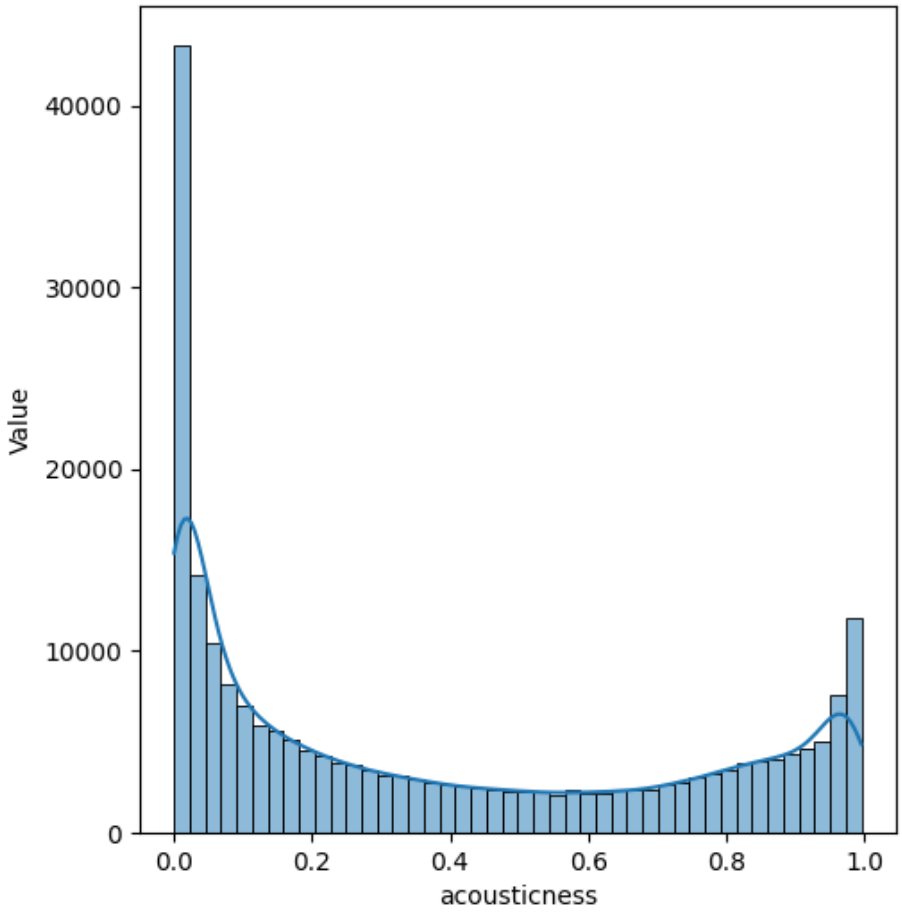


Figura 4.30. Diagrama de cajas y bigotes de “Acousticness”

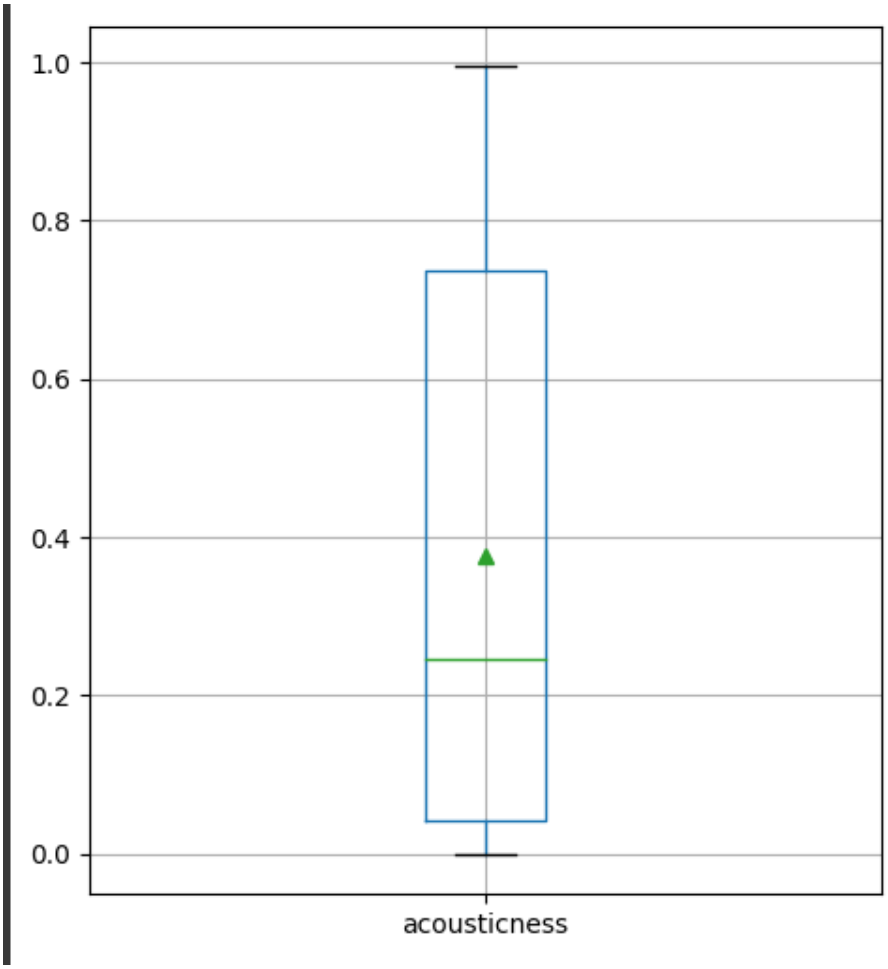


Figura 4.31. Diagrama de cajas y bigotes sobre acousticness

Datos estadísticos

Media 0.377160

Desviación estándar 0.356205

mínimo 0.000000

25% 0.042000

50% 0.246000

75% 0.737000

Máximo 0.996000

Es una medida del 0 al 1 que representa que tanta seguridad hay de que la canción haya sido tocada de forma acústica, como podemos la media implica que la mayoría de las canciones no fueron tocadas de forma acústica, sin embargo, sabemos que si hay bastante presencia de canciones acústicas en diferentes valores puesto que no obtuvimos outliers en este diagrama.

Instrumentalness

Figura 4.32. Histograma de instrumentalness

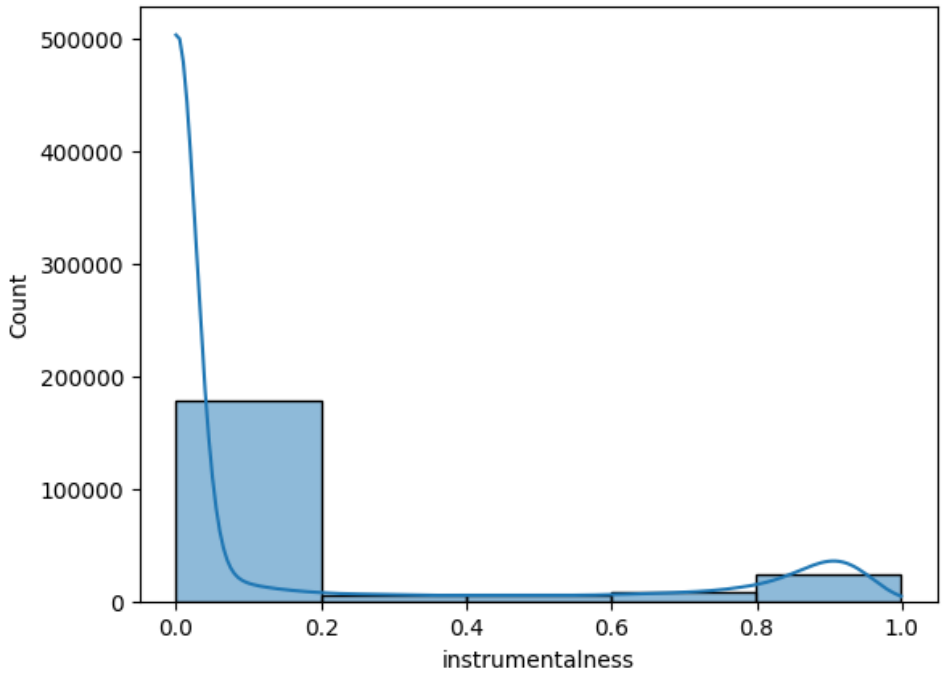
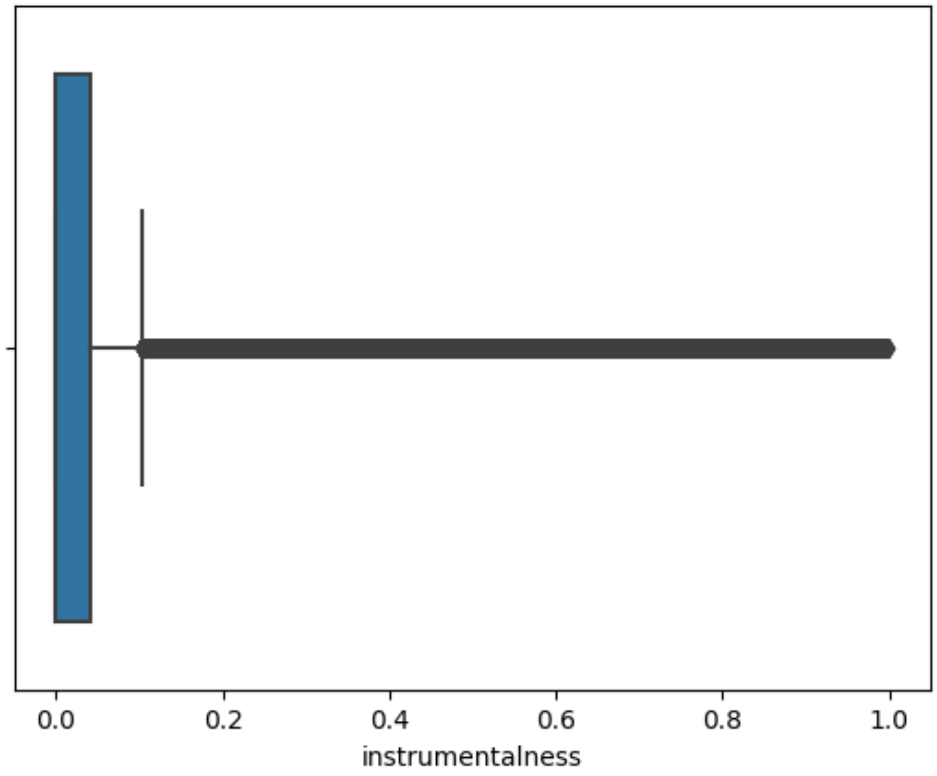


Figura 4.33. Diagrama de cajas y bigotes de instrumentalness



Datos estadísticos

Promedio: 0.152236

Desviación estándar: 0.306533

Mínimo: 0.0

25%: 0.0

50%: 0.000044

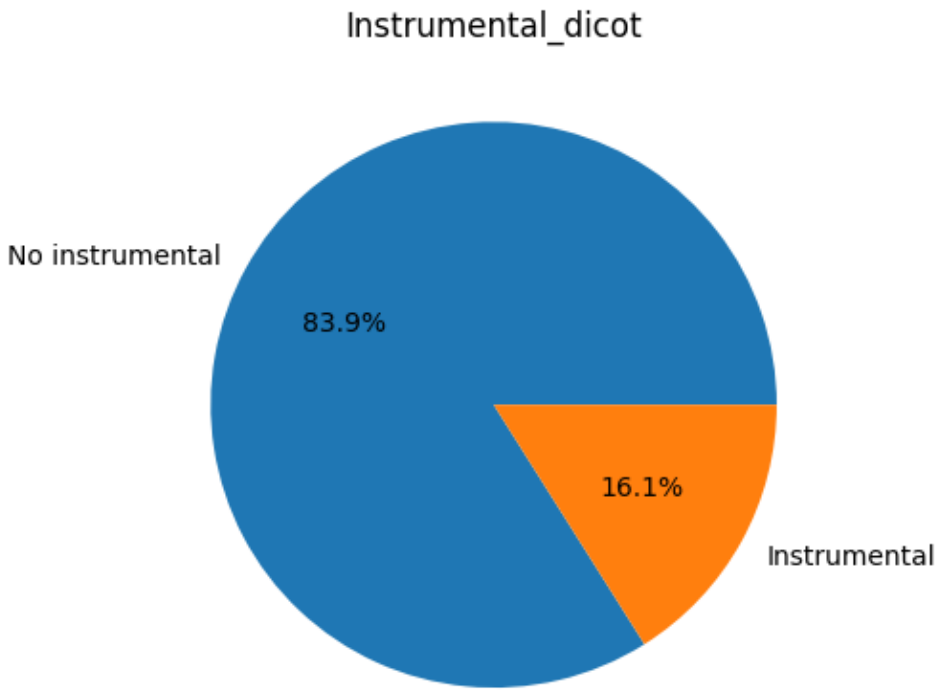
75%: 0.0415

Máximo: 0.999

Los valores de instrumentalness detectan si una canción es de exclusivamente de contenido instrumental conforme se acerca al 1 y conforme se acerca al 0, es porque lleva contenido vocal. De forma más específica, los valores mayores a 0.5 buscan representar canciones instrumentales.

En nuestra muestra, la gran mayoría de observaciones se interpretan como de contenido no instrumental, ya que se acercan demasiado al 0. Sin embargo, se detecta una gran cantidad de outliers en el diagrama de cajas, y que la distribución es bimodal en el histograma. Para tratarlo, creamos una nueva columna dicotómica "instrumental_dicot" donde True significa que la canción es instrumental.

Figura 4.34. Gráfica de pastel de instrumental_dicot



Tempo

Figura 4.35. Histograma de tempo

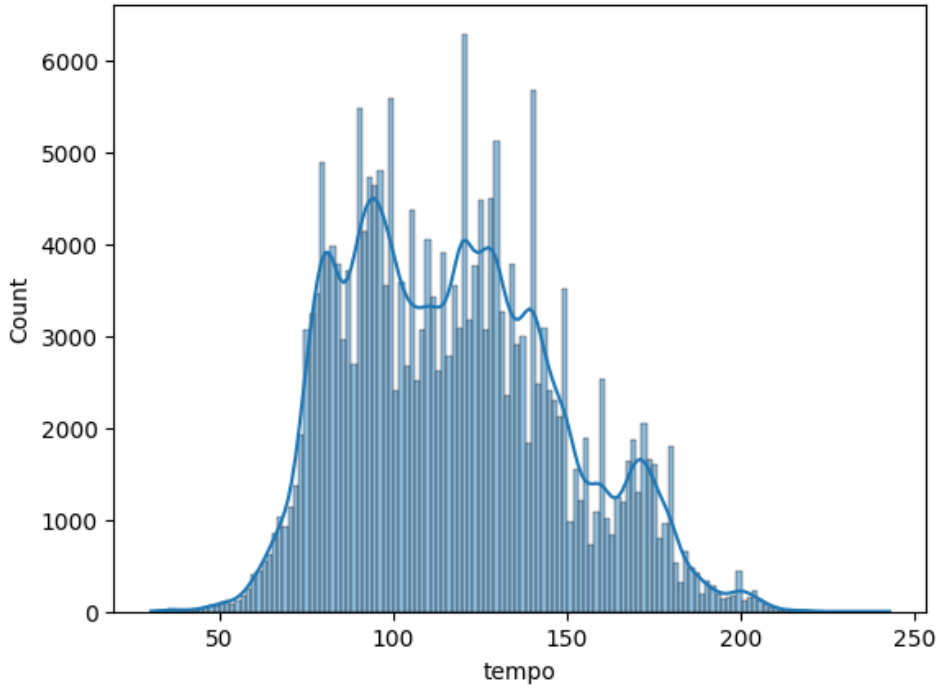
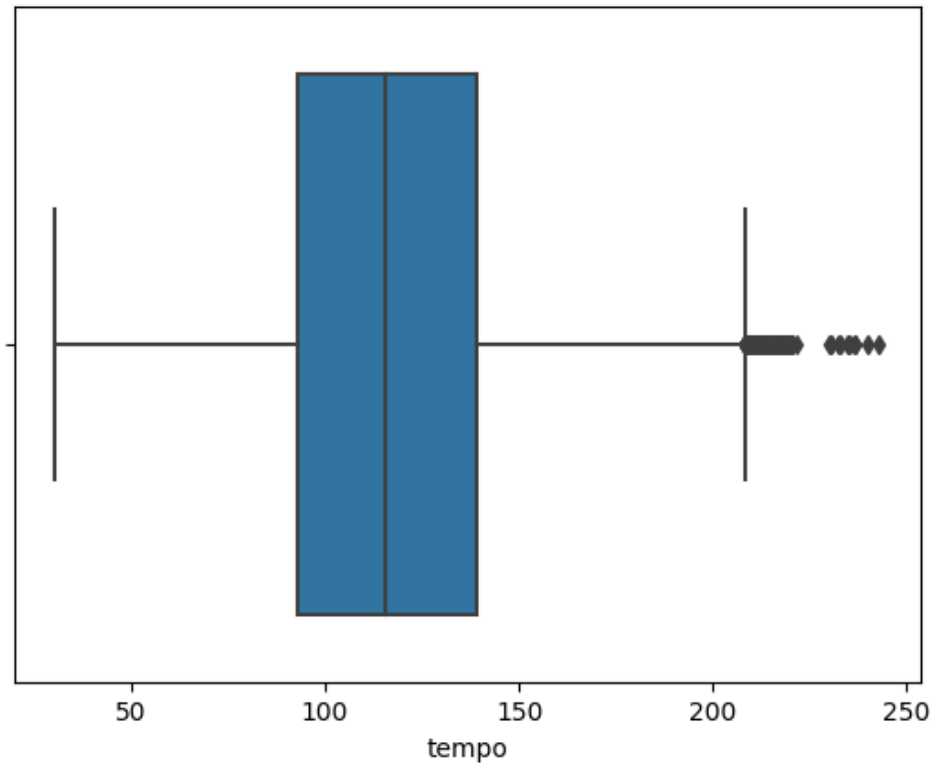


Figura 4.36. Diagrama de cajas y bigotes de tempo



Datos estadísticos

Promedio: 117.488022

Desviación estándar: 30.924366

Mínimo: 30.379

25%: 92.691

50%: 115.404

75%: 138.864

Máximo: 242.903

La mayoría de los datos obtenidos de esta columna se encuentran cercanos al promedio, a excepción de ciertas observaciones que, al ser mucho mayores, se encuentran alejados del resto. Estos valores más altos se ven reflejados como valores atípicos en el diagrama de cajas y bigotes (Figura 4.36).

4.1.4.1 Valores atípicos

Figura 4.37. Histograma con transformación log10 de la columna speechiness

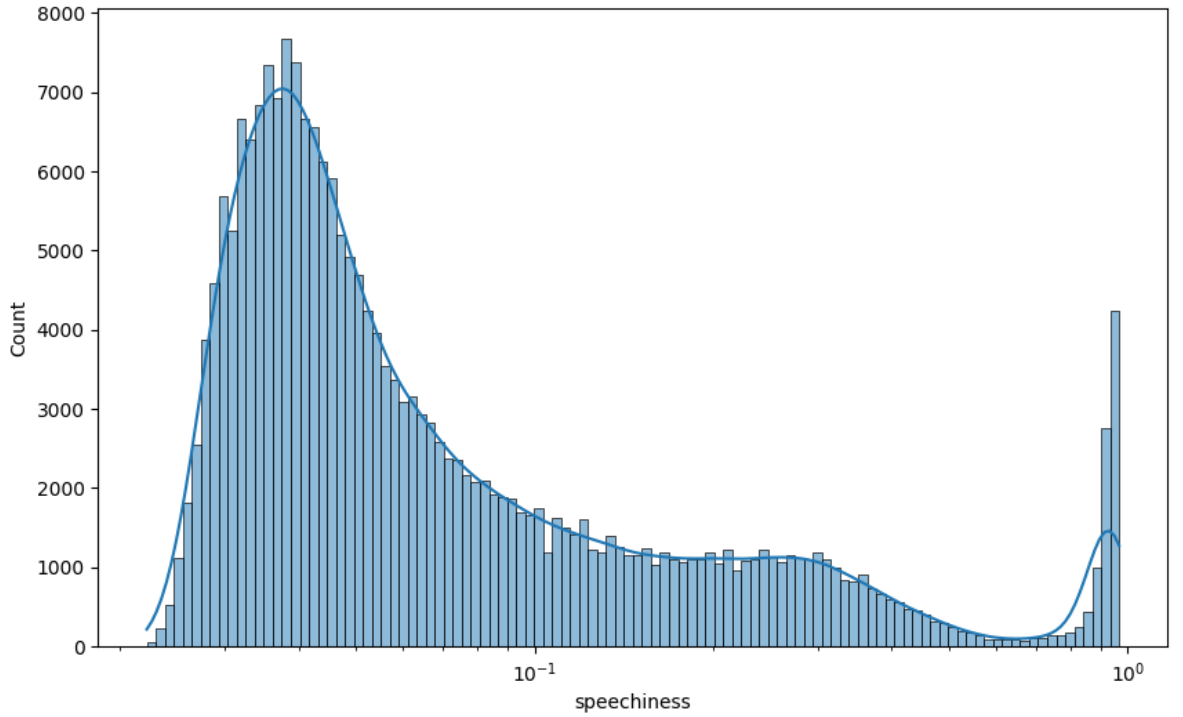


Figura 4.38. boxplot con transformación log10 de la columna speechiness

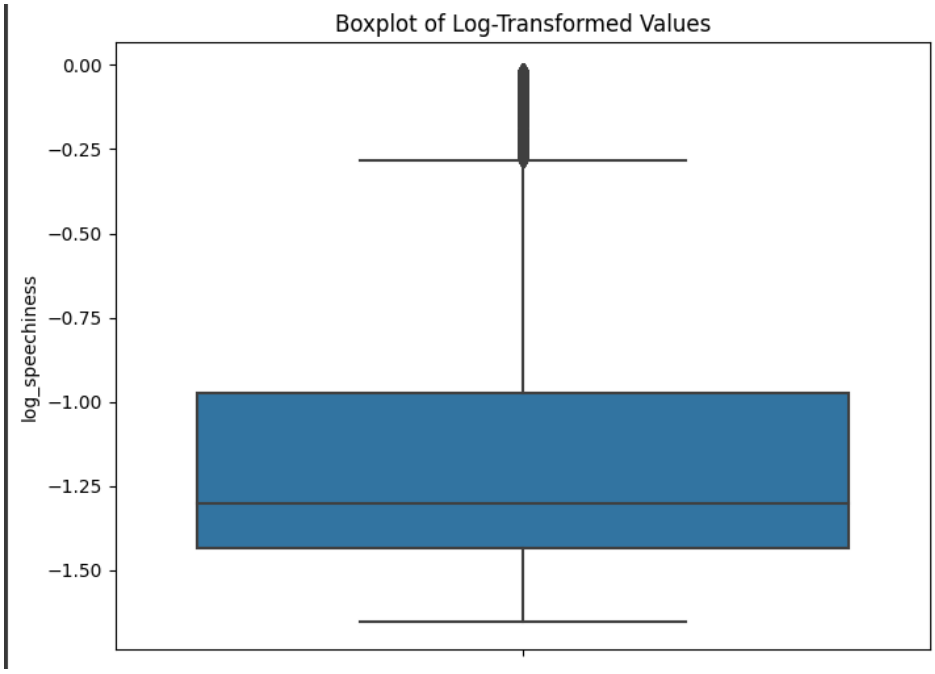


Figura 4.39. Histograma de instrumentalness sin valores atípicos

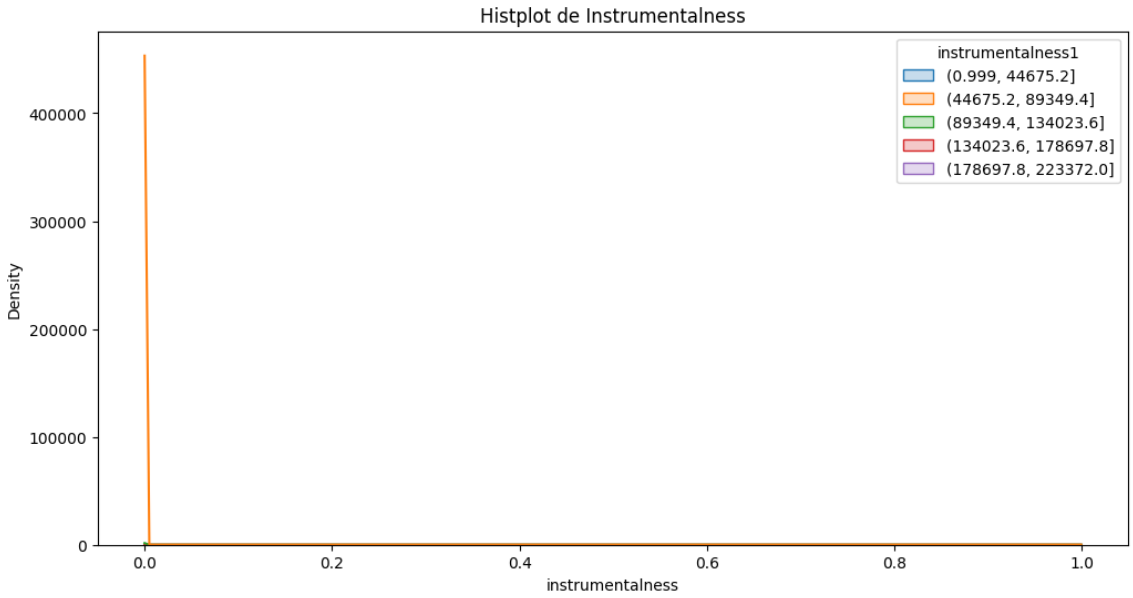
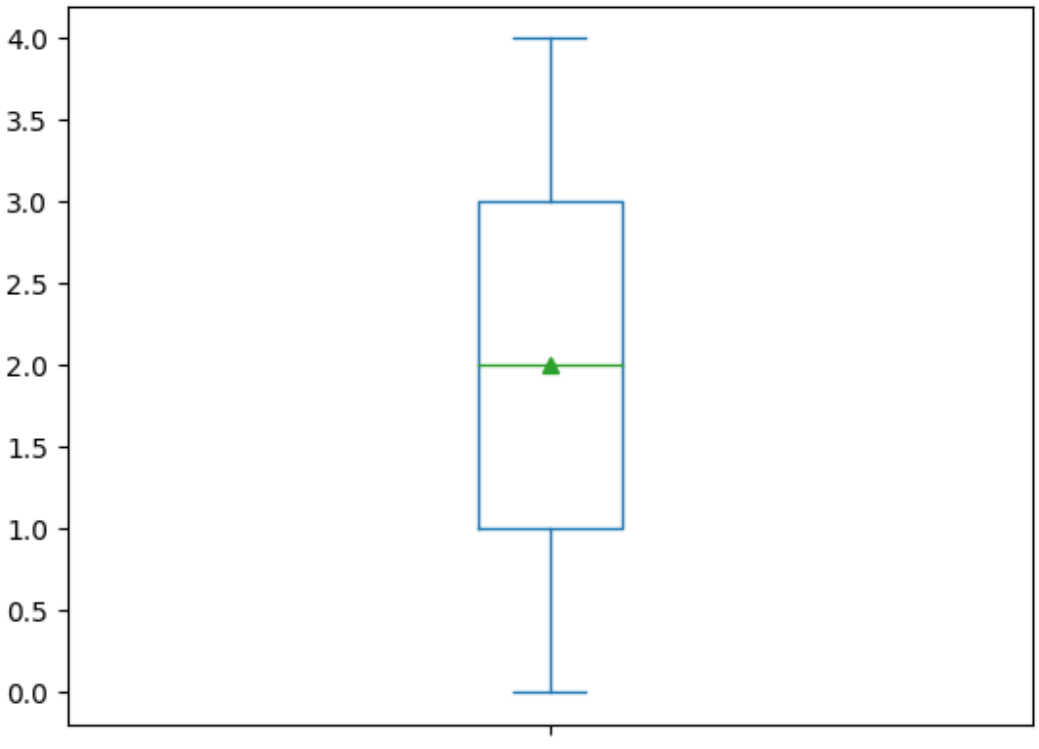


Figura 4.40. Diagrama de cajas y bigotes de instrumentalness sin valores atípicos



La gráfica con los valores de instrumentalness se encuentran extremadamente cargados hacia la izquierda ya que las observaciones de la muestra, casi en su totalidad, presentaron valores muy cercanos a 0.

Figura 4.41. Histograma de tempo sin valores atípicos

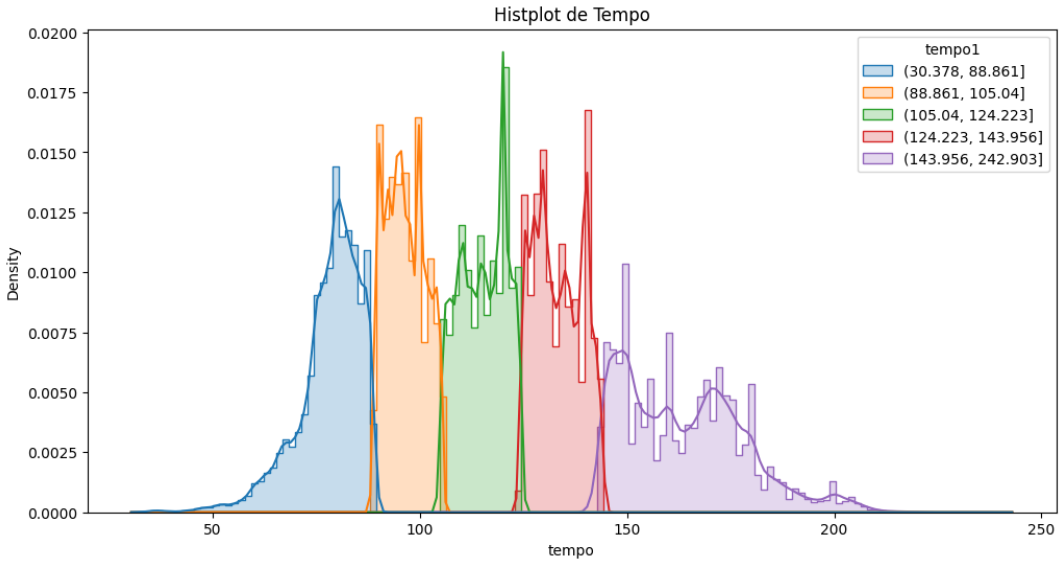
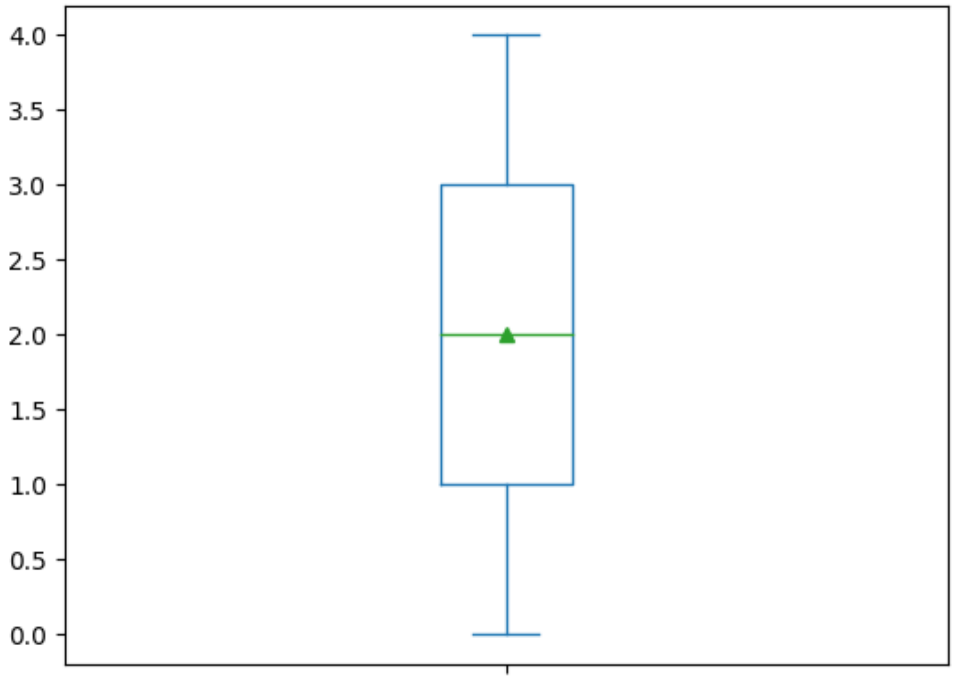


Figura 4.42. Diagrama de cajas y bigotes de tempo sin valores atípicos



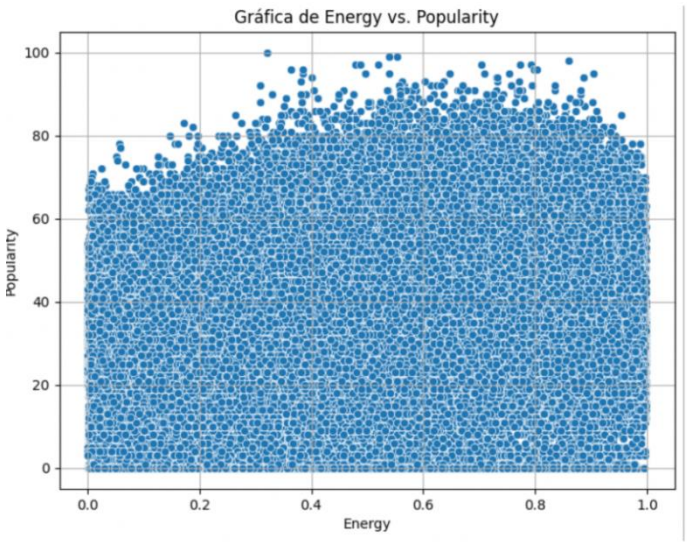
Los datos de la columna tempo se encuentran agrupados en el centro, sin embargo, presentan varios picos entre cada uno de los rangos que forman la gráfica. No sigue una distribución normal.

4.2 Análisis bivariado descriptivo y relacional

El análisis bivariado fue realizado con el propósito de responder las preguntas de investigación, las cuales son las siguientes:

¿Existe alguna relación entre la popularidad de una canción y su energía?

A esta pregunta obtuvimos una respuesta un tanto ambigua, realmente nos resultaría difícil establecer una relación dada la información obtenida, pero podemos observar el siguiente grafico



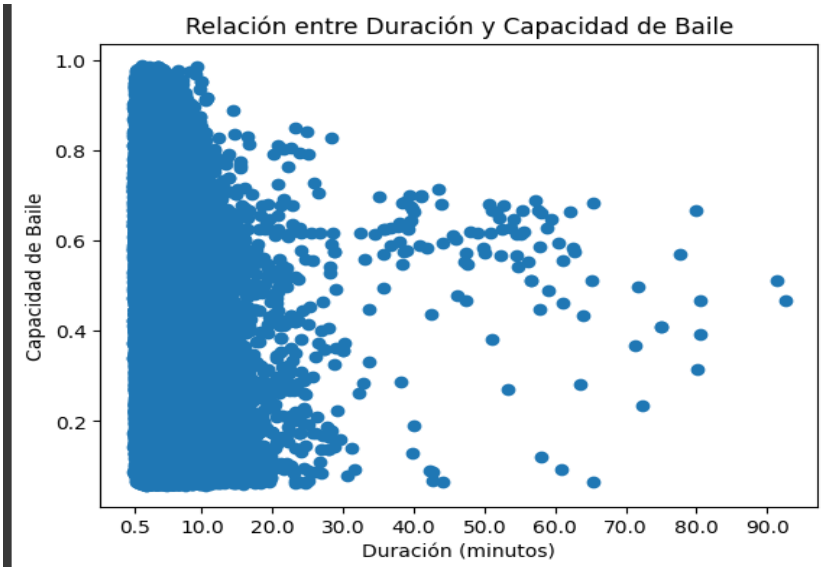
popularity_quartiles	Q1	Q2	Q3	Q4
energy_quartiles				
Q1	23103	17119	9803	5842
Q2	10456	13253	15927	16424
Q3	9041	12584	15958	18070
Q4	13775	14544	13688	13785

Figura 4.43. Grafica de energy con popularity y división por cuartiles

¿Cuál es la relación entre la duración de una canción y su capacidad de baile?

En este análisis, se adoptó un enfoque diferente para examinar la situación. A pesar de que inicialmente no parecía haber una relación clara entre las variables, una vez que se llevó a cabo un análisis detallado en diferentes cuartiles, se reveló un patrón significativo. Este análisis más refinado nos proporcionó una visión más precisa y matizada de la dinámica en juego. Contrario a las apariencias iniciales, emergió la evidencia de una relación inversa débil entre las variables en cuestión. Al desglosar los datos en diversos cuartiles, se pudo observar que, en términos generales, existe una tendencia clara: a medida que la duración disminuye, la capacidad de baile tiende a aumentar. Este hallazgo sugiere que la relación entre estos dos elementos no es directa ni uniforme, sino que está influenciada por matices que se manifiestan en diferentes rangos de duración. Este tipo de análisis detallado es esencial para desentrañar patrones sutiles que podrían pasar desapercibidos en una evaluación superficial. Estos resultados proporcionan una base sólida para comprender mejor la naturaleza de la relación entre la duración y la capacidad de baile, permitiendo

así tomar decisiones más informadas y diseñar estrategias más efectivas en función de estos hallazgos reveladores.




```
Coeficiente de correlación de Pearson entre Q1 de duration y Q1 de danceability: 0.0036800858540580648
Coeficiente de correlación de Pearson entre Q2 de duration y Q2 de danceability: -0.014464791338199463
Coeficiente de correlación de Pearson entre Q3 de duration y Q3 de danceability: 0.007918834870180448
Coeficiente de correlación de Pearson entre Q4 de duration y Q4 de danceability: -0.0025665426364911417
Coeficiente de correlación de Pearson entre Q5 de duration y Q5 de danceability: -0.012085153017723392
```

Figura 4.44. Gráfica de relación entre duración en minutos y danceability. Y coeficientes de correlación por cuartiles

Realicé un análisis bivariado utilizando dos variables continuas, para lo cual fue necesario emplear el coeficiente de correlación de Pearson. Obtuvimos un valor de -0.12, indicando que no existe una relación lineal fuerte entre la duración de una canción y su capacidad de ser bailable. Este resultado sugiere que, a medida que aumenta la duración, disminuye la capacidad de baile, evidenciando una relación inversa débil, ya que el incremento en una variable se asocia con la disminución de la otra.

Cabe destacar que la duración de las canciones se transformó a minutos, oscilando entre 0.26 minutos como la duración más corta y 92.5 minutos como la más extensa. Este rango de duración nos permitió dar respuesta a nuestra pregunta de análisis: "¿Cuál es la relación entre la duración de una canción y su capacidad de baile?". En consecuencia, podemos afirmar que, aunque no existe una relación lineal fuerte, sí hay una relación inversa débil entre ambas variables.

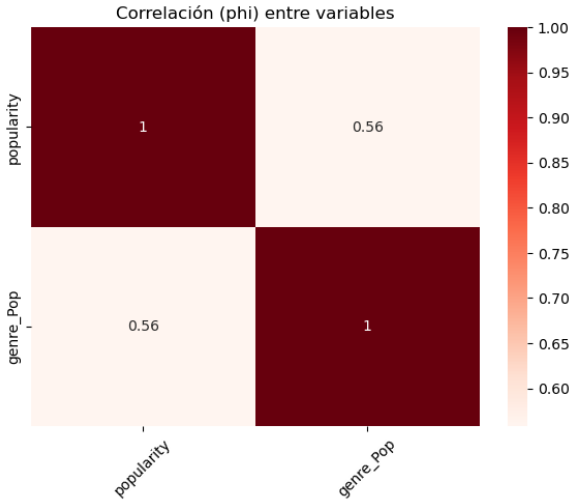


Figura 4.45. Heatmap de genre_pop y Popularity

¿Hay una relación entre la popularidad de una canción y el género pop

Esta respuesta implicó un proceso de investigación exhaustivo, ya que, al analizar inicialmente el gráfico, no resultaba evidente de inmediato si existía una relación lineal o inversa, las cuales son las relaciones más comunes en este tipo de análisis. Después de una investigación más detallada, se llegó a la conclusión de que la naturaleza de la relación es, de hecho, asociativa. Esta revelación subraya la complejidad y la sutileza de las relaciones presentes en los datos. Mientras que las relaciones lineales e inversas son típicamente más visibles, la relación asociativa implica una conexión más compleja y no sigue un patrón uniforme. En este contexto, la investigación desempeñó un papel crucial para desentrañar la verdadera naturaleza de la relación, destacando la necesidad de abordar las complejidades inherentes a través de métodos más avanzados y un análisis más profundo.

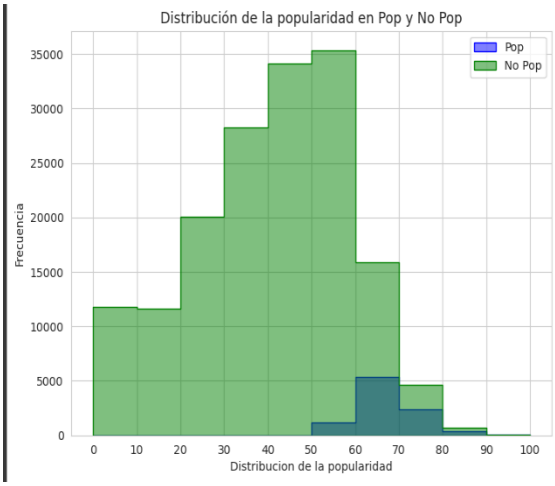


Figura 4.45B. Distribución de popularidad dividido entre canciones del género pop y no pop

¿Existe una relación lineal entre la valencia y la bailabilidad de las canciones de rock?

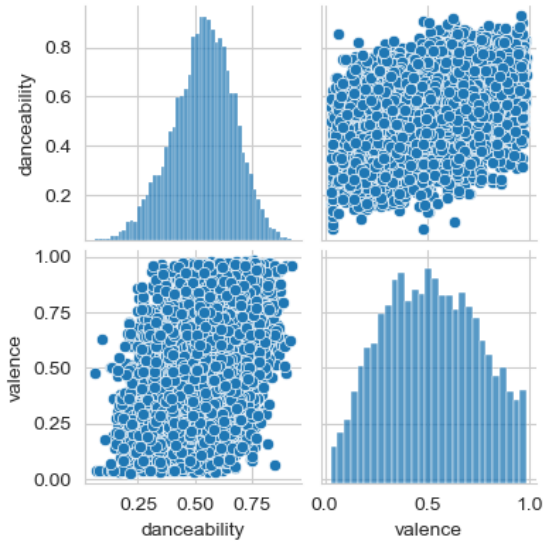


Figura 4.46. Pairplot entre las variables danceability y valence para las canciones del género rock

Observando la anterior gráfica, nos dimos cuenta de que podríamos modelar esta relación como una relación lineal positiva, con la cual se pueda describir la bailabilidad a partir de la valencia.

OLS Regression Results

Dep. Variable:	danceability	**R-squared:**	0.202
Model:	OLS	**Adj. R-squared:**	0.202
Method:	Least Squares	**F-statistic:**	2349.
Date:	Mon, 27 Nov 2023	**Prob (F-statistic):**	0.00
Time:	20:41:14	**Log-Likelihood:**	6569.4
No. Observations:	9272	**AIC:**	-1.313e+04
Df Residuals:	9270	**BIC:**	-1.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4041	0.003	133.262	0.000	0.398	0.410
valence	0.2595	0.005	48.463	0.000	0.249	0.270

Omnibus:	71.939	**Durbin-Watson:**	2.027
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	73.510
Skew:	-0.218	**Prob(JB):**	1.09e-16
Kurtosis:	2.994	**Cond. No.**	5.53

Figura 4.47. Summary del modelo creado con mínimos cuadrados ordinarios

Utilizamos el método ols de la librería statsmodels y obtuvimos los siguientes hallazgos:

- El estadístico de R cuadrado tuvo un valor bajo de 0.202, por lo que el modelo solo es apto para esta proporción de las observaciones

- La hipótesis nula de la prueba de Jarque-Bera propone que los residuos siguen una distribución normal. Sin embargo, con el valor menor a 0.05 en Prob(JB), dicha hipótesis se rechaza y concluimos que los residuos no son normales.

- El valor del estadístico Durbin-Watson se encuentra muy cerca de 2, por lo que podemos decir que no hay autocorrelación entre las variables

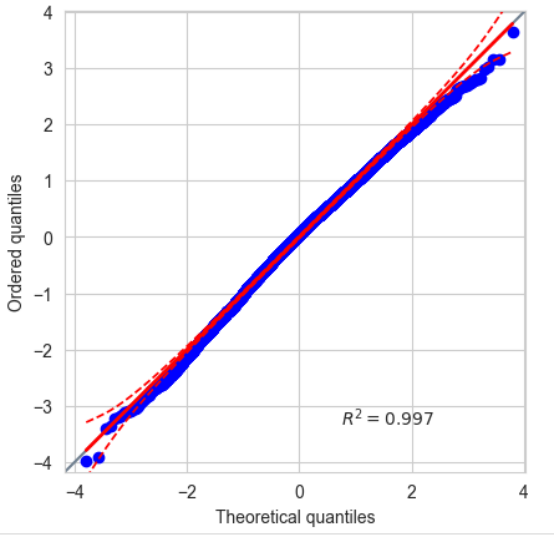


Figura 4.48. Diagrama qqplot de los residuos del modelo. Las observaciones se salen del rango permitido para interpretarlos como normales

Como conclusión, podemos decir que, a pesar de que existe una relación entre danceability y valence, el modelo obtenido no es viable para describir esta relación con la presición necesaria. Por esta razón, decidimos no continuar con este par de variables para diseñar un modelo predictivo.

4.2.1 Entre dos variables categóricas (nominales u ordinales)

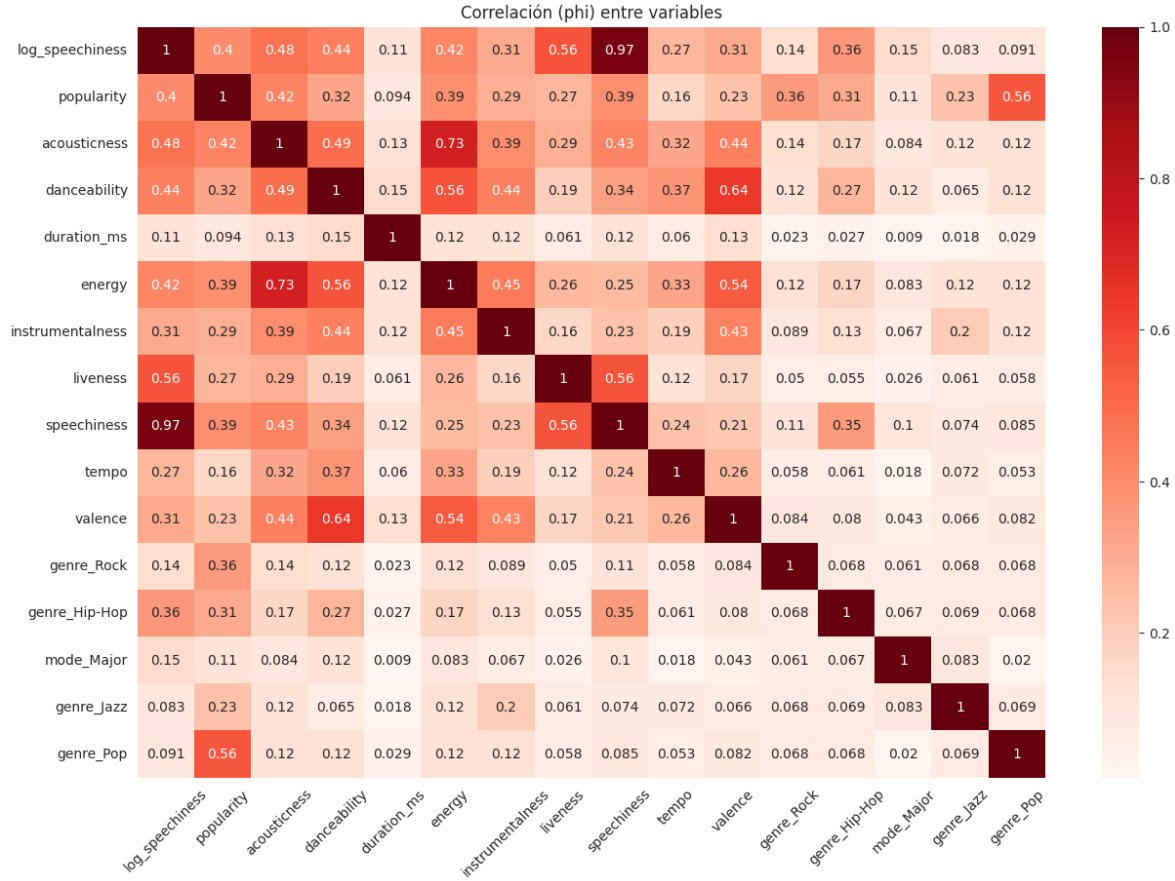


Figura 4.49. Heatmap de correlación entre variables

En este heatmap, se evidencian algunas relaciones notables que saltan a la vista, como la fuerte conexión entre la popularidad y el género pop. Esta asociación clara podría indicar una tendencia significativa en la preferencia del público para este género en particular. Sin embargo, llama la atención una curiosidad: la falta de relación aparente con la duración en milisegundos (duration_ms). Resulta intrigante observar que la duración de la canción no presenta correlación discernible con ninguna de las otras variables representadas en el heatmap. Este hallazgo puede llevarnos a considerar que la duración de una canción podría ser un factor independiente en términos de su relación con otros atributos, o podría sugerir la presencia de influencias externas que afectan la duración de manera más aleatoria., podemos observar más relaciones, pero son bastantes para resaltarlas individualmente

4.2.2 Entre una categórica (nominal u ordinal) y una numérica (de intervalo o razón)

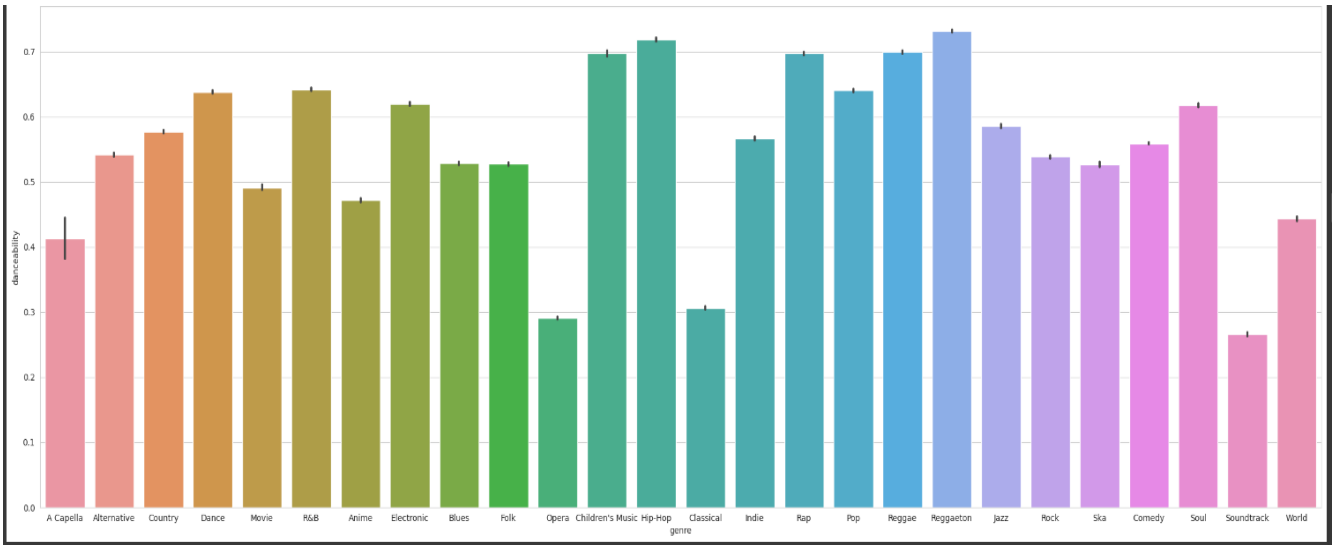
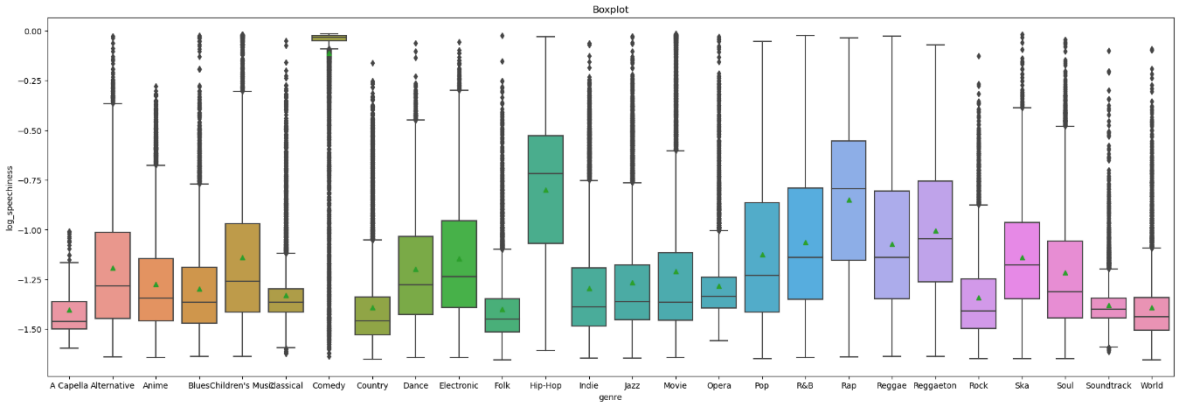


Figura 4.50. Gráfico de distribución de danceability por cada género

En este gráfico, se llevó a cabo una comparación entre los géneros musicales y la bailabilidad, y los resultados son reveladores. Notamos que los géneros como hip-hop y reguetón destacan como aquellos que tienden a tener canciones altamente bailables. Esta observación concuerda intuitivamente con la naturaleza rítmica y enfocada en el baile inherente a estos estilos musicales, respaldando la coherencia entre la percepción subjetiva y los datos cuantitativos. Por otro lado, resulta lógico encontrar que géneros como clásica o soundtrack se sitúan entre los que poseen menor bailabilidad. Estos géneros suelen caracterizarse por composiciones más complejas, con enfoque en la expresión artística y narrativa, en lugar de la accesibilidad para el baile.

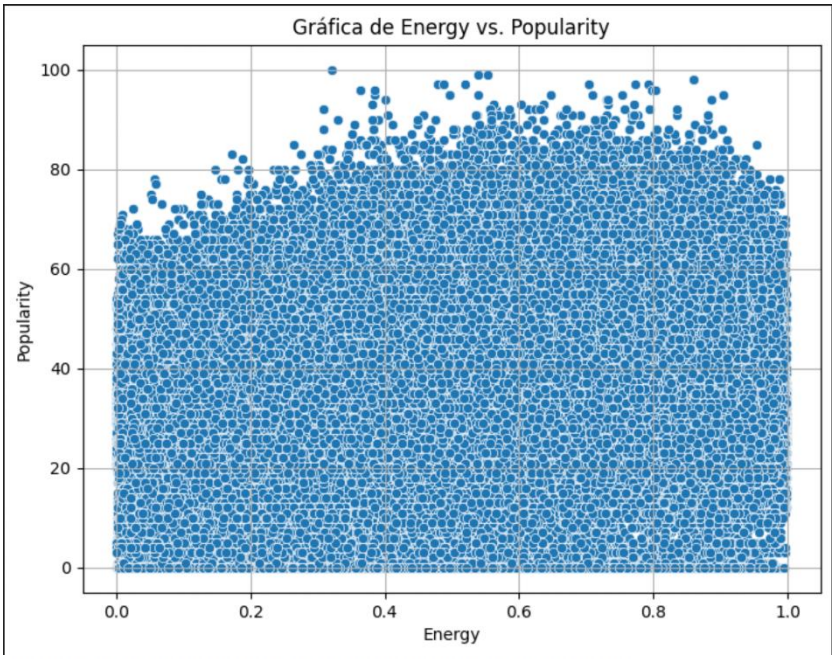
Figura 4.50B Boxplot para cada género de log_speechiness



La speechiness probablemente guarda algún tipo de relación con los géneros musicales. Si revisamos los boxplots para cada género, algunos tienen evidentemente mayor speechiness que otros. Evidentemente los 3 géneros con mayor speechiness son Comedy (que ni siquiera es música), Rap y HipHop.

Entre los menos hablados está la acapella, el country (sorpresivamente), la música de películas (soundtrack) y la clásica.

4.2.3 Entre dos numéricas (de intervalo o razón)



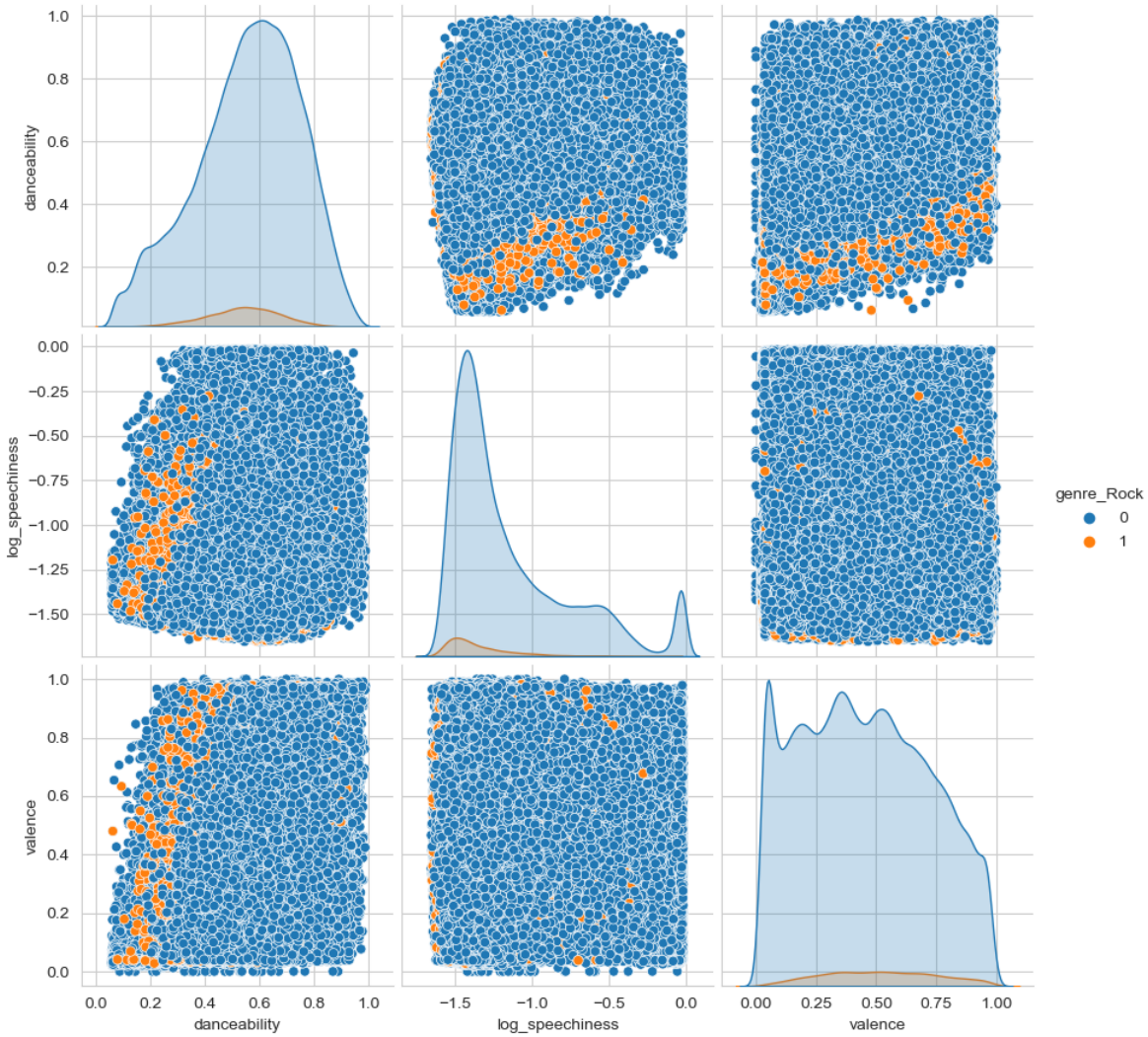
popularity_quartiles	Q1	Q2	Q3	Q4
energy_quartiles				
Q1	23183	17119	9803	5842
Q2	18456	13253	15927	16424
Q3	9041	12584	15958	18070
Q4	13775	14544	13688	13785

Figura 4.51. Gráfica de dispersión entre energy y popularity y división intercuartil

En este análisis, se enfrenta el desafío de lidiar con una muestra significativamente extensa, superando las 200,000 entradas. Esta amplitud de datos puede complicar la interpretación y limitar la capacidad de análisis detallado. La complejidad inherente a una muestra tan grande a menudo dificulta la obtención de conclusiones claras y detalladas. Dada esta situación, se optó por utilizar el análisis intercuartílico como una herramienta para abordar la vastedad de los datos. Esta metodología se centra en las medidas estadísticas que se encuentran entre los cuartiles, permitiendo identificar patrones generales sin sumergirse en detalles específicos que podrían perderse en la inmensidad del conjunto de datos.

4.3 Otros análisis exploratorios multivariados (opcional)

Figura 4.52 Pairplot de comparación de variables dividido por canciones que son o no de Rock



Aquí podemos observar el comportamiento del modo en valencia, speechiness y bailabilidad, siendo la aparente relación lineal entre valencia y danzabilidad del género rock, la que nos llevó a formular la pregunta ¿Existe una relación lineal entre valencia y danzabilidad para canciones del género rock?.

4.4 Conclusiones del capítulo

La conclusión principal que derivamos de este análisis es la importancia de reconsiderar la elección de trabajar con muestras tan extensas. Nos dimos cuenta de que muchos de nuestros análisis quedaban prácticamente obsoletos o resultaban ininteligibles debido a la naturaleza errática del comportamiento de los datos cuando se manejan conjuntos tan grandes.

Este fenómeno añade una capa adicional de complejidad que complica significativamente diversos enfoques, como el análisis de duración frente a bailabilidad o la relación entre energía y popularidad. Aunque puede existir un patrón claro en estos casos, la cantidad masiva de muestras obstaculiza la capacidad de extraer conclusiones significativas.

Una conclusión adicional que podemos extraer es que, aunque es posible la obtención de datos a partir de muestras extensas, se requiere un mayor nivel de conocimiento en áreas como probabilidad y estadística, así como habilidades avanzadas en análisis de datos.

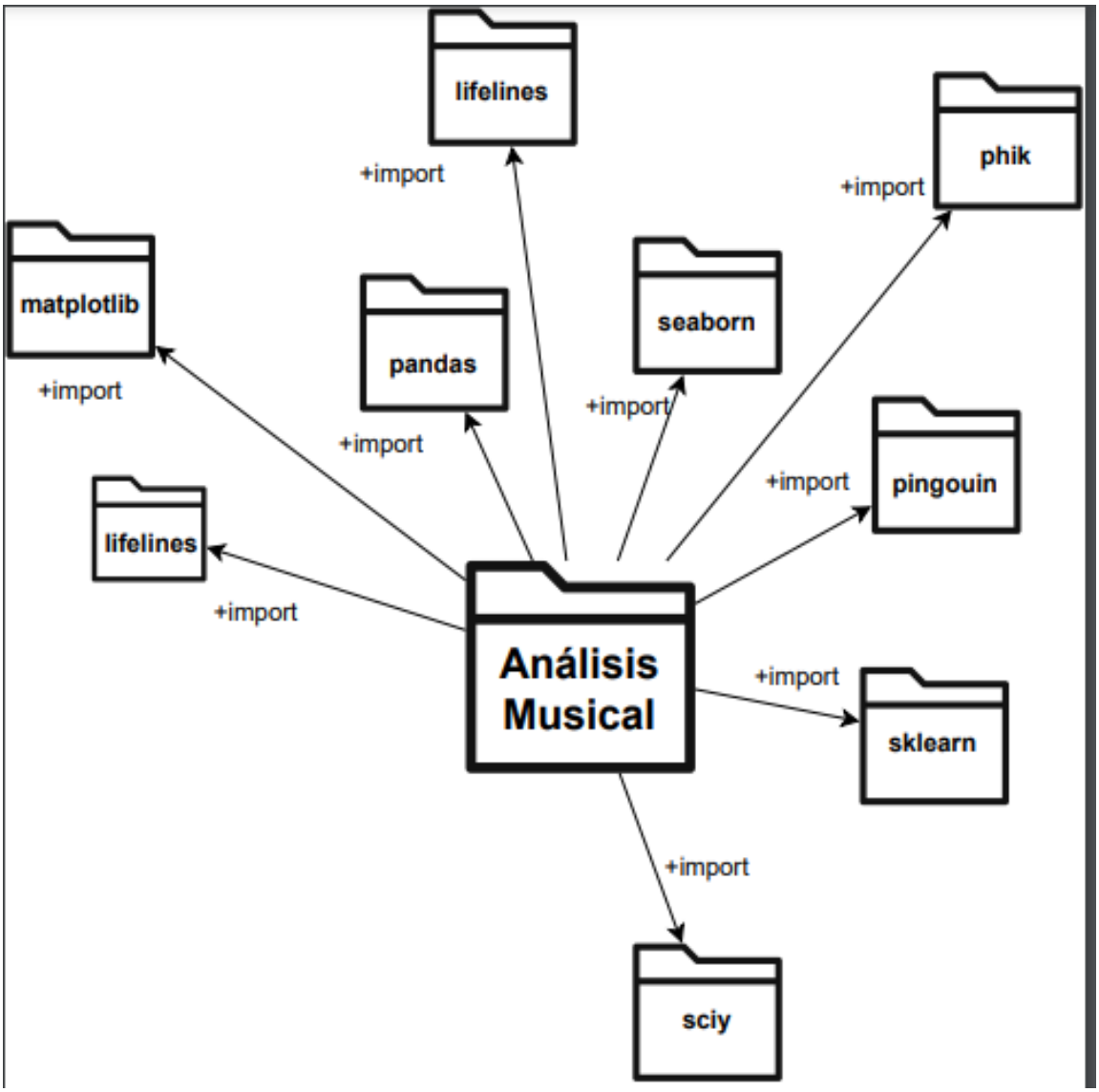
Durante el proceso, nos encontramos con situaciones en las que fue necesario recurrir al apoyo de la maestra o realizar búsquedas exhaustivas en internet para determinar ciertas relaciones entre variables. Este hallazgo subraya la necesidad de un enfoque interdisciplinario y un continuo desarrollo de habilidades para afrontar desafíos en el manejo de grandes conjuntos de datos.

A pesar de las dificultades encontradas, la experiencia resultó ser fascinante y educativa. Nos llevó a reconocer la importancia de una gestión cuidadosa y reflexiva de grandes cantidades de datos para obtener conclusiones significativas en el ámbito del análisis musical. Esta reflexión profunda resalta la naturaleza dinámica y desafiante de la investigación basada en datos, alentándonos a explorar nuevas estrategias y enfoques en futuros proyectos analíticos.

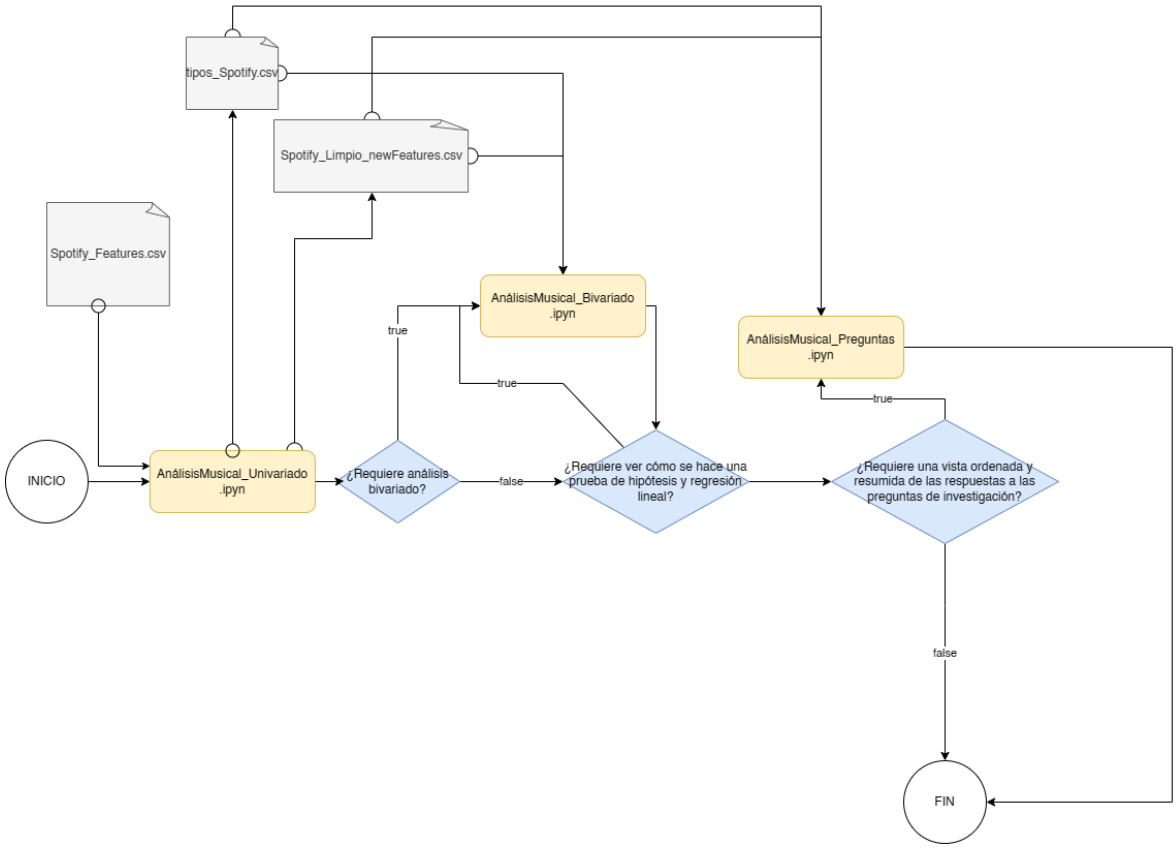
5 Implementación

5.1 Diagramas de paquetes de UML

Figura 5-1 Diagrama de paquetes UML



5.2 Diagrama de flujo de notebooks



5.3 Conclusiones del capítulo

En el transcurso de este capítulo, nos encontramos en la necesidad de recurrir a la maestra en varias ocasiones para obtener orientación sobre las mejores bibliotecas a utilizar en nuestro proyecto. Inicialmente, teníamos la percepción de que la cantidad de bibliotecas disponibles sería considerablemente mayor de lo que finalmente se utilizó.

Este descubrimiento resaltó la importancia de aprovechar los conocimientos adquiridos durante las clases para discernir y seleccionar las bibliotecas más adecuadas para abordar eficazmente nuestro proyecto. La orientación proporcionada por la maestra fue invaluable en el proceso de navegación a través de las bibliotecas disponibles, permitiéndonos centrarnos en aquellas que mejor se alineaban con los objetivos y requisitos específicos de nuestro análisis musical.

Este ejercicio no solo fortaleció nuestra comprensión de las herramientas disponibles, sino que también destacó la importancia de la selección estratégica de recursos para maximizar la eficiencia y la calidad del trabajo. A medida que avanzamos en el proyecto, pudimos aplicar de manera efectiva los conocimientos adquiridos en clase para trabajar con las bibliotecas seleccionadas de la mejor manera posible.

Este proceso de aplicación práctica no solo consolidó nuestras habilidades técnicas, sino que también subrayó la importancia de una toma de decisiones informada y estratégica al elegir las herramientas adecuadas para abordar desafíos específicos en el análisis de datos musicales.

En última instancia, este capítulo no solo fue una lección sobre la variedad de recursos disponibles, sino también sobre la importancia de la adaptabilidad y la toma de decisiones fundamentada en el contexto del proyecto.

6 Conclusiones

En el estado actual de la investigación, si se decide continuar con el proyecto, sugerimos considerar algunas recomendaciones clave. En primer lugar, sería beneficioso samplear la muestra con la que se trabaja, ya que más de 200,000 muestras pueden complicar la obtención de respuestas estadísticas significativas. Además, proponemos utilizar gráficos distintos, ya que los de dispersión pueden no ser los más adecuados cuando se manejan conjuntos de datos tan extensos. Por último, alentaríamos a indagar con diferentes preguntas, ya que aún es posible recabar más información valiosa.

En otro ámbito, durante el desarrollo del proyecto, nos encontramos en la necesidad de buscar orientación de la maestra en varias ocasiones para seleccionar las bibliotecas más apropiadas. Inicialmente, esperábamos enfrentarnos a una gama más amplia de bibliotecas, pero este descubrimiento resaltó la importancia de utilizar los conocimientos adquiridos en clase para discernir y elegir las herramientas más efectivas para nuestro análisis musical.

La guía proporcionada por la maestra fue crucial en la navegación por las opciones disponibles, permitiéndonos centrarnos en aquellas que mejor se alineaban con nuestros objetivos. Este proceso no solo fortaleció nuestra comprensión de las herramientas, sino que también resaltó la importancia de la selección estratégica de recursos para maximizar la eficiencia y calidad del trabajo. A medida que avanzamos, pudimos aplicar con éxito los conocimientos adquiridos, consolidando nuestras habilidades técnicas y destacando la necesidad de decisiones informadas y estratégicas en la selección de herramientas para abordar desafíos específicos.

Observamos que la cantidad masiva de datos puede hacer que muchos análisis sean obsoletos o ininteligibles debido a la naturaleza errática de los datos. También reconocimos que, a pesar de la posibilidad de obtener datos a partir de muestras extensas, se necesita un mayor nivel de conocimiento en probabilidad, estadística y análisis de datos.

A pesar de las dificultades encontradas, la experiencia fue fascinante y educativa, esta reflexión destaca la naturaleza dinámica y desafiante de la investigación basada en datos, alentándonos a explorar nuevas estrategias y enfoques en futuros proyectos analíticos.

Contestando a la pregunta de: ***¿Cree que con lo que ha desarrollado en su proyecto ha cumplido con los objetivos de un Análisis de Datos Exploratorio?***

Sí, considero que he logrado cumplir con los objetivos establecidos para un Análisis de Datos Exploratorio (ADE) de manera satisfactoria. A continuación, destacaré algunos aspectos clave:

- **Aplicación de Conocimientos de Programación:**

Se han utilizado eficientemente herramientas de programación en Python, incluyendo bibliotecas como numpy, pandas, matplotlib, seaborn, statsmodels y scikit-learn. Estas herramientas han facilitado la manipulación de datos, la visualización y la aplicación de modelos estadísticos.

- **Recopilación y Estructuración del Dataset:**

Se realizó una recopilación de datos exhaustiva y se estructuró un dataset que permitió abordar las preguntas de investigación de manera adecuada. Se llevó a cabo una limpieza efectiva de observaciones incorrectas y transformaciones necesarias para facilitar la interpretación gráfica.

- **Aplicación de Conocimientos de Probabilidades y Estadísticas:**

Durante la etapa de exploración y preparación de datos, se aplicaron sólidos conocimientos en Probabilidades y Estadísticas. Esto incluyó análisis descriptivo, formulación y pruebas de hipótesis, así como análisis de regresión con fines explicativos, proporcionando una base estadística robusta.

- **Visualización de Datos:**

Se llevaron a cabo visualizaciones efectivas durante la exploración de datos. Estas visualizaciones fueron fundamentales para formular y evaluar hipótesis, brindando una comprensión más profunda de las relaciones entre las variables.

- **Documentación y Comunicación de Resultados:**

Los resultados del proceso de análisis exploratorio se documentaron de manera exhaustiva. Se enfatizó la redacción de hallazgos y conclusiones, facilitando la comprensión para audiencias no técnicas.

7 Referencias

<Instale Mendeley (gratuito) en su navegador, así como la versión desktop. Asegúrese de que su Word tenga el plugin para Mendeley.

Inserte referencias bibliográficas (libros, artículos, materiales de clase, sitios web, tutoriales, etc.) y un apartado como este de Referencias. Use APA o IEEE>

BPMN Specification - Business Process Model and Notation. (n.d.). Retrieved June 24, 2022, from https://www.bpmn.org/

Debuse, J. C. W., de la Iglesia, B., Howard, C. M., & Rayward-Smith, V. J. (2001). Building the KDD Roadmap. In *Industrial Knowledge Management* (pp. 179–196). https://doi.org/10.1007/978-1-4471-0351-6_12

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science \& Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

KUMAR, A. (2022). A quick guide to . . . Bivariate correlation. Retrieved May 25, 2022, from https://www.analyticsvidhya.com/blog/2022/02/a-quick-guide-to-bivariate-analysis-in-python/

pandas development team, T. (2020). *pandas-dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.3509134

Patil, P. (2018). What is Exploratory Data Analysis? | by Prasad Patil | Towards Data Science. Retrieved May 25, 2022, from TowardsScience website: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Trochim, W. M. K. (2020). Structure of Research | Research Methods Knowledge Base. Retrieved May 24, 2022, from https://conjointly.com/kb/structure-of-research/

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

Wirth, R., & Wirth, R. (2000). CRISP-DM: Towards a standard process model for data mining. *PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON THE PRACTICAL APPLICATION OF KNOWLEDGE DISCOVERY AND DATA MINING*, 29--39.

8 Anexo

Tabla 8-1Operaciones válidas por tipo de dato

Operation	Nominal	Ordinal	Interval	Ratio
Equality	✔	✔	✔	✔
Order		✔	✔	✔
Add / subtract			✔	✔
Multiply / divide				✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Arithmetic mean			✔	✔
Geometric mean				✔

Fuente: https://matthewrenze.com/articles/the-four-subtypes-of-data-in-data-science/