Informe de Fundamentos de la Ciencia de Datos

“COVERS DE UNA DÉCADA ICÓNICA DEL SIGLO PASADO (Amazon Music)”

Integrantes

- Videla, Braian.
- Perna, Ignacio Agustin.

Introducción

Para comenzar, analizaremos de qué se trata nuestro conjunto de datos. Tenemos 980 muestras de canciones de toda la década del 70. extraídas de la plataforma **Amazon Music** con 17 variables que las evalúan. Inicialmente haremos un análisis exploratorio de los datos para comprender su estructura y características generales. Esto incluye identificar tipos de datos, valores faltantes o atípicos, registros repetidos, valores nulos, etc. También analizaremos la distribución de las variables mediante estadísticas descriptivas y visualizaciones. Este proceso nos permitirá detectar patrones, relaciones y problemas en los datos.

Posteriormente plantearemos posibles hipótesis generadas a partir del análisis exploratorio sobre variables que nos puedan interesar y por cómo se pueden estar relacionando. Utilizaremos pruebas adecuadas para contrastar las hipótesis planteadas, evaluando si los datos respaldan o refutan las afirmaciones y finalmente haremos una conclusión general de cada una de ellas.

Análisis exploratorio

Descripción de las variables

Track: el título de la canción.

Artist: el intérprete o grupo que grabó la canción.

Duration: la duración de la canción, medida en minutos y segundos.

Time Signature: la métrica musical de la canción, indica el número de pulsaciones por compás.

Danceability: una medida de qué tan adecuada es una pista para bailar, basada en el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general.

Energy: una medida de intensidad y actividad en la canción, donde los valores más altos indican una pista más enérgica.

Key: la tonalidad musical en la que está compuesta la canción, representada por un número entero.

Loudness: el volumen promedio de la canción, medido en decibelios(dB).

Mode: la modalidad de la pista, indica si la canción está en tono mayor o menor.

Speechiness: una medida de la presencia de palabras habladas en una pista, valores más altos indican cualidades más parecidas al habla.

Acousticness: una medida de la calidad acústica de la pista, valores más altos indican una probabilidad de ser acústica.

Instrumentalness: una medida que indica presencia de voces, valores más altos representan pistas más instrumentales.

Liveness: una medida de la probabilidad de que la pista se haya interpretado en vivo, valores más altos indican más ruido de audiencia.

Valence: una medida de la positividad musical de la pista, valores más altos indican música más positiva o alegre.

Tempo: la velocidad o ritmo de la pista, medida en pulsaciones por minuto(BPM).

Popularity: una puntuación que refleja la popularidad de la pista, generalmente basada en los recuentos de transmisiones y otras métricas.

Year: el año en el que se lanzó la canción.

Tipo de Datos

Con el objetivo de adentrarnos en el análisis exhaustivo de los datos, necesitamos entender cuáles son las características estructurales de nuestro dataset.

Del método info de pandas, sabemos que por lo menos a primera vista, no tenemos valores nulos en nuestro conjunto, y que tenemos un **11,76%** de datos que son Cualitativos, Categóricos y Nominales porque no tienen un orden definido: **Track** y **Artist**.

Vemos que la variable **Duration** está catalogada como lo que podría ser un dato cualitativo, pero la analizaremos más adelante. Por ahora lo tomaremos como en realidad debería ser: un dato cuantitativo.

Luego, tenemos que el **29,41%** son datos Cuantitativos, Numéricos y Discretos que representan cantidades o medidas que pueden contarse de manera específica y no toman valores intermedios: **Time Signature**, **Key**, **Mode**, **Popularity** y **Year**.

Y por último, tenemos un **58,82 %** de datos Cuantitativos, Numéricos y Continuos que representan cantidades o medidas que no pueden contarse fácilmente ya que pueden tomar valores como fracciones o decimales: **Danceability**, **Energy**, **Loudness**, **Speechiness**, **Acousticness**, **Instrumentalness**, **Liveness**, **Valence** y **Tempo**.

Descripción estadística de las variables

Usando el método describe() de Pandas, podemos observar en nuestras variables, detalles estadísticos como:

- **Count:** la cantidad de valores no nulos en la columna. En este caso, todos tienen 980 valores.
- **Mean**: el promedio de los valores en la columna.
- **Std**: Desviación estándar. Mide la dispersión de los datos alrededor de la media. Una mayor desviación indica mayor variabilidad en los datos.
- **Min**: el valor mínimo en la columna.
- **25%**: Primer cuartil. El valor por debajo del cual se encuentra el 25% de los datos, también conocido como el percentil 25.
- **50%**: Segundo cuartil. El valor central que separa la mitad superior e inferior de los datos. La mediana.
- **75%**: Tercer cuartil. El valor por debajo del cual se encuentra el 75% de los datos, también conocido como el percentil 75.
- **Max**: El valor máximo en la columna.

En dicho método podremos observar que las variables Track y Artist, al tener datos cualitativos nominales, la librería Pandas las elimina del análisis descriptivo porque se encarga de tomar datos estadísticos, lo que es imposible con este tipo de datos.

Time Signature: Sus valores se encuentran en un rango de 1 a 5 con un desvío estándar cercano a cero, con un primer cuartil ya alcanzando las 4 pulsaciones por compás indicando que la gran mayoría de nuestras canciones tienen una alta cantidad de pulsaciones por compás.

Danceability: Sus valores se encuentran en un rango de 0 a 1, con un desvío estándar bajo, y con un promedio por encima del 50% dejando a saber que nuestras canciones tienen en su mayoría una buena bailabilidad.

Energy: Sus valores se encuentran en un rango de 0 a 1, con un desvío estándar bajo. Vemos también que el valor mínimo de ésta variable es de una magnitud 116 veces menor que el promedio, indicando que tenemos por lo menos un outlier. Y con un promedio que nos dice que la gran mayoría de nuestros temas son enérgicos.

Key: Sus valores están en un rango de 0 a 12, con un desvío estándar medianamente elevado, lo que nos indica que los valores de esta variable están muy dispersos alrededor de la media, es decir, que existe una amplia variabilidad en los tonos musicales en el conjunto de canciones analizadas.

Loudness: Como es una medida en decibelios, su rango es de -100 a 0. En este caso, vemos que el 75% de los datos se encuentran más cercanos al límite superior teórico (donde los valores cercanos a 0 son las canciones más ruidosas), por lo que tenemos la mayoría de canciones con un nivel de ruido elevado.

Mode: Variable dicotómica donde sus posibles valores son 0 o 1 indicando si la canción está en tono mayor o menor.

Speechiness: El rango se mueve entre 0 y 1. Notamos que la media baja indica que la presencia de palabras habladas en nuestras muestras por lo general es baja, fijándonos también en el valor del tercer cuartil que es muy similar. Notamos también la presencia de algún outlier al tener un valor máximo 12 veces más grande que la media.

Acousticness: El rango se mueve entre 0 y 1. La media de 0.33 indica que, en promedio, hay una presencia acústica moderada en las pistas. Y teniendo un mínimo con valor 15 mil veces menor que la media, habrá outliers mínimos.

Instrumentalness: El rango va de 0 a 1. Vemos que la mayor cantidad de canciones se concentran en valores muy bajos y en su mayoría son ceros o cercanos al mismo, por lo tanto nuestro conjunto se caracteriza por tener canciones con poca presencia de instrumentales. La disparidad entre el valor máximo y el 75% nos indica la presencia de outliers.

Liveness: El rango va de 0 a 1. Tendremos la mayoría de los valores cercanos a 0, por lo que habrá poco ruido de audiencia, es decir, menor probabilidad de que se haya interpretado en vivo.

Valence: El rango está entre 0 y 1. Con la mayoría de pistas consideradas alegres pero tenemos un claro outlier por el lado de una pista musical completamente triste con valor mínimo 62 mil veces menor que la media.

Tempo: Su rango se encuentra entre 50 y 212 pulsaciones por minuto. Una desviación estándar elevada nos indica que nuestras pistas varían bastante en ésta característica.

Popularity: Su rango se ubica entre 0 y 100 puntos de popularidad. La mayoría de nuestras canciones pueden considerarse populares con un promedio puntuado de 53%.

Year: El rango de los años está entre 1970 y 1979.

Limpieza del conjunto de datos

Registros repetidos

Queremos trabajar con datos limpios y claros para analizar nuestro conjunto, por lo que deberíamos buscar registros repetidos, para no incluirlos dentro de nuestro trabajo.

En primer lugar tomamos como clave de los registros la dupla Artista-Canción y las agrupamos con la función Group By de Pandas para ver si conduce a algún problema de repetición de registros.

Analizando los resultados obtuvimos 980 filas, que es la cantidad inicial de filas que teníamos previamente, por lo que podemos concluir que **no tenemos registros repetidos** en nuestro Dataset.

Valores nulos

Para seguir puliendo, vamos a buscar valores nulos para saber cómo tratarlos próximamente: eliminarlos, transformarlos o analizar la razón de la nulidad de los mismos.

Ejecutando una sumatoria de valores con el método isnull() de Pandas que marca la presencia de valores “NULL” observamos que esta suma de nulos en todos los registros nos devuelve 0, por lo que podemos **afirmar la ausencia de valores nulos** en todo el conjunto.

Ejecutamos también una sumatoria de valores con el método isna() de Pandas que marca la presencia de valores “NaN” (Not a Number) y en todas las variables obtuvimos una suma de 0, por lo que **no tenemos valores NaN** en nuestro conjunto.

Para terminar de asegurarnos… como anteriormente cuando describimos estadísticamente las variables no había ninguna de las cuantitativas que presentara valores extraños que pueda considerarse nulos, vamos a agrupar los valores solamente por cualitativas y encontrar posibles nulos o errores de carga.

Sabiendo que no tenemos valores NULL o NaN, como la librería Pandas no nos muestra todos los datos, los imprimimos todos iterando para comprobar que no tenemos uno o más datos cualitativos “Unknown” o cualquier nombre extraño. Dicha impresión estará dada por el artista y la cantidad de canciones que éste tiene. Gracias al método items() que convierte la serie devuelta por el método value_counts() en una vista de pares (índice, valor).

Luego de repasar toda la lista, comprobamos los datos de los Artistas “10cc”,”GQ” y “M” solo por tener nombres extraños y asegurarnos de que no sean muestras basura y resultaron tener datos válidos por lo que no habría necesidad de eliminar ni transformar ninguna variable.

Conversión de tipos

Al principio habíamos dicho que la variable **Duration** la íbamos a tratar como un dato cuantitativo porque es la medición de la duración de una canción. Así que, para poder incluirla en el análisis y tratar con ella, vamos a convertir en un dato numérico aplicando un reemplazo de ":" (razón por la que Pandas la tomaba como un String) por un "." y convertirla a un tipo float.

Completando la descripción estadística, esta es una variable que en este espacio muestral tendrá un rango de 0 a 30 minutos por canción. Con un promedio de 3.6 minutos y un máximo de 26 minutos.

Por otro lado, notamos que la variable **Loudness** que mide el volumen promedio de la canción, tiene valores negativos y la gran mayoría de las variables tienen un rango de 0 a 1. Como siempre es conveniente trabajar con rangos iguales en ambos lados, y no queremos perder la información de la cantidad de decibelios, vamos a invertir su rango teórico de -100 a 0, a un intervalo entre 0 y 1. Y para hacerlo, le aplicaremos a la columna una función de valor absoluto donde convertimos todo su rango negativo en uno positivo entre 0 y 1 dividiéndolos por 100.

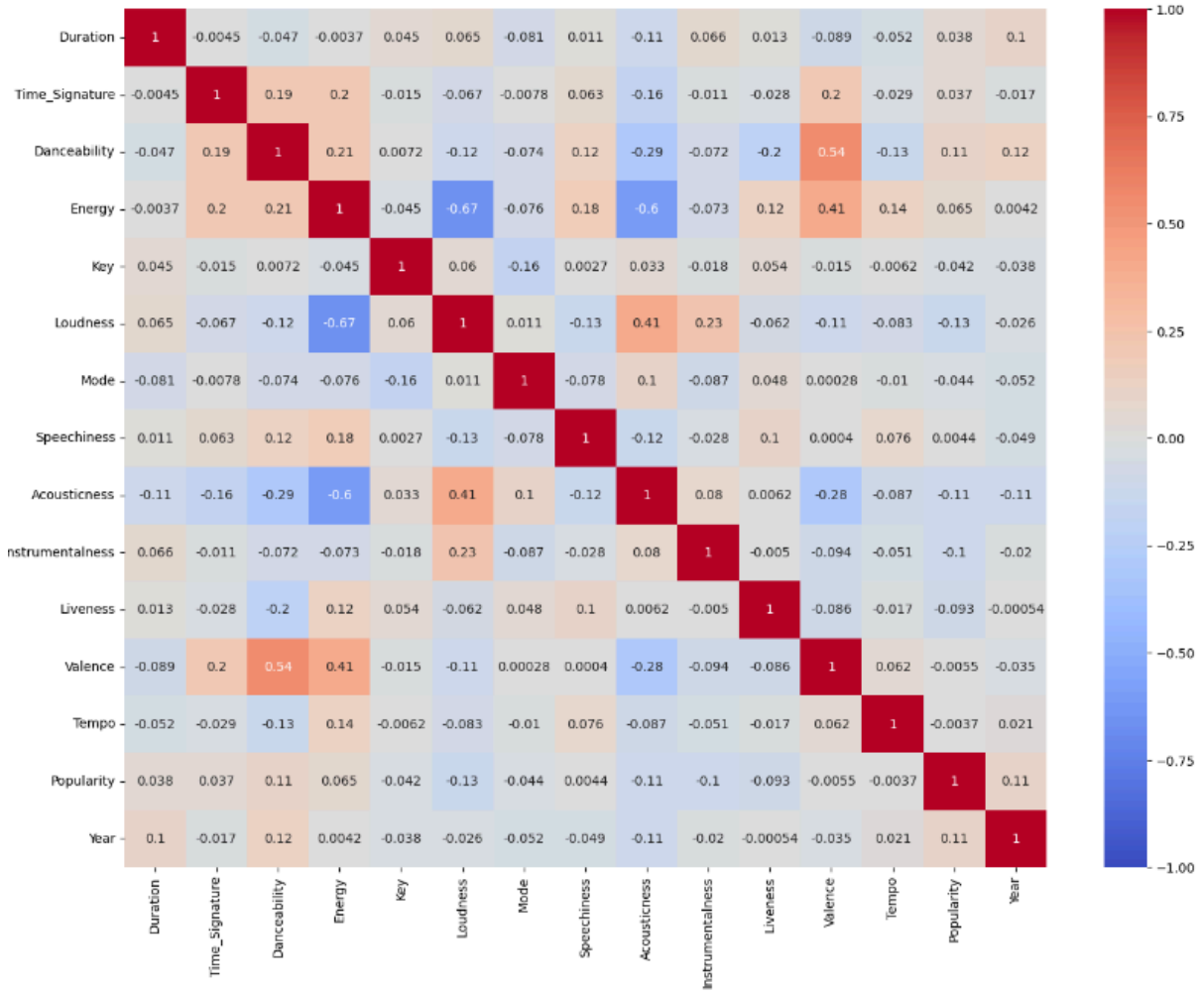
Matriz de Correlación

Llegamos al punto donde ya tenemos todos los datos básicos y necesarios de cada variable particular, por lo que ahora queremos estudiar el comportamiento bivariado de todas las variables cuantitativas de nuestro conjunto. Para eso, vamos a utilizar la **Matriz de Correlación**, pero antes debemos quitar de nuestro DataFrame los datos cualitativos porque no existe una relación numérica entre dos palabras.

Para esto **eliminamos las columnas Track y Artist** y luego lanzamos la matriz de correlación entre todas las variables con el método Corr() de Pandas .

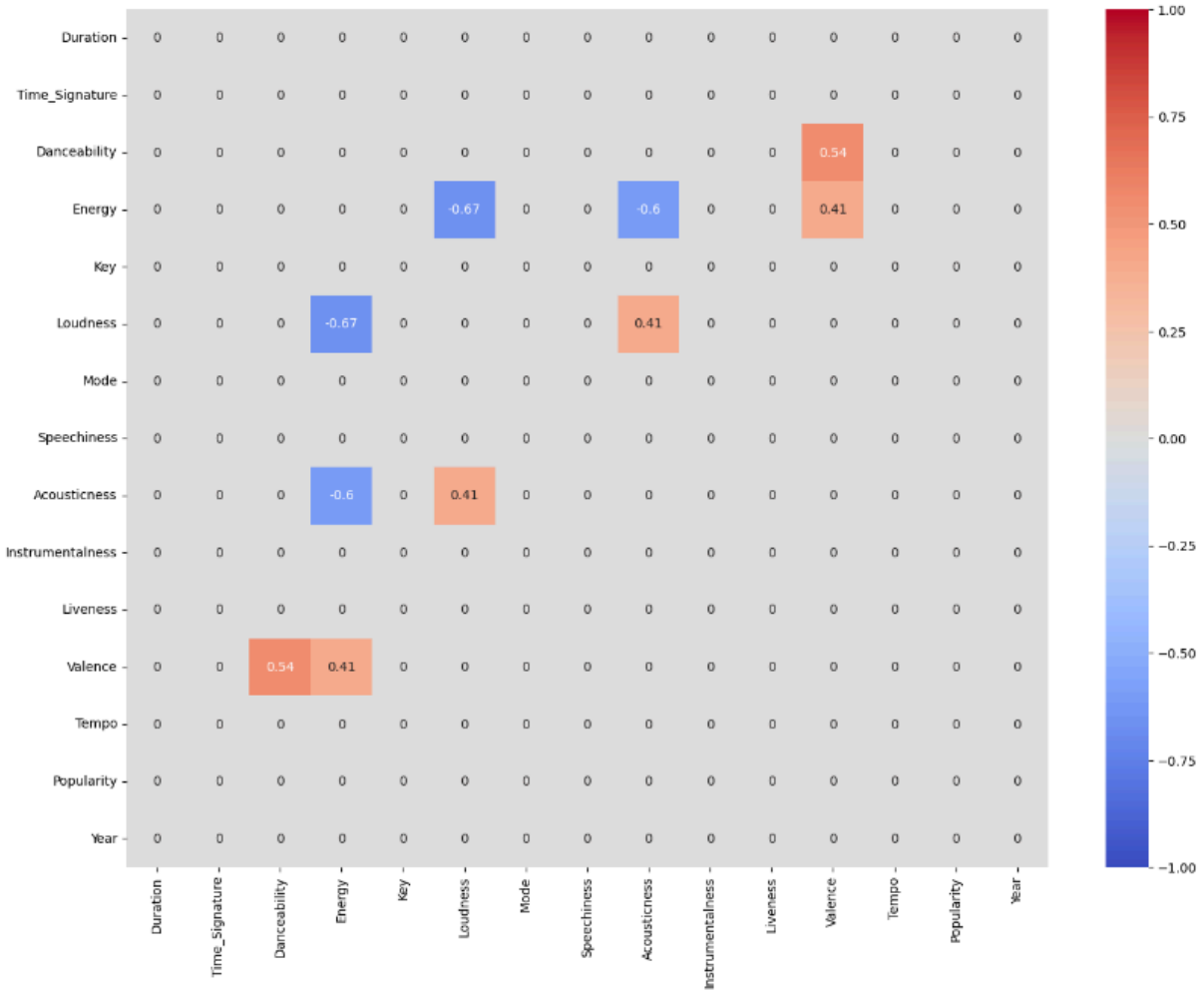
Representamos la matriz gráficamente para identificar las correlaciones elevadas o bajas, asignando colores a cada correlación. Esto lo haremos con la característica de Heatmap de la librería "Seaborn" montada sobre "Matplotlib".

Las tonalidad de color **rojo** representan **valores cercanos a 1** y tonalidades de color **azul** representan **valores cercanos a -1**.



Vamos a pasar en limpio esta matriz quitándole información difusa como la correlación entre mismas variables. Y establecer una condición para mostrar solo aquellos coeficientes que su valor sea digno de estudio. En este caso, basándonos en el estándar aceptado actualmente, no tenemos ni un solo coeficiente que llegue al **valor absoluto de 0.7**, por lo que vamos a establecer un **límite en 0.4** y estudiar el comportamiento entre aquellas variables que lo cumplan.

Igualmente, se estudiará más adelante el motivo de la poca o casi nula relación entre la mayoría de las variables.



Finalmente, las siguientes variables fueron las que cumplieron con el límite de **correlación de Pearson** que impusimos:

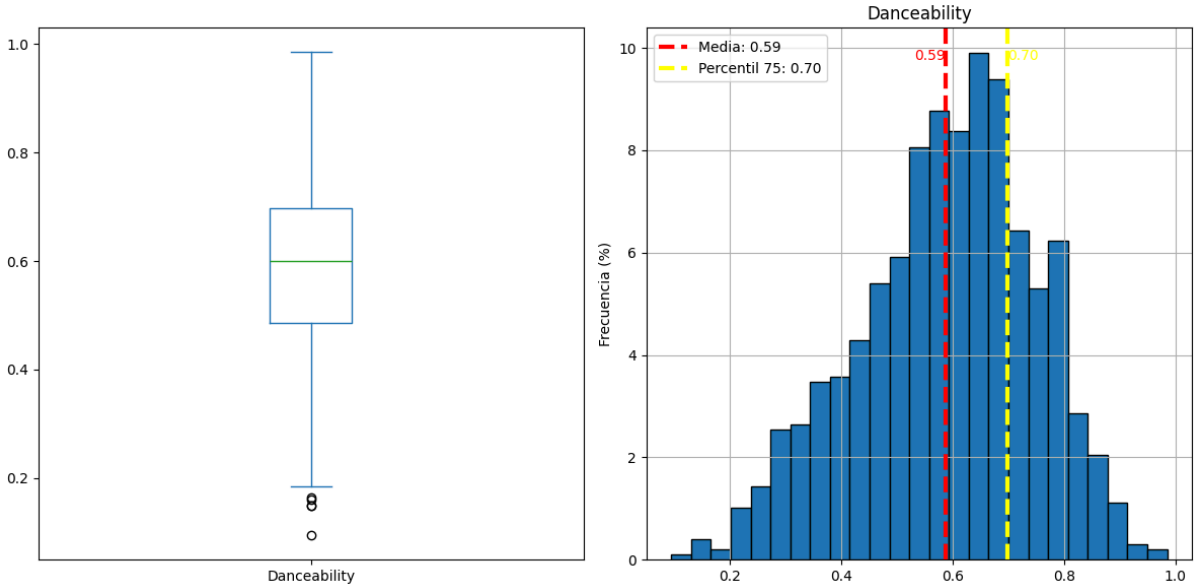
- **Danceability - Valence:** 0.54
- **Energy - Loudness:** -0.67
- **Energy - Acousticness:** -0.6
- **Energy - Valence:** 0.41
- **Loudness - Acousticness:** 0.41

Estudio univariado

Antes de hacer el **análisis bivariado**, vamos a observar en profundidad individualmente cada variable para comprenderlas bien antes de compararlas con otra. De cada variable analizaremos su **boxplot** y su **histograma** para conocer su distribución y varianza, además de **comprobar su normalidad** mediante el Test de Shapiro-Wilk, donde debemos obtener un p-valor mayor al nivel de significancia $\alpha = 0,05$.

Por lo que investigamos, Pandas utiliza una medida de K=0 para indicar una distribución normal en el estudio de Curtosis.

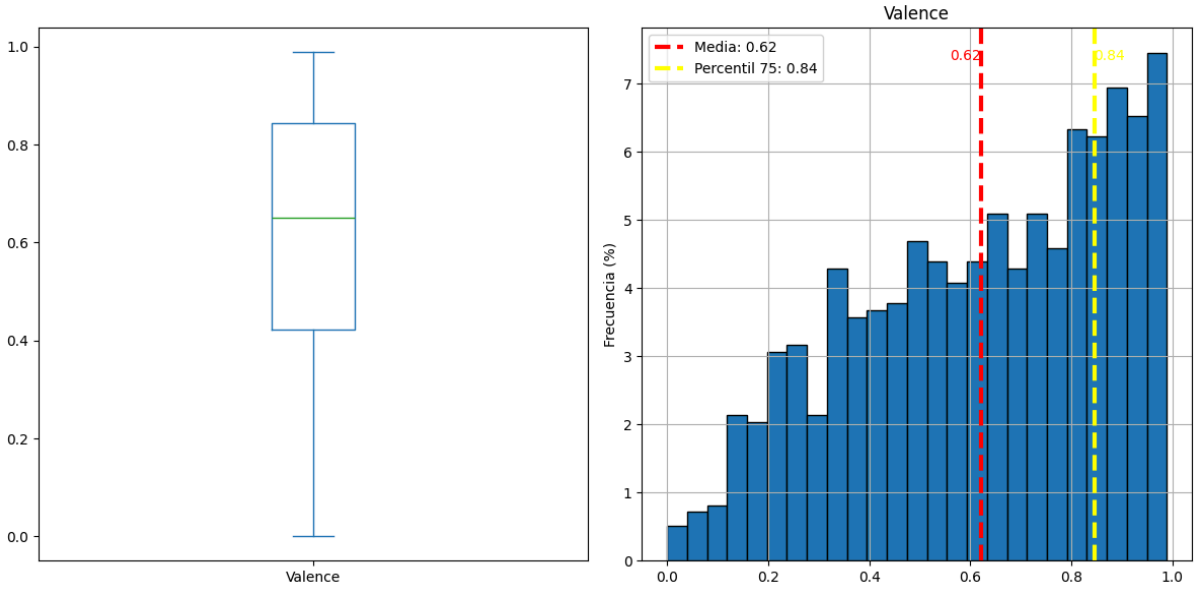
Danceability



- Coeficiente de asimetría: -0.34
- Curtosis: -0.29
- Distribución **no normal** por Test de Shapiro-Wilk

Tenemos una distribución no normal con un ligero sesgo a la izquierda, coeficiente de asimetría negativo indicando que los valores extremos están a la izquierda de la media. Su cola de distribución es ligera al tener una curtosis negativa y además cercana al 0. Poca presencia de outliers al fijarnos en su boxplot.

Valence



- Coeficiente de asimetría: -0.39

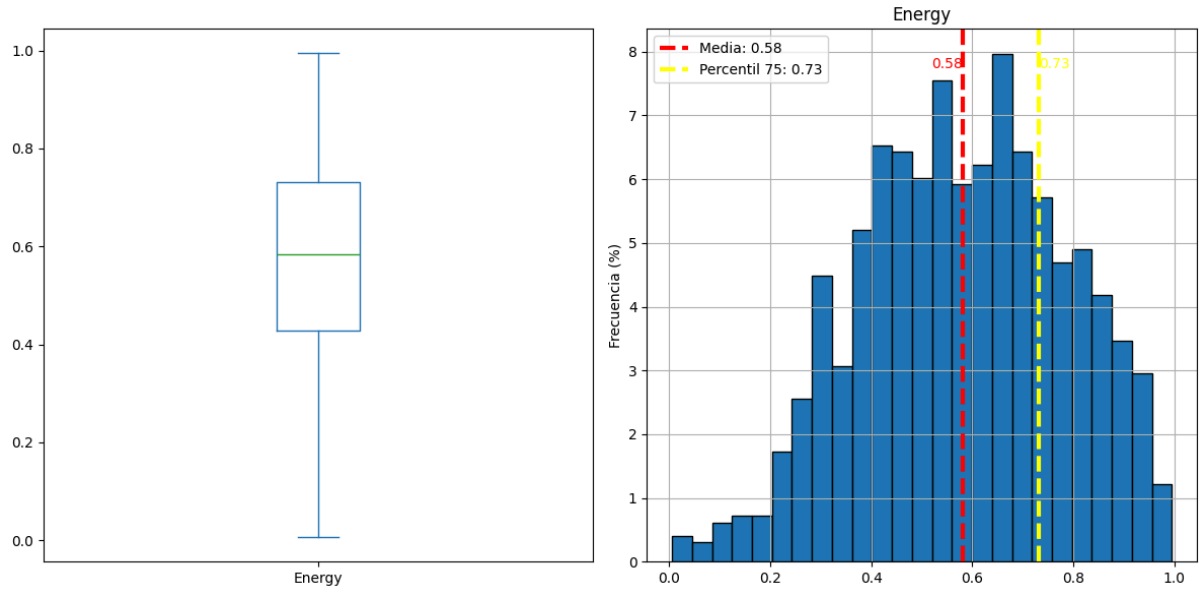
- Curtosis: -0.95
- Distribución **no normal** por Test de Shapiro-Wilk

Observamos principalmente en el boxplot que ésta variable tiene un sesgo a izquierda porque el 50% de sus datos están inclinados a sus valores máximos, con un coeficiente de asimetría negativo indicando que sus valores extremos están ubicados a la izquierda de la media. Difícil de notar en el histograma por la alta frecuencia de sus valores en la muestra. Y como habíamos notado anteriormente, el valor mínimo de **Valence** tiene una diferencia tan gigantesca con su media que no logra notarse en el boxplot y menos en el histograma.

Notamos que la mayor frecuencia de las muestras se encuentran en su máximo y alrededores cercanos, lo que siempre a primera vista puede parecer extraño. Así que vamos a analizar todos sus valores máximos a ojo para notar si hay cargas incorrectas o tenemos alguna especie de datos basura que nos están desplazando las distribuciones y dando una mala lectura de los análisis posteriores.

Al analizar todos los valores entre un rango de 0.7 y 1 podemos afirmar que no estamos en presencia de datos mal cargados, que podría ocurrir si hubiera muchísimas muestras con un mismo valor de Valence o con el máximo posible (lo que podría indicar un error de carga). Por suerte, este no es nuestro caso porque los valores repetidos que encontramos sólo se extienden por 4 o 5 muestras, pero el resto son todos valores diferentes.

Energy

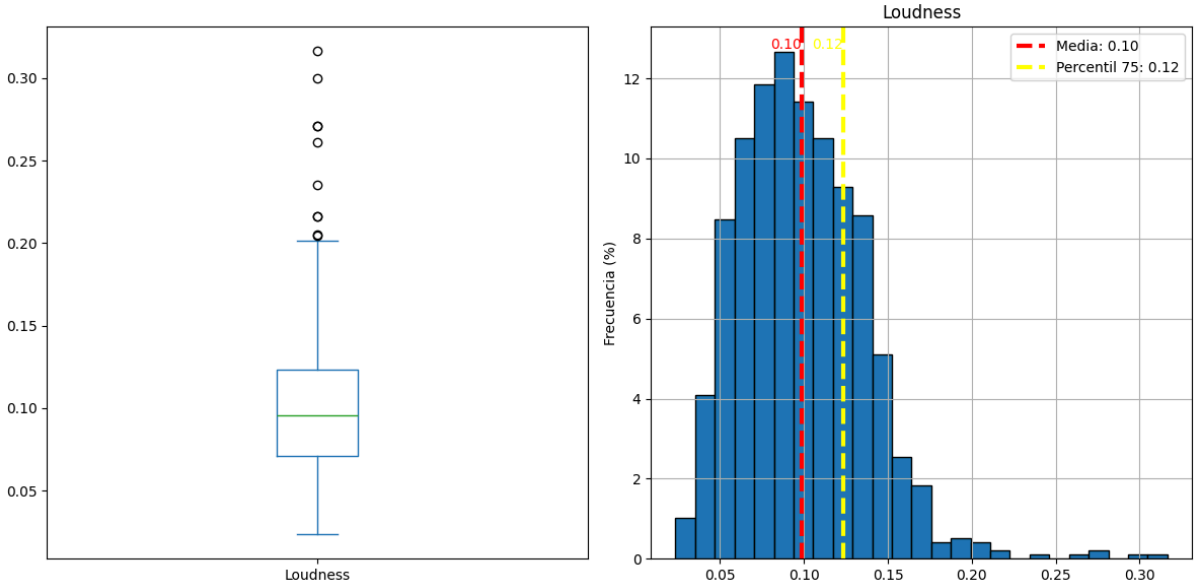


- Coeficiente de asimetría: -0.14
- Curtosis: -0.58
- Distribución **no normal** por Test de Shapiro-Wilk

Profundizando lo que habíamos notado al principio del trabajo con **Energy**, el boxplot y el histograma nos confirman que se trata de una distribución normal, donde con un desvío estándar cercano a cero, la gran mayoría de sus valores se concentran alrededor de la media y el coeficiente de asimetría negativo nos indica que los extremos estarán en los

valores bajos de la distribución como es el caso del valor mínimo que al igual que en el caso de **Valence** no puede representarse gráficamente.

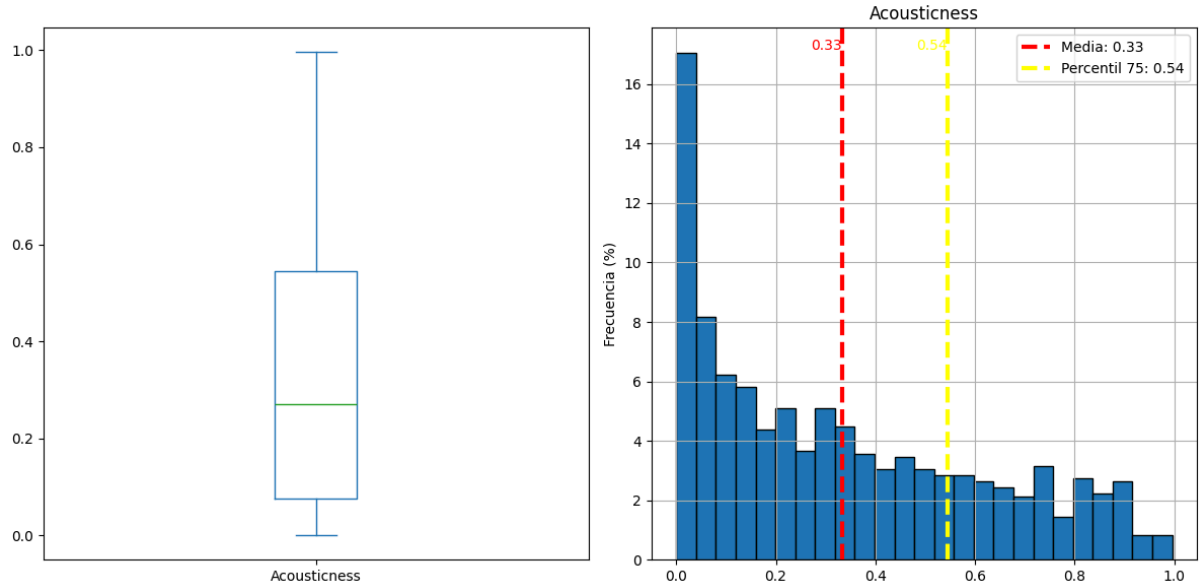
Loudness



- Coeficiente de asimetría: 0.94.
- Curtosis: 2.67
- Distribución **no normal** por Test de Shapiro-Wilk

Observamos gráficamente que la distribución de **Loudness** con un **coeficiente de asimetría** positivo tiene un sesgo a derecha, donde podemos ver que hay muchos valores extremos repartidos por derecha de la media pero con poca frecuencia de valores altos. Con una **Curtosis** elevada sabemos que ésta variable tiene una cola pesada significando la presencia de valores atípicos, lo que confirmamos al ver su boxplot con presencia de **outliers a partir de 0.20** o -20 decibelios.

Acousticness



- Coeficiente de asimetría: 0.59
- Curtosis: -0.84
- Distribución **no normal** por Test de Shapiro-Wilk

Podemos ver que la variable **Acousticness** tiene un sesgo a derecha, con el 75% de los valores tirando hacia valores bajos y con una gran presencia de extremos mínimos en la muestra de pistas. Con una curtosis negativa y baja indica una cola ligera que significa ausencia de valores atípicos. Y al tener coeficiente de asimetría positivo tiene la cola más larga a la derecha.

Como ocurría con **Valence**, pero a la inversa, notamos que tiene una alta frecuencia de valores mínimos en el cero y sus alrededores muy por encima del resto de muestras, por lo que vamos a estudiar un poco esto para asegurarnos de que no tengamos datos mal cargados que nos estén haciendo ruido en el análisis.

Luego de ordenar los valores de forma ascendente y mostrar las 980 filas para poder comprobar que no tengamos problemas, las inspeccionamos a ojo y podemos afirmar que no tenemos datos mal cargados en nuestro dataset ya que prácticamente todos los datos que nos encontramos son únicos y los repetidos no se extienden por más de 3 o 4 muestras. Así que podemos continuar con el análisis sin modificar esta variable.

Análisis bivariado

Una vez analizadas las variables individualmente, procedemos a ver cómo se relacionan éstas dependiendo una de la otra.

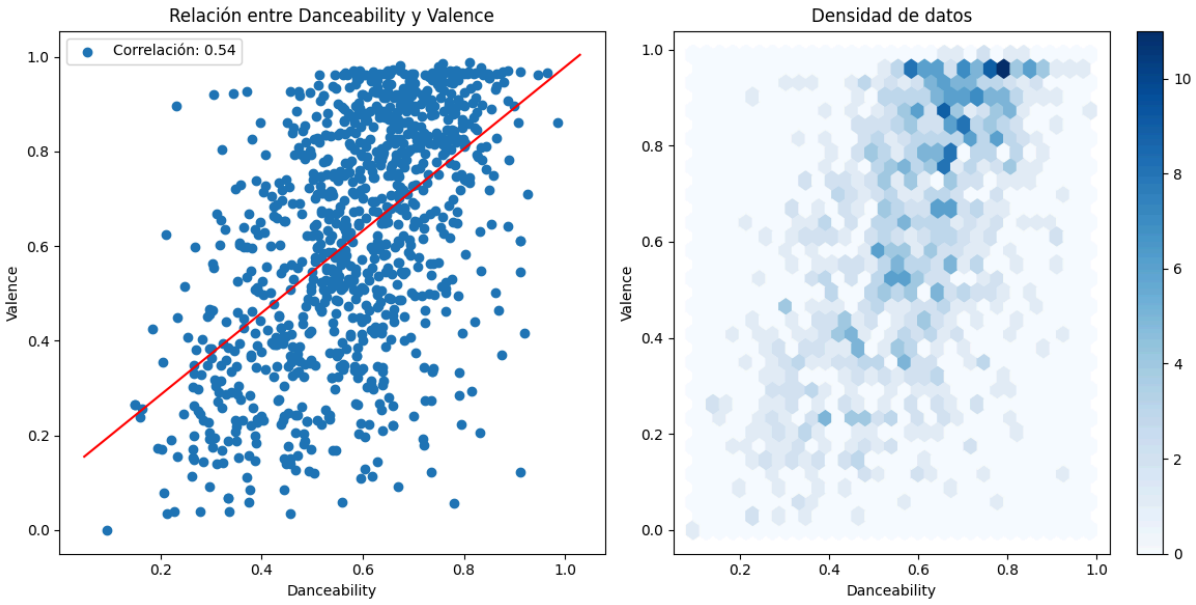
Para empezar siempre es conveniente hacer un **gráfico de dispersión** o **Scatter Plot** para ver particularmente los datos representados de cada variable y valorar su influencia o impacto.

Incluímos una **línea de mejor ajuste** o “line of best fit” para marcar visualmente la tendencia lineal de la correlación de las variables.

También generamos un **diagrama de hexágonos** o **Hexbin plot** para representar la densidad de los datos cuando tenemos muchos puntos agrupados o concentrados en una zona o intervalo.

Danceability - Valence

Vamos a explorar la relación que hay entre cuán adecuada es una pista para bailar y qué tan alegre es la misma.



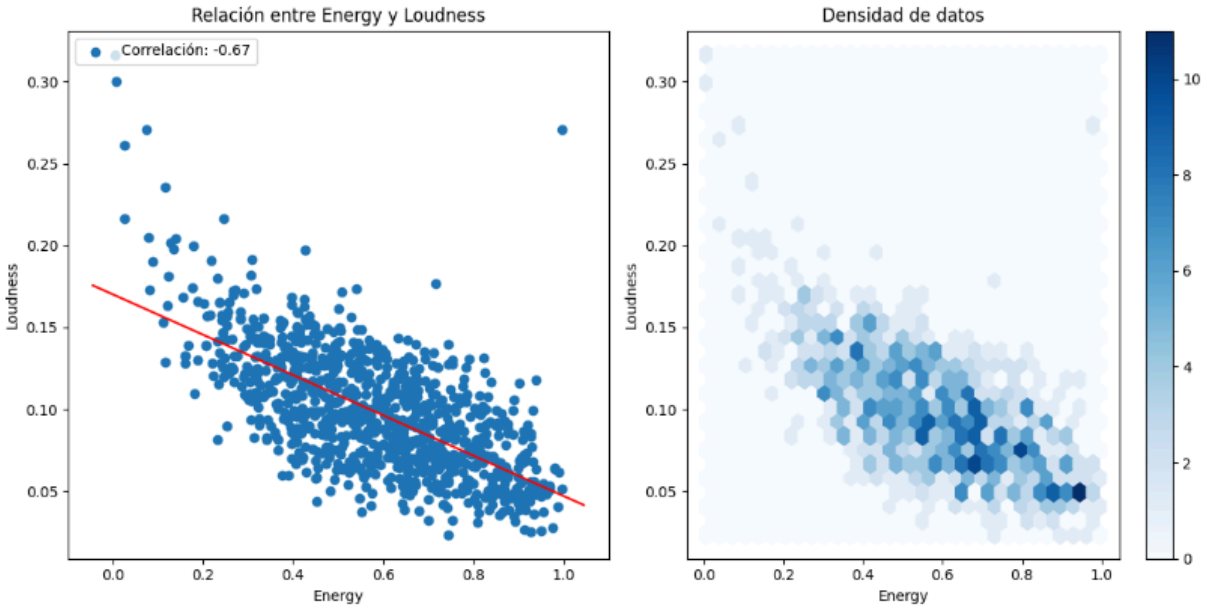
A simple vista podemos suponer que no hay ninguna relación entre que tan “buena” es una canción para danzar y su positividad musical fijándonos en el **scatter plot** por la gran dispersión que hay en los datos, donde para un mismo valor de **Danceability** tenemos distintos valores de **Valence** en casi todo su rango. Gracias a la herramienta de la línea de mejor ajuste y el diagrama de hexágonos podemos profundizar un poco más.

Con la **línea de mejor ajuste** marcada sabemos que la relación tiene una tendencia positiva (pendiente > 0), indicando que a medida que una pista es más adecuada para bailar (basada en el tempo, la estabilidad del ritmo, la fuerza, etc) tiene una positividad musical cada vez más alta.

Con el **Hexbin plot** conocemos que hay una concentración de valores casi máximos de **Valence** en el rango de 0.6 a 0.8 de **Danceability**, lo que nos confirma un poco lo anterior; una canción adecuada para bailar tiene una tendencia a ser muy alegre o viceversa.

Energy - Loudness

Vamos a explorar la relación que hay entre la medida enérgica de una pista y el volúmen promedio de la misma.



Como para el caso particular de la variable Loudness tuvimos que invertir su rango, pasando de negativo a positivo y de una escala de [-100, 0] a [0, 1], entonces el análisis del gráfico de la correlación en general debe hacerse invertido, donde **menores valores** de Loudness representan **mayor volumen**. Por lo tanto, también a fines prácticos, la correlación de Pearson es positiva (mayores valores de uno con mayores valores de otro).

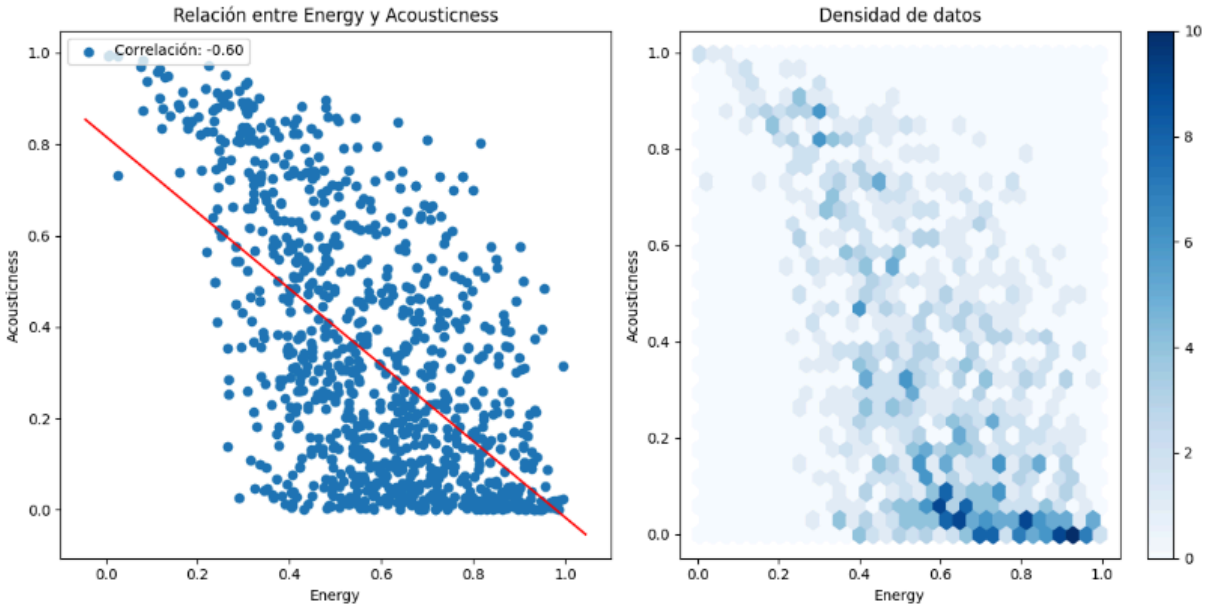
Gracias a la **línea de mejor ajuste**, si bien su pendiente es negativa, sabemos que la tendencia de los datos es positiva porque en la correlación donde hay mayor volumen de la canción, más enérgica es. Y vemos en el **Scatter plot** que los puntos están concentrados sobre la línea indicando que es una correlación fuerte, lo que igualmente podíamos suponer de antemano al tener un valor de **correlación de Pearson de 0,67**.

Entonces, apoyándonos también en el **Hexbin plot** de la derecha, hay una fuerte tendencia que si las canciones presentan un volumen muy elevado (cercano a 0 decibelios), será más enérgica y con mucha actividad. Esto se puede ver con la densidad de los datos donde tenemos una concentración alta que se va acumulando de manera progresiva tal como la línea de ajuste lo indica, en el rango de los **-15 a 0 decibelios** de volumen (valores altos teniendo en cuenta que los decibelios se miden en un rango de -100 a 0) y en el rango de **0.5 a 1 valores de Energy**.

Como dato extra notamos que tenemos un outlier en la correlación para el valor más alto de Energy, tenemos un valor “relativamente bajo” en nuestra escala de volúmenes cercano a -27 decibelios.

Energy - Acousticness

Vamos a explorar la relación que hay entre la medida enérgica de una pista y su calidad acústica, o que tan acústica es.



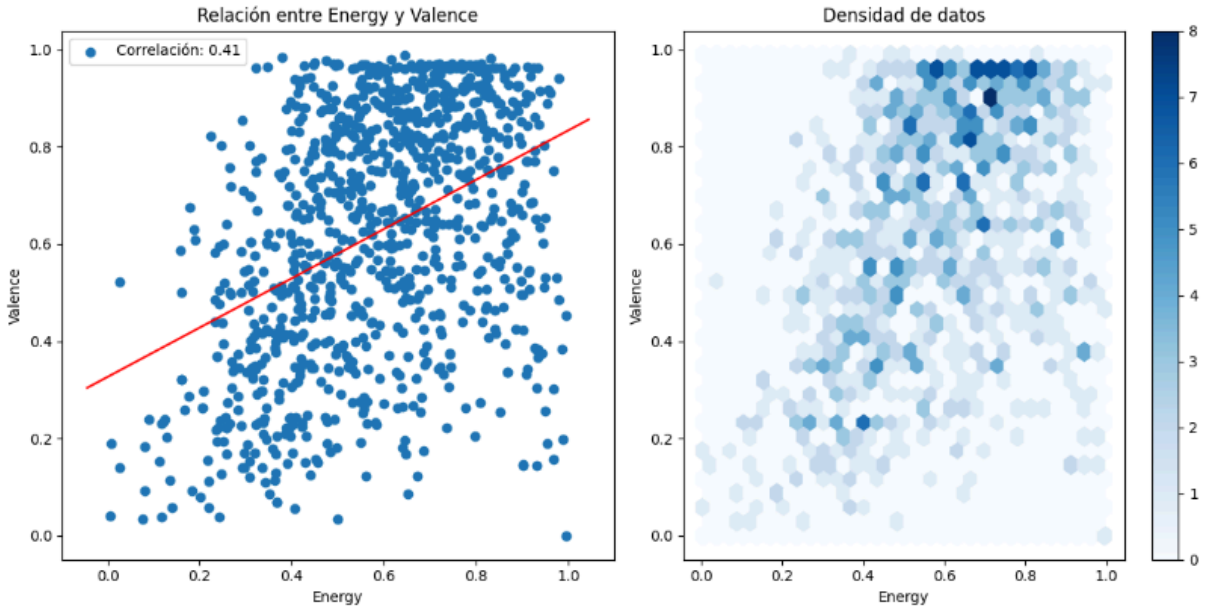
Si observamos a priori la línea de mejor ajuste, podemos destacar que la tendencia de la relación es que a mayores valores de una variable, menores valores que la otra porque su pendiente es negativa.

En este caso a mayores valores de **Energy**, menores valores de **Acousticness**, lo que traducido sería: a medida que una canción empieza a tener un tono más enérgico, se vuelve cada vez menos acústica y eso se representa bien en ambos gráficos.

Si bien puede notarse en el **scatter plot** que los puntos están bastantes dispersos, ayudándonos con el **Hexbin plot** podemos notar que en nuestra muestra hay una tendencia de aglomeración de datos de Acousticness en el rango de 0.6 a 1 de Energy, confirmando la conclusión anterior de la tendencia que hay pero además significando que ya prácticamente si una pista tiene un **valor enérgico mayor a 0.6** entonces se puede considerar **nula acústicamente.**

Energy - Valence

Vamos a explorar la relación que hay entre la medida enérgica de una canción y su positividad musical.



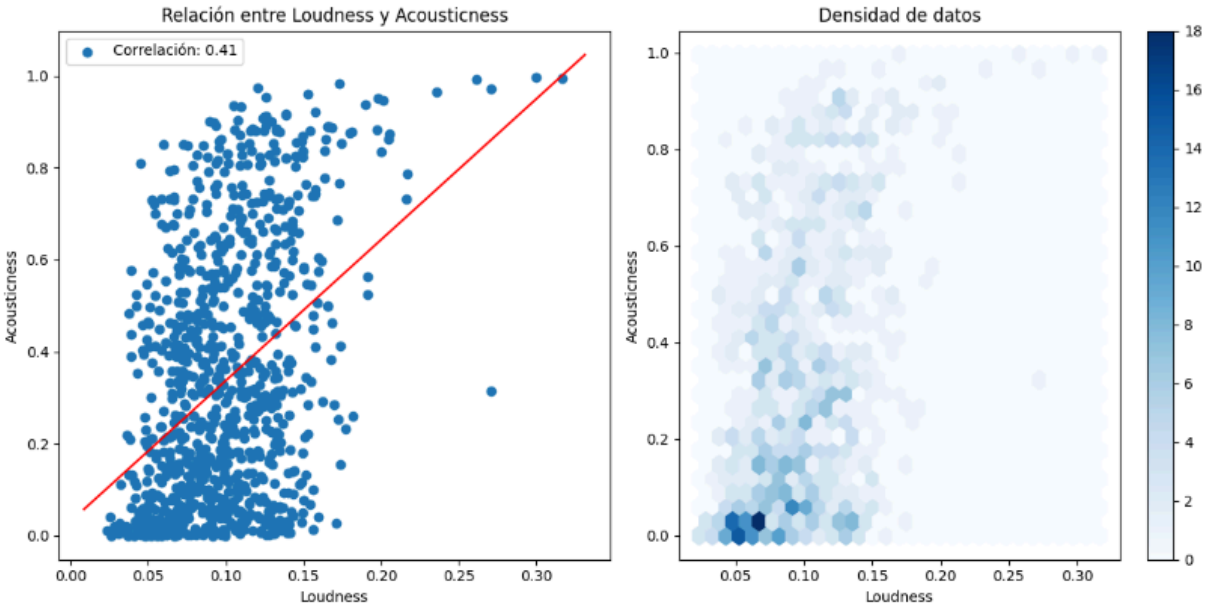
Nuestro valor de **correlación de Pearson** para ésta ocasión es de **0.41**, osea que se encuentra en el límite que decidimos poner para estudiar dichas correlaciones debido a que no tenemos ninguna que alcance el estándar aceptado de 0.7. Por lo tanto, como se considera un valor bajo se comprueba gráficamente que los datos están bastante dispersos, teniendo para un mismo valor de Energy, muchos valores de Valence y viceversa.

Podríamos solo con ver en el **scatter plot** deducir que no existe relación alguna entre la medida enérgica de una pista y que tan alegre es o que no vale la pena el análisis. Pero el **diagrama de hexágonos** nos devela que hay una concentración de datos en el rango de 0.5 a 0.8 de Energy y 0.9 a 1 de Valence, lo que nos indica que por lo menos en nuestra muestra hay una tendencia de pistas enérgicas con valores medio-altos y casi máximos de positividad musical.

Para concluir, en nuestro dataset hay una tendencia que marca si una pista es medianamente enérgica, es **probable** que sea una canción muy alegre. Pero no descartemos del análisis los “outliers” que nos dicen si una canción tiene nula energía o actividad puede ser medianamente alegre. Y por el lado contrario, si es totalmente enérgica puede ser muy triste.

Loudness - Acousticness

Vamos a explorar la relación entre el volúmen promedio de la canción y su calidad acústica.



Nuevamente vale aclarar, que para el caso particular de la variable Loudness tuvimos que invertir su rango, pasando de negativo a positivo y de una escala de [-100, 0] a [0, 1], entonces el análisis del gráfico y de la correlación en general debe hacerse invertido, donde **menores valores** de Loudness representan **mayor volumen.** Por lo tanto, también a fines prácticos, la correlación de Pearson es negativa (menores valores de uno con mayores valores de otro).

Con la **línea de mejor ajuste** teniendo una pendiente positiva notamos que hay una tendencia entre menores valores de Loudness y menores valores de Acousticness, aunque no es algo fácilmente distinguible en el gráfico de dispersión porque si agarramos el rango de -15 a 0 decibelios de volumen podemos emparejar distintos valores de medida acústica en la totalidad de su rango.

No nos sirve de mucho analizar el espacio vacío que se encuentra en los valores de -30 a -15 decibelios de volumen porque no tiene que ver su relación de medida acústica sino con nuestra muestra general de 980 canciones donde tenemos muy pocas que se encuentran en ese rango, con un mínimo de -31 decibelios y un promedio de -9.86. Lo mismo se puede notar en la **correlación Energy-Loudness**.

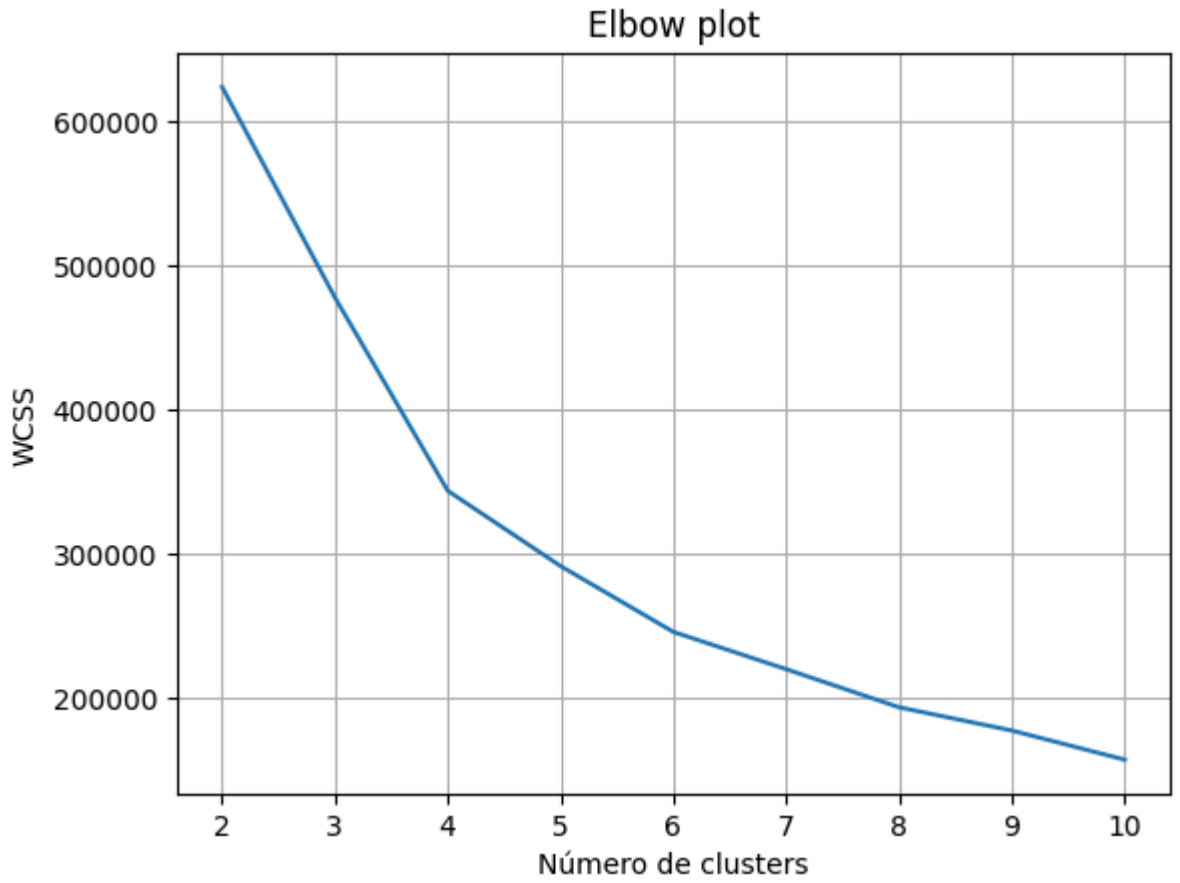
Podemos destacar igualmente gracias al Hexbin plot que las canciones con un volumen muy elevado (valores cercanos a 0 de Loudness) son casi nulas acústicamente. Pero esto solo es notable en nuestra muestra porque si nos ponemos a pensar, por ejemplo, el caso de las orquestas, el volúmen que utilizan es muy elevado y son de las melodías más acústicas que podemos encontrar en el mundo de la música.

Entonces por lo explicado anteriormente, y teniendo en cuenta que la correlación de Pearson es de 0.41 (rozando el límite establecido), no tenemos una tendencia de relación fuerte en este caso.

Clustering

Para finalizar nuestro análisis exploratorio de los datos, es una buena idea aplicar **Clustering** en este momento, porque es una técnica basada en la formación de grupos o “clusters” en las muestras para que luego nosotros podamos buscar similitudes intra-cluster y extra-cluster que nos permitan darnos cuenta de comportamientos en los que podemos profundizar más adelante.

Utilizaremos **clustering basado en particionamiento** y para eso necesitamos el valor de K-Means necesario para tener una cantidad de agrupaciones adecuadas. Así que usaremos de ayuda el **Elbow plot** o gráfico del codo que realiza un análisis sobre la cantidad de clusters y la distancia entre las muestras para encontrar el número óptimo donde la distancia comienza a decrecer lentamente.

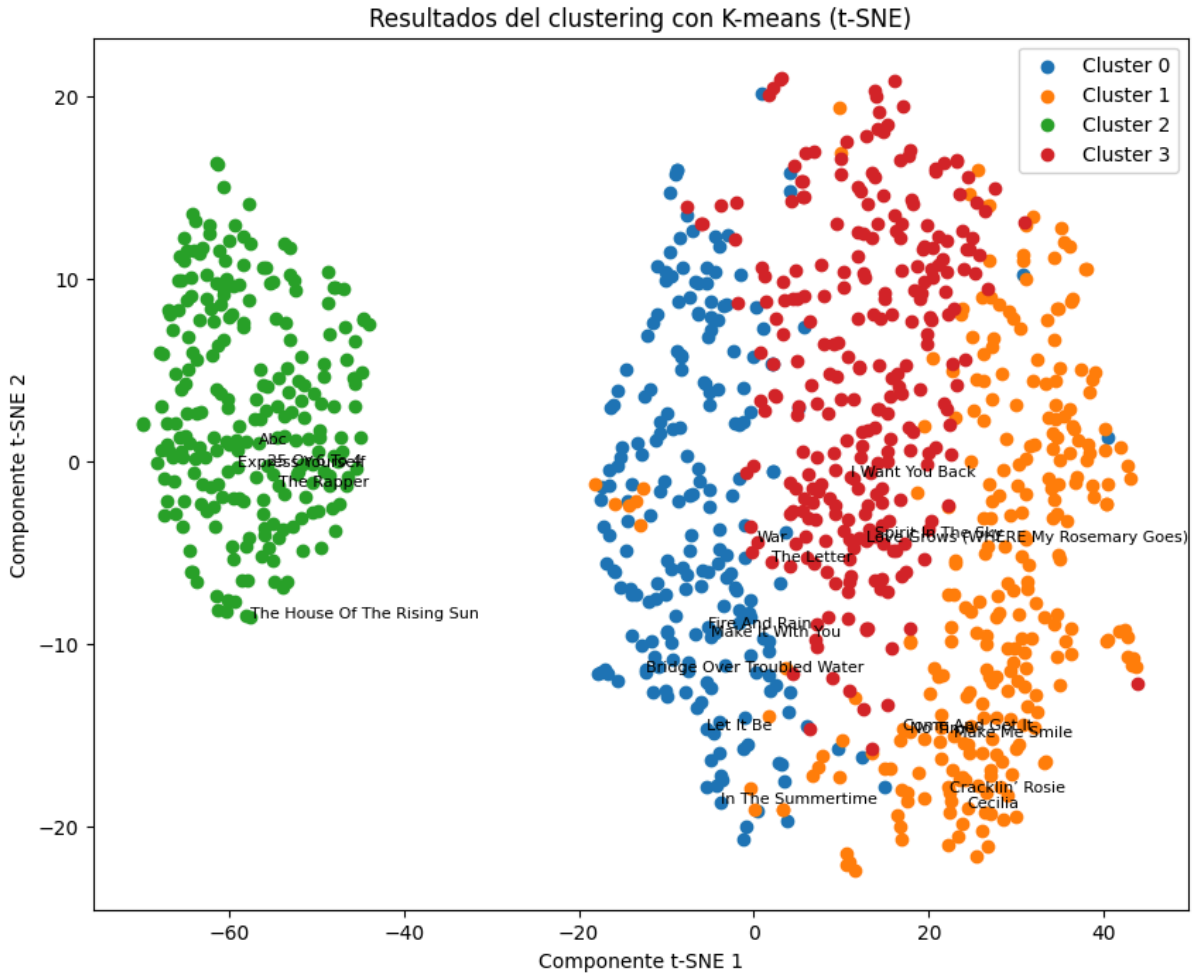


Al ver el resultado, entramos en duda sobre qué cantidad de clusters nos daría mejor resultado, si 4 o 5. Así que aplicamos una validación mediante el **Coeficiente de Silueta** que tiene en cuenta la combinación de la **cohesión** (cercanía de muestras intra-cluster) y la **separación** (distancia extra-cluster) y obtuvimos lo siguiente:

- K=4 con un puntaje de 0.1734
- K=5 con un puntaje de 0.1616

Si bien ambos puntajes son bajos, porque lo ideal es cercano a 1, y la diferencia es corta, tenemos un mejor resultado implementando 4 clusters. Así que elegimos un K=4 y continuamos.

Decidimos agregar a los clusters el nombre de algunas canciones para identificarlas rápidamente en el gráfico.



Observamos dos grandes grupos principales considerablemente separados y dentro de éstos los clusters están bien delimitados salvo por algunas excepciones que pueden mejorar o empeorar el gráfico dependiendo la cantidad de veces que ejecutemos el código por el modelo de entrenamiento de T-SNE.

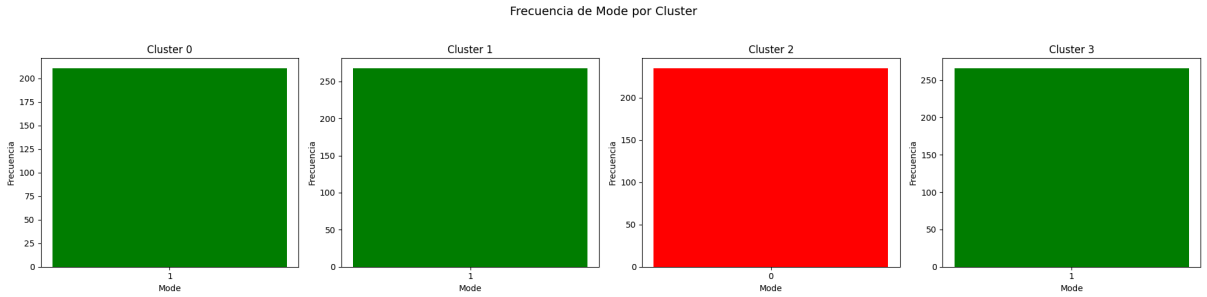
Ahora vamos a analizar los datos estadísticos que contiene cada cluster formado para poder mirarlos con atención e intentar llegar a alguna conclusión de las diferencias formadas. Los ordenaremos según como están formados los grupos.

Conclusión de clustering

Para estudiar las características de cada cluster decidimos generar los boxplot para cada variable continua y unos histogramas de frecuencia para las categóricas. Y luego de analizar detalladamente las comparaciones entre los dos tipos de variables de cada cluster,

podemos concluir que la única diferencia entre los grupos de los **clusters 0, 1, 3** y el **cluster 2** es su variable **Mode** donde en el primer grupo mencionado su valor es de 1, indicando que toda las canciones que lo conforman tienen un tono mayor, a diferencia del segundo grupo mencionado donde todas sus pistas tienen un valor de 0, indicando un tono menor. Y eso justifica la distancia extra-cluster que encontramos.

En el resto de características, todos nuestros clusters comparten el mismo rango práctico y hasta en algunos casos, tienen el mismo valor.



Como **Mode** es una variable dicotómica, creímos conveniente pintar sus histogramas de frecuencia con un **verde** si la canción está en un **tono mayor** y con un color **rojo** si la canción está en un **tono menor**.

Conclusión del análisis exploratorio

En conclusión, luego de estudiar todas las características de nuestras variables, hacer un control de limpieza de nuestro conjunto y estudiar las distribuciones y comportamientos de aquellas involucradas en relaciones lineales, podemos afirmar que estamos ante un dataset de canciones populares (aunque por poco), con ritmos aptos para ser bailables, un tanto enérgicas, bastante ruidosas y alegres, diferenciadas entre ellas por sus tonalidades altas o bajas. Principalmente con **distribuciones no normales** en las únicas variables con correlaciones lineales, algo que no nos beneficia a la hora de estudiarlas porque los resultados pueden tender a ser engañosos o difíciles de interpretar por culpa de los sesgos en las distribuciones con mucha frecuencia de valores extremos.
Al ya conocer normalidades, ya sabemos por qué técnicas o tests vamos a tener que inclinarnos próximamente.

Ya conociendo toda la información básica de nuestras características, estamos en condiciones de plantear algunas hipótesis.

Planteamiento de hipótesis

1. **Las canciones más adecuadas para bailar, tienden a ser más alegres.**

2. **El volúmen de una canción y su calidad acústica pueden determinar su nivel de intensidad y actividad.**

3. **Una canción con altos grados de actividad es más alegre.**

4. **Si una canción tiene popularidad alta, es muy probable que se haya interpretado en vivo al tener mayor ruido de audiencia o viceversa.**

5. **Las canciones más adecuadas para bailar son más populares que otras teniendo en cuenta la época.**

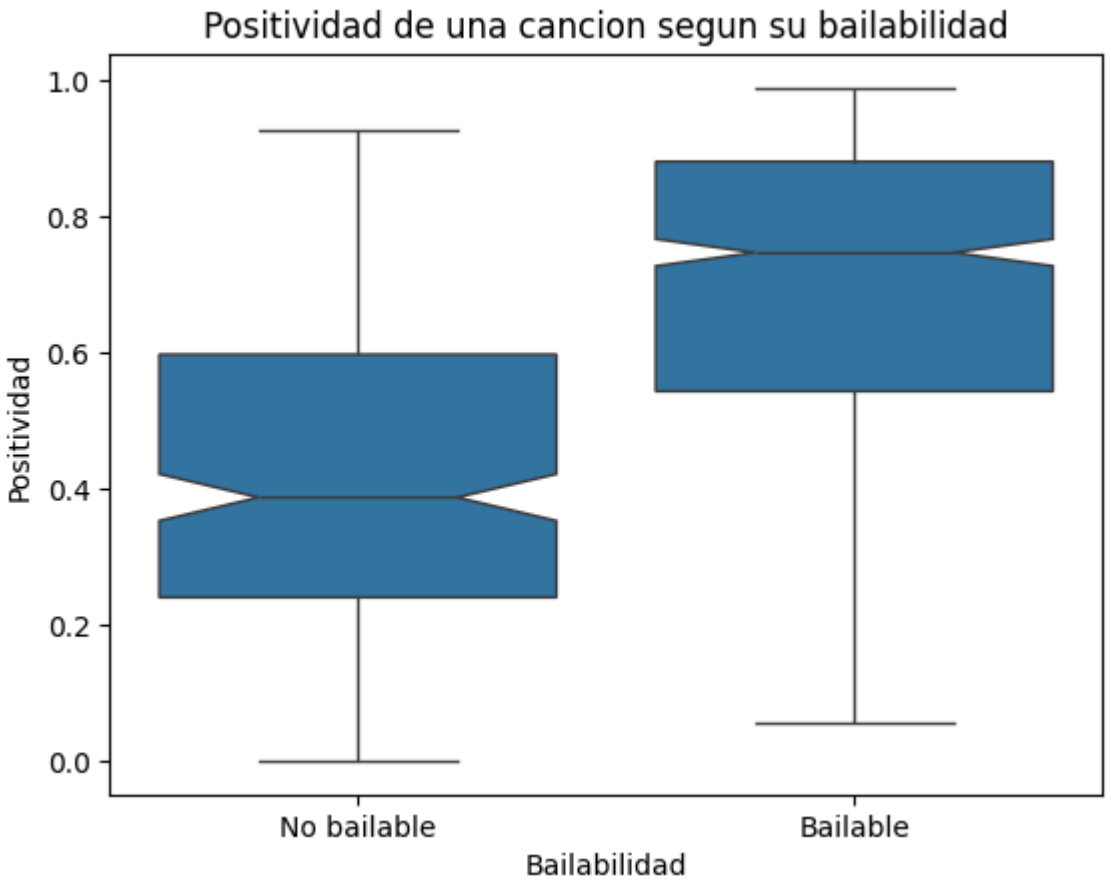
6. **La popularidad de una canción está conformada por una serie de características comunes que eran de buen gusto multitudinario en la época de los años 70.**

Validación de hipótesis

Hipótesis 1

Nuestra curiosidad inicial surgió de lo que encontramos en el análisis exploratorio (más específicamente en el análisis bivariado) respecto a la bailabilidad de una canción y su positividad musical, marcando una tendencia positiva en los datos. Por lo tanto ahora vamos a validar si estábamos en lo correcto cuando pensamos que si un tema es adecuado para bailar debería ser alegre.

Para esto, como ya sabemos desde la exploración de datos que tanto Danceability como Valence no se distribuyen normal, resta validar el supuesto de homocedasticidad para saber que test No Paramétrico implementar. Y por sus boxplot podemos deducir que algo de significancia estadística podemos tener al no compartir rangos ni intervalos de sus medianas en el 50% de sus datos.




```
Test de Levene para positividad musical: Estadístico=0.088, p-valor=0.767
```

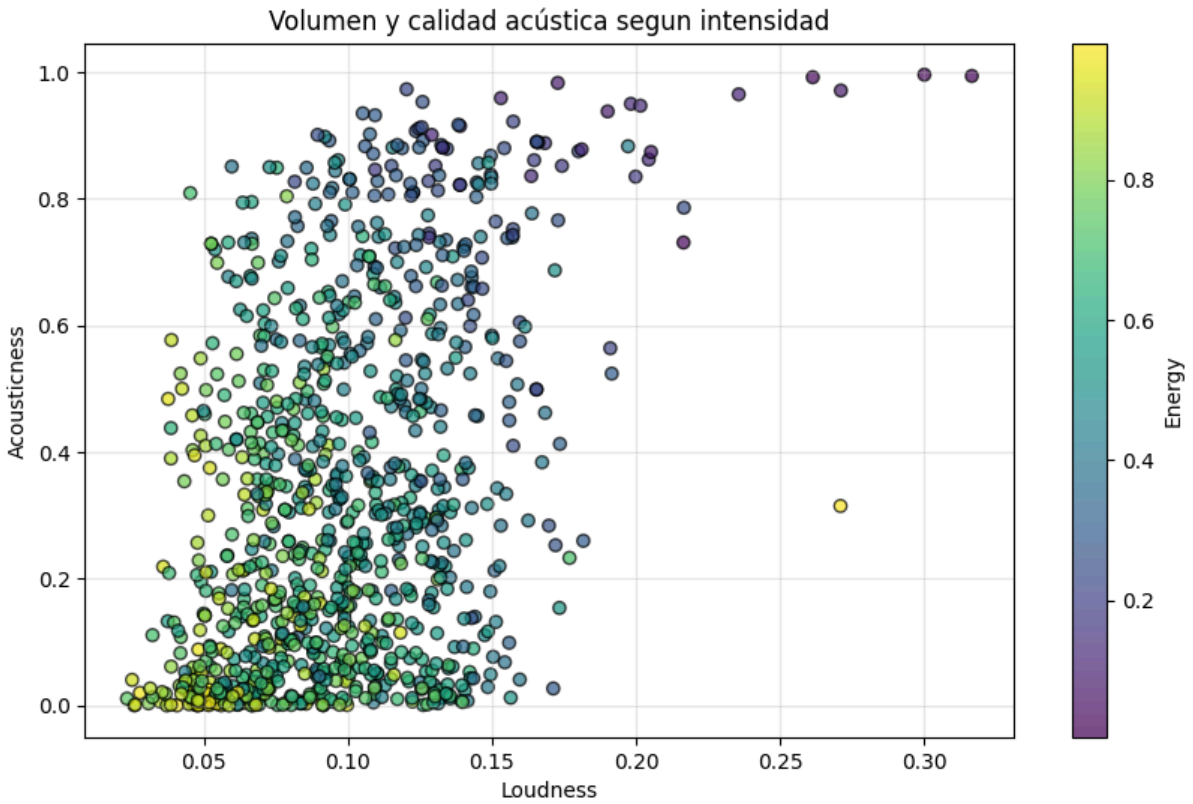
```
Test de Mann Whitney U para positividad musical: Estadístico=153004.000, p-valor=0.000
```

Como nuestros conjuntos tienen homogeneidad de varianzas, hicimos un test de Mann Whitney U y obtuvimos que SI hay una gran significancia estadística como supusimos, así que fuimos a hacer el mismo test a una cola para comprobar nuestra hipótesis y el resultado fue el siguiente:
Se rechazó la hipótesis nula, probando así la hipótesis alternativa propuesta, con un p-valor ínfimo de `2.064869822831447e-49`. Entonces, estamos en condiciones de afirmar que validamos nuestro supuesto: **Las canciones adecuadas para bailar tienden a ser más alegres**.

Hipótesis 2

El volumen y la calidad acústica de una canción son suficientes para determinar su nivel de intensidad y actividad?.

Para responder a esa pregunta y validar nuestra hipótesis, las variables de nuestro dataset que nos ayudan a hacerlo son **Loudness, Acousticness y Energy.** Y para empezar vamos a graficar un scatter plot utilizando a Energy como nuestro mapa de color para intentar encontrar alguna posible correlación entre estas 3 características.



Observamos que se puede encontrar un patrón, donde los mayores valores de Energy están concentrados en los menores valores de Loudness y Acousticness, y van decreciendo a medida que los valores de Loudness y Acousticness aumentan. Significando que por lo pronto parece haber una relación marcando que a medida que una canción tiene un volumen muy elevado y su calidad acústica es baja, más intensa es; y por el otro lado, a medida que una pista comienza a tener una calidad acústica elevada con un volumen que disminuye, su intensidad baja progresivamente.

Pero esto es sólo un gráfico así que vamos a implementar una regresión lineal múltiple para saber si nuestras dos características principales son capaces de determinar la intensidad y actividad de una canción.

OLS Regression Results

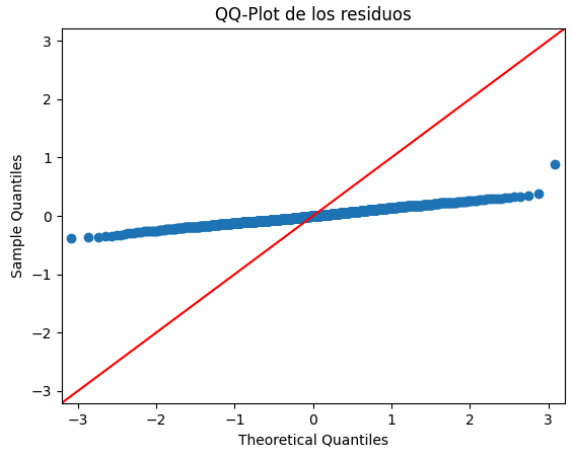
Dep. Variable:	Energy	R-squared:	0.580
Model:	OLS	Adj. R-squared:	0.580
Method:	Least Squares	F-statistic:	675.8
Date:	Tue, 26 Nov 2024	Prob (F-statistic):	5.51e-185
Time:	00:02:27	Log-Likelihood:	601.18
No. Observations:	980	AIC:	-1196.
Df Residuals:	977	BIC:	-1182.
Df Model:	2		
Covariance Type:	nonrobust		

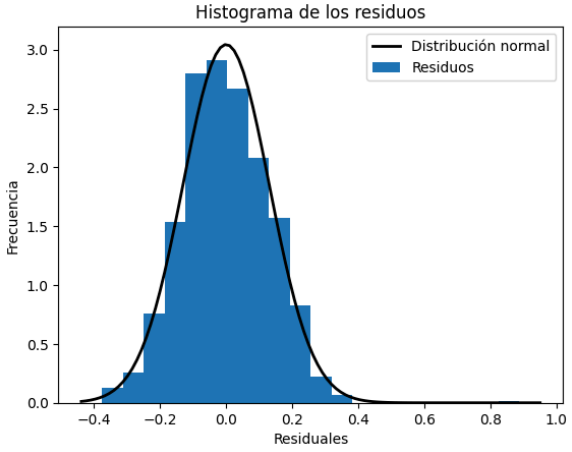
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5809	0.004	138.575	0.000	0.573	0.589
Loudness	-0.1032	0.005	-22.497	0.000	-0.112	-0.094
Acousticness	-0.0799	0.005	-17.413	0.000	-0.089	-0.071

Omnibus:	57.049	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	129.520
Skew:	0.338	Prob(JB):	7.50e-29
Kurtosis:	4.648	Cond. No.	1.54

Luego de estandarizar los datos y entrenar el modelo, obtuvimos un coeficiente de determinación de 0.58 indicando que **nuestro modelo entrenado explica el 58% de la varianza de nuestra variable dependiente Energy**, lo cual está bastante bien. Además tenemos un p-valor ínfimo así que nuestro modelo tiene una mayor varianza explicada que la varianza residual, volviendolo **significativo**.

Pero todo esto no tiene validez al comprobar que nuestros residuos no siguen una distribución normal, fijándonos en el p-valor del Test de Omnibus que rechaza la hipótesis nula de normalidad de los datos.





Por lo tanto, **no tenemos evidencia para validar nuestra segunda hipótesis.**

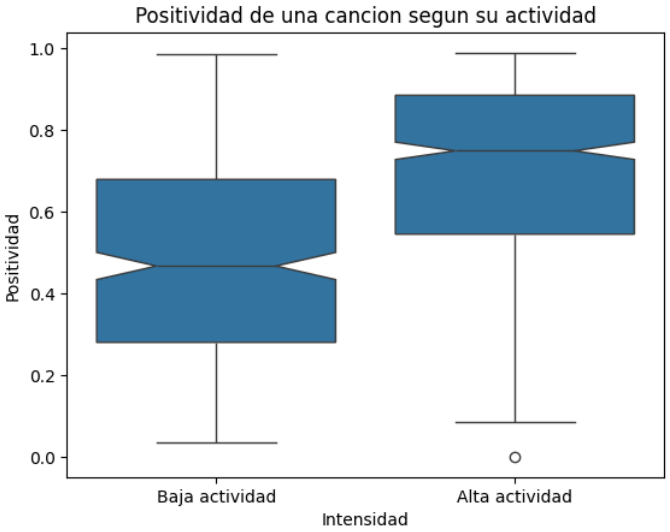
Hipótesis 3

Esta tercera hipótesis está motivada por la curiosidad de encontrar patrones en común que puedan hacer a la alegría de una canción, partiendo del buen resultado de la hipótesis 1 donde comprobamos que la danzabilidad sirve en este caso.

Para esta validación discretizamos la variable Energy generando grupos de canciones con alta y baja actividad.

Validación de supuestos

Vimos en el estudio bivariado que la correlación lineal entre estas dos variables es aceptable pero tenemos que comprobar los resultados con un test definitivo. Y en el estudio univariado comprobamos que ninguna se distribuye normal, así que completando los supuestos obtuvimos un **p-valor = 0.002** en el **Test de Levene** por lo que no tenemos homogeneidad de varianzas entre los grupos.



Kruskal-Wallis

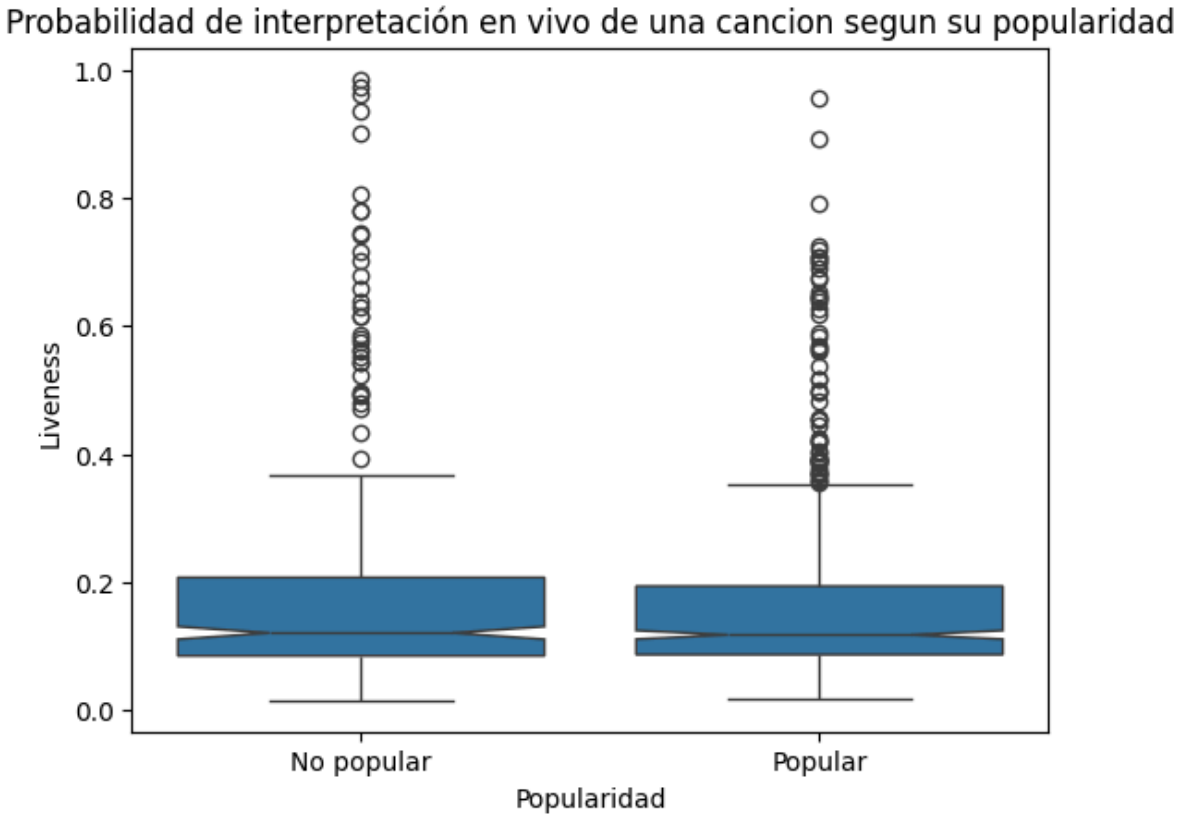
El rango de variación de la mediana no se interpone entre los grupos indicando una posible diferencia estadística que confirmamos en el test de **Kruskal-Wallis**, al no cumplir con ningún supuesto, con un p-valor mínimo de **1.3485355698272232e-35** rechazando su hipótesis nula de no diferencia estadística significativa.

Ya sabemos ahora que hay una diferencia confirmada en la positividad musical de una canción con baja actividad y otra con alta actividad, pero para conocer qué grupo es más alegre que otro usamos una comparación de medianas porque Kruskal-Wallis es sensible a ellas y obtuvimos que **las canciones con un alto grado de actividad musical son más alegres que las canciones poco intensas.**

Hipótesis 4

Vamos a validar nuestra cuarta hipótesis, y para eso vamos a utilizar las variables **Popularity** y **Liveness**.

Queremos averiguar si la popularidad de una canción tiene incidencia en su probabilidad de haberse interpretado en vivo, motivados por el hecho de que nos parece un razonamiento lógico, que los artistas elijan sus temas más reconocidos para sus recitales. Y para esto, utilizamos la misma idea que en la hipótesis 1 pero con la popularidad de una canción separándola en canciones populares y no populares respecto de su probabilidad de haber sido interpretada en vivo.



Ya por el boxplot comparativo, podemos saber que no vamos a tener normalidad en nuestros conjuntos por el intervalo de confianza de la mediana de ambos grupos y los sesgos que tienen ambos debido a lo que habíamos visto en la descripción estadística de Liveness donde el 75% de sus datos no superan el 20% de probabilidad.

```
Test de Shapiro-Wilk para canciones populares: Estadístico=0.738, p-valor=0.000
Test de Shapiro-Wilk para canciones no populares: Estadístico=0.715, p-valor=0.000
```

```
Test de Levene para Liveness: Estadístico=3.425, p-valor=0.065
```

No tenemos normalidad pero sí una homocedasticidad aceptable, así que implementamos un test no paramétrico de **Mann Whitney**.

```
Test de Mann Whitney U para Liveness: Estadístico=110799.500, p-valor=0.633
```

Al obtener un p-valor muy por encima del nivel de significancia $\alpha = 0,05$, no tenemos suficiente evidencia para rechazar la hipótesis nula del test, por lo tanto **no tenemos significancia estadística entre ambos grupos** estudiados, significando que **las canciones populares o no populares no influyen en su probabilidad de tocarse en vivo,** rechazando nuestra hipótesis planteada.

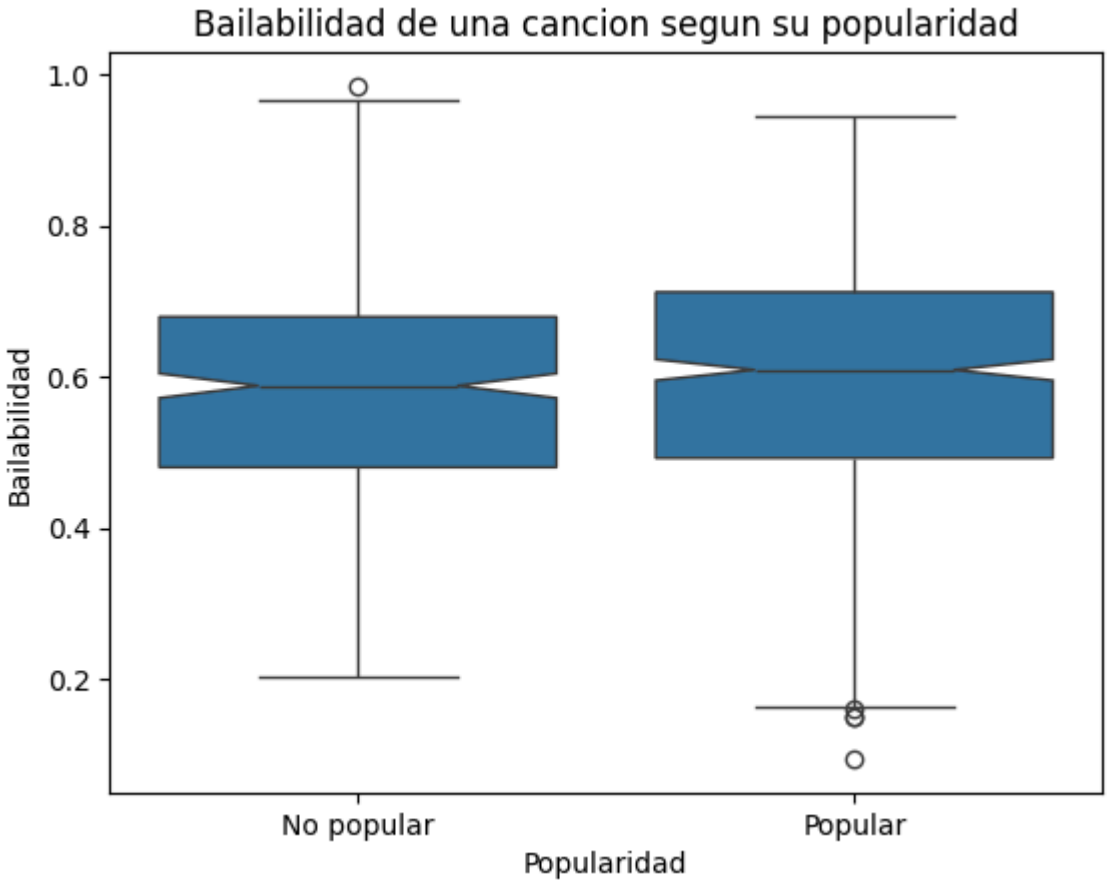
Las razones por las que no se relacionan en absoluto estas dos características pueden ser varias, comenzando por el hecho de que no tenemos contacto con el creado del Dataset y/o muestreo, lo que nos dificulta en gran medida el análisis porque no sabemos de qué forma fue evaluada la popularidad de una canción. Y tampoco sabemos en qué año fueron tomadas estas muestras, lo cual influye también en la estrategia de medida de popularidad. Sabemos que si el dataset fue creado en los años cercanos, lo más seguro es que la popularidad se haya medido sólo con las visualizaciones o vistas que tiene ese mismo tema en Youtube, pero depende del año puede haber sido también por la cantidad de búsquedas en internet, entre tantas otras formas.

Hipótesis 5

Queremos validar la quinta hipótesis. Donde planteamos que las canciones más adecuadas para ser bailadas tienen una tendencia a tener un buen nivel de popularidad. Y esto lo podemos comprobar usando las variables **Danceability** y **Popularity**.

Para simplificar nuestro análisis, vamos a aplicar una transformación a nuestro Dataset donde renombramos la variable **Danceability por Bailabilidad** y a la variable **Popularity** la renombramos por **Popular** y sus valores ahora serán 1 o 0 dependiendo si la popularidad de la canción es mayor a 50 o no.

Mostramos el **boxplot** de ambos grupos de popularidad según su bailabilidad para conocer sus distribuciones y el comportamiento de sus varianzas.



Por lo que podemos ver, tenemos muy poca diferencia entre canciones populares y no populares porque **ambos boxplot comparten todo el mismo rango de bailabilidad** y sus **rangos intercuartil Q1-Q3 son muy similares**. Incluso la mediana de sus valores son muy similares. Respecto a la varianza, fijándonos en el acogotamiento de las cajas, son prácticamente idénticas así que ya obtenemos un spoiler de que las diferencias entre estos grupos no van a ser significativas.

Para validar nuestra hipótesis, tenemos que validar primero la **normalidad** de los datos usando **Shapiro-Wllk** y la **homocedasticidad** de varianzas usando un test de **Levene**, para luego poder saber si podemos aplicar un test paramétrico como el Test-t o alguno no paramétrico.

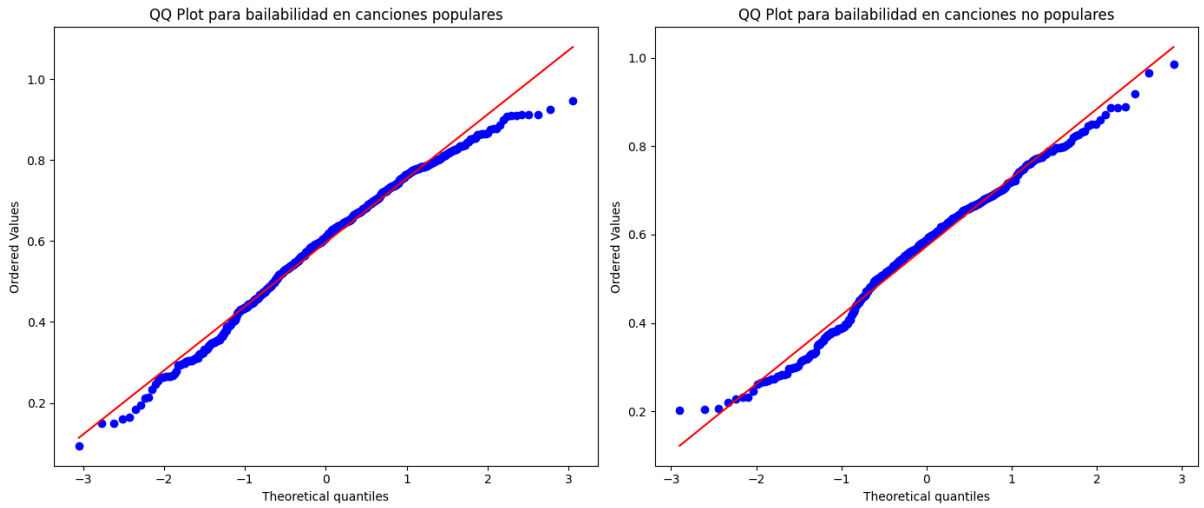
Validamos normalidad

Usando el test de **Shapiro-Wilk** la hipótesis nula es la normalidad de los datos, así que si el p-valor que obtenemos es menor al nivel de confianza **α=0.05** estaremos ante datos que no respetan una distribución normal y deberemos utilizar tests no paramétricos para validar nuestra hipótesis.

Luego de ejecutar **Shapiro** obtuvimos el siguiente resultado:

```
Test de Shapiro-Wilk para canciones populares: Estadístico=0.986, p-valor=0.000
Test de Shapiro-Wilk para canciones no populares: Estadístico=0.986, p-valor=0.001
```

Al obtener esos p-valor, **se rechaza la hipótesis nula** y tendremos que implementar un test no paramétrico, pero igualmente generamos un QQ-Plot para visualizar el comportamiento de los datos respecto de una distribución teórica perfecta.



Si bien el **QQ-Plot** se utiliza más que nada para tamaños de muestra mucho más grandes que el nuestro, podemos notar que ninguna de las dos categorías sigue una distribución normal porque no se adecúan exactamente a la recta de distribución teórica perfecta.

Validamos homocedasticidad

Ya no podemos realizar un test paramétrico al no cumplir con el requerimiento de normalidad, pero igualmente necesitamos validar homocedasticidad para saber que test no paramétrico implementaremos.

Utilizamos el **test de Levene** para validar la homogeneidad de varianzas donde la hipótesis nula es que las varianzas de los grupos son aproximadamente iguales. Y obtuvimos el siguiente resultado:

```
Test de Levene para bailabilidad: Estadístico=0.239, p-valor=0.625
```

Obtuvimos un p-valor mayor a 0.05 por lo que tenemos homocedasticidad en nuestros datos.

Ya validamos que **no tenemos normalidad** en nuestros datos, pero si **tenemos homocedasticidad**, por lo que lo ideal para nosotros es realizar un test **Mann Whitney U** porque sólo requiere homocedasticidad al comparar medianas.

Mann Whitney U

Este método también es preferible ante la presencia de valores atípicos como nos sucede con las canciones populares.

La hipótesis nula de este test es que no hay significancia estadística entre los grupos estudiados. Y luego de ejecutarlo obtuvimos el siguiente resultado:

```
Test de Mann Whitney U para bailabilidad: Estadístico=122798.000, p-valor=0.021
```

Como vemos el p-valor nos dio menor que nuestro nivel de significancia, por lo que se **rechaza la hipótesis nula** significando que si tenemos una diferencia estadística significativa en nuestros grupos. Así que ahora vamos a tener que profundizar un poco más para saber de qué tanta diferencia estamos hablando y de qué tipo es la misma.

Para eso vamos a hacer el test a una cola de **Mann Whitney U** donde le pasamos como parámetro un **alternative greater** indicandole en este caso que calcule si la diferencia de bailabilidad de canciones populares es mayor que la de canciones no populares. Y el resultado que obtuvimos fue el siguiente:

```
Test de Mann-Whitney U a una cola (populares > no populares): Estadístico=122798.000, p-valor=0.010
```

- **Hipótesis nula:** Las canciones populares no tienen mayor bailabilidad que las canciones no populares.

Obtuvimos un p-valor menor a 0.05 que rechaza nuestra hipótesis nula y acepta nuestra hipótesis alternativa propuesta. Por lo tanto, pudimos validar nuestra quinta hipótesis concluyendo que las **canciones más adecuadas para bailar son más populares que el resto**.

Hipótesis 6

Queremos comprobar la validez de nuestra sexta hipótesis donde planteamos que la popularidad de una canción está dada por un conjunto de características que eran de gusto multitudinario en la época estudiada. Para lograr esto primero tenemos que buscar la forma de conocer si existen esas características y cuáles son. Así que decidimos utilizar métodos de reducción de dimensionalidad y extrapolarlos con la popularidad para observar su comportamiento.

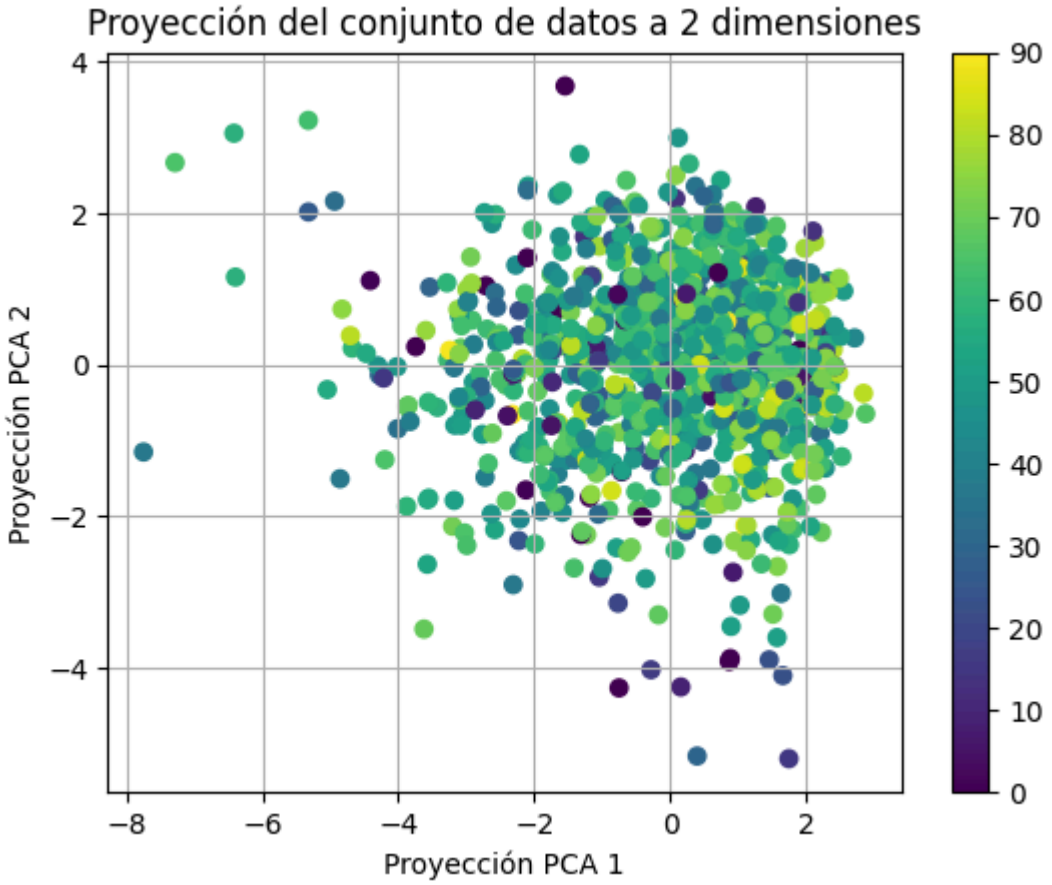
PCA

Empezamos utilizando el método de **Análisis de Componentes Principales** porque encontrar relaciones lineales en nuestros datos sería un camino fácil y rápido para darle un veredicto a nuestra hipótesis.

Para aplicar la técnica, **quitamos la variable Popularity** para utilizarla después como una tercera dimensión y que no genere ruido en el gráfico. Y vemos que redujimos el espacio de **14 dimensiones a 2**.

Luego **estandarizamos los datos** porque PCA trabaja con los desvíos estándar y la variabilidad que arrastra cada componente principal, entonces si no estandarizamos,

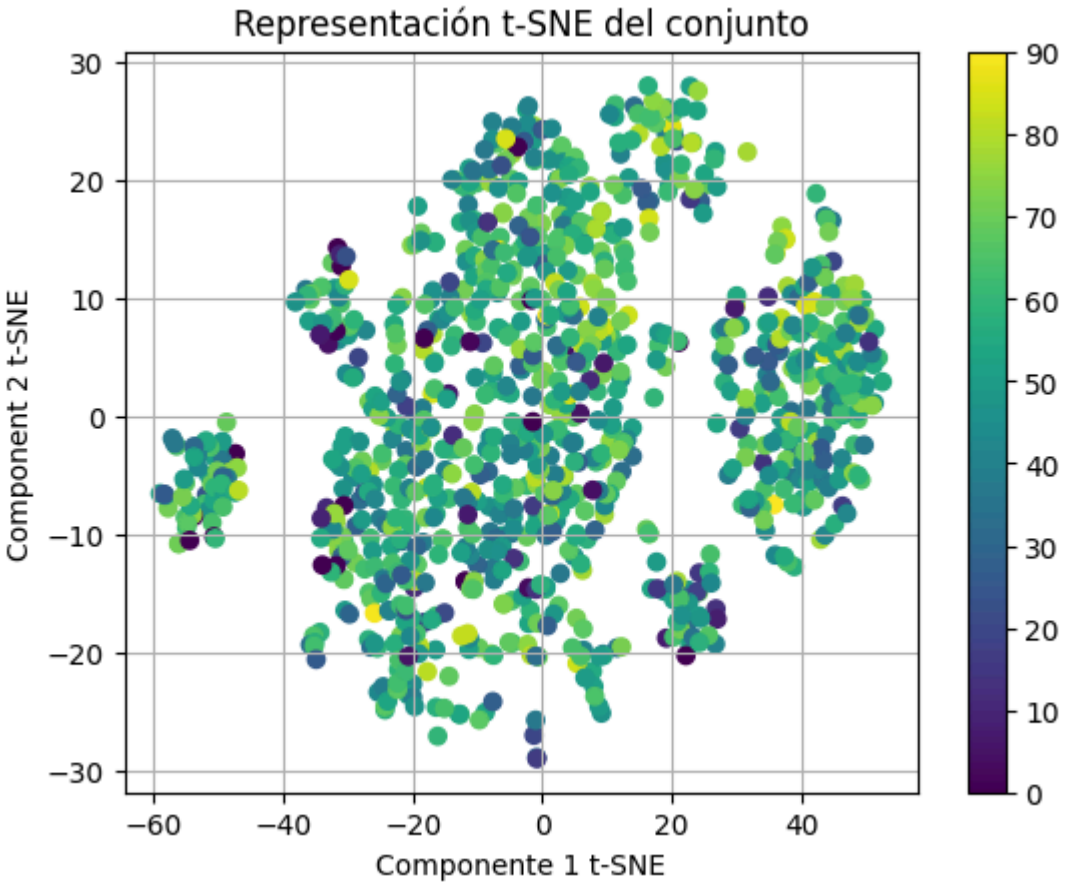
cualquier variable con una varianza muy elevada nos puede dar una falsa lectura del análisis.



Al revelar la varianza explicada por las dos componentes principales, vimos que era demasiado baja de ambas partes (**19%** y **10%,** respectivamente), lo que indica que este enfoque no logra capturar una proporción significativa de la información presente en los datos, no pudiendo concluir nada respecto a la linealidad de los mismos. Entonces probaremos utilizar un método no lineal como T-SNE y UMAP, que son más adecuados para capturar estructuras complejas en los datos.

T-SNE

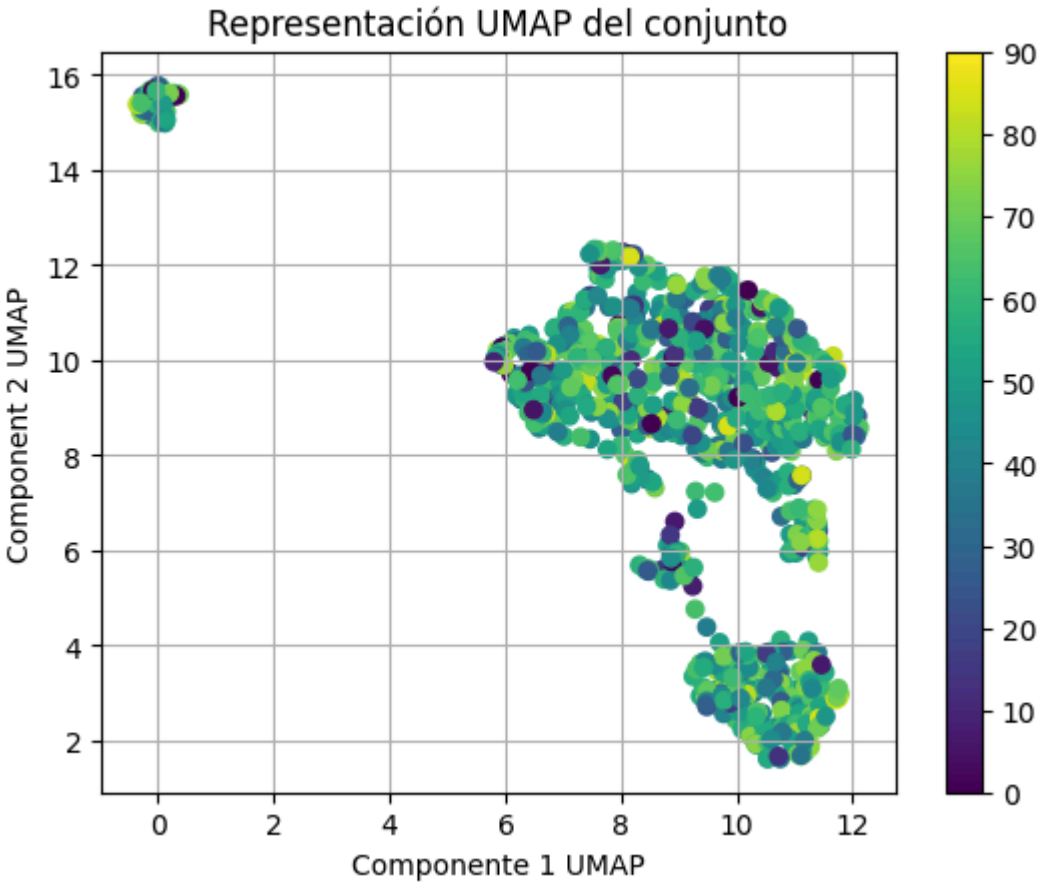
Recurrimos a métodos como T-SNE para la **reducción no lineal de la dimensionalidad** y el modelado de la variedad de los datos.



Apenas se distinguen algunos grupos formados debido a que hay determinadas cantidades de canciones que comparten características pero sigue habiendo una grandísima dispersión de colores dentro de estos grupos indicando que el formado de estos no está nada relacionado con la popularidad, así que en este caso **T-SNE no nos aportó nada de información relevante a la hipótesis**.

UMAP

Esta técnica es la única alternativa que nos queda para aplicar una reducción no lineal al no obtener ningún resultado que nos sirva con PCA y T-SNE.



Notamos que se formaron algunos grupos pero como la dispersión de valores de Popularity (graduación de colores) sigue igual de vigente que en PCA, podemos afirmar que **no existe relación alguna entre un determinado grupo de factores para conformar la popularidad de una canción**.

Por lo tanto… nuestra **sexta hipótesis es errónea**.

Regresión Lineal Múltiple

Si bien ya redujimos dimensionalidad y notamos que ningún conjunto de características atrae relación con ciertos valores de popularidad, igualmente vamos a ajustar un modelo de regresión lineal múltiple para este caso sólo por la curiosidad de saber, seguramente, qué tan malo puede resultar un modelo de predicción con las características mencionadas antes.

Primero utilizamos el mismo dataframe estandarizado que usamos en PCA pero recuperando la información del nombre de las columnas que en su momento no nos servían pero ahora si.

Luego de la estandarización, entrenamos el modelo con una constante como referencia agregada a los datos y la variable Popularity como nuestra independiente.

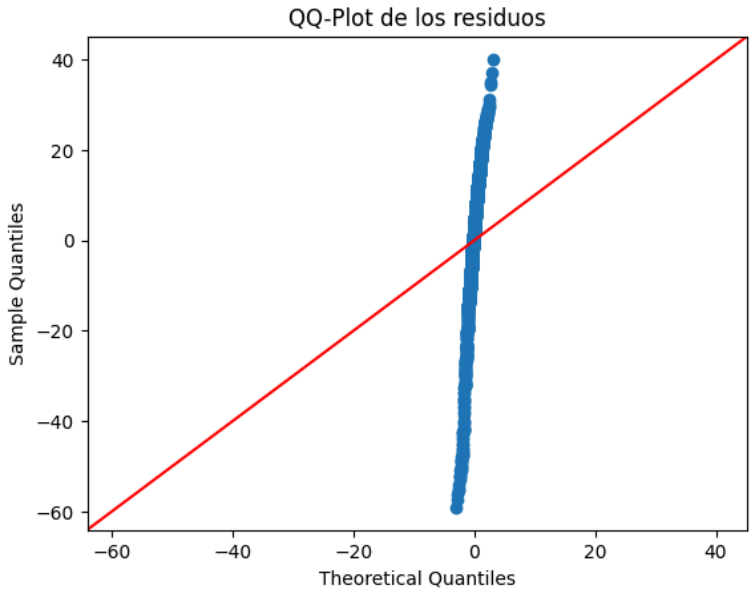
OLS Regression Results

Dep. Variable:	Popularity	R-squared:	0.059
Model:	OLS	Adj. R-squared:	0.045
Method:	Least Squares	F-statistic:	4.318
Date:	Sun, 10 Nov 2024	Prob (F-statistic):	1.76e-07
Time:	15:27:37	Log-Likelihood:	-4207.1
No. Observations:	980	AIC:	8444.
Df Residuals:	965	BIC:	8517.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.2898	0.570	93.487	0.000	52.171	54.408
Duration	0.6288	0.588	1.070	0.285	-0.525	1.782
Time_Signature	0.4396	0.593	0.741	0.459	-0.725	1.604
Danceability	1.9401	0.751	2.582	0.010	0.466	3.415
Energy	-0.4551	1.019	-0.447	0.655	-2.455	1.545
Key	-0.7052	0.583	-1.210	0.227	-1.849	0.438
Loudness	-1.9674	0.839	-2.346	0.019	-3.613	-0.321
Mode	-0.6538	0.591	-1.107	0.269	-1.813	0.505
Speechiness	-0.3629	0.598	-0.607	0.544	-1.536	0.811
Acousticness	-0.9137	0.746	-1.225	0.221	-2.378	0.550
Instrumentalness	-1.4764	0.600	-2.459	0.014	-2.654	-0.298
Liveness	-1.3965	0.602	-2.321	0.021	-2.577	-0.216
Valence	-1.6917	0.785	-2.155	0.031	-3.232	-0.151
Tempo	0.0347	0.594	0.058	0.953	-1.132	1.201
Year	1.3834	0.589	2.349	0.019	0.228	2.539

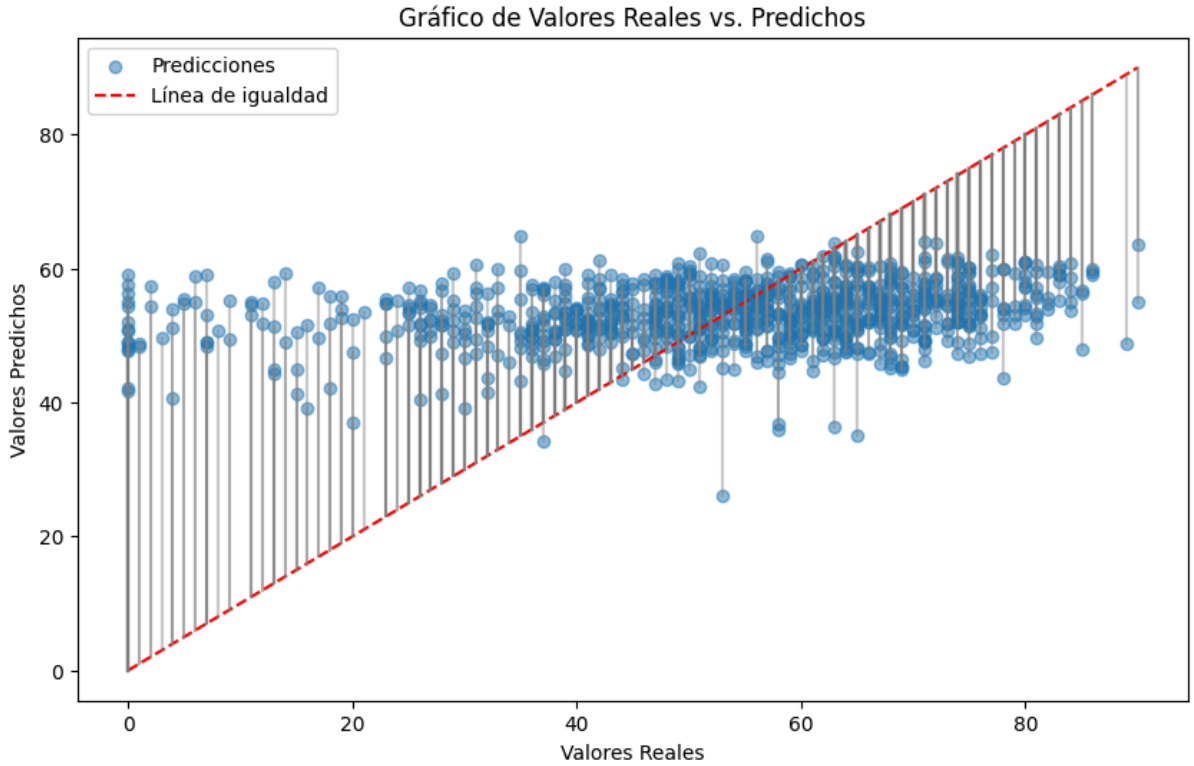
Lo primero que notamos es que el **coeficiente de determinación** R-squared es **ridículamente bajo**, significando que nuestro modelo explica solamente casi el 6% más de la varianza en los datos que un modelo trivial basado en tomar la media de todas las muestras. Cuando en un caso ideal ese número debería estar cerca de 1 equivalente al 100% de la varianza.

Ya sabemos que nuestro modelo es muy malo pero igualmente graficamos la normalidad de los residuos.




```
Test de Shapiro-Wilk para residuos: Estadístico=0.956, p-valor=0.000000
```

Efectivamente como suponíamos no tenemos nada de normalidad en nuestros residuos.

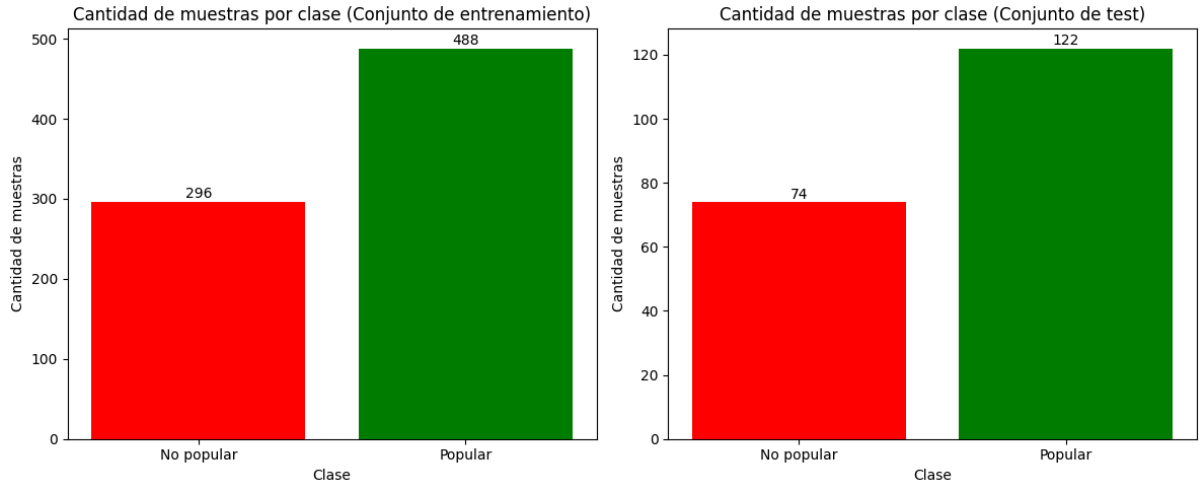


Aquí graficamos el error de nuestro modelo, donde la línea roja son los valores reales y los puntos en el plano son los valores predichos. Y vemos que hay bastante distancia entre los puntos y la recta, lo que indica un error muy elevado al estar los puntos tan distanciados y no seguir ningún patrón parecido a la recta.

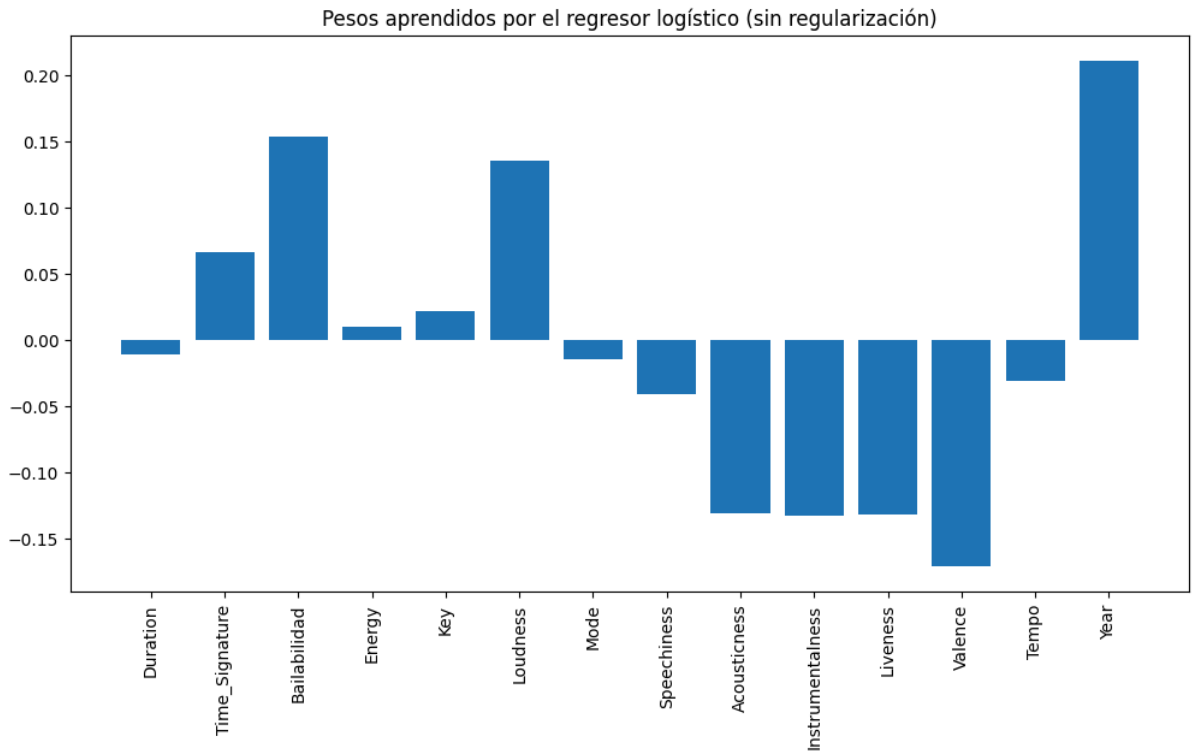
Regresión Logística

Al no obtener buenos resultados utilizando regresión lineal, vamos a simplificar un poco el objetivo y en vez de intentar generar un modelo para predecir la popularidad como variable continua, recurrimos a métodos como la regresión logística para ver si podemos predecir el valor de la **Popularidad como variable categórica** (la que transformamos al principio del Test Paramétrico) de una muestra en base al resto de las características del conjunto.

En este caso, reciclamos el Dataset que utilizamos en los test paramétrico y no paramétricos, transformando la variable Duration como al principio del informe, y asignando la variable dependiente e independientes para nuestro modelo. Luego **estratificamos** los datos reservando un **20% de los mismos para el conjunto de prueba** y manteniendo la misma proporción de clases, las cuales comprobamos graficando la cantidad de muestras de cada conjunto.



Luego de estandarizar los datos y ajustar nuestro modelo de regresión logística, graficamos las diferencias de pesos que tiene cada variable independiente respecto a su inferencia sobre nuestra variable dependiente “Popular” en un modelo evaluado únicamente sobre los datos de entrenamiento, sin regularización de ningún tipo y en comparación con el mismo modelo regularizado haciendo que sea más simple y menos sobre ajustado.

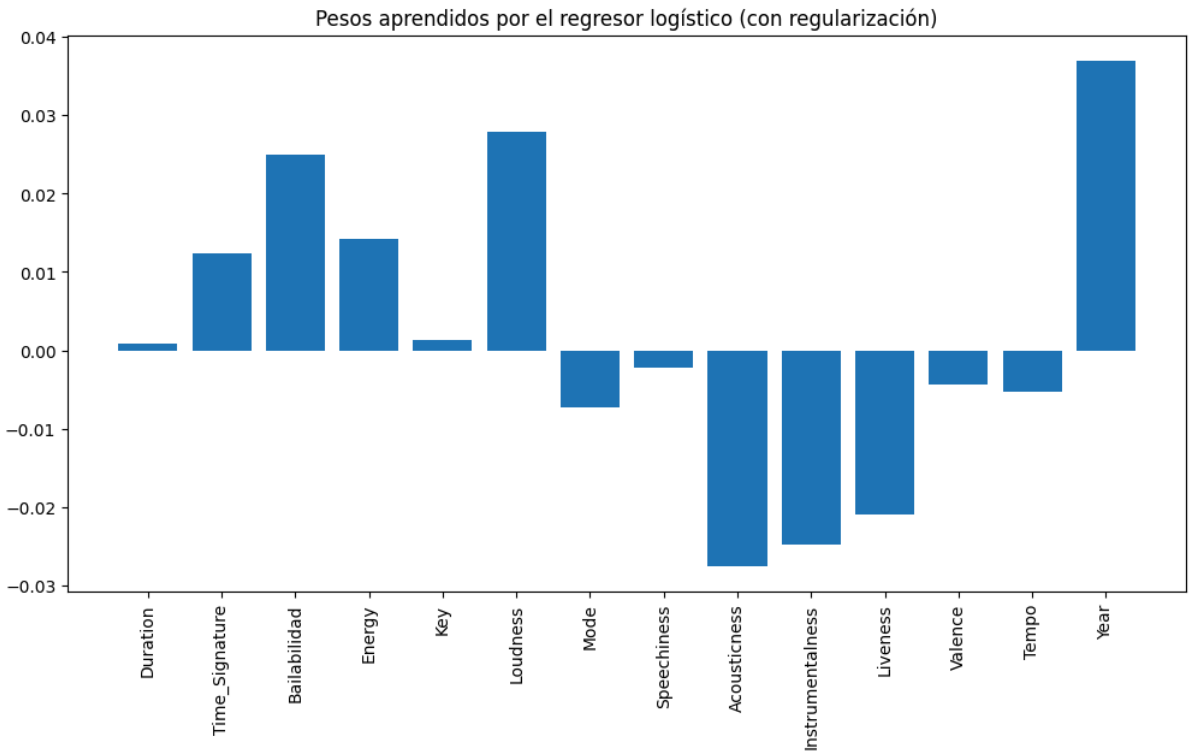


En este primer modelo base, vemos que se le da un peso mayoritario a las características del año de una canción, la bailabilidad y su positividad musical (Valence) por sobre otras.

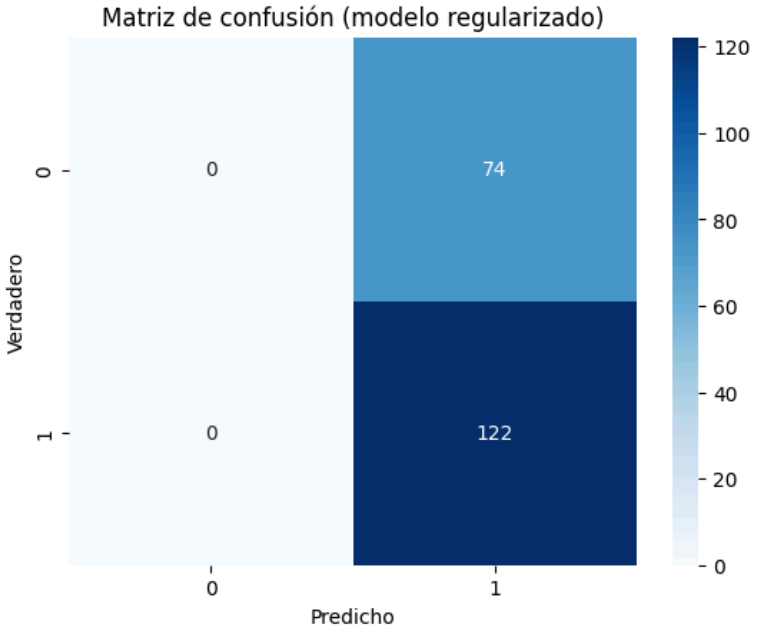
Procedemos a mejorar la validación del modelo usando el método de **validación cruzada** dividiendo el conjunto de entrenamiento en 5 partes iguales y re-entrenando el modelo 5 veces usando 4 partes y 1 dedicada a evaluación. Obteniendo así una exactitud de predicción del **casi 61%** con un desvío de subconjuntos muy bajo. Pareciendo que nuestro modelo podría ser aceptable.

Luego de esto, regularizamos nuestro modelo y le marcamos un límite de iteraciones para que tenga margen de aprendizaje y ajuste. Obtuvimos una nueva exactitud promedio de **62,24%** aproximadamente con un valor de **regularización C muy bajo**, lo que es una buena señal. Pero si bien obtuvimos un mejor rendimiento que en el modelo base sin regularización, esta diferencia no termina siendo significativa para nuestro conjunto.

También probamos diferentes valores de **hiperparámetros**, cada uno de ellos más mínimo y **todos convergían en el valor 62,24%** encontrando ahí un pico de rendimiento en nuestro conjunto de datos.



Por último, hacemos una **matriz de confusión** para conocer cuántos valores de nuestro conjunto de entrenamiento predice correctamente nuestro mejor modelo ajustado y obtenemos lo siguiente:



Teniendo como objetivo la variable **Popular** que nos indicaba con un 1 si la canción es popular en la década del 70 (con un valor mayor a 50) o un 0 en caso contrario, notamos que nuestro mejor modelo que lograba atinar con un 62% de exactitud, tiene una **tendencia clara a predecir por canciones populares** únicamente, cuando ya comprobamos que en nuestro conjunto de entrenamiento tenemos un total de 296 canciones no populares y 74 en el conjunto de test.

En conclusión… **nuestro modelo de regresión logística no sirve para predecir la categoría popular de una canción según el resto de características del dataset**.

Conclusiones

Recibimos un conjunto de datos que creemos pudimos exprimir lo máximo posible usando todos los contenidos de las materia, desde identificación de tipos de datos y relacionales, hasta la visualización y el planteamiento de hipótesis en base a ello y sus validaciones con todas las técnicas estudiadas.

Nuestro conjunto de 980 canciones de la década del 70’ se conforma de distribuciones no normales en la gran mayoría de los casos y con poca relación entre sus características, lo cual nos perjudica a la hora de plantear hipótesis interesantes y más aún a las conclusiones que llegamos con las mismas. Porque a la hora de hacernos preguntas sobre el conjunto pensamos cuáles pueden ser las características más interesantes para estudiarlas y decidimos enfocarnos en dos ramas: la positividad musical de una canción y su índice de popularidad. Estas dos creemos que le pueden dar una información útil al lector futuro de este informe interesado por estas muestras.

Llegamos a la conclusión de que la positividad de una canción está dada en gran parte por su capacidad de que la gente pueda bailarla y su intensidad del ritmo (la cual descubrimos que no se genera teniendo en cuenta el volumen y la calidad acústica), donde un tema alegre se conforma de una buena capacidad de baile y una alta intensidad en los sonidos. Y por el otro lado, descubrimos que un buen nivel de popularidad en la época estaba afectado por las canciones que eran bailables y por ningunos otros factores, donde también la popularidad de una canción no era información importante que los artistas usaran para elegir su repertorio a interpretar en vivo.

Algunos de los problemas que tuvimos fueron a la hora de plantear las hipótesis porque no tenemos contacto con el creado del Dataset y la forma en que fueron obtenidos los datos de cada característica, entonces no sabemos por ejemplo en que se basan los valores de la alegría de una canción y popularidad, porque cambiaría bastante la forma de ver el trabajo si conocemos que provienen de algún dato objetivo de la época o es un número subjetivo que le asignó la persona que juntó todas estas muestras.

En fin, el desarrollo de todo este trabajo nos resultó muy interesante principalmente por la implementación y práctica de todos y cada uno de los contenidos vistos en la materia.