Informe de Fundamentos de la Ciencia de Datos

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

Integrantes

- Videla, Braian.
- Perna, Ignacio Agustin.

Nuestro Dataset

Para comenzar, analizamos 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. Así que estamos en condiciones de comenzar con el análisis exploratorio de nuestros datos averiguando cuales son estas variables, sus tipos, datos nulos o faltantes, etc..

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 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 medusa 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.

Instrumentaleness: 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.

Planteamiento de hipótesis

Al principio planteamos algunas hipótesis lineales de correlaciones directas que surgen de nuestra curiosidad después de haber visto los tipos de datos con los que estamos tratando, para comprobar si realmente es como pensábamos o estábamos equivocados en cuanto al significado de las variables.

1. La medida que indica el número de pulsaciones por compás está directamente relacionada con el ritmo de la pista medida en pulsaciones por minuto que a su vez está relacionado con la intensidad de la canción. A mayor pulsaciones por compás, mayor las pulsaciones por minuto, y mayor actividad. Esta hipótesis la podemos validar usando **Time_Signature**, **Tempo**, **Energy**.

2. Si una canción es muy adecuada para bailarla, tenderá a ser más alegre. Esta hipótesis la podemos validar usando **Danceability** y **Valence**..

3. Una canción con volumen muy elevado, debe ser más enérgica. Esta hipótesis la podemos validar usando **Loudness** y **Energy**.

4. Si una pista tiene muchas probabilidades de haber sido interpretada en vivo, entonces es muy probable que tenga un volumen elevado. Esta hipótesis la podemos validar usando **Liveness** y **Loudness**.

5. Si una canción tiene popularidad alta, es muy probable que se haya interpretado en vivo al tener mayor ruido de audiencia o viceversa. Esta hipótesis la podemos validar usando **Popularity** y **Liveness**.

6. Las canciones más adecuadas para bailar son más populares que otras teniendo en cuenta la época. Esta hipótesis se puede validar usando a **Danceability** y **Popularity**.

7. Una pista muy instrumental tiene valores directamente opuestos a la cantidad de palabras habladas y viceversa. Esta hipótesis la podemos validar usando **Instrumentalness** y **Speechiness**.

La siguiente hipótesis está pensada para tratar con la totalidad de las variables y enriquecer nuestro análisis esperando llegar a conclusiones interesantes.

8. 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.

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, por lo que podríamos estar observando una distribución normal porque los datos se encuentran cerca de la media.

Danceability: Sus valores se encuentran en un rango de 0 a 1, con un desvío estándar cercano a cero, obteniendo la misma conclusión anterior.

Energy: Sus valores se encuentran en un rango de 0 a 1, con un desvío estándar cercano a cero. 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.

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, por lo que tenemos la mayoría de canciones con un nivel de ruido elevado.

Mode: Al ser una variable discreta, no tiene sentido analizar promedios, desvíos o cuartiles. 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 de 0,059923 indica que la presencia de palabras habladas en una pista por lo general es baja. El primer y tercer cuartil refuerzan esta idea con datos similares(0.031 y 0.038). Con desvío estándar cercano a 0 podemos suponer distribución normal. 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. Se puede ver una concentración significativa de pistas con niveles bajos de elasticidad. Y teniendo un mínimo con valor 15 mil veces menor que la media, habrá, outliers por el lado de la pista poco acústica.

Instrumentales: El rango va de 0 a 1. Vemos que la mayor cantidad de calores se concentran en valores muy bajos y en su mayoría son ceros o cercanos al mismo. 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. Podemos ver por el valor de la media y la relativa paridad entre cuartiles, que parece ser una distribución normal. Aunque 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 216 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. Vemos que en ésta década la popularidad variaba bastante, la mayoría tiende a ser medianamente reconocida, pero con algunos casos atípicos de gran popularidad y de baja popularidad.

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 con una iteración 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.

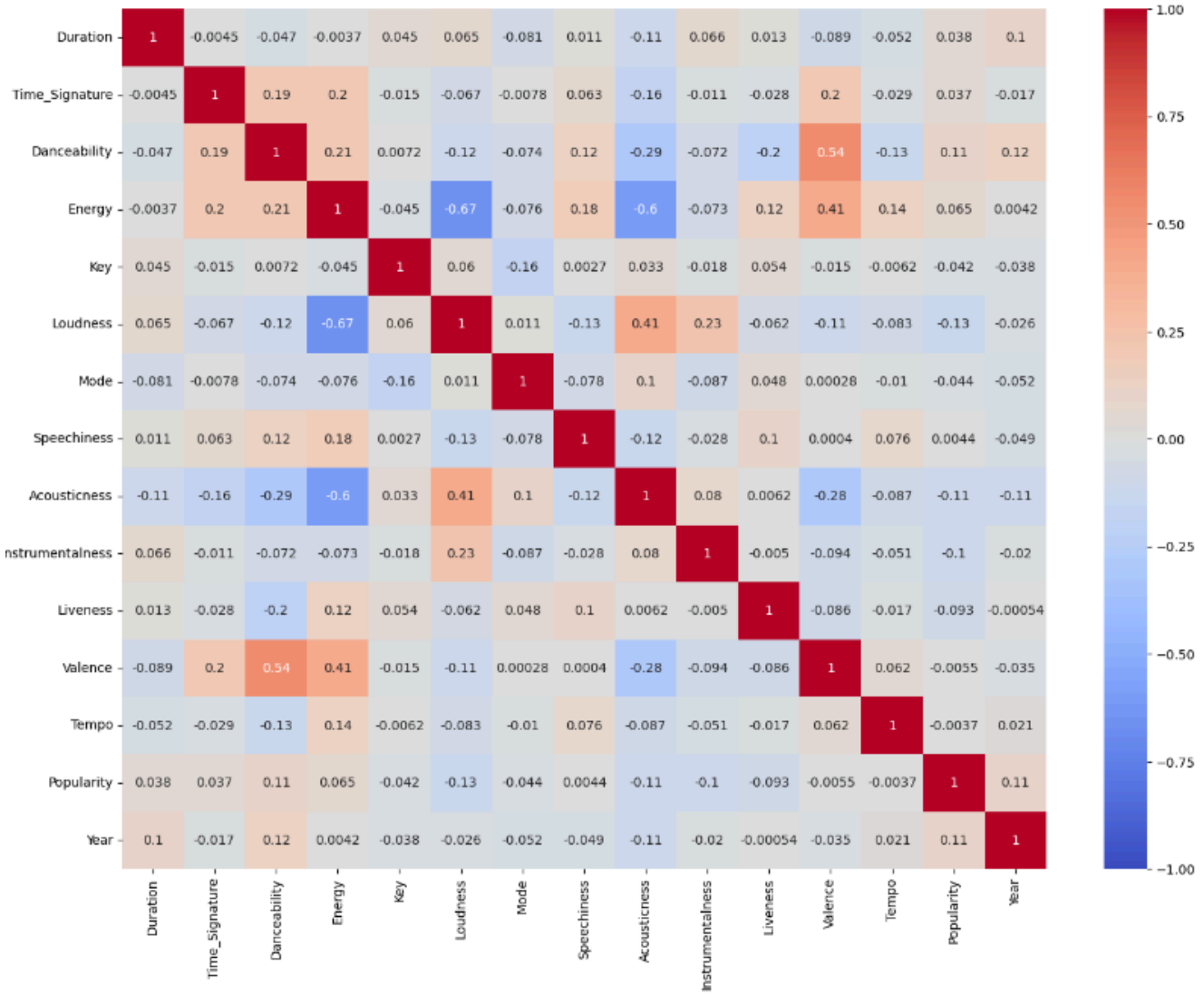
Matriz de Correlación

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 .

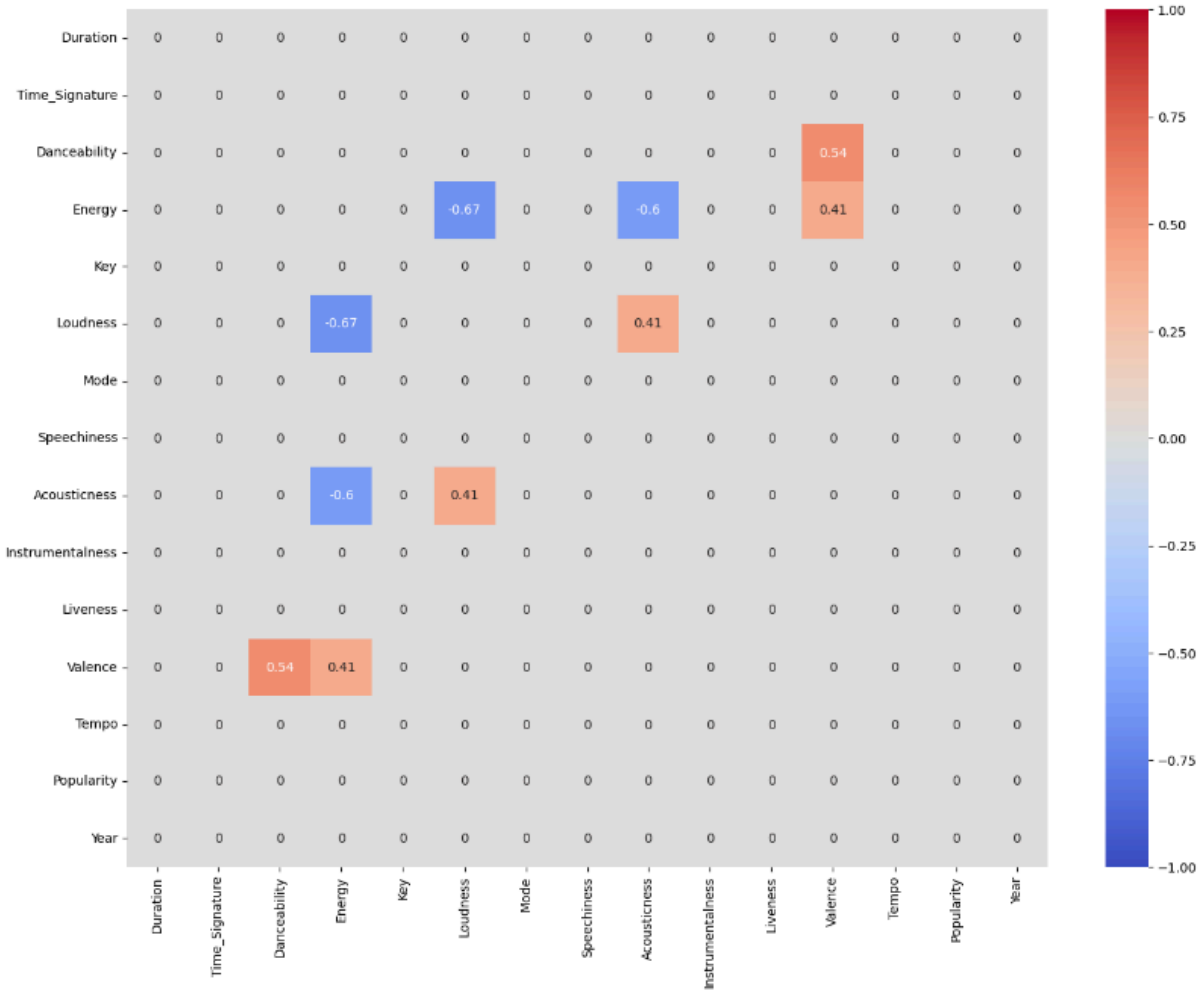
Representamos la matriz gráficamente para identificar las correlaciones elevadas o bajas, asignando colores a cada correlación. Esto lo haremos con la característica de Heatmap de la librería "Seaborn" montada sobre "Matplotlib".

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:

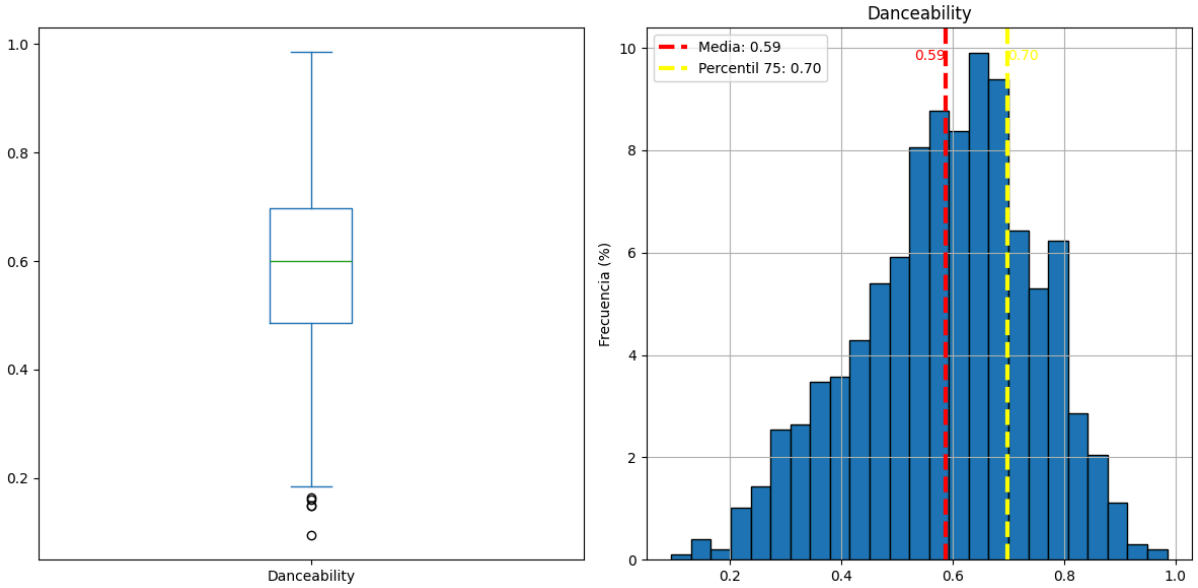
- **Danceability - Valence:** 0.54
- **Energy - Loudness:** 0.67
- **Energy - Acousticness:** 0.6
- **Energy - Valence:** 0.41
- **Loudness - Acousticness:** 0.41

Estudio de correlaciones

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.

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

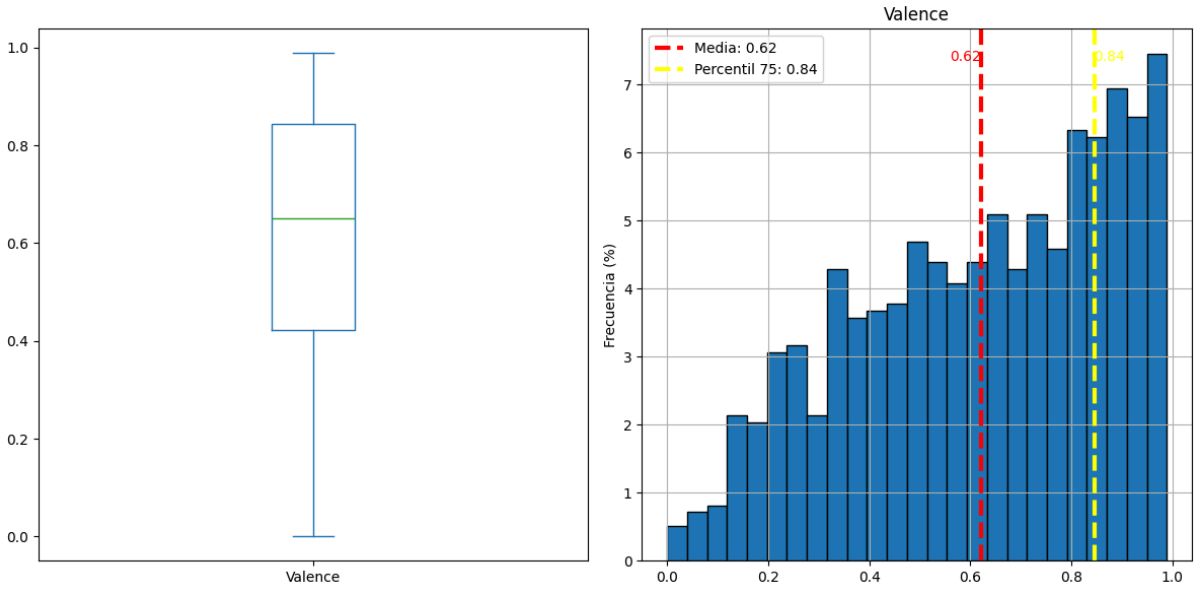
Danceability



- Coeficiente de asimetría: -0.34
- Curtosis: -0.29

Lo que vemos tanto en el boxplot como en el histograma de la variable **Danceability** es lo que habíamos supuesto en la descripción estadística de las variables, se trata prácticamente de una distribución 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



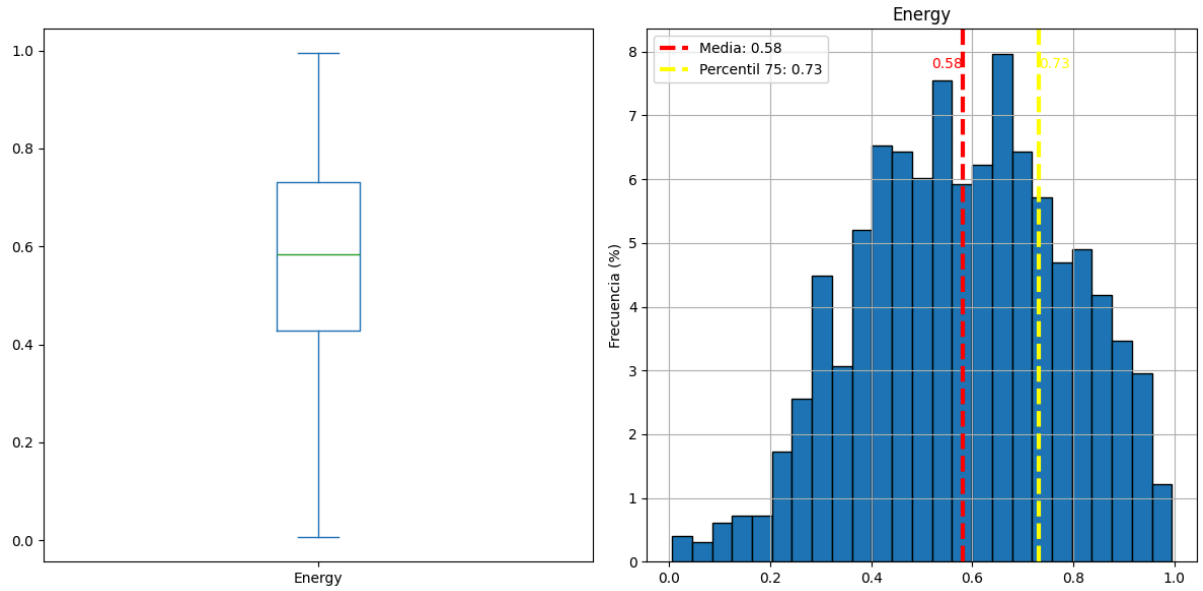
- Coeficiente de asimetría: -0.39
- Curtosis: -0.95

Observamos principalmente en el boxplot que ésta variable tiene una distribución normal 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

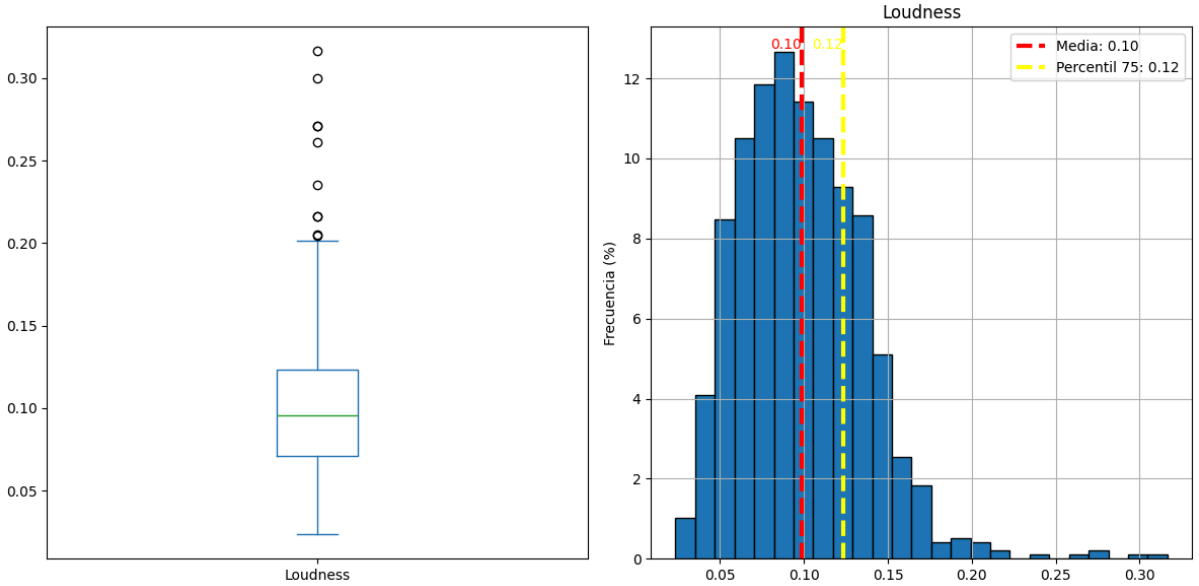


- Coeficiente de asimetría: -0.14
- Curtosis: -0.58

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.

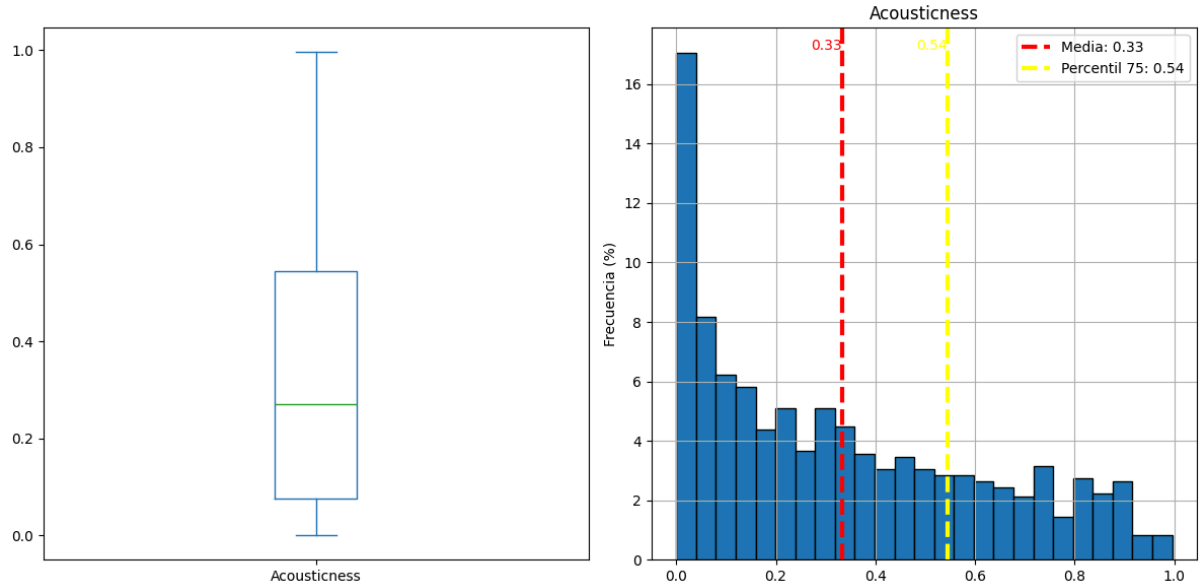
Loudness



- Coeficiente de asimetría: 0.94.
- Curtosis: 2.67

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

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 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.

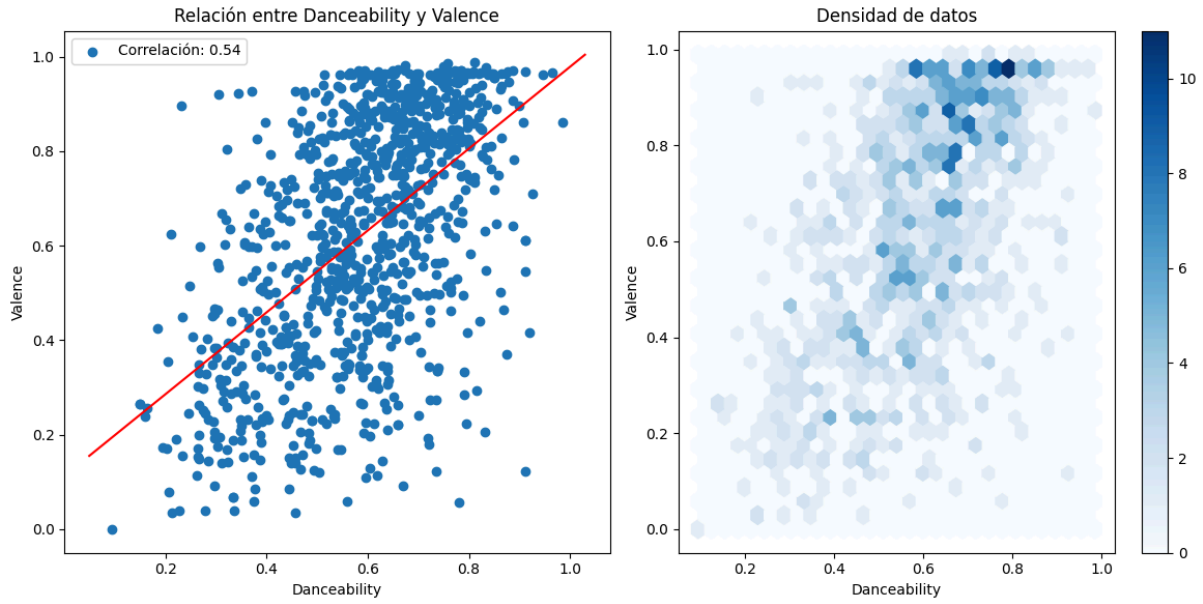
Para empezar siempre es conveniente hacer un **gráfico de dispersión** o **Scatter Plot** para ver particularmente los datos representados de cada variable y valorar su influencia o impacto.

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

Esta correlación que pudimos obtener nos viene genial para validar nuestra segunda hipótesis que planteamos… así que vamos a estudiar la relación que hay entre cuán adecuada es una pista para bailar y qué tan alegre es la misma.



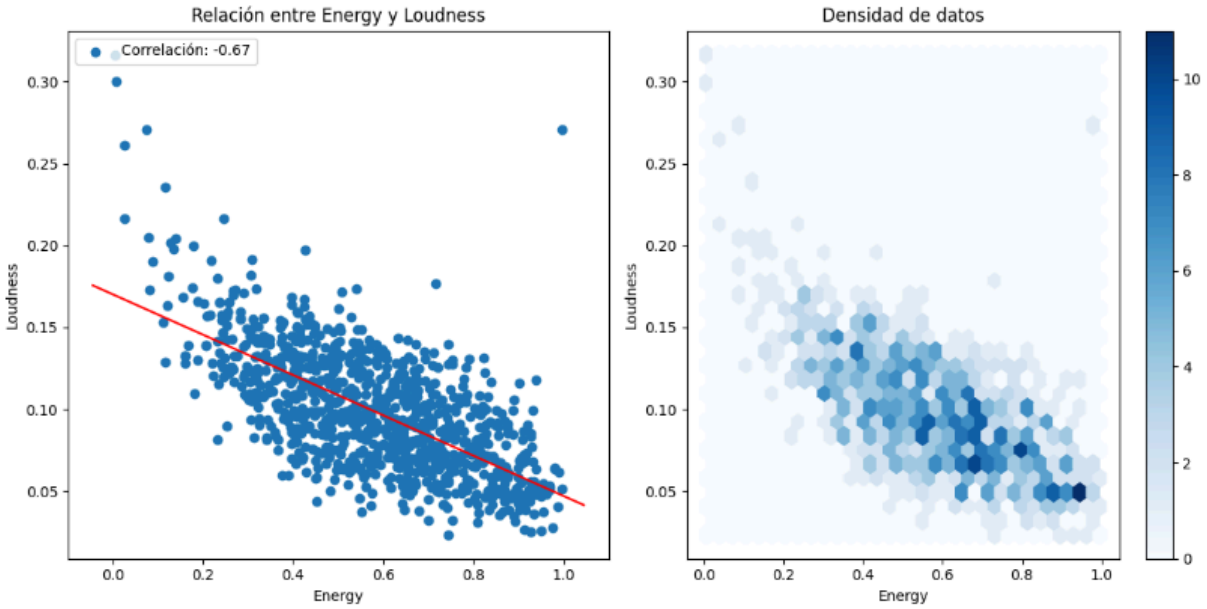
A simple vista podemos suponer que no hay ninguna relación entre que tan “buena” es una canción para danzar y su positividad musical fijándonos en el **scatter plot** por la gran dispersión que hay en los datos, donde para un mismo valor de **Danceability** tenemos distintos valores de **Valence** en 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. Por lo tanto, **validamos nuestra hipótesis**.

Energy - Loudness

Nuevamente, utilizaremos este mismo estudio de correlación de Pearson para validar nuestra tercera hipótesis: Vamos a estudiar 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).

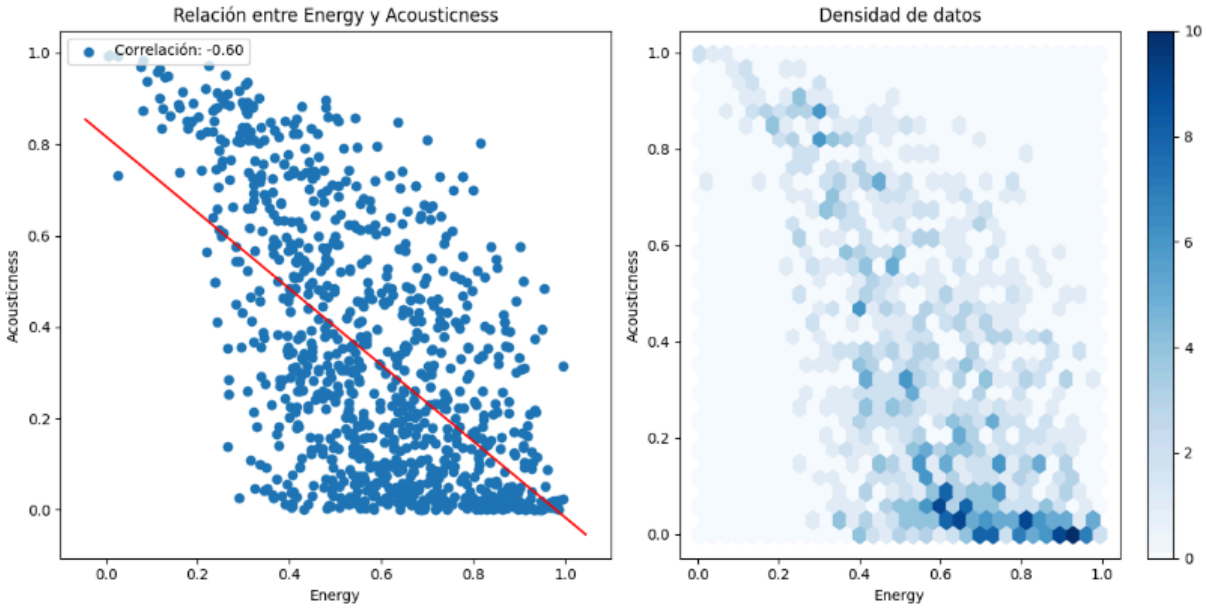
Gracias a la **línea de mejor ajuste**, si bien su pendiente es negativa, sabemos que la tendencia de los datos es positiva porque en la correlación donde hay mayor volumen de la canción, más enérgica es. Y vemos en el **Scatter plot** que los puntos están concentrados sobre la línea indicando que es 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, estamos en condiciones de afirmar que si las canciones presentan un volumen muy elevado (cercano a 0 decibelios), será más enérgica y con mucha actividad, **validando nuestra hipótesis**. 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 estudiar la relación que hay entre la medida enérgica de una pista y su calidad acústica, o que tan acústica es.



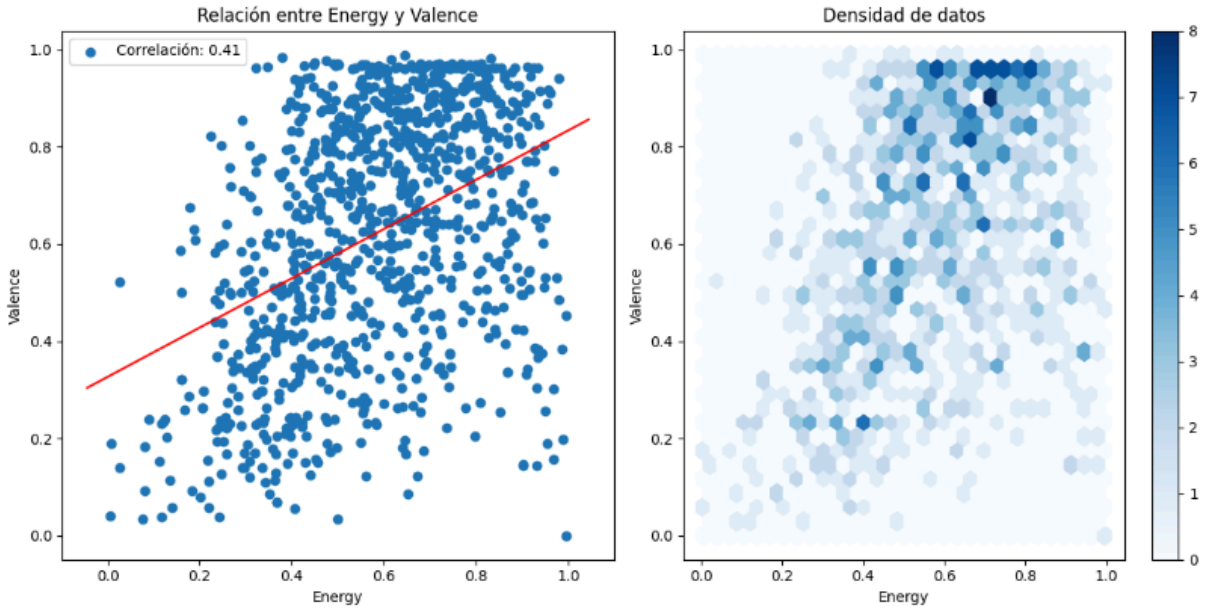
Si observamos a priori la línea de mejor ajuste, podemos destacar que la tendencia de la relación es que a mayores valores de una variable, menores valores que la otra porque su pendiente es negativa.

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 estudiar la relación que hay entre la medida enérgica de una canción y su positividad musical.



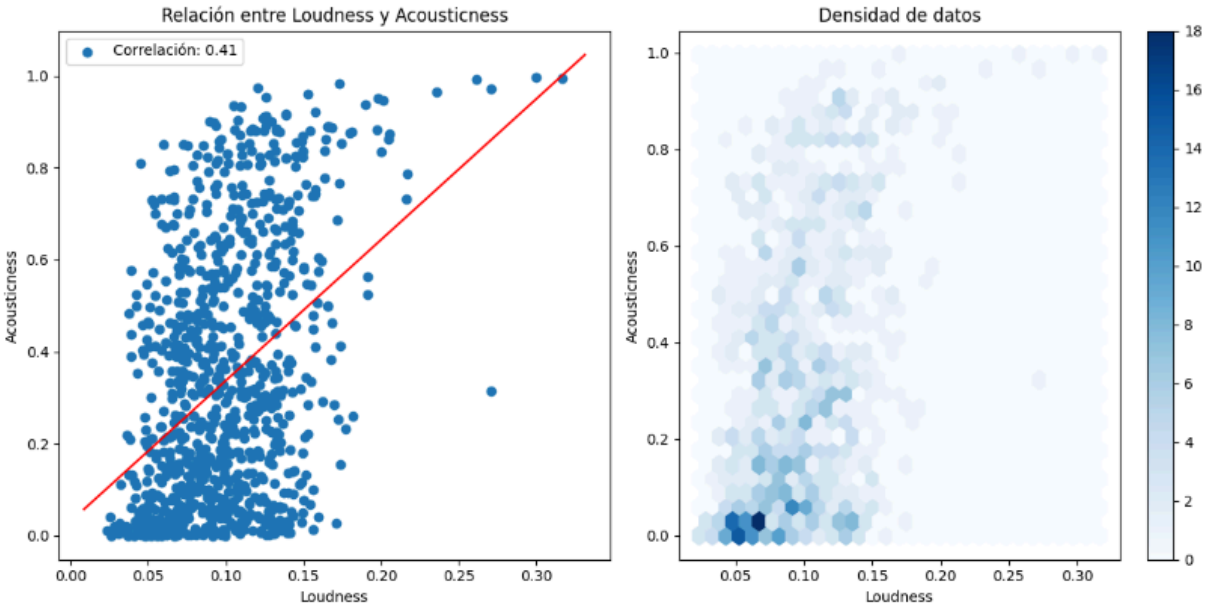
Nuestro valor de **correlación de Pearson** para ésta ocasión es de **0.41**, osea que se encuentra en el límite que decidimos poner para estudiar dichas correlaciones debido a que no tenemos ninguna que alcance el estándar aceptado de 0.7. Por lo tanto, como se considera un valor bajo se comprueba gráficamente 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 estudiar 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) podríamos asegurar en términos generales que no existe una relación (por lo menos lineal) con un volumen promedio de una canción y su nivel acústico.

Validación del resto de hipótesis

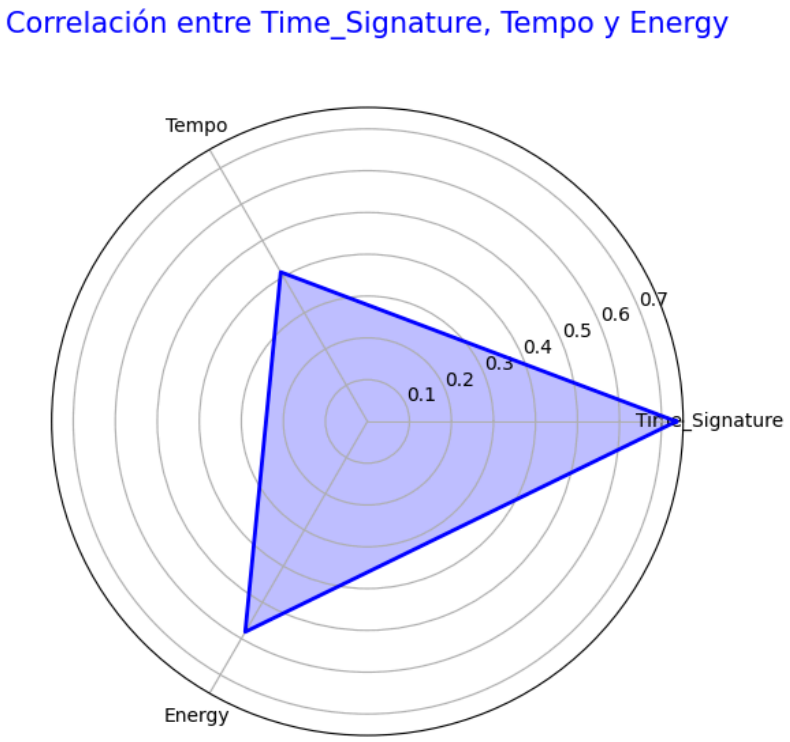
Ahora que ya estudiamos y comprobamos 2 de nuestras 8 hipótesis al tener una correlación por encima de lo que marcamos como aceptable, vamos a dedicarnos a validar todas las restantes de ahora en adelante, con diferentes técnicas.

Time_Signature - Tempo - Energy

Vamos a validar nuestra primera hipótesis.

Si nos fijamos en la descripción de las variables**,** estas 3 tienen todas diferente rango entre ellas, por lo que para estudiar cómo se comportan a la vez vamos a tener que aplicar alguna transformación, y decidimos que sea la **normalización** para que todas las características vivan en el rango [0,1].

Para representarlas gráficamente y estudiar su comportamiento utilizamos un gráfico radial, generando sus radios internos según el rango de las variables normalizadas.



Podemos observar que valores altos de Time_Signature corresponden a valores medios de Tempo y valores medio-altos de Energy, por lo que se puede deducir que no existe una relación lineal como nosotros habíamos pensado en un principio. Para que exista, en el gráfico radial o de “araña” un mismo valor de Time_Signature debería corresponder con el mismo valor de las otras dos variables.

Este resultado nos ayudó a darnos cuenta que estábamos equivocados en el concepto de Time_Signature. Nosotros teníamos la idea de que esta variable indicaba la misma medida que la variable Tempo con las pulsaciones de una canción por minuto pero llevada a una escala de hasta 4 segundos. Gracias al gráfico radial, tuvimos que investigar más de la

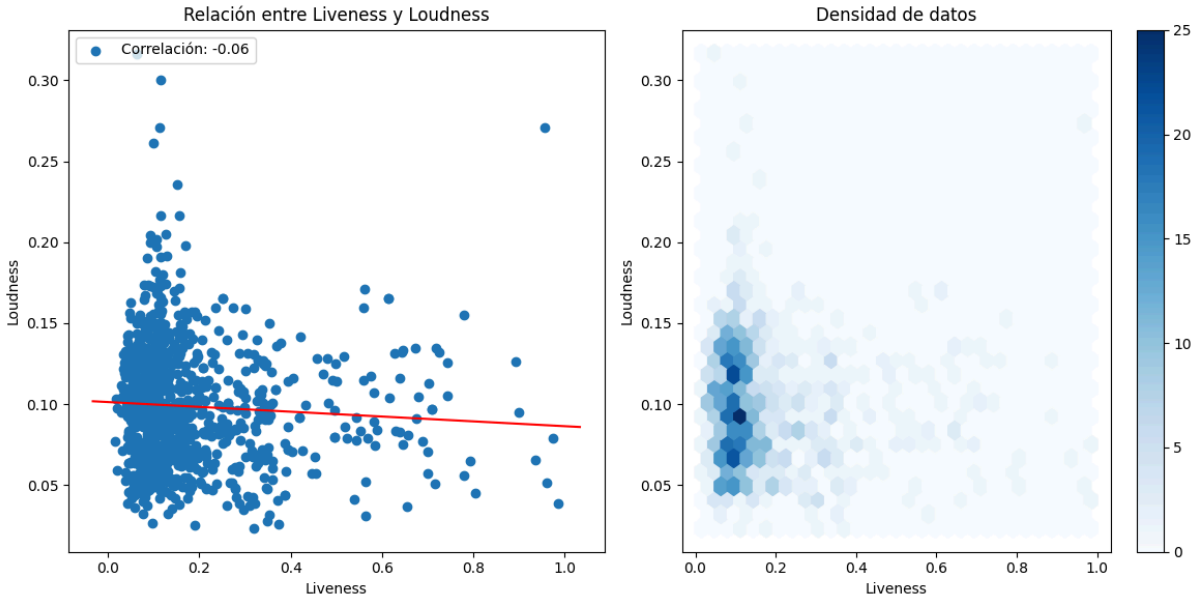
relación entre las pulsaciones por compás y las pulsaciones por minuto. Y encontramos que Time_Signature indica la estructura rítmica de la pista, marcando hasta un máximo de 4 pulsaciones por compás (4/4) y Tempo indica la velocidad a la que se tocan esas pulsaciones, por lo que no necesariamente se tiene una relación directa entre las mismas.

Ejemplo: Una pista puede tener 4 pulsaciones por compás y 50 pulsaciones por minuto haciendo un Tempo bajo con un Time_Signature considerablemente rápido.

Fuente: https://es.wikipedia.org/wiki/Comp%C3%A1s_(m%C3%BAsica)

Liveness - Loudness

Vamos a validar nuestra cuarta hipótesis.



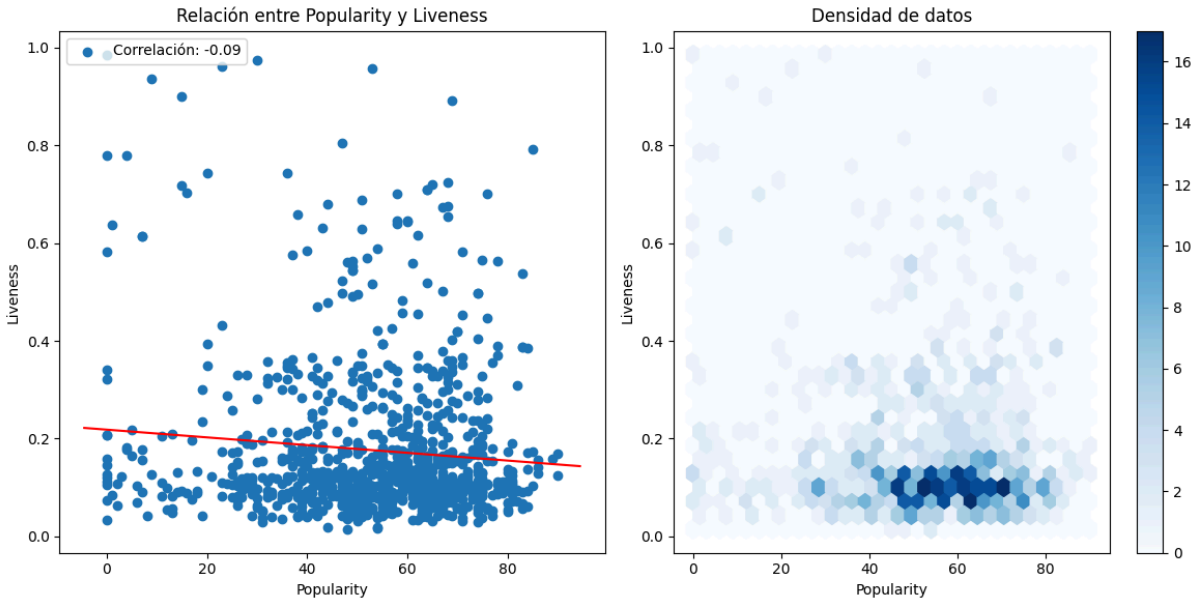
Viendo el **scatter plot** junto con el valor de la **correlación de Pearson**, la verdad que contrariamente a lo que imaginábamos, no hay relación alguna con el volúmen de una canción y su probabilidad de que se haya interpretado en vivo.

En un principio, nos planteamos esta hipótesis creímos que los artistas o bandas iban a elegir preferentemente sus pistas más enérgicas o sonoras para interpretarlas en vivo. Y como vimos anteriormente en la relación Energy-Loudness, la intensidad de una canción está relacionada con su volúmen. Pero por lo visto, nos equivocamos.

Una de las razones de la poca relación entre estas variables puede ser que en la década del 70, unos de los géneros más populares eran el R&B y el Jazz, que no se caracterizan por un volumen alto, y en la mayoría de los casos es todo lo contrario.

Popularity - Liveness

Vamos a validar nuestra quinta hipótesis.

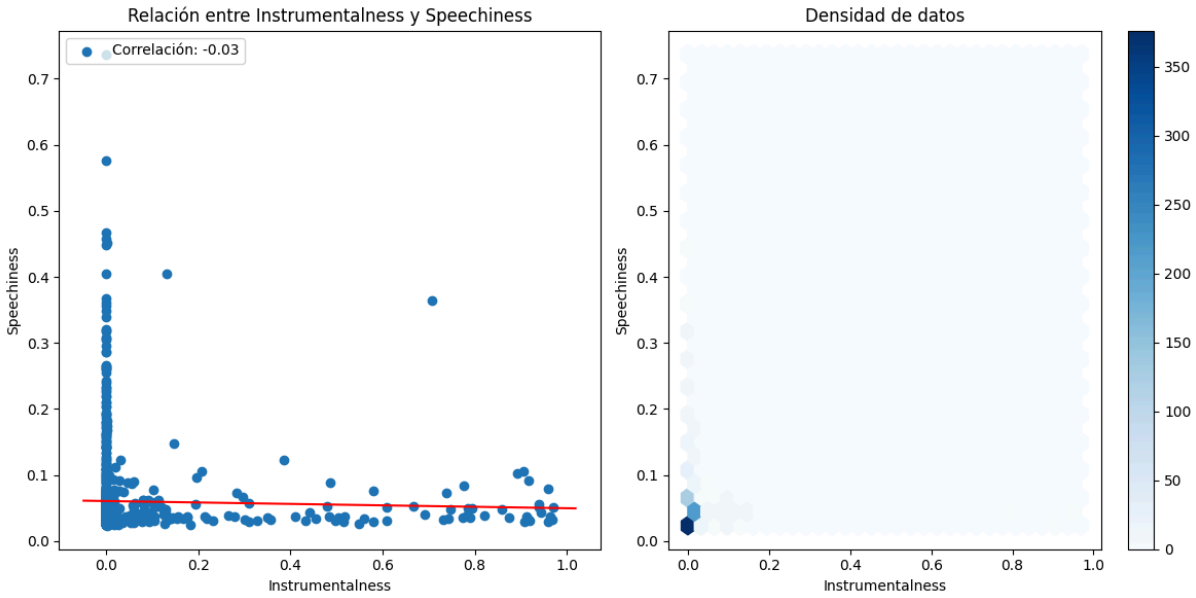


Tenemos prácticamente el mismo caso que la hipótesis anterior, con un valor de **correlación de Pearson cercano a cero**. Entonces lo que creíamos en un principio sobre si una canción tiene mucha popularidad probablemente se haya tocado en vivo, **queda refutado por los datos**.

Las razones por las que no se relacionan en absoluto estas dos variables 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.

Instrumentalness - Speechiness

Vamos a validar nuestra séptima hipótesis.



Este es por el momento el valor de **correlación de Pearson más bajo que obtuvimos**, así que **nuevamente refuta nuestra hipótesis** donde creíamos que si una pista tenía una presencia más instrumental tendría menos presencia de palabras hablas en ella.

Popularidad vs resto de características

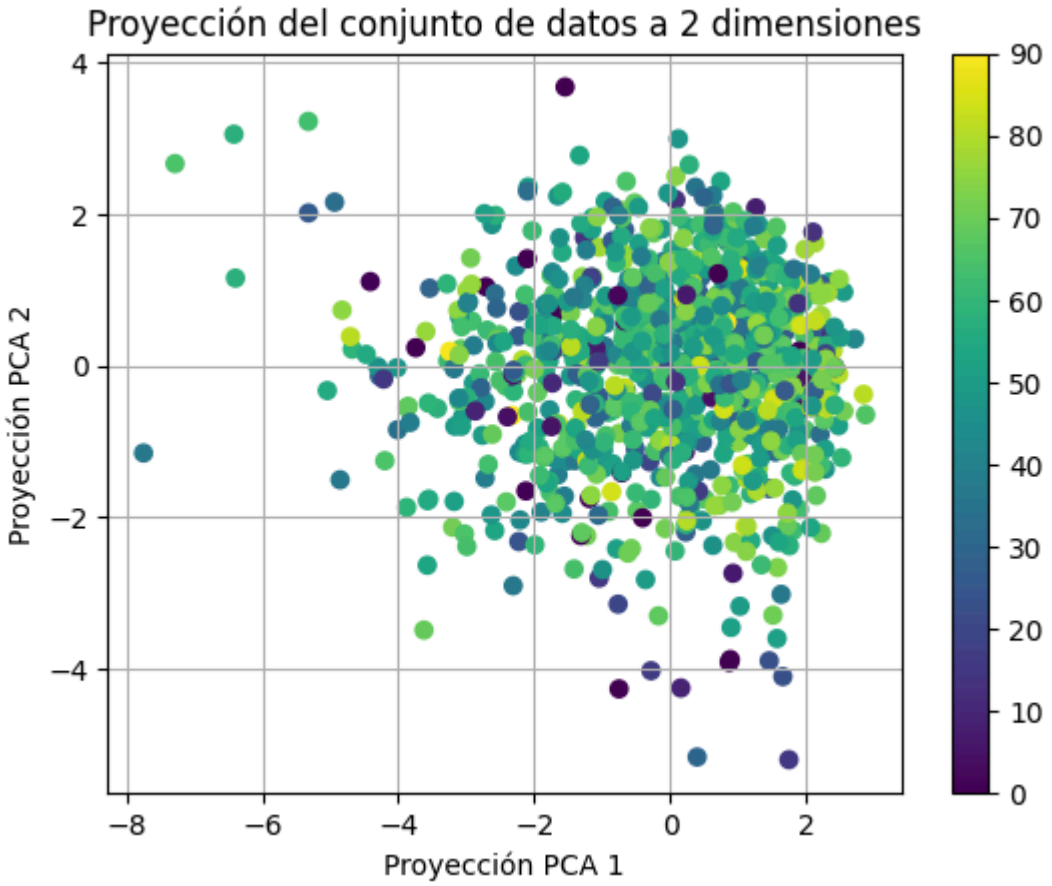
Queremos comprobar la validez de nuestra octava hipótesis donde planteamos que la popularidad de una canción está dada por un conjunto de características que eran de gusto multitudinario en la época estudiada. Para lograr esto primero tenemos que buscar la forma de conocer si existen esas características 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**.

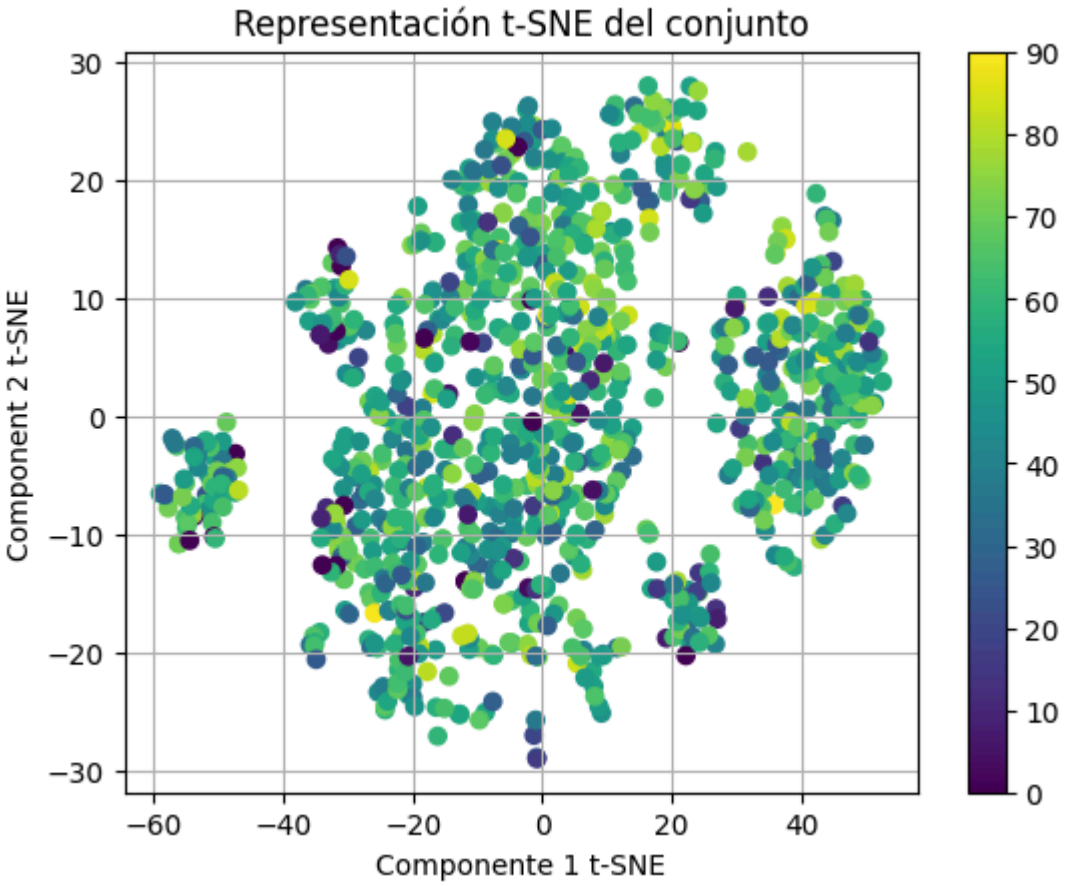
Luego **estandarizamos los datos** porque PCA trabaja con los desvíos estándar y la variabilidad que arrastra cada componente principal, entonces si no estandarizamos, cualquier variable con una varianza muy elevada nos puede dar una falsa lectura del análisis.



Luego de utilizar a **Popularity** como nuestra tercera dimensión, notamos que no hay una relación lineal que se puede deducir entre la popularidad de una canción y el resto de factores porque no se observan patrones de colores en el gráfico. Cada punto de color representando una popularidad alta o baja se encuentra en todos los lugares del rango de las **componentes principales**, los cuales además ninguno de ellos se lleva la mayor parte de la **variabilidad con un 19% y 14% de los datos**.

T-SNE

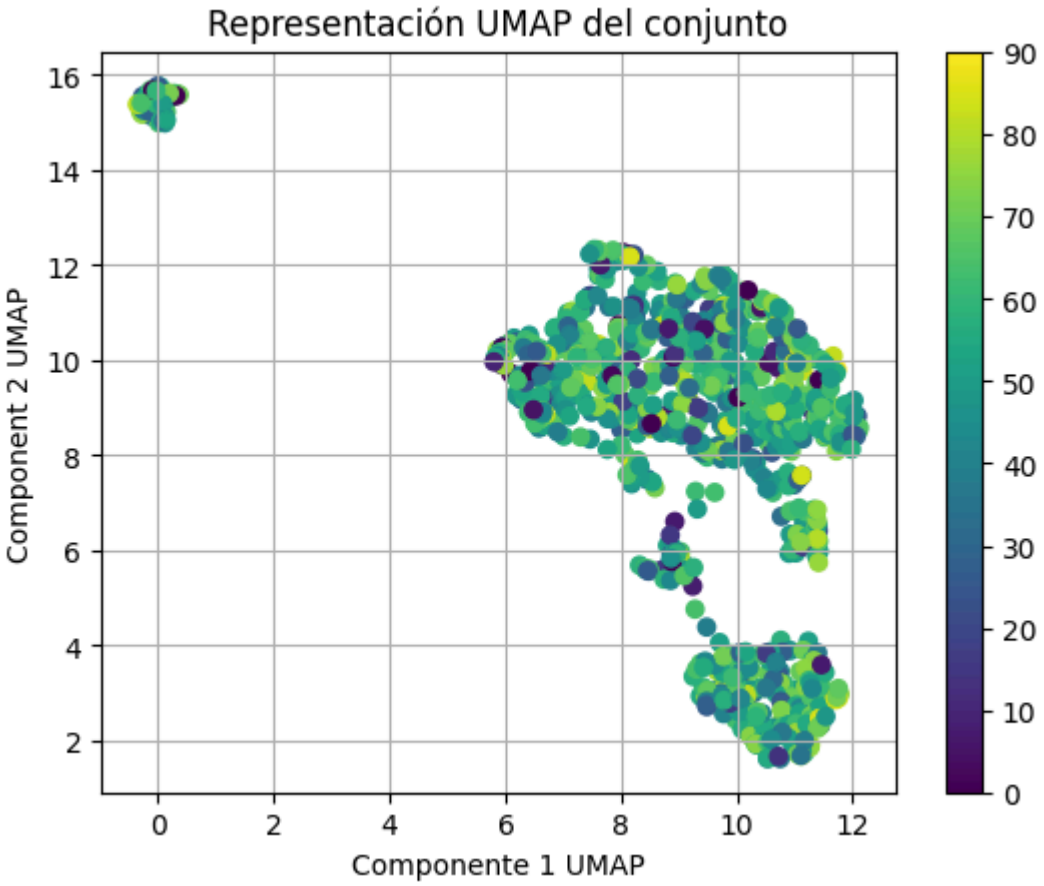
Ahora que comprobamos que no hay linealidad en el análisis de nuestra hipótesis, tenemos que recurrir a métodos como TSNE para la **reducción no lineal de la dimensionalidad** y el modelado de la variedad de los datos.



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.



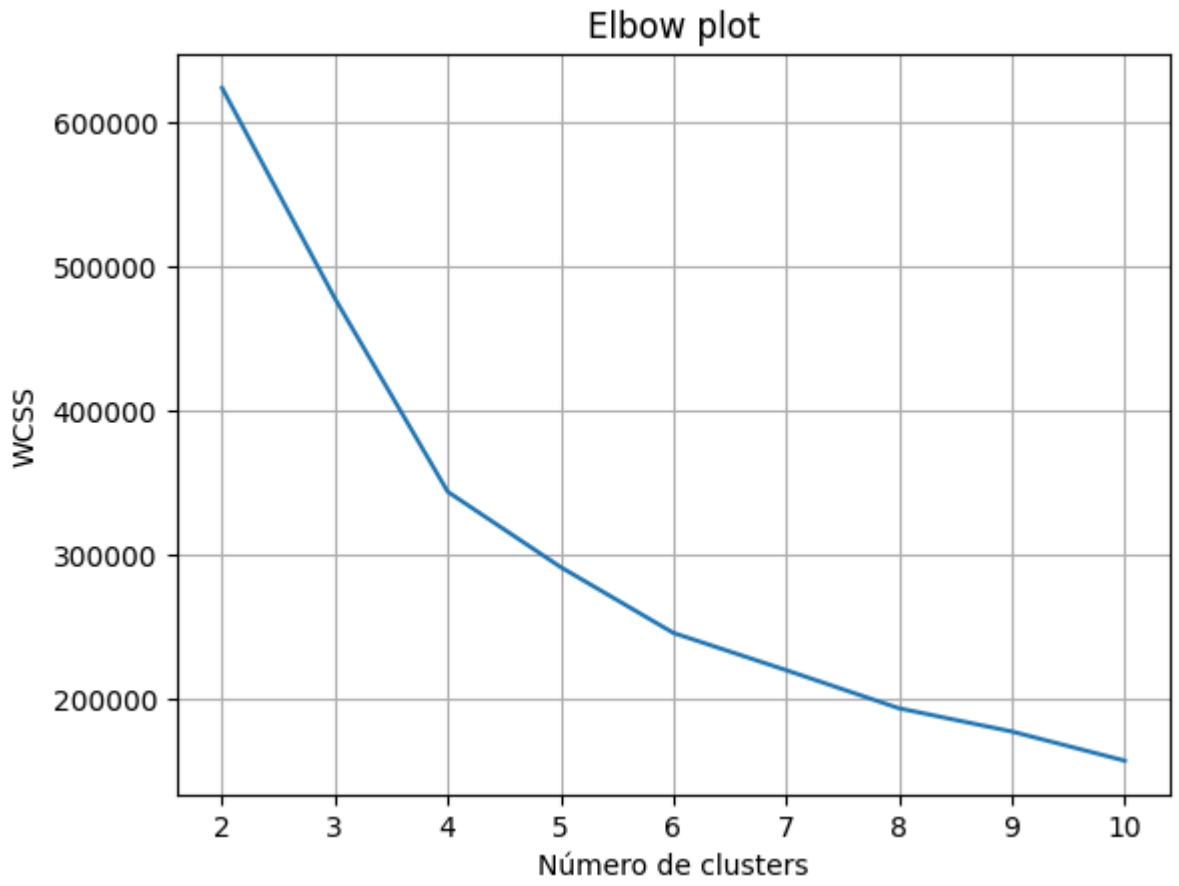
Notamos que se formaron algunos grupos pero como la dispersión de valores de Popularity (graduación de colores) sigue igual de vigente que en PCA, podemos afirmar que **no existe relación alguna entre un determinado grupo de factores para conformar la popularidad de una canción**.

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

Clustering

La gran mayoría de nuestras hipótesis fueron rechazadas cuando las evaluamos con los datos y todas las técnicas que conocemos. Entonces 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 aporten a algunas conclusiones sólidas y enriquezcan nuestro trabajo.

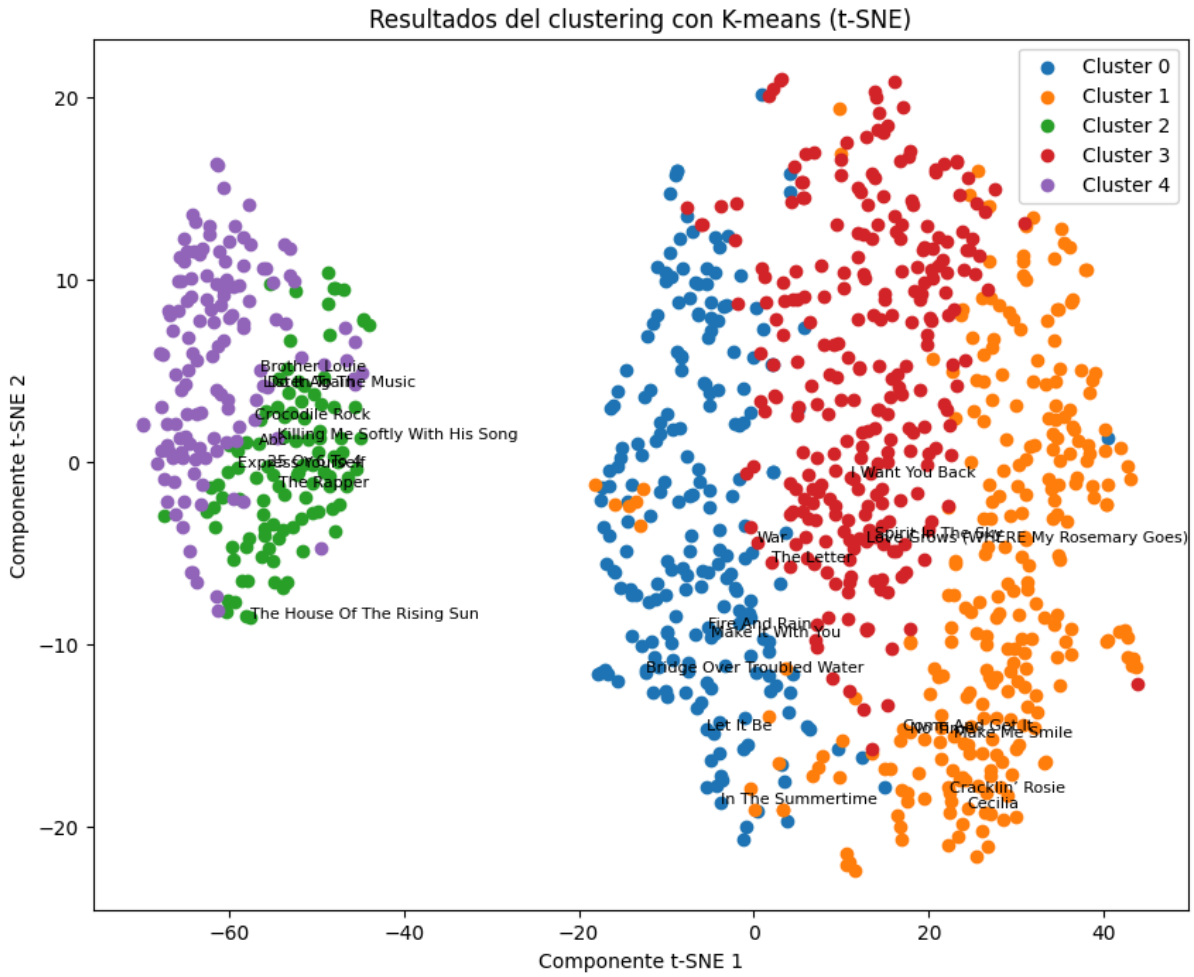
Utilizaremos **clustering basado en particionamiento** y para eso necesitamos el valor de K-Means necesario para tener una cantidad de agrupaciones adecuadas. Así que usaremos de ayuda el **Elbow plot** o gráfico del codo que realiza un análisis sobre la cantidad de clusters y la distancia entre las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



En nuestro caso **elegimos un K = 5** pero podríamos haber tomado un K = 4 igualmente.

Queda entrenar el algoritmo y generar los clusters.

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.

Cluster 0

	Time_Signature	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Popularity	Year
count	212.000000	212.000000	212.000000	212.000000	212.000000	212.0	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000
mean	3.867925	0.484302	0.357396	4.943396	-12.391193	1.0	0.045063	0.684667	0.045123	0.189147	0.415492	115.262292	49.537736	1974.122642
std	0.353063	0.150740	0.133700	3.185662	3.936084	0.0	0.050766	0.185351	0.166447	0.170215	0.207942	31.689960	19.971864	2.869030
min	3.000000	0.149000	0.005320	0.000000	-31.646000	1.0	0.024100	0.032300	0.000000	0.037200	0.034600	61.530000	0.000000	1970.000000
25%	4.000000	0.357500	0.281500	2.000000	-14.303000	1.0	0.029800	0.567500	0.000001	0.101000	0.242750	86.156750	40.000000	1972.000000
50%	4.000000	0.488000	0.357000	5.000000	-12.264000	1.0	0.033300	0.721500	0.000052	0.125000	0.387000	114.532000	53.000000	1974.000000
75%	4.000000	0.589500	0.440250	7.000000	-9.653250	1.0	0.038825	0.835250	0.001913	0.185750	0.559250	138.424750	63.000000	1977.000000
max	5.000000	0.965000	0.816000	11.000000	-3.862000	1.0	0.467000	0.996000	0.968000	0.985000	0.973000	205.747000	89.000000	1979.000000

Cluster 1

	Time_Signature	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Popularity	Year
count	268.000000	268.000000	268.000000	268.000000	268.000000	268.0	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000
mean	3.973881	0.625526	0.678623	1.574627	-8.498332	1.0	0.064464	0.205015	0.039885	0.168823	0.709004	121.932179	54.742537	1974.902985
std	0.219099	0.141765	0.158704	1.632428	2.976240	0.0	0.068432	0.187209	0.154738	0.146159	0.221480	24.291751	17.524860	2.805477
min	1.000000	0.212000	0.267000	0.000000	-17.340000	1.0	0.024800	0.000022	0.000000	0.016600	0.055800	63.059000	0.000000	1970.000000
25%	4.000000	0.531500	0.560750	0.000000	-10.619250	1.0	0.032525	0.046850	0.000000	0.083800	0.560750	106.923500	47.000000	1973.000000
50%	4.000000	0.632500	0.680500	1.000000	-8.140500	1.0	0.041550	0.154500	0.000014	0.116000	0.766500	120.608000	57.000000	1975.000000
75%	4.000000	0.735000	0.803000	2.000000	-6.325250	1.0	0.062450	0.336250	0.001310	0.194250	0.894250	132.213500	67.000000	1977.000000
max	4.000000	0.985000	0.995000	5.000000	-2.340000	1.0	0.576000	0.729000	0.920000	0.962000	0.989000	211.270000	90.000000	1979.000000

Cluster 3

	Time_Signature	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Popularity	Year
count	265.000000	265.000000	265.000000	265.000000	265.000000	265.0	265.000000	265.000000	265.000000	265.000000	265.000000	265.000000	265.000000	265.000000
mean	3.962264	0.613589	0.636396	8.120755	-9.289464	1.0	0.059184	0.227142	0.036527	0.185269	0.696936	118.226868	53.554717	1974.339623
std	0.190917	0.139354	0.154333	1.738747	3.090653	0.0	0.058509	0.187452	0.137556	0.168030	0.213665	25.160585	18.070093	2.742368
min	3.000000	0.221000	0.265000	5.000000	-18.168000	1.0	0.023200	0.000668	0.000000	0.015000	0.144000	72.269000	0.000000	1970.000000
25%	4.000000	0.529000	0.531000	7.000000	-11.464000	1.0	0.030800	0.063100	0.000000	0.081100	0.543000	100.178000	43.000000	1972.000000
50%	4.000000	0.632000	0.634000	8.000000	-9.138000	1.0	0.039200	0.180000	0.000023	0.117000	0.749000	114.751000	57.000000	1974.000000
75%	4.000000	0.703000	0.743000	9.000000	-6.916000	1.0	0.061200	0.348000	0.001920	0.214000	0.878000	133.000000	67.000000	1976.000000
max	4.000000	0.912000	0.989000	11.000000	-3.144000	1.0	0.452000	0.810000	0.959000	0.900000	0.981000	203.812000	85.000000	1979.000000

Cluster 2

	Time_Signature	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Popularity	Year
count	104.000000	104.000000	104.000000	104.000000	104.000000	104.0	104.000000	104.000000	104.000000	104.000000	104.000000	104.000000	104.000000	104.000000
mean	3.913462	0.556963	0.570663	7.500000	-10.275115	0.0	0.077572	0.374169	0.101904	0.160917	0.572710	119.191529	50.663462	1972.182692
std	0.464478	0.178542	0.208107	2.946118	4.366042	0.0	0.099518	0.292657	0.247497	0.137480	0.262844	30.047305	18.025662	1.783225
min	1.000000	0.094200	0.116000	0.000000	-27.090000	0.0	0.025500	0.000109	0.000000	0.030900	0.000010	68.482000	8.000000	1970.000000
25%	4.000000	0.442000	0.417250	5.000000	-12.553250	0.0	0.034100	0.090650	0.000002	0.086525	0.399250	94.989000	39.000000	1971.000000
50%	4.000000	0.566000	0.551500	8.000000	-10.306500	0.0	0.046800	0.312000	0.000230	0.119000	0.580000	115.972000	52.000000	1972.000000
75%	4.000000	0.682000	0.734500	10.000000	-6.973000	0.0	0.061225	0.591000	0.014950	0.177250	0.817250	139.855250	64.000000	1973.000000
max	5.000000	0.911000	0.995000	11.000000	-2.515000	0.0	0.737000	0.973000	0.970000	0.957000	0.969000	207.266000	90.000000	1978.000000

Cluster 4

	Time_Signature	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Popularity	Year
count	131.000000	131.000000	131.000000	131.000000	131.000000	131.0	131.000000	131.000000	131.000000	131.000000	131.000000	131.000000	131.000000	131.000000
mean	3.969466	0.649626	0.638366	5.236641	-9.402977	0.0	0.062167	0.206871	0.051467	0.164647	0.659382	119.510260	57.938931	1976.923664
std	0.172713	0.135770	0.184859	3.551451	3.316989	0.0	0.055546	0.219318	0.151473	0.137326	0.243207	24.793623	16.071837	1.752441
min	3.000000	0.271000	0.129000	0.000000	-20.149000	0.0	0.024800	0.000215	0.000000	0.018800	0.085100	53.986000	7.000000	1972.000000
25%	4.000000	0.587500	0.503500	2.000000	-11.819500	0.0	0.033450	0.030250	0.000002	0.073950	0.460000	106.422500	48.000000	1976.000000
50%	4.000000	0.669000	0.646000	5.000000	-9.223000	0.0	0.043300	0.111000	0.000220	0.111000	0.708000	116.735000	59.000000	1977.000000
75%	4.000000	0.739000	0.795000	8.000000	-6.785500	0.0	0.062600	0.313500	0.011850	0.206000	0.887500	129.279000	71.000000	1978.000000
max	4.000000	0.889000	0.968000	11.000000	-3.556000	0.0	0.348000	0.947000	0.794000	0.701000	0.985000	202.297000	86.000000	1979.000000

Conclusión de clustering

Luego de analizar detalladamente las descripciones estadísticas de cada cluster, podemos concluir que la única diferencia entre los grupos de los **clusters 0, 1, 3** y los **clusters 2, 4** 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.

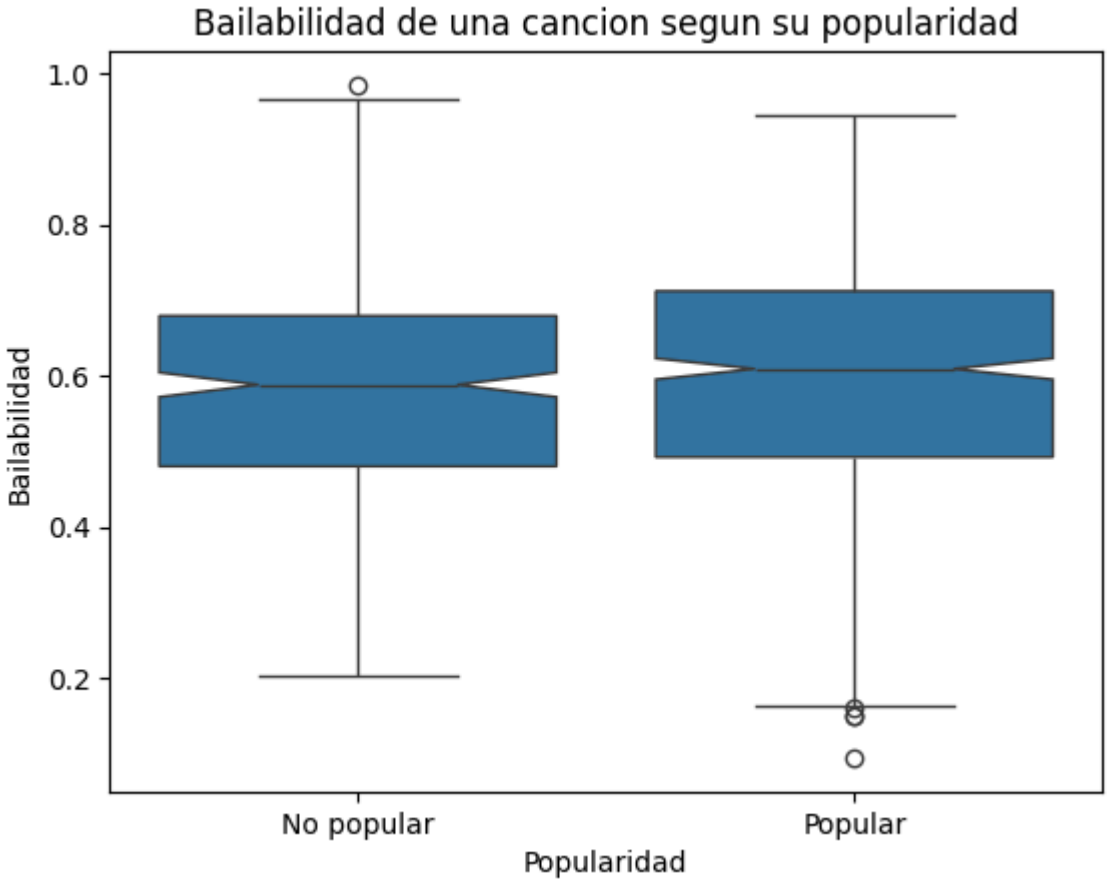
En el resto de características, todos nuestros clusters comparten el mismo rango práctico y hasta en algunos casos, tienen el mismo valor.

Test Paramétrico

Nos queda validar la última hipótesis restante de todas las planteadas: la sexta. Donde planteamos que las canciones más adecuadas para ser bailadas tienen una tendencia a tener un buen nivel de popularidad.

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 iguales. 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.

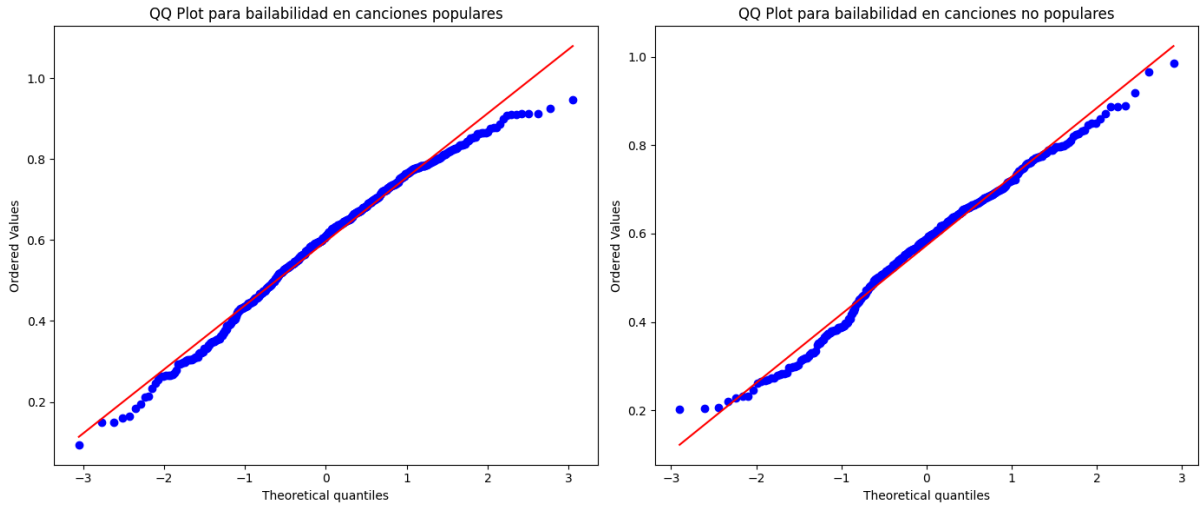
Validamos normalidad

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** importado de la librería **Scipy.stats** 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.

Test No Paramétrico

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 poquito 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
```

Obtuvimos un p-valor menor a 0.05, rechazando nuestro supuesto inicial, nos revela que **las canciones NO POPULARES tienen una mayor bailabilidad que las canciones POPULARES**.

Esto nos contradice fuertemente nuestra sexta hipótesis probando que estábamos muy equivocados.

Regresión lineal

Siguiendo con la línea de nuestra octava hipótesis, si bien ya comprobamos que no tenemos ninguna relación lineal de la popularidad de una canción con el resto de variables cuando aplicamos la técnica de PCA, 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.

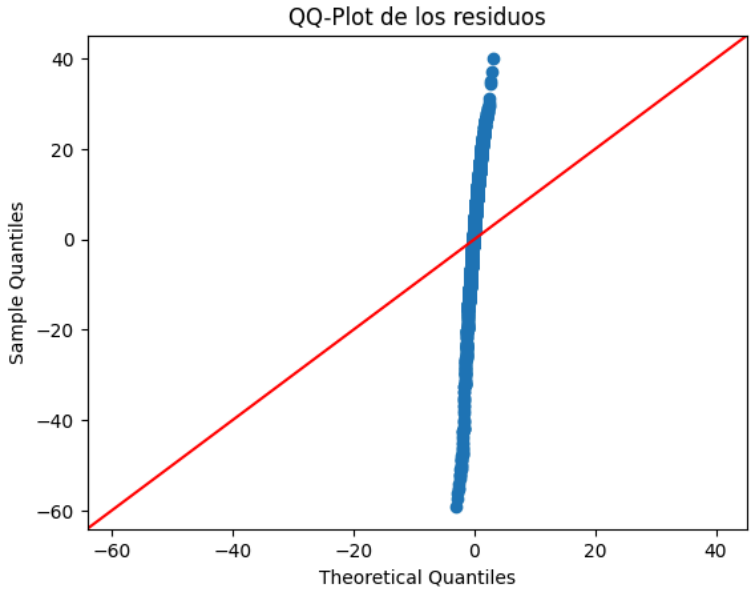
Primero utilizamos el mismo dataframe estandarizado que usamos en PCA pero recuperando la información del nombre de las columnas que en su momento no nos servían pero ahora si.

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
==============================================================================
```

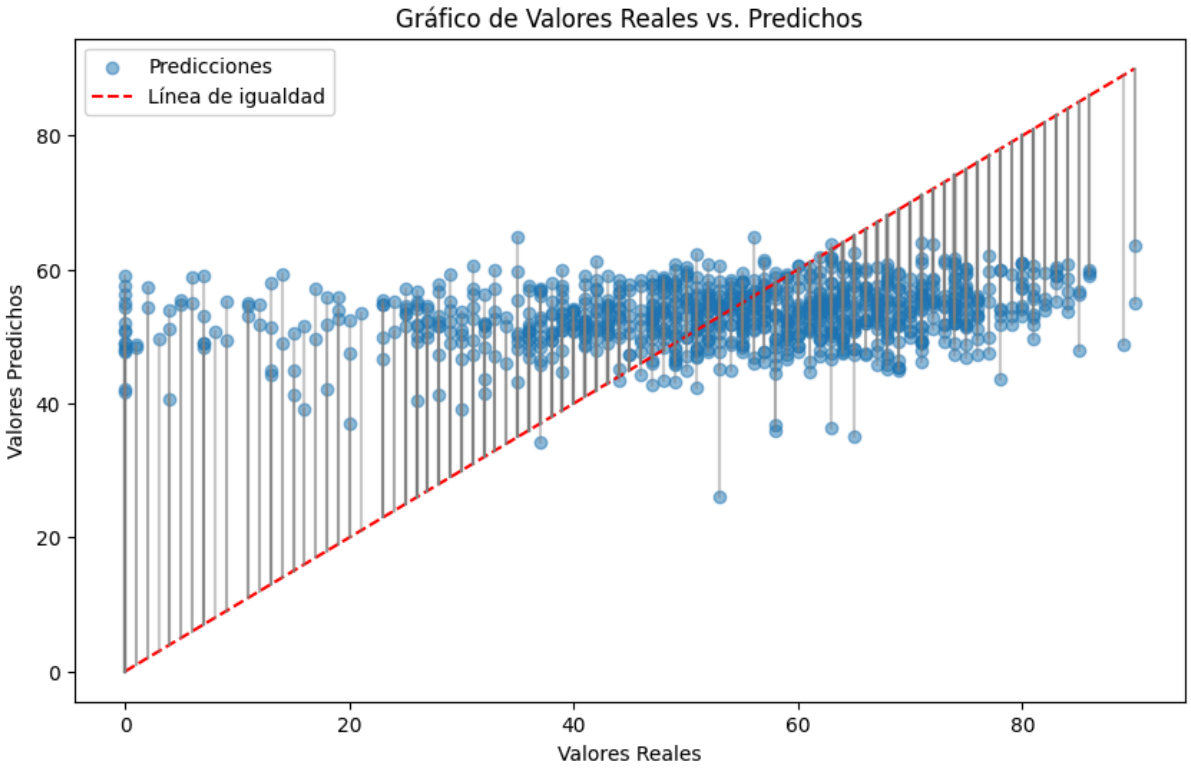
Lo primero que notamos es que el **coeficiente de determinación** R-squared es **ridículamente bajo**, significando que nuestro modelo explica solamente casi el 6% más de la varianza en los datos que un modelo trivial basado en tomar la media de todas las muestras. Cuando en un caso ideal ese número 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 predecidos. Y vemos que hay bastante distancia entre los

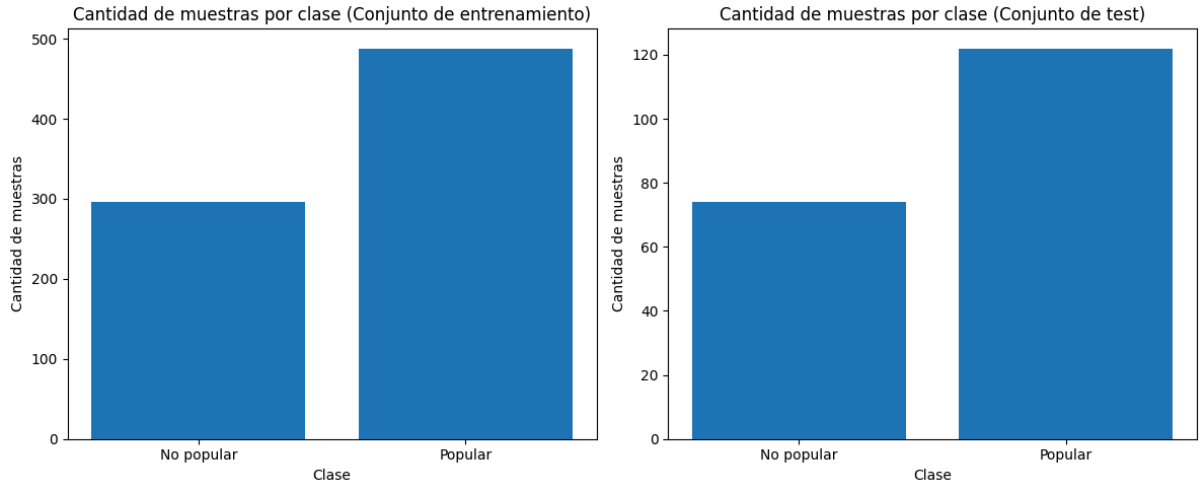
puntos y la recta, lo que indica un error muy elevado al estar los puntos tan distanciados y no seguir ningún patrón parecido a la recta.
Como ya comprobamos que **no tenemos relaciones lineales** en nuestros datos a lo largo de este trabajo, no vale la pena intentar modelar otro modelo de regresión lineal, así que **vamos a tener que aplicar alguna técnica de regresión no lineal**.

Regresión no lineal

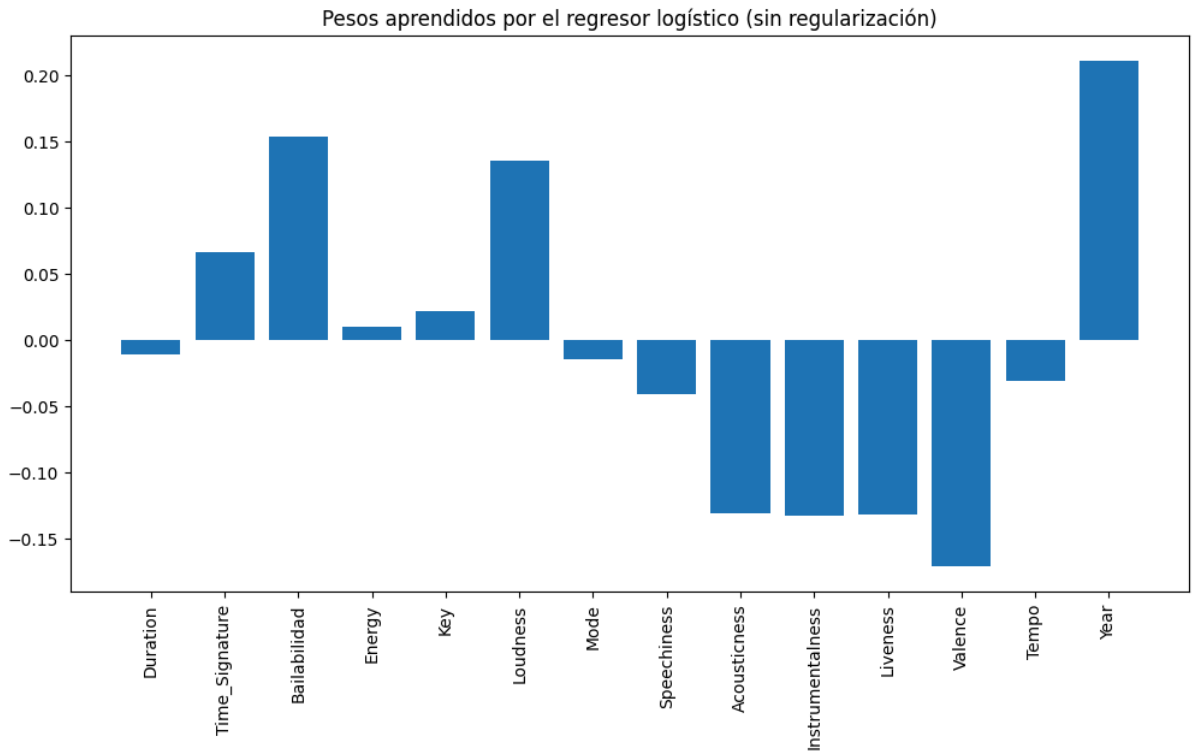
Regresión Logística

Al no obtener buenos resultados utilizando regresión lineal, recurrimos a métodos como la regresión logística para ver si podemos predecir el valor de la variable categórica **Popular** (la que transformamos al principio del Test Paramétrico) de una muestra en base al resto de característica 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.

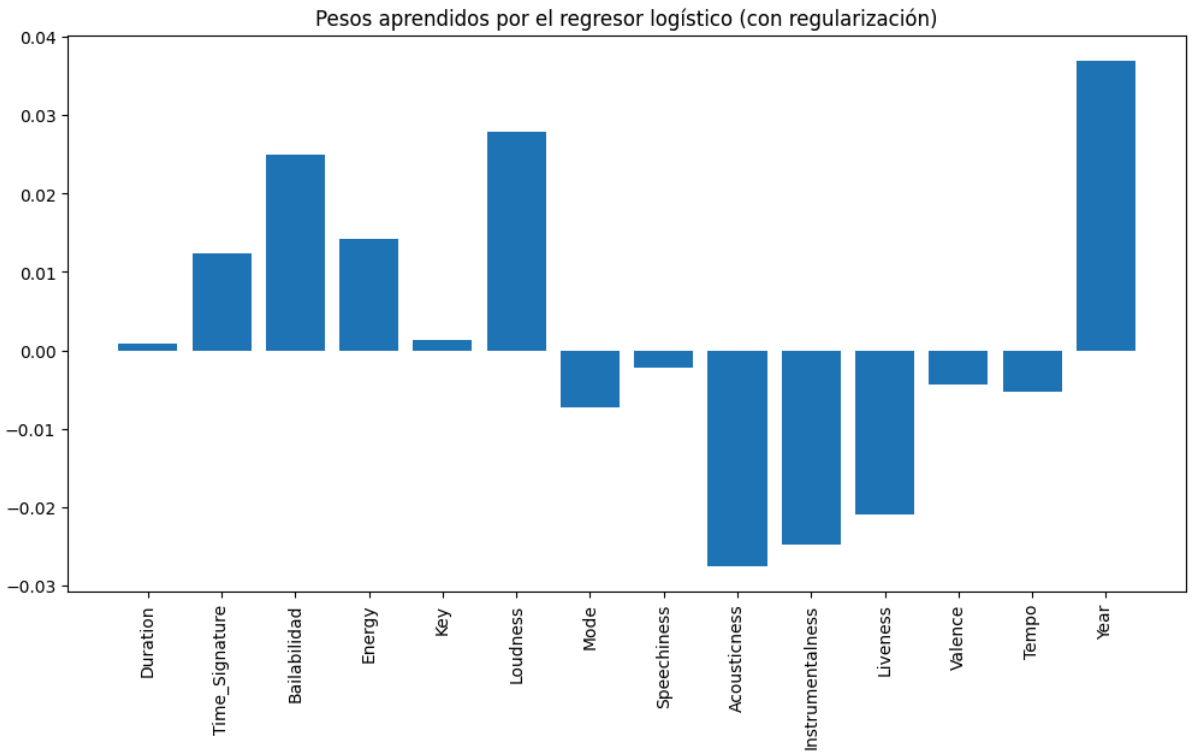


En este primer modelo base, vemos que se le da un peso mayoritario a las características del año de una canción, la bailabilidad y su positividad musical (Valence) por sobre otras.

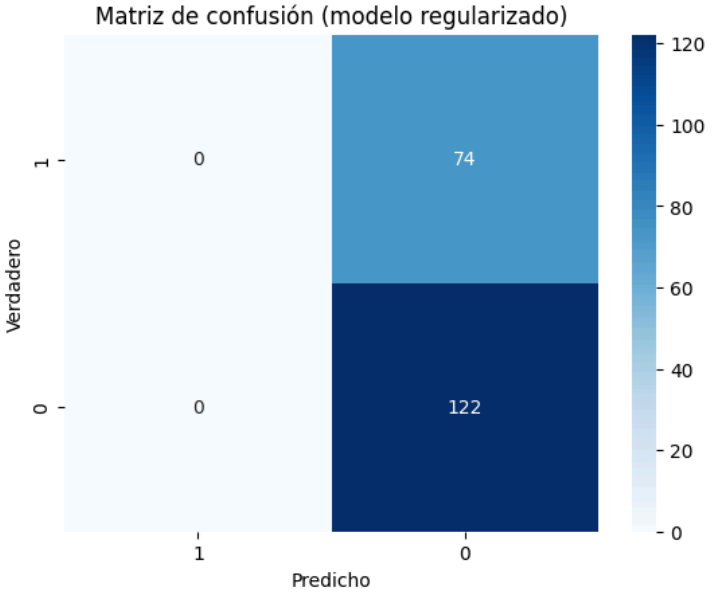
Procedemos a mejorar la validación del modelo usando el método de **validación cruzada** dividiendo el conjunto de entrenamiento en 5 partes iguales y re-entrenando el modelo 5 veces usando 4 partes y 1 dedicada a evaluación. Obteniendo así una exactitud de predicción del **casi 61%** con un desvío de 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 no populares** únicamente, cuando ya comprobamos que en nuestro conjunto de entrenamiento tenemos un total de 488 canciones populares.

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