Universidad Nacional del Centro de la
Provincia de Buenos Aires

FACULTAD DE CIENCIAS EXACTAS

Ingeniería de Sistemas

Trabajo Práctico Especial - Análisis Exploratorio

Fundamentos de la Ciencia de Datos

GRUPO 7

Agustín Buralli
Sofía Anabel Todesco
Tomás Antúnez Monges

ÍNDICE

Introducción

Se nos pidió hacer un análisis sobre un conjunto de datos acerca de covers de canciones de la década de los 90's para, de esta manera, obtener conclusiones que sirvan como información útil para indagar sobre qué factores juegan en que una canción sea popular, o acústica, etc; y cómo estas variables interactúan entre sí.

Contaremos qué herramientas utilizamos para el desarrollo del trabajo, estudiaremos el significado de cada variable contenida en el dataset. Realizaremos un análisis exploratorio de los datos, analizando las distribuciones de cada variable y realizando limpieza y correcciones en caso de ser necesarias.

Luego de conocer bien el dataset, comenzamos a plantearnos hipótesis que podrían resultar de interés, para luego analizarlas a fondo y concluir si son ciertas o no.

También buscaremos relaciones sin una hipótesis en mente, quizás los datos nos puedan mostrar algo que no intuimos naturalmente.

Finalmente desarrollaremos las conclusiones del trabajo realizado.

Materiales

Utilizamos PyCharm, un IDE para trabajar con el lenguaje de programación Python. Dentro de PyCharm trabajamos con una integración para utilizar Jupyter Notebook, que permite combinar código en Python con texto en Markdown. Para los análisis utilizamos distintas librerías como Pandas, Scikit-learn, Seaborn, Matplotlib, Numpy y Scipy, entre otras. Además, compartimos las actualizaciones del trabajo mediante un repositorio remoto de GitHub (plataforma basada en la web que utiliza Git para alojar y gestionar proyectos de software).

Centrándonos en los datos, recibimos un dataset con las siguientes características: 980 observaciones(canciones) de 536 artistas diferentes, con 17 características que contienen la siguiente información:

- ★ **Track:** nombre de la canción.
- ★ **Artist :** nombre del artista o intérprete.
- ★ **Duration:** duración en minutos de la canción.
- ★ **Time Signature:** número de pulsaciones por compás.
- ★ **Danceability:** medida de qué tan bailable es la canción (entre 0 y 1).
- ★ **Energy:** medida de que tan enérgica es la canción (entre 0 y 1).
- ★ **Key:** clave o tonalidad de la canción, número entero.
- ★ **Loudness**: volumen de la canción, en decibelios.
- ★ **Mode:** tono mayor o menor (0 o 1, respectivamente).
- ★ **Speechiness:** medida de presencia de palabras habladas en las canciones, valores altos indican una alta presencia de estas palabras.
- ★ **Accoustiness:** mide qué tan acústica es la pista.
- ★ **Instrumentales:** mide la presencia de instrumentos en las canciones, valores más altos indican una canción con mayor parte instrumental.
- ★ **Liveness** : Mide la presencia de público en el track (gritos, silbidos, aplausos) (de 0 a 1).
- ★ **Valence**: medida de la positividad de la canción, valores más altos indican presencia de melodías más alegres.
- ★ **Tempo**: velocidad de la canción, medida en beats por minuto (bpm).
- ★ **Popularity:** puntuación de la canción. Mide la popularidad de la misma.
- ★ **Year**: año de lanzamiento del cover. (1991-1998)

Métodos / Resultados

Análisis del conjunto de datos

Lo primero que hicimos luego de analizar bien de qué se trata el dataset fue comenzar con la limpieza de los datos.

Aparecieron un par de cosas extrañas como la repetición de ceros en Instrumentalness y algunos valores repetidos en Speechiness.

Correcciones

Uno de los problemas que tuvimos es que el formato de duración no era manejable para comparaciones. Cada fila poseía un valor en string minutos:segundos. Optamos por crear una nueva columna con los valores correspondientes en segundos, solo bastó con separar el string en dos según él ':' y multiplicar los minutos por 60 y sumarle la parte en segundos.

Análisis de distribuciones

♦ **Duration** → Una vez que la corregimos, nos encontramos con que la mayoría de las canciones duran entre 150 y 300 segundos (2:30 y 5:00 minutos). Había algunos outliers que superaban los 500 segundos pero los eliminamos para que no nos afecten la distribución de la variable.

♦ **Time_Signature** → Observamos que la gran mayoría de las canciones tienen 4 pulsaciones por compás, es un compás muy común. Aquí dejamos la siguiente página[1] que contiene información de otros compases también.

♦ **Danceability** → Realizamos un boxplot inicial donde visualizamos algunos outliers. Luego generamos un histograma donde vimos que la distribución es semejante a una normal.

♦ **Energy** → Podemos observar mediante un histograma de la variable una curva bastante sesgada hacia la derecha, lo que indica una gran presencia de canciones energéticas.

♦ **Key** → Es una variable cuantitativa discreta que presenta valores de 0 a 11, cada valor representa la clave de la canción a la que corresponde, las claves respectivamente son: Do, Do#, Re, Re#, Mi, Fa, Fa#, Sol, Sol#, La, La# y Si.

La clave es como la nota en la que se centra la canción, una nota de descanso o de retorno. Siempre se vuelve a esa nota y si no lo hiciera sonaría incompleto [(4)(5)].

Realizamos un histograma y descubrimos que las Key con más canciones son Do# y Sol, y la que menos tiene Re#.

Después de toda la investigación que realizamos, descubrimos que no es muy bueno el análisis de Key sin analizarlo junto al Mode o Tono, ya que una clave suena distinta según el tono que sea. Cuando el tono es menor suena más triste, y cuando es mayor suena más alegre.

A partir de esto también realizamos un histograma diferenciado por Mode, y vimos que de las canciones de tono mayor había más con clave Sol, Do# y Do, que casualmente coincide con el histograma anterior. Por otro lado, de las canciones en tono menor había más con clave Si, Fa y La#. Así vemos que las canciones más tristes por así decirlo se encuentran con esas claves.

♦ **Loudness** → Como es una variable cualitativa continua, hicimos un histograma. Nos dió un sesgo bastante pronunciado a izquierda, indicando que la mayoría de las canciones tienen un sonido de entre -20 a 0 decibelios, exceptuando algunos casos particulares.

♦ **Instrumentalness** → Nos pareció que se debe discretizar los valores de instrumentalness ya que hay una excesiva cantidad de valores nulos en la variable así que la separamos en 3 grupos, por un lado los valores igual a 0, otro grupo para los inferiores o iguales a 0.5 y otro para los mayores a 0.5 y menores o iguales a 1. Existen muy pocas canciones con valores de instrumentalness alto.

♦ **Liveness** →. Vemos una distribución muy sesgada hacia la derecha, lo que quiere decir que hay muchas canciones con poca presencia de la audiencia. Hay un par de tuplas que podrían ser posibles outliers arriba de 0.9 pero realmente no nos parece significante.

♦ **Valence** → La curva se ve bastante normal, hay un pico justo en el centro de la distribución lo que indica una gran cantidad de canciones con una positividad media.

♦ **Year**→ Bastante equilibrado entre los diferentes años, siendo 1991 el año en que se lanzó la mayor cantidad de covers.

♦ **Tempo** → En el histograma y, según información que encontramos en esta fuente[(2)], la mayoría de las canciones tienen un ritmo medio tendiendo a rápido.

♦ **Popularity** → Mediante un histograma, observamos una curva un poco sesgada a izquierda, indicando valores medianamente altos de popularidad de las canciones. También se puede apreciar una concentración de canciones con popularidad entre 0 y 5. No conocemos la medida en la que está basada esta variable(me gustas en apps de streaming, vistas mensuales, etc).

♦ **Speechiness:** Es una variable cuantitativa continua. Observamos que el 75% de los datos tiene valores debajo de 0.1. Esto significa que la mayoría de las canciones no tienen casi palabras habladas. Pensando en la medida de "palabras habladas" se nos viene a la mente que una canción con valores altos sería una canción de rap, por ejemplo. Tampoco hay ninguna que sea totalmente acapella, ya que el máximo llega a 0.529.

♦ **Mode** → Variable categórica, hay muchas más canciones escritas en escala mayor que en escala menor, según esta fuente[(3)] una canción compuesta por una escala mayor da una sensación de alegría, mientras que a las que se conforman por escalas menores se les atribuye sentimientos más tristes o depresivos (interesante para plantear una hipótesis al respecto).

♦ **Acousticness** → Vimos que el 75% de los datos se encuentran en el rango de 0.00 a 0.34 aprox. mientras que el otro 25% se encuentra entre 0.34 y 0.99, por lo tanto, se puede apreciar un sesgo a derecha debido a ese porcentaje de datos restante. Aunque la mayoría de los datos son pequeños, hay cierta dispersión y unos cuantos valores más altos que contribuyen a que la desviación estándar sea un poco más alta en comparación con la media.

Planteamiento de hipótesis

Luego del análisis exploratorio de los datos, hemos decidido plantearnos las siguientes hipótesis para analizarlas más a fondo:

Hipótesis 1: La presencia de audiencia en vivo está relacionado con la popularidad de la canción.

Hipótesis 2: La cantidad de canciones lanzadas por año es debido o se vio reflejada también en la popularidad.

Hipótesis 3: La medida de que tan bailable es una canción puede depender de si la pista es muy enérgica o poco enérgica

Hipótesis 4: La positividad de la canción dependerá del tono (Mode) en que se haya tocado, si en escala mayor o menor. La escala mayor se asocia a canciones con ritmos más positivos, mientras que la menor, a lo contrario.

Hipótesis 5: Cuanto más positiva es una canción, mayor tempo tiene.

Hipótesis 6: Las canciones que no son de 4 pulsaciones por compás son menos populares.

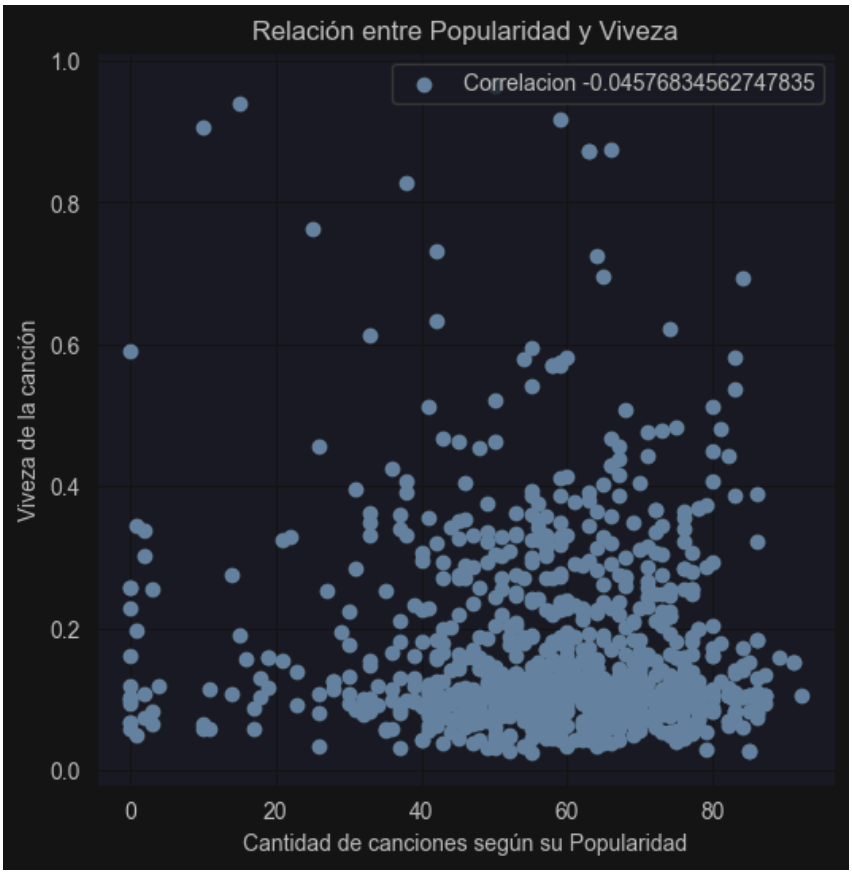
Hipótesis 7: La medida de que tan ruidosa es una canción depende de que tan enérgica sea.

Hipótesis 8: Energy, liveness y speechieness tienen una relación con el loudness de la canción.

Hipótesis 9: Las canciones más instrumentales son más acústicas que las cantadas.

Hipotesis 1: Liveness vs Popularity

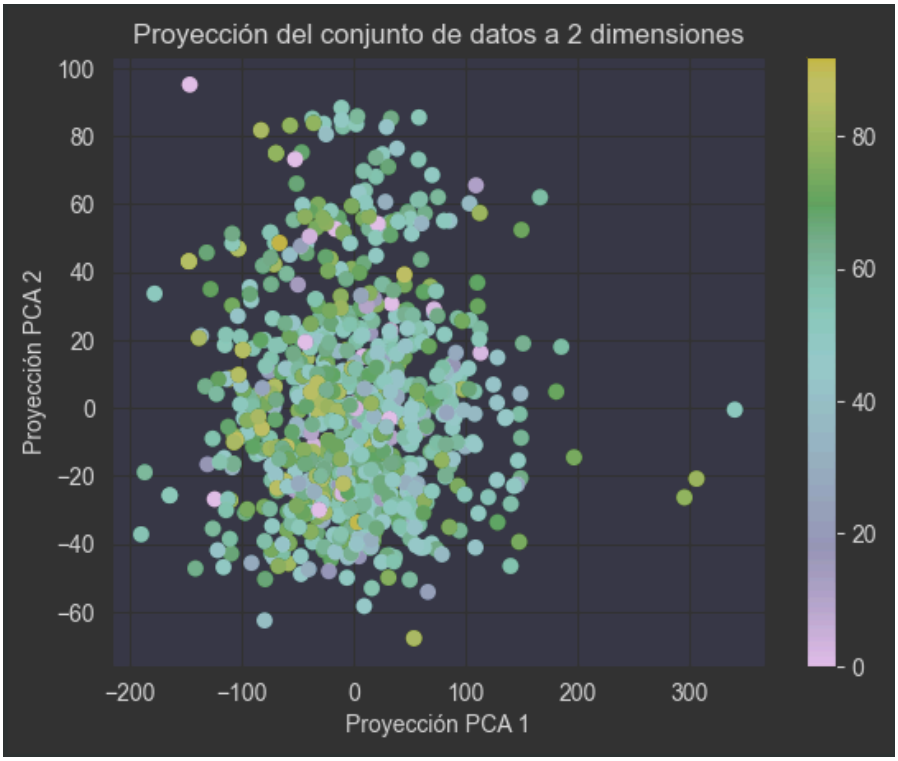
Lo que a primeras se nos ocurrió para analizar la relación entre Liveness y Popularity fue realizar un análisis bivariado. Un scatter plot acompañado por la correlación de pearson que resultó en una nube de puntos y un coeficiente muy cercano a cero. Este primer acercamiento descartó nuestra teoría de una relación lineal entre ambas variables, ya que lo que naturalmente pensamos fue: 'A mayor popularidad del cover mayor será su presencia de audiencia en vivo, no?'.

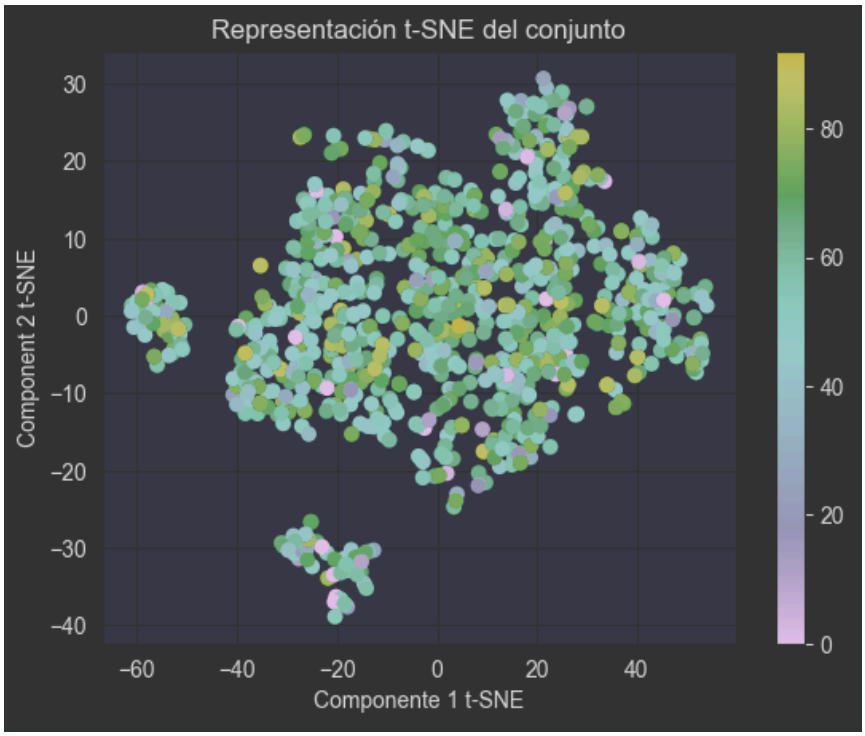


Con el objetivo de no rendirnos tan fácilmente supusimos que quizás la relación no fuera entre ellas dos sino que interviniera alguna/s variables más, así que nos lanzamos al análisis multivariado. Primero probamos con componentes principales, descartamos todas las variables que no fueran cuantitativas continuas y sacamos popularity (variable con la que luego pintaremos).

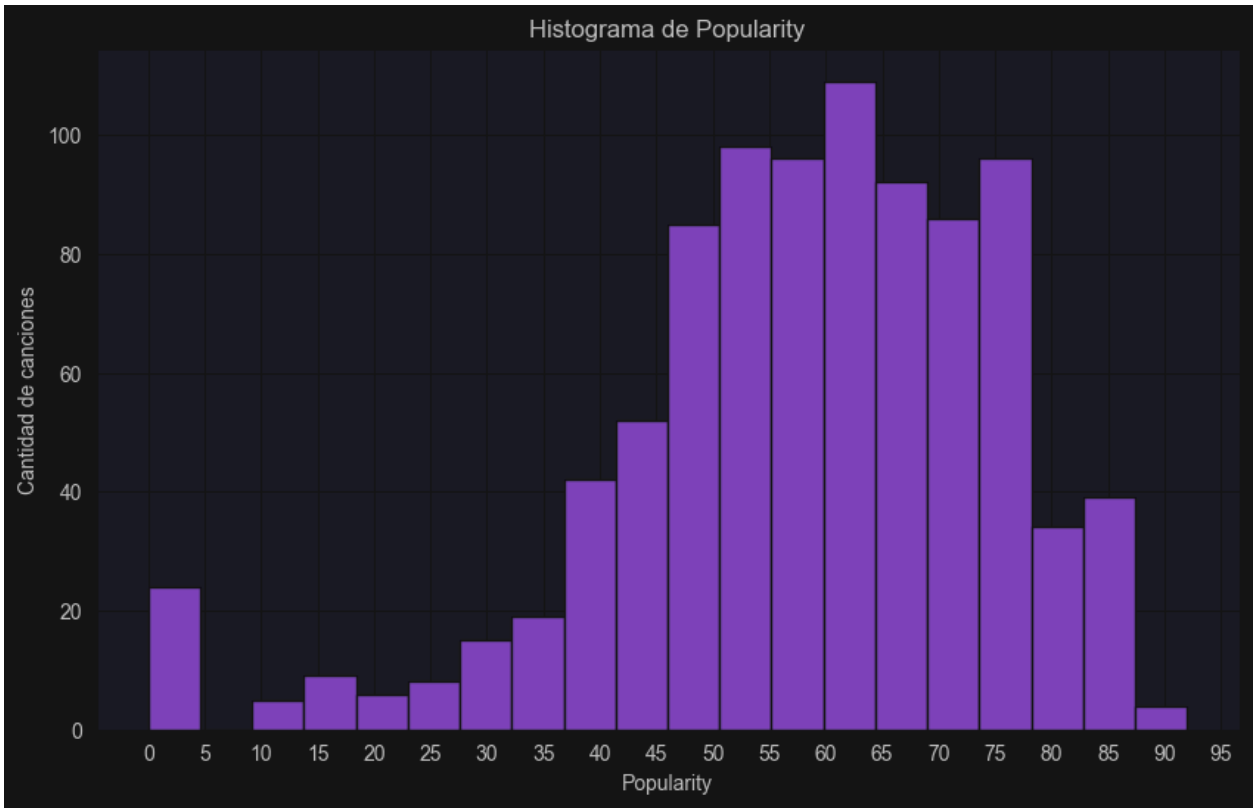
Nada, nube de puntos, ninguna agrupación evidente. Sorprendentemente nos dio una varianza explicada muy buena (un 98% de la varianza explicada) así que estábamos seguros que

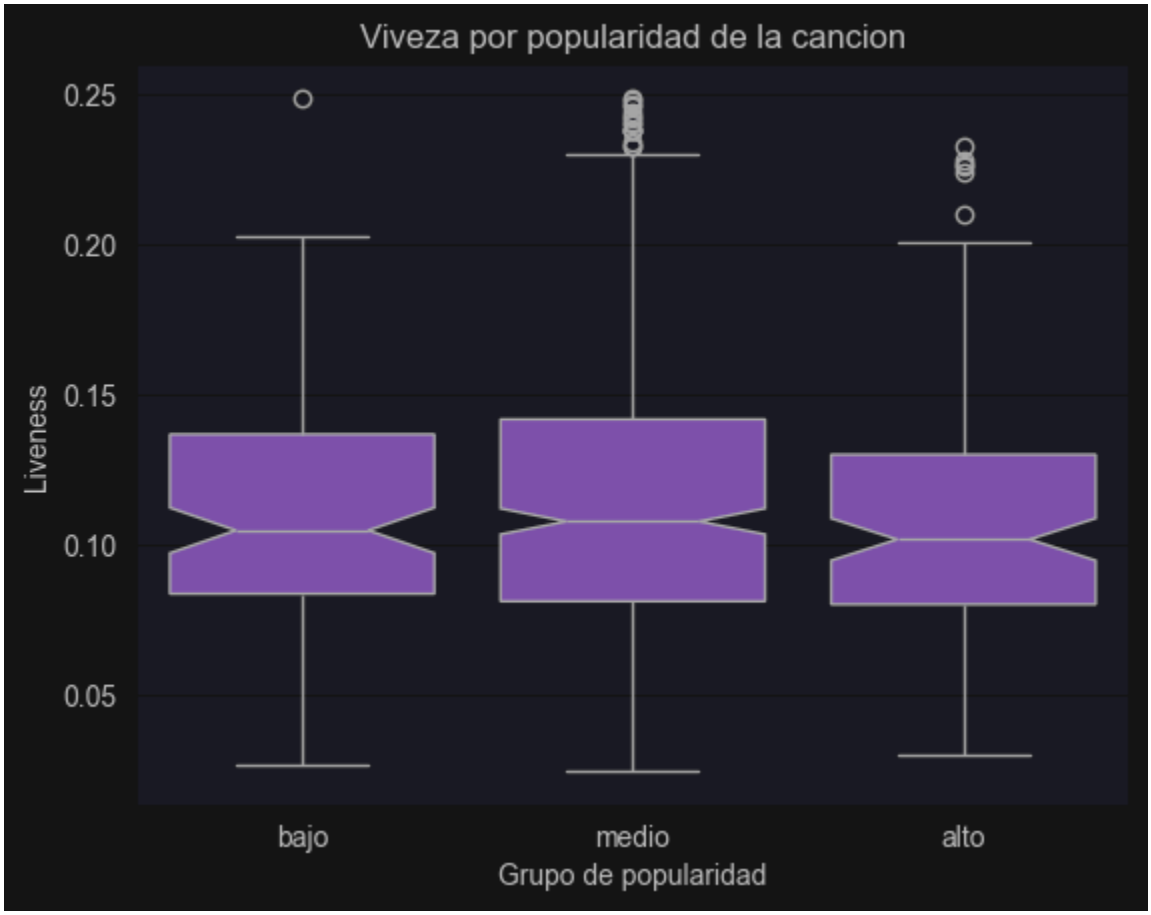
lo que estábamos viendo era correcto. Lo mismo pasó con t-SNE, aunque se veían 2 sub grupos muy pequeños en el gráfico, no compartían una relación entre ellos según la popularidad.





Todo esto nos llevó a preguntarnos, ¿realmente existe una diferencia significativa en el Liveness entre las canciones más populares y las menos populares?. Separamos las canciones en tres grandes grupos según la moda, en base al histograma obtenido. Graficamos los boxplot y *voilà*!, no parece haber a simple vista una diferencia entre los grupos, pero sí que había un montón de outliers!. Pero bueno, tampoco lo vamos a hacer a ojo el análisis, primero analizamos normalidad, en ambos grupos un p valor cercano a 0 así que se rechaza la hipótesis de que son normales, luego homocedasticidad y se acepta la hipótesis de que son homocedásticas (p valor aprox 0.22). Procedemos con Mann Whitney y no da un p valor aprox. de 0.62 por lo que no podemos rechazar la hipótesis nula y no podemos asegurar de que exista una diferencia significativa entre los valores de liveness de ambos grupos de canciones.

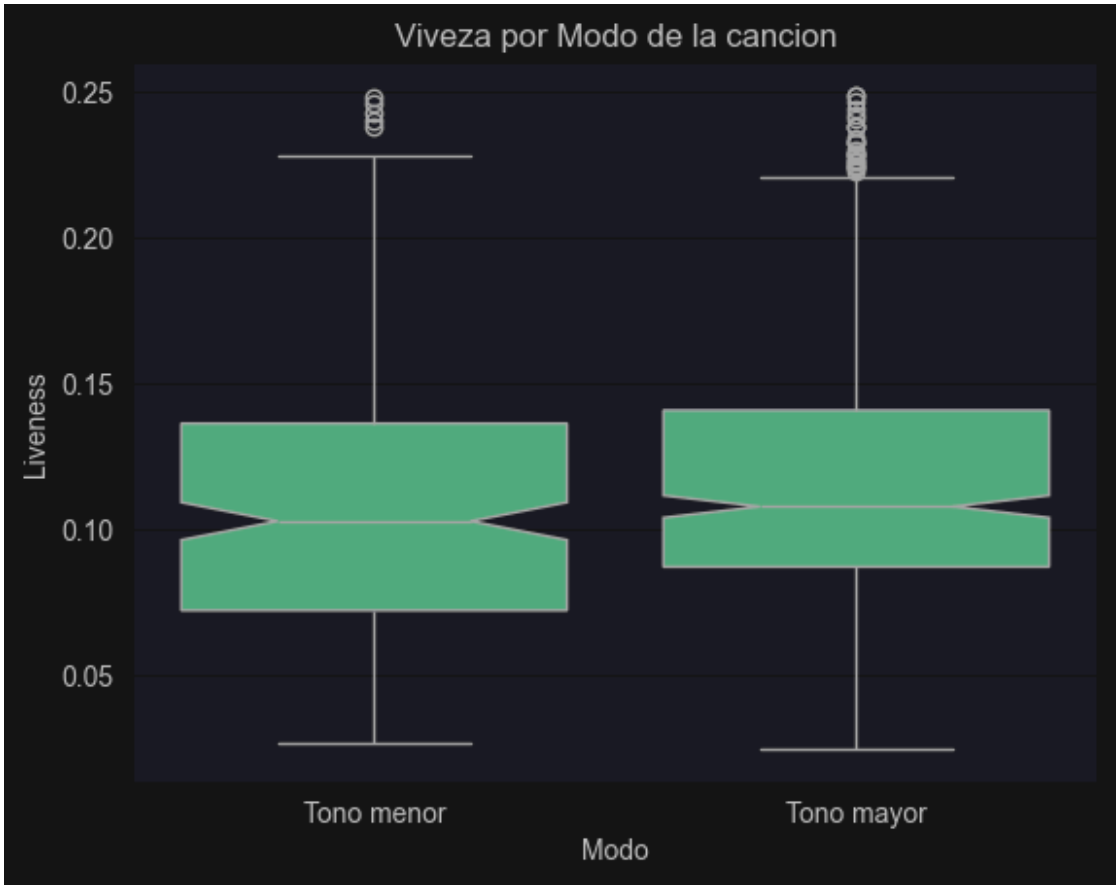




Entre lágrimas y llantos tuvimos una idea, que tal si existe una relación entre el ruido de la audiencia y el tono de la canción. Supongamos que canciones con tonos mayores tendrían más gritos y emoción de parte de la audiencia, mientras que las canciones con un tono menor o más tristes tendrían menos ruido de parte de la audiencia.

Para comenzar con el proceso, no fue necesario armar los grupos ya que la variable Mode (Tono mayor o menor) solo puede tomar dos valores posibles, 0 y 1, siendo 0 para el tono menor y 1 para el tono mayor. Graficamos box plots y a simple vista los “acogotamientos” parecen solaparse un poco, por lo que el pronóstico era malo. Nos enfocamos en los grupos extremos, el de más y el de menos popularidad, para poder ver con claridad que pasa entre las canciones más populares y las menos populares, ya que nos parece que el grupo medio no está aportando contraste. Usamos Shapiro para analizar normalidad y resultó que ninguno lo es, luego Levene para homocedasticidad en el grupo superior e inferior, y si lo son, Mann Whitney U será

entonces. El pronóstico negativo era incorrecto, nuestro p valor fue lo suficientemente chico(0.01) como para rechazar la hipótesis nula y poder asegurar que el valor de liveness entre ambos grupos es significativamente diferente. Y respaldando nos de los boxplot podemos ver que las canciones de tono mayor tienen mayor presencia de público, que era lo que esperábamos.

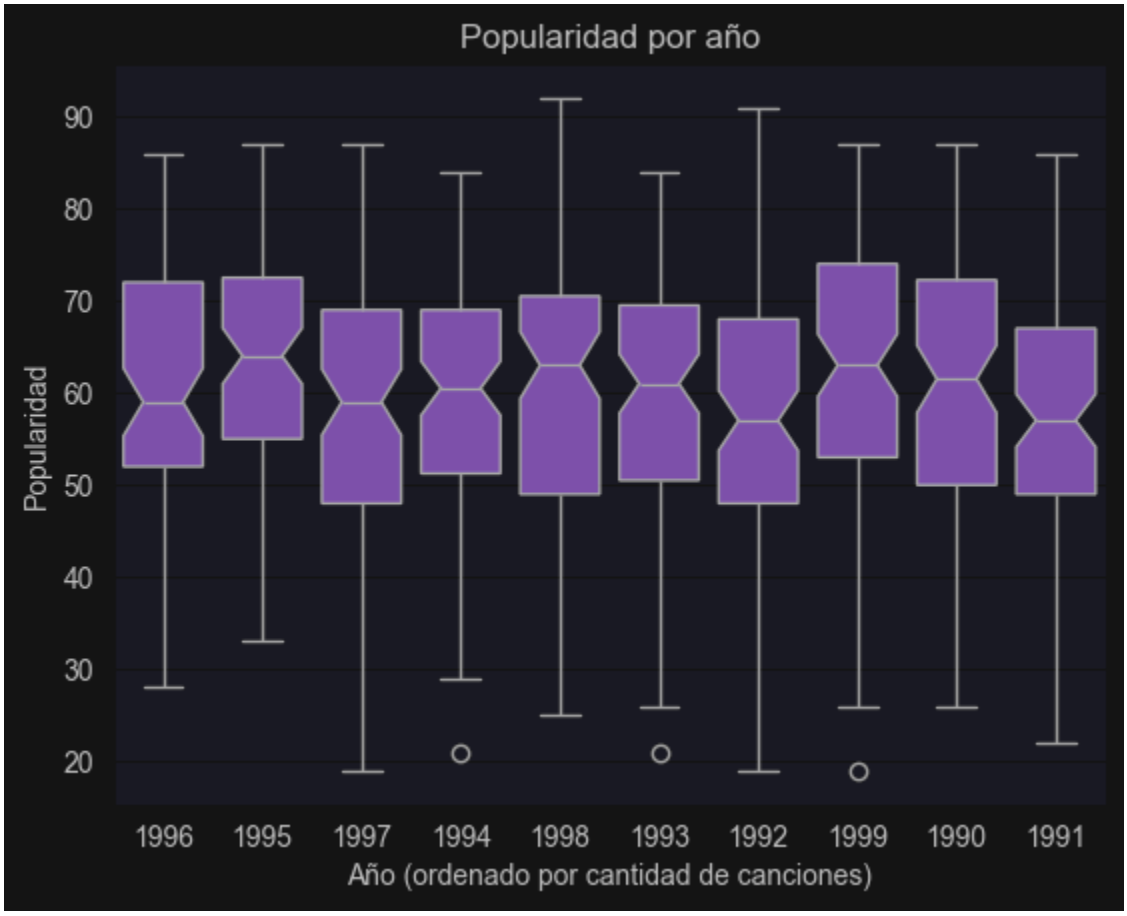


Hipotesis 2: Year vs Popularity

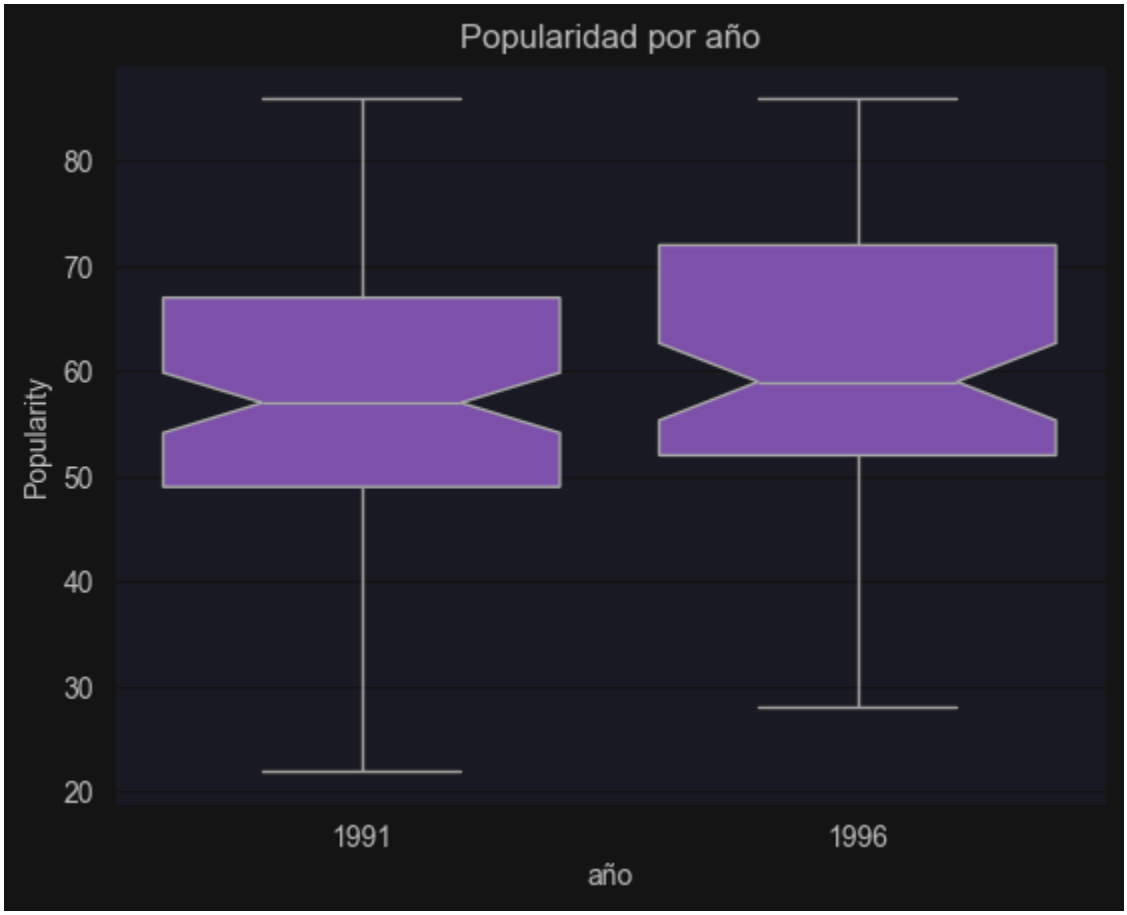
Una de las hipótesis que planteamos es que a cuanto más canciones fueran lanzadas en un año menos populares serían las canciones del mismo ¿porque habrian más cosas para escuchar?. Bueno, es una hipótesis!, veamos qué dice la estadística.

Lo que se nos ocurrió para verlo gráficamente son varios boxplot correspondientes a cada variedad de año en el dataset indicando el valor de popularidad y, muy importante , ordenados según la cantidad de canciones en sentido ascendente, a la izquierda los años con menos canciones creciendo la cantidad hacia la derecha. Si nuestra hipótesis tendría cierto grado de certeza tendríamos que ver que los boxplot de izquierda deberían estar más altos que los de la

derecha “a menor cantidad de covers lanzados mayor sería la popularidad en promedio de las canciones ese mismo año”. Obvio que no podía ser tan idealista todo, no había un patrón evidente en el gráfico, no parecen seguir ninguna tendencia.



Como entre los diferentes años no se evidencia ningún comportamiento en específico respecto de la popularidad y el número de canciones lanzadas, nos propusimos comparar los dos casos extremos, con el fin de simplificar el análisis. Armamos una copia del dataset dejando solo los años de interés, graficamos los dos box plot y parece ser que la relación es al revés de lo que pensábamos!, 1991 que es el año con más canciones lanzadas parece tener canciones un poco más populares, pero… ¿qué tanto más populares?.



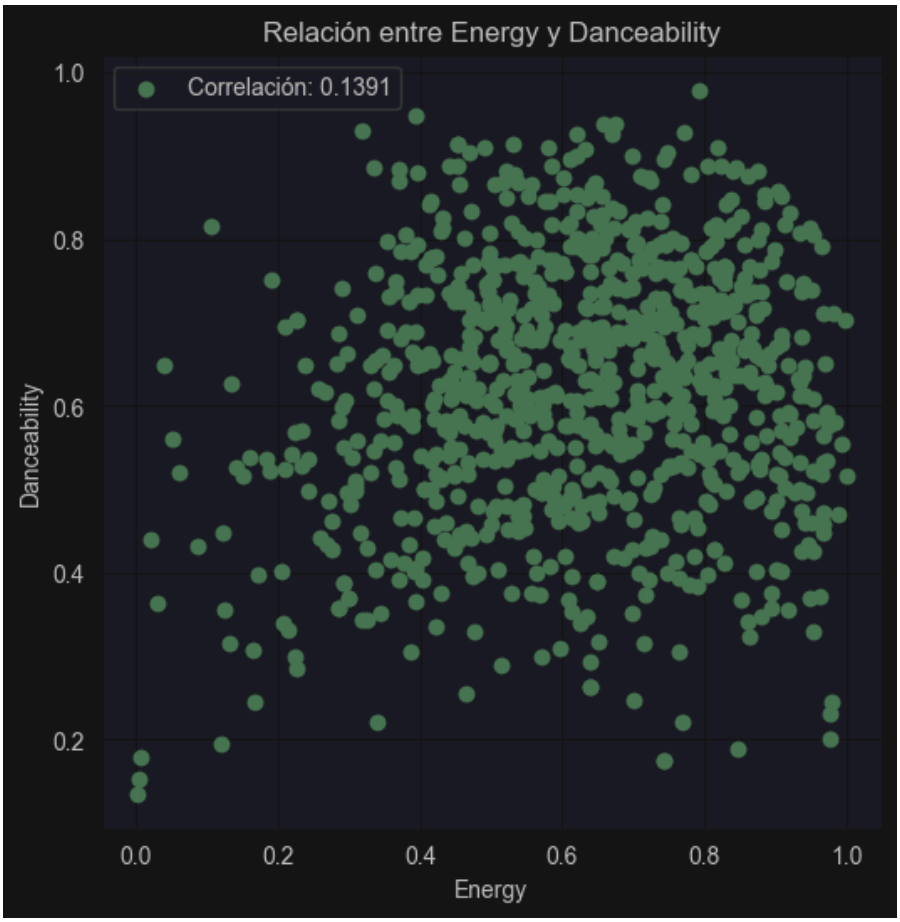
Analizamos normalidad con Shapiro y nos da que son normales, ¡increible!. Luego, homocedasticidad y tampoco nos falló por suerte. Todo parecía propicio para un test-t así que lo aplicamos. El test t nos falló por muy poquito, nos dio un p valor del 0.059, casi 0.05. Pero bueno, nada indica que haya una diferencia significativa entre la popularidad de las canciones de ambos años.

Hipotesis 3: Danceability vs Energy

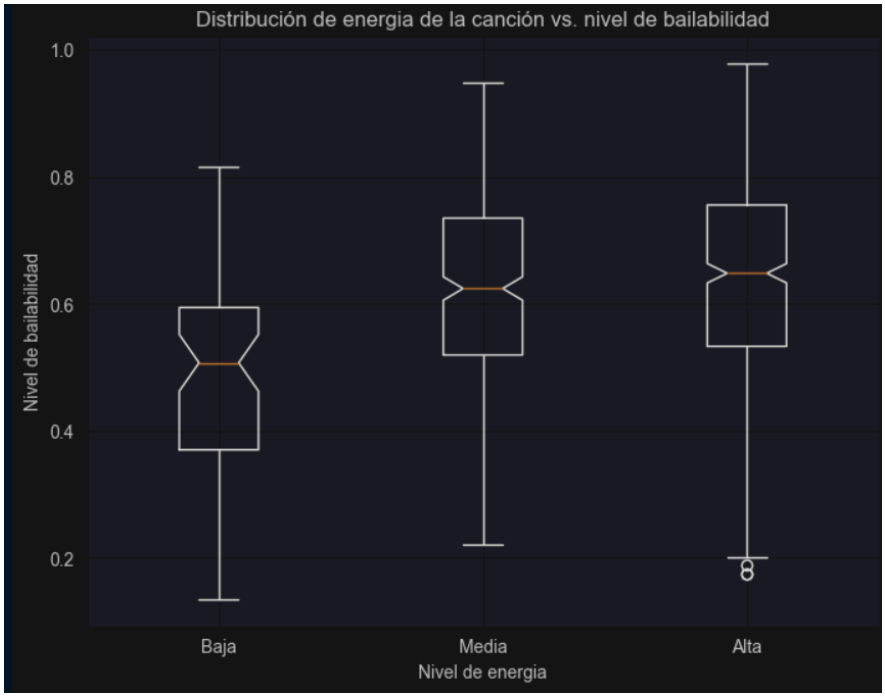
Otra hipótesis interesante que se nos ocurrió fue la posible relación entre la medida de cuán bailable es una canción y el nivel de energía de la pista.

Para este análisis, consideramos adecuado incorporar un gráfico de dispersión (Scatter-Plot) para observar de forma más clara una posible correlación lineal entre ambas

variables. Al analizar el gráfico, encontramos una correlación muy baja, en torno a 0.13, donde se podía apreciar una nube de puntos sin una tendencia clara.



No nos dimos por vencidos y decidimos explorar cómo se comporta el nivel de energía en distintos rangos de bailabilidad. Para ello, dividimos la variable Energy en tres grupos característicos: baja energía (valores inferiores al 30%), media energía (valores entre 30% y 60%), y alta energía (valores de 60% en adelante). Luego, comparamos la variabilidad de los grupos respecto a los cambios en Danceability utilizando un gráfico de cajas (Box-Plot), como el siguiente:



En el gráfico se observa una diferencia aparentemente notable entre los valores de energía bajos, medios y altos. Sin embargo, esta diferencia no es tan evidente entre los grupos de media y alta energía. Para validar nuestra hipótesis de que estas diferencias eran realmente significativas, decidimos probarlo utilizando un test de ANOVA o, en caso de no cumplirse los supuestos, un test de Kruskal-Wallis.

Primero, intentamos con ANOVA, para lo cual es necesario validar los siguientes supuestos:

- **Independencia de las muestras**: La observación de un grupo no debe influir en las observaciones de otro grupo.
- **Normalidad en las distribuciones de los datos**: Las variables deben distribuirse de forma normal.
- **Homocedasticidad (igualdad de varianzas)**: Los grupos deben tener varianzas similares.

Asumimos que el muestreo se llevó a cabo de forma aleatoria, por lo que podríamos considerar que existe independencia entre las observaciones de los grupos. Solo nos quedaba validar los otros dos supuestos:

- ➔ **Normalidad**: Probamos con dos métodos, el test de Shapiro-Wilk y el QQ-plot, y ambos arrojaron valores inferiores a 0.05, lo que indica que la variable Energy no se distribuye de forma normal. Dado que no se cumple este supuesto, descartamos ANOVA.
- ➔ **Homocedasticidad**: También evaluamos la homocedasticidad de las varianzas mediante el test de Levene, y encontramos que las varianzas de los grupos son homogéneas (aunque no es necesario para Kruskal, esta bueno igual analizarlo).

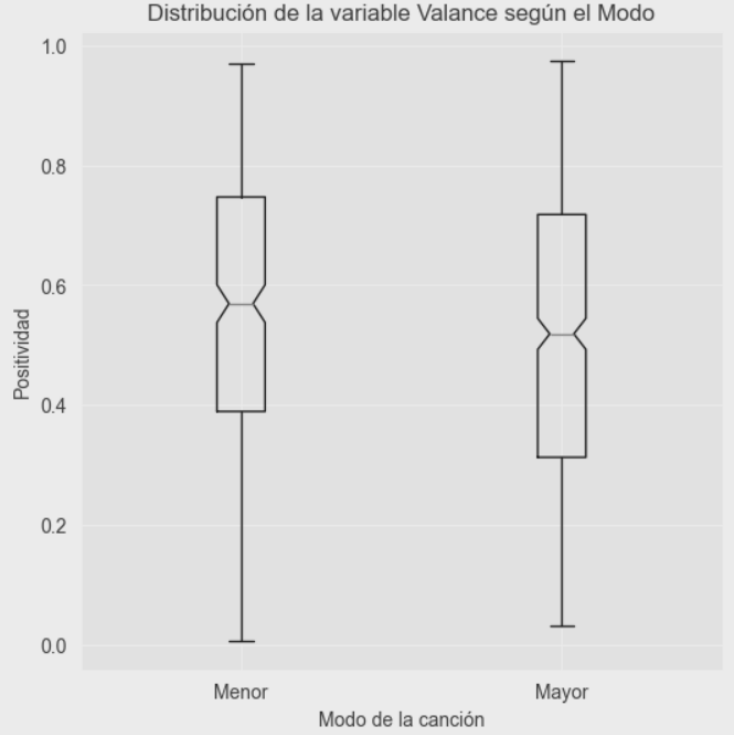
Finalmente, aplicamos el test de Kruskal-Wallis, cuyo resultado fue menor a 0.05 (aproximadamente 0). Por lo tanto, podemos afirmar que existen diferencias significativas entre los grupos. Esto sugiere que para niveles más bajos de energía, las canciones tienden a ser menos bailables, mientras que a mayores niveles de energía, las canciones presentan características que las hacen más bailables. Por ende, se comprueba nuestra hipótesis.

Hipotesis 4: Valence vs Mode

Proseguimos con el análisis de la relación entre el nivel de positividad de la canción y el tono en el que se compuso (mayor o menor). Para ello, dividimos la variable Valence según el Mode (0 para menor, 1 para mayor).

Al observar los valores de media y mediana, notamos que en promedio las canciones en modo menor tienen un Valence (positividad) ligeramente superior, aunque la diferencia parece pequeña. La desviación estándar es similar en ambos grupos (0.23 para menor y 0.25 para mayor), lo que indica que la variabilidad es comparable.

Mediante el siguiente boxplot, se puede apreciar esta diferencia de la que hablamos:



Dado que las medias y medianas no difieren significativamente a simple vista y que tenemos dos grupos, una prueba de Mann-Whitney U o un Test T podría ayudarnos a confirmar si estas diferencias son significativas.

Probamos la normalidad de los datos con el método de Shapiro-Wilk, pero este mostró que Valence no sigue una distribución normal, por lo que descartamos el test t.. Como la homocedasticidad nos dió un valor de 0.078, podemos afirmar que son estadísticamente similares (cosa que a simple vista ya podíamos inferir del análisis del desvío de los grupos).

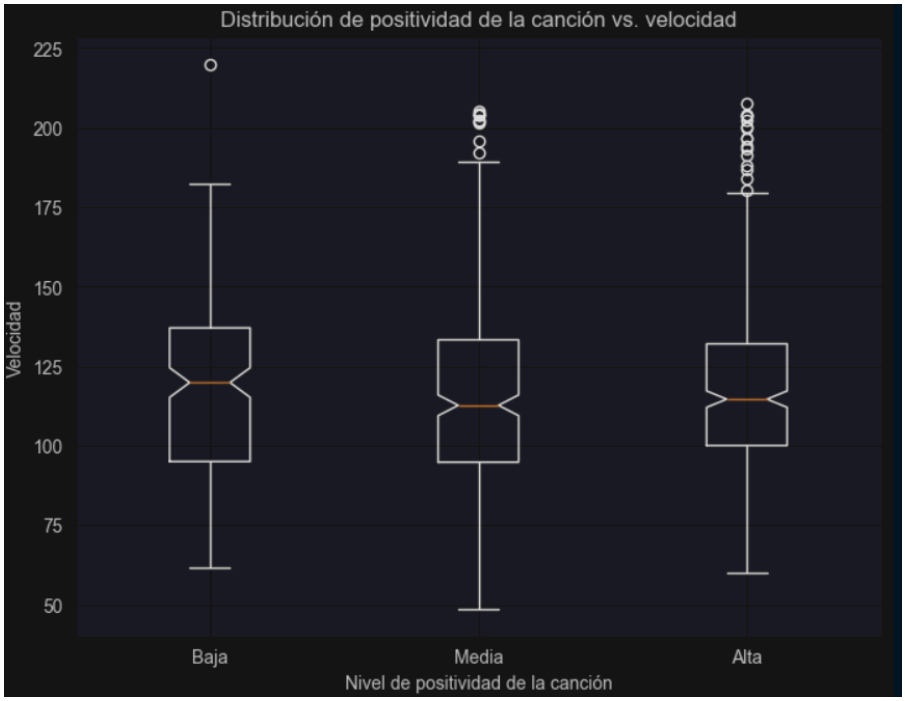
Luego realizamos la prueba de Mann-Whitney U, y el resultado fue sorprendente: la prueba indicó que las diferencias entre las canciones en modo mayor y menor son estadísticamente significativas.

Esto sugiere que, en general, las canciones en escala menor tienden a ser más positivas, lo que implica que podrían ser percibidas como más positivas o alegres que las canciones en escala mayor. Este hallazgo contradice nuestra hipótesis inicial, lo cual nos dejó sorprendidos🤯.

Hipotesis 5: Valence vs Tempo

Podría parecer intuitivo pensar que, dependiendo del nivel de positividad de la canción (Valence), una canción puede ser más rápida o más lenta (Tempo). ¿Será esto cierto? Vamos a averiguarlo…🧐

Primero, realizamos un gráfico de dispersión (Scatter-Plot) para ver si había algún indicio de relación lineal entre las variables. A primera vista, solo se observó una nube de puntos sin dirección ni sentido alguno, por lo que no pudimos concluir nada al respecto. Decidimos entonces separar Valence en grupos para analizar el tempo promedio dentro de cada grupo y observar posibles diferencias. De manera similar al análisis de Danceability vs Energy, dividimos Valence en tres grupos según el nivel de positividad: baja positividad (menor a 0.3), positividad media (de 0.3 a 0.6) y alta (superior a 0.6). Luego, generamos un boxplot para visualizar y comparar los grupos:



A partir de la información del boxplot, parece no haber una diferencia significativa entre los grupos. Para confirmar esto, recurrimos a una prueba estadística como ANOVA o Kruskal-Wallis.

Primero, intentamos verificar la normalidad en los grupos mediante el test de Shapiro-Wilk. Los resultados indicaron que la distribución de Valence no es normal, por lo que descartamos el uso de ANOVA y optamos por aplicar Kruskal-Wallis.

Además, como análisis extra, probamos la homocedasticidad de las varianzas mediante el test de Levene, que indicó que las varianzas entre los grupos pueden considerarse iguales.

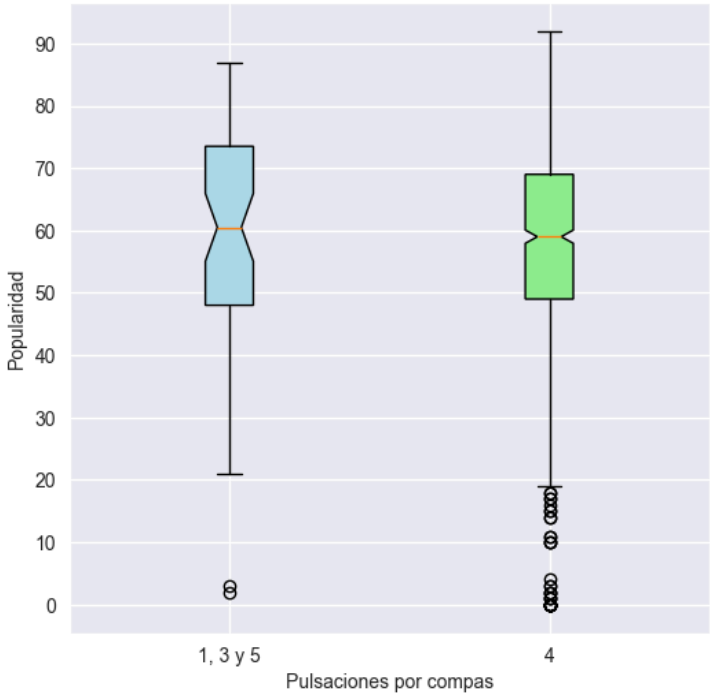
El test de Kruskal-Wallis reveló que no hay suficiente evidencia para rechazar la hipótesis nula (p-valor igual a 0.336), lo cual sugiere que no existen diferencias significativas entre los grupos. En consecuencia, la hipótesis de que, a medida que aumenta el tempo de la canción, también aumenta la positividad, no es válida.

Hipotesis 6: Time Signature vs Popularity

La hipótesis que planteamos es que las canciones que no son de 4 pulsaciones por compás son menos populares que las de 4.

Los posibles valores de Time Signature son {1, 3, 4, 5} y de Popularity hay valores discretos entre 0 y 100. Esta hipótesis surgió de que hay muchísimas canciones con Time Signature = 4, entonces pensamos que las canciones con compás de 4 son más populares que las demás, ya que debe haber alguna razón de porque casi todas son así.

Para el análisis estadístico, al principio pensamos en realizar un análisis bivariado para poder comprender qué tipo de relación posee las pulsaciones por compás de la canción con su popularidad. Agrupamos las canciones de 1, 3, 5 pulsaciones y las de 4 por separado y graficamos un boxplot para compararlas.



Al realizar el gráfico podemos ver que no se ven diferencias significativas entre ambos grupos. Las canciones de 4 pulsaciones por compás tienen una gran variación, allí se encuentra la canción menos popular y la canción más popular. Por otro lado, parece que las canciones que no son de 4 pulsaciones son un poco más populares.

Para verificar estadísticamente si las diferencias son significativas o no decidimos hacer algún test de hipótesis. Primero planteamos la posibilidad de realizar un test t, y comenzamos a verificar si se cumplen o no los supuestos.

Para la independencia, realizamos una prueba de chi-cuadrado y nos dio que si son independientes. Para la normalidad hicimos un test de Shapiro Wilk y nos dió que no son normales. Debido a esto ya no podemos realizar el test t, pero quizás podamos hacer el de Mann Whitney U.

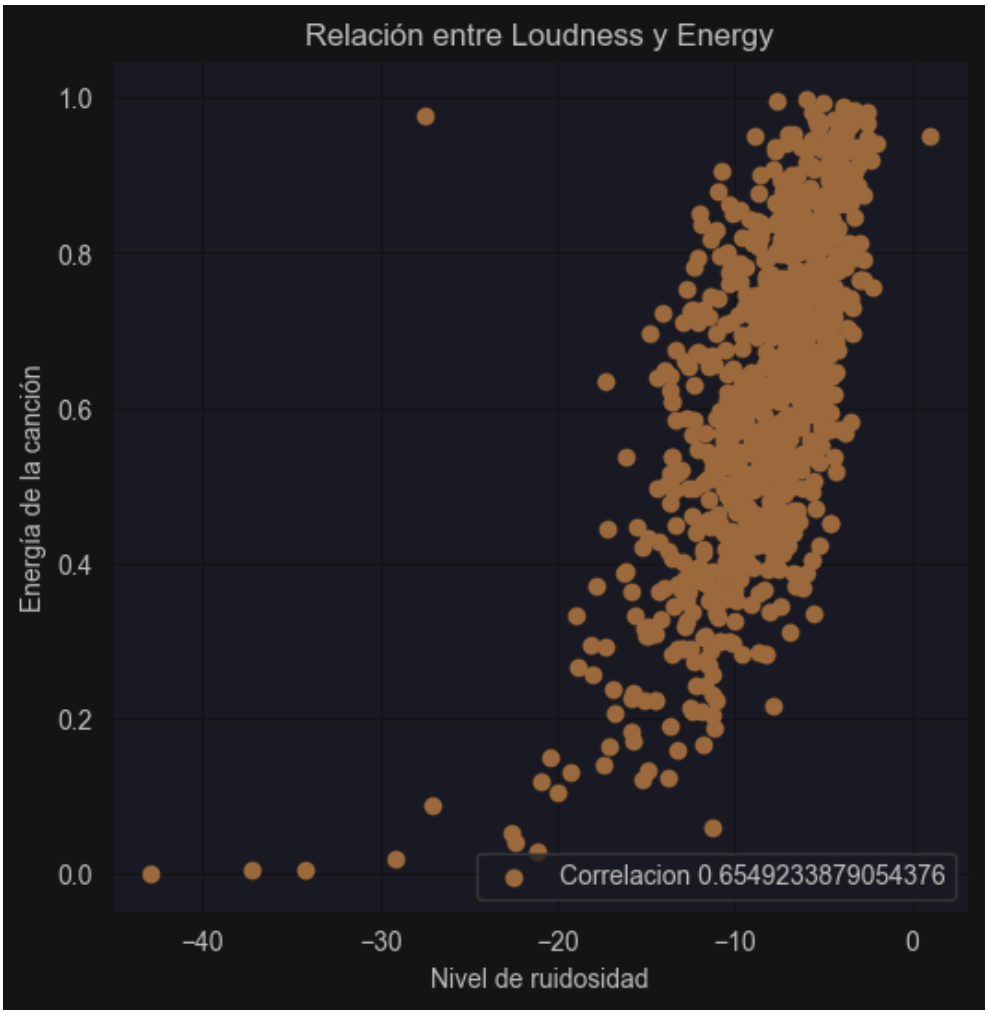
Probamos la homocedasticidad con el test de Levene y nos dio que sí son homocedásticos, y por lo tanto pudimos realizar el test de Mann-Whitney U. Al realizarlo nos dio un p-valor de 0,543. Al ser mayor que 0,05 aceptamos la hipótesis nula del test que nos dice que no hay diferencia significativa entre ambos grupos.

Como resultado, podemos decir que las canciones de 4 pulsaciones por compás son semejantes en popularidad a las canciones de 1, 3 y 5 pulsaciones por compás.

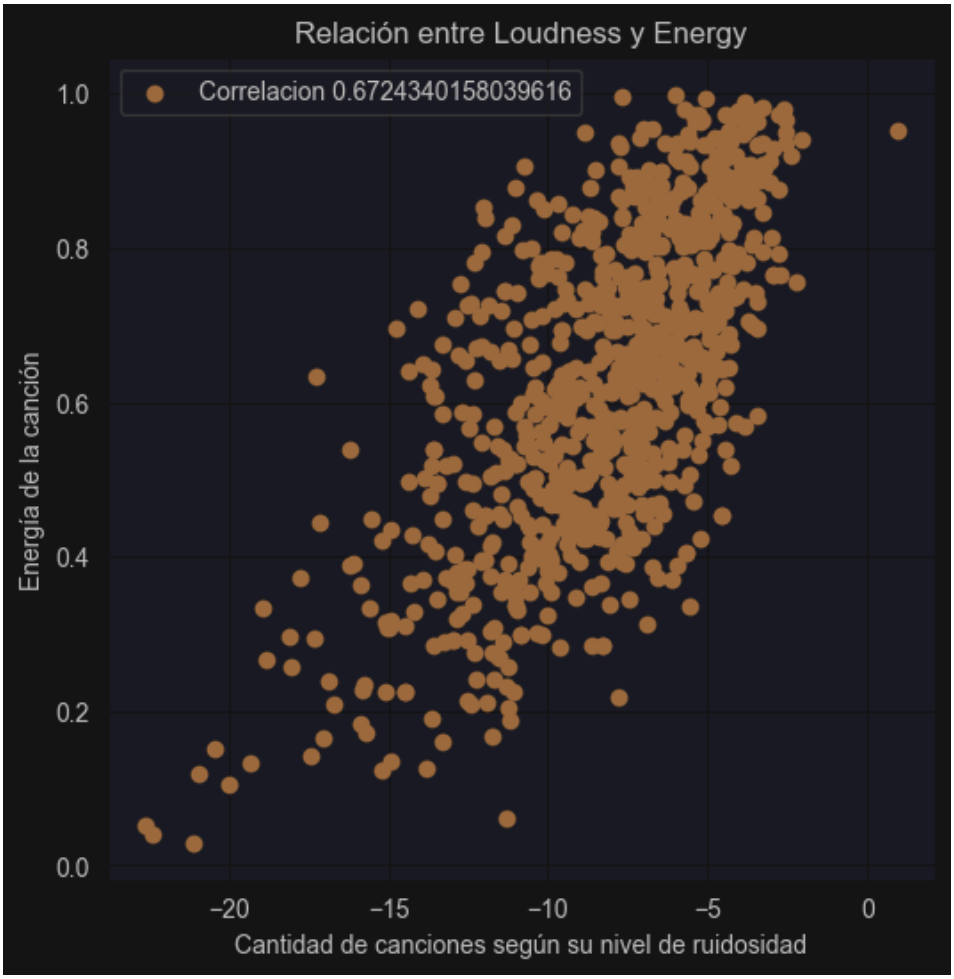
Hipotesis 7: Loudness vs Energy

Como siguiente hipótesis, se nos ocurrió analizar la posible relación entre dos variables que, en principio nos pareció muy clara: La ruidosidad de la canción con el nivel de energía que transmite.

A diferencia de los anteriores análisis, hicimos un Scatter Plot que sorprendentemente no parecía una nube (¡Por fín😎👍!), este tenía una forma cuadrática tirando a una especie de lineal, con una correlación lineal medianamente alta (0.65). Este es el gráfico que nos dió:



Como se puede observar es bastante lineal, decidimos eliminar los valores de loudness que tenían pinta de posibles outliers menores a -25 para que no alterarán la distribución de la variable.



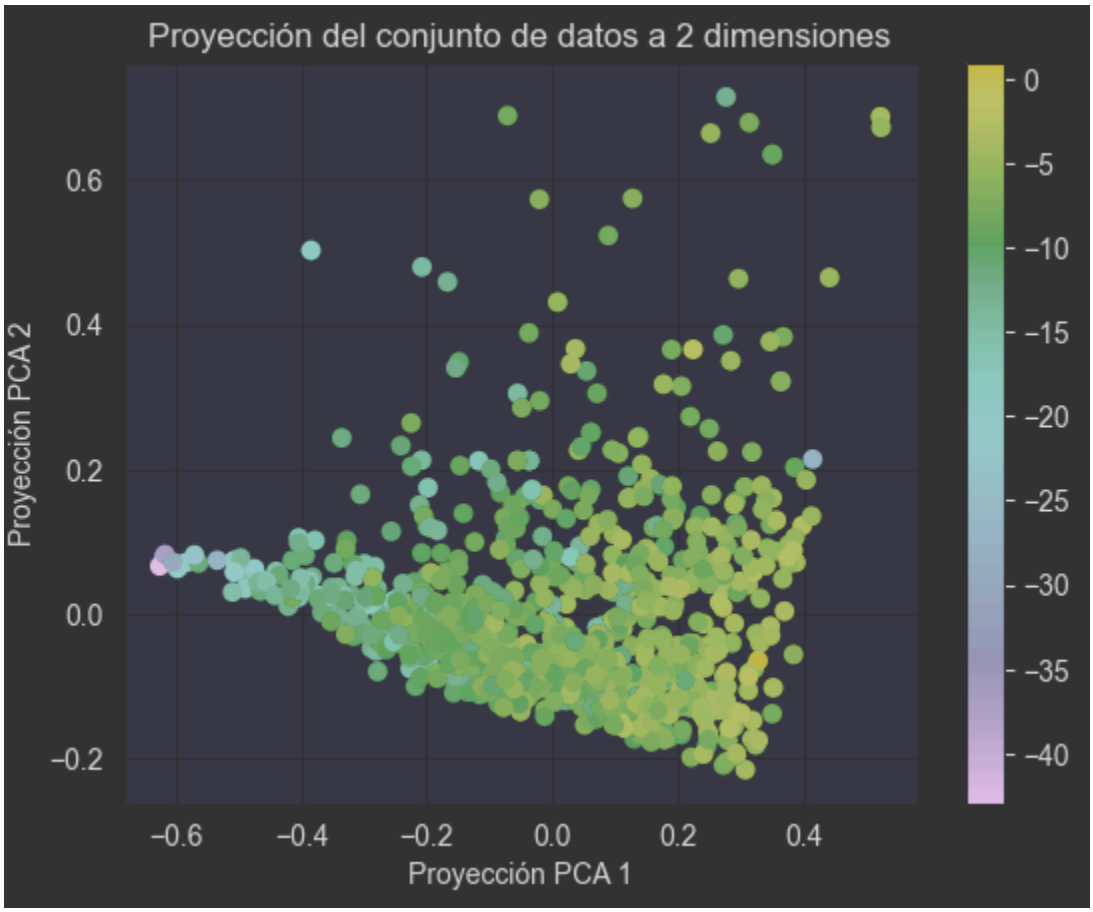
Acá podemos observar de forma más clara la tendencia lineal al eliminar los outliers. Sin embargo, para ver si esta correlación es estadísticamente significativa, decidimos optar por realizar una prueba de hipótesis usando una prueba de correlación de Pearson.

Como el p-valor dió muy por debajo de los 0.05 indica que la correlación es estadísticamente significativa entre las variables, es decir, la relación no se dio de forma aleatoria e indica una relación real entre ellas. Por lo tanto, esto implica que las canciones con mayor nivel de ruidosidad tienden a ser también las que presentan niveles de energía más altos. Se comprueba la hipótesis.

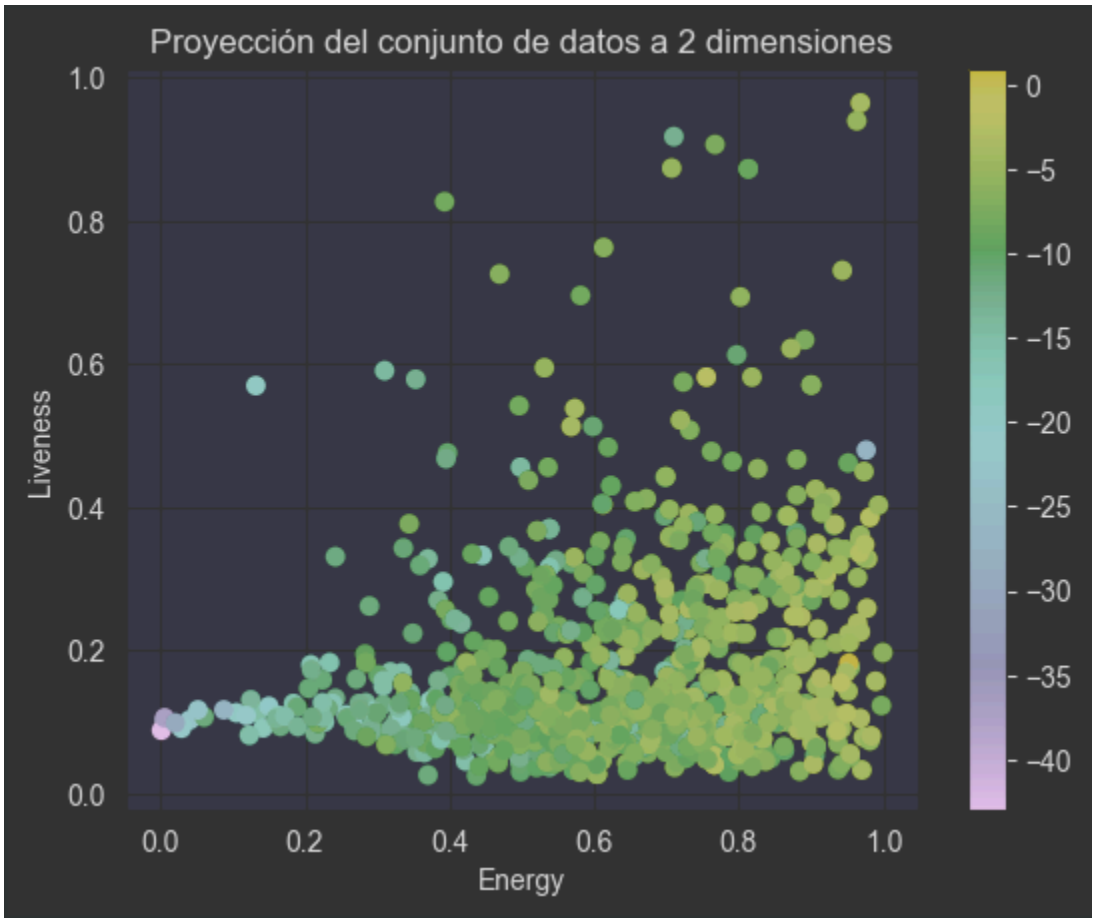
Hipótesis 8: Energy, Liveness y Speechieness tienen una relación con el Loudness de la canción

Originalmente esta hipótesis surgió de la pregunta ¿qué hace que una canción tenga valores más alto de decibelios? o de loudness mejor dicho, y nos pusimos a pensar qué variables podrían estar relacionadas y condicionando el valor de loudness del cover. La primera sería energy, nos resulta natural pensar que cuanto más enérgica sea la canción, el track tenga un volumen más alto. La segunda sería liveness, porque a mayor presencia de gritos y aplausos la canción tendría un valor de loudness más alto. Y tercero speechiness, aunque quizás un poco discutible, cuanto más presencia haya de una voz cantando en el track más podría fluctuar el valor del ruido de la canción. ¡Sin más dilación vayamos al análisis!.

Lo que primero hicimos fue realizar PCA, y pintarlo por loudness para ver si detectamos algún patrón o grupos.

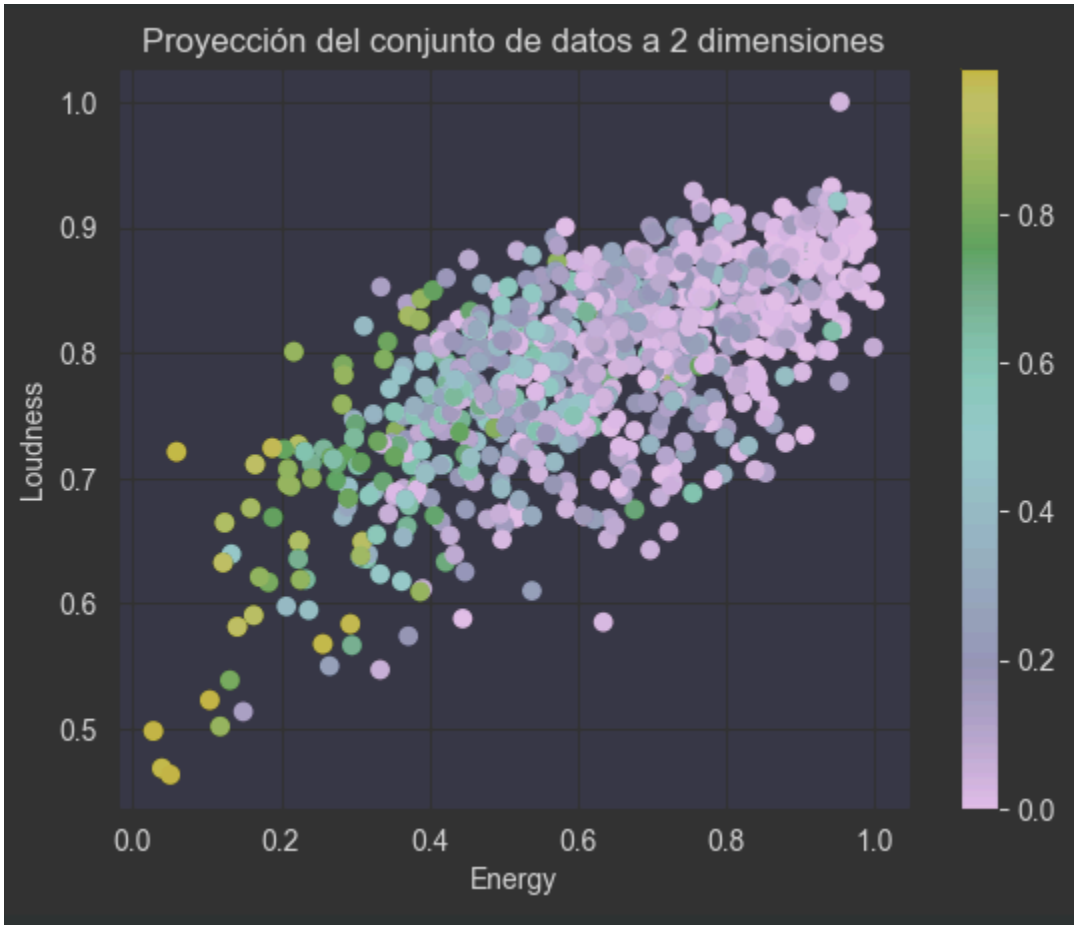


Por lo visto, hay un patrón bastante prometedor y se observa un gradiente de colores. La varianza explicada es de 90% aproximadamente, bastante buena. Lo más interesante es que en el primer componente, la que más influye es Energy con un aporte de 0,97, y en el segundo componente Liveness con 0,97 también, comparado con lo que aportan el resto de las variables es mucho!. Esto quiere decir que speechiness estaría un poco de más, por lo que decidimos bajar de dimensionalidad el análisis, de forma de analizar solamente Energy y Liveness, y cómo se relacionan con Loudness. Realizamos un scatter y pintamos, ¡NO VAMOS A HACER PCA si ya está en dos dimensiones!.



Lo que podemos ver es que el patrón o la forma no cambió demasiado del análisis por PCA cuando estaba speechiness, probablemente debido a lo poco que aportaba en la reducción de componentes, pero… ¿por qué también se siguen manteniendo los colores?. Todo esto nos llevó a cuestionarnos las cosas. Resulta que Energy y Loudness son las dos variables más relacionadas en todo el dataset, con una correlación del 0,65, no es una correlación alta pero la

más alta entre sus pares (“En un reino de ciegos, el tuerto es rey”). Esta relación fue analizada anteriormente, esta hipótesis llevó a otra cuestión, y es a qué variable influencian energy y loudness. Esta fue la sorpresa que nos encontramos, realizamos el scatter, fuimos pintando por las diferentes variables y nos encontramos que tienen una fuerte relación con la acústica de la canción.

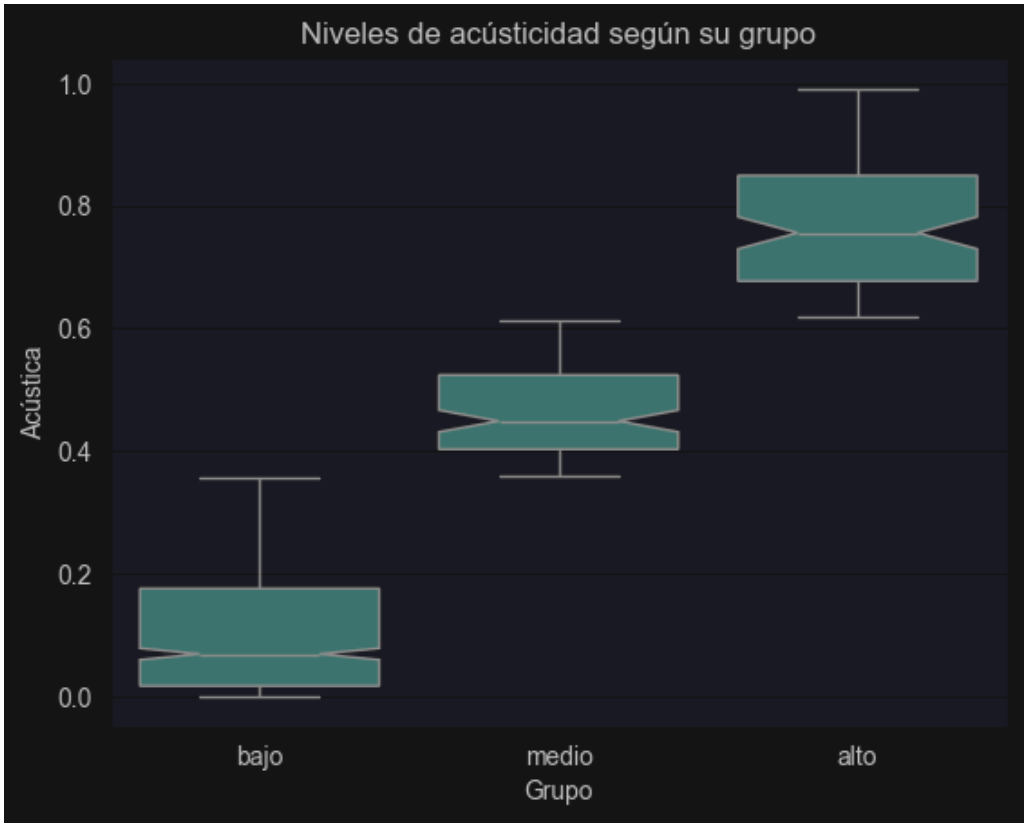


Como vemos en el gráfico, en la parte inferior, o en los valores inferiores de energy y loudness observamos que se concentran los valores más altos de acústica. Tiene mucho sentido lo que estamos viendo, las canciones menos energéticas y ruidosas son las más acústicas, quizás porque son más tranquilas. Nosotros nos imaginamos una canción con guitarra y cada acorde resuena en la sala en donde se grabó, algo bien tranquilo. Por ejemplo, no creemos que Welcome To The Jungle de los Guns ‘n Roses sea muy acústica, porque es una canción con mucho ruido y muy energética!. Y si nos ponemos a hilar muy fino, como vimos en el análisis bivariado de los distintos pares de canciones, la energía y el nivel de acústicidad tienen una relación decreciente

con una correlación del -0.65. Cómo energy y loudness tienen una correlación creciente de 0.65, consecuentemente energy y loudness tienen una relación conjunta con la acústica de la canción. Un embrollo importante pero tiene mucho sentido. Lo podríamos explicar como una especie de transitividad entre las variables.

Todas estas conclusiones nos llevaron a pensar, que más puede estar condicionando la acústica de la canción. Previamente dimos el ejemplo de la guitarra tocando lentamente y resonando cada acorde, y esto nos podría llevar a pensar que el nivel de instrumentalización de la canción tiene que ver con el nivel de acústica de la canción, pero… ¿qué pasa con los covers que son más cantados que instrumentalizados?. ¡Lo veremos en la siguiente hipótesis!. Lo más importante es darle un p valor a todo esto, vayamos a eso.

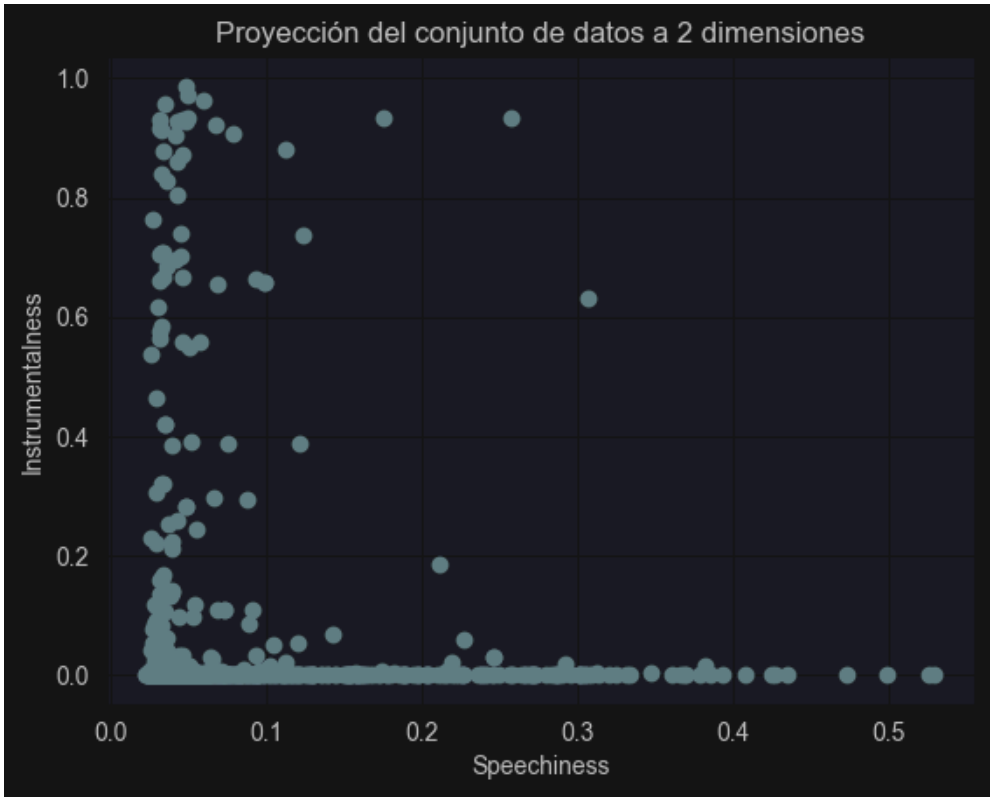
Lo primero sería dividir los grupos según los colores que estamos viendo, decidimos dividir la escala en 3 y graficar los boxplot en base a la acústica (si, ya sabemos que los del grupo de acústica mayor van a tener mayor acústica que los de acústica menor, pero no sabemos si la diferencia es significativa!). Graficamos boxplot y…



Como se puede ver paso lo esperado, pero lo más importante a destacar es que a simple vista las distancias son bastante grandes, ahora bien, ¿qué test utilizamos para corroborar la hipótesis?. Tenemos tres grupos por lo que tendríamos que usar ANOVA si tuviéramos normalidad y homocedasticidad, sino Kruskal. Lamentablemente no son normales los grupos, con Shapiro nos da un p valor de 0 en los tres grupos y los QQ-Plot dan cosas espantosas, así que ANOVA no es una opción. Por lo que decidimos utilizar Kruskal que permite generalizar a más de 2 grupos dándonos un p valor de 0, por lo que la hipótesis es correcta, es decir, que la diferencia entre los tres grupos de acústica es significativa, y como consecuencia también podemos decir que las canciones con menos loudness y energy son más acústicas que el resto.

Hipótesis 9: Las canciones más instrumentales son más acústicas que las que son más cantadas.

Antes de saltar a cualquier análisis sobre la acústica, nos pareció interesante y necesario analizar la relación entre la instrumentalización y la presencia de canto en la canción. Por lo tanto, lo que primero realizamos es un scatter.

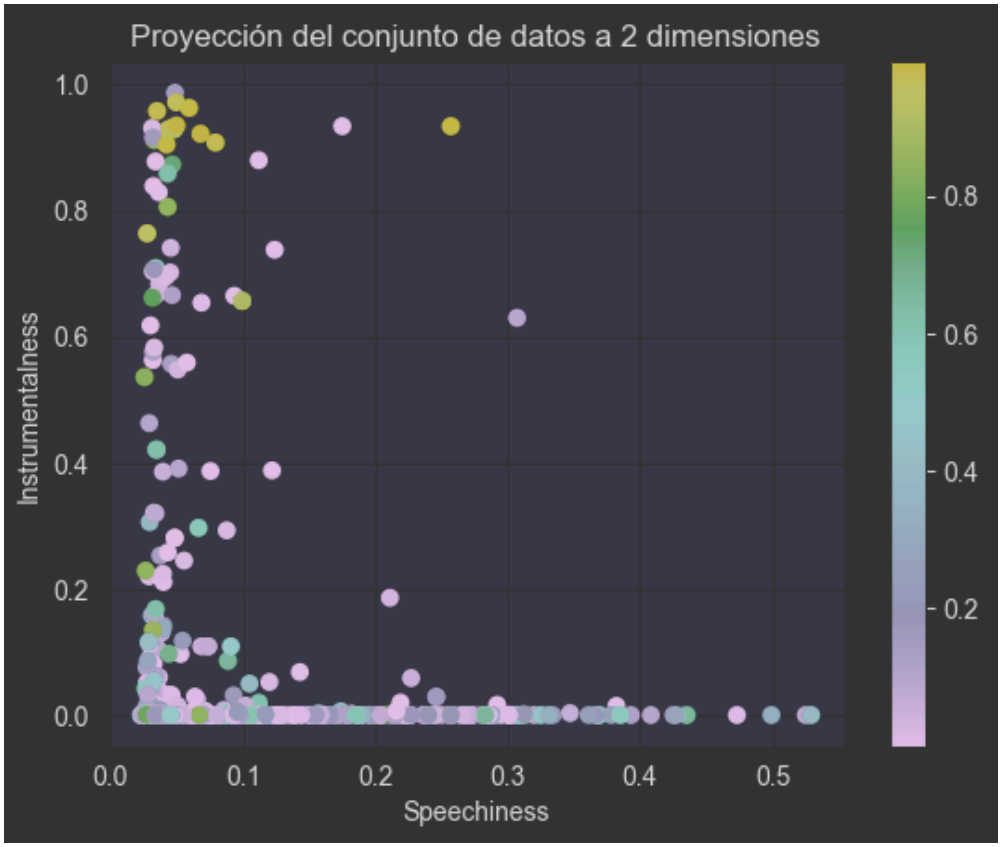


Muy raro el scatter, no hay ningún valor en el medio o muy pocos, solo hay un punto que posee bastante de instrumentalness (0.6) y un poco de speechiness (0.3) (prestar atención que el eje de speechiness está graficado hasta 0.5). Analizando los valores que toman los puntos nos podemos dar cuenta que, o se tienen valores altos de instrumentalness o se tienen valores altos de speechiness, no de ambos simultáneamente. ¿Qué nos quiere decir todo esto?, nos quiere decir que probablemente exista alguna relación entre ambas por cómo fueron medidas, y adelantando un poco, si lo pensamos mejor podría hasta llegar a resultar intuitivo.

Reflexionando, si instrumentalness contempla la presencia de instrumentos en el track y speechiness la presencia de un cantante a lo largo de la misma, es natural pensar que si tengo más instrumentos sonando, el cantante se va a escuchar menos o va a tener menos espacio en la pista. Si pasara lo contrario, es decir, hubiera más presencia de cantos, habría menos espacio para instrumentos o estos pasarían a estar en un plano secundario en el track. Quizás el asunto no sea porque en un momento se canta y en otro se toca sino por el solapamiento de ambos, cuando alguien está cantando se suele querer que sea la voz lo que resalte del track y por lo tanto lo que se escuche más alto.

El gráfico nos podría indicar que ambas variables casi podrían ser modeladas en una variable dicotómica, “O se tiene una cosa o la otra”.

Ahora bien, qué pasa con la acústica de las canciones que es lo que habíamos planteado al principio, veámoslo con el mismo scatter pero pintado por la acústica de las canciones.



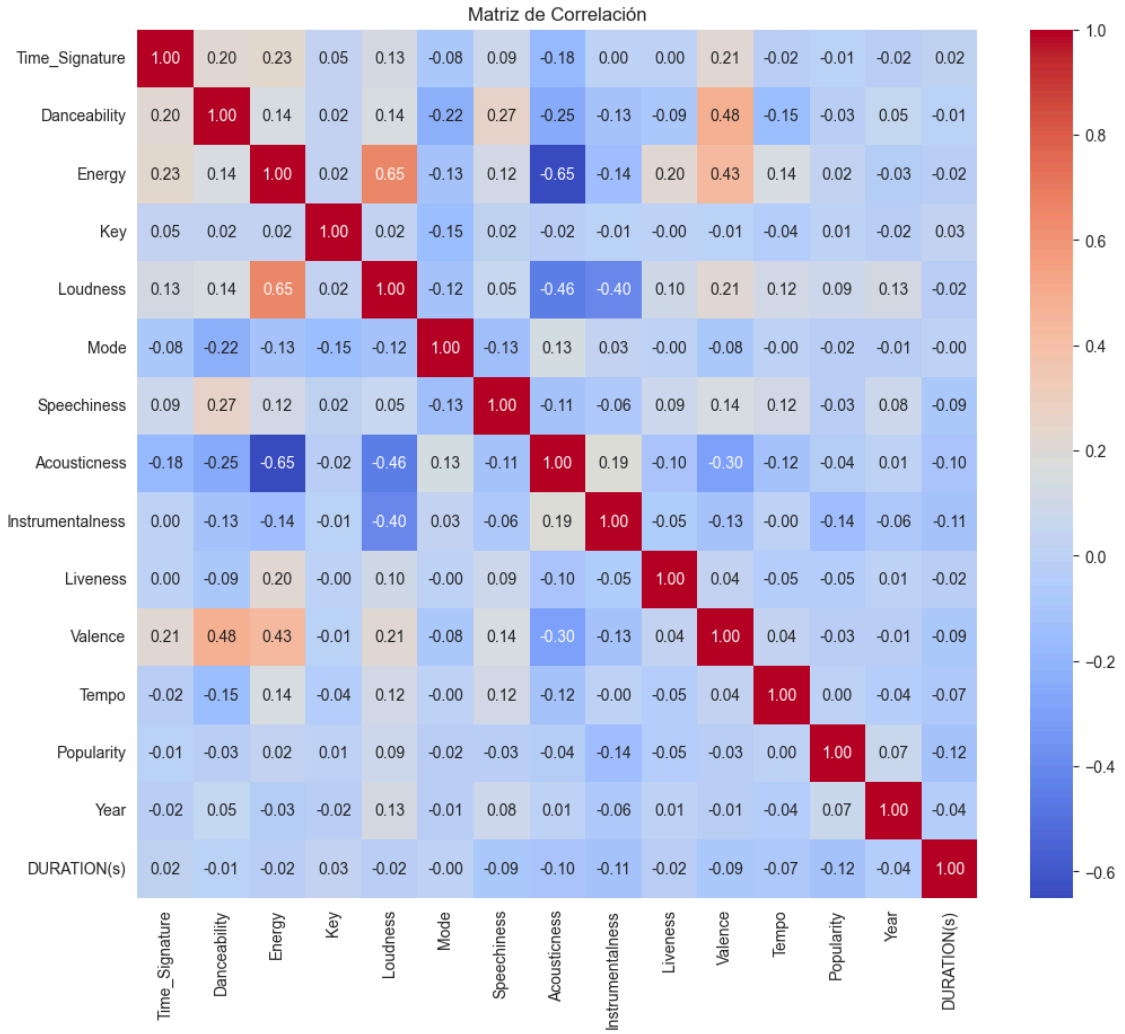
No se podría decir con precisión si las canciones más instrumentales son más acústicas, porque también por arriba de 0,6 de instrumentalización hay valores bajos. Lo que sí podemos asegurar es que las canciones más acústicas están en los valores más altos de instrumentalización, allí podemos ver canciones con un valor de acústica superior al 0.8.

Búsqueda de relaciones

Para analizar este conjunto de datos podemos aplicar técnicas que permitan identificar patrones de popularidad, afinidades entre canciones, y características clave que distinguen unas canciones de otras. Por ejemplo, usando regresión o algunos métodos de clustering.

Aplicación de regresión para el conjunto de datos

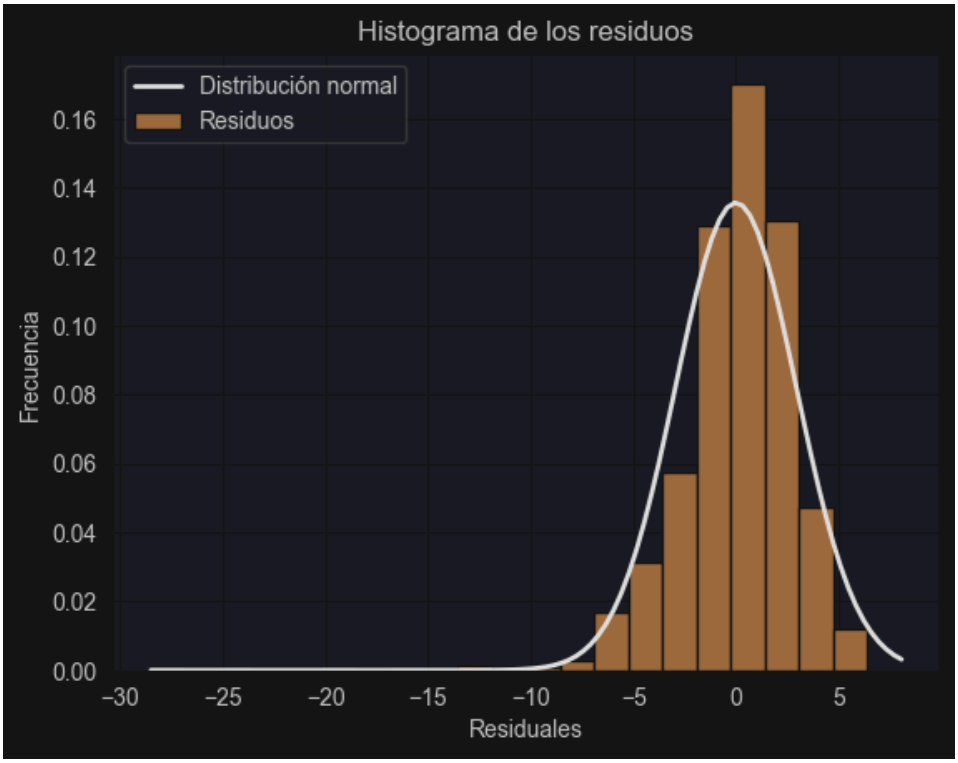
El análisis parte de la idea de predecir la popularidad de una canción en función de sus características. Sin embargo, la matriz de correlación inicial mostró que ninguna variable tenía una relación significativa con popularity, descartando la posibilidad de un modelo de regresión para esa variable.



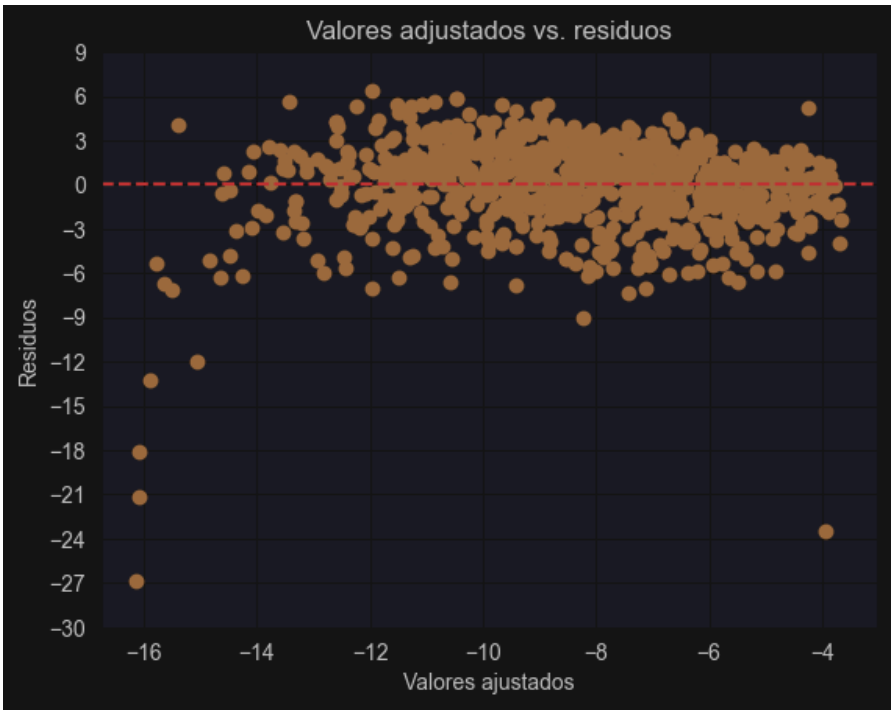
Luego, se reorientó el enfoque hacia la relación entre Energy y Loudness, variables que presentan una correlación significativa (0.65). Un modelo de regresión lineal simple usando Energy como variable independiente sugiere que es posible predecir Loudness de manera razonable.

Resultados del modelo:

1. **R-squared** (42.9%): Indica qué proporción de la varianza en la variable dependiente (o respuesta) es explicada por el modelo de regresión lineal. En este caso, el modelo explica el 42.9% de la varianza total de Loudness.
2. **Significancia del modelo:** Dice si la varianza explicada por el modelo (gracias a las variables independientes) es significativamente mayor que la varianza residual (o el "ruido" que no puede ser explicado por el modelo). Si el valor de F es alto, implica que la varianza explicada es considerablemente mayor que la varianza residual, lo que sugiere que el modelo es significativo. Como “alto” es algo que por lo general es relativo entre una cosa y otra, interesa estudiar el p-valor asociado al estadístico. Dicho valor resultó significativamente bajo ($p << 0.001$), indicando que el modelo es válido.
3. **Coeficiente de Energy** (12.4991): Por cada incremento unitario en Energy, Loudness se incrementa en 12.5 unidades. El p-valor de este coeficiente indica cuánta presencia tiene Energy en la predicción de Loudness. Si es menor a 0.05 (para significancia del 5%), esto nos indica que efectivamente aporta a la definición de la variable. Cómo nos dió un p-valor bajo ($p << 0.001$), se confirma la fuerte asociación entre ambas variables.
4. Supuesto no cumplido: **Normalidad de residuos**
    a. **Con histograma de residuos vs normal:** Los histogramas de los residuos mostraron una distribución ligeramente asimétrica con sesgo a izquierda. Esto indica un desvío de la normalidad



b. **Gráfico de valores ajustados vs. residuos**: Aunque se observó una distribución en banda, existen errores significativos fuera del rango esperado de [-3, 3], violando los supuestos de homocedasticidad.



Por lo tanto, como no se cumplió el supuesto de normalidad de los residuos, es imposible continuar y el modelo de regresión lineal propuesto dejó de ser válido

Exploración de Nuevas Variables:

Se propusieron otras características potencialmente influyentes en Loudness:

Tempo: Ritmos más rápidos podrían correlacionarse con un mayor volumen percibido.
Danceability: Canciones más rítmicas y bailables pueden tener mayores niveles de Loudness.
Instrumentalness: Canciones puramente instrumentales tienden a ser menos ruidosas.
Key y Mode: Algunos modos y tonalidades pueden influir en la percepción de volumen.
Valence: Canciones más alegres podrían estar asociadas con producciones más "fuertes"

Los scatter plots de estas variables con Loudness dieron como resultado la ausencia de una relación lineal fuerte con la variable dependiente. Esto descarta la viabilidad de un modelo de regresión lineal múltiple en este contexto.

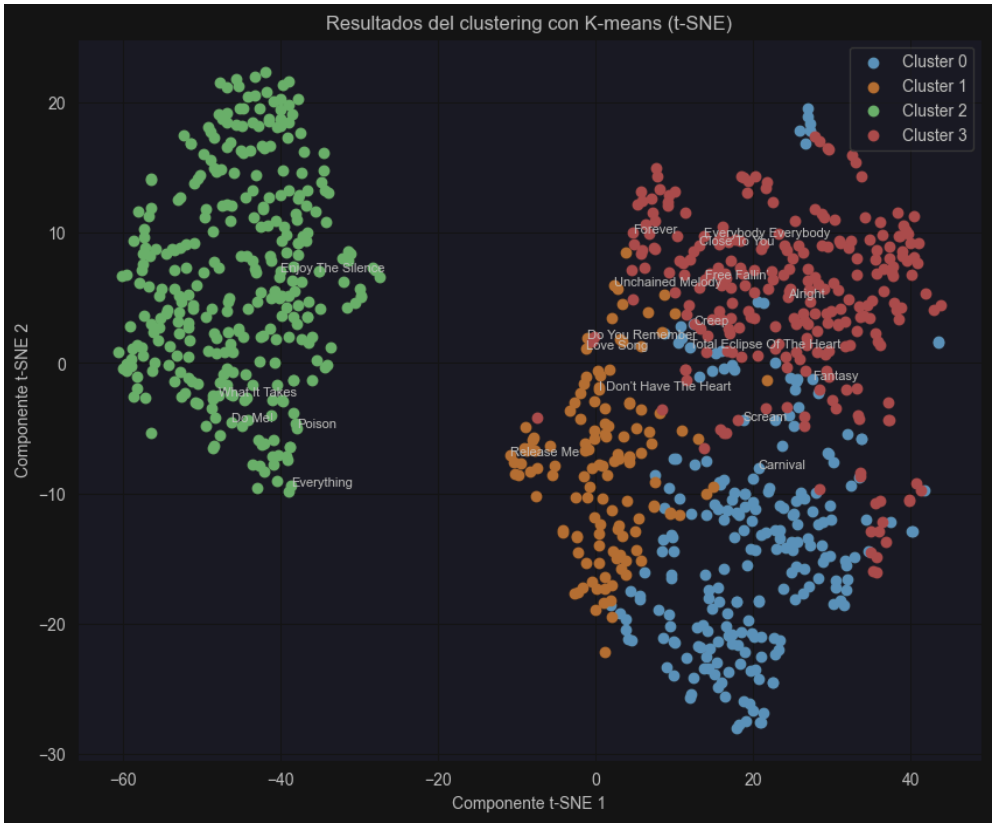
Aplicación de clustering para el conjunto dado

El objetivo del análisis es agrupar canciones según características clave (Danceability, Energy, Tempo, Acousticness, Instrumentalness, Valence, etc) mediante técnicas de clustering. Esto permite identificar patrones comunes entre canciones, facilitando la detección de subgéneros musicales.

Análisis con K-Means

Se comenzó utilizando el algoritmo K-means con 4 clusters. Este método busca centroides que representen mejor los diferentes agrupamientos. Como tenemos dimensionalidad alta, utilizamos t-SNE para proyectar los datos a dos dimensiones.

Visualización de los Clusters obtenidos:



El gráfico muestra cómo las canciones se agrupan en torno a los cuatro clusters, destacando que el **Cluster 2** se separa significativamente del resto. Este patrón indica diferencias sustanciales en las características de las canciones dentro de ese grupo.

Luego, mediante diversos gráficos de barras, se calcularon y mostraron los promedios de cada atributo para comparar diferencias significativas entre los grupos:

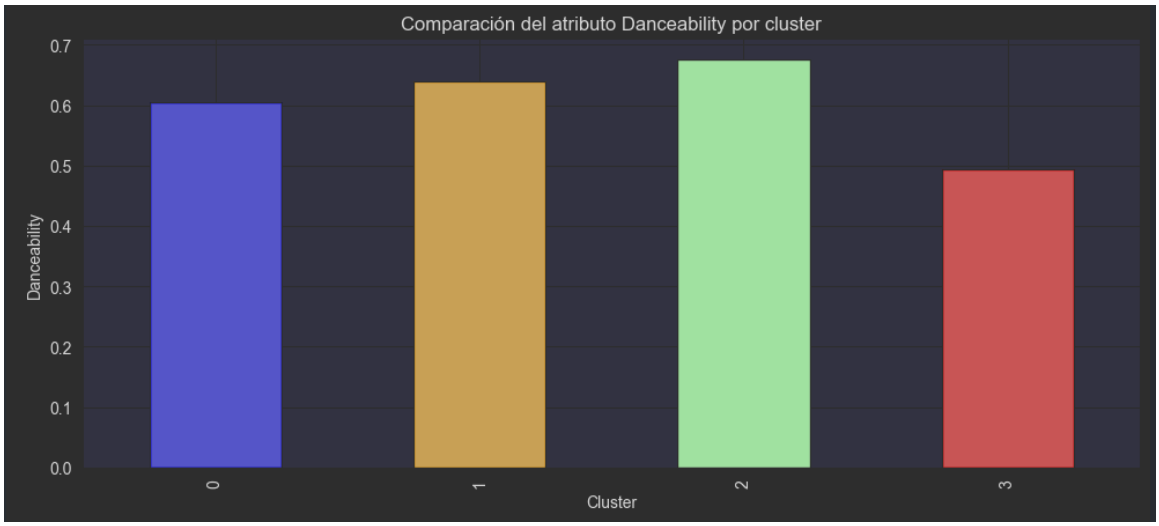
Cluster 0: Canciones instrumentales, enérgicas, bailables con tempos medio-altos. Podrían ser canciones de rock y electrónica.

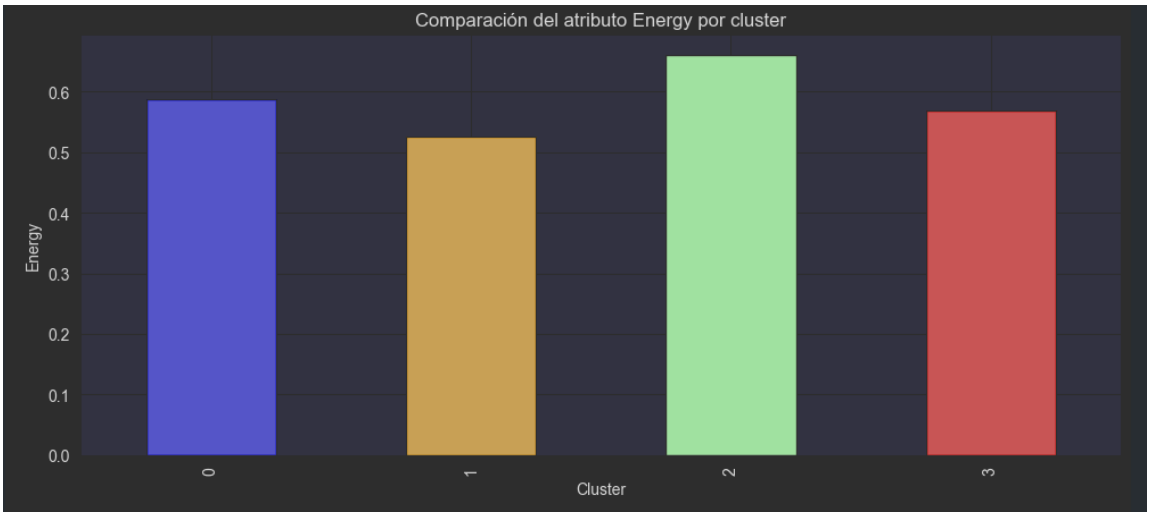
Cluster 1: Canciones acústicas, bailables y de tempo medio-bajo, como baladas.

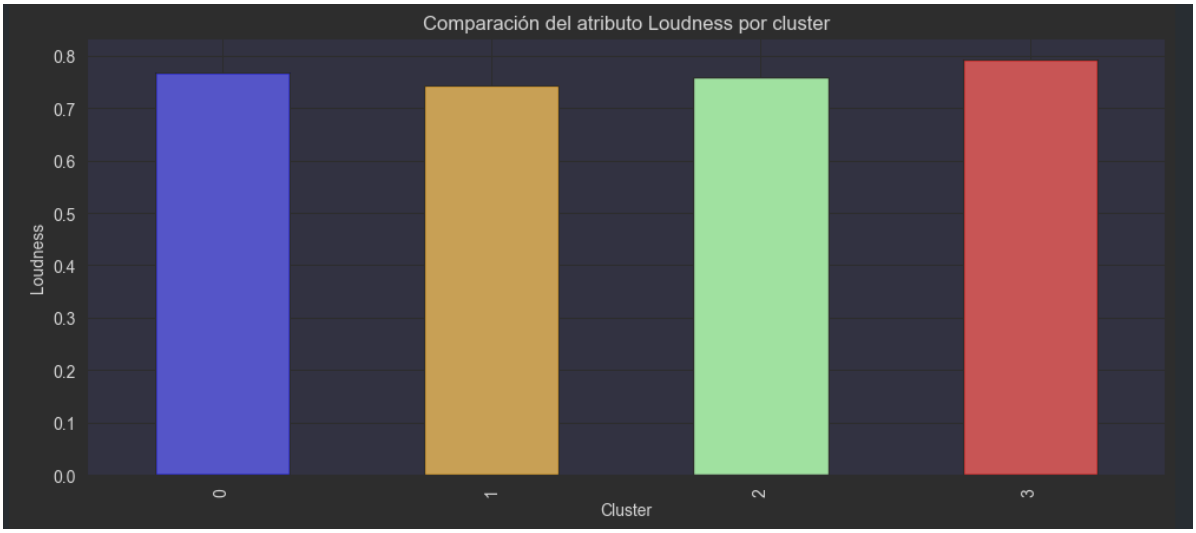
Cluster 2: Canciones rápidas, positivas, y habladas, como rap, hip-hop.

Cluster 3: Canciones acústicas, populares, con tempos medios, ejemplos: folk, jazz

Nota: Cada columna del gráfico representa un Cluster, desde 0 a 3 yendo de izquierda a derecha. Se aprecian las diferencias de los atributos entre agrupaciones, en promedio

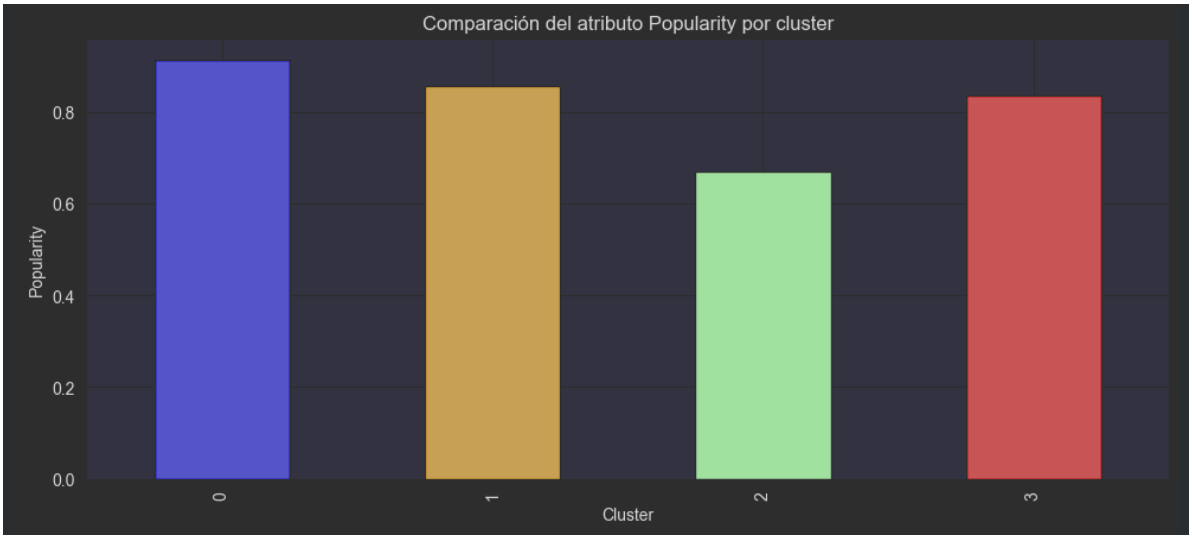




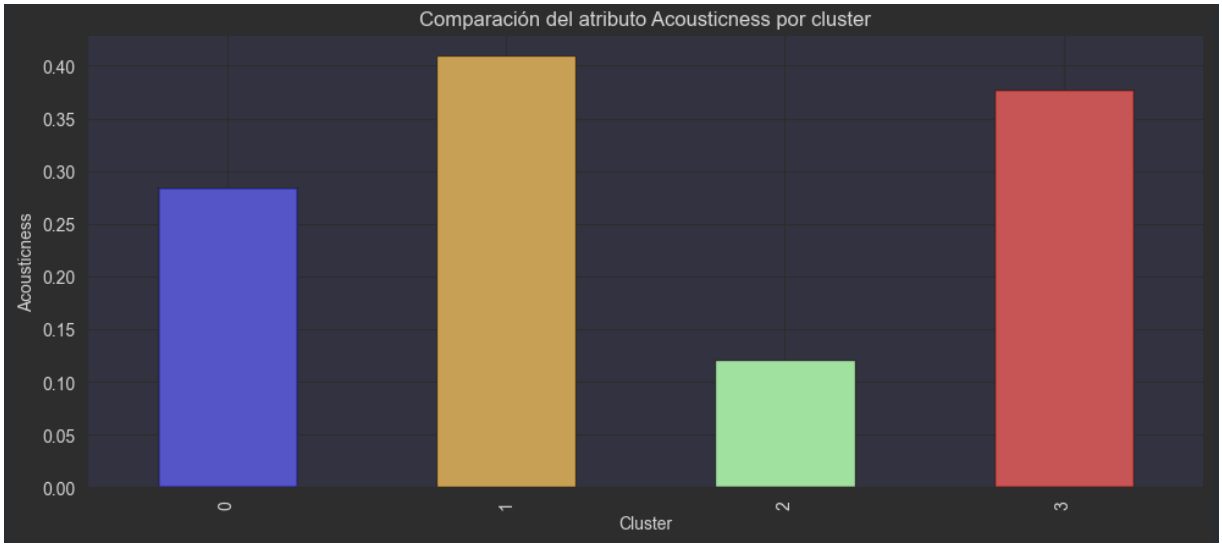

Comparación del atributo Loudness por cluster
Loudness
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
2
3
Cluster


Comparación del atributo Valence por cluster
Valence
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
Cluster


Comparación del atributo Tempo por cluster
Tempo
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
Cluster

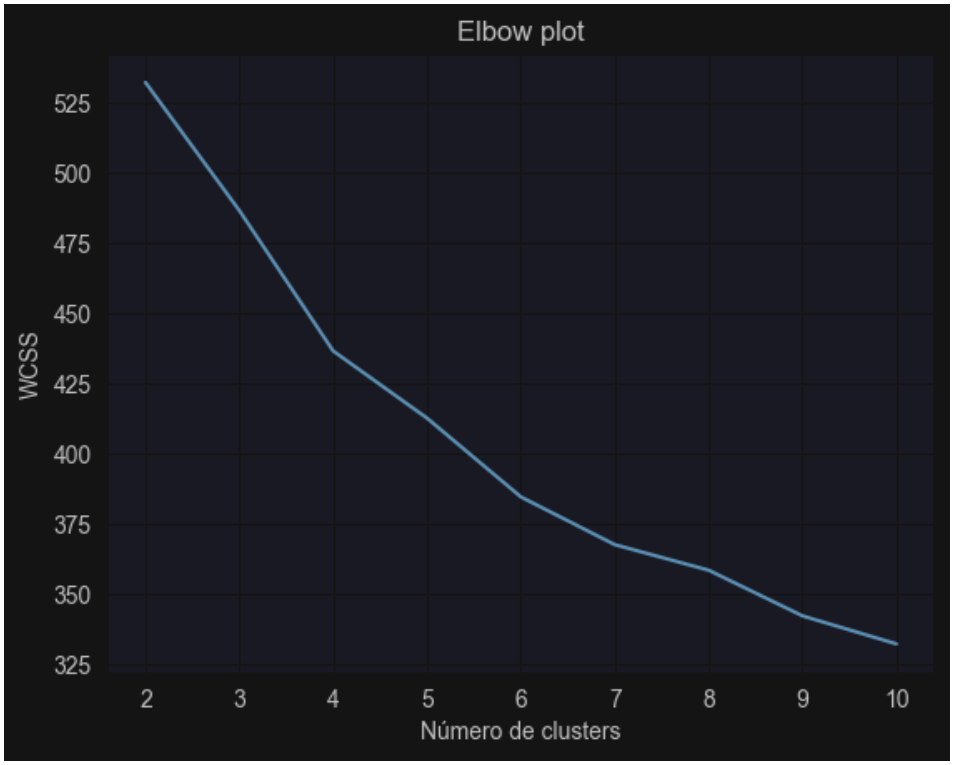

Comparación del atributo Popularity por cluster
Popularity
0.0
0.2
0.4
0.6
0.8
0
1
2
3
Cluster


Comparación del atributo Instrumentalness por cluster
Instrumentalness
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
1
2
3
Cluster



Este análisis confirma que cada cluster representa características específicas.

Para asegurarnos que la elección de K=4 fue una solución correcta. Decidimos armar un Elbow plot.



El "codo" del gráfico nos indica que esa elección fue la adecuada. Como se puede observar, a medida que aumenta el valor de K, disminuye la distancia intra-clusters, está "igualdad" es más pronunciada cuando se supera dicho valor.

Análisis con Clustering Jerárquico

Para mejorar la agrupación y evitar la elección arbitraria del número de clusters, se aplicó ***clustering jerárquico aglomerativo***. Esto permite identificar la cantidad óptima de grupos mediante un **dendrograma**.

Visualización del Dendrograma (acotado para mejor visualización a los 100 clusters más significativos):

Dendrograma.png

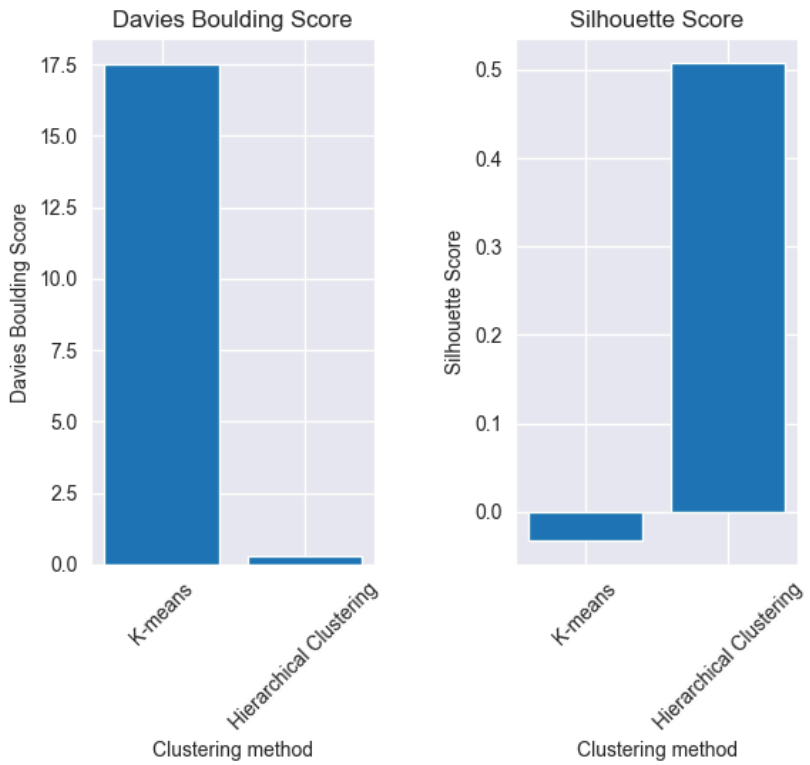
El dendrograma revela que canciones como *Runaway*, *November Rain* y *Feels So Good* forman un cluster separado, posiblemente por su larga duración (cercana a 10 minutos).

Evaluación de soluciones

Se evaluaron las soluciones mediante el **Índice de Davies-Bouldin** y el **Coeficiente de Silueta**:

Davies-Bouldin Score: Indica qué tan bien separados están los clusters (menor es mejor).
Silhouette Score: Mide la coherencia interna de los clusters (mayor es mejor).

Ambos indicadores muestran que el clustering jerárquico supera a K-means en términos de separación y cohesión.



El análisis muestra que el *clustering* jerárquico logra formar grupos más claros y con más sentido que K-means. Esto resulta súper útil si se quiere, por ejemplo, organizar canciones por características comunes, ya sea para armar playlists personalizadas o para identificar géneros automáticamente de una manera más precisa.

Conclusiones

Como pudimos observar, la mayoría de las hipótesis que planteamos se comprobaron estadísticamente. Podemos concluir lo siguiente acerca del comportamiento de las variables del dataset:

- No hay evidencia suficiente para decir que el año de lanzamiento de la canción se relacione con que tan popular es.
- Qué tan bailable es una canción depende del nivel de energía que transmite.
- Las canciones en escala menor son más positivas que las canciones en escala mayor.
- La positividad de la canción no influye en la velocidad de la pista.
- Las canciones que tienen pulsaciones por compás distintas de 4 no son necesariamente menos populares.
- Las canciones más ruidosas son también las que transmiten más energía.
- La energía y el ruido del track del cover tienen un impacto en el nivel acústico de la canción.
- El número de canciones lanzadas por año NO tiene influencia en la popularidad de las canciones de cada año.
- La presencia de público en el track no necesariamente significa que el track sea popular.
- El tono de la canción (menor o mayor) no tiene relación con la presencia de público (ruidos y/o aplausos).
- Es un misterio como la popularidad fue medida, ni si tiene alguna relación con alguna/s variables. Habría que mandarles el dataset a los que lo armaron para que nos expliquen cómo la midieron.

Respecto del trabajo en sí, nos gustó mucho poder trabajar con el dataset de canciones y compadecemos a nuestros pares que recibieron el dataset de la materia fecal. A medida que realizamos el análisis fuimos viendo qué canciones eran más populares, alegres o acústicas y las escuchábamos. Nos encontramos con canciones que teníamos olvidadas, algunos clásicos y joyas por ahí. Aprendimos también, un montón de terminologías y conceptos de la teoría musical producto de tener que informarnos para poder interiorizarnos mejor con el contexto.

Tristemente, quedamos un poco decepcionados, nuestra idea inicial era armar y encontrar la nueva fórmula para armar el futuro "**Hit del verano"**, pero no pudimos encontrar que es lo que hace que un cover sea más popular que otro. Podríamos habernos hecho ricos, lanzar algún single que se llamara "Amo la ciencia de datos", "Median Rhapsody", "Regresión lineal en abril" y por último "Todo comenzó explorando (datos)" de Marama. Se nos ocurren muchos más, pero por las dudas nos detenemos. Seguiremos siendo estudiantes y no productores musicales.

El arranque del trabajo fue todo un tema, lo que más nos complicó fue Github y sincronizar nuestros archivos entre todos, nos encontramos con que para que no tuviéramos que estar haciendo "merge" constantemente teníamos que primero limpiar la notebook antes de hacer pull y push. Hubiera estado bueno que el actual representante de la parte práctica de la materia nos lo hubiera comentado, nos hubiera ahorrado mucha tensión en el grupo y tiempo (Todo bien con vos Nacho, sin tus colab hubiéramos estado perdidos).

Nos arrepentimos de un par de cosas de cómo hicimos el trabajo. Primero, de haber realizado el clustering tarde, probablemente pudimos haber armado hipótesis alrededor de los diferentes "géneros" de canciones que encontramos. Segundo, de no haber planteado una hipótesis alrededor de si los artistas definen o se relacionan con la popularidad del cover, porque encontramos mucho Madonna y Mariah Carey. Igualmente no nos hubiera servido para lanzar la mejor canción del siglo porque ninguno del grupo es alguna de ellas dos.

Nada más que decir, gracias por haber llegado hasta acá y esperamos que les haya sido leve la lectura. Disculpen la informalidad que presentamos en ciertos puntos, pero nos parece que es una forma de demostrar que esto fue escrito por seres humanos, que tienen sentimientos, humor y posturas distintas. "*¡Sean felices!*" ~ I. Orlando .

Referencias

(1) *Compases musicales: Conceptos importantes y tipos de compases.* (2024, November 1). Skoove. Retrieved November 5, 2024, from https://www.skoove.com/blog/es/compases-musicales-conceptos-importantes-y-tipos-de-compases/

(2) *El tempo y la música*. (n.d.). hacercanciones.com. Retrieved November 6, 2024, from https://hacercanciones.com/tutorial/el-tempo-y-la-musica/

(3) *Diferencia entre escalas mayores y menores*. (n.d.). ArtsMúsica. Retrieved November 6, 2024, from https://www.artsmusica.net/teoria-musical/diferencia-entre-escalas-mayores-y-menores/

(4) Piano From Scratch. (2020, June 27). *What does 'in the key of' mean? // Beginner music theory for piano*. YouTube. Retrieved November 6, 2024, from https://www.youtube.com/watch?v=h2k-NDPTL4o

(5) Altozano, J. (2017, November 27). *Qué son las TONALIDADES. Tonalidad mayor y menor. | Jaime Altozano*. YouTube. Retrieved November 6, 2024, from https://www.youtube.com/watch?v=o6aOC3rERF0