Universidad Nacional del Centro de la

Provincia de Buenos Aires

**FACULTAD DE CIENCIAS EXACTAS**

Ingeniería de Sistemas



**Trabajo Práctico Especial**

**Fundamentos de Ciencia de Datos**

**GRUPO 30**

Goncalves Dias Lucia: [lgoncalvesdias@alumnos.exa.unicen.edu.ar](mailto:lgoncalvesdias@alumnos.exa.unicen.edu.ar)

Redolatti Emilia: [eredolatti@alumnos.exa.unicen.edu.ar](mailto:eredolatti@alumnos.exa.unicen.edu.ar)

ÍNDICE

**Introducción 2**

**Información del Data Set 3**

**Exploración del dominio 4**

**Análisis inicial 5**

**Hipótesis 6**

**Conclusión 12**

**Bibliografía 13**

# 

# 

# Introducción

Este informe presenta un análisis exploratorio de un conjunto de datos que abarca canciones de una década clave en la historia musical del siglo XX: los años 90. Este período estuvo marcado por una notable diversidad musical, impulsada por la consolidación de géneros como el grunge, el hip-hop, el eurodance, y la expansión del pop y el rock alternativo. Fue una época de innovaciones en el estilo y la producción musical, reflejando profundos cambios culturales y tecnológicos.

Para este análisis, exploramos el dominio de los datos, su distribución, y los valores predominantes, entre otros aspectos. A partir de esta comprensión inicial, formulamos preguntas orientadas a caracterizar las canciones de la década y sus particularidades. Empleando técnicas de visualización, estadística descriptiva y modelos predictivos, este estudio busca describir la estructura musical de los 90, abordando los comportamientos de las variables tanto de manera individual como en sus interrelaciones.

Información del Data set

El conjunto de datos, está formado por 919 canciones, que serían las muestras, también incluye una variedad de variables que nos permitirán profundizar en estas cuestiones. Las variables consideradas son:

* Track: Título de la canción.
* Artist: Intérprete o grupo que grabó la canción.
* Duration: Duración de la canción, medida en minutos y segundos.
* Time\_Signature: Métrica musical de la canción, que indica el número de pulsaciones por compás.
* Danceability: Medida de qué tan adecuada es la canción para bailar, basada en el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad.
* Energy: Medida de la intensidad y actividad en la canción, donde valores más altos indican una pista más enérgica.
* Key: Tonalidad musical en la que está compuesta la canción, representada por un número entero.
* Loudness: Volumen promedio de la canción, medido en decibelios (dB).
* Mode: Modalidad de la canción, que indica si está en tono mayor o menor.
* Speechiness: Medida de la presencia de palabras habladas en la canción; valores más altos indican cualidades más parecidas al habla.
* Acousticness: Medida de la probabilidad de que la canción sea acústica; valores más altos indican una mayor presencia de sonidos naturales.
* Instrumentalness: Medida de la presencia de voces en la canción; valores más altos representan canciones más instrumentales.
* Liveness: Probabilidad de que la canción haya sido interpretada en vivo; valores más altos indican mayor ruido de audiencia.
* Valence: Medida de la positividad musical de la canción; valores más altos indican música más alegre o positiva.
* Tempo: Velocidad o ritmo de la canción, medida en pulsaciones por minuto (BPM).
* Popularity: Puntuación que refleja la popularidad de la canción, generalmente basada en el número de reproducciones y otras métricas.
* Year: Año de lanzamiento de la canción.

# 

# Exploración de dominio

Para poder comenzar con el análisis primero hay que tener en cuenta el contexto de los datos que vamos a estudiar. El conjunto de datos, que abarca 919 canciones interpretadas por 536 artistas de la década de los 90, nos sitúa en un período de gran transformación en la industria musical. Los años 90 fueron una época marcada por la innovación tecnológica y una notable diversidad de estilos y géneros, los cuales redefinieron el sonido y la distribución de la música.

Durante esta década, el avance de la tecnología digital no solo transformó la producción y distribución musical, sino que también cambió la forma en que el público consumía y experimentaba la música. Este cambio impulsó la popularidad de estilos que hasta entonces no formaban parte del mainstream, como el ska punk, el reggae fusión, el punk rock, el pop punk, el new wave, el alternative rock y el pop rock. La inclusión de estos géneros reflejó una era de fusión y eclecticismo que permeó la cultura popular, creando una mezcla sonora que resonó en todas las edades y contextos sociales.

Esta pluralidad de estilos contribuyó a la creación de un sonido característico en los 90, donde los géneros alternativos y de fusión no sólo coexisten con el rock, el pop, el hip-hop y el R&B, sino que también contribuyeron a redefinir el concepto de música popular. Cada género dejó una huella única en la cultura popular, y muchos de estos sonidos siguen influyendo en la música contemporánea.

El propósito de este análisis exploratorio es examinar cómo las métricas sonoras y la popularidad de las canciones de los 90 reflejan esta diversidad y complejidad. Analizaremos las características de las canciones en términos de su energía, ritmo, tonalidad, y otros factores que contribuyeron a su recepción y éxito en esta época de cambios rápidos y variados en la industria musical.

# Análisis Inicial

Para comenzar con el análisis de datos, realizamos un exhaustivo análisis exploratorio de cada variable en el conjunto de datos. Esta etapa nos permitió examinar los valores más frecuentes y el rango de cada variable, lo que nos ayudó a verificar si estos valores reflejaban de manera adecuada las características que representan.

Durante este análisis exploratorio, pudimos identificar posibles valores atípicos e inconsistencias en los datos, lo cual nos llevó a una fase de limpieza. En esta etapa, tuvimos el objetivo de corregir registros incorrectos, eliminar duplicados y asegurar que todas las métricas reflejaran con precisión los atributos sonoros y estructurales de las canciones. En general, encontramos que las variables se ajustaban a los rangos esperados, por lo que no fue necesario realizar una limpieza exhaustiva. También pudimos notar que las mayorías de las variables tienen una distribución sesgada, lo que nos generó la duda si eran outliers, pero al investigar los rangos que podían tomar, todos estos valores que resultan atípicos en la distribución estaban dentro de los rangos y acordes con la canción a la que pertenecían. Sin embargo, detectamos valores fuera de rango en la variable *Loudness*, por lo que decidimos excluir algunos registros de esta variable para asegurar la coherencia de los datos. Además, no encontramos valores nulos ni falsos nulos que pudieran afectar el análisis.

Esta fase inicial fue fundamental para preparar los datos de forma adecuada, de modo que pudieran ser usados en los análisis posteriores y para garantizar la fiabilidad de los resultados del estudio.

Además, identificamos las variables que podrían ser de interés para un análisis más profundo en el futuro. La variable de *popularidad* fue especialmente relevante, ya que este indicador refleja el nivel de aceptación y preferencia del público por una canción. Nos pareció importante investigar cómo otras características musicales, como el ritmo, el volumen, la energía o la tonalidad, podrían estar relacionadas con la popularidad. Este análisis permitirá comprender mejor qué atributos contribuyen a que una canción sea más atractiva y popular entre el público.

Por otro lado, también analizamos las relaciones significativas entre las diferentes variables, lo que podría aportar información valiosa sobre las características de las canciones. Nos planteamos preguntas sobre qué métricas son las más importantes para cada atributo, cuáles influyen en la *energía*, en la *bailabilidad*, y en otros aspectos relevantes para la música. Este enfoque nos permitirá identificar patrones y correlaciones que podrían ser útiles para futuros análisis y predicciones.

# 

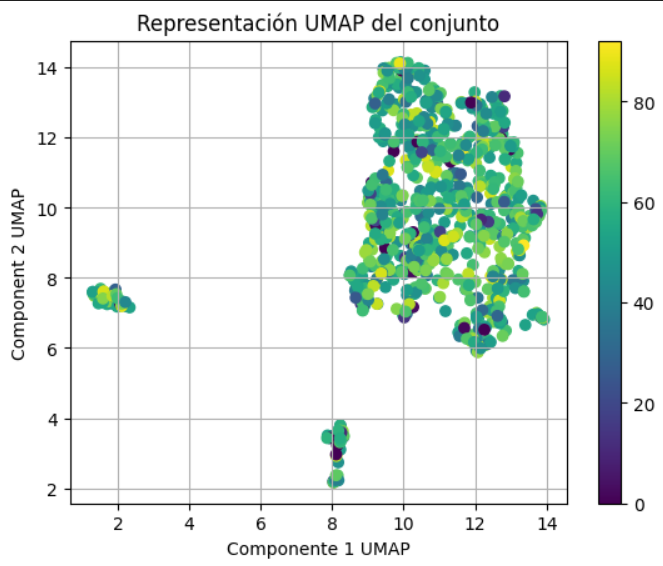
# Hipótesis

**Hipótesis 1: La popularidad se relaciona con el resto de las características**

Dado que la Popularidad es una de las variables centrales en el análisis de las canciones, planteamos una hipótesis general para investigar su relación con otras características y explorar si es posible explicarla a partir de mediciones como Acousticness, Loudness, Energy, entre otras.

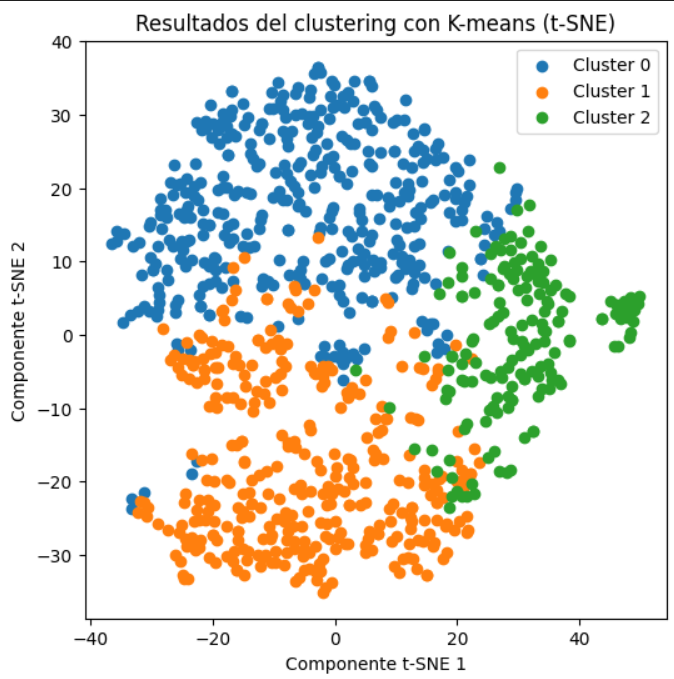
Para comenzar, aplicamos Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos. Sin embargo, observamos que las muestras se dispersaron ampliamente sin formar ningún patrón claro, lo que sugirió una baja correlación lineal entre las variables. A raíz de esto, decidimos utilizar t-SNE, una técnica más adecuada para detectar relaciones no lineales. Con t-SNE, las muestras se agruparon en tres conjuntos distintos, aunque aún no logramos identificar un patrón claro relacionado con la popularidad.

Finalmente, empleamos UMAP, que permitió definir los grupos con mayor precisión, aunque nuevamente sin revelar patrones evidentes en términos de popularidad. Sin embargo, encontramos que la popularidad muestra una variabilidad considerable incluso con valores muy distintos de estas características, lo que sugiere que no existen patrones claros que asocian niveles específicos de popularidad con ciertos valores de otras variables.



Ahora se ven grupos más diferenciados pero los colores están mezclados, por lo que puede ser una indicación de que la popularidad no se asocia fuertemente con las variables utilizadas en el análisis, es decir que la popularidad no está directamente relacionada con una única combinación de características. Por lo tanto, la hipótesis se rechaza.

Además, al analizar la variabilidad de la popularidad dentro del conjunto de datos, observamos que las muestras tendían a agruparse. Esto nos llevó a intentar clasificar las canciones sin considerar la popularidad, con el objetivo de identificar posibles agrupaciones basadas únicamente en las características musicales, sin que estuvieran influenciadas por el nivel de popularidad.

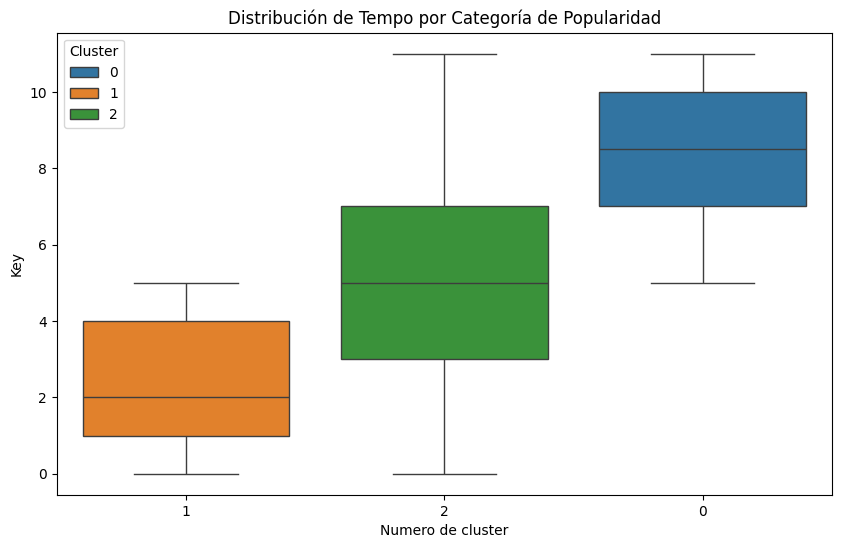


Entonces exploramos el conjunto de datos dividiendo las muestras en tres grupos, donde la variable Key (tonalidad) varía ampliamente. Este hallazgo nos llevó a formular una segunda hipótesis.

**Hipótesis 2: Los grupos de canciones presentan diferencias significativas en la tonalidad predominante de sus composiciones.**

En este análisis exploramos la posible influencia de la tonalidad a la hora de agrupar las canciones, basándonos únicamente en sus características, sin tener en cuenta la popularidad. Para ello, generamos un subconjunto del dataset al que le agregamos una columna llamada “Cluster”, que contiene el valor del cluster al que pertenece cada canción.

Con el fin de analizar la distribución de la variable "Key" en cada grupo, realizamos un gráfico de boxplot, donde se observa que existe una diferencia notable entre los clusters.



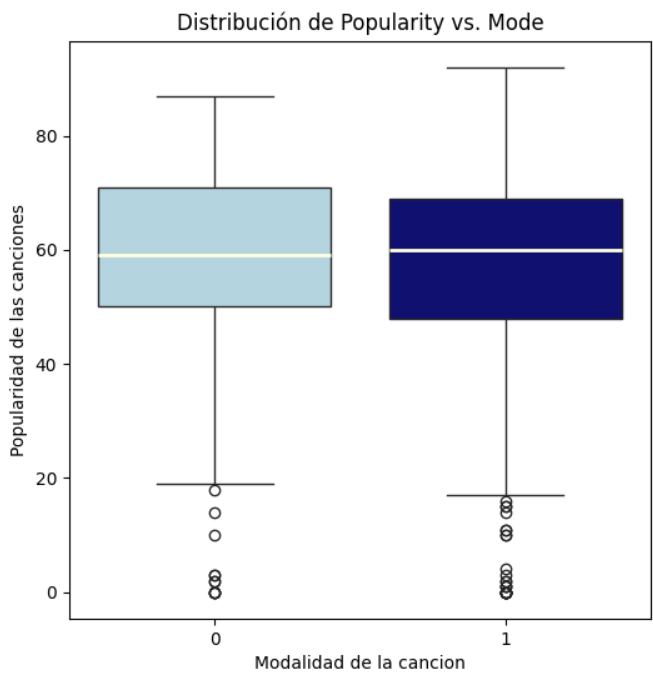
En el gráfico se puede ver que el Cluster 1 tiene una menor variabilidad en la tonalidad, centrando sus valores en tonos bajos de Key. Esto indica que este grupo de canciones tiende a utilizar tonalidades más graves o se concentra en un rango reducido de tonalidades. El Cluster 2, por su parte, muestra una mayor variabilidad en la tonalidad, con un rango mucho más amplio de valores de Key, lo que sugiere que este grupo incluye canciones con una mayor diversidad de tonalidades. Finalmente, el Cluster 0 presenta una distribución intermedia en cuanto a la variable Key, con una tendencia hacia tonalidades más altas que el Cluster 1, pero con menos variabilidad que el Cluster 2.

Para formalizar esta observación, realizamos un test paramétrico. Primero, llevamos a cabo el test de Shapiro-Wilk y de Levene para verificar si se cumplían los supuestos necesarios para una prueba paramétrica. Al comprobar que el conjunto de datos no cumplía los supuestos de normalidad y homocedasticidad, decidimos realizar un test de Kruskal-Wallis. La hipótesis nula de este test es que no existen diferencias significativas entre los grupos en cuanto a la tonalidad predominante (Key). El resultado del test arrojó un p-valor de 0, lo que indica que se rechaza la hipótesis nula, y por lo tanto, podemos concluir que existen diferencias significativas en la tonalidad de las canciones agrupadas en cada cluster, lo que respalda nuestra hipótesis inicial. Entonces la hipótesis se acepta.

**Hipótesis 3: Las canciones más populares están en tono mayor**

En la primera hipótesis dejamos fuera la variable "Mode" debido a su naturaleza dicotómica, por lo que no analizamos su relación con la popularidad. Además, teniendo en cuenta que la mayoría de las canciones en el dataset están en tono mayor y que muchas de las muestras presentan valores medios o altos de popularidad, planteamos la hipótesis de que las canciones más populares podrían estar predominantemente en tono mayor.

Para realizar una primera observación, construimos un gráfico de boxplot para analizar cómo se comporta la variable "Popularidad" en función del tono (mayor o menor). A simple vista, no parece haber una diferencia significativa en la distribución de la popularidad entre las canciones en tono mayor y menor.



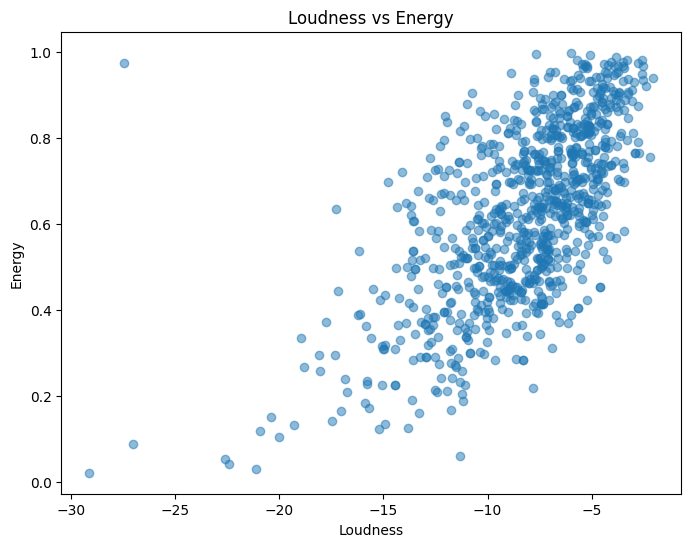
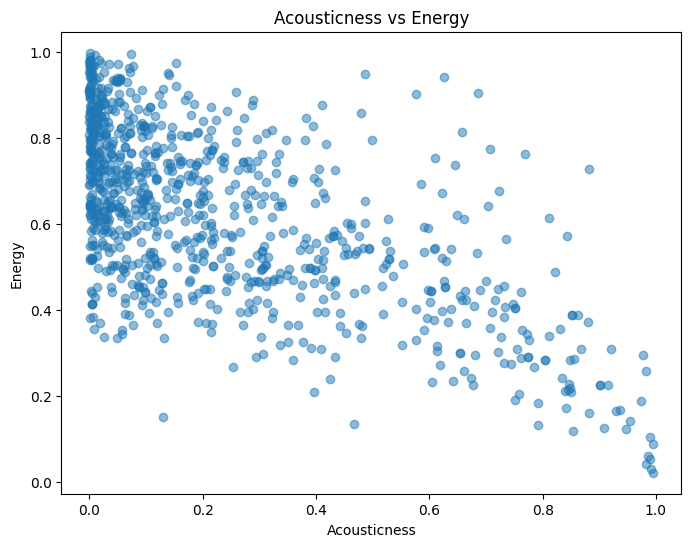
Sin embargo, para confirmar esto formalmente, realizamos un test de Mann-Whitney U. El resultado de este test arrojó un p-valor de 0.756, lo que indica que las diferencias en la popularidad de las canciones entre el tono mayor y el tono menor no son estadísticamente significativas. En consecuencia, se rechaza la hipótesis de que las canciones más populares estén en tono mayor.

|  |
| --- |
| El estudio de las hipótesis 1 y 3 relacionadas a la popularidad nos lleva a pensar que no es posible definir la popularidad de las canciones únicamente en función de sus distintas características. Esto resulta coherente, ya que, al explorar el dominio de los datos, encontramos que el conjunto corresponde a canciones de la década de los 90. Esta época estuvo marcada por una gran transformación en la industria musical, lo cual sugiere que las canciones populares fueron cambiando y presentando distintas características a lo largo de la década.  Esta observación nos plantea la posibilidad de haber enfocado el análisis en los cambios de estas características a lo largo del tiempo, utilizando la variable Year para observar cómo evolucionan las tendencias y los elementos musicales que definieron la popularidad en diferentes momentos de los 90. |

**Hipótesis 4: Energy se puede predecir a partir de Loudness y Acousticness**

En un análisis preliminar, observamos que la correlación entre Energy y Loudness, así como entre Energy y Acousticness, es moderada a fuerte y significativa. Específicamente, cuando Loudness aumenta, Energy también tiende a aumentar, mientras que Acousticness y Energy muestran una relación inversa: cuando una crece, la otra disminuye.

Este hallazgo sugiere que Energy podría ser predecible en función de Loudness y Acousticness. Sin embargo, observamos que aunque ambas variables presentan una tendencia lineal con respecto a Energy, esta relación no es estrictamente lineal, lo cual nos hace dudar de si un modelo lineal será suficiente para capturar la variabilidad de Energy solo en función de estas dos variables.



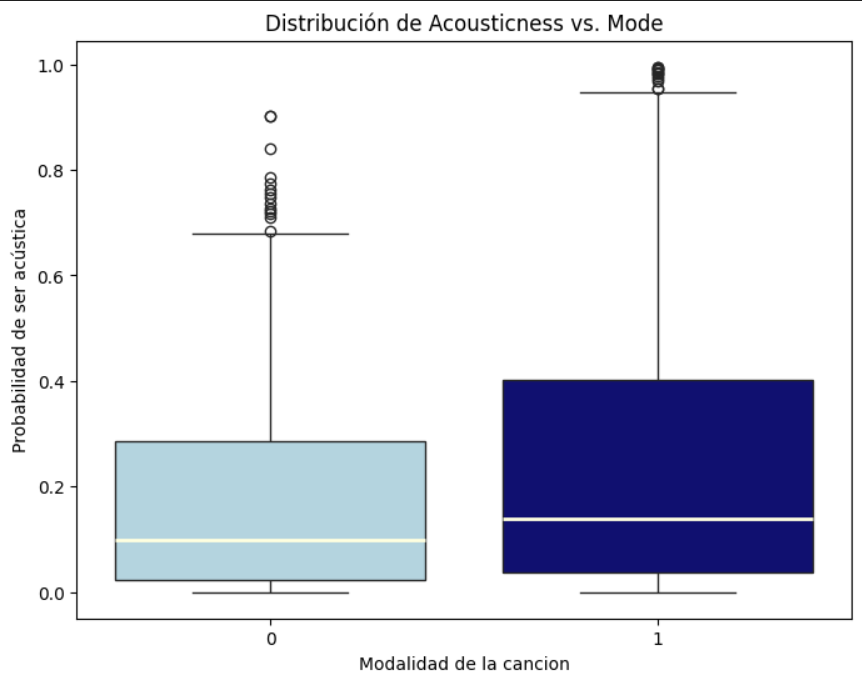
Para probar esta hipótesis, construimos un modelo de regresión lineal usando Loudness y Acousticness como predictores. Este modelo obtuvo un coeficiente de determinación (R²) de 0.58, lo cual indica un ajuste moderado a los datos. No obstante, las diferencias entre los valores observados y predichos de Energy aún son considerables en varios casos, por lo que decidimos añadir una tercera variable, Valence, al modelo.

Al incluir Valence, el modelo mostró una ligera mejora, alcanzando un R² de 0.629, lo que sugiere un mejor ajuste, aunque todavía hay discrepancias significativas entre los valores predichos y observados.

En conclusión, nuestra hipótesis se acepta parcialmente: si bien el modelo mejorado ofrece una descripción moderada de Energy en función de Loudness, Acousticness y Valence, aún deja espacio para mejorar. Sería útil explorar modelos más complejos o no lineales que puedan capturar la relación de manera más precisa y reflejar la variabilidad de Energy con mayor fidelidad.

**Hipótesis 5: Las canciones en tono mayor son más acústicas que las canciones en tono menor**

La hipótesis se basa en el análisis de la distribución de la variable "acousticness" según la modalidad de las canciones (mayor o menor), que realizamos en el análisis inicial, observamos que las canciones en tono mayor (modalidad 1) tienden a presentar valores de "acousticness" más altos que aquellas en tono menor (modalidad 0). En el gráfico, se aprecia que la mayoría de las canciones en tono menor tienen valores de "acousticness" por debajo de 0.4, lo que sugiere que estas canciones son, en promedio, menos acústicas. En contraste, las canciones en tono mayor abarcan un rango más amplio de "acousticness" y, en general, muestran una tendencia hacia valores más altos, indicando una mayor probabilidad de ser acústicas. Esta observación nos lleva a proponer que, en general, las canciones en tono mayor tienen una mayor probabilidad de ser acústicas que las canciones en tono menor.



**Hipótesis 6: El alto nivel de Acousticness se debe a bajos valores de Valence y Energy**

Esta última hipótesis surgió al analizar en detalle los valores promedio de las variables en los grupos definidos en la primera hipótesis. La media (o promedio) es el valor que representa el "centro" o "promedio" de un conjunto de datos; es decir, la media de una variable nos da una idea del nivel típico o común de esa variable en un grupo de datos.

Al observar las medias de Acousticness, Energy y Valence en estos grupos, notamos que en dos de ellos los niveles promedio de Acousticness son bajos, mientras que en un tercer grupo la media de Acousticness es alta. Al mismo tiempo, en este grupo con alta Acousticness, los valores promedio de Energy y Valence son bajos. Esto sugiere una posible relación entre un alto nivel de Acousticness (una mayor presencia de cualidades acústicas en la música) y bajos niveles de Energy (intensidad o actividad de la canción) y Valence (positividad o alegría de la canción).

# Conclusión

Este informe ha proporcionado un análisis exhaustivo de las características sonoras y su relación con la popularidad de las canciones de la década de los 90, un periodo de transformación en la industria musical. A lo largo de este estudio, se han explorado diversas hipótesis con el fin de entender los factores que influyen en el éxito de estas canciones.

Se encontró que la tonalidad predominante (Key) desempeña un papel crucial en la clasificación de las canciones, ya que, al dividir el conjunto de datos en diferentes grupos, se evidenció una diferencia significativa en la tonalidad de las canciones de cada grupo, especialmente en cuanto a la variabilidad del Key. Este descubrimiento refuerza la idea de que la tonalidad puede ser un factor determinante en la clasificación y el estilo de las canciones de esta época.

En cuanto a la popularidad, la hipótesis que sugería que está relacionada con otras características, como la energía, el volumen o la acústica, fue rechazada. No se pudo identificar una relación fuerte entre estas variables y la popularidad, lo que nos lleva a concluir que la popularidad de las canciones no puede ser atribuida únicamente a un conjunto específico de características. En su lugar, es el resultado de una combinación más compleja de factores que van más allá de las variables analizadas.

Una de las hipótesis que recibió respaldo parcial fue la de que la energía de las canciones (Energy) puede predecirse a partir de variables como el volumen (Loudness) y la acústica (Acousticness). Si bien el modelo propuesto mostró una capacidad moderada para predecir la energía, las discrepancias observadas sugieren que la relación no es completamente lineal, lo que abre la posibilidad de mejorar la predicción mediante el uso de modelos más complejos o no lineales.

Finalmente, la hipótesis que afirmaba que las canciones en tono mayor son más acústicas que las canciones en tono menor fueron respaldadas por los datos, ya que las canciones en tono mayor tendieron a mostrar valores más altos de acústica, lo que sugiere una mayor probabilidad de ser acústicas. Sin embargo, sería necesario validar esta relación con pruebas estadísticas adicionales para obtener conclusiones más firmes.

Este análisis refleja la complejidad de los factores que influyen en la popularidad de las canciones y destaca la importancia de considerar múltiples variables, así como su evolución a lo largo del tiempo. La década de los 90, marcada por su diversidad estilística y la innovación tecnológica, dio lugar a un panorama musical heterogéneo, lo que provocó que las características de las canciones populares variarán considerablemente. En futuras investigaciones, sería interesante explorar cómo la popularidad de las canciones evolucionó a lo largo de los años dentro de esta década, lo que podría proporcionar más información sobre las tendencias musicales de la época.

Bibliografía

<https://www.campuslasmusas.com/industria-musical-90/>

<https://www.noventeros.es/la-era-dorada-de-la-musica-de-los-90/>

<https://espectaculos-excelsior.com/artistas-y-canciones-mas-influyentes-de-la-musica-de-los-90/>