Informe del Trabajo Práctico Especial de Fundamentos de la Ciencia de Datos

Facultad de Ciencias Exactas de la Universidad Nacional del Centro de la Provincia de Buenos Aires

Grupo 1, integrantes:

* Dimuro, Agustín Nicolás
* Grillo, Agustín German
* Padilla, Tomás Agustín

## Índice

[Índice 2](#_Toc182513711)

[Introducción 3](#_Toc182513712)

[Materiales 4](#_Toc182513713)

[Explicación de las variables 4](#_Toc182513714)

[Preprocesamiento de los datos 5](#_Toc182513715)

[Análisis de los datos 6](#_Toc182513716)

[Métodos de agrupamiento o Clustering 8](#_Toc182513717)

[Métodos / Resultados 12](#_Toc182513718)

[Las canciones que tienen valores más altos de positividad y alegría hoy en día son las más populares. 12](#_Toc182513719)

[Las canciones compuestas con 4 pulsaciones por compas durante la década del 1970 actualmente son más populares que aquellas que fueron compuestas con un distinto número. 14](#_Toc182513720)

[Existe una relación entre las canciones que son bailables y las que son instrumentales. 15](#_Toc182513721)

[Las canciones en la primera mitad de la década tienen diferencias significativas con respecto al estilo, contra las de la segunda mitad. 18](#_Toc182513722)

[Regresión lineal 24](#_Toc182513723)

[Modelo Predictivo de la variable Energy 27](#_Toc182513724)

[Conclusiones 31](#_Toc182513725)

[Referencias 33](#_Toc182513726)

# Introducción

Durante el transcurso del siguiente informe trabajaremos sobre un conjunto de datos proveniente de Amazon Music, el cual contiene observaciones sobre canciones de la década de los años 1970 al 1979. Realizaremos sobre este conjunto el filtrado de datos los cuales puedan ser erróneos para luego poder utilizarlos con el fin de buscar respuestas a ciertas hipótesis que puedan ser útiles para obtener información acerca del comportamiento de las canciones en esa época. Si los datos nos lo permiten, trataremos de crear un modelo de regresión lineal el cual para utilizarlo como una herramienta que nos permita entender los vínculos de diferentes variables independientes contra otra variable. A su vez, queremos tratar de generar un modelo el cual nos ayude a predecir los valores de la variable elegida.

Las hipótesis que planteamos y que vamos a estar analizando durante el transcurso del siguiente informe son las siguientes.

* Las canciones que tienen valores más altos de positividad y alegría hoy en día son las más populares.
* Las canciones compuestas con 4 pulsaciones por compas durante la década del 1970 actualmente son más populares que aquellas que fueron compuestas con un distinto número.
* Existe una relación entre las canciones que son bailables y las que son instrumentales.
* Las canciones que tienen mayor grado de instrumentalidad son las que más duración tienen.
* Las canciones que son más habladas son las que tienen un menor volumen promedio.
* Las canciones que fueron grabadas durante un concierto en vivo tienen valores de energía más altos.
* Las canciones en la primera mitad de la década tienen diferencias significativas con respecto al estilo, contra las de la segunda mitad. Poniendo el foco en el análisis sobre sus valores de energía, el grado de que tan acústica es la canción, las pulsaciones por minuto y el volumen promedio de la canción en decibeles.

# Materiales

## Explicación de las variables

El conjunto de datos con el que vamos a estar trabajando durante todo el proyecto contiene 17 variables, las cuales explican diferentes características sobre cada canción de los años 1970.

1. Track: el título de la canción.
2. Artist: el intérprete o grupo que grabó la canción.
3. Duration: la duración de la canción, medida en minutos y segundos.
4. Time\_Signature: la métrica musical de la canción, indica el número de pulsaciones por compás.
5. 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.
6. Energy: una medida de intensidad y actividad en la canción, donde los valores más altos indican una pista más enérgica.
7. Key: la tonalidad musical en la que está compuesta la canción, representada por un número entero.
8. Loudness: el volumen promedio de la canción, medido en decibelios (dB).
9. Mode: la modalidad de la pista, indica si la canción está en tono mayor o menor.
10. Speechiness: una medida de la presencia de palabras habladas en una pista, valores más altos indican cualidades más parecidas al habla.
11. Acousticness: una medida de la calidad acústica de la pista, valores más altos indican una mayor probabilidad de ser acústica.
12. Instrumentalness: una medida que indica la presencia de voces, valores más altos representan pistas más instrumentales.
13. Liveness: una medida de la probabilidad de que la pista se haya interpretado en vivo, valores más altos indican más ruido de audiencia.
14. Valence: Una medida de la positividad musical de la pista, valores más altos indican música más positiva o alegre.
15. Tempo: la velocidad o ritmo de la pista, medida en pulsaciones por minuto (BPM).
16. Popularity: una puntuación que refleja la popularidad de la pista, generalmente basada en los recuentos de transmisiones y otras métricas.
17. Year: el año en que se lanzó la canción.

## Preprocesamiento de los datos

Al momento de realizar la exploración de los datos que estaban dentro del dataset encontramos que estaban en una muy buena condición, es decir, que poseían una alta confiabilidad en lo que pretenden representar de la población. Esto se debe a que no logramos identificar valores erróneos o incorrectos ni señales que indiquen que puedan ser falsos. Tampoco observamos posibles outliers ni valores fuera de rango.

A pesar de lo mencionado, nos encontramos con la columna “Popularity” la cual, en una primera impresión nos daba a entender que representaba la popularidad que había tenido la canción durante la década. Pero al tratar de contrastar ese significado con lo observado en los datos nos dimos cuenta de que era imposible que la variable mencionada represente la popularidad en la década de 1970 ya que, canciones que están situadas en los primeros puestos en los rankings de popularidad de canciones de esa década tenían valores mucho menores que canciones las cuales fueron mucho menos populares. Por esto decidimos buscar una nueva explicación para los datos almacenados en esta variable y nos encontramos en Kaggle con un dataset de Spotify el cual poseía una columna “Popularity” la cual los valores calculados en ella se mueven entre 0 y 100, al igual que en dataset nuestro, y que los valores generados para cada canción son calculados gracias a un algoritmo el cual toma en cuenta la cantidad total de reproducciones de la canción y que tan recientes son esas reproducciones. Dado que no encontramos cómo funciona el algoritmo que calcula la popularidad de las canciones en Amazon Music, asumimos que usa uno similar o igual a Spotify, por lo que la variable “Popularity” la analizaremos como la popularidad actual de las canciones.

Para facilitar el posterior análisis de los datos decidimos modificar los valores de algunas variables. Una de estas es la columna “Duration” la cual poseía los valores de la duración de la canción en minutos y segundos, pero decidimos modificar a que solo sea almacenado en segundos dado que puede generar problemas a la hora de interpretar los resultados el que esta columna posea dos unidades.

Otra de las columnas que modificamos fue “Loudness” ya que, luego de buscar en distintos sitios web especializados en música, información sobre cómo se calcula el volumen promedio de la canción en decibelios, siempre era mostrado como un valor entero y no con múltiples valores decimales. Por lo tanto, decidimos quitarle esos valores decimales a la columna y dejarlos como valores enteros.

De igual manera que con la columna “Loudness”, la columna “Tempo”, que mide la velocidad o el ritmo de la pista medido en pulsaciones por minuto, posee valores que están representados con números que poseen múltiples decimales. Al comparar los datos con los sitios web especializados en música, los valores que estaban almacenados en el dataset eran correctos, solo que no poseían la parte decimal. Por lo tanto, decidimos modificar esta columna para que solo almacene números enteros.

Por último, nos percatamos que la columna “Instrumentalness” tiene muchos ceros, por lo que pensamos que podía representar un valor nulo en vez de un valor posible en los datos. Investigamos, pero no encontramos una explicación de cómo fue calculado ese valor, por lo tanto, decidimos seleccionar canciones al azar que posean el valor de la columna mencionada en cero y nos pusimos a escucharlas, para luego compararlas con canciones que tengan valores distintos a cero y que sean significativamente distintos a cero. Con nuestra poca experiencia en análisis musical notamos que las canciones con el valor de la columna en cero tenían una gran cantidad de presencia de voces, mientras que las que tenían valores distintos a cero se podía notar una mayor presencia de los instrumentos. Por este motivo decidimos confiar en los datos proporcionados por la columna dado que no encontramos una forma convincente y robusta de refutarlos.

Como paso final, realizamos un Profile Report el cual no nos fue de gran utilidad a la hora de analizar los datos y buscar posibles hipótesis o soluciones. Por ende, decidimos no utilizarlo para la confección de este informe y lo eliminamos de la Jupyter Notebook.

## Análisis de los datos

Luego de realizar, sobre el conjunto de datos, todas las transformaciones mencionadas previamente, pasamos a investigarlos en búsqueda de posibles variables las cuales representes información que pueda llegar a ser relevante. Gracias a esto nos interesamos por la variable “Popularity” ya que la podemos usar como un parámetro de que tan bien envejecieron las canciones. Luego se nos ocurrió que las canciones más alegres y positivas podrían ser las que mejor envejecieron dado que habrían dejado buenos recuerdos en las memorias de las personas que las escuchaban en la década de 1970, por lo que podría ser más posible que no solo las escuchen ellos hoy en día, sino que también se las hayan mostrado a personas más jóvenes las cuales no tuvieron la posibilidad de escucharlas en su lanzamiento.

A su vez, estando también relacionado con la popularidad, se nos generó la duda de si las canciones que tienen cuatro pulsaciones por compás hoy en día son más escuchadas que las que no. Esta duda está basada en que los géneros musicales más escuchados en la actualidad, como el reggaetón o el pop, también comparten la característica de tener cuatro pulsaciones por compás. También notamos que esta métrica es la más utilizada con gran diferencia a lo largo de toda la década, por lo que si las canciones que no poseen cuatro pulsaciones por compás son las más populares hoy en día sería una buena forma para luego poder predecir valores de popularidad de canciones que no están en el conjunto de datos.

Otra variable la cual nos pareció interesante de analizar es “Danceability” ya que representa que tan adecuada es una canción para ser bailada. Gracias a que también está disponible en el conjunto de datos una variable que representa que tan instrumental es una pista, pensamos que puede llegar a existir una relación entre ambas, es decir, que las canciones más bailables son las más instrumentales.

Como ya estábamos trabajando con la variable “Instrumentalness”, también nos surgió la duda de si tienen una relación con la duración de las canciones. Esta duda surge de que en la década de 1970 se empezó a experimentar con distintos sonidos y formas de hacer música, por lo que había muchas canciones las cuales poseen solos, como por ejemplo de guitarra, lo que aumentaría la duración de la canción con respecto a las que no los poseen.

Al analizar la variable “Speechiness”, que es una medida de la presencia de palabras habladas en las canciones, pensamos que podía tener una relación con el volumen promedio de las canciones. Esto se nos ocurrió ya que canciones las cuales son mucho más habladas que cantadas suelen ser mucho más tranquilas y acompañadas de instrumentos suaves o incluso minimalistas, generando un ambiente más calmo.

Adicionalmente pensamos que la variable “Energy”, que es una medida de la intensidad y actividad de la canción, tiene potencial para distintos análisis. Una idea que se nos ocurrió fue que tenga una relación con “Liveness”. Esto puede estar dado por que en los recitales o conciertos en vivo se suelen tocar las canciones más movidas para que el público salte y baile. Otro análisis que se nos ocurrió para realizar con esta variable es que, gracias a su alta correlación con otras variables del conjunto de datos, podríamos tratar de plantear una regresión lineal con el fin de poder entender si esas variables influyen en “Energy”. A su vez, podríamos realizar un modelo que trate de predecir los valores de esta variable al ingresar nuevas canciones que no estén incluidas en el conjunto de datos.

Posteriormente, tuvimos la idea de que podría existir una diferencia significativa con respecto a los valores de intensidad y actividad de una canción, el grado de que tan acústicas son, las pulsaciones por minuto y el volumen promedio de las canciones en decibeles al analizarlo entre lustros. Es decir, queremos corroborar que hay una diferencia en el estilo musical de las canciones que fueron lanzadas entre 1970-1974 y las canciones que fueron lanzadas en 1975-1979.

## Métodos de agrupamiento o Clustering

Previo a empezar a plantear hipótesis quisimos realizar Clustering con el fin de encontrar grupos de datos en nuestro conjunto de datos los cuales nos puedan guiar hacia las posibles hipótesis. Para ello, decidimos aplicar primero un algoritmo basado en particiones llamado K-Means.

Antes de correr el algoritmo K-Means separamos los datos cuantitativos del resto de columnas que poseían otros tipos de datos, como por ejemplo el nombre de la canción o en que año fue lanzada. Luego decidimos escalar los datos con la finalidad de que estos tengan la misma varianza y el mismo desvío, con el fin de que no haya datos que pesen más que otros.

Una vez aplicado el escalamiento, corrimos el algoritmo con una cantidad de clusters aleatoria para después generar el Elbow Plot o gráfico del Codo con el fin de poder elegir el mejor número de clusters.

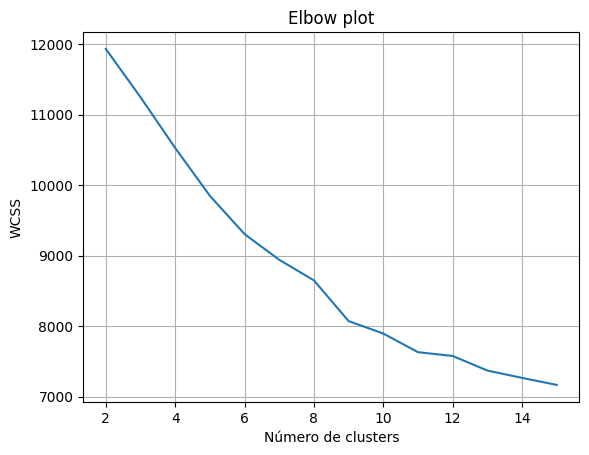


Figura 8: gráfico Elbow Plot luego de realizar K-Means. Muestra la relación entre la cantidad de clusters y la suma de distancias al cuadrado entre todas las observaciones

Gracias al análisis de este gráfico concluimos que puede haber múltiples valores que pueden ser ideales para la cantidad de clusters, pero nos quedaremos con ocho por el momento.

Para poder visualizar mejor si se formaron grupos gracias a los clusters decidimos reducir la dimensionalidad de los datos mediante el uso de PCA. Para poder realizar este proceso volvimos a escalar los datos cuantitativos y lo proyectamos en dos variables. Pero al analizar la varianza explicada por estas dos variables nos dimos cuenta que entre ambas no llegan al 30% del total de la varianza, por lo tanto, descartamos el uso de este procedimiento.

Como todavía queríamos ver mediante un gráfico si se habían generado o no grupos, decidimos aplicar tanto t-SNE como UMAP para reducir la dimensionalidad de nuestro conjunto de datos. Luego de realizar los pasos necesarios para la ejecución de ambos algoritmos, decidimos pintar las muestras generadas en los scatterplots dependiendo de a que cluster generado por K-Means pertenecían.

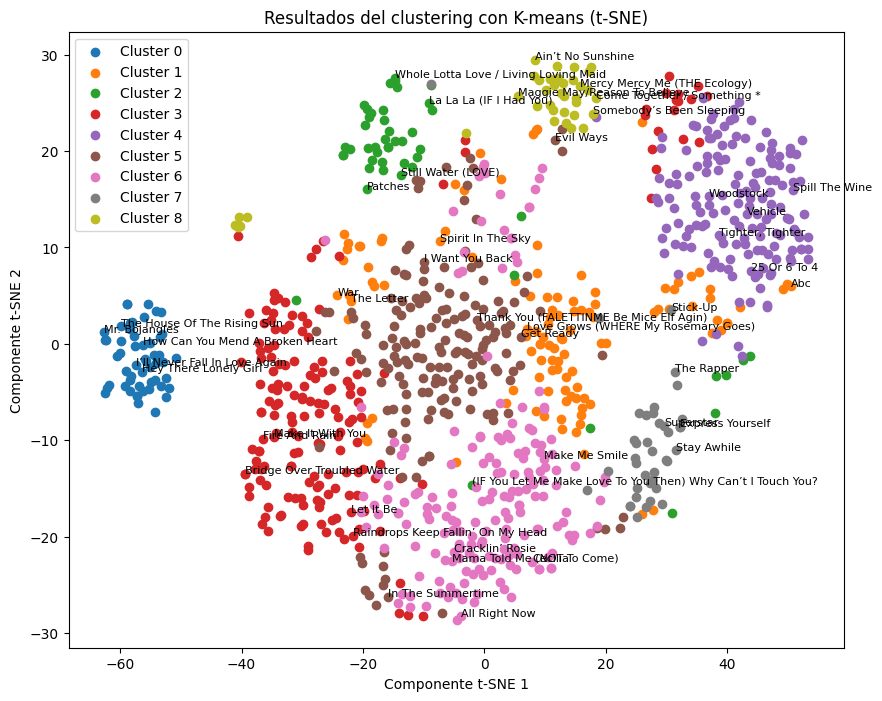


Figura 9: Scatterplot generado luego de aplicar reducción de dimensionalidad con t-SNE. Los colores representan a que cluster, generado con el algoritmo K-Means, pertenecen.

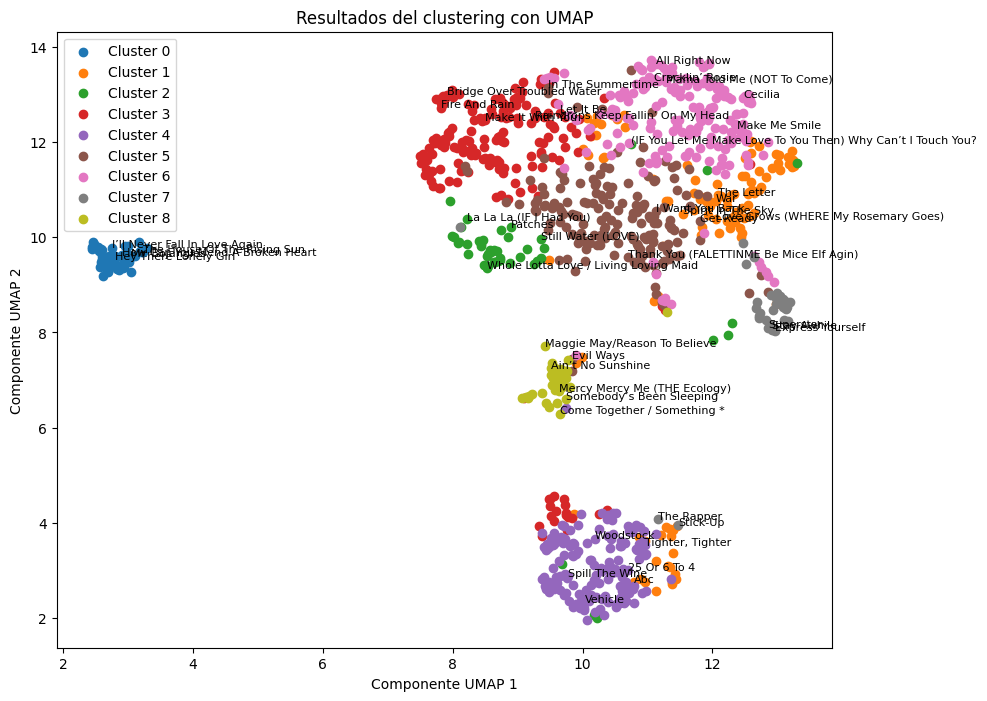


Figura 10: Scatterplot generado luego de aplicar reducción de dimensionalidad con UMAP. Los colores representan a que cluster, generado con el algoritmo K-Means, pertenecen.

Posterior al análisis de ambos gráficos, pudimos observar que había algunos clusters en lo que podría haber comportamiento de grupos, pero que a nivel general no se puede ver grupos claros y distinguidos. Para tratar de solucionar este problema tratamos de utilizar múltiples cantidades de clusters pero no difería en gran medida de lo observado en los gráficos presentados previamente, Por este motivo descartamos el uso de clustering basado en particionamiento.

Sin embargo, todavía queríamos verificar que no existan comportamiento de grupos por lo que aplicamos un algoritmo de clustering jerárquico aglomerativo. Para esto corrimos el algoritmo con los mismos datos con los que hicimos K-Means, utilizando las distancias euclídeas, y obtuvimos el siguiente dendograma.

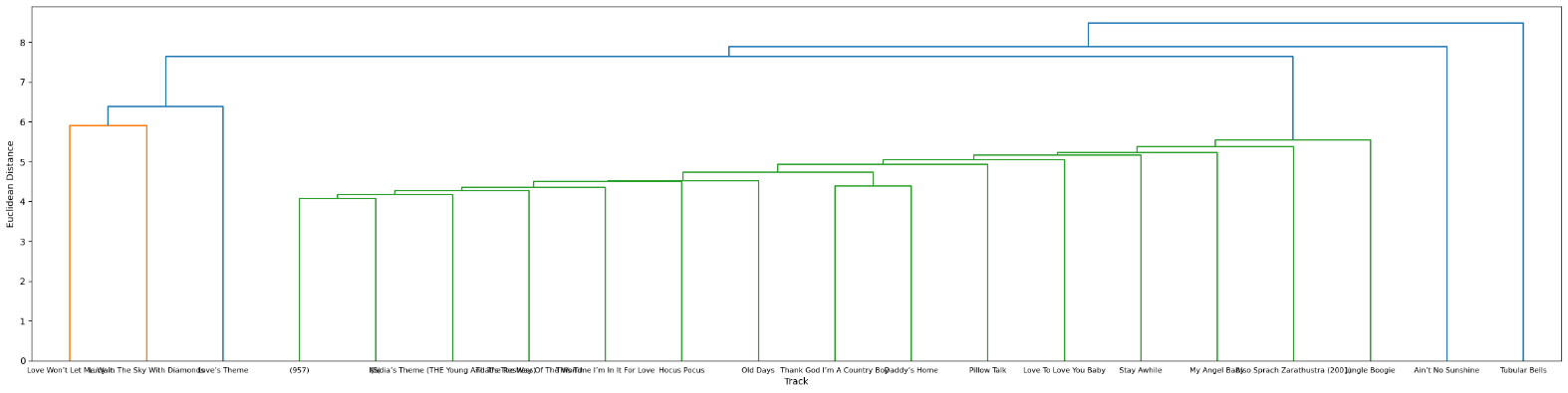


Figura 11: Dendograma que muestra el resultado de realizar un algoritmo de clustering jerárquico aglomerativo.

Al analizar este dendrograma, y luego de cortarlo en múltiples valores concluimos que no podemos observar comportamientos de grupos en este conjunto de datos. A esta conclusión llegamos gracias a que, mirando el gráfico, se nota como se agregan, en la mayoría de los casos, de a una canción a un cluster que posee una gran cantidad de canciones. Por este motivo decidimos que no utilizaríamos clustering para analizar los datos ni para plantear posibles hipótesis.

# Métodos / Resultados

### Las canciones que tienen valores más altos de positividad y alegría hoy en día son las más populares.

Para poder analizar el conjunto de datos decidimos separarlos en dos. Filtramos los datos de manera tal que podamos separar las canciones que son más alegres y positivas por sobre las que son menos, tomando como el corte las que tengan un valor mayor a “0.5” en la columna “Valence”. Con esta división logramos tener dos subconjuntos de datos con los cuales trataremos de compararlos para poder verificar si las canciones que son más alegres y positivas hoy en día son más populares que las que son menos.

Como primer paso para empezar a analizar los datos decidimos crear un gráfico de boxplot para cada subconjunto y ver cómo se distribuye la popularidad en cada uno.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figura 1: muestra la distribución de la variable “Popularity” para las canciones menos alegres y positivas (a izquierda) y las que son más (a derecha)

Cuando observamos el gráfico podemos concluir que la distribución de popularidad es muy similar en ambos subconjuntos de datos. Pero solo con el boxplot no podemos proveer una respuesta robusta para saber si la hipótesis que fue planteada es verdadera o no. Para poder asegurarnos estadísticamente de ello trataremos de realizar un test paramétrico el cual nos ofrezca una respuesta.

Para poder realizar un t-test necesitamos que los datos de ambos subconjuntos cumplan con los supuestos obligatorios para que el resultado ofrecido por el test sea confiable. Uno de estos supuestos es que los datos se distribuyen de manera normal, por lo tanto, debemos utilizar algún método que nos ayude a comprobarlo, como por ejemplo Kolmogórov-Smirnov o QQ-plot (solo gráfico). Nosotros decidimos utilizar el test de Shapiro Wilks, el cuál compara la distribución de los datos de nuestro subconjunto en búsqueda de saber si se distribuyen de manera normal. Este test ofrece un p-valor el cual si es mayor a 0,05 indica que se acepta la hipótesis nula, la cual propone que la distribución es normal. Al realizar este test sobre cada uno de nuestros subconjuntos de datos, el algoritmo nos devolvió como resultado un p-valor para ambos casos de 0,000, es decir, que nuestros datos no se distribuyen de manera normal. Por ende, descartamos la idea de trabajar con test paramétricos y vamos a buscar cumplir supuestos de algún test no paramétrico.

Luego de descartar la posibilidad de trabajar con test paramétricos, decidimos tratar de verificar los supuestos de Mann-Whitney o U-test. Este supone que los datos de ambos subconjuntos son homocedásticos, es decir, que tienen varianzas iguales. Para poder comprobar esto tenemos múltiples test que nos dan una mano para verificar este supuesto. El que elegimos para utilizar es el test de Levene. La hipótesis nula que propone es que los datos son homocedásticos, es decir, que si nuestro p-valor es mayor a 0,05 nuestros datos cumplirían con el supuesto de igualdad de varianzas. Al realizar esta prueba sobre nuestros subconjuntos nos arrojó un p-valor de 0,093, por lo tanto, nuestros datos cumplen el supuesto de homocedasticidad.

Gracias a que se cumplen todos los supuestos estadísticos necesarios para realizar el U-test, decidimos utilizarlo. Para la hipótesis nula decidimos plantear que las canciones más positivas y alegres no son más populares actualmente, por lo que configuramos el test de tal manera que compare los subconjuntos de datos en búsqueda de que el subconjunto asociado a las canciones más alegres y populares tengan una popularidad actual mayor. Esto implica que si el test de Mann-Whitney nos otorga un p-valor mayor a 0,05, no hay razones para afirmar que las canciones más alegres y positivas son más populares hoy en día. Luego de realizar la prueba, nos arrojó un p-valor de 0,963, por lo tanto, concluimos que con nuestros datos no podemos afirmar que las canciones más alegres y positivas tienen una popularidad mayor actualmente.

### Las canciones compuestas con 4 pulsaciones por compas durante la década del 1970 actualmente son más populares que aquellas que fueron compuestas con un distinto número.

Como los estilos musicales actuales más populares generalmente tienen cuatro pulsaciones por compás, decidimos analizar si la popularidad de las canciones de la década de 1970 que están en el conjunto de datos es mayor que las canciones que poseen un número distinto de pulsaciones por compás.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente Para poder realizar esta comparación decidimos dividir el conjunto de datos en dos subconjuntos menores los cuales se dividen según el criterio de que posean, o no, cuatro pulsaciones por compás. Esto nos generó dos subconjuntos los cuales utilizaremos para lograr darle una respuesta a nuestra hipótesis. Para empezar con el análisis decidimos plantear para cada subconjunto un boxplot el cual muestre la distribución de la popularidad en cada uno de ellos.

Figura 2: muestra la distribución de la variable “Popularity” para las canciones con cuatro pulsaciones por compás (a derecha) y las que poseen un número distinto (a izquierda)

Al observar el gráfico vimos que no parece haber una gran diferencia entre las distribuciones de popularidad en ambos subconjuntos. Pero solo con el método gráfico presentado anteriormente no podemos llegar a ninguna conclusión sobre la validación de los datos. Por lo tanto, decidimos tratar de realizar un t-test.

Para poder realizar un test paramétrico debemos cumplir los supuestos de normalidad y de homocedasticidad en nuestros subconjuntos de datos. Buscamos obtener el p-valor para poder verificar el supuesto de normalidad utilizando como herramienta el test de Shapiro Wliks explicado anteriormente. Luego de aplicar el algoritmo en cada subconjunto, nos otorgó como resultado que ninguna de las distribuciones de ambos subconjuntos de datos es normal, por lo que descartamos el uso de algún test paramétrico.

Por consecuencia, nos orientamos a utilizar un test no paramétrico y elegimos el U-test de Mann-Whitney. Para poder realizarlo debemos comprobar el supuesto de homocedasticidad, el cual propone la igualdad de varianzas entre ambos subconjuntos. Para poder afirmar que nuestros datos cumplen con el supuesto realizamos el test de Levene, el cual nos arrojó un p-valor de 0,319. Por lo tanto, nuestros subconjuntos de datos son homocedásticos y podemos realizar un U-test.

Como nuestra hipótesis propone que las canciones con cuatro pulsaciones por compás son más populares actualmente que las canciones que poseen un número distinto, decidimos plantear como hipótesis nula que no hay razones para afirmar que las canciones con un cuatro en la columna “Time\_Signature” son más populares que las demás. Para ello configuramos el Mann-Whitney para que compare si los valores del subconjunto que tienen las canciones con 4 pulsaciones por compás tienen más popularidad que las que poseen un número distinto, y obtuvimos un resultado, gracias al análisis del p-valor, que no hay razones para afirmar que las canciones de la década de 1970 tienen más popularidad por tener cuatro pulsaciones por compás.

### Existe una relación entre las canciones que son bailables y las que son instrumentales.

Mientras que analizábamos los datos se nos ocurrió la idea de verificar si podía llegar a existir una relación entre las variables “Instrumentalness” y “Danceability” ya que nosotros asociamos que, en la mayoría de los casos, las canciones que poseen mayor presencia de instrumentos suelen ser las que son más bailables. Para poder comprobar esta hipótesis decidimos dividir nuestro conjunto de datos en cuatro subconjuntos. Estos surgen de separar los datos según las canciones que son más bailables, que las definimos como las que posean un valor mayor a 0,5 en la columna “Danceability”, y las canciones instrumentales, las cuales las definimos como cualquier valor mayor a cero en la columna “Instrumentalness”. Gracias a realizar este procedimiento tendríamos por separado las canciones bailables e instrumentales, las bailables y no instrumentales, las instrumentales y no bailables y las que no son ni bailables ni instrumentales.

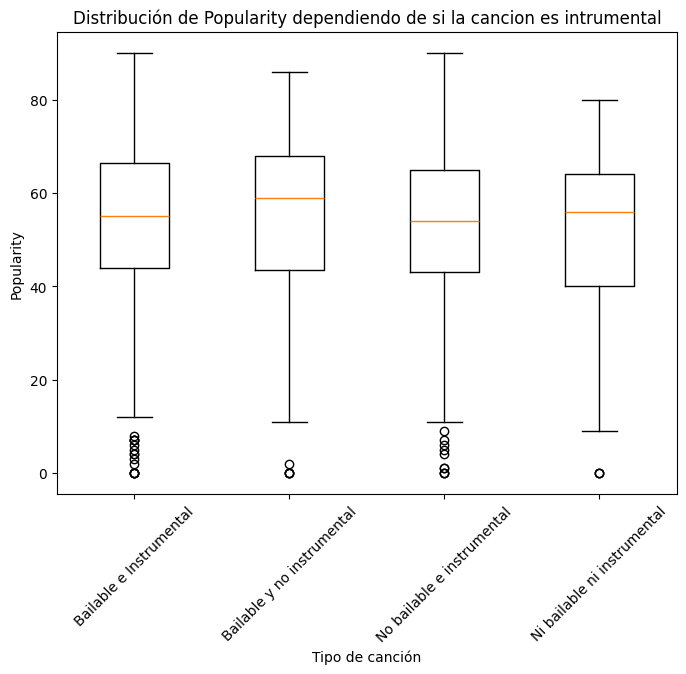
 Una vez que terminamos de particionar los datos para el análisis quisimos visualizar mediante un boxplot para cada uno de los subconjuntos de datos, la distribución de la popularidad ya que era una variable que veníamos utilizando previamente para los demás análisis. La idea de realizar estos boxplots era poder visualizar claramente que haya cuatro subconjuntos de datos, pero no tratar sacar una conclusión respecto a la hipótesis planteada.

Figura 3: muestra la distribución de la variable “Popularity” para las canciones pertenecientes a los distintos grupos dependiendo de que tan bailable e instrumental sea la canción

Luego de generado el gráfico, decidimos plantear un test paramétrico para poder obtener un resultado para nuestra hipótesis. Decidimos que queríamos utilizar un test ANOVA para que nos ofrezca una respuesta sobre la hipótesis. Por lo tanto, el paso siguiente es comprobar los supuestos para poder realizar el test mencionado, que serían que los datos tienen distribución normal, que son homocedásticos, que los datos son lineales y que son independientes. Al realizar el test de Shapiro Wilks sobre nuestros subconjuntos de datos, nos dio un p-valor menor a 0,05 en todos, por lo que se nos cayó el supuesto de normalidad.

Para poder buscar una respuesta a nuestra hipótesis, decidimos realizar un test Chi-Cuadrado en búsqueda de conocer si nuestras variables eran o no independientes. Con este objetivo, creamos la tabla de contingencia, agregándole los datos de nuestros subconjuntos, que luego utilizara el test para ofrecernos un p-valor. Luego de creada, podría ser vista de la siguiente manera:

|  |  |  |
| --- | --- | --- |
| Canción | Instrumental | No instrumental |
| Bailable | 495 | 220 |
| No bailable | 200 | 65 |

Posterior a la creación de la tabla de contingencia decidimos que nuestra hipótesis nula en este caso es que las variables son independientes, por lo tanto, al pedirle al test Chi-Cuadrado el p-valor, si este es menor a 0,05 no diría que existe una relación entre canciones instrumentales y bailables. Para conseguir el estadístico, la prueba genera una tabla de frecuencias esperadas la cual se ve de la siguiente forma:

|  |  |  |
| --- | --- | --- |
| Canción | Instrumental | No instrumental |
| Bailable | 507,066 | 207,933 |
| No bailable | 187,933 | 77,066 |

Por último, le solicitamos al test que nos otorgue el p-valor, el cual es 0,067. Al ser mayor a 0,05, cae en la región de aceptación de la hipótesis nula. Por lo tanto, concluimos que con los datos disponibles no podemos afirmar que hay una relación entre las canciones que son bailables e instrumentales.

### Las canciones en la primera mitad de la década tienen diferencias significativas con respecto al estilo, contra las de la segunda mitad.

Luego de realizar múltiples análisis sobre los datos pertenecientes al conjunto de datos, se nos ocurrió que podría haber una diferencia entre las canciones fueron lanzadas en el primer lustro de la década de 1970. Cómo esta década tuvo muchas innovaciones musicales, decidimos seguir adelante con la hipótesis, pero nos encontramos con el impedimento de sobre qué variables podíamos realizar el análisis.

Para solucionar el problema planteado anteriormente, observamos las variables disponibles en el conjunto de datos y se nos ocurrió que podíamos detectar diferencias en estilos marcados en ambos lustros mediante el análisis de la intensidad y actividad de las canciones, de la velocidad o ritmo de la pista y de el volumen promedio de la canción. Estos datos los obtendremos de las variables “Energy”, “Tempo” y “Loudness”. Adicionalmente, en sintonía con lo mencionado sobre las innovaciones musicales en la década de 1970, se empezaron a utilizar más sonidos artificiales, por lo tanto, creímos que debíamos agregar al análisis la variable “Acousticness”, la cual es una medida de la calidad acústica de la música.

Separamos nuestro conjunto de datos entre las canciones pertenecientes a la primera mitad de década y al segundo lustro. Con el fin de poder analizar mejor las diferencias entre lustros, decidimos realizar por separado el análisis de cada variable.

#### ¿Existen diferencias en la intensidad y actividad de las canciones entre los lustros?

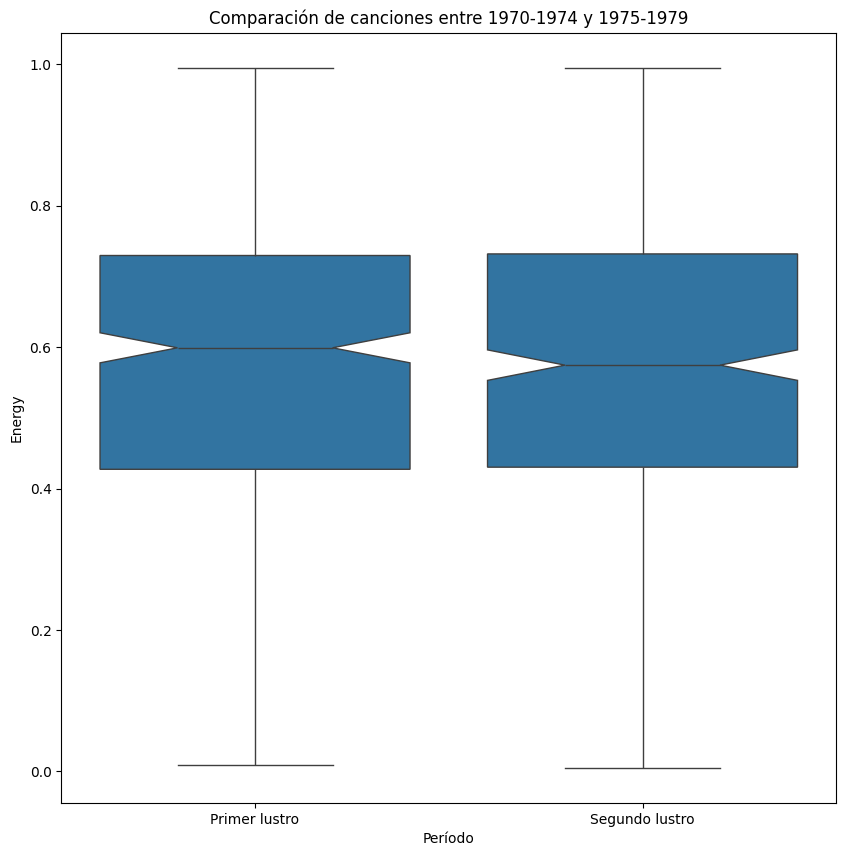
Como primer paso para poder responder a la hipótesis decidimos agarrar nuestro conjunto de datos, separados por lustro, y dejar solamente en cada subconjunto la columna asociada a la energía de las canciones de cada canción perteneciente a ese lapso de tiempo. Para poder tener un primer contacto con estos datos, decidimos crear un boxplot el cual muestra la distribución de la variable “Energy” en cada mitad de década.

Figura 4: muestra la distribución de la variable “Energy” para las canciones pertenecientes la primera mitad de década (a izquierda) y a la segunda mitad (a derecha)

Posterior a la creación del gráfico, nos pusimos a analizar y a simple vista no podemos observar una diferencia entre ambas mitades de década. Sin embargo, para poder proveer una respuesta de forma efectiva, tomamos la opción de realizar un test paramétrico.

Previo a la realización del t-test, elegido para buscar una resolución a la hipótesis, debemos verificar que se cumplan los supuestos necesarios para poder utilizar este test. Por lo tanto, debemos analizar distribución normal en los datos de cada subconjunto e igualdad de varianzas u homocedasticidad. Decidimos probar primero el supuesto de normalidad mediante el uso del test de Shapiro Wilks, el cual al utilizarlo en nuestros subconjuntos de datos nos arrojo p-valores menores a 0,05 para ambos, por lo que la distribución de nuestros datos no es normal. Por ende, descartamos el uso de test paramétricos para resolver esta hipótesis.

Como consecuencia de que no cumplimos el supuesto de normalidad, decidimos inclinarnos hacia el uso del U-test de Mann-Whitney como herramienta para poder verificar la hipótesis. Para ello todavía debemos cumplir con el supuesto de homocedasticidad. Por lo tanto, decidimos aplicar un test de Levene sobre nuestros datos, el cual arrojo un p-valor de 0,972, confirmando la igualdad de varianzas de nuestros datos.

Gracias a que cumplimos con el supuesto de homocedasticidad, decidimos aplicar el U-test de Mann-Whitney. Para ello planteamos como hipótesis nula que intensidad y actividad de las canciones es idéntica en ambos lustros, es decir, que no existen diferencias notorias en la columna “Energy” para justificar una diferencia entre ambas mitades de década. Luego de realizado el test, nos devolvió un p-valor igual a 0,793, por lo tanto, concluimos que no hay razones para decir que entre la primera mitad de década y la segunda hay diferencias con respecto a la variable “Energy”.

#### ¿Existen diferencias en la calidad acústica de las canciones entre los lustros?

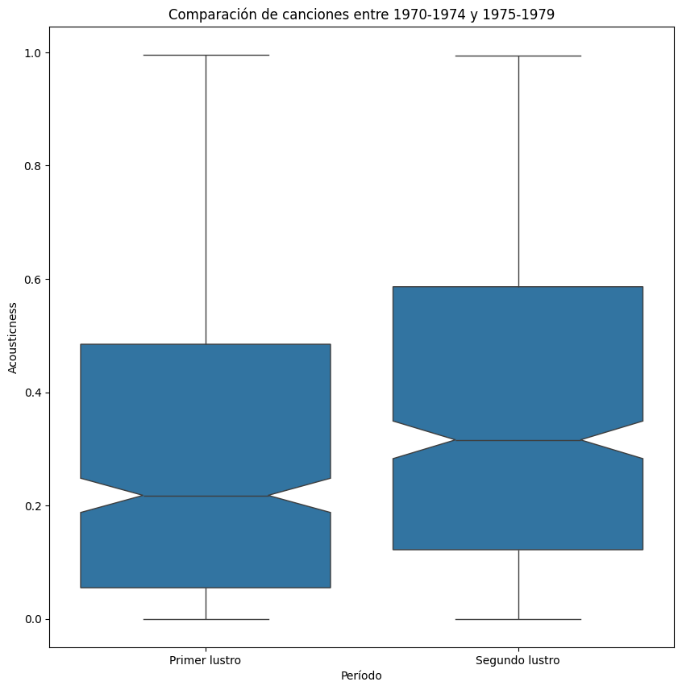
 En primer lugar, para poder iniciar el análisis en búsqueda de un resultado para la hipótesis planteada, decidimos generar dos subconjuntos de datos en los cuales solo este la variable “Acousticness” asociada a la mitad de década correspondiente. Luego quisimos observar cómo se distribuía la variable mencionada en cada lustro, así que decidimos crear un boxplot.

Figura 5: muestra la distribución de la variable “Acousticness” para las canciones pertenecientes la primera mitad de década (a izquierda) y a la segunda mitad (a derecha)

#### Cuando nos pusimos a analizar el gráfico, vimos que las canciones del segundo lustro tenían una media un poco mayor a las correspondientes con las de la primera mitad de década. Pero esto no nos pareció suficientemente distinto como para poder pensar que existiría una diferencia entre ambas mitades de década. Para poder llegar a una conclusión más sólida decidimos aplicar un test paramétrico.

Posteriormente decimos aplicar un t-test en búsqueda de una respuesta para la hipótesis. Pero previo a la aplicación de esta prueba debíamos cumplir los supuestos necesarios para poder realizarla. Al momento de verificar que ambos subconjuntos cumplían con el supuesto de normalidad, nos inclinamos por utilizar como herramienta el test de Shapiro Wilks, el cual nos proveyó de un p-valor menor a 0,05 para ambos subconjuntos. Por ende, descartamos el uso del t-test dado que no cumplimos con el supuesto de normalidad.

Como consecuencia, decidimos cambiar el t-test de Student por el U-test de Mann-Whitney. Pero para poder realizarlo todavía debemos verificar que se cumpla el supuesto de homocedasticidad. Para ello utilizamos el test de Levene, el cual nos otorgó un p-valor de 0,435. Por lo tanto, nuestros datos poseen igualdad de varianzas.

Por último, decidimos plantear nuestra hipótesis nula como que no hay motivos para encontrar diferencias en la columna “Acousticness” entre ambos lustros. Es decir, si nuestro estadístico cayese en la región de aceptación de la hipótesis nula, confirmaría lo que supusimos luego de ver los gráficos boxplot de que no hay una diferencia significativa entre ambas mitades de década. Para poder corroborarlo realizamos el U-test que nos otorgó un p-valor igual a 0,000. Por lo tanto y para nuestra sorpresa, se rechaza la hipótesis nula por lo que hay razones para afirmar que las medidas acústicas cambiaron a lo largo de la década.

#### ¿Existen diferencias en la velocidad o ritmo de las canciones entre los lustros?

De igual manera que para las anteriores dos variables, separamos en dos subconjuntos de datos la variable “Tempo” dependiendo de a que lustro corresponde. Luego creamos un gráfico boxplot que nos muestre la distribución de la variable dependiendo de a que mitad de década pertenece.

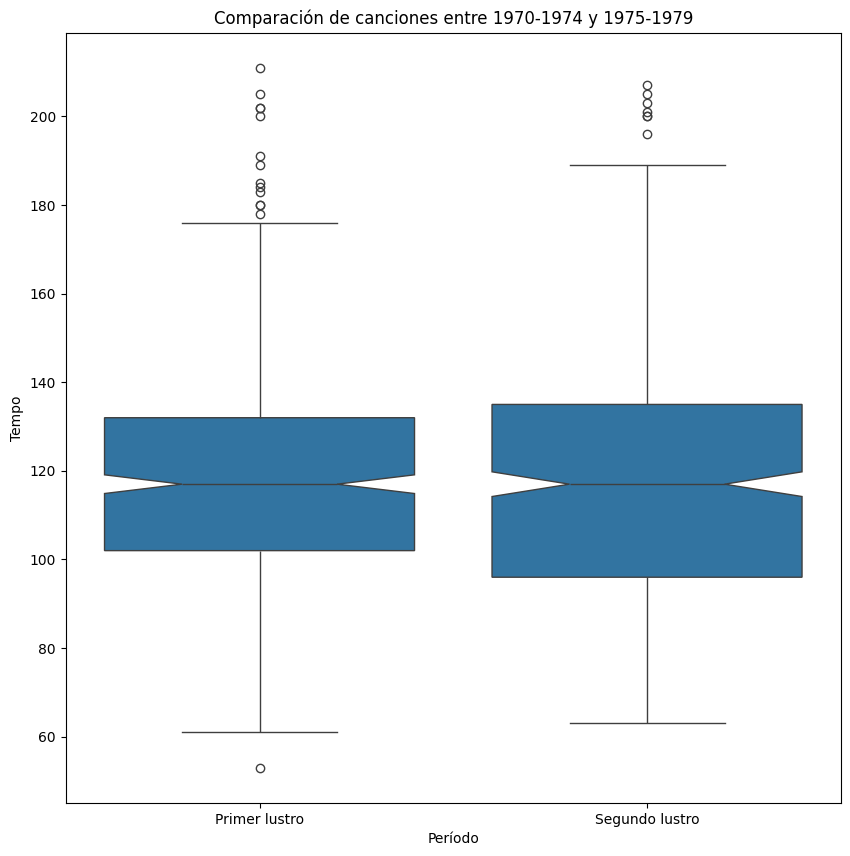


Figura 6: muestra la distribución de la variable “Acousticness” para las canciones pertenecientes la primera mitad de década (a izquierda) y a la segunda mitad (a derecha)

Una vez realizado el análisis del gráfico boxplot, supusimos que, a pesar de que los datos en el boxplot asociado al primer lustro están más concentrados hacia el centro, las medias son muy similares por lo que probablemente no observemos diferencias entre ambos lustros. Pero para poder proveer una respuesta mucho mas sólida y confiable a la hipótesis, decidimos hacer un test paramétrico.

Por ende, para la realización de un t-test, debemos comprobar que nuestros subconjuntos de datos cumplan los supuestos necesarios para poder utilizar esa prueba. Para ello realizamos el test de Shapiro Wilks con la finalidad de comprobar la normalidad de nuestros datos. Al observar el p-valor arrojado por el test, como era menor a 0,05 concluimos que no se cumple el supuesto de normalidad de los datos, por lo tanto, descartamos el uso de test paramétricos.

Consecuentemente, decidimos cambiar el t-test por el U-test de Mann-Whitney. Pero para poder utilizarlo todavía debemos verificar que se cumpla es supuesto de igualdad de las varianzas. Para ello utilizaremos el test de Levene, el cual nos arrojó un p-valor igual a 0,001. Por lo tanto, también descartamos el uso del U-test dado que los datos no cumplen con el supuesto de homocedasticidad.

Como último recurso y dado que no requiere cumplir ningún supuesto, decidimos utilizar el test de Kruskal-Wallis el cual nos permitirá comparar los valores de la columna “Tempo” en cada lustro. Para ello planteamos como hipótesis nula que las canciones de la década del 1970 no tienen grandes diferencias con respecto a la variable analizada. Luego de realizado el test, nos arrojó un p-valor igual a 0,841, el cual cae en la región de aceptación de la hipótesis nula. Por lo tanto, concluimos que no hay razones para afirmar que la velocidad o el ritmo de las canciones cambio durante la década.

#### ¿Existen diferencias en el volumen promedio de las canciones entre los lustros?

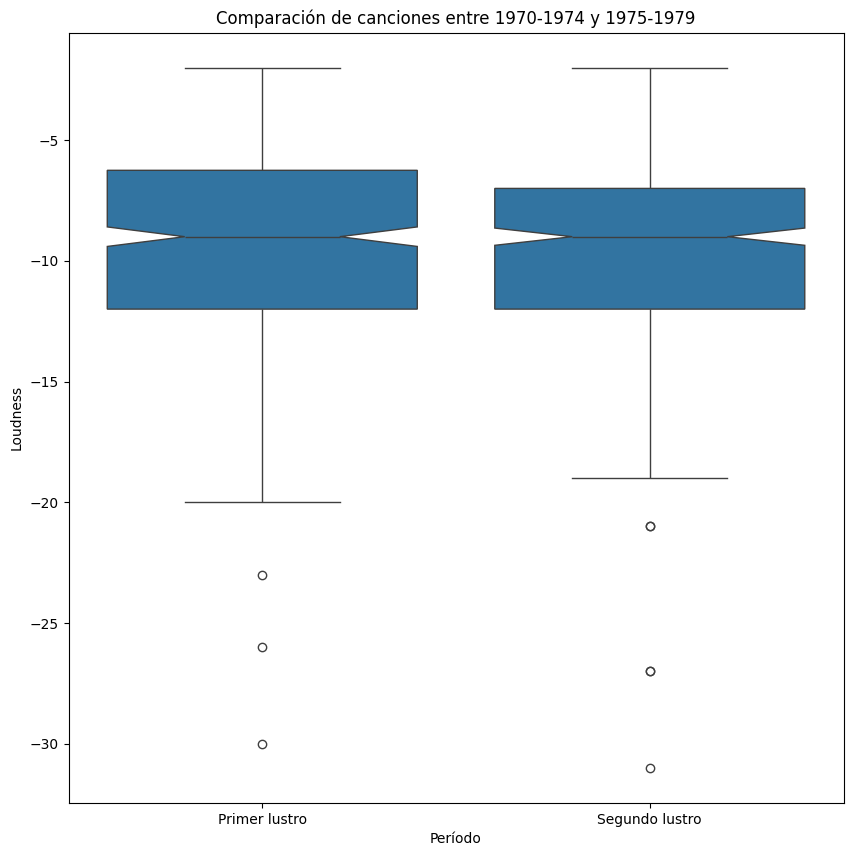
 De igual forma que con las tres hipótesis anteriores, subdividimos en dos subconjuntos de datos según a que década está asociado cada valor de la columna “Loudness” con el fin de facilitarnos el análisis. Luego decidimos ver como se distribuía la variable mencionada en cada subconjunto, por lo que creamos un boxplot para cada uno.

Figura 7: muestra la distribución de la variable “Loudness” para las canciones pertenecientes la primera mitad de década (a izquierda) y a la segunda mitad (a derecha)

De manera similar que lo ocurrido con el análisis del grafico para la variable “Tempo”, vimos que las medias de ambas distribuciones parecían estar sobre los mismos valores, por lo que supusimos que lo más probable es que no haya diferencias entre ambos lustros. Pero para poder aportar una respuesta estadística más robusta, decidimos tratar de aplicar un test paramétrico.

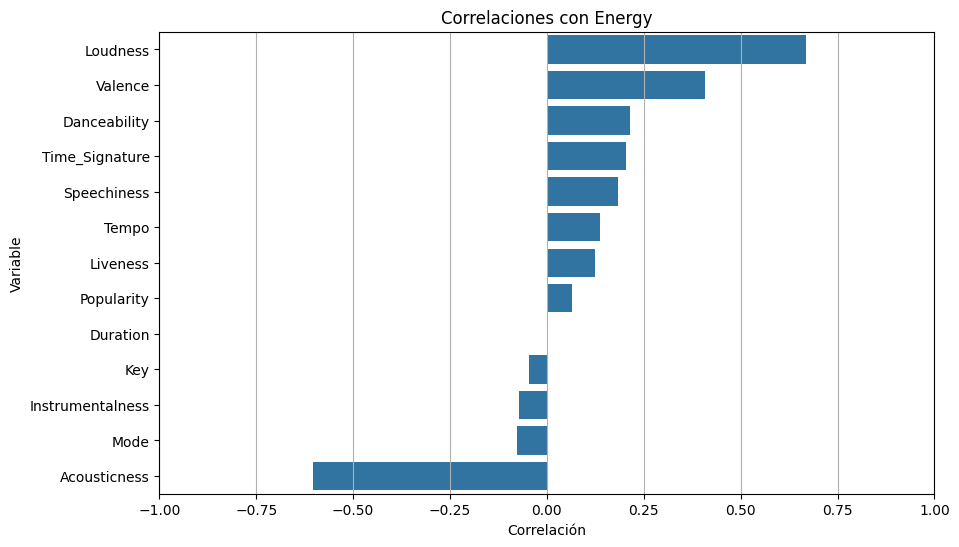
De esta forma, tratamos de realizar un t-test. Pero previo a la realización de este debemos cumplir los supuestos necesarios. Uno de estos es el supuesto de normalidad el cual testearemos con el test de Shapiro Wilks. En este caso, el p-valor que arrojó para ambos es menor a 0,05 por lo que no se cumple el supuesto. Por culpa de ello, descartamos el uso de algún test paramétrico.

Por ende, decidimos tratar de utilizar un U-test de Mann-Whitney. Pero este posee un supuesto el cual requiere homocedasticidad de los datos. Para ello utilizamos un test de Levene el cual nos proveyó de un p-valor el cual era 0,478. Como el p-valor es mayor a 0,05 podemos concluir que ambos subconjuntos tienen igualdad de varianzas y que podemos utilizar el U-test.

Por último, decidimos que nuestra hipótesis nula seria que las canciones de la década del 1970 no tienen grandes diferencias en “Loudness” al ser divididas por lustro. Para comprobar esto realizamos el test, el cual nos otorgó un p-valor igual a 0,155. Como es mayor a 0,05 podemos aceptar la hipótesis nula, por lo que no podemos afirmar que hay diferencias entre las canciones de ambos lustros de la década de 1970 con respecto al volumen de las canciones.

### Regresión lineal

Luego de realizar todos los análisis de las hipótesis quisimos ver si podíamos realizar una regresión lineal sobre la variable “Popularity”, dado que todavía teníamos la duda de sí podía ser explicada por una combinación de variables. Pero al ver que las correlaciones una a una con todas las demás variables del conjunto de datos era muy baja, decidimos orientar el análisis sobre la variable “Energy” dado que las correlaciones con otras variables eran mucho mayores.

Figura 15: gráfico de barras que muestra la correlación de la variable “Energy” con todas las demás variables cuantitativas del conjunto de datos.

Como se puede observar en el gráfico, “Loudness”, “Acousticness”, “Valence” y “Danceability” son las variables con mayor correlación una a una con la variable elegida. Por ende, vamos a buscar predecir “Energy” en base a las otras mencionadas previamente. Realizamos el escalado de las variables para ninguna pese por demás en la regresión. Luego planteamos el modelo de regresión lineal y obtuvimos sus resultados.

Al momento de analizar cómo quedo nuestra regresión lineal, lo primero que miramos es el R cuadrado, el cual representa la varianza de la variable “Energy” y da como significado los puntos o muestras que caen exactamente sobre la línea de regresión. En nuestro caso, el valor es de 0,651, el cuál es muy bueno por lo que podemos estar en presencia de un buen modelo de regresión lineal. Para asegurar esto nos fuimos a fijar el p-valor asociado al estadístico F, el cual es considerablemente bajo, lo que indica que el modelo en su conjunto es significativo. Otro motivo por el cual podemos afirmar que el modelo es bueno es que los p-valores asociados a cada una de las variables independientes que componen los coeficientes Bi son mucho menores a 0,001, por lo que indican que hay una asociación fuerte entre esas variables y “Energy”.

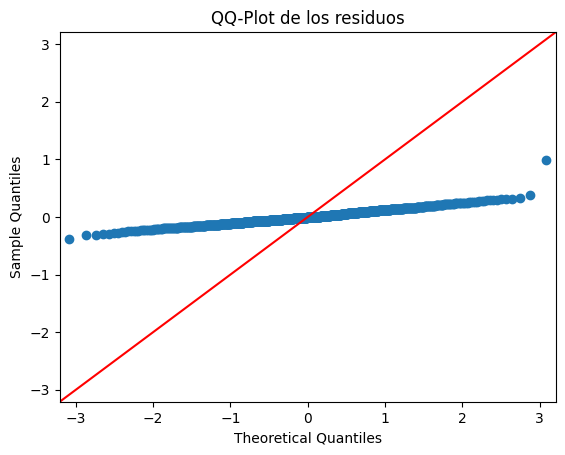
Pero para poder terminar de afirmar que nuestro modelo de regresión lineal es bueno debe cumplir con un supuesto, normalidad de los errores. Por lo tanto, decidimos analizarlo generando un QQ-Plot el cuál muestra que tan cercana esta nuestra distribución de errores a una distribución normal. Para poder suponer que son normales debemos ver que la línea diagonal roja que cruza de abajo a la izquierda hacia la esquina superior derecha esté tapada por los puntos azules que representan la distribución de nuestros errores.

Figura 16: QQ-Plot que muestra la distribución de los errores (puntos azules) del modelo de regresión lineal comparado con una distribución normal (línea roja).

Luego del análisis del QQ-Plot podemos darnos cuenta que la distribución de los residuos no esta ni aproximada a parecerse a una distribución normal. Por este motivo concluimos que nuestro modelo de regresión lineal, a pesar de ser muy bueno en algunos aspectos, no es confiable por culpa de la distribución de sus residuos. Lo más lógico en este caso es descartarlo y que no sea utilizado.

### Modelo Predictivo de la variable Energy

En base a lo que sabemos y lo aprendido, se nos plantea la posibilidad de crear un modelo predictor de la variable “Energy”, a través de un regresor lineal. Para eso utilizamos los datos que teníamos previamente limpios de nuestro conjunto de datos. Separamos los datos en tres grupos de canciones, estratificándolos en base a la variable Energy (alta, media y baja). Luego procedimos con la clasificación en datos de entrenamiento, validación y testeo. Para mantener la consistencia en las pruebas y futuras replicaciones fijamos el “random\_state”. En total nos quedamos con 617 datos de entrenamiento, 69 de validación y 294 de Test.

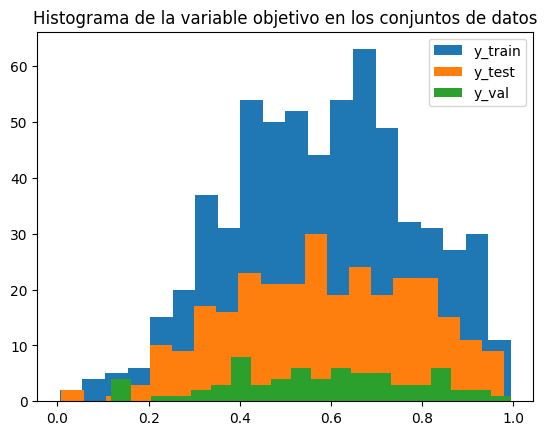


Figura 12: histograma que muestra cómo se distribuyen los datos de entrenamiento, validación y testeo.

En el histograma de la variable objetivo podemos observar cómo se distribuyen en azul los datos de entrenamiento, en naranja los de test y en verde los de validación. Se puede observar que, además de que concuerdan el tamaño de cada histograma para cada grupo de datos, podemos ver que sus distribuciones son muy parecidas entre sí, por lo tanto, concluimos que la división de los datos es correcta.

Después de todo lo procedido, escalamos las variables independientes con el fin de que tengan la misma varianza y desvío para que no haya variables que pesen por demás. Guardamos el escalador para poder hacer la entrega junto con el modelo predictivo, dado que, si no se utiliza exactamente el mismo escalador, se pueden generar discrepancias dado que los datos se pueden modificar de distintas formas.

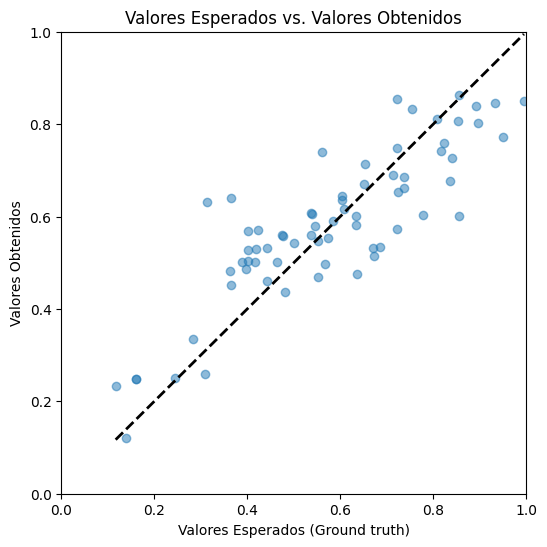
Para poder crear el modelo utilizamos un Regresor Lineal. Lo entrenamos con los datos de entrenamiento escalado y los valores de Energy asignados al entrenamiento.

Figura 13: Scatterplot que muestra la relación entre los valores obtenidos y esperados.

Lo que podemos observar de la figura es los valores obtenidos de cada muestra y lo que se espera de cada uno es la línea punteada que atraviesa el gráfico. Si las muestras están más acercadas a la línea significa que están más acercados a estar correctamente predichos.

Decidimos implementar un Ridge Regression para ver si con un regularizador podíamos obtener un error menor. Del mismo modo, decidimos plantear un LASSO, aplicando el mejor Alpha encontrado cuando hicimos el Ridge Regression, con el mismo objetivo de buscar minimizar los errores. Para nuestra sorpresa, encontramos que el mejor Alpha era 0,001, lo que implicaría que el regresor lineal convencional es muy bueno y no necesita ser regularizado.

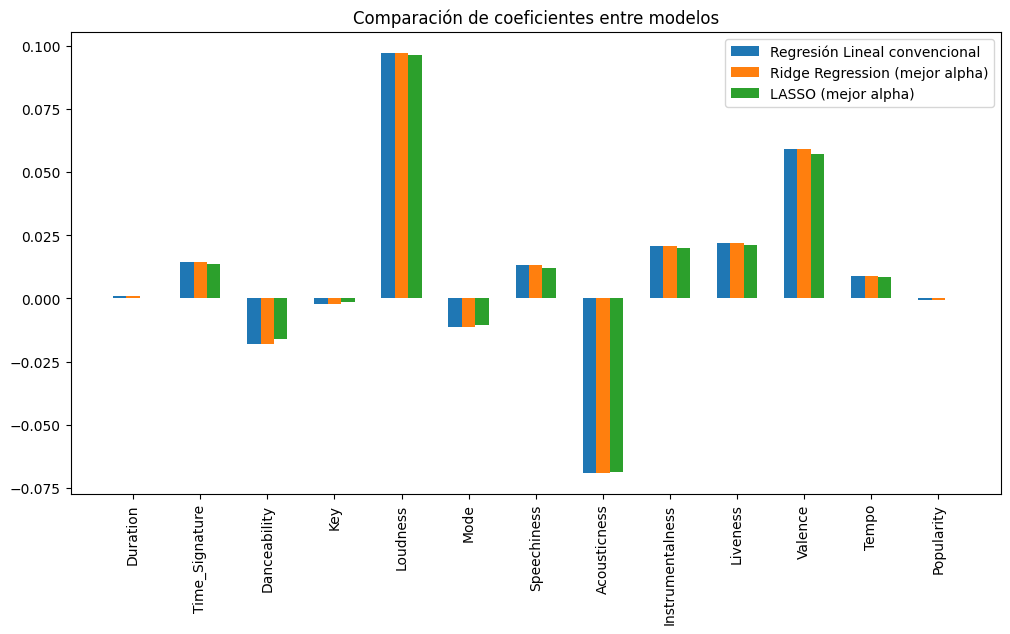


Figura 14: gráfico de barras que muestra la relación entre los coeficientes según la Regresión Lineal Convencional, con Ridge Regression y LASSO.

En el gráfico se puede observar todos los coeficientes que se generan con cada forma de crear la predicción. En este se puede observar que los coeficientes son prácticamente iguales sin importar cual de todos los métodos se aplicó. Esto esta dado ya que el mejor Alpha elegido para ambos métodos es 0,001. Por lo tanto, se puede intuir que el regresor lineal convencional es muy bueno y no necesita de un regularizador.

Posteriormente, calculamos el MAE (error absoluto medio) para los datos de validación aplicando el mejor Alpha tanto para LASSO como para Ridge Regression. os resultados son 0,08726 en el modelo convencional, 0,08726 para Ridge y 0,08784 para LASSO. Es decir, todos poseen un MAE idéntico, por lo que la afirmación de que el regresor lineal convencional es la mejor toma fuerzas, pero todavía queda probarlo con los datos de testeo.

Habiendo probado con los datos de validación pasamos a hacerlo con los datos de test, los cuales generaron los siguientes resultados. El MAE nos dio 0,08514 en el modelo convencional, lo mismo para el Ridge, y 0,08547 para LASSO. Como los resultados generados con los datos de testeo siguen la misma línea que los generados al hacerlo con los datos de validación, concluimos que efectivamente la mejor forma para hacer predicción sobre la variable “Energy” es con un regresor lineal convencional ya que este es el más simple.

Adicionalmente concluimos que el regresor lineal es bueno ya que el error que infieren en los datos predichos no es grande. Por lo tanto, es aceptable cometer ese error teniendo en cuenta que el regresor es simple. Para una mejor comparación sería necesario probarlo con muchos más datos, pero, con lo visto hasta el momento, nos parece un buen predictor.

# Conclusiones

Al finalizar este análisis, logramos explicar que, en norma general, la popularidad actual de las canciones de la década de los años 1970 no se puede explicar de manera simple utilizando medidas como la positividad y alegría o pulsaciones por compás. A su vez, se puede ver, a raíz del análisis de las hipótesis, si hacemos la comparación entre canciones de la primera mitad de la década y las de la segunda mitad en búsqueda de diferencias claras en el estilo de las mismas, solo podemos decir que varió la calidad acústica de las pistas. Por ende, las canciones durante toda la década no tienen diferenciaciones claras de estilo si comparamos por lustros.

Mientras realizábamos el trabajo fuimos notando algunas cuestiones que nos limitaron el análisis de los datos, como por ejemplo que las variables están poco correlacionadas, particularmente la variable de popularidad tiene muy poca relación con las demás y esto hace que no se pueda exprimir tanto con ella como posible variable objetivo. Además, dichas variables tampoco tenían una distribución normal, a la que le podamos hacer un Test-t o un ANOVA, lo que nos permita utilizar la totalidad de las herramientas aprendidas durante la cursada.

A su vez, el conjunto de datos que se nos proporcionó tiene algunas características buenas que nos gustaría destacar. Esto está dado gracias a que no tuvimos presencia de nulos ni valores inválidos en ninguna de las columnas. Esto nos facilitó el trabajo ya que no tuvimos que dedicarle tanto tiempo a la limpieza de datos. Otra buena característica del dataset es que se trata un tema interesante y que genera motivación para investigarlo y analizarlo, al cual una persona puede llegar a tener cierto conocimiento del domino sin ser especialista en la materia.

Otra buena experiencia que tuvimos al hacer este trabajo fue al realizar el modelo de regresión lineal y el predictor. Esto nos genero motivación a tratar de realizar un buen modelo de regresión lineal ya que pudimos identificar una variable la cual podía ser explicada mediante otras y que, a pesar que se nos terminó cayendo por incumplir el supuesto de normalidad de los residuos, al analizar el propio modelo vimos que podía llegar a ser muy bueno. Con respecto al predictor, fuimos con menos esperanzas dado que el regresor lineal no nos había salido como esperábamos, pero nos llevamos una grata sorpresa al ver que habíamos generado un regresor lineal convencional bueno que podía predecir los valores de la variable “Energy” de buena manera y sin necesidad de un regularizador.

En conclusión, este trabajo nos pareció muy interesante ya que comenzamos a adentrarnos en la Ciencia de Datos, que es un campo en el que nunca habíamos explorado. Otro motivo por el que nos gustó este trabajo es que, a diferencia de otros que hemos hecho en el pasado, no hay una receta para seguir y se valora más cuestiones como la creatividad o el buen criterio a la hora de elegir que investigar o demostrar.

# Referencias

* Diapositivas y Google Colabs propuestos por la Cátedra de Fundamentos de la Ciencia de Datos de la Facultad de Ciencias Exactas de la Universidad Nacional del Centro de la Provincia de Buenos Aires
* Dataset mencionado para la obtención de la explicación de el significado de la variable “Popularity”: [🎹 Spotify Tracks Dataset](https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset)
* Lista de sitios web mencionados, utilizados para corroborar valores de las variables
  + GetSongBPM: <https://getsongbpm.com/>
  + Tunebat: <https://tunebat.com/>