Informe TPE Fundamentos de la Ciencia de Datos

Integrantes:

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

## Índice

[Índice 2](#_Toc182475227)

[Introducción 3](#_Toc182475228)

[Materiales 4](#_Toc182475229)

[Preprocesamiento de los datos 4](#_Toc182475230)

[Análisis de los datos 5](#_Toc182475231)

[Métodos / Resultados 8](#_Toc182475232)

[Las canciones que tienen valores más altos de positividad y alegría hoy en día son las más populares. 8](#_Toc182475233)

# 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 cuál 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

## 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.

## 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

# 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 genero 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 1: 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 arrojo 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.

Conclusiones

Referencias