Fundamentos de la Ciencia de Datos
TPE 2024

Grupo 13

Miserendino, Enzo
Pereyra Wagner, Agustina

Introducción

Con el objetivo de dominar las técnicas y algoritmos para el análisis de datos, e identificar las más apropiadas para problemas específicos, se llevó a cabo un análisis exploratorio de datos sobre un conjunto que incluye 998 canciones de una década del siglo pasado (Spotify).

El procedimiento comenzó con la limpieza y la transformación de los datos en el dataset, y prosiguió con el análisis univariado de los datos, donde se plantearon las primeras hipótesis bivariadas basadas en ideas racionales que surgieron a partir del conocimiento de las características presentadas.

Las hipótesis fueron descartadas y validadas mediante el uso de diversos gráficos, entre ellos gráficos de dispersión y mapas de calor. Además, durante su análisis comenzaron a surgir nuevas hipótesis, tanto bivariadas como multivariadas que también fueron tratadas con los métodos que se consideraron pertinentes.

La hipótesis más importante se generó alrededor de la interrogante: ¿Existen características que contribuyan a la popularidad de una canción? Esta permitió generar un amplio estudio que requirió el uso de variadas técnicas y conceptos aprendidos.

Materiales

Se hizo uso de Visual Studio para el desarrollo del código implementado; también de GitHub para facilitar la colaboración de proyectos y el material informativo aportado por la cátedra de la materia.

Desarrollo

Análisis Exploratorio

El análisis comenzó con una exploración general de las canciones y las características contempladas para cada una de ellas. Entre las variables se encuentran: el título de la canción (Track), el intérprete o grupo (Artist), la duración en minutos y segundos (Duration), y atributos técnicos como la métrica musical (Time_Signature), el ritmo (Tempo), la tonalidad (Key), el volumen promedio (Loudness), y las medidas de características subjetivas como la facilidad para bailar

(Danceability), la energía (Energy), y la positividad (Valence). También se incluyen variables como la popularidad de las canciones (Popularity) y el año de lanzamiento (Year).

Algunas de estas características provienen del ámbito musical, por lo que se debió investigar el significado de términos específicos como “Loudness”, luego de indagar en el tema para los valores que toma esta variable, se encontró que, en audio digital, el punto de referencia para medir dB es el "nivel máximo posible sin distorsión", conocido como 0 dBFS. Todo lo que esté por debajo de este nivel se representa con valores negativos. Por ejemplo: -5 dB indican una canción muy fuerte, y -20 dB una canción significativamente más suave.

El conjunto de datos incluyó: 9 variables cuantitativas continuas (Danceability, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo), 4 cuantitativas discretas (Popularity, Time_Signature, Year), 2 categóricas nominales (Track, Artist), una dicotómica (Mode: indica si la tonalidad es mayor o menor), una categórica ordinal (Key) y una variable cuantitativa continua denominada Duration, que representa la duración de las canciones en formato minutos:segundos.

Realizando un análisis univariado de los datos con sus resúmenes estadísticos, se pudieron interpretar las tendencias de las canciones del dataset. Algunas observaciones fueron:

- Las canciones típicamente llevan un compás de 4 tiempos.
- La distribución es normal para Danceability, Energy y Loudness por lo que las canciones son moderadamente bailables y energéticas ademas de ser uniformes en volumen.
- Suelen contar con baja presencia de habla, baja acústica y presencia de voces (en base al sesgo a derecha en Speechiness, Instrumentalness y Acousticness).
- Tienden a ser positivas.
- La popularidad es muy diversa.

Se hallaron posibles outliers en Speechiness (el valor máximo es considerablemente más alto que el 75%, lo que sugiere que hay al menos una canción con una presencia muy alta de voz hablada), Acousticness (el valor máximo es notablemente alto en comparación al resto) e Instrumentalness (el máximo es muy elevado con respecto a los demás valores). No obstante como no se conocen los rangos posibles que podrían adoptar dichos valores, no se consideró correcto deducirlos como errores y fueron interpretados como variaciones normales dentro del conjunto de datos pudiendo ser indicativos de canciones con características inusuales.

Para poder emplear cada una de las técnicas aprendidas, fue fundamental realizar primero una limpieza de los datos. En un principio se consideró y descartó la posible existencia de valores faltantes en atributos (verificando que no hubieran Nan, ceros sin sentido o algún otro tipo de nulo), mismo para datos inconsistentes y ruidosos. Por otro lado, se encontraron columnas de datos donde se consideró apropiado aplicar transformaciones: en primer lugar se convirtió en una variable numérica continua a Duration (pasando la duración a segundos), para facilitar cálculos, comparaciones y su uso en futuros análisis; luego, como se hallaron datos en la columna Artist que contenían más de un artista por canción, se tomó la decisión de crear una nueva columna categórica nominal llamada “Featuring”, la cual utiliza encoding binario asignando el valor 1 si la canción pertenece a más de un artista y 0 en caso contrario.

Posteriormente se realizó una búsqueda de elementos repetidos en todo el dataset, donde se encontró únicamente una canción completamente repetida y nueve canciones “repetidas” que compartían todas las características con su repetida, excepto la de Year. Se tomó la decisión de eliminar únicamente la que estaba completamente repetida, debido a que no aportaba un gran cambio la eliminación de las demás y no se podía asegurar que la canción era realmente la misma.

Validación de las Hipótesis

Haciendo uso únicamente del conocimiento de los atributos de cada canción y sus significados, se plantearon 3 hipótesis bivariadas de base, con el fin de confirmarlas o descartarlas a lo largo del estudio:

1. Mientras mayor sea el **Loudness**, mayor será el **Liveness**: se planteó esta hipótesis bajo la suposición de que, al ser interpretada en vivo, el ruido de la audiencia podría incrementar los decibelios de la canción.
2. Mientras mayor sea el **Acousticness**, menor será la **Energy**: esta se generó bajo la idea de que mientras más acústica sea una canción, menos intensa y enérgica podría ser.
3. Mientras mayor sea la **Energy**, mayor será el **Valence**: especulando que, mientras más intensa y enérgica sea la canción, más alegre podría resultar.

Validación de la Hipótesis 1

El análisis bivariado de datos comenzó con la visualización de un Profile Report que permitió inspeccionar las dispersiones de los valores de las variables con Scatter Plots para detectar si existía una correlación entre cada par. Gracias a eso, se pudieron detectar relaciones entre *Danceability y Valance* (este gráfico presentó una correlación positiva aunque un poco dispersa), *Energy y Loudness* (las muestras tienden a incrementar en conjunto) y *Energy y Acousticness* (la dispersión mostró una tendencia decreciente, de forma que al aumentar los valores de energy, los de acousticness se decrementan y viceversa).

El Scatter Plot realizado para ***Loudness y Liveness*** resultó en muestras dispersas sin una tendencia notoria, indicando que no había una relación entre ambas. Para analizar a fondo antes de descartar y/o generar nuevas hipótesis se creó un mapa de calor que permitió observar el coeficiente de correlación lineal para cada par de variables. Como se encontró únicamente una correlación alta (es decir, mayor que 0.7 o menor que -0.7) se consideró la posibilidad de reducir el umbral, pero la opción quedó descartada cuando se observó que el umbral debía ser de 0.5 para encontrar más correlaciones, un valor débil que no iba a permitir realizar predicciones confiables. Loudness y Liveness no presentaron un coeficiente de correlación alto, de manera que no existe una relación evidente entre el volumen promedio de la canción (loudness) y la probabilidad de que haya sido interpretada en vivo (liveness). La **hipótesis 1** quedó finalmente descartada.

Validación de la Hipótesis 2

Con respecto a la **Hipótesis 2**, el Scatter Plot presentó una dispersión analizable entre ***Acousticness y Energy*** dado que los puntos mostraban una tendencia descendente (aunque un poco dispersa), indicando que a medida que aumentaban los valores de energy, los de acousticness se decrementaban y viceversa. Para profundizar el análisis, se observó el coeficiente de correlación entre las variables en el mapa de calor, sin embargo este no resultó significativo, por lo que a pesar de haber generado un gráfico de dispersión con un patrón notable, la **hipótesis 2** quedó descartada. No es cierto que mientras más acústica sea una canción, menos enérgica es.

Validación de la Hipótesis 3

El Scatter Plot realizado para ***Energy y Valence*** resultó en muestras dispersas sin un patrón claro, además, el coeficiente de correlación observado en el mapa de calor fue bajo, razón por la cual la

hipótesis 3 quedó descartada, indicando que el comportamiento de la variable energy en las canciones no se encuentra relacionado con el de valence.

Al inspeccionar detenidamente el mapa de calor para Popularity y no encontrar coeficientes de correlación mayores a 0.7 y notar que sus gráficos de dispersión no seguían una tendencia concisa, se encontró que no había una correlación fuerte entre la popularidad y el resto de características, esto generó una nueva interrogante: ¿Existen características que contribuyan a la popularidad de una canción?

Se generó una nueva hipótesis:

4. La **Popularity** no está relacionada con el resto de características de las canciones.

Validación de la Hipótesis 4

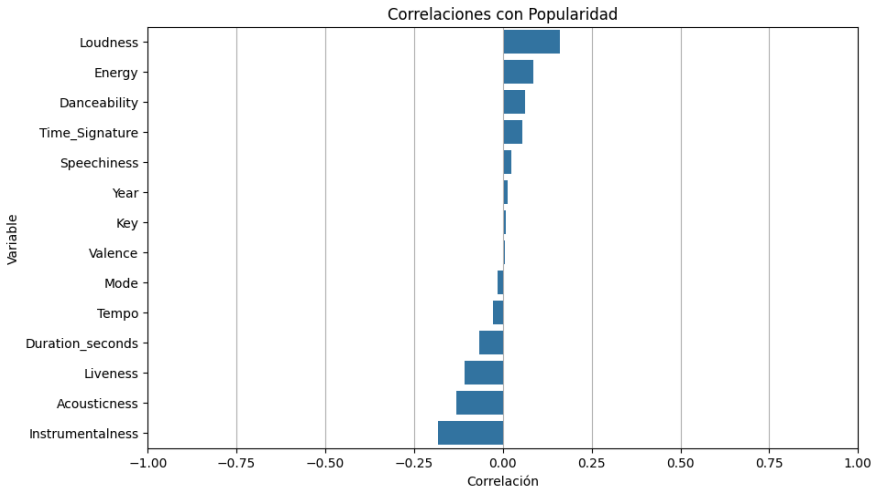
Para corroborar la **Hipótesis 4** se empezó observando las canciones más y menos populares de cada año con un gráfico de barras. Durante su generación se observaron varias con popularidad cero, lo cual resultó extraño, como una especie de error en la medición. Por esa razón y con el fin de tener una mejor visualización se creó un dataset sin las canciones con valor 0 de popularidad. Posteriormente se escucharon las mismas con el fin de analizar si existía alguna diferencia de estilo o género entre las más y menos populares. Por lo que se pudo notar, las canciones menos populares parecían más lentas y románticas, mientras que por otro lado las más populares eran en su mayoría movidas, lo que podría relacionarse con el auge de las canciones pop y rock características de los años 80. [Referencia 1]. Estas observaciones motivaron el análisis multivariado para investigar si realmente había características cuantitativas que distinguieran a las canciones según su popularidad.

Lo primero fue estandarizar los valores para que las escalas no fueran demasiado diferentes, de manera que todas presentaran una varianza semejante.

Lo siguiente fue reducir la dimensionalidad del dataset para conservar únicamente las características más relevantes aplicando PCA. Además, para ver si realmente habían características que diferenciaban a las canciones más y menos populares, se tomó a “Popularity” como la variable a estudiar en relación al resto y se pintó el gráfico en base a esta.

Al observar que PCA producía una gran pérdida de información, se probaron otros métodos: T-SNE y UMAP. Sin embargo, los gráficos salientes no parecieron mostrar mejoras significativas con respecto al primero, siendo que los colores que representaban a la popularidad se encontraban dispersos en variadas zonas. Esto llevó a pensar que la **Hipótesis 4** cobraba más sentido, dado que no existía una separación clara entre canciones populares y no populares en el espacio reducido.

Para ir más a fondo, se realizó un gráfico de barras con las correlaciones entre “Popularity” y cada una de las demás características, este con el fin de observar de manera clara cuáles eran las más relacionadas con la popularidad: Instrumentales en primer lugar y Loudness en el segundo [Img 1].



[Img 1] Correlaciones con Popularidad en gráfico de barras.

Para un análisis más amplio se llevó a cabo un gráfico de dispersión entre Instrumentalness y Popularity. Como resultado, se obtuvo una distribución de puntos que no contribuyó con el análisis, siendo que no se distinguía una tendencia determinada.

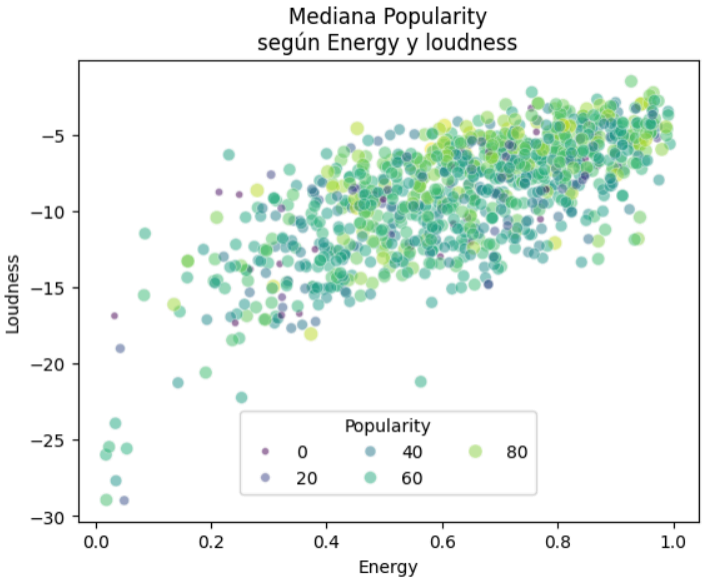
A pesar de esto, se decidió implementar Regresión Lineal Simple con esta variable, para respaldar la **hipótesis 4** con más seguridad.

El resultado fue que Instrumentales explicaba solo el 3.4% de la varianza de Popularity, lo cual resultó muy bajo, implicando que el modelo no ajustaba adecuadamente los datos.

Se buscó verificar la distribución de los residuos con Shapiro Wilk para comprobar si el sumary arrojado tenía validez, para que la tuviera los residuos debían seguir una distribución normal (p-valor>0.05). Como el p-valor fue menor que el umbral, dió como resultado el rechazo de la hipótesis nula “los datos provienen de una distribución normal” lo cual en conjunto al histograma de distribución permitieron concluir que la distribución no era normal. Se concluyó que no había evidencia suficiente para apoyar la conjetura de que la instrumentalidad se relaciona significativamente con la popularidad.

Recordando los resultados en la matriz de correlaciones, había un único coeficiente de correlación alto: entre Energy y Loudness. Se planteó entonces, la idea de que quizás a pesar de no estar fuertemente relacionadas Popularity y Energy, podría generarse un gráfico más claro para el análisis si esta nueva característica se tenia en cuenta.

El nuevo Gráfico de Dispersión mostró una leve tendencia de popularidad en la zona superior de la dispersión [Img 2].



[Img 2] Gráfico de dispersión Energy y Loudness vs Popularity

Para ver en profundidad cómo afectaban las variables a la popularidad se realizó Regresión Lineal Múltiple incluyendo a Instrumentalness, Loudness y Energy, con la finalidad de entender el impacto combinado de las tres características en la variable de interés Popularity.

Luego de ajustar el modelo y observar su Summary, el coeficiente de determinación R^2 arrojó un resultado de 5%, indicando que las variables adicionales no aportaron información útil al modelo, concluyendo que no es posible afirmar que las variables analizadas explican la popularidad.

Antes de sacar conclusiones con este modelo, se corroboró que la distribución de los residuos fuera normal, para ello se observaron los resultados de las variables Prob(Ómnibus) y Prob(JB), si ambos valores son cercanos a 0, los residuos no son normales; este fue el caso ocurrido, por lo que no se pudo validar lo arrojado por el modelo.

Dado que, nuevamente no se halló evidencia suficiente para confirmar una relación entre la popularidad y las variables Instrumentalness, Loudness y Energy, se implementó como último recurso agregar al modelo todas las variables cuantitativas del dataset, para ver si así podía explicarse de mejor manera el comportamiento de la popularidad a través de sus variables. Este último modelo, dió un coeficiente de determinación de 8.5%, lo cual se percibió mejor que lo obtenido anteriormente, pero aún así insignificante. Nuevamente se verificaron los valores de Prob(Ómnibus) y Prob(JB) para probar la normalidad de los residuos y ambos valores dieron muy cercanos a 0. Por lo tanto este modelo también se rechazó.

Ante la imposibilidad de demostrar la existencia de variables capaces de explicar el comportamiento de la popularidad, se rechazó la **hipótesis 4**.

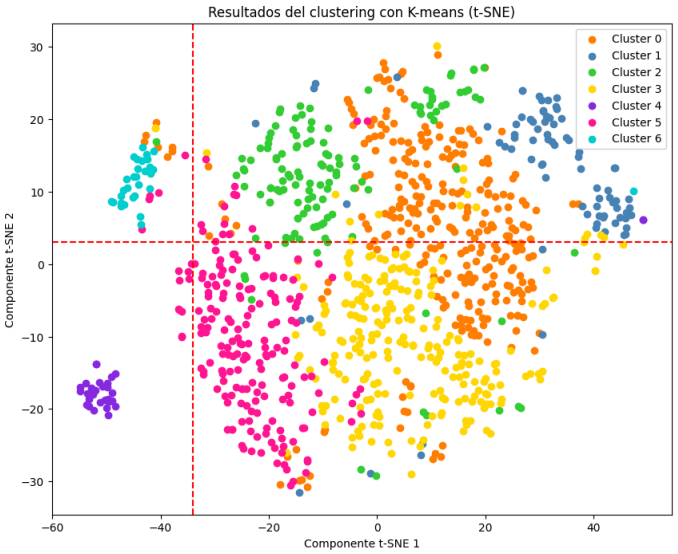
Cuando se realizaron los gráficos para visualizar la proyección del conjunto de datos en dos dimensiones (gráfico de salida T-SNE) se percibió una concentración amplia de puntos por un lado, y pequeñas agrupaciones de puntos separadas de ellos. Es por esto, que se consideró interesante hacer un análisis detallado sobre la separación de las muestras, agrupando los datos por clusters para luego comparar las distribuciones de las características en cada uno.

Se intentó agrupar los datos según sus similitudes haciendo uso del Clustering Jerárquico. Este generó tres grupos, sin embargo, uno de ellos estaba conformado por una canción y otro por dos, lo

que provocó que estos grupos no fueran representativos del dataset. Por este motivo se buscó otra alternativa para obtener clusters, el algoritmo K-means.

Dada la necesidad de utilizar un valor k para implementar el algoritmo K-means, se utilizó Elbow plot, esta técnica se basa en observar el gráfico de línea generado, el valor óptimo se encuentra analizando en qué valor de k la línea comienza a decrecer de forma diferente, más gradual, siendo este el valor más óptimo. El gráfico retornó k=7 como el mejor k.

Posteriormente se llevó a cabo la visualización del resultado en un Scatter-Plot, donde, por ser alta la dimensionalidad en las características, se trasladó a dos dimensiones utilizando T-SNE [Img 3].



[Img 3] Gráfico de clusters identificados por color

Se observaron entonces 7 clusters diferentes representados cada uno con su respectivo color y tres agrupaciones de datos: una en la que formaron parte la mayoría de los datos del dataset, una aislada a la izquierda y otra aislada en menor medida.

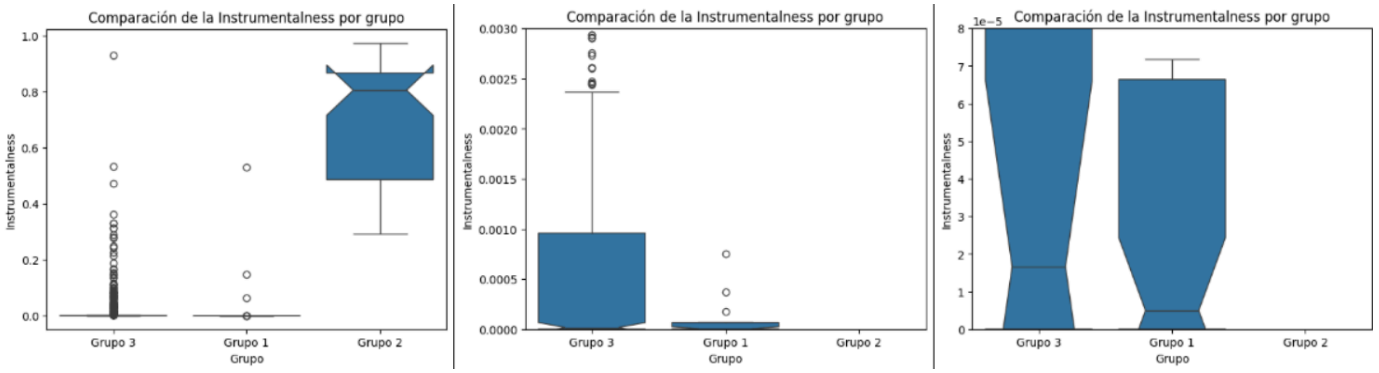
A pesar de que los métodos Davis Bouldin y Silhouette Score favorecieron el Clustering Jerárquico (dado que el índice de Davis Bouldin dió más bajo y el de Silhouette Score generó un valor más cercano a 1) la decisión final fue mantener los clusters obtenidos con K-means, ya que estos separaban los datos de manera más equitativa y representativa.

Como los grupos aislados podrían representar canciones con características distintas en relación al resto, se realizó un análisis sobre cada grupo. Para este análisis, se creó un dataset con una nueva columna “Grupo” de valores categóricos nominales cuya función fue clasificar cada canción con “Grupo 1”, “Grupo 2” o “Grupo 3”. Esto permitió estructurar los datos para un análisis comparativo de las características en cada grupo.

Se realizó un análisis más detallado utilizando Box Plots para cada grupo y variable, esto fue con el fin de ver en qué variable se diferenciaban más los grupos. En particular, se observó una variabilidad considerable entre los grupos cuando el análisis se hizo a partir de Instrumentales [Img 4] siendo que existía una clara distinción en los rangos de los valores y se contaba con la presencia de outliers.

Por esta diferencia visual, se interpretó que la separación entre los grupos observada en el Scatter-Plot podría deberse a la amplia diferencia entre los valores que toma esta variable para cada grupo:

- El grupo 1 presentó una gran cantidad de valores atípicos y la mayoría de sus datos rondaban el valor 0.
- El grupo 2 contenia valores significativamente mayores en comparación con los demás grupos.
- El grupo 3 (el que englobaba la mayor cantidad de datos) presentó la mayor dispersión de valores (también cercanos a 0) y outliers.



[Img 4] Box Plot de Instrumentalness para cada grupo, ampliado con fines visuales

De esta manera Instrumentalness se consideró una característica potencialmente diferenciadora para realizar un análisis más profundo y explorar en qué medida las diferencias ayudaban a separar los grupos.

Surgió entonces, la nueva hipótesis:

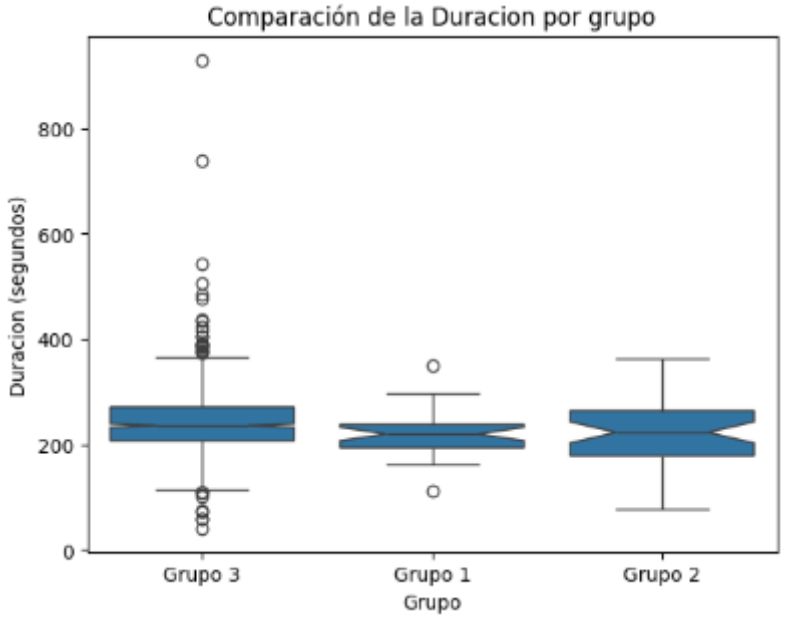
5. La variable **Instrumentalness** podría ayudar a distinguir subconjuntos dentro del conjunto de datos.

Validación de la Hipótesis 5

Para corroborar la **Hipótesis 5**, se procuró hacer un test t, este es un test paramétrico, es decir, asume que los datos siguen una distribución determinada. Es por esto que primero fue necesario verificar que los grupos tuvieran una distribución normal, con este fin se aplicó el test de Shapiro Wilk. Como el p-valor de todos los grupos dió menor que 0.05, se rechazó la hipótesis nula (“los datos provienen de una distribución normal”) y se concluyó que los datos no tenían una distribución normal. El siguiente paso consistió en verificar con el test de Levene si el **Grupo 2** era homocedástico (varianza similar entre los grupos) en relación con los **Grupos 1 y 3**. Esta hipótesis también se rechazó por generar un p-valor menor a 0.05. Por lo que no se pudo implementar un test t. Con esto demostrado, optamos por realizar el test de Kruskal-Wallis por ser un test no paramétrico ideal para comparar más de dos grupos cuando los datos no son normales. El resultado de este test arrojó un p-valor menor a 0.05 por lo que se rechazó la hipótesis nula (“no hay diferencias significativas entre las distribuciones de los grupos”). En consecuencia, se concluyó que hay una diferencia significativa de Instrumentalness entre el Grupo 2 y los otros dos grupos por lo que se verificó la veracidad de la **Hipótesis 5**.

Cuando se realizaron Box Plots para ver en qué variable se diferenciaban más los grupos, se encontró que, a diferencia de Instrumentalness, la variable Duration generaba Box Plots similares para cada grupo [Img 5]. En base a esta observación surgió una nueva hipótesis:

6. Todas las canciones cuentan con una **duración** similar.



[Img 5] Box Plots de Duration para cada grupo

Validación de la Hipótesis 6

Para demostrar la **Hipótesis 6**, se comenzó testeando la normalidad de los grupos con el test de Shapiro-Wilk cuyos resultados indicaron que el grupo 1 y 2 presentaban una distribución normal, y que el grupo 3 no. Para examinar esta normalidad se generaron QQ-Plots para los dos primeros grupos, con el fin de tener una manera visual de demostrar la normalidad. En ambos gráficos se vió como los puntos se acercaban a la recta, confirmando que ambos grupos seguían efectivamente una distribución normal.

A continuación, se buscó verificar mediante el test de Levene que los grupos 1 y 2 fueran homocedásticos (que tuvieran varianzas similares), como los resultados indicaron que los grupos eran en efecto homocedásticos por generar un p-valor>0.05, se habilitó la implementación de un test t para comparar la duración de las canciones entre estos dos grupos. El test t arrojó un p-valor mayor que 0.05, por lo que indicó que no había evidencia suficiente para rechazar la hipótesis nula, en consecuencia, no se encontró una diferencia significativa entre la duración de las canciones de los grupos 1 y 2.

Para el caso del grupo 3, por no tener una distribución normal, se procedió a utilizar un test no paramétrico, el test de Mann-Whitney U adecuado para comparar dos grupos. Antes, se verificó la homocedasticidad (aunque no es una condición estricta) utilizando el test de Levene, que arrojó un p-valor mayor a 0.05 indicando que el grupo 3 era homocedástico con los grupos 1 y 2.

Con estos supuestos probados se llevó a cabo el test Mann-Whitney U, el cual arrojó un p-valor menor que 0.05 para la comparación entre el Grupo 1 y Grupo 3, significando una gran diferencia entre la duración de las canciones. Para la comparación entre el Grupo 2 y Grupo 3 se observó un p-valor mayor que 0.05, de manera que no fue suficiente para rechazar la hipótesis nula (“los grupos

tienen la misma distribución"), sugiriendo que no hay una diferencia significativa entre ambos grupos.

Finalmente, la hipótesis 6 no se pudo validar en su totalidad, si bien entre los grupos 1 y 2 y los grupos 2 y 3 no hubo una diferencia significativa en la Duration, entre los grupos 1 y 3 sí, por lo que no es verdadera la **Hipotesis 6**.

Conclusión

El dataset presentó pocas características correlacionadas, muchas de las hipótesis que se pensaron en un principio resultaron descartadas.

Aunque en un principio no fue de nuestro agrado encontrarnos con que hipótesis que tenían mucha lógica, como "mientras más intensa y enérgica sea la canción, más alegre podría resultar", no contaban con un respaldo válido, a medida que fuimos desarrollando el informe entendimos que eso es algo cotidiano en el estudio de datos reales y que ese es el motivo por el cual es necesario realizar un análisis antes de dar una afirmación por cierta. Incluso, las conclusiones que se obtuvieron a lo largo del informe podrían no estar completamente justificadas por no tener en cuenta el momento en el que se obtuvieron los datos, el lugar y los métodos utilizados. Este podría ser el caso, por ejemplo, del análisis de la popularidad, donde podríamos considerar que no encontrar diferencias entre las características de las canciones más y menos populares se debe a que la muestra se constituye por el conjunto de canciones más populares de toda la década, siendo que las menos populares estuvieron aun así entre las 997 más populares de los 80s (esto lo pensamos en base a que algunas de las canciones menos populares son canciones conocidas al día de hoy).

A nivel personal, comenzamos resolviendo el trabajo con objetivos determinados y a medida que desarrollábamos nos encontrábamos con incógnitas que despertaban nuestro interés, de manera que el análisis se iba haciendo dinámico y nos resultó llevadero. Consideramos que logramos llevarnos un amplio aprendizaje para el análisis de datos.

Referencias

[Referencia 1] https://seniaevents.com/viaje-nostalgico-80-musica/