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 graficos, entre ellos graficos 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

El análisis comenzó con un paneo general de las canciones y las características que se contemplan para cada una de ellas. Estas últimas son propias del mundo musical, por lo que se necesitó investigar sobre el significado de términos como “Loudness” y “Tempo”.

Se encontraron 9 tipos de datos cuantitativos continuos, 4 cuantitativos discretos, 2 categóricos nominales, una dicotómica y una característica “Duration” categórica que contiene la duración de la canción en formato minutos : segundos.

Realizando un análisis univariado de los datos con sus resúmenes estadísticos, se hallaron posibles outliers en Speechiness, Acousticness e Instrumentalness. No obstante, no se consideró correcto deducirlos como errores y fueron interpretados como variaciones normales dentro del conjunto de datos, bajo el criterio de que no se conocen los rangos posibles que podrían adoptar dichos valores.

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

1. Mientras mayor sea el **Loudness**, mayor será el **Liveness**: se supone 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**: bajo la idea de que mientras más acústica sea una canción,menos intensa y enérgica es.
3. Mientras mayor sea la **Energy**, mayor será el **Valence**: se especula que, mientras más intensa y enérgica sea la canción, más alegre podría resultar.

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, mismo para datos inconsistentes y ruidosos. Por otro lado, se encontraron columnas de datos donde se consideró apropiado aplicar transformaciones, en primer lugar se utilizó encoding para convertir Duration a un dato numérico. Además, como se hallaron datos en Artist que contenían más de un artista por canción, se tomó la decisión de agregar una columna en el dataset “Featuring”, categórica nominal, cuyos valores son 0 o 1 dependiendo de si la canción pertenece a más de un artista o no .

Posteriormente se realizó una búsqueda de elementos repetidos en todo el dataset, donde se encontró únicamente una canción repetida dos veces y nueve canciones que compartían todas las características con su repetida, excepto la de Year. Se tomó la decisión de eliminar únicamente la que estaba completamente repetida.

El análisis bivariado de datos comenzó con la visualización de un Profile Report que permitió inspeccionar las dispersiones de los valores de las variables para detectar si existía una relación o correlación entre ellas. Gracias a eso, se pudo interpretar desde un principio que valores como Danceability y Valance, Energy y Loudness, Energy y Acousticness tienen algún tipo de relación. También se encontró que de las hipótesis planteadas en un principio, únicamente la **Hipótesis 2** presentó una dispersión analizable.

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. Además, 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. 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 .

En base a lo analizado anteriormente, quedaron descartadas **Hipótesis 1**, **Hipótesis 2** e **Hipótesis 3** y se encontró una correlacion fuerte entre Energy y Loudness. Además, al inspeccionar detenidamente el mapa de calor y los gráficos de dispersión, 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.

Para corroborar la **Hipótesis 4** se empezó observando las canciones más y menos populares de cada año con un gráfico de barras, luego se escucharon con el fin de analizar si existía alguna diferencia de estilo o género entre ambas, y 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]

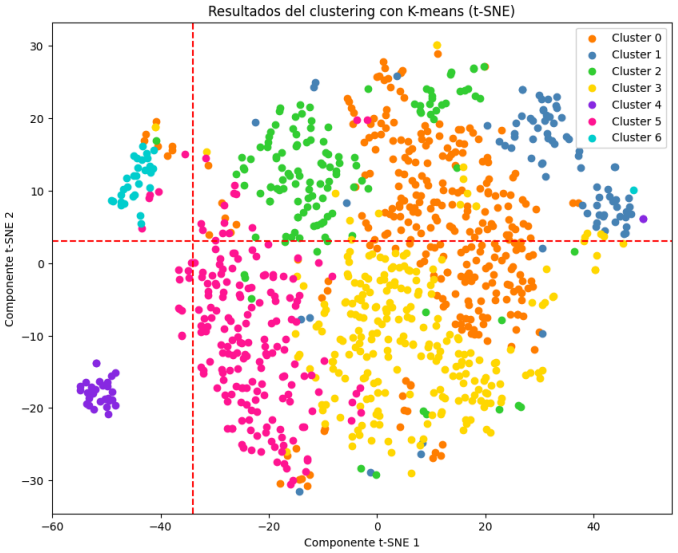
De esta manera, se comenzó con el análisis multivariado. Lo primero fue reducir la dimensionalidad del dataset para conservar únicamente las características más relevantes, y estandarizar los valores para que las escalas no fueran demasiado diferentes, de manera que todas presentaran una varianza semejante.

Lo siguiente fue aplicar PCA para ver qué características diferenciaban a las canciones más y menos populares, para esto se tomó a "Popularity" como la variable a estudiar en relación al resto.

Al observar que PCA producía una gran pérdida de información, se probaron otros métodos: T-SNE y UMAP. Sin embargo, los gráficos 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. Se retomó el desarrollo de esta hipotesis más adelante.

Al hacer un análisis más profundo en los gráficos mencionados anteriormente 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.

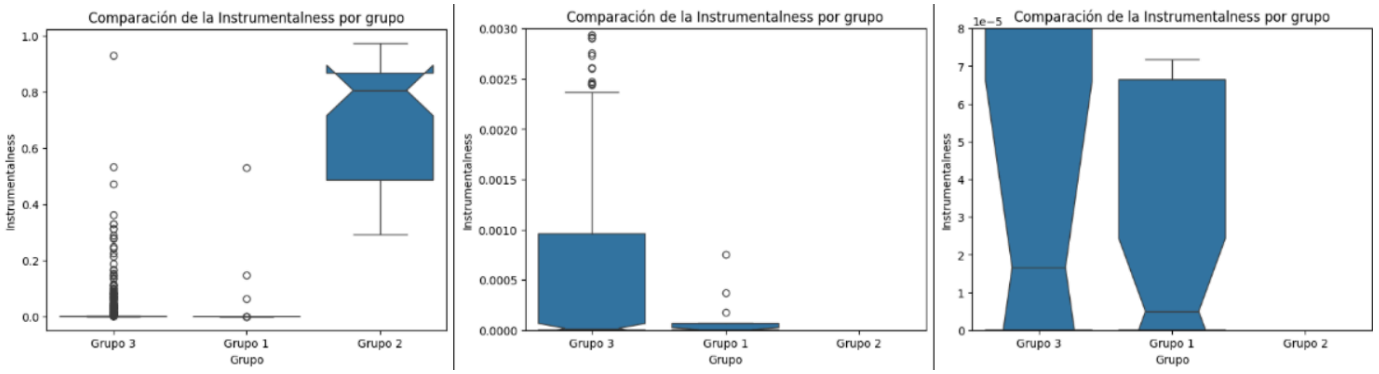
Dada la necesidad de utilizar un valor k para implementar el algoritmo K-means, se utilizó Elbow plot que retornó k=3 y k=7 como los mejores valores para k. A partir de esto, se realizaron los análisis correspondientes con ambos valores, los cuales llevaron a la conclusión de que k=7 era el indicado. Esta elección se hizo en base a 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 1].



[Img 1] 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 y otra aislada en menor medida. Como los grupos aislados podrían representar canciones con características distintas en relación al resto, se realizó un análisis sobre cada grupo. En 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". De esta manera los datos quedaron organizados por grupo, permitiendo realizar un análisis general para las características en cada uno.

Se implementó un Box Plot para cada grupo y por cada variable, pero la mayor diferencia se encontró en la variable Instrumentalness. Por esto, se interpretó que la separación entre los grupos era debido a la amplia diferencia entre los valores definidos para esta variable. 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. Por último, el grupo 3 (el que englobaba la mayor cantidad de datos) presentó la mayor dispersión de valores (también cercanos a 0) y outliers. [Img 2]



[Img 2] Box Plot de Instrumentalness para cada grupo, ampliado con fines visuales

Por otra parte, se intentó agrupar los datos según sus similitudes haciendo uso del Clustering Jerárquico. Este generó tres grupos, pero uno de ellos estaba conformado por una canción y otro por dos, razón por la cual no fue considerado como el método definitivo. Se implementaron igualmente los métodos de Davis Boulding y Silhouette Score y se graficaron sus resultados con Gráficos de

Barras que fueron a favor del Clustering Jerárquico. La decisión final fue conservar de todas formas los grupos obtenidos con K-means, por separar los datos de forma más equitativa y racional.

Surgieron entonces, a partir de los grupos obtenidos, dos hipótesis nuevas .

5. El Grupo 2 se destaca y diferencia del resto por su **Instrumentalness**.
6. Los tres grupos tienen canciones con una duración similar.

Para corroborar la **Hipótesis 5**, se intentó hacer un test t. Para ello, primero se verificó que los grupos tuvieran una distribución normal aplicando el test de shapiro. Como el p-valor de todos los grupos dió menor que 0.05, se rechazó la hipótesis nula y se concluyó que los datos no tenían una distribución normal. Luego se examinó si el grupo 2 era homocedástico en relación con los grupos 1 y 3. Esta hipótesis también se rechazó por el p-valor menor a 0.05, por lo que no se pudo implementar el test de Whitney Mann U. Con esto demostrado, no quedó otra opción que realizar un test de Kruskal-Walis. El resultado de este test arrojó un p-valor menor a 0.05 en ambos casos por lo que se rechazó la hipótesis. Finalmente, 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**.

Para demostrar la **Hipótesis 6**, se volvió a testear la normalidad de los grupos. El test de Shapiro-Wilks arrojó que el grupo 1 y 2 tenían distribución normal, y que el grupo 3 no. Para examinar esta normalidad se ejecutó un QQ-Plot para los dos primeros grupos, con el fin de tener una manera más visual de demostrar la normalidad. En ambos gráficos se vió como los puntos se acercaban a la recta.

Se demostró con el test de levene que los grupos 1 y 2 eran homocedásticos, lo cual permitió que se implemente un test t. Este último arrojó un p-valor de 0.531, el cual es mayor que 0.05 por lo que no hubo 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, al no tener una distribución normal, se probó la homocedasticidad con los grupos 1 y 2 con el fin de observar si se podía hacer un test de Mann-Whitney U. En ambos casos, el test de Levene arrojó un p-valor mayor a 0.05 por lo que el grupo 3 resultó homocedástico con los grupos 1 y 2. Con estos supuestos probados se llevó a cabo el test Mann-Whitney U. Este último test arrojó un p-valor de 0.019 para el grupo 1 y un p-valor de 0.069 para el grupo 2. Por lo que se rechazó la hipótesis nula en el primer caso, por lo tanto, los grupos 1 y 3 se consideraron significativamente diferentes, y en el segundo no hubo evidencia suficiente para rechazar la hipótesis nula, de manera que los grupos 2 y 3 no fueron significativamente diferentes.

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, entre los grupos 1 y 3 sí, por lo que no es verdadero que los tres grupos tienen una duración de las canciones similar.

Retomando la **Hipótesis 4**, 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. Surgió entonces una nueva hipótesis:

7. Mientras menos **Instrumentales** tenga una canción, más popular es.

Para comprobar esta hipótesis se realizó 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, en consecuencia la hipótesis se debilitó. A pesar de esto, se decidió implementar regresión lineal simple con esta variable.

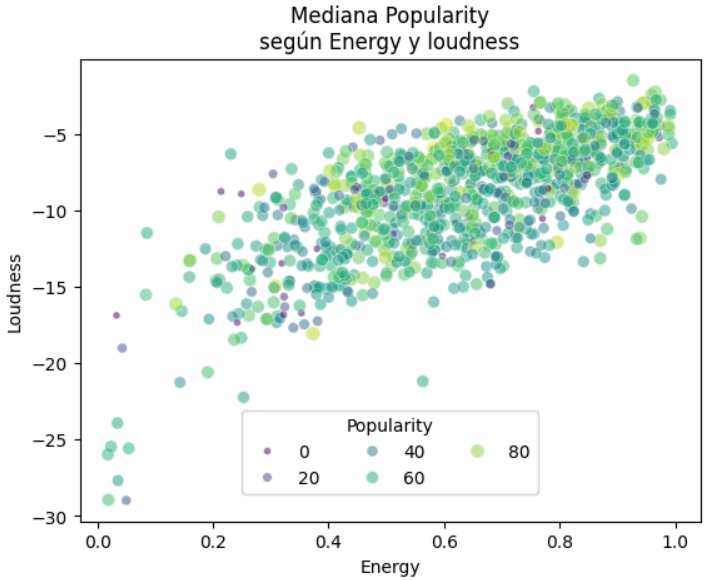
El resultado fue que Instrumentales explicaba solo el 3.4% de la varianza explicada, lo cual resultó muy bajo, implicando que el modelo no ajustaba adecuadamente los datos.

Se buscó verificar la distribución de los residuos para comprobar si el sumary arrojado tenía validez, dando como resultado el rechazo de la hipótesis nula, lo cual en conjunto al Histograma de distribución permitieron concluir que la distribución de los datos no era normal. De esta manera, queda completamente descartada la **Hipótesis 7**, siendo que Instrumentalness no explica la popularidad de la canción.

Sin embargo, no había sido contemplada hasta el momento la correlación fuerte entre Energy y Loudness visualizada en el mapa de calor. 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. De esta manera surgió la hipótesis:

8. Mientras más **Lodness** y **Energy** tenga una canción, más popular es.

El nuevo Gráfico de Dispersión mostró una leve tendencia de popularidad en la zona superior de la dispersión [Img 3]. Para ver en profundidad cómo afectaban estas variables a la popularidad se realizó Regresión Lineal Múltiple incluyendo a la anterior Loudness y Energy, esto permitió entender el impacto combinado de las tres características en la variable de interés Popularity.



[Img 3] Gráfico de dispersión Energy y Loudness vs Popularity

Luego de ajustar el modelo y observar su Summary, se llegó a la conclusión de que no es posible afirmar que las variables analizadas explican la popularidad. Esto fue porque el coeficiente de determinación arrojó un resultado de 5%, indicando que los datos no se ajustaron de manera precisa y las variables adicionales no aportaron información útil al modelo.

Antes de sacar conclusiones con este modelo, se debe corroborar que la distribución de los residuos es normal, para ello se observó los resultados de las variables Prob(Ómnibus) y Prob(JB), si ambos valores son cercanos a 0, los residuos no son normales, este es el caso por lo que no pudimos validar lo arrojado por el modelo.

Se descartó de esta manera la **Hipótesis 8**.

Finalmente, se agregaron al modelo todas las variables cuantitativas del dataset al modelo, para ver si así se puede explicar 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, provocando a su vez el descarte de la **Hipótesis 4**.

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/