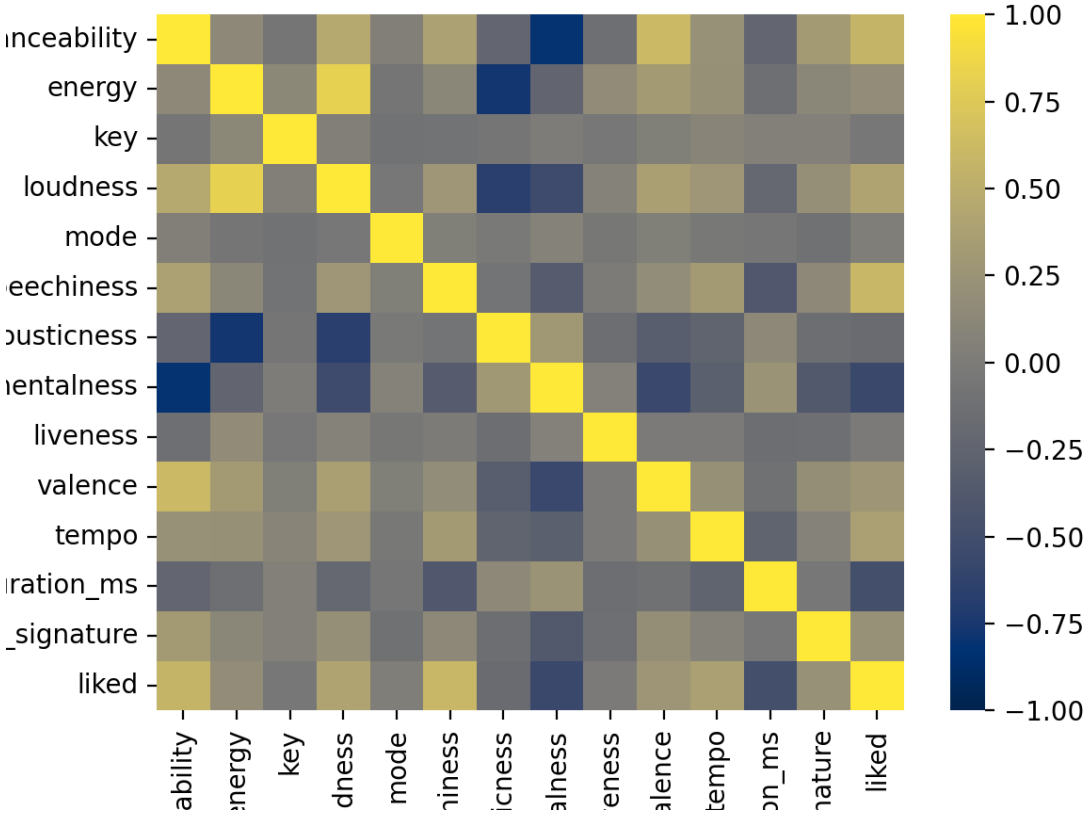
Actividad Evaluable: Patrones con K-means

el conjunto de datos usados en la actividad provienen de Spotify, específicamente son los diferentes valores que se le asignan a cada canción que existe dentro de las bases de datos de Spotify. Este mismo se obtuvo de el siguiente link
https://www.kaggle.com/bricevergnou/spotify-recommendation
Este dataset fue creado por un tercero usando el API de Spotify, por lo que las canciones existentes muestran la opinión del creador, y representa 100 canciones que le gustan y 95 que no.

En el dataset existen 195 datos, y 14 variables.
Estas variables son de los siguientes tipos:

danceability	float64
energy	float64
key	int64
loudness	float64
mode	int64
speechiness	float64
acousticness	float64
instrumentalness	float64
liveness	float64
valence	float64
tempo	float64
duration_ms	int64
time_signature	int64
liked	int64

Mapa de Calor



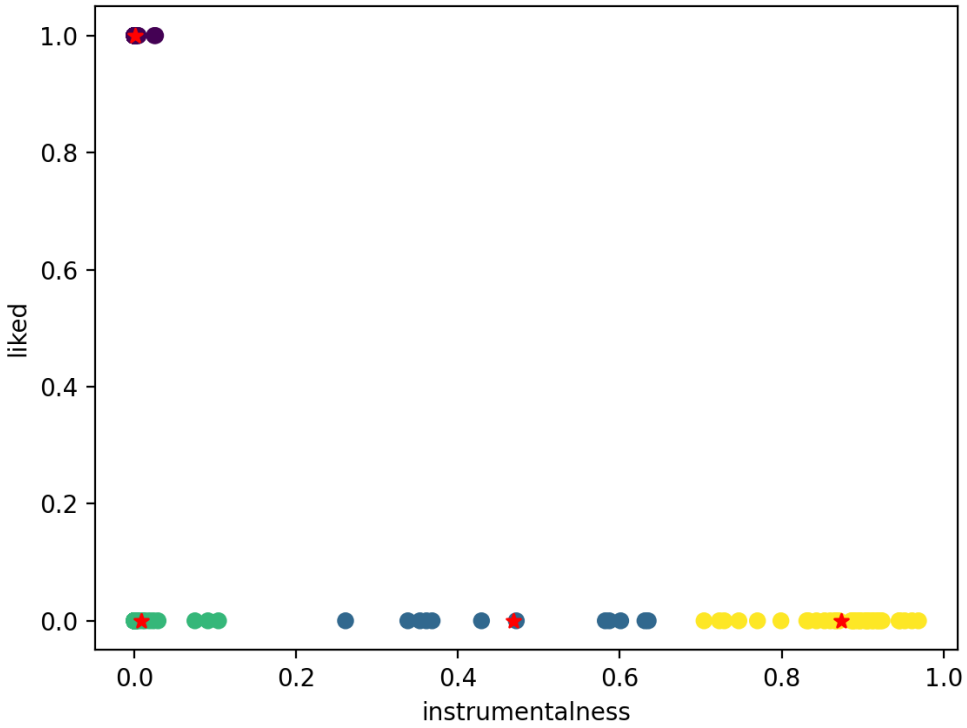
El mapa de calor muestra datos significativos en relación a el interés que el el creador del dataset tiene con cada variable. La última columna que muestra a la variable ‘liked’ sugiere que a el creador le interesa específicamente canciones que se puedan bailar, o en otras palabras con un índice de ‘danceability’ alto, pero no tan significativamente alto. Le agradan canciones con un alto grado de ‘speechiness’, y aquellas que tienen un ‘tempo’ alto.
Aún más interesante que esto son las variables que no le agradan al creador.
Le agradan canciones con un ‘duration_ms’ bajo, es decir de corta duración, le agradan canciones con poca instrumentación o ‘instrumentalness’.

Las variables anteriormente mencionadas son las que muestran más claramente el gusto que el creador del dataset tiene sobre las canciones.

Patrones con K-means

basándose en los valores obtenidos durante la creación del mapa de calor, la variables que afectan de manera más directa que tanto le gustan al creador del dataset fueron identificadas y en esta ocasión se eligió la variable “instrumentalness” para responder la siguiente pregunta, ¿qué tipo de instrumentación le agrada al creador del dataset?.

previo a la creación del gráfico se le asignó un 4 al valor de k, ya que al revisar visualmente el dataset y por acciones previas parecen existir cuatro diferentes grupos de valores, es decir cuatro diferentes géneros musicales ó por lo menos cuatro diferentes tipos de instrumentación en canciones. A continuación se muestra el gráfico generado.



Los resultados obtenidos son mucho más drásticos de lo que originalmente se había pensado, además de la gran diferencia entre canciones agradables y no agradables según el gráfico, se muestra que la instrumentación no es en su totalidad la variable que decida qué tan agradable es la canción para el creador, ya que existe un grupo de similar tamaño en este caso de color verde, el cual no es agradable. Finalmente la respuesta obtenida es que al creador del dataset le agrada la música con baja o cero instrumentación o ‘instrumentalness’.
Donde se agrupan la mayoría de los datos parece ser que los centros no son tan representativos como podrían haberlo sido, en el caso de los puntos de color azul, parece que hay dos sub-grupos que podrian tener sus propios centros, sin embargo, esto no afecta los resultados obtenidos ó los resultados a la pregunta planteada al inicio de la investigación. Si hubiera muchos outliers en el análisis de cajas y bigotes los centros y los datos agrupados con estos serían mucho menos reconocibles y posiblemente dificultaría el análisis de los datos encontrados ya que todo sería mucho menos objetivo.