Actividad Evaluable: Análisis final

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 oficial de Spotify, por lo que las canciones existentes muestran la opinión del creador, y representan 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

Análisis básico y familiarización con el dataset

De todas las variables existentes se decidió elegir la energía y la duración para hacer un análisis ligeramente más complejo.

Energía:

-El rango de los valores en la columna de energía es de 0.0024 - 0.996

-Basándose en la media, mediana, y desviación estándar el tipo de música elegida para el dataset varía en gran parte en cuanto a su energía, es probable que esto sea ya que el creador eligió música que no le gustaba, así que el gran rango muestra dos opuestos de estilo de música.

Duración:

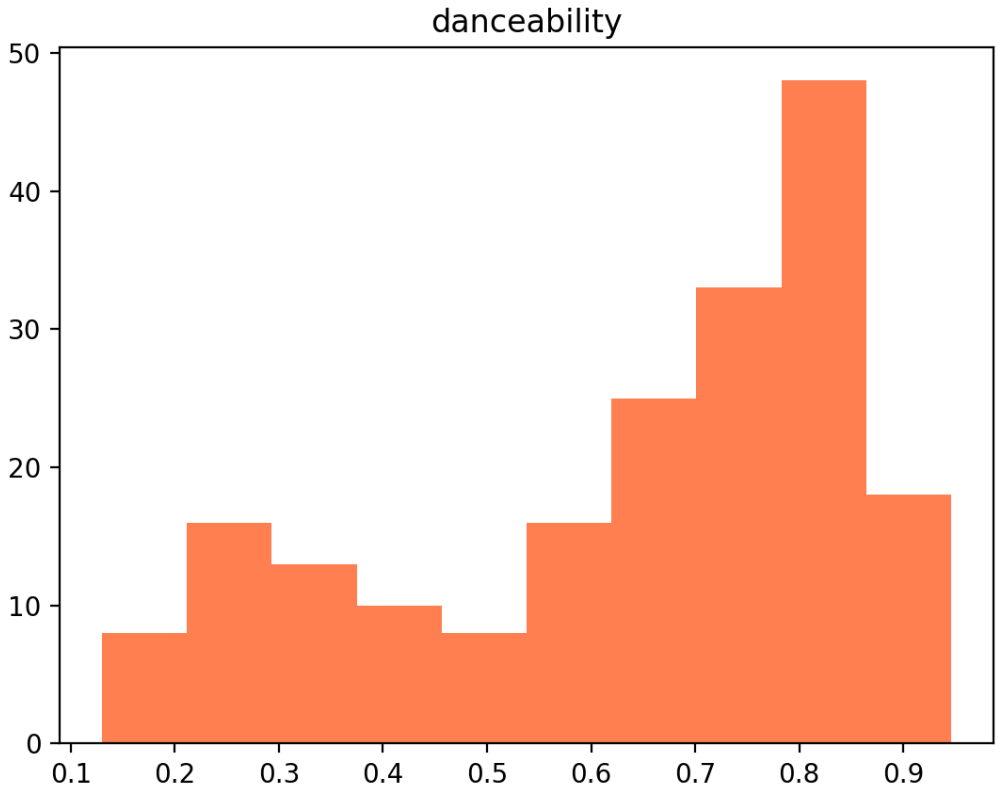
-El rango de los valores en la columna de duración es de 77203 - 655213

-Basándose en la duración, la cual está en milisegundos, la variación no es tan grande como la variable mencionada anteriormente, sin embargo, la diferencia en tiempos simplemente puede ser explicado por la gran diferencia de tiempos que las canciones tienden a tener, y no necesariamente significa algo más profundo. Con la información recolectada parece ser que la duración no afecta que tanto le gusta la canción al creador de el dataset.

Análisis de datos inicial e identificación de variables relevantes

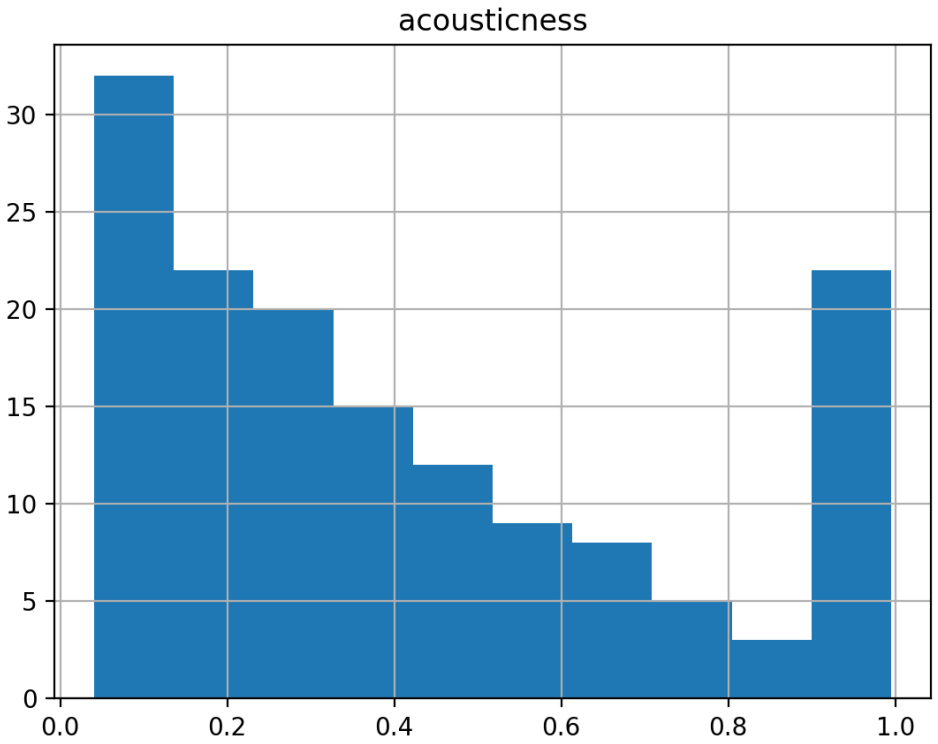
En esta sección se decidió utilizar las variables “danceability” y “acousticness” ya que son dos partes de la música muy relevantes al determinar si al creador de la lista le interesan las canciones y son partes importantes de toda la música en general.

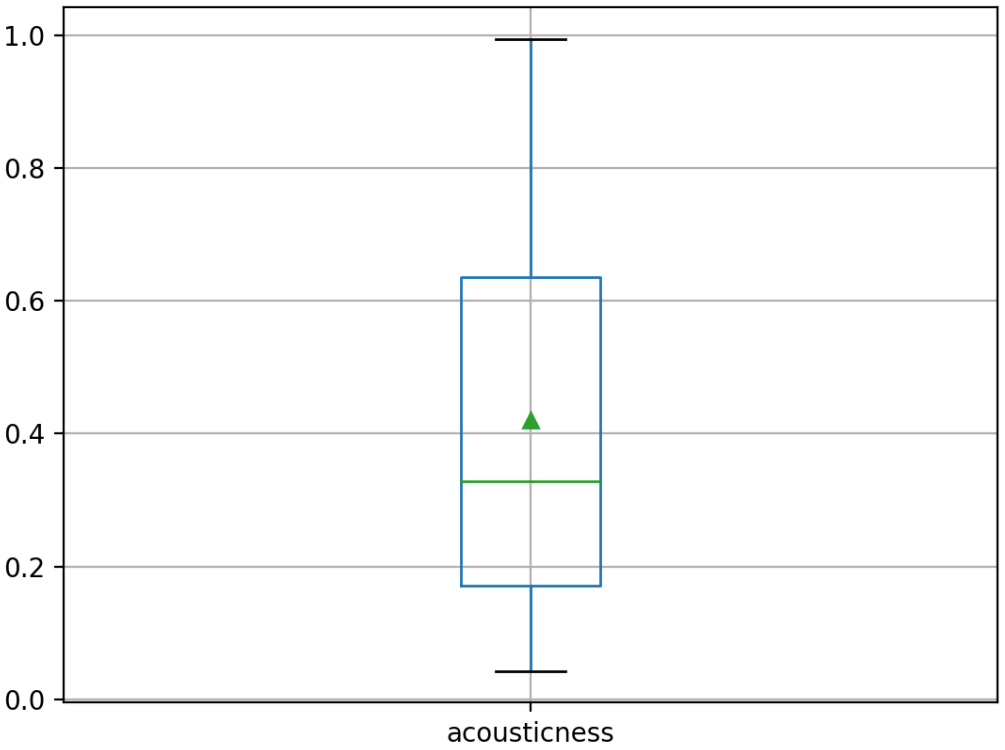
Danceability



las gráficas anteriores muestran el histograma y el diagrama de bigotes de la variable elegida, en este caso ‘danceability’, en esta variable existe una variación ligera lo cual es interesante especialmente notando que la mitad de las canciones son aquellas que el creador del dataset no le parecen atractivas. Esto sugiere que ‘danceability no afecta mucho la decisión del creador sobre su interés en la canción específica.

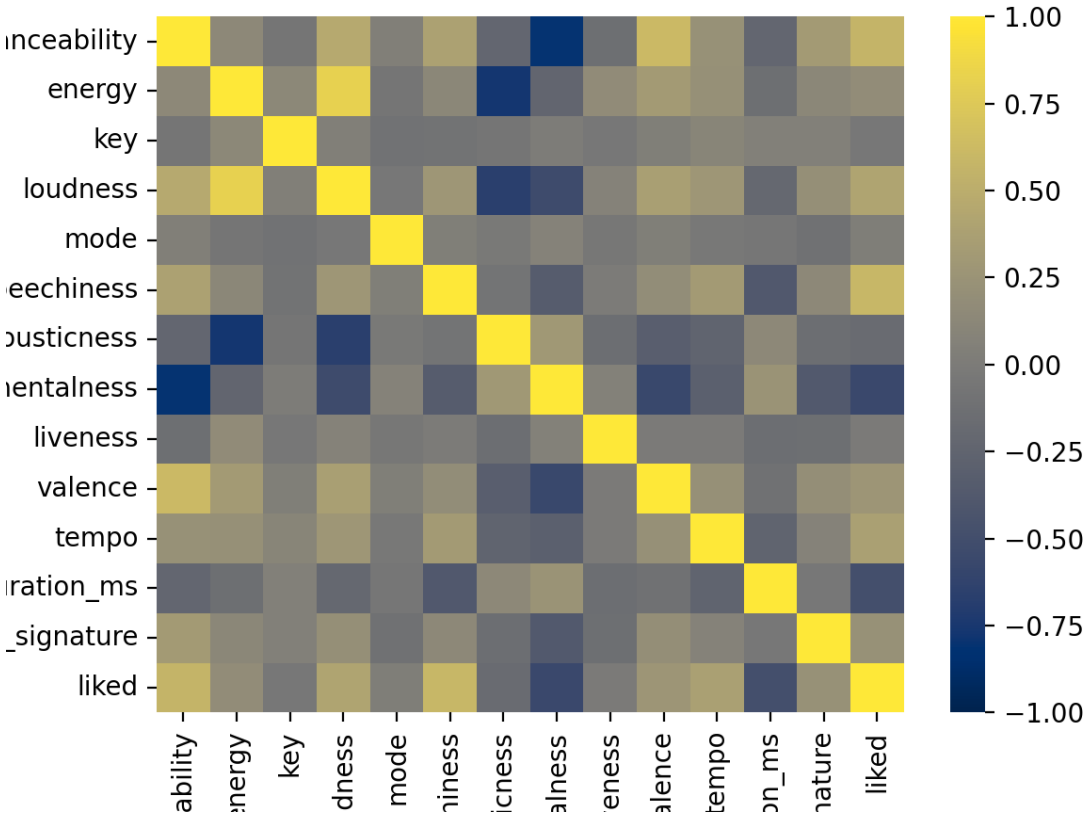
Acousticness





las gráficas anteriores muestran el histograma y el diagrama de bigotes de la variable elegida, en este caso ‘danceability’, esta variable a diferencia de la anterior muestra datos interesantes especialmente en el histograma, ya que dentro de la variación que existe se pueden notar dos diferentes barras con una gran cantidad de datos, tomando en cuenta la cantidad de canciones positivas y negativas que el autor del dataset eligió, esta variable es significativa al decidir si una canción le será de agrado al creador del dataset.

Mapa de Calor



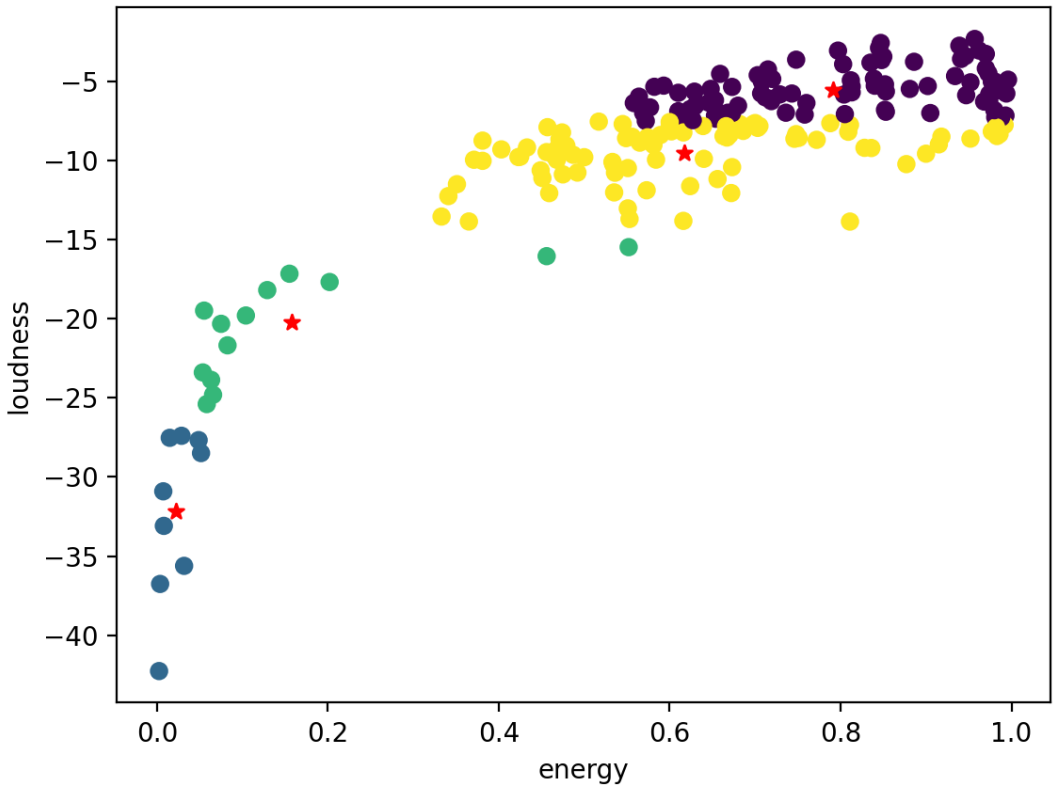
Por último, el mapa de calor muestra datos significativos en relación a el interés que el el creador del dataset tiene con cada variable. La última columna que muestra a la variable ‘liked’ sugiere que a el creador le interesa específicamente canciones que se puedan bailar, o en otras palabras con un 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.

Análisis de datos avanzado y derivación de conclusiones en base a resultados

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 “energy” y “loudness” para responder la siguiente pregunta, ¿qué tanto afecta el tempo de la canción a la energía?.

previo a la creación del gráfico se le asignó un 4 al valor de k, ya que se conoce que el creador del dataset eligió cuatro diferentes géneros musicales para agregar al dataset. Además de esto, es posible por simple inspección manual y gracias al pre procesamiento de los datos, ver que existen cuatro distintos rangos de valores.



Los resultados se apegan a las expectativas existentes, se pueden visualizar cuatro diferentes grupos, los grupos están divididos en dos secciones, con los azules y verdes teniendo baja energía y poco volumen. Al saber que el creador del dataset eligió ligeramente más canciones que le agradan, y al ver el gráfico anterior, es muy probable que las canciones con una energía alta sean las que le agradan más.
la respuesta a la pregunta planteada al inicio de esta investigación se puede fácilmente observar, hay una relación directa entre la energía de una canción y su volumen ó ‘loudness’.

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 verde, parece que hay dos valores se acercan más a el grupo amarillo; sin embargo, los resultados obtenidos ó los resultados a la pregunta planteada al inicio de la investigación no se ven afectados por esta variabilidad. 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.

Conclusión

En conclusión, los datos recabados son significativos y contienen un gran valor que nos permite obtener conclusiones no solo de los gustos del creador del dataset, si no de la música en general, como se hizo en la sección previa. Los análisis visuales apoyado de gráficos y ciencia de datos, junto con el análisis visual de los datos iniciales es sumamente interesante y un reporte de una magnitud mucho mayor analizando la música se podría crear fácilmente.Se ha podido resolver todas las preguntas y dudas planteadas a lo largo de la actividad e incluso se ha generando nuevas preguntas.