

Hits don't lie

¿Comparten características
las canciones más populares?

Resumen

Se analizarán las características musicales de las canciones y la relación que estas tienen con su popularidad en una base de datos de Spotify.


Introducción

La música es, además de un arte, un negocio, y Spotify y las demás compañías de streaming utilizan millones de datos en sus operaciones diarias. Algunos de estos datos pueden obtenerse en la API de Spotify para utilizarlos para nuevas aplicaciones o analizarlos. Claro está, los datos se presentan con restricciones, tanto en el tipo de dato que se ofrece como en las limitaciones de cantidad de registros.

La popularidad de las canciones tiene muchos factores, desde la popularidad anterior del artista, hasta cómo ha sido promocionada o publicitada, pero muchos de los hits de nuestra cultura parecen tener características en común.

Con los datos que contiene el dataset en el que se centra este estudio, mi hipótesis es la siguiente: *existen características musicales comunes a las canciones más populares de la base de datos.*

Metodología

Se descargó el dataset  Spotify Tracks Dataset de Kaggle.com, que estaba compuesto de 114 000 registros y 21 columnas en un archivo CSV. El dataset se limpió y analizó con Python, ayudado de bibliotecas para el análisis y la visualización de datos como Pandas, Matplotlib y Seaborn.

Se comienza aplicando métodos para observar la composición del dataset, la información que contiene, cómo está distribuida y los valores estadísticos básicos.

1 of .sort_values(by='popularity', ascending=False).head(25)																				
✓ 5/5																				
	track_id	artists	album_name	track_name	popularity	duration_ms	explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	track_genre
20001	3nqQ0yQOWXESLDFHNG	Sam Smith/Kim Petras	Unholy (feat. Kim Petras)	Unholy (feat. Kim Petras)	100	156943	False	0.714	0.472	2	-7.375	1	0.0864	0.01300	0.000005	0.2660	0.238	131.121	4	dance
81051	3nqQ0yQOWXESLDFHNG	Sam Smith/Kim Petras	Unholy (feat. Kim Petras)	Unholy (feat. Kim Petras)	100	156943	False	0.714	0.472	2	-7.375	1	0.0864	0.01300	0.000005	0.2660	0.238	131.121	4	pop
51664	2rTmW7RDmQ2Bk7m7e5w	Bizarrap/Quevedo	Quevedo: Bizarrap Music Sessions, Vol. 52	Quevedo: Bizarrap Music Sessions, Vol. 52	99	198937	False	0.621	0.782	2	-5.548	1	0.0440	0.01250	0.033000	0.2300	0.550	128.033	4	hip-hop
88410	5wm28F9dyIghVOK37BIC4u	Manuel Turizo	La Bachata	La Bachata	98	162637	False	0.835	0.679	7	-5.329	0	0.0364	0.58300	0.000002	0.2180	0.850	124.980	4	reggae
30003	4uJG5SRvCk48mYERFj3cK	David Guetta/Bebe Rexha	I'm Good (Blue)	I'm Good (Blue)	98	175238	True	0.561	0.965	7	-3.673	0	0.0343	0.00383	0.000007	0.3710	0.304	128.040	4	edm
68303	5wm28F9dyIghVOK37BIC4u	Manuel Turizo	La Bachata	La Bachata	98	162637	False	0.835	0.679	7	-5.329	0	0.0364	0.58300	0.000002	0.2180	0.850	124.980	4	latino
89411	5wm28F9dyIghVOK37BIC4u	Manuel Turizo	La Bachata	La Bachata	98	162637	False	0.835	0.679	7	-5.329	0	0.0364	0.58300	0.000002	0.2180	0.850	124.980	4	reggaeton
81210	4uJG5SRvCk48mYERFj3cK	David Guetta/Bebe Rexha	I'm Good (Blue)	I'm Good (Blue)	98	175238	True	0.561	0.965	7	-3.673	0	0.0343	0.00383	0.000007	0.3710	0.304	128.040	4	pop
67356	5wm28F9dyIghVOK37BIC4u	Manuel Turizo	La Bachata	La Bachata	98	162637	False	0.835	0.679	7	-5.329	0	0.0364	0.58300	0.000002	0.2180	0.850	124.980	4	latin
20006	4uJG5SRvCk48mYERFj3cK	David Guetta/Bebe Rexha	I'm Good (Blue)	I'm Good (Blue)	98	175238	True	0.561	0.965	7	-3.673	0	0.0343	0.00383	0.000007	0.3710	0.304	128.040	4	dance
68304	1HhW5LamUGeU4ozQ2SXZ	Bad Bunny	Un Verano Sin Ti	Tií Me Preguntó	97	243716	False	0.650	0.715	5	-5.198	0	0.2530	0.09930	0.000291	0.1260	0.187	106.672	4	latino
89407	65q78F9Qa75NF8vV5C9g	Bad Bunny/Chencho Corleone	Un Verano Sin Ti	Me Porto Bonito	97	178567	True	0.911	0.712	1	-5.105	0	0.0817	0.09010	0.000027	0.0933	0.425	92.005	4	reggaeton
68305	65q78F9Qa75NF8vV5C9g	Bad Bunny/Chencho Corleone	Un Verano Sin Ti	Me Porto Bonito	97	178567	True	0.911	0.712	1	-5.105	0	0.0817	0.09010	0.000027	0.0933	0.425	92.005	4	latino
67358	65q78F9Qa75NF8vV5C9g	Bad Bunny/Chencho Corleone	Un Verano Sin Ti	Me Porto Bonito	97	178567	True	0.911	0.712	1	-5.105	0	0.0817	0.09010	0.000027	0.0933	0.425	92.005	4	latin
67359	1HhW5LamUGeU4ozQ2SXZ	Bad Bunny	Un Verano Sin Ti	Tií Me Preguntó	97	243716	False	0.650	0.715	5	-5.198	0	0.2530	0.09930	0.000291	0.1260	0.187	106.672	4	latin
89405	65q78F9Qa75NF8vV5C9g	Bad Bunny/Chencho Corleone	Un Verano Sin Ti	Me Porto Bonito	97	178567	True	0.911	0.712	1	-5.105	0	0.0817	0.09010	0.000027	0.0933	0.425	92.005	4	reggae
89405	1HhW5LamUGeU4ozQ2SXZ	Bad Bunny	Un Verano Sin Ti	Tií Me Preguntó	97	243716	False	0.650	0.715	5	-5.198	0	0.2530	0.09930	0.000291	0.1260	0.187	106.672	4	reggaeton
89407	1HhW5LamUGeU4ozQ2SXZ	Bad Bunny	Un Verano Sin Ti	Tií Me Preguntó	97	243716	False	0.650	0.715	5	-5.198	0	0.2530	0.09930	0.000291	0.1260	0.187	106.672	4	reggae

La columna *track_genre* elimina, ya que esta variable no está ligada al track, sino al artista, por lo que los registros se duplican. Por ejemplo, un mismo artista puede estar catalogado como pop y dance, por lo que cada vez que sus canciones aparezcan en el dataset aparecerán dos veces, con todos los datos idénticos excepto por el género.

Variables del dataset

Spotify obtiene estos valores realizando un análisis de audio.

track_id - *artists* - *album_name* - *track_name*

Estas cuatro variables son objects (strings) y sirven para identificar el track y a su artista.

popularity - entero, 0 a 100

Es un valor entero que va de 0 a 100, siendo el 100 el más popular. La popularidad “es calculada por un algoritmo y se basa, mayormente, en la cantidad de reproducciones que ha tenido el track y qué tan recientes fueron esas reproducciones” (*Web API Reference*, n.d.)

duration_ms - float

La duración del track en milisegundos.

explicit - booleana

True representa que la letra de la canción es explícita, False, que no lo es (o se desconoce).

key - entero, -1 a 11

La escala musical en la que está la canción. Va del 0 al 11, comenzando por do. Si no se detectó la escala, se representa con -1.

loudness - float, -60 a 0

El volumen general de un track en decibelios. Se promedia en todo el track y es útil para comparar de forma relativa qué tan fuerte es el volumen un track. Típicamente tiene un rango de -60 a 0.

mode - entero, 1 o 0

Indica si el track está en modo mayor o menor. Mayor es 1 y menor es 0.

tempo - float

El tempo en pulsaciones por minuto (BPM). Es la velocidad o el ritmo que tiene una canción.

time_signature - entero, 3 a 7

El compás estimado. Cuántos tiempos o pulsos tiene cada compás. Va de 3 a 7, que indica compases de 3/4 a 7/4.

danceability - float, 0 a 1

Qué tanailable es un track.

energy - float, 0 a 1

Representa intensidad y actividad. Los tracks energéticos son rápidos y ruidosos (el death metal, por ejemplo).

speechiness - float, 0 a 1

Detecta la presencia de discurso hablado en el track. Cuanto más se acerque el track a una grabación de voz (como un pódcast o un audiolibro), más cerca de 1.

acousticness - float, 0 a 1

Cuanto más cerca de 1, más probable que el track sea acústico.

instrumentalness - float, 0 a 1

Predice si un track no incluye voces. Cuanto más cerca de 1, más probable es que sea instrumental.

liveness - float, 0 a 1

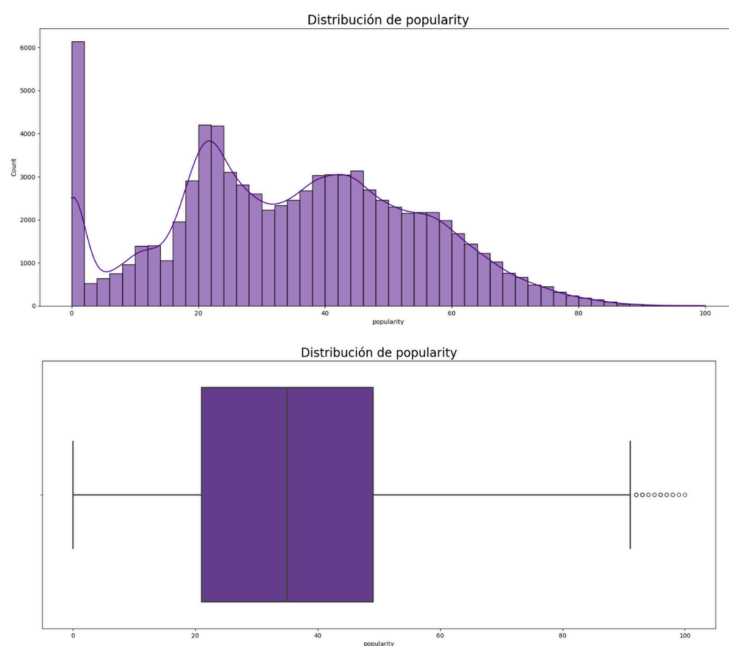
Detecta la presencia de público en la grabación.

valence - float, 0 a 1

Describe qué tan positivo y alegre es un track.

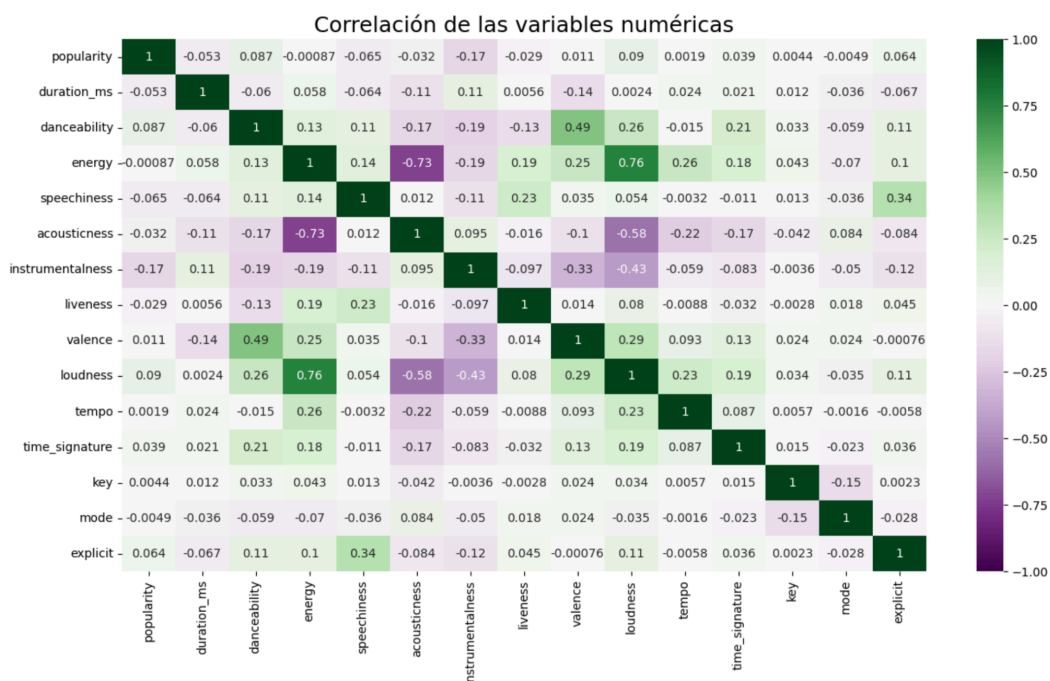
Análisis exploratorio de datos

Se analiza la distribución de la variable *popularity*, ya que es central para este estudio.



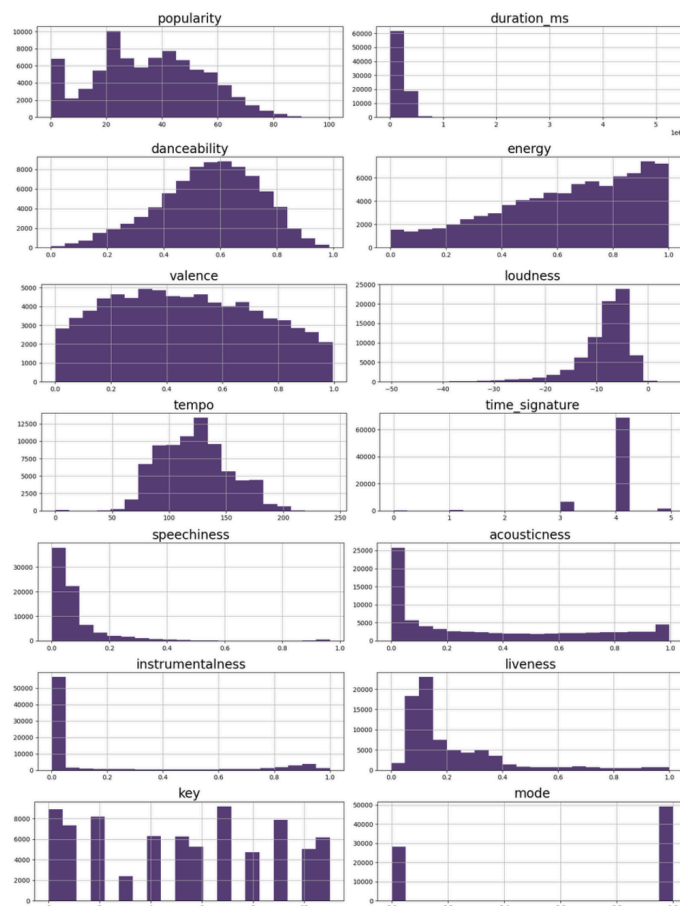
Aunque hay un gran sesgo a la izquierda debido a muchos valores 0, esto tiene sentido por cómo se asigna el puntaje de popularidad, por lo tanto, los outliers eran esperables y se conservarán.

Se realiza una matriz de correlación. La variable *popularity* no parece tener correlación con las demás, casi todos los valores tienden a 0.

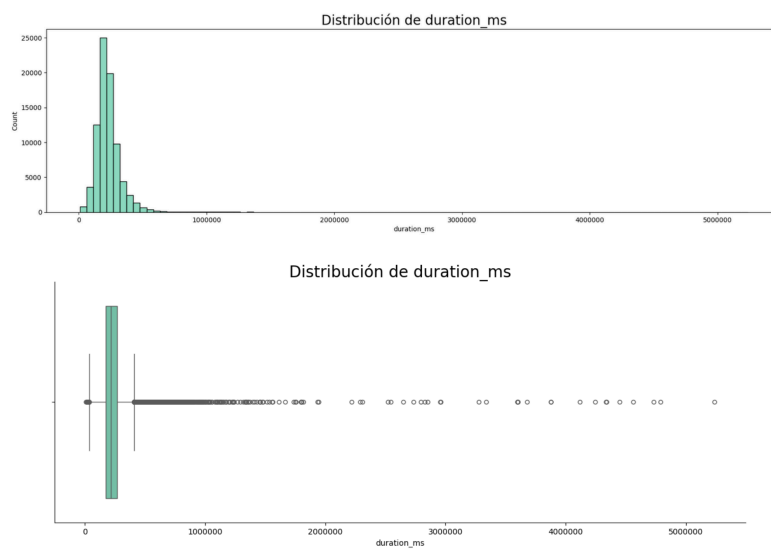


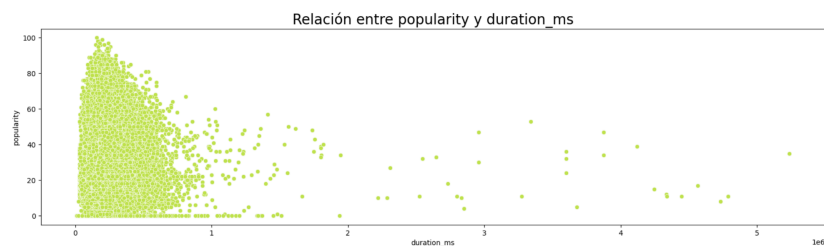
Sí se observan una correlación entre la variable *loudness* y *energy* (lo que tiene sentido, son el volumen y la energía o intensidad de un track), y una correlación inversa entre *energy* y *acousticness* (si el track es acústico, tiene menos energía). Otras correlaciones menos pronunciadas ocurren entre variables que también se pensaría que están relacionadas: *valence* (cuando más alto el valor, más positivo es el track) y *danceability* (cuando más alto, más bailable), entre otras.

Se realizan histogramas de todas las variables numéricas para observar la distribución de sus valores:

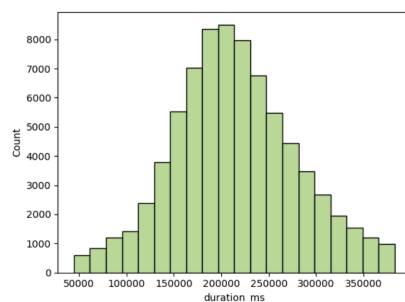


Luego de visualizar la distribución de todas las variables, se aprecia que *duration_ms* tiene una distribución sesgada a la izquierda y se constata que hay outliers que podrían distorsionar el análisis.

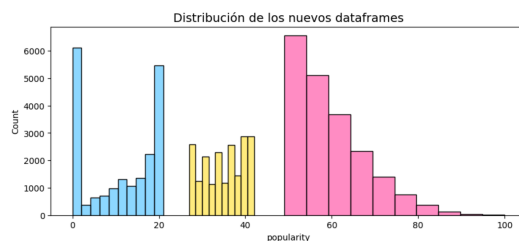




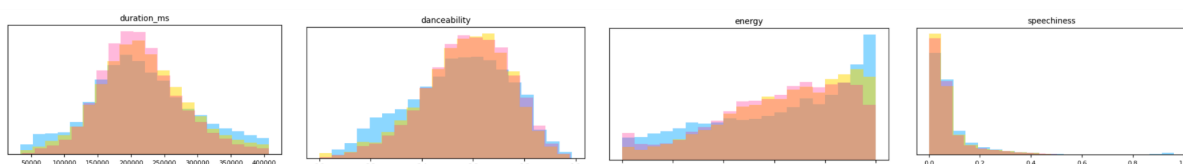
Como no se observa una relación entre las variables *popularity* y *duration_ms*, se eliminan los outliers y la nueva distribución de *duration_ms* es normal y las duraciones de los tracks son más ajustadas a las duraciones que por lo general tienen las canciones. Duración media: 3:36, duración mínima: 0:33, duración máxima: 6:47.

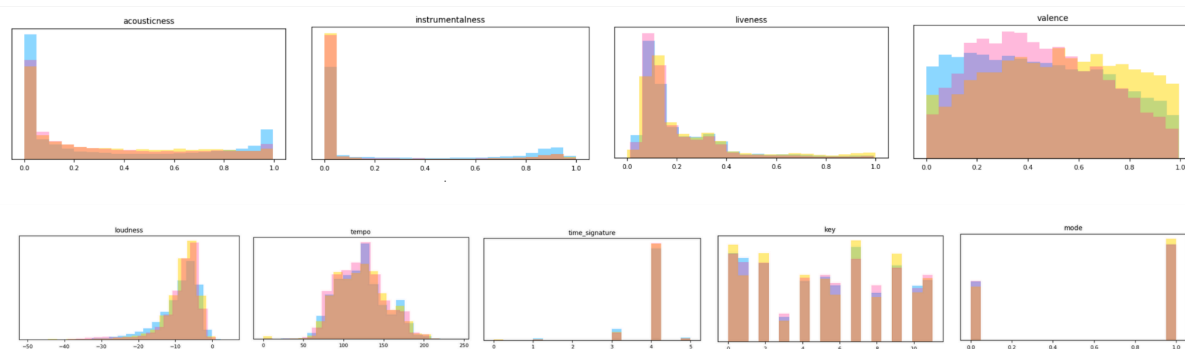


Para analizar la distribución de los valores de las variables de acuerdo a la popularidad, se segmenta el dataframe en tres: el cuantil más alto, el más bajo, y un segmento con la misma cantidad de valores alrededor de la mediana. De esta forma, se pueden comparar las características de las canciones más populares con las menos y detectar diferencias en las demás variables.

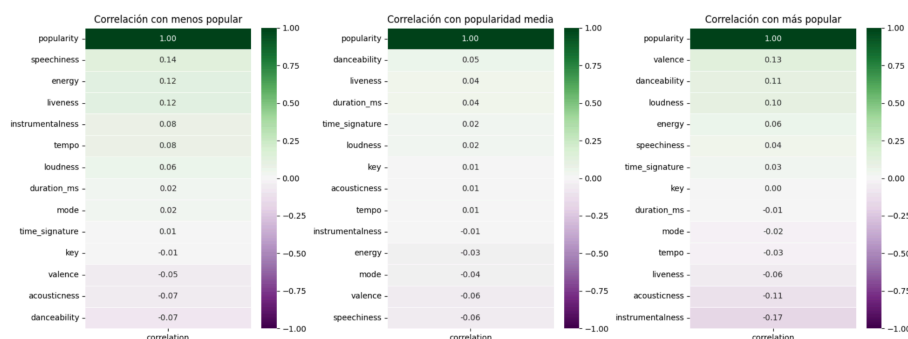


Se vuelven a graficar todas las variables superponiendo los tres dataframes segmentados, y aunque se observan algunas leves diferencias, no es suficiente para decir que las canciones más populares tengan una u otra característica.





Se vuelven a graficar las correlaciones de *popularity* y las demás variables en los dataframes segmentados y no se observa ninguna correlación.



Resultados

La hipótesis no puede comprobarse. Las correlaciones de las variables numéricas con *popularity* tienden a 0 y no presentan valores significativos que permitan sacar conclusiones sobre la influencia de una variable sobre otra. Las características que comparten las canciones más populares no son significativamente diferentes a las que comparten las canciones menos populares.

Discusión

La metodología que se utilizó para crear el dataset no parece haber sido la más acertada, ya que al primero determinar un listado de géneros y luego descargar 1000 tracks de cada uno de ellos, queda sesgada la muestra, pues si nos basáramos solo en la popularidad de las canciones, seguramente la cantidad de canciones por género sería más variada.

Por otro lado, podría considerarse un beneficio que la cantidad de tracks no esté limitada a la popularidad del género, ya que podría darse el caso de que hubiera muchas más canciones de pop o reggaeton, dos géneros populares, y no sabríamos con certeza si estamos analizando las características de las canciones populares o las de esos géneros que tienen más representación en la muestra.

En un futuro, podrían descargarse los datos directamente de la API de Spotify y volver a realizar este análisis con un dataset con información más completa y menos sesgada por el género musical.

Conclusión

El análisis del dataset arroja que la hipótesis no se confirma y, si bien hay características que sí comparten algunas canciones, no puede afirmarse que esto esté relacionado con la popularidad de estas.

Referencias bibliográficas

Spotify Tracks Dataset. (n.d.). Kaggle. Consultada el 3 de febrero de 2025, en

<https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset>

Web API Reference. (n.d.). Web API Reference | Spotify for Developers. Consultada el 3 de febrero de 2025, en

<https://developer.spotify.com/documentation/web-api/reference/get-several-tracks>