# Libraries

### Librerías necesarias para ejecutar los bloques de códigos

# Exploration Data Initial

## Information Data Initial

| | Unnamed: 0 int64<br>0 - 9 | track_id object<br>5SuOikwiRy... 10%<br>4qPNDBW1i... 10%<br>8 others 80% | artists object<br>Gen Hoshino 10%<br>Ben Woodw... 10%<br>8 others 80% | album_name obj...<br>We Sing. We ... 20%<br>Comedy 10%<br>7 others 70% | track_name object<br>Comedy 10%<br>Ghost - Aco... 10%<br>8 others 80% | popularity int64<br>55 - 82 | duratic<br>149610 |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 5SuOikwiRyPMVo... | Gen Hoshino | Comedy | Comedy | 73 | |
| 1 | 1 | 4qPNDBW1i3p13... | Ben Woodward | Ghost (Acoustic) | Ghost - Acoustic | 55 | |
| 2 | 2 | 1iJBSr7s7jYXzM8... | Ingrid Michaelson... | To Begin Again | To Begin Again | 57 | |
| 3 | 3 | 6lfxq3CG4xtTiEg... | Kina Grannis | Crazy Rich Asian... | Can't Help Falling... | 71 | |
| 4 | 4 | 5vjLSffimiIP26QG... | Chord Overstreet | Hold On | Hold On | 82 | |
| 5 | 5 | 01MVOl9KtVTNfF... | Tyrone Wells | Days I Will Reme... | Days I Will Reme... | 58 | |
| 6 | 6 | 6Vc5wAMmXdKIA... | A Great Big World... | Is There Anybody... | Say Something | 74 | |
| 7 | 7 | 1EzrEOXmMH3G... | Jason Mraz | We Sing. We Dan... | I'm Yours | 80 | |
| 8 | 8 | 0IktbUcnAGrvD0... | Jason Mraz;Colbi... | We Sing. We Dan... | Lucky | 74 | |
| 9 | 9 | 7k9GuJYLp2Azq... | Ross Copperman | Hunger | Hunger | 56 | |

10 rows, 21 cols 10 / page Page 1 of 1

## Información del dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113999 entries, 0 to 113998
Data columns (total 21 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        113999 non-null  int64
 1   track_id          113999 non-null  object
 2   artists           113999 non-null  object
 3   album_name        113999 non-null  object
 4   track_name        113999 non-null  object
 5   popularity        113999 non-null  int64
 6   duration_ms       113999 non-null  int64
 7   explicit          113999 non-null  bool
 8   danceability      113999 non-null  float64
 9   energy            113999 non-null  float64
 10  key               113999 non-null  int64
 11  loudness          113999 non-null  float64
 12  mode              113999 non-null  int64
 13  speechiness       113999 non-null  float64
 14  acousticness      113999 non-null  float64
 15  instrumentalness  113999 non-null  float64
 16  liveness          113999 non-null  float64
 17  valence           113999 non-null  float64
 18  tempo             113999 non-null  float64
 19  time_signature    113999 non-null  int64
 20  track_genre       113999 non-null  object
dtypes: bool(1), float64(9), int64(6), object(5)
memory usage: 17.5+ MB
```

## Estadística descriptiva del DataSet

| | Unnamed: 0 float... | popularity float64 | duration_ms floa... | danceability float... | energy float64 | key float64 | loudne |
|---|---|---|---|---|---|---|---|
| cou... | 113999 | 113999 | 113999 | 113999 | 113999 | 113999 | |
| me... | 56999.42192 | 33.23882666 | 228031.1534 | 0.566800643 | 0.6413832705 | 5.309125519 | -8 |
| std | 32909.24346 | 22.30495908 | 107296.0577 | 0.1735428253 | 0.2515301126 | 3.559999216 | 5 |
| min | 0 | 0 | 8586 | 0 | 0 | 0 | |
| 25% | 28499.5 | 17 | 174066 | 0.456 | 0.472 | 2 | |
| 50% | 56999 | 35 | 212906 | 0.58 | 0.685 | 5 | |
| 75% | 85499.5 | 50 | 261506 | 0.695 | 0.854 | 8 | |
| max | 113999 | 100 | 5237295 | 0.985 | 1 | 11 | |

8 rows, 15 cols 10 / page Page 1 of 1

## Total de filas y columnas

```
Los datos de Spotify contienen un total de 113999 filas y 21 columnas
```

```
De los cuales los datos de Spotify, el total de canciones es 89740 y el total de generos es 114
```

## Tamaño del DataSet

```
(113999, 21)
```

## Columnas del DataSet

```
Index(['Unnamed: 0', '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'],
      dtype='object')
```

## Tipo de dato de las columnas

```
Unnamed: 0            int64
track_id             object
artists              object
album_name           object
track_name           object
popularity            int64
duration_ms           int64
explicit               bool
danceability        float64
energy              float64
key                   int64
loudness            float64
mode                  int64
speechiness         float64
acousticness        float64
instrumentalness    float64
liveness            float64
valence             float64
tempo               float64
time_signature        int64
track_genre          object
dtype: object
```

## Cantidad del total de datos

```
float64    9
int64      6
object     5
bool       1
Name: count, dtype: int64
```

## ¿Existen valores nulos?

```
Unnamed: 0          False
track_id            False
artists             False
album_name          False
track_name          False
popularity          False
duration_ms         False
explicit            False
danceability        False
energy              False
key                 False
loudness            False
mode                False
speechiness         False
acousticness        False
instrumentalness    False
liveness            False
valence             False
tempo               False
time_signature      False
track_genre         False
dtype: bool
```

```
Unnamed: 0          0
track_id            0
artists             0
album_name          0
track_name          0
popularity          0
duration_ms         0
explicit            0
danceability        0
energy              0
key                 0
loudness            0
mode                0
speechiness         0
acousticness        0
instrumentalness    0
liveness            0
valence             0
tempo               0
time_signature      0
track_genre         0
dtype: int64
```

## Revisión de fila con datos nulos

| | **Unnamed: 0** int64 | **track_id** object | **artists** object | **album_name** obj... | **track_name** object | **popularity** int64 | **duratic** |
|---|---|---|---|---|---|---|---|

0 rows, 21 cols 100 / page Page 0 of 0

## Eliminación de fila con datos nulos

### Revisión del tamaño del DataSet

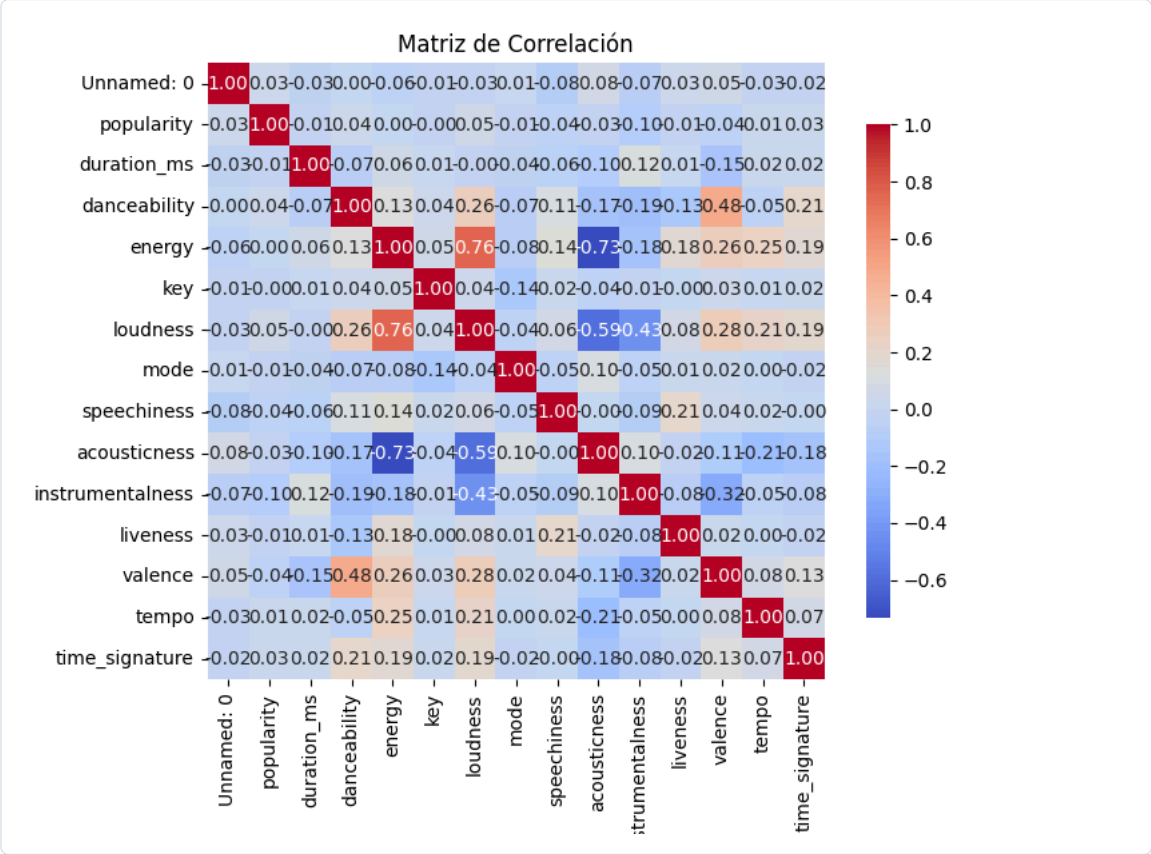
```
(113999, 21)
```

## Valores Duplicados

Revisión de valores duplicados

```
False    113999
Name: count, dtype: int64
```

### Correlation Matrix




### Matriz:

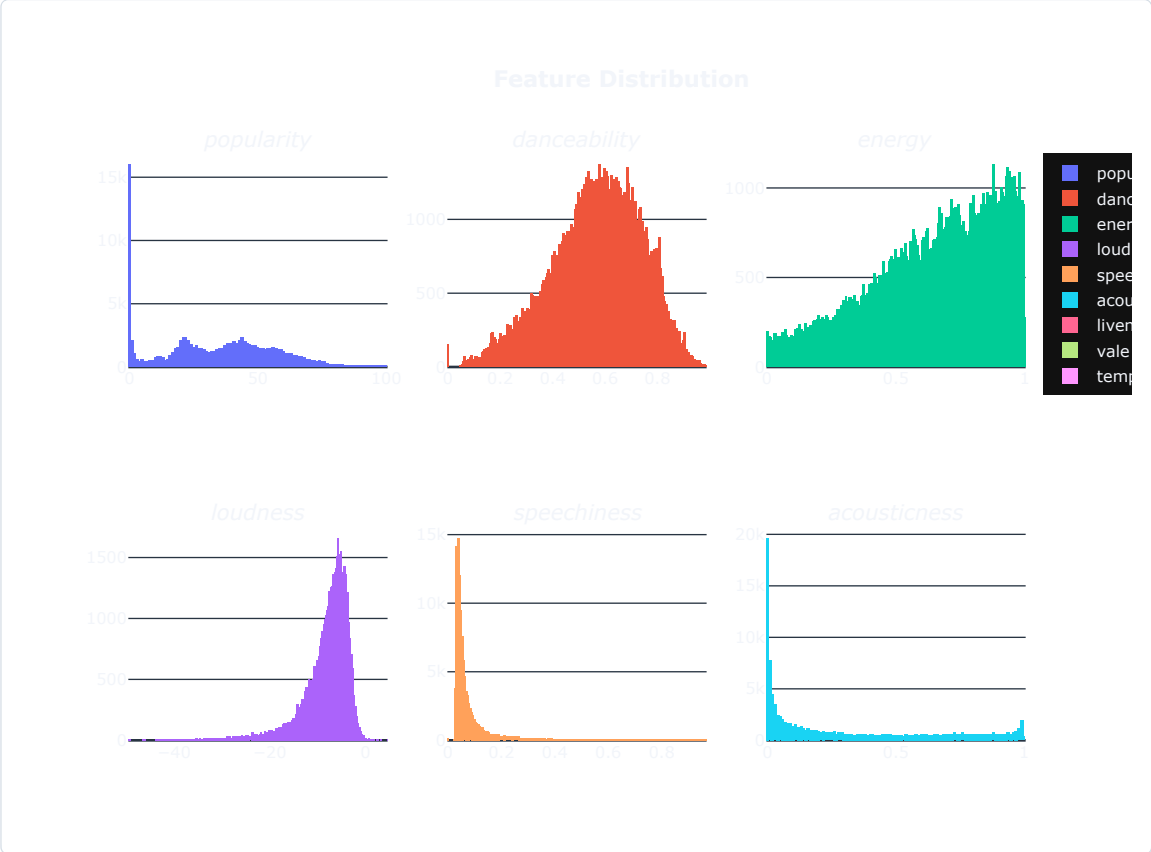
La correlación mostrada sirve para demostrar que algunos campos del dataset son correlación positiva (1) o correlación negativa (-1), en caso de tener valor 0, no hay correlación lineal entre las variables:

Para ello se utilizaron solo los campos de valores númericos, en donde las celdas rojas indican que existe una correlación positiva y las celdas azules lo contrario.

1) El campo 'Danceability' tiene una fuerte correlación positiva con un valence de 0.48, lo que sugiere que el campo 'valence' esta implicitamente relacionado.

2) El campo 'energy' tiene una fuerte correlación negativa con un valence de -0.73, lo que sugiere que el campo 'acousticness' esta directamente relacionado.

## Histogramas




### El histograma mostrada sirve para mostrar la distribución de los campos mas valiosos del Dataset, donde se ha sacado las siguientes concluciones:

1) 'Popularity' tiene la distribución sesgada hacia el lado izquierdo, lo que indica que la mayoria de las canciones del Dataset tienen una popularidad baja.

2) 'Danceability' tiene una distribución normal (Forma de campana) con la mayoria de las canciones, esto sugiere que la mayoría de las canciones son bailables.

3) 'Energy' tiene una distribución asimétrica hacia el lado derecho, lo que indica que la mayoria de las canciones tienen valores altos de energia y por lo tanto, menos canciones con energia baja.

4) 'Loudness' tiene una distribución alta hacia el lado derecho, lo que indicaria que la mayoria de canciones son fuertes en terminos de volumen.

5) 'Speechiness' tiene una distribución alta hacia el lado izquierdo, lo que indica que la mayoria de canciones no tienen contenido hablado.

6) 'Acousticness' esta distribución similar al anterior indicaria que la mayoria de canciones tienen bajos niveles acusticos (Es decir, producido digitalmente).

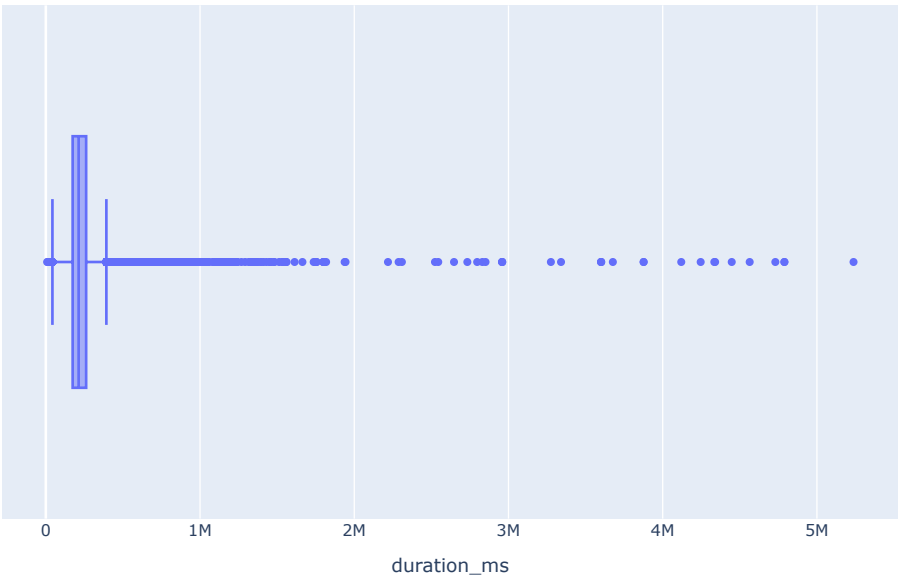
7) 'Liveness' esta distribución con valores altos en el lado izquierdo sugiere que la mayoria de las canciones tienen poca presencia de sonido en vivo.

8) 'Valence' esta distribución es bastante uniforme, lo que indica que la muestra de emociones abarca una amplia gama transmitida en las canciones.

9) 'Tempo' esta distribución destacada por tener varios picos (principalmente entre 120-130 BPM) indicaria que el rango común entre generos populares como el pop o el rock abarcados por este tempo.

## Outlinders

Busqueda de outliders en la duración de las canciones




El siguiente diagrama de caja enfocado para la variable 'duration_ms' ayudara para resumir la distribución del conjunto de datos.

Para esto se explicara los conceptos principales entregados por el diagrama:

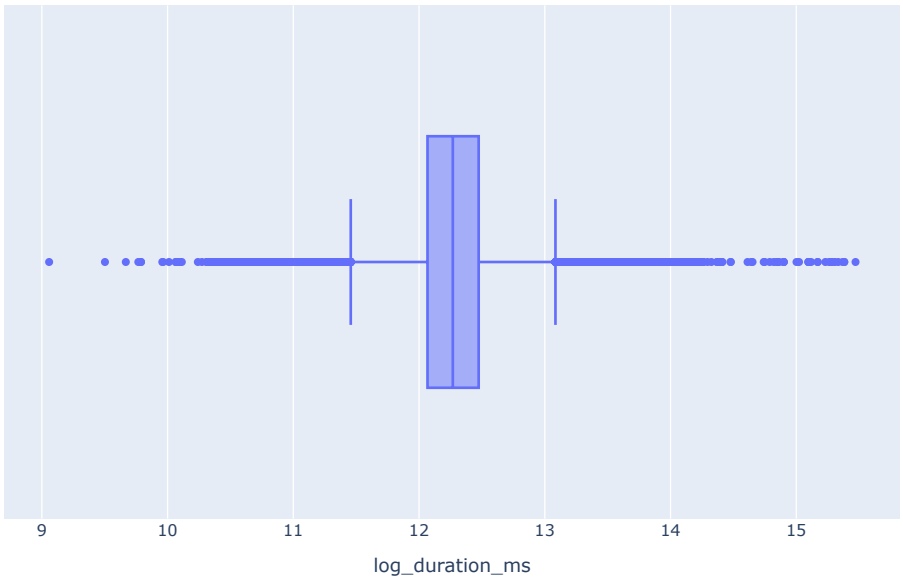
1) Mediana: El valor que divide los datos en 2 mitados, representando el punto medio de duración de las canciones el cual seria de '212.906' que traducido a minutos y segundos seria de '3 minutos y 33 segundos'.

2) Quartiles:
a) El borde inferior (Q1) muestra que la duración de las canciones del 25% del dataset traducidos a minutos y segundos seria de '2 minutos y 54 segundos'.

b) El borde superior (Q3) muestra que la duración de las canciones del 75% del dataset traducidos a minutos y segundos seria de '4 minutos y 21 segundos'.

3) Outliers: Representaria valores atípicos que son mucho mas largos que la duración promedio.

Transformación log de duration_ms

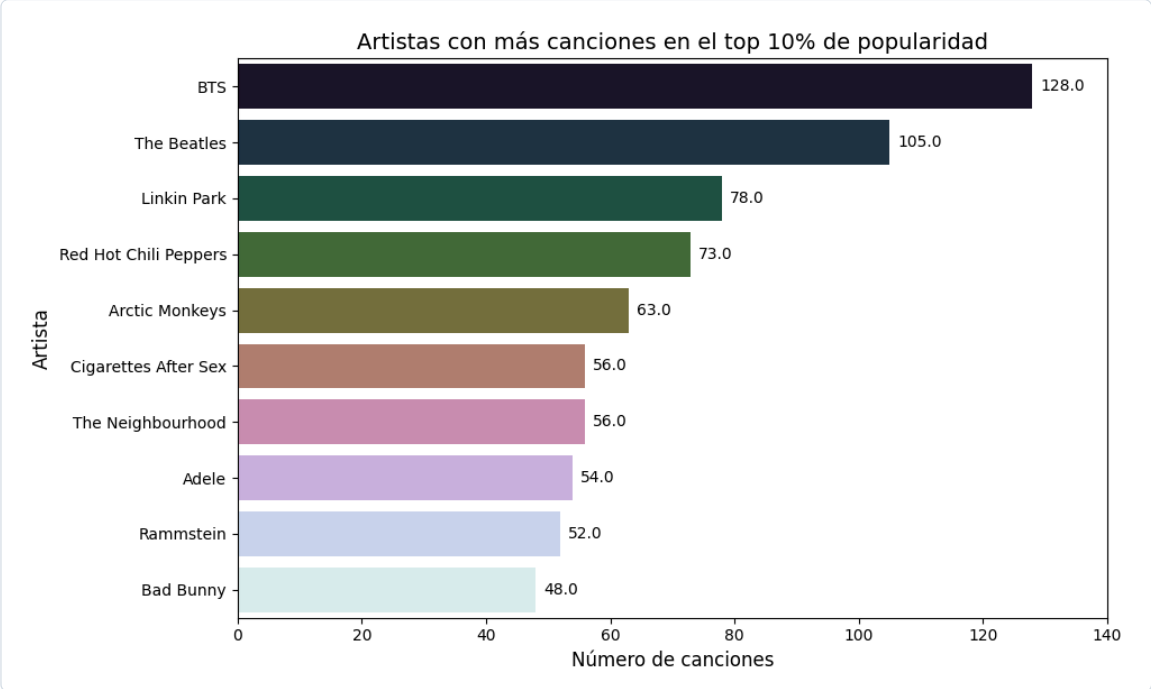



# Preguntas de negocio

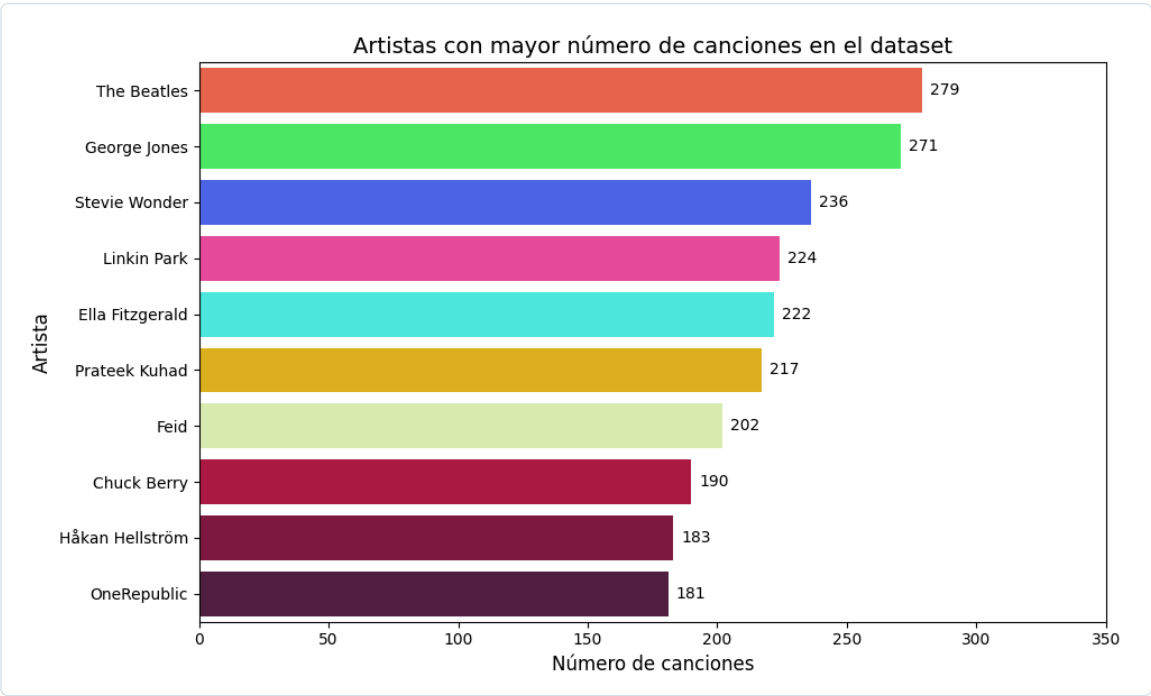
### 1) ¿Cuántas canciones hay por género de música?

```
A continuación se mostraran el nombre de las columnas y el total de sus canciones:

          Género  Número de canciones
0       acoustic                 1000
1       afrobeat                 1000
2       alt-rock                 1000
3    alternative                 1000
4        ambient                 1000
..           ...                  ...
109       techno                 1000
110       trance                 1000
111     trip-hop                 1000
112      turkish                 1000
113  world-music                 1000

[114 rows x 2 columns]
Cada genero de música tiene un total de 1000 canciones
```

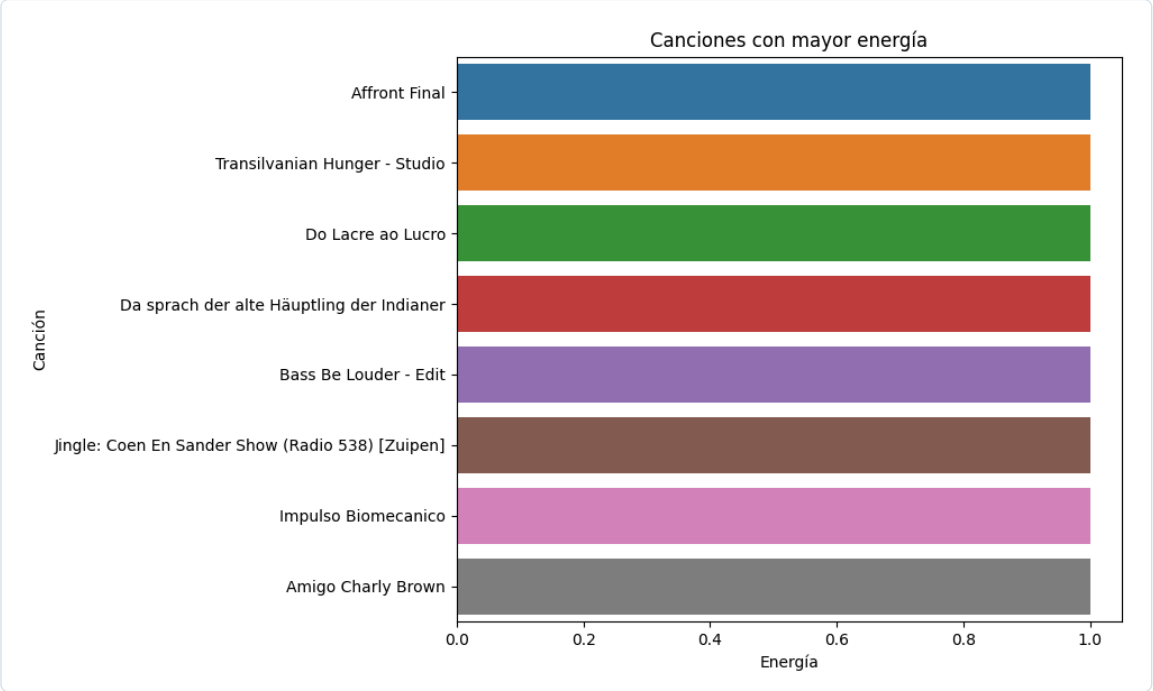
### 2) ¿Qué artista tiene más canciones en el top 10% de popularidad?




### 3) ¿Qué género tiene la mayor cantidad de canciones explícitas?(hacer que se muestre el numero exacto del resultado)



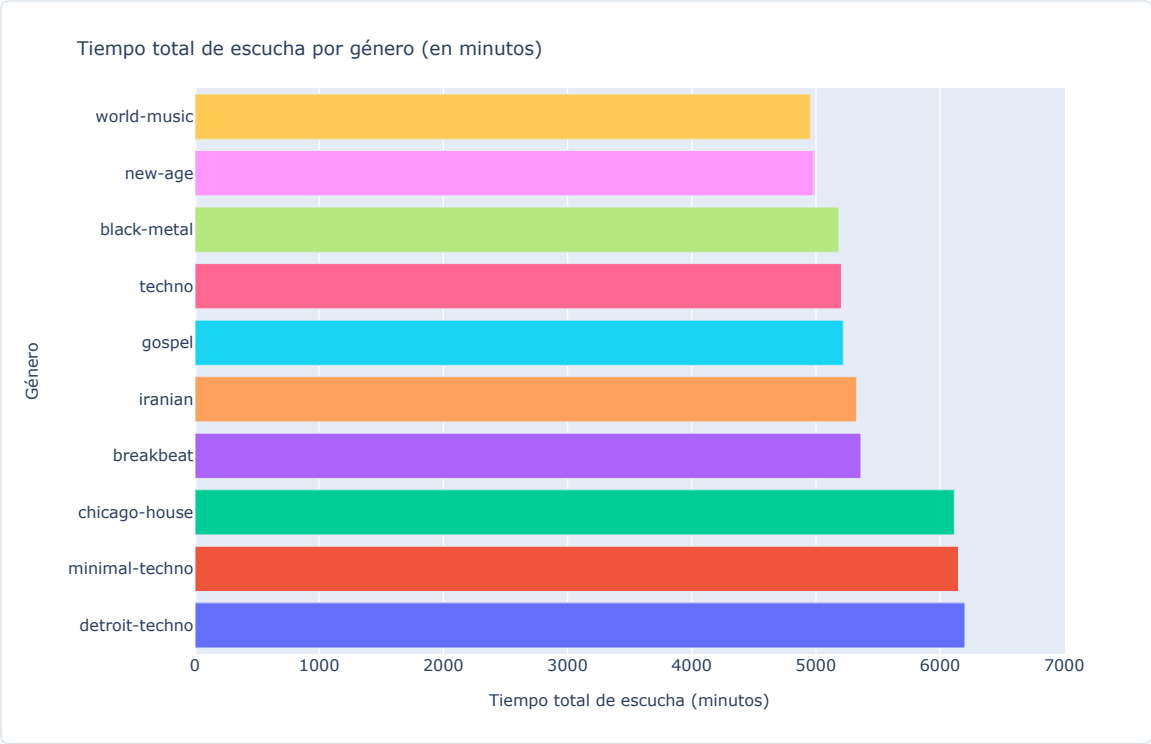

### 4) ¿Cuáles son las canciones con la mayor energía en el dataset?



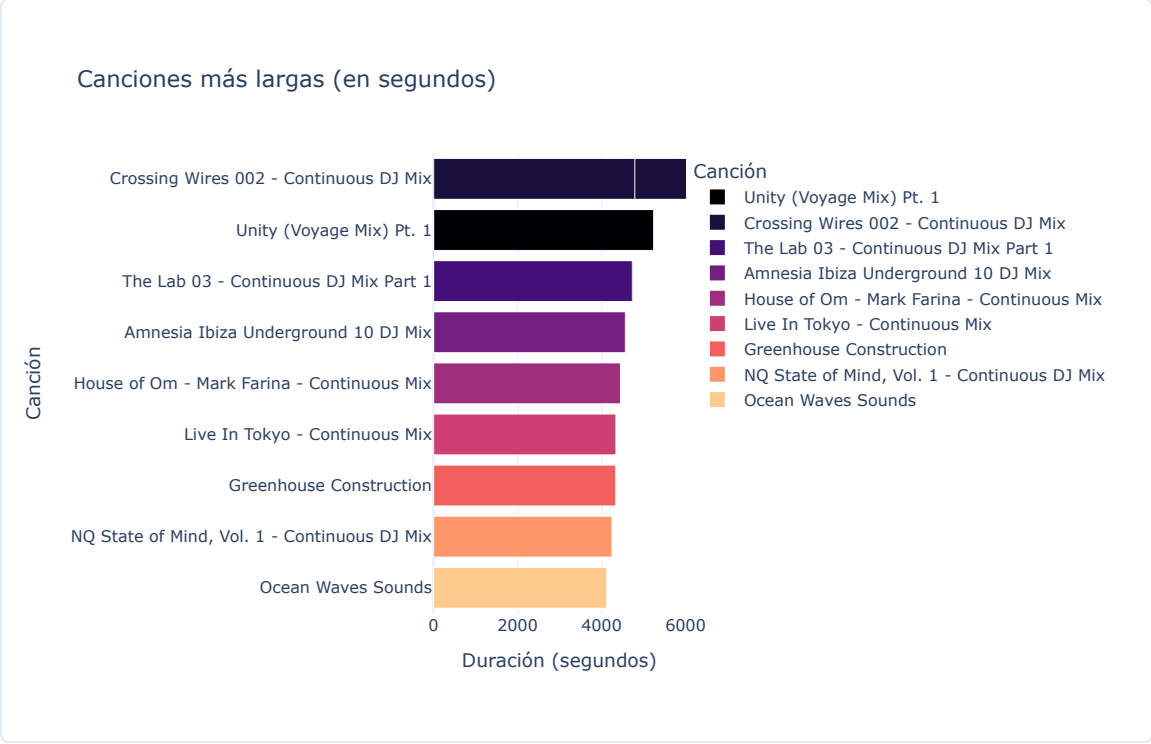

### 5) ¿Qué artista tiene el mayor número de canciones en el dataset?

### 6) ¿Cuál es el tiempo total de escucha  de todas las canciones en un género específico?

### 7): ¿Cuál es el género con la duración promedio más corta de las canciones?




### 8) ¿Qué artista tiene la canción más larga en términos de duración?



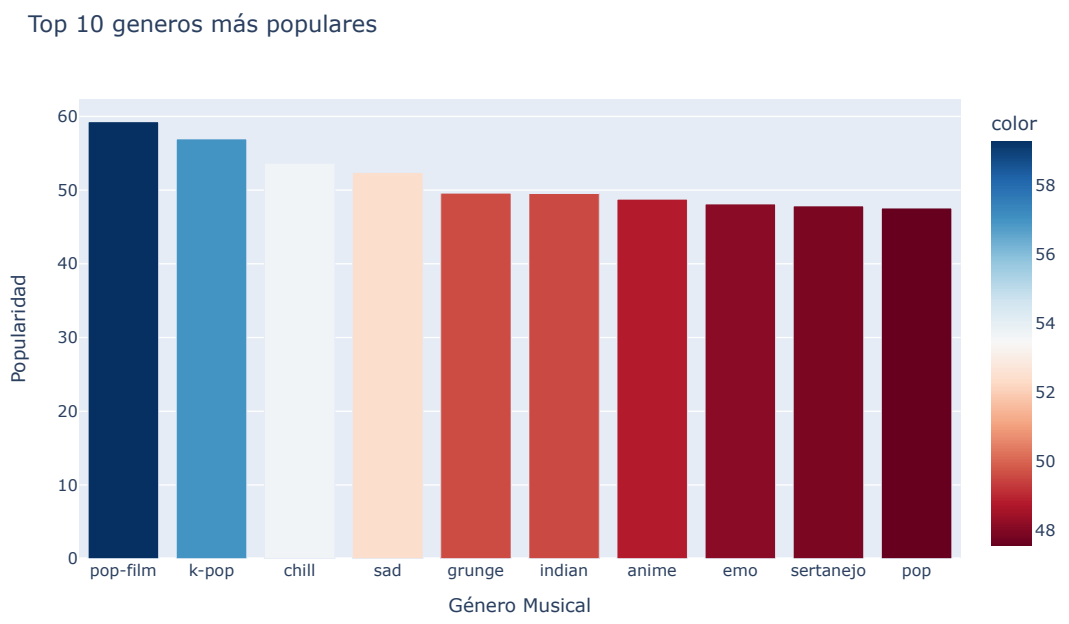

## Sound_features

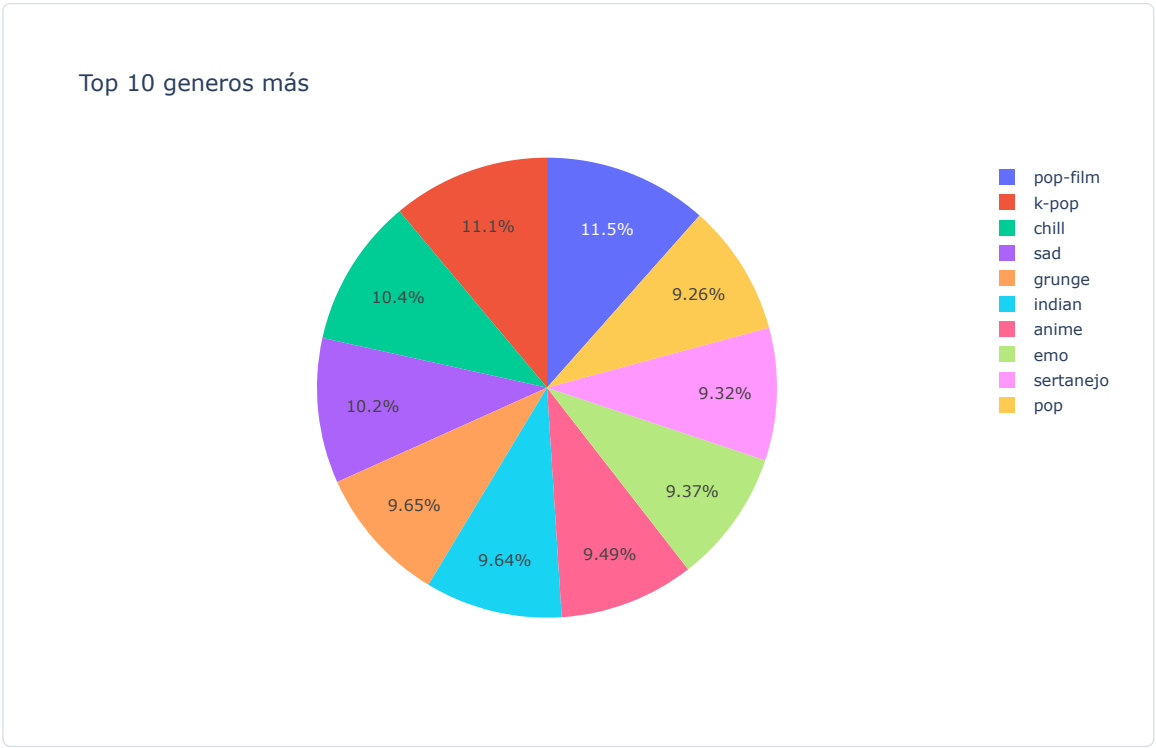
Se creo una

```
['acousticness',
 'danceability',
 'energy',
 'instrumentalness',
 'liveness',
 'valence']
```

# Top Genre

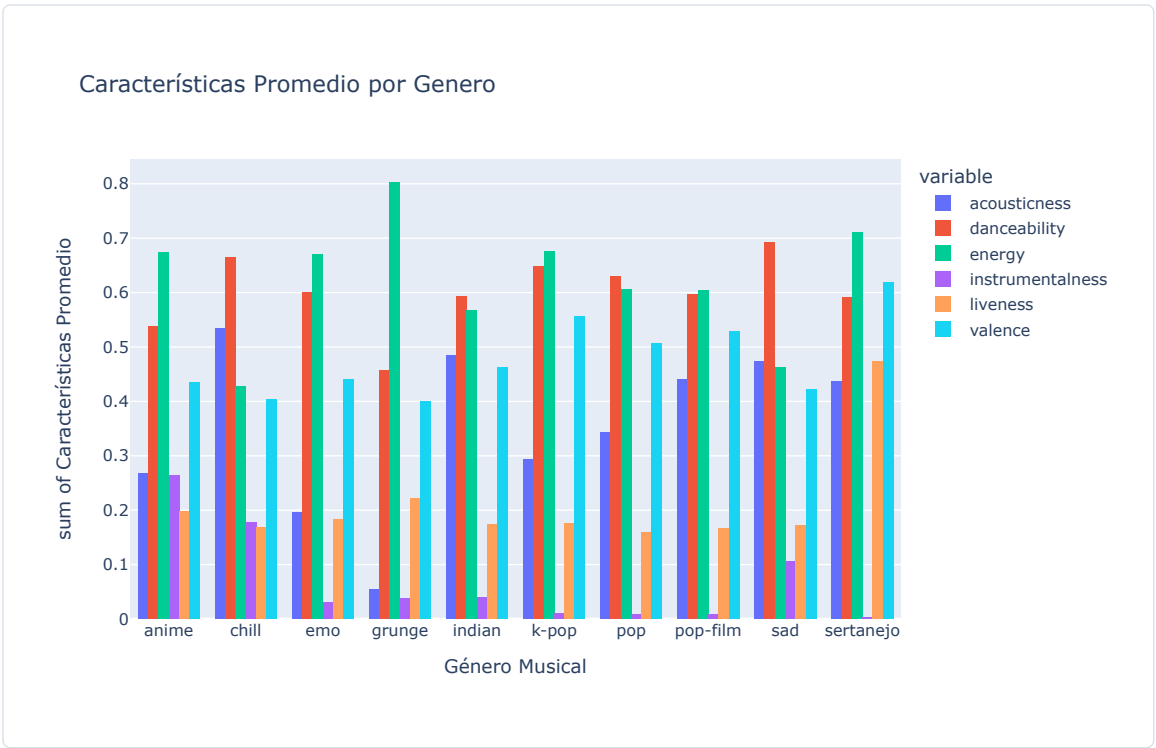
```
track_genre
pop-film     59.283000
k-pop        56.952953
chill        53.651000
sad          52.379000
grunge       49.594000
indian       49.539000
anime        48.772000
emo          48.128000
sertanejo    47.866000
pop          47.576000
Name: popularity, dtype: float64
```





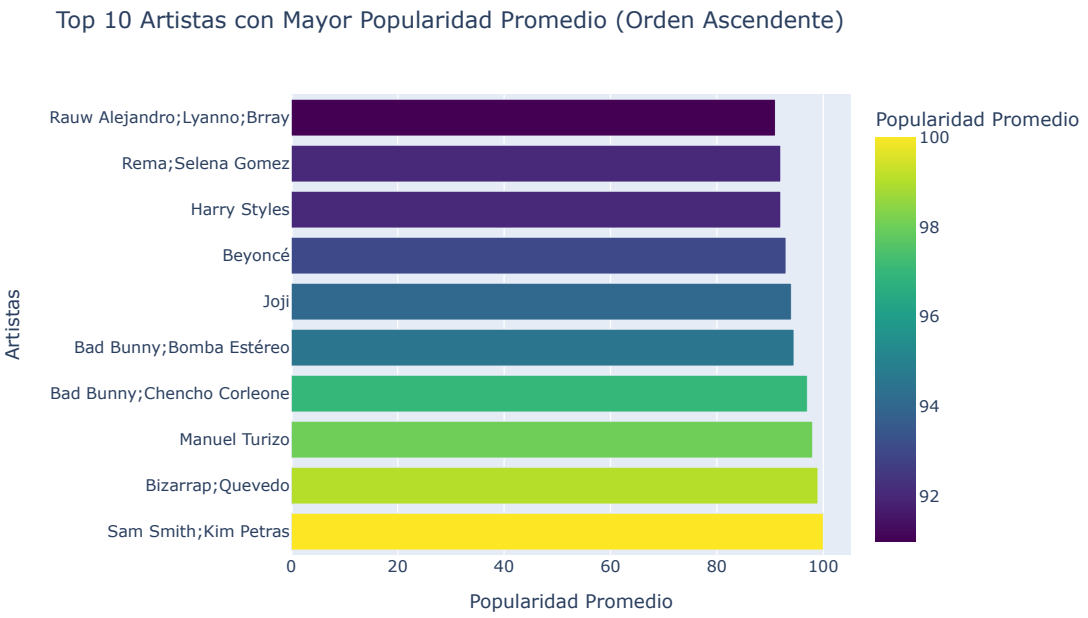

## Scatter

Grafico para encontrar algún patrón que determine porque x genero es mejor.

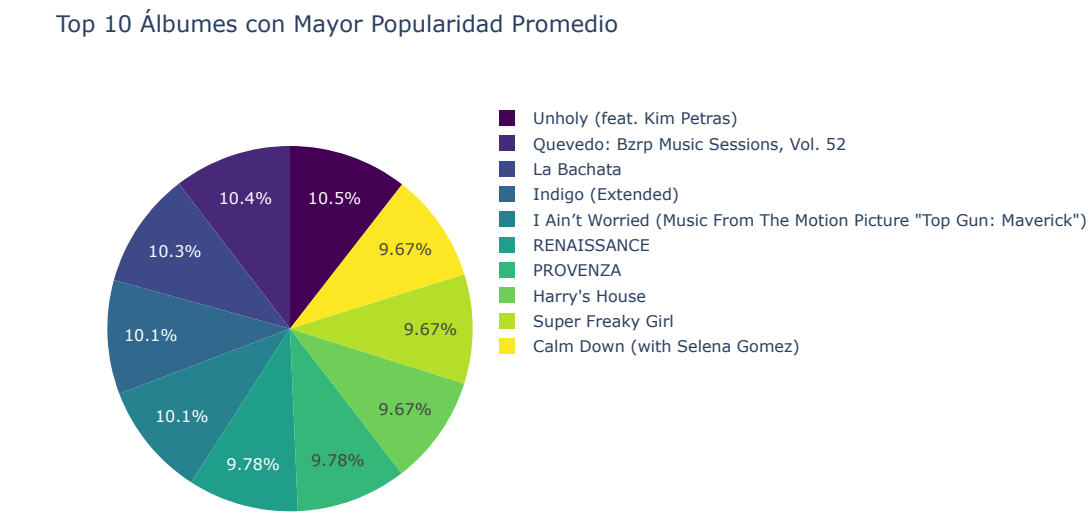



No se encontró ningún patrón para determinar cual es el mejor genero, todo depende del gusto musical de las personas

# Top Artists


Top 10 Artistas con Mayor Popularidad Promedio (Orden Ascendente)
Rauw Alejandro;Lyanno;Brray
Rema;Selena Gomez
Harry Styles
Beyoncé
Joji
Bad Bunny;Bomba Estéreo
Bad Bunny;Chencho Corleone
Manuel Turizo
Bizarrap;Quevedo
Sam Smith;Kim Petras
Artistas
0
20
40
60
80
100
Popularidad Promedio
Popularidad Promedio
100
98
96
94
92

# Top Album


Top 10 Álbumes con Mayor Popularidad Promedio
10.5%
9.67%
9.67%
9.67%
9.78%
9.78%
10.1%
10.1%
10.3%
10.4%
Unholy (feat. Kim Petras)
Quevedo: Bzrp Music Sessions, Vol. 52
La Bachata
Indigo (Extended)
I Ain't Worried (Music From The Motion Picture "Top Gun: Maverick")
RENAISSANCE
PROVENZA
Harry's House
Super Freaky Girl
Calm Down (with Selena Gomez)