Musical Analytics Spotify 2010-222

Bootcamp Ciencia de Datos

Misael Vergara Huerta

Introducción

Actualmente existe una recopilación abundante de datos por parte de las empresas de streaming, principalmente para conocer a sus consumidores y así poder mejorar los servicios que ofrece, un ejemplo de esto se observa en los servicios musicales como Spotify, por tal motivo es necesario aprovechar y realizar un análisis de datos con respecto a las canciones mas populares del periodo 2010 al 2022.

El conjunto de datos puede ser descargado del siguiente enlace: https://www.kaggle.com/code/akouaorsot/musical-analytics-spotify-2010-2022/notebook

Procedimiento

#Cargar la paquetería necesaria para llevar a cabo el análisis.

- import pandas as pd
- import polars as pl
- import altair as alt
- import seaborn as sns
- import matplotlib.pyplot as plt
- import py arrow

Cargamos el conjunto de datos y eliminamos las columnas que no usaremos:

- Spotify_data = pl.read_csv('playlist_2010to2022.csv')
- Spotify_data = spotify_data.drop(['playlist_url','track_id','time_signature'])
- Spotify_data.head (5)

Breve descripción de datos.

¿Qué tamaño tiene los datos?

- Con data.shape nosotros podeos conocer las dimensiones o tamaño de nuestro conjunto de datos el cual consta de 2300 filas y 23 columnas.

¿Cuáles son los tipos de datos?

- Los datos con los que cuenta este archivo son de tipo str, int y float.

¿Qué información contiene cada columna?

Playlist_url: es una liga que nos direcciona a la playlist

Year: nos indica el año de la canción

track_id: es el id de la canción

track_name: el nombre de la canción

track_popularity: indica que tan popular es una canción donde los valores van de 0 a 100

album: Indica el álbum al que pertenece la canción

artist_id: es el id del artista

artist_name: nombre del artista

artist_genres: género en el que se encuentra el artista

artist_popularity: La popularidad del artista donde los valores son del 0 a 100.

Acústica: Una medida de confianza de 0.0 a 1.0 de si la pista es acústica. 1.0 representa una alta confianza en la que la pista es acústica.

Danzabilidad: La danzabilidad describe lo adecuado que es una pista para bailar basada en una combinación de elementos musicales, incluyendo tempo, estabilidad del ritmo, fuerza de ritmo y regularidad general. Un valor de 0.0 es menos bailable y 1.0 es el más bailable.

Energía: La energía es una medida de 0,0 a 1,0 y representa una medida perceptual de intensidad y actividad. Normalmente, las pistas energéticas se sienten rápidas, ruidosas y ruidosas. Por ejemplo, Death Metal tiene alta energía, mientras que un preludio de Bach obtiene bajas en la escala. Las características perceptivas que contribuyen a este atributo incluyen rango dinámico, ruido percibido, timbre, tasa de inicio y entropía general.

Instrumentalidad: Predece si una canción no contiene voces. Los sonidos de "Oh" y "aah" se tratan como instrumentales en este contexto. Las pistas de rap o palabras habladas son claramente "vocales". Cuanto más se acerque el valor de instrumentalidad es a 1.0, mayor es la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0,5 están destinados a representar pistas instrumentales, pero la confianza es mayor a medida que se acerca el valor 1.0.

Clave: La llave en la que está la pista. Los enteros mapean a los lanzamientos usando notación estándar de clase de tono. Por ejemplo. 0 = C, 1 = C, 2 = D, y así es. Si no se detectó ninguna llave, el valor es de -1.

Vive: Detecta la presencia de un público en la grabación. Los valores de vida más altos representan una mayor probabilidad de que la pista se realizó en vivo. Un valor por encima de 0.8 proporciona una fuerte probabilidad de que la pista esté en vivo.

Loudness: El ruido general de una pista en decibelios (dB). Los valores de ruido se promedian en toda la pista y son útiles para comparar la intensidad relativa de las pistas. La olencia es la cualidad de un sonido que es la principal correlación psicológica de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db.

Modo: El modo indica la modalidad (mayor o menor) de una pista, el tipo de escala a partir del cual se deriva su contenido melódico. Mayor está representado por 1 y menor es 0.

Lanzaz de discursos Speechiness detecta la presencia de palabras habladas en una pista. La más exclusivamente habla-como la grabación (por ejemplo, talk show, audioliario, poesía), más cerca de 1.0 el valor de atributo. Valores por encima de 0,66 describen pistas que probablemente se hacen enteramente de palabras habladas. Los valores entre 0.33 y 0,66 describen temas que pueden contener tanto música como en el habla, ya sea en secciones o en capas, incluyendo casos como música rap. Los valores por debajo de 0.33 probablemente representan la música y otras pistas no-ecérech-likes.

Tempo: El ritmo total estimado de una pista en latidos por minuto (BPM). En terminología musical, el tempo es la velocidad o el ritmo de una pieza determinada y deriva directamente de la duración media del ritmo.

Valence: Una medida de 0.0 a 1.0 que describe el positivo musical transmitida por un tema. Las pistas con alta valencia suenan más positivas (por ejemplo, feliz, alegre, eufórico), mientras que las pistas con baja valencia suenan más negativas (por ejemplo, triste, deprimido, enfadado).

Duration_ms: duración de la cancion en milisegundos.

Time_signature: indica el Compas de la canción.

¿Hay valores nulos en los datos?

- En este conjunto de datos a primera vista solo hay un valor faltante, sin embargo, explorando nuevamente el archivo podemos encontrar 24 valores nulos en la columna de “artis_genres”, por lo tanto, se tomó la decisión de clasificarlos como “sin género”.

Para el análisis de datos la información a utilizar.

Las columnas que se utilizaran son: year, track_name, track_popularity, album, artist_name, artist_genres, artist_popularity, danceability, energy, key, loudness, mode, speechiness acousticness, instrumentalness, liveness, valence, tempo, duration_ms, time_signature. Se descartan las columnas track_id, playlist_url y artis_id ya que no proporcionan información relevante para el análisis de datos.

Utilizando la siguiente función, permite eliminar las columnas que no necesitaremos:

- spotify_data.drop(["playlist_url","track_id","artist_id","time_signature"])

A continuación, daremos respuesta a los siguientes cuestionamientos:

Ordenamos nuestro conjunto de datos por popularidad del artista

- Sort_spotify_data = Spotify_data.sort ('artist_popularity', descending = TRUE)
- Sort_spotify_data.head()

¿Cuáles son los artistas más escuchados?

```
top_artist = sort_spotify_data.group_by(["artist_name","artist_popularity"],
maintain_order=True).agg(pl.count())

top_artist.head(10)
```

Tabla 1. Top 10 de artistas más escuchados

artist_name	artist_popularity	count
str	i64	u32
"Taylor Swift"	100	31
"Bad Bunny"	94	11
"Drake"	94	32
"The Weeknd"	94	15
"Travis Scott"	93	6
"KAROL G"	90	3
"Kanye West"	89	18
"Lana Del Rey"	89	7
"Post Malone"	89	15
"21 Savage"	88	2

¿Cuáles son los artistas menos escuchados?

Tabla 2. Top 10 de artistas menos escuchados

artist_name	artist_popularity	count
str	i64	u32
"Bodyrockers"	38	1
"OPM"	38	1
"Dirty Vegas"	37	1
"Kandi"	36	1
"Liberty X"	36	1
"BBMAK"	35	1
"DJ Pied Piper …	35	1
"Holly Valance"	34	1
"Rachel Stevens…	31	1
"Matt Cardle"	29	1

¿Cuáles son las canciones más escuchadas?

```
top_songs =
sort_spotify_data_track.select(["artist_name","track_name","track_popularity"]).filter(pl.col("trac
k_popularity")  <= 100)

top_songs.head(10)
```

Tabla 3. Top 10 canciones más escuchadas.

artist_name	track_name	track_popularity
str	str	i64
"Taylor Swift"	"Cruel Summer"	100
"David Guetta"	"I'm Good (Blue…	93
"Taylor Swift"	"august"	93
"Taylor Swift"	"Anti-Hero"	93
"OneRepublic"	"I Ain't Worrie…	92
"Rema"	"Calm Down (wit…	92
"The Weeknd"	"Starboy"	92
"The Weeknd"	"Blinding Light…	92
"Bizarrap"	"Quevedo: Bzrp …	91
"Coldplay"	"Yellow"	91

¿Cuáles son las canciones menos escuchadas?¶

Tabla 4. Top 10 canciones menos escuchadas.

artist_name	track_name	track_popularity
str	str	i64
"Billie Eilish"	"Your Power"	1
"Black Eyed Pea…	"Shut Up"	1
"Ariana Grande"	"positions"	0
"BLACKPINK"	"How You Like T…	0
"BTS"	"Dynamite"	0
"BTS"	"ON (Feat. Sia)…	0
"Fatman Scoop"	"Be Faithful"	0
"Lorde"	"Solar Power"	0
"Machine Gun Ke…	"my ex's best f…	0
"Pop Smoke"	"What You Know …	0

¿Cuál es genero de música más escuchado?

```
top_genres = sort_spotify_data.group_by(["artist_genres"], maintain_order=True).agg(pl.count())

top_genres = top_genres.sort("count", descending = True)

top_genres = top_genres.with_columns(
   ((pl.col("count") / 2300) * 100).round(2).alias("porcentaje")
)

top_genres.head(19)
```

artist_genres	count	porcentaje
str	u32	f64
"pop"	556	24.17
"dance pop"	252	10.96
"rap"	137	5.96
"hip hop/rap"	126	5.48
"hip hop"	109	4.74
"r&b"	104	4.52
"pop rap"	68	2.96
"trap"	64	2.78
"pop rock"	60	2.61
"electro house"	57	2.48
"house"	49	2.13
"reggaeton"	42	1.83
"rock alternati…	42	1.83
"rock"	38	1.65
"nu metal"	33	1.43
"pop punk"	33	1.43
"europop"	33	1.43
"country"	30	1.3
"british soul"	29	1.26

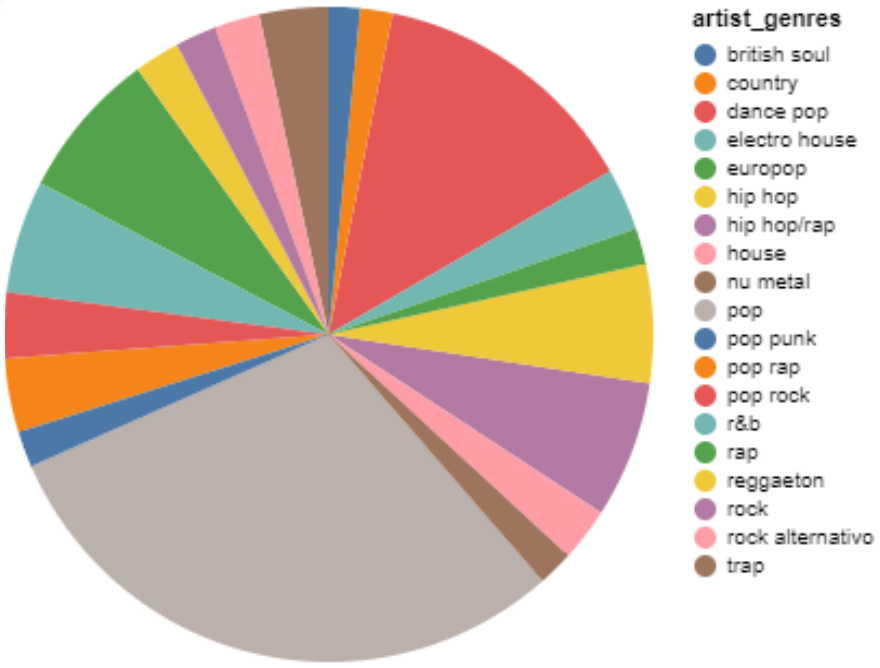

Ilustración 1. Grafica de distribución con respecto al género musical más escuchado.

¿Existe correlaciones?

```
plt.figure(figsize=(12, 10))
sns.heatmap(sopotify_heatmap, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix')
plt.tight_layout()
plt.show()
```

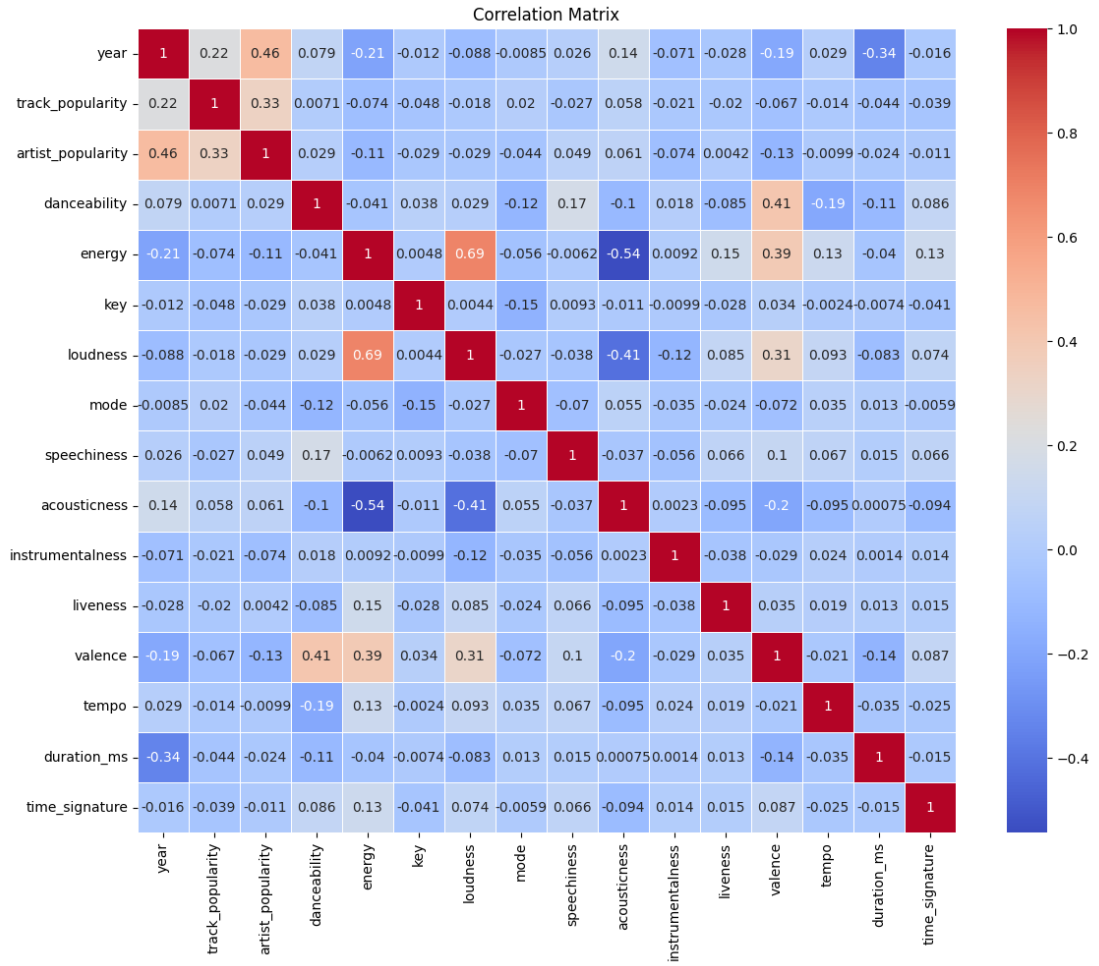

Ilustración 2. Matriz de correlación.

A partir del análisis de correlación se observa que en su mayoría son de carácter negativo, el objetivo principal de este análisis era encontrar la relación entre las variables para saber si la popularidad de la canción o del artista dependía de las características (danceability, energy, key, loudness, mode, speechiness acousticness, instrumentalness, liveness, valence, tempo, duration_ms)

Conclusiones

Mediante el análisis de datos se puede obtener una información detallada, que posterior mente puede ser utilizada para conocer a los usuarios y futuros clientes, promoviendo elementos populares o con mayor reproducción de contenido, como es en el caso de los servicios de streaming que posterior mente son utilizados para agregar mas servicios y que los usuarios tengan un mayor nivel de consumo en la plataforma.