# Ejercicio Intermedio Módulo 3

## Los Datos:

- Varias estadísticas de la versión musical en spotify, incluido el número de streams;
- Número de visualizaciones del vídeo musical oficial de la canción en youtube.

Las variables que tendremos en este *dataframe* son:

Incluye 26 variables para cada una de las canciones recogidas de spotify. Estas variables se describen brevemente a continuación:

- `Track`: nombre de la canción, tal y como aparece en la plataforma Spotify.
- `Artist`: nombre del artista.
- `Url_spotify`: Url de la canción.
- `Album`: el álbum en el que se encuentra la canción en Spotify.
- `Album_type`: indica si la canción se publica en Spotify como single o dentro de un álbum.
- `Uri`: un enlace de Spotify utilizado para encontrar la canción a través de la API.
- `Danceability`: describe lo adecuada que es una canción para bailar basándose en una combinación de elementos musicales como el tempo, la estabilidad del ritmo, la fuerza del compás y la regularidad general. Un valor de 0,0 es el menos bailable y 1,0 el más bailable.
- `Energy`: es una medida de 0,0 a 1,0 y representa una medida perceptiva de intensidad y actividad. Normalmente, las pistas energéticas son rápidas, ruidosas y ruidosas. Por ejemplo, el death metal tiene mucha energía, mientras que un preludio de Bach tiene una puntuación baja en la escala. Las características perceptivas que contribuyen a este atributo incluyen el rango dinámico, el volumen percibido, el timbre, la velocidad de inicio y la entropía general.
- `Key`: la clave de la pista. Los números enteros se asignan a los tonos utilizando la notación estándar Pitch Class. Por ejemplo, 0 = C, 1 = C♯/D♭, 2 = D, y así sucesivamente. Si no se detecta ninguna tonalidad, el valor es -1.
- `Loudness`: la sonoridad global de una pista en decibelios (dB). Los valores de sonoridad se promedian en toda la pista y son útiles para comparar la sonoridad relativa de las pistas. La sonoridad es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db.
- `Speechiness`: detecta la presencia de palabras habladas en una pista. Cuanto más exclusivamente hablada sea la grabación (por ejemplo, un programa de entrevistas, un audiolibro, poesía), más se acercará a 1,0 el valor del atributo. Los valores superiores a 0,66 describen pistas que probablemente estén compuestas en su totalidad por palabras habladas. Los valores entre 0,33 y 0,66 describen pistas que pueden contener tanto música como voz, ya sea en secciones o en capas, incluyendo casos como la música rap. Los valores por debajo de 0,33 representan probablemente música y otras pistas no habladas.

- `Acousticness`: una medida de confianza de 0,0 a 1,0 para determinar si la pista es acústica. 1,0 representa una confianza alta en que la pista es acústica.
- `Instrumentalness`: predice si una pista no contiene voces. Los sonidos "ooh" y "aah" se consideran instrumentales en este contexto. Las pistas de rap o spoken word son claramente "vocales". Cuanto más se acerque el valor de instrumental a 1,0, mayor será la probabilidad de que la canción no contenga voces. Los valores superiores a 0,5 representan pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0.
- `Liveness`: detecta la presencia de público en la grabación. Los valores más altos representan una mayor probabilidad de que la pista se haya interpretado en directo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista sea en directo.
- `Valence`: una medida de 0,0 a 1,0 que describe la positividad musical que transmite una pista. Las pistas con una valencia alta suenan más positivas (por ejemplo, felices, alegres, eufóricas), mientras que las pistas con una valencia baja suenan más negativas (por ejemplo, tristes, deprimidas, enfadadas).
- `Tempo`: el tempo global estimado de una pista en pulsaciones por minuto (BPM). En terminología musical, el tempo es la velocidad o el ritmo de una pieza determinada y se deriva directamente de la duración media de los tiempos.
- `Duration_ms`: duración de la pista en milisegundos.
- `Stream`: número de streams de la canción en Spotify.
- `Url_youtube`: url del vídeo enlazado a la canción en Youtube, si lo tiene.
- `Title`: título del videoclip en Youtube.
- `Channel`: nombre del canal que ha publicado el vídeo.
- `Views`: número de vistas.
- `Likes`: número de me gusta.
- `Comments`: número de comentarios.
- `Description`: descripción del vídeo en Youtube.
- `Licensed`: Indica si el vídeo representa contenido con licencia, lo que significa que el contenido fue subido a un canal vinculado a un socio de contenido de Youtube y luego reclamado por dicho socio.
- `official_video`: valor booleano que indica si el vídeo encontrado es el vídeo oficial de la canción.

## Los ejercicios planteados son:

1. Importa Pandas
2. Exploración del conjunto de datos:

- Carga el fichero, al cargarlo te aparecerá una columna llamada "Unnamed: 0". Carga el dataset sin que aparezca esta columna. Además realiza un análisis exploratorio que incluya:
    - Cuando leas el fichero, veras que no podemos ver todas las columnas, utiliza el comando correcto para poder visualizarlas todas cuando hacemos un head
    - Primeras 5 filas del *dataframe*
    - Últimas 5 filas del *dataframe*
    - 10 filas aleatorias del *dataframe*
    - ¿Cuántas filas y columnas tenemos en el *dataframe*
    - ¿Cuáles son los tipos de los datos de cada columna del *dataframe*
    - ¿Cuántos valores nulos tenemos por columna?
    - ¿Tenemos filas duplicadas en el *dataframe* ?
    - Muestra los principales estadísticos para las columnas numéricas del *dataframe*
    - Muestra los principales estadísticos para las columnas categóricas del *dataframe*

3. Preparación de los datos:
    - Los nombres de las columnas empiezan con mayúsculas, pon todos los nombres de las columnas en minúsculas.
    - Haciendo el análisis exploratorio te deberías haber dado cuenta de que algunas de las variables (`danceability`, `energy`, `key`, `loudness`, `speechiness`, `acousticness`, `instrumentalness`) no son del tipo que deberían. Esto es debido a que los decimales están establecidos como comas y no con puntos. Crea una función que nos permita cambiar esas comas por puntos para que los datos tengan el tipo correcto.
        - Después de haber hecho los cambios, chequea los tipos de datos. ¿Son ya del tipo correcto? En caso de que no, crea otra función o modifica la anterior para que sean de tipo *float*.
    - Hay algunas canciones cuyo título está en mayúsulas. Crea una función para que todos los títulos estén en minúscula. Haz lo mismo para las columnas de "title" y "album".
    - Algunos de los artistas tienen símbolos raros, en concreto el símbolo `$`. Crea una función que los reemplaze pos "s".
4. Filtrado de datos:
    - ¿Cuáles son los valores únicos de la columna "album_type"? Crea tres *dataframes* diferentes, uno para cada tipo de "album_type". ¿Cuántas canciones tenemos en cada tipo?
    - Usando el *dataframe* de los albumes que hemos creado en el ejercicio anterior. ¿Cuál es la media y la desviación estándar de "danceability", "acousticness" y "speechiness" de cada artista?. Muesrtra los resultados en tres dataframes diferentes, uno para cada

métrica ("danceability", "acousticness" y "speechiness"). Ordena los resultados de mayor a menor en base a la media.

- Se quiere estudiar si existe una relación entre la cantidad de visitas que reciben los videos de música de un artista en una plataforma de streaming y la cantidad de likes que estos videos obtienen. Por lo tanto, queremos contestar a la siguiente pregunta: ¿Son aquellos artistas que tienen más visitas los que más likes tienen? Para solucionar este ejercicio deberéis:
  - Agrupa por artista y calcular la media de visitas y de likes.
  - Quedate con los 10 artistas que más visitas han tenido.
  - Quedate con los 10 artistas que más likes han tenido.
  - **BONUS** Haced lo mismo para los datos de tipo "single".