Reporte de calidad de datos API Spotify – Taylor Swift

Dentro de las seis dimensiones, se concluye lo siguiente:

1. Completeness

Aspectos destacados

- Existe un 12.9% de valores nulos para album_name, seguido de 1.5% para track_name y audio_features.energy con un 1.3%, para los demás, se encuentra por debajo del 1% del total de registros.

- El album_name en porcentaje es mayor, al ser repetitivo en sus n-números de discos por cada álbum. Pero es de resaltar, que 8 track_name están vacíos del total de 539 canciones únicas.

- Las columnas (27) se encuentran a completitud, y existe en zona raw una coherencia entre los datos del json en sus diccionarios y forma de entregarse al dataset.csv

2. Uniqueness

Aspectos destacados

- Existe 18 registros duplicados, los cuales modifican campos como el número de album_total_tracks. El porcentaje del total de registros no duplicados es del 96.6%.

3. Timeliness

No existe un campo en el registro de datos que permita ver la actualización de datos por parte de Spotify sobre registros como popularidad que puede variar con el input de los usuarios. Este campo, podría permitir dentro de la calidad de datos, ver en cuanto tiempo fue su última actualización.

4. Validity

Aspectos destacados

- El campo explicit que debe ser booleano (TRUE/FALSE) pero en la revisión existen 4 casos con el valor No y 1 con Si, se debe realizar un reemplazo de los valores y dejarlos en su formato deseado.
- El formato de "release_date" en los datos se encuentra como tipo entero y la API sugiere tipo string. Los datos se encuentran con formato tipo AAAA-MM-DD, pero existe una diferencia entre la documentación de la API y los datos que entrega, dado que recomienda tener formato tipo AAAA-MM, y los datos están en AAAA-MM-DD, en este caso tocaría encontrar un camino para estandarizar y alinearse con la documentación de la API y los datos recibidos (ver foto posterior)

5. Accuracy

Aspectos destacados

- En la columna "duration_ms" existen 5 registros inadecuados en relación con el campo de tiempo (ms). Los valores negativos no tienen representación en el tiempo y los valores menores a 60000,

fueron 10, 1000 y 3000 ms, que representan menos de 3 segundos en una canción, lo cual no es adecuado para el ámbito musical.

- Para la columna "track_popularity" existen valores negativos y mayores a 100. El rango que ofrece la documentación de Spotify va de los 0 a 100, todo valor por fuera, se considerará anormal.

- Para "track_name" se encontraron valores como: This Love (Taylorâ€™s Version) que contienen "â€™" caracteres extraños, por problemas de comas, tildes, acentos y no tenerlo en cuenta al llevarlos de json-csv-pandas por el formato. Estos cambios deben realizarse, para tener calidad en el nombre de la canción.

- Para "audio_features.acousticness" existen dos valores -0.000537,-0.003540 por debajo del rango de 0 y 3 por encima de 1 (1.5, 2 y 5.0)

- Para "audio_features.instrumentalness" 7.28x-06, valor que no está en formato tipo número para su notación científica.

- Para "artist_id" con valor "06HL4z0CvFAxyc27GX" este dato, no responde a la API para la cantante Taylor Swift (ver imagen posterior). Este es un hecho relevante, dado que es la principal llave para consultar a la cantante en la API de Spotify. El valor que llama a su nombre es: "06HL4z0CvFAxyc27GXpf02"

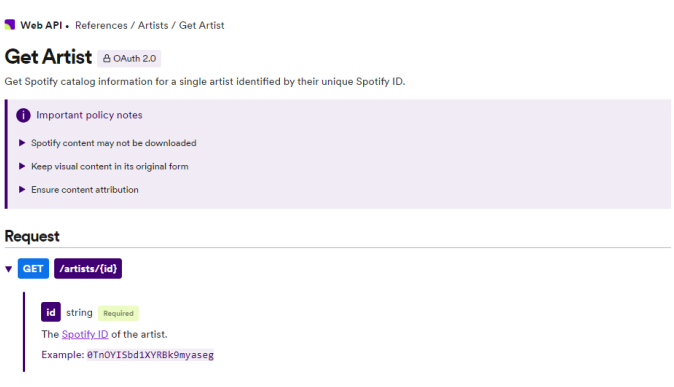
- La columna de "artist_popularity" en la documentación de la API de Spotify el rango va de 0 a 100. El único valor que trae los datos es de 120, valor que no corresponde, de hecho, taylor tiene un artist_popularity de: 100. En este caso, tocaría modifica el valor o truncarlo por encima de este.

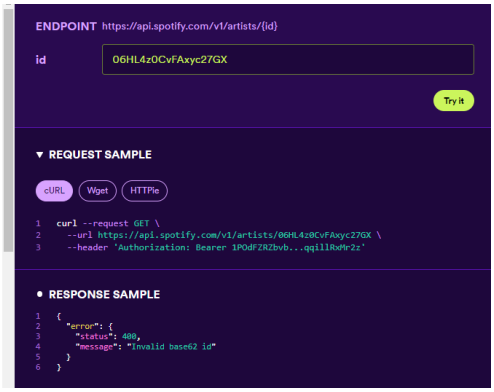
- En la columna "album_release_date" la fecha "2027-05-26" para el álbum: "Midnights (The Til Dawn Edition)" es incorrecta, dado que nos encontramos en 2024, la fecha correcta es: "2023-05-26". Aparte la fecha "1989-10-24" del álbum "Taylor Swift" fue lanzado el "2006-10-24", claramente teniendo un error en la casilla del año de 1989 a 2006.

- La columna "album_total_tracks" la documentación recomienda valores enteros, pero se encuentra un valor diferente "Thirtheen". Este dato, no concuerda con el tipo de dato usados en esta columna.

- Valores nulos se pueden revisar en el encabezado de "Completeness"

- Otro: Status 400 error por no encontrar el artist_id de Taylor Swift





6. Consistency

Coherencia del tipo de datos:

- df Type: asignación del tipo de datos al leer el csv.

- API Type: definición a partir de documentación de la API de Spotify del tipo de dato.

Ver tabla de propuesta a partir de las especificaciones de la API de Spotify en su documentación.

Nombre columna	df Type	API Type
0 disc_number	int64	integer
1 duration_ms	int64	integer
2 explicit	object	bool
3 track_number	int64	integer
4 track_popularity	int64	integer
5 track_id	object	string
6 track_name	object	string
7 audio_features.danceability	float64	float
8 audio_features.energy	float64	float
9 audio_features.key	float64	integer
10 audio_features.loudness	float64	float
11 audio_features.mode	float64	integer
12 audio_features.speechiness	float64	float
13 audio_features.acousticness	float64	float
14 audio_features.instrumentalness	object	float
15 audio_features.liveness	float64	float
16 audio_features.valence	float64	float
17 audio_features.tempo	float64	float
18 audio_features.id	object	string
19 audio_features.time_signature	float64	integer
20 artist_id	object	string
21 artist_name	object	string
22 artist_popularity	int64	integer
23 album_id	object	string
24 album_name	object	string
25 album_release_date	object	string
26 album_total_tracks	object	integer

Otros aspectos

1. Llaves únicas

Se encontraron dentro del json, las claves únicas para crear topologías y construcción de un data modelling a partir de las diferentes tipo de request a la API de Spotify.

- track_id
- audio_features.id
- album_id

2. Ordenar las columnas y ajuste de nombres:

Ajustado al json y la estructura del diccionario, seguir la secuencia para el nombre de las columnas:

- artista.[*] > álbum.[*] > track.[*] > audio_features.[*]

De esa misma forma el orden de las columnas de mayor a menor importancia, pasando primero por artista y terminando en audio_features.

3. **Columnas que probablemente no generen valor**

En audio_features se encuentran audio_features.liveness y audio_features.valence, datos que son muy específicos para un contexto de entender algunos comportamientos de la canción sobre algo puntual. Sobre un posterior análisis, se puede considerar tener o no, algunas columnas de audio_features, que entrega el dataset.

Anexo: punto2_SCO_R5.ipynb, cuaderno en Python, donde se encuentran todos los hallazgos a partir de los datos.