Reporte de calidad de datos

Prueba técnica

JULIÁN ORREGO CASTAÑEDA
Data Quality Engineer

1. Resumen

El presente reporte contiene los principales hallazgos de calidad obtenidos al analizar los datos de la artista Taylor Swift, descargados a través de la API de Spotify. Para esto, se realizó un análisis general del dataset e individual para cada variable, clasificando los hallazgos en cuatro dimensiones: Completitud, Validez, Precisión y Unicidad.

Entre los principales resultados se tiene la presencia de valores nulos, especialmente en los nombres de los álbumes, con un porcentaje de registros nulos o vacíos del 11,5% respecto al total de registros; además de la presencia de registros duplicados.

Igualmente, se identificaron errores de formato (valores expresados como números y texto, cadenas vacías, expresiones en notación científica) y de consistencia (uso de minúsculas y mayúsculas en los campos tipo texto y cantidad de posiciones decimales en los campos numéricos), principalmente en los campos relacionados con características de audio.

Adicionalmente, se encontraron valores atípicos como fechas de lanzamiento posteriores a la fecha actual o inferiores a la fecha de nacimiento de la artista, valores negativos, al igual que valores fuera del rango establecido según la documentación de la API.

Por último, el informe se acompaña de un archivo en formato .html con el análisis técnico y código empleado para la identificación de los hallazgos descritos.

Contenido

COMPLETITUD

2. Datos nulos

Los datos exportados constan de 27 campos, de los cuales se identificó que 12 de ellos presentan valores nulos o vacíos, es decir el 44% con respecto al total de campos. No obstante, solo el campo “album_name” contiene un porcentaje significativo respecto al total de registros (11.5%). La siguiente tabla presenta los campos cuyo porcentaje de valores nulos es superior al 1%.

Tabla 1. Valores nulos o vacíos

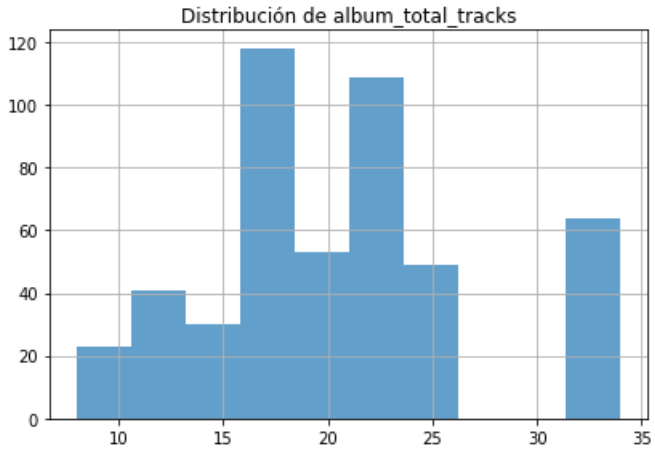
Nombre campo	Número de registros nulos o vacíos	Porcentaje de registros nulos o vacíos respecto al total de registros
album_name	62	11.50%
track_id	8	1.48%
track_name	7	1.30%

Adicionalmente, el campo “track_name” permite identificar las características de audio de la canción al relacionarse con el campo “audio_features.id”. Por tanto, aunque el porcentaje de valores nulos es bajo, se puede ver afectado el principio de integridad referencial al no tener un valor para poder relacionarlo con las demás características.

3. Registros vs canciones por álbum

Se realizó un análisis de la distribución de las canciones por álbum (campo album_total_tracks), como se evidencia en la figura 1, encontrando álbumes con más de 30 canciones, los cuales pertenecen a versiones extendidas, por lo cual tienen concordancia.

Figura 1. Distribución campo “album_total_tracks”



Sin embargo, dicho análisis permitió identificar que, para el álbum “Red (Taylor's Version)”, la cantidad de registros (canciones) no corresponde con el indicado en el campo “album_total_tracks”. Por tanto, se realizó el análisis para todo el dataset (una vez eliminados los registros duplicados), evidenciando que, en total, son 6 álbumes para los cuales la cantidad de registros no coincide con el número de canciones reportadas. Los hallazgos se presentan en la tabla 2.

Tabla 2. Diferencias entre las canciones indicadas por álbum y el total de registros para ese álbum

Álbum	Cantidad de canciones reportadas (Album_total_tracks)	Cantidad de registros en el dataset para dicho álbum
Lover	18	19
Red (Taylor's Version)	34	30
Taylor Swift	13	15
evermore	10	15
reputation	15	16
Midnights (The Til Dawn Edition)	24	23

Lo anterior implica que existe la posibilidad de tener errores en la información del campo “album_total_tracks” o que no se cuenta con la información completa sobre todos los álbumes.

VALIDEZ Y PRECISIÓN

Dados los hallazgos de calidad identificados en el dataset, se procede a agrupar las dimensiones de validez y precisión debido a que, como lo expresa el DAMA UK Working Group[1], para que exista precisión es necesario que los datos sean válidos y presenten el valor correcto respecto a formato y rangos.

4. Consistencia de los datos

Se realizó un análisis del formato de los datos, encontrando lo siguiente:

Formato decimal

Las características de audio (campos identificados como audio_features) cuyo formato es decimal, no presentan una estructura definida, encontrando registros que contienen uno, dos o hasta cuatro decimales y que pertenecen al mismo campo. En caso de que dichas cifras decimales posean un nivel de significancia, esto puede afectar futuros modelos o análisis y se recomienda normalizar los datos.

Formato string

Respecto a los campos “track_name” y “album_name”, cuyo formato es string, se encontró que no existe un patrón para el uso de minúsculas y mayúsculas, esto es, existen valores ingresados totalmente en minúscula, mayúscula inicial en cada palabra, etc. Se recomienda normalizar este campo para facilitar el uso de modelos de procesamiento del lenguaje natural (NLP).

4.1. Casos específicos

Campo “audio_features.acousticness”

Para esta columna en particular se identificó que, a pesar de ser de tipo numérico, presenta registros tipo texto, expresados como cadenas vacías (‘’).

[1] The six primary dimensions for data quality assessment (2013). DAMA UK Working Group

Igualmente, según la documentación de Spotify, los valores de este campo deben estar entre 0 y 1. No obstante, se encontraron valores negativos y valores superiores a 1.

Campo “explicit”

Este campo es de tipo booleano, por lo cual debería aceptar solo dos valores (False o True). Sin embargo, presenta un error respecto al idioma y está aceptando también los valores Si y No.

Campo “album_total_tracks”

Contiene el total de canciones asociadas a cada álbum en el dataset. Sin embargo, acepta valores tanto numéricos como tipo texto. Por ejemplo, en vez del valor “13”, puede recibir el valor “Thirteen”. Se recomienda definir restricciones para aceptar solo valores numéricos, de modo que se puedan hacer comparaciones entre álbumes de manera más sencilla.

Campo “audio_features.instrumentalness”

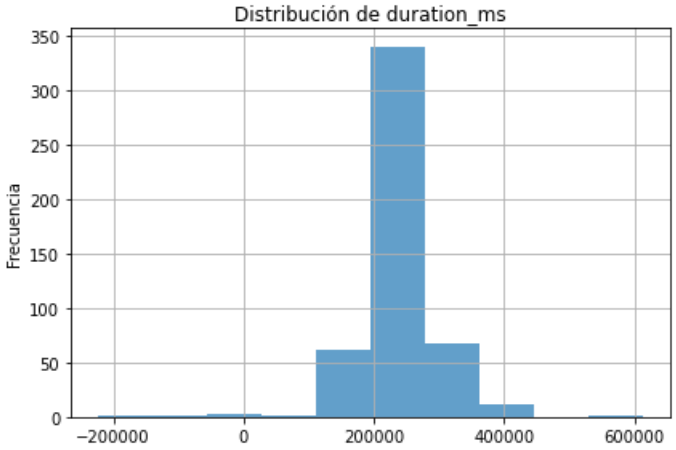
Adicional a no tener una estructura definida para la cantidad de decimales, este campo presenta números almacenados en formato de notación científica (por ejemplo 2.01e-05), e incluso cadenas de texto como “7.28x-06”.

Campo “duration_ms”

Contiene el total de canciones asociadas a cada álbum en el dataset. Sin embargo, acepta valores tanto numéricos como tipo texto. Por ejemplo, en vez del valor “13”, puede recibir el valor “Thirteen”. Se recomienda definir restricciones para aceptar solo valores numéricos, de modo que se puedan hacer comparaciones entre álbumes de manera más sencilla.

De igual forma, se hallaron valores atípicos como se muestra en la figura 2, con duraciones negativas y valores extremos (pistas con una duración muy larga o corta).

Figura 2. Distribución campo “duration_ms”



Las duraciones cercanas a 10 minutos (600000 milisegundos) concuerdan con las pistas que presentan versiones extendidas. Por el contrario, las canciones con valores cercanos a cero (pistas de menos de 10 segundos) son errores en los registros.

Campo “track_popularity”

Acorde con la documentación de Spotify, el valor para este campo debe estar entre 0 y 100, entre más alto el valor, mas popular es la canción. Sin embargo, se encontraron valores menores a cero (valores negativos) y superiores a 100.

Campo “album_release_date”

Este campo hace referencia a la fecha de lanzamiento del álbum y, aunque presenta un formato consistente, se hallaron 2 valores atípicos.

En primer lugar, se identificó que el álbum “Midnights (The Til Dawn Edition)” tiene como fecha de lanzamiento el 26/05/2027, valor superior al año de elaboración del reporte. Igualmente, se observó que el álbum “Taylor Swift” se lanzó el 24/10/1989, lo cual no concuerda con la fecha de nacimiento de la artista (13/12/1989).

En este sentido, se recomienda establecer restricciones al momento de ingresar la información en este campo.

UNICIDAD

5. **Registros duplicados**

Se identificaron 36 registros duplicados (18 registros únicos), equivalentes al 3,3% del total de registros (calculado sobre los registros únicos). Los valores corresponden a los álbumes “Lover” y “Midnights (The Til Dawn Edition)”.

REFERENCIAS

- Defining Data Quality Dimensions (2013). DAMA Working Group
- ISO/IEC 25012
- Spotify web API – Track
- Spotify Web API - Album
- Spotify Web API - Artist
- Spotify Web API - Track Audio Features