Reporte de calidad de datos: Spotify API

Data Quality Engineer - Prueba Técnica

Jhon Jairo Silva Zabala

Estudiante Ingeniería de Sistemas y Computación

A continuación se resumen los resultados del proceso de validación de calidad de datos para el dataset que fue proporcionado por la empresa. El proceso consistió en la identificación de anomalías dentro del dataset siempre alrededor de las seis dimensiones de calidad de datos recomendadas.

Identificación de valores nulos.

La completitud del dataset es una medida que representa el número de entradas no nulas. En la evaluación de valores no nulos se hace un conteo del total de entradas que no representen un valor faltante y se obtiene un porcentaje de estos sobre el total de registros.

Identificación de columnas con valores nulos

Column name	Nulls
track_id	8
track_name	7
audio_features.danceability	2
audio_features.energy	2
audio_features.key	1
audio_features.loudness	2
audio_features.speechiness	1
audio_features.acousticness	1
audio_features.liveness	1
audio_features.tempo	1
audio_features.time_signature	1
album_name	62

Tabla 1. Número de valores nulos por columnas. **Nota:** Las columnas que no aparecen no contienen valores nulos.

Con estos valores se procede a calcular el porcentaje de completitud para cada una de las columnas, teniendo en cuenta que el dataset tiene un total de **539** registros o filas, para cada columna un porcentaje de completitud perfecto, esto es 100%, representa una columna que no contiene ninguna entrada nula. Los resultados se muestran en la Figura 1.

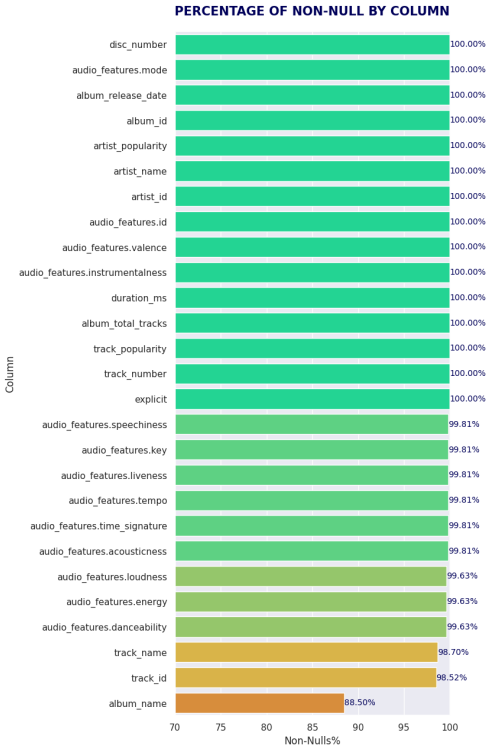

Figura 1. Porcentaje de valores no nulos por columna.

Filas con campos nulos

También es importante calcular el porcentaje de filas que no contienen ningún valor nulo, en muchos casos si una fila contiene algún valor nulo en alguno de sus campos se descarta su uso. Antes de presentar el porcentaje encontrado, se puede visualizar a continuación en la Figura 2 dónde se encuentran los valores nulos en el dataset, esto se realiza a través de la librería missingno, el gráfico muestra el dataset en forma tabular donde los espacios blancos muestran la ubicación de los valores nulos.

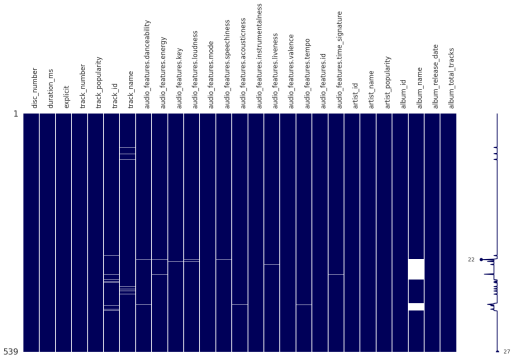

Figura 2. Matriz de completitud donde se muestra la ubicación de los valores nulos (espacios blancos).

Se encontró que existen **74 registros (filas)** que en alguno de sus campos tienen al menos una entrada que es nula.

539 Filas	**86.27%** De las filas no contienen campos cam nulos, esto son **465** filas.
74 Filas con nulos	

Por último se sabe que el tamaño del dataset es de 539 x 27 = **14553** entradas, de estas **99.39%** son entradas no nulas, lo que corresponde a **14464** entradas, este es el porcentaje que representa la completitud final del dataset sin embargo se recomienda siempre revisar las estadísticas anteriores.

99.39 % entradas no nulas, esto son 14.464 entradas de 14.553

Verificación de periodos de tiempo.

En la verificación de periodos de tiempo se consideró la única columna de tipo datetime del dataset, esta es album_release_date. Según **Spotify Web API** la columna representa la fecha en la que fue lanzado el álbum por primera vez. Por fuentes externas se sabe que el primer álbum de Taylor Swift fue lanzado el 24 de Octubre de 2006[1] homónimo a la artista y el último fue lanzado el 27 de Octubre de 2023[2] con el 1989 (Taylor 's Version). En la Figura 3 se ve en qué periodos están distribuidas las fechas de lanzamiento.

Figura 3. Distribución temporal de las canciones de acuerdo a la fecha de lanzamiento de los álbumes.

Se puede ver que existen fechas fuera del rango de tiempo definido. Por tanto se procede a contar el número de fechas que están fuera de este rango para calcular un porcentaje de consistencia temporal.

539 Fechas	**15** < 2006-10-24	**24** > 2023-10-27
500 Fechas en el rango correcto	**88.68%** De las fechas se encuentran en el rango correcto.	

Validación de registros únicos

Se desea encontrar el número total de registros duplicados, se debe tener en cuenta que no es suficiente contar si dos registros son iguales pues debido a inconsistencias encontradas en el dataset es mejor validar la unicidad mediante llaves primarias. Contando únicamente registros duplicados se encontraron **18** registros duplicados, estos son registros exactamente iguales.

Selección de llave primaria

Se selecciona una de las columnas como llave primaria, esto se realiza con el propósito de tener un identificador único para cada registro y así poder contar posteriormente los registros no únicos. Los identificadores posibles son:

- track_id: cada canción se identifica con una cadena de 22 caracteres, se descarta su uso debido a que en el análisis de valores nulos
- **audio_features.id**: cada canción tiene unas características de audio asociadas, estas características contienen un ID de 22 caracteres, se puede usar como identificador para los registros.

Conteo de llaves primarias repetidas

Se realiza un conteo de los registros que tienen llave primaria duplicada se usa el método duplicated() de Pandas para la columna, audio_features.id, se encontró que existen **20** registros que tienen la misma llave, de los cuales 18 registros son los mismos identificados anteriormente. Los 2 registros adicionales que aparecen contando las llaves primarias presentan inconsistencias en los datos, estos son identificados por llave primaria.

- **1BxfuPKGuaTgP7aM0Bbdwr:** La columna explicit contiene el valor **"No"** su duplicado el valor **"False"** por esto no se identifica como registro duplicado sin embargo es evidente que son el mismo registro.

[1] Según Wikipedia y otras fuentes fecha del primer album de Taylor Swift: https://es.wikipedia.org/wiki/Anexo:Discograf%C3%ADa_de_Taylor_Swift

[2] Según Wikipedia y otras fuentes fecha del último album de Taylor Swift: https://es.wikipedia.org/wiki/Anexo:Discograf%C3%ADa_de_Taylor_Swift

- **1ZY1PqizIl78geGM4xWlEA:** La columna track_id contiene un valor nulo su duplicado el ID de la canción por esto no es detectado como registro duplicado pero todos los demás campos son iguales..

En total se concluye que existen 20 registros duplicados.

Evaluación de formatos correctos

Mediante la documentación de **Spotify Web API** se establece un marco de trabajo para validar que los registros se encuentren en los formatos, rangos, tipos y categorías correctas.

Consistencia en los tipos de datos

Se identificó el tipo actual de dato y el tipo deseado, se resume en la Tabla 2.

Column Name	Current Data Type	Convert to
disc_number	int64	int64
duration_ms	int64	int64
explicit	object	bool
track_number	int64	int64
track_popularity	int64	int64
track_id	object	object
track_name	object	object
audio_features.danceability	float64	float64
audio_features.energy	float64	float64
audio_features.key	float64	int64
audio_features.loudness	float64	float64
audio_features.mode	int64	int64
audio_features.speechiness	float64	float64
audio_features.acousticness	float64	float64
audio_features.instrumentalness	object	float64
audio_features.liveness	float64	float64
audio_features.valence	float64	float64
audio_features.tempo	float64	float64
audio_features.id	object	object
audio_features.time_signature	float64	int64
artist_id	object	object
artist_name	object	object
artist_popularity	int64	int64
album_id	object	object
album_name	object	object
album_release_date	object	datetime
album total tracks	object	int64

Tabla 2. Tipos de datos detectados y tipos de datos deseados para cada columna

Para las columnas explicit, audio_features.key, audio_features.instrumentalness, audio_features.time_signature, album_release_date y album_total_tracks los tipos de datos no son detectado correctamente. A continuación en la figura 4 se presentan las inconsistencias halladas.

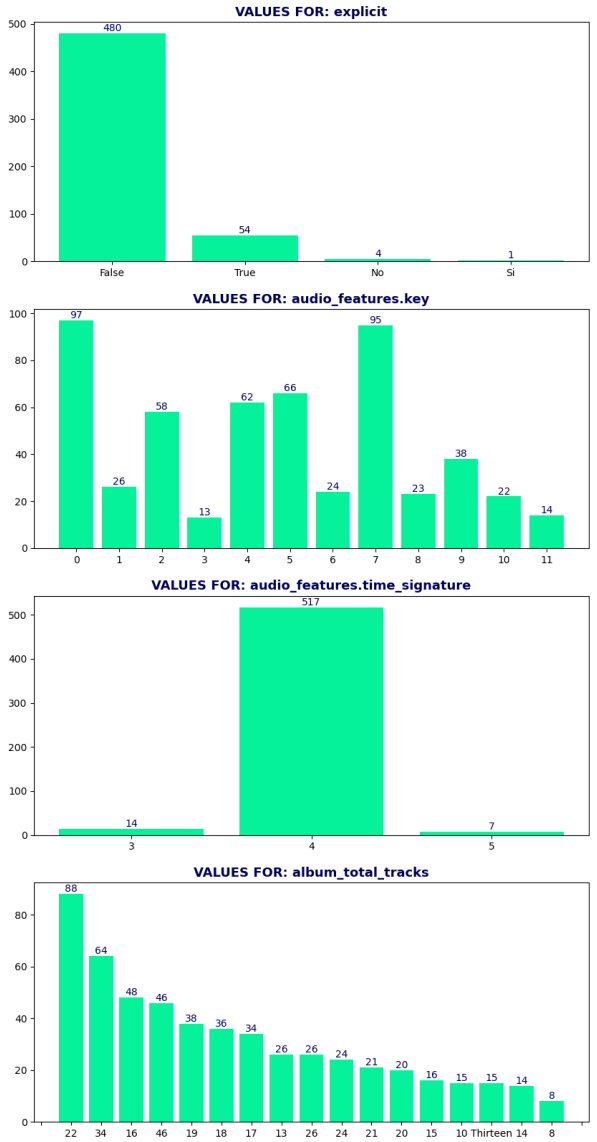

Figura 4. Conteo de valores para las columnas explicit, album_total_tracks, audio_features.key y audio_features.time_signature.

Las columnas explicit y album_total_tracks tienen entradas no consistentes, en explicit hay valores como "Si" y "No" que no representan un valor booleano, mientras para la columna album_total_tracks hay una cadena de texto "Thirteen" lo que no es un entero. Las columnas audio_features.key y audio_features.time_signature no contienen valores inconsistentes que impidan su conversión a int64.

Para la columna album_release_time se convirtió a tipo datetime de Pandas todas las entradas fueron convertidas exitosamente por lo que las fechas se encuentran en el formato adecuado. Por último la columna audio_features.instrumentalness se encontró un valor inconsistente este es la cadena "7.28X-06".

El porcentaje de consistencia de tipos se muestra a continuación en la figura 5 para cada columna.

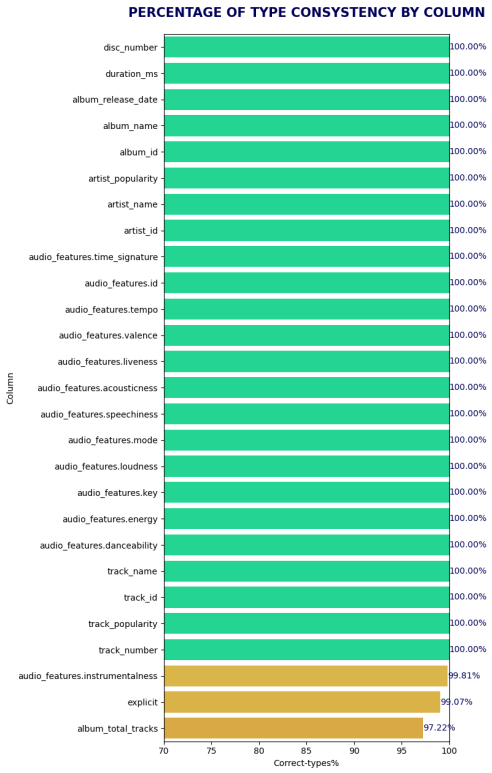

Figura 5. Porcentaje de consistencia de tipos para cada columna.

Consistencia en los rangos

De acuerdo a la documentación de **Spotify Web API** se identificó que los valores para algunas columnas estaban restringidos, se resumen a continuación en la Tabla 3.

Column Name	Mínimo	Máximo
disc_number	1	inf
duration_ms	0	inf
track_number	1	inf
track_popularity	0	100
audio_features.danceability	0	1
audio_features.energy	0	1
audio_features.key	-1	11
audio_features.loudness	-60	0
audio features.mode	0	1
audio_features.speechiness	0	1
audio_features.acousticness	0	1
audio_features.instrumentalness	0	1
audio_features.liveness	0	1
audio_features.valence	0	1
audio_features.tempo	0	
audio_features.time_signature	3	7
artist_popularity	0	100
album_total_tracks	1	inf

Tabla 3. Columnas con restricciones en los rangos.

Las columnas duration_ms, track_popularity y audio_features.acousticness contienen valores fuera de los rangos presentados, a continuación en la Figura 6 se presenta el resultado para todas las columnas, para las columnas sin rango el porcentaje se establece en 100%.

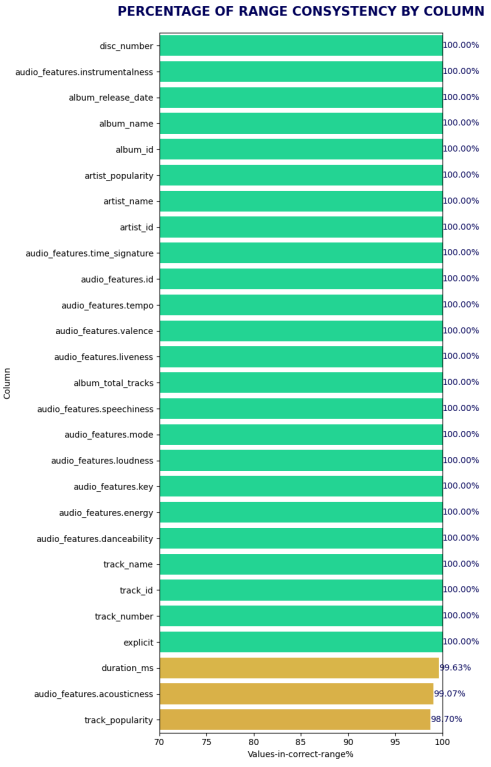

Figura 6. Porcentaje de consistencia de rangos para cada columna.

Consistencia en las categorías

Dos columnas representan categorías estas son audio_features.key y audio_features.mode. Los valores que representan se muestran a continuación.

- **audio_features.key:** esta columna representa la clave o nota en la que está la canción, por lo que se puede considerar una categoría. Las categorías que puede tomar esta columna son las siguientes:
  - **0**: C
  - **1**: C♯/D ♭
  - **2**: D
  - **3**: D♯/E ♭
  - **4**: E
  - **5**: F
  - **6**: F♯/G ♭

    - **7**: G
    - **8**: G♯/A ♭
    - **9** :A
    - **10**: A♯/B ♭
    - **11**: B

- **audio_features.mode:** el modo de la canción, puede ser mayor o menor, por lo que se puede considerar una categoría. Las categorías que puede tomar esta columna son las siguientes:
    - **0**: menor
    - **1**: mayor

Al convertir estos valores a tipo category de Pandas todas las categorías fueron reconocidas exitosamente pues no existen inconsistencias en el formato de las entradas por tanto se concluye el porcentaje de consistencia de categorías es 100%.

Conclusiones

Los resultados finales para las dimensiones evaluadas se resumen a continuación, estos representan el porcentaje de entradas que no se encuentran dentro de las anomalías.

Completeness

99.39% completo

Existen 14553 entradas de estas 14464 son valor no nulos.

12 columnas

tienen al menos un valor nulo, el porcentaje de completitud está por encima del **85%** para todas las columnas.

74 filas

tiene al menos uno de sus campos con valores nulos, el porcentaje de filas con todos los valores no nulos es **86.27%**

Timeliness

88.68% fechas

se encuentran dentro del rango temporal adecuado, determinado por las fechas de lanzamiento de los álbumes de la artista.

15 fechas

por debajo del límite inferior.

24 fechas

por encima del límite superior

Uniqueness

96.29% filas

Un total de 20 registros son duplicados identificados por la llave primaria seleccionada.

Validity

99.76% datos válidos

que cumplieron con consistencias de tipo, categóricas , rangos correctos y formato correctos.

Consistency

100% entradas

coinciden con el dataset extraído del fichero .json

References

DAMA UK Working Group. (2013, October). *THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT*. DAMA UK Working Group.

Gartner. (2021, July 14). *How to Improve Your Data Quality*. Gartner. Retrieved January 7, 2024, from https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality

IBM. (2022, February 24). *Data quality dimensions*. IBM. Retrieved January 06, 2024, from https://www.ibm.com/docs/fr/iis/11.7?topic=results-data-quality-dimensions

Anéxo

Repositorio de **GitHub**.