REPORTE DE CALIDAD DE DATOS - GRUPO R5

Este reporte busca dar a conocer los resultados obtenidos al realizar un análisis de la calidad de los datos contenidos en el archivo 'dataset.csv', el cual contiene información sobre álbumes de Taylor Swift, sus canciones y sus características. Este archivo cuenta con 539 registros y 27 columnas o variables.

En la verificación del tipo de variable o dato se encontró que cinco variables presentaban inconsistencias. Estas se aprecian en la Tabla 1.

Tabla 1. Variables con tipo de variable incorrecto.

VARIABLE	TIPO EN ARCHIVO	TIPO CORRECTO
explicit	object - str	bool
audio_features.key	float	int
audio_features.instrumentalness	object - str	float
audio_features.time_signature	float	int
album_total_tracks	object - str	int

Al evaluar la ausencia de datos se obtuvo que 12 variables tienen valores nulos, estas variables se evidencian en la Tabla 2. También se encontraron 18 registros duplicados.

Tabla 2. Variables con valores o datos nulos.

VARIABLE	CANTIDAD DE VALORES NULOS
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

Realizando la validación exhaustiva del rango y tipo de las variables, se utilizó la información de referencia que se muestra en la Tabla 3, extraida de la página *https://developer.spotify.com/documentation/web-api/reference/*

Tabla 3. Referencia de rango y tipo de variable.

VARIABLE	RANGO Y TIPO
disc_number	Entero mayor o igual a 1
duration_ms	Entero mayor a 0
explicit	Booleano True ó False
track_number	Entero
track_popularity	Entero entre 0 y 100
track_id	Cadena de carácteres
track_name	Cadena de carácteres
audio_features.danceability	Float entre 0 y 1
audio_features.energy	Float entre 0 y 1
audio_features.key	Entero entre -1 y 11
audio_features.loudness	Float usualmente entre -60 y 0
audio_features.mode	Entero 1 ó 0
audio_features.speechiness	Float máximo 1
audio_features.acousticness	Float entre 0 y 1
audio_features.instrumentalness	Float máximo 1
audio_features.liveness	Float máximo 1
audio_features.valence	Float entre 0 y 1
audio_features.tempo	Float positivo
audio_features.id	Cadena de carácteres
audio_features.time_signature	Entero entre 3 y 7
artist_id	Cadena de carácteres
artist_name	Cadena de carácteres
artist_popularity	Entero entre 0 y 100
album_id	Cadena de carácteres
album_name	Cadena de carácteres
album_release_date	Cadena de carácteres
album_total_tracks	Entero mayor a 0

En esta validación de anomalías se encontró que:

Dos canciones presentan duración negativa (duration_ms) y otras tres tienen muy poca duración, en la Tabla 4 se evidencia esto.

Tabla 4. Canciones con duración extraña.

track_id	duration_ms
4eTXfpHxhxVofrBUjAhPMg	-107133
7gJtmLyPTwKzhGzMBXtuXH	-223093
5PjfMmF06QtxTPZBZHdhoZ	10
7mFiEij8AXPUZB7aKLbUlQ	1000
7BFc7ffruhZ4Hecnqf5xju	3000

La variable explicit debería ser de tipo booleano, pero resulta ser de tipo carácter por cinco canciones que tienen las palabras ‘Si’ ó ‘No’. Estas se muestran en la Tabla 5.

Tabla 5. Canciones con anomalías en la variable explicit.

track_id	explicit
0PurA4JVJ8YQgSVopY8fn6	Si
3RaT22zZsxVYxxKR7TAaYF	No
76mOLcXOjOEhyY4mMF1l3r	No
1BxfuPKGuaTgP7aM0Bbdwr	No
0kqC185OTcnuhbz4T7g5RB	No

La variable track_popularity cuenta con 7 registros fuera del rango entre 0 y 100, estos se muestran en la Tabla 6.

Tabla 6. Canciones con anomalías en la variable track_popularity.

track_id	track_popularity
45R112Jz5hQeKglTXgSXzs	-69
4g2c7NoTWAOSYDy44l9nub	-70
5jQI2r1RdgtuT8S3iG8zFC	-85
0V3wPSX9ygBnCm8psDIegu	-92
0heeNYlwOGuUSe7TgUD27B	-75
2r9CbjYgFhtAmcFv1cSquB	-71
7BFc7ffruhZ4Hecnqf5xju	152

Verificando la presencia de duplicados de track_id, se encontraron 19 valores y se determinó que el segundo en la Tabla 7 es uno de los que presenta anomalía en la variable explicit. En términos generales, un track_id pertenece al álbum Midnights (The Til Dawn Edition) y los 18 restantes pertenecen al álbum Lover, por lo que estos últimos se asocian a los 18 duplicados encontrados al inicio del análisis.

Tabla 7. Registros con track_id duplicados

track_id	Álbum
214nt20w5wOxJnY462klLw	Lover
1BxfuPKGuaTgP7aM0Bbdwr	Lover
2dgFqt3w9xIQRjhPtwNk3D	Lover
2YWtcWi3a83pdEg3Gif4Pd	Lover
6RRNNciQGZEXnqk8SQ9yv5	Lover
2Rk4JlNc2TPmZe2af99d45	Lover
3pHkh7d0lzM2AldUtz2x37	Lover
3RauEVgRgj1IuWdJ9fDs70	Lover
1dGr1c8CrMLDpV6mPbImSI	Lover
1LLXZFeAHK9R4xUramtUKw	Lover
1fzAuUVbzlhZ1lJAx9PtY6	Lover
3xYJScVfxByb61dYHTwiby	Midnights (The Til Dawn Edition)
1SymEzIT3H8UZfibCs3TYi	Lover
5hQSXkFgbxjZo9uCwd11so	Lover
12M5uqx0ZuwkpLp5rJim1a	Lover
43rA71bccXFGD4C8GOpIIN	Lover
4AYtqFyFbX0Xkc2wtcygTr	Lover
4y5bvROuBDPr5fuwXbIBZR	Lover
1SmiQ65iSAbPto6gPFlBYm	Lover

En cuanto a track_name, se determinaron 352 registros con nombre de canción que se repetían más de una vez, aunque existen nombres de canciones que pertenecen a diferentes álbumes. Además, entre estos registros que tenían nombre de canción repetido, se encontraron 18 registros duplicados por completo.

Se determinaron anomalías en la variable audio_features.acousticness donde los valores estaban fuera del rango entre 0 y 1, en la Tabla 8 se evidencia lo mencionado.

Tabla 8. Canciones con anomalías en la variable audio_features.acousticness.

track_id	audio_features.acousticness
0108kcWLnn2HlH2kedi1gn	5
1OcSfkeCg9hRC2sFKB4IMJ	-0.000537
3FxJDucHWdw6caWTKO5b23	-0.003540
1oR4MUBpyNrAViC8wPNpfm	1.5
7CzxXgQXurKZCyHz9ufbo1	2

Para la variable audio_features.instrumentalness se halló una anomalía, pues aunque ésta representa números del tipo float entre 0 y 1, estos están expresados en notación científica del tipo 1.02e-06. A partir de esto se encontró un registro que en lugar de tener la letra ‘e’, tiene la letra ‘x’ y esto causa que la variable tome el tipo carácter. Esto se muestra en la Tabla 9.

Tabla 9. Registro con anomalía en audio_features.instrumentalness.

track_id	audio_features.instrumentalness
0Jlcvv8IykzHaSmj49uNW8	7.28x-06

Al revisar la variable audio_features.id se encontraron 20 canciones con este valor duplicado, una es del álbum Midnights (The Til Dawn Edition), 18 pertenecen al álbum Lover y otra al álbum reputation. Estos se muestran en la Tabla 10.

Tabla 10. Registros con duplicados en audio_features.id

audio_features.id	audio_features.id (cont.)
1SmiQ65iSAbPto6gPFlBYm	2Rk4JlNc2TPmZe2af99d45
43rA71bccXFGD4C8GOpIlN	6RRNNciQGZEXnqk8SQ9yv5
3xYJScVfxByb61dYHTwiby	4y5bvROuBDPr5fuwXbIBZR
1dGr1c8CrMLDpV6mPbImSI	12M5uqx0ZuwkpLp5rJim1a
5hQSXkFgbxjZo9uCwd11so	2YWtcWi3a83pdEg3Gif4Pd
3RauEVgRgj1IuWdJ9fDs70	2dgFqt3w9xIQRjhPtwNk3D
3pHkh7d0lzM2AldUtz2x37	214nt20w5wOxJnY462klLw
1BxfuPKGuaTgP7aM0Bbdwr	1SymEzIT3H8UZfibCs3TYi
1LLXZFeAHK9R4xUramtUKw	4AYtqFyFbX0Xkc2wtcygTr
1ZY1PqizII78geGM4xWlEA	1fzAuUVbzlhZ1lJAx9PtY6

Además de lo anterior, la canción Cruel Summer es un registro duplicado por completo pero uno de ellos tiene un error en explicit y su track_id es 1BxfuPKGuaTgP7aM0Bbdwr. Lo mismo sucede con la canción Gorgeous al ser un registro duplicado por completo, solo que uno de ellos tiene un error en track_id, su audio_features.id es 1ZY1PqizII78geGM4xWIEA.

Continuando con el análisis, se identificó una anomalía en la variable 'artist_popularity', pues todos los registros tienen el valor 120 cuando el máximo es 100.

En la variable 'album_id' se identificaron cinco álbumes con inconsistencias entre el número de canciones que tiene cada álbum según la variable 'album_total_tracks' y el conteo de los diferentes 'album_id'. Estas anomalías se presentan a continuación en la Tabla 11.

Tabla 11. Inconsistencias entre conteo de album_id y album_total_tracks

Número de álbum en tabla	album_id	album_total_tracks	album_id_count
1	1NAmidJlEaVgA3MpcPFYGq	18	36
2	6kZ42qRrzov54LcAk4onW9	34	30
3	6DEjYFkNZh67HP7R9PSZvv	15	16
4	2Xoteh7uEpea4TohMxjtaq	10	15
5	5eyZZoQEFQWRHkV2xgAeBw	13	15

Con respecto a la tabla anterior, el álbum 1 se trata del álbum 'Lover' duplicado, lo que confirma lo evidenciado anteriormente donde la mayoría de los duplicados encontrados hacen referencia a este álbum. El álbum 2 parece tener una anomalía en album_total_tracks o tal vez faltan canciones. En el álbum 3 se encontró que el registro adicional corresponde a la canción 'Gorgeous', pues es duplicado pero uno de los dos no tiene track_id. Por último, el caso del álbum 4 y 5 es que podría haber una anomalía en album_total_tracks, puesto que su valor es menor a la cantidad de album_id y no hay duplicados en cada uno de estos 2 álbumes..

Pasando a la variable album_name, se encontraron dos álbumes sin nombre que podrían dar a entender que es una anomalía o que estos dos álbumes fueron eliminados, y por esto quedan sin nombre según las definiciones de las variables en la página de referencia mencionada al introducir la Tabla 3 al inicio del reporte.

La fecha de lanzamiento del álbum contenida en 'album_total_tracks' presenta dos anomalías, el álbum 'Taylor Swift' tiene como fecha de lanzamiento 1989-10-24 y es inconsistente porque la fecha de nacimiento de la artista fue luego en 1989-12-13. La segunda anomalía es en el álbum Midnights (The Til Dawn Edition), por el hecho de presentar como fecha de lanzamiento 2027-05-26 cuando en realidad fue lanzado el 2023-05-26.

Por último, al analizar la variable album_total_tracks se encontró que el álbum ‘Taylor Swift’ tiene escrito en inglés el número de canciones del álbum, o sea, ‘Thirteen’ en lugar de 13. Eso causa que la variable sea de tipo carácter y no tipo entero como debería ser.