Prueba técnica Karen Daniela López

Análisis de calidad de datos

Como primer paso para realizar el análisis de calidad de datos propongo realizar un diccionario de variables, ya que esto va a permitir conocer los tipos de datos, rangos y en general va a brindar un panorama sobre lo que podría o no permitirse en la revisión de calidad de datos.

Variable	Descripción	Tipo de dato
disc_number	Número del disco	Entero
duration_ms	Longitud de la pista en milisegundos	Entero
explicit	Si la pista tiene o no letras explícitas	Booleano
track_number	Número de la pista	Entero
track_popularity	Popularidad de la pista. Debe estar entre 0 y 100	Entero
track_id	ID de la pista en Spotify	Texto
track_name	Nombre de la pista	Texto
audio_features.danceability	Medida de 0.0 a 1.0 que describe qué tan adecuada es la pista para bailar	Decimal
audio_features.energy	Medida de 0.0 a 1.0 representa una medida perceptiva de intensidad y actividad	Decimal
audio_features.key	Tonalidad en la que se encuentra la pista. Rango (-1,-11)	Entero
audio_features.loudness	La sonoridad general de una pista en decibelios (dB). Rango (-60 , 0)	Decimal
audio_features.mode	El modo indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. Mayor está representado por 1 y menor es 0.	Entero
audio_features.speechiness	detecta la presencia de palabras habladas en una pista. Cuanto más	Decimal

	exclusivamente hablada sea la grabación	
audio_features.acousticness	Una medida de confianza de 0.0 a 1.0 sobre si la pista es acustica	Decimal
audio_features.instrumentalness	Medida que predice si una pista contiene o no voces	Decimal
audio_features.liveness	Detecta la presencia de audiencia en la grabación	Decimal
audio_features.valence	Una medida de 0.0 a 1.0 que describe la positividad musical transmitida por una pista.	Decimal
audio_features.tempo	El tempo general estimado de una pista en pulsaciones por minuto (BPM)	Decimal
audio_features.id	ID de la pista en spotify	Texto
audio_features.time_signature	Un compás estimado. Rango entre 3,7	Entero
artist_id	ID del artista en Spotify	Texto
artist_name	Nombre del artista	Texto
artist_popularity	La popularidad del artista. El valor estará entre 0 y 100, siendo 100 el más popular. La popularidad del artista se calcula a partir de la popularidad de todas sus pistas.	Entero
album_id	El ID del álbum en spotify	Texto
album_name	Nombre del álbum. En caso de que se elimine el álbum, el valor puede ser un NA	Texto
album_release_date	Fecha en que se lanzó el álbum	Fecha
album_total_tracks	Número de pistas en el álbum	Entero

Valores faltantes

Una vez realizado el diccionario de datos se procede a realizar un análisis de datos faltantes:

Para conocer la completitud de los datos, se calcula el porcentaje de datos completos para cada variable, esto es:

Número de filas completas/ número total de filas

Una vez realizado el cálculo se encontró:

disc_number	100.000000
duration_ms	100.000000
explicit	100.000000
track_number	100.000000
track_popularity	100.000000
track_id	98.515770
track_name	98.701299
audio_features.danceability	99.628942
audio_features.energy	99.628942
audio_features.key	99.814471
audio_features.loudness	99.628942
audio_features.mode	100.000000
audio_features.speechiness	99.814471
audio_features.acousticness	99.814471
audio_features.instrumentalness	100.000000
audio_features.liveness	99.814471
audio_features.valence	100.000000
audio_features.tempo	99.814471
audio_features.id	100.000000
audio_features.time_signature	99.814471
artist_id	100.000000
artist_name	100.000000
artist_popularity	100.000000
albums.album_id	100.000000
albums.album_name	88.497217
albums.album_release_date	100.000000
albums.album_total_tracks	100.000000

Las variables con un porcentaje menor a 100% tienen observaciones faltantes. Los campos **track_id** y **track_name** presentan datos faltantes a pesar de ser un datos obligatorios. En los campos relacionados con las características de audio también se encuentran algunos datos faltantes, sin embargo estos no son campos obligatorios, por último, hay datos faltantes en **album_name**, este también es un campo obligatorio, sin embargo, los valores ausentes pueden ser porque el álbum ha sido eliminado.

Valores duplicados

Se encontraron 36 filas idénticas dentro del dataset, también se encontró 46 duplicados para la columna track_id.

Tipos de datos

A continuación, se verifica que los tipos de datos correspondan a lo que está en la documentación.

```
disc_number                      int64
duration_ms                      int64
explicit                        object
track_number                     int64
track_popularity                 int64
track_id                        object
track_name                      object
audio_features.danceability    float64
audio_features.energy          float64
audio_features.key             float64
audio_features.loudness        float64
audio_features.mode              int64
audio_features.speechiness     float64
audio_features.acousticness    float64
audio_features.instrumentalness  object
audio_features.liveness        float64
audio_features.valence         float64
audio_features.tempo           float64
audio_features.id                object
audio_features.time_signature  float64
artist_id                        object
artist_name                      object
artist_popularity                int64
albums.album_id                 object
albums.album_name               object
albums.album_release_date       object
albums.album_total_tracks       object
```

Variables como explicit, audio_features.key, audio_features.instrumentalness, audio_features.time_signature, album_release_date y album_total_tracks no tienen el tipo de dato que corresponde, esto puede ser debido a que las variables tienen niveles que no corresponden, más adelante se podrá identificar.

Niveles de las variables

A continuación, un análisis a los niveles de las variables que no tienen el tipo que corresponde según la documentación.

- **explicit** : para la variable se encontraron los siguientes niveles ['False' 'True' 'Si' 'No'] cómo es una variable de tipo booleano lo correcto sería cambiar ‘Si’ por True y ‘No’ por False.
- **audio_features.key:** Los niveles son [7. 0. 11. 8. 5. 1. 4. 3. 9. 2. 10. 6. nan] se puede identificar que no hay niveles de la variable incorrectos, asi que solo sería cambiar el tipo entero.
- **audio_features.instrumentalness:** Dentro de los niveles de la variable se encuentra el valor '7.28x-06' el cual no está definido como un número.
- **audio_features.time_signature:** Los niveles son [4. 3. 5. nan], no hay niveles incorrectos.
- **album_release_date:** Los niveles están bien, se debe cambiar el tipo de dato.
- **album_total_tracks:** Dentro de los niveles se encuentra 'Thirteen' el cual debe ser reemplazado por el número 13.

Rangos de las variables

Mínimos por columna:		Máximos por columna:	
disc_number	1.00000	disc_number	2.000
duration_ms	-223093.00000	duration_ms	613026.000
track_number	1.00000	track_number	46.000
track_popularity	-92.00000	track_popularity	152.000
audio_features.danceability	0.24300	audio_features.danceability	0.897
audio_features.energy	0.11800	audio_features.energy	0.949
audio_features.key	0.00000	audio_features.key	11.000
audio_features.loudness	-17.93200	audio_features.loudness	-1.909
audio_features.mode	0.00000	audio_features.mode	1.000
audio_features.speechiness	0.02310	audio_features.speechiness	0.912
audio_features.acousticness	-0.00354	audio_features.acousticness	5.000
audio_features.liveness	0.03350	audio_features.liveness	0.931
audio_features.valence	0.03740	audio_features.valence	0.943
audio_features.tempo	68.09700	audio_features.tempo	208.918
audio_features.time_signature	3.00000	audio_features.time_signature	5.000
artist_popularity	120.00000	artist_popularity	120.000

Las variables que presentan valores por fuera de su rango son: **duration_ms, track_popularity, audio_features.acousticness y artist_popularity.**

En la variable album_release_date se encontraron los siguientes niveles: ['2023-10-27' '2023-10-26' '2023-07-07' '2027-05-26' '2022-10-22' '2022-10-21' '2021-11-12' '2021-04-09' '2021-01-07' '2020-12-11' '2020-11-25' '2020-08-18' '2020-07-24' '2019-08-23' '2017-11-10' '2017-11-09' '2014-01-01' '2012-10-22' '2010-10-25' '2010-01-01' '2008-11-11' '2008-06-28' '1989-10-24']

Algunas de las inconsistencias son: album con fecha “1989-10-24” ya que el primer álbum de la cantante fue lanzado en 2006, también se puede ver una fecha de

lanzamiento que aún no ocurre (2027-05-06), adicionalmente, se encuentran fechas seguidas '2023-10-27' '2023-10-26' y '2022-10-22' '2022-10-21'.

Dentro de la variable duration_ms también se identificó que hay algunos valores muy pequeños que no corresponde a la duración de las pistas.