Data Quality Engineer – Prueba Técnica

Reto: Spotify API

Reporte de calidad de datos

R5

Oscar Mario Mariño Arias

06/01/2024

Tabla de contenido

Tabla de figuras

Tabla de tablas

Data Quality Engineer – Prueba Técnica
Reto: Spotify API

Reporte de Calidad de datos

R5

1. Introducción:

A partir de un archivo proveniente de la API de Spotify, se realizó un análisis exhaustivo de datos enfocado en encontrar las anomalías del dataset, para posteriormente realizar una evaluación y categorización de estos. Dicho trabajo fue llevado a cabo haciendo uso de Python y la librería Pandas.

2. Metodología:

Para realizar la evaluación de la calidad de los datos se adoptó el enfoque mencionado en DAMA UK (2018):

1. Identificar los datos que deben ser evaluados en calidad.
2. Definir las dimensiones de calidad a emplear y el peso de estas para la evaluación de los datos.
3. Definir los valores y rangos que representan buena y mala calidad de los datos para las diferentes dimensiones.
4. Aplicar los criterios de evaluación al set de datos.
5. Analizar y definir si la calidad de los datos es aceptable.
6. Cuando sea el caso, adoptar medidas correctivas para mejorar la calidad de los datos y evitar futuras recurrencias.

3. Dimensiones de calidad de datos:

A continuación, se describen las dimensiones utilizadas para la evaluación de la calidad de los datos.

Completitud (Completeness): Hace referencia a la proporción de los datos almacenados con relación a su totalidad. Medida de ausencia de valores sin información (nulos).

Unicidad (Uniqueness): Dimensión que hace alusión a si los registros en el set de datos son de carácter único.

Validez (Validity): Los datos son válidos si se ajustan a la sintaxis de su definición, tanto en el formato, como el tipo y rango.

Precisión (Accuracy): Hace referencia al grado en el cual los datos representa correctamente el evento que están describiendo.

Consistencia (Consistency): Es la ausencia de diferencias en la comparación de dos o más registros con relación a la definición del campo.

Es importante mencionar que de acuerdo con DAMA UK (2018), las diferentes dimensiones utilizadas para la evaluación de datos pueden tener diferentes pesos, lo que a su vez repercute en la medida de calidad obtenida para los datos. Para beneficio del siguiente reporte, las dimensiones serán tomadas con igual

peso a la hora de evaluar la información suministrada. Así mismo, todas las características del archivo "dataset.csv" serán analizadas.

4. Rangos de evaluación:

Para la evaluación de los datos y definición de su calidad se fijaron los siguientes rangos:

Tabla 1. Rangos de evaluación.

Rango	Calidad
100 – 96	Buena
<96	Aceptable
<90	Regular
<85	Mala

Dependiendo de la cantidad de errores los datos serán definidos dentro de un rango, lo cual dará como resultado la calidad de estos.

5. Anomalías:

A partir del análisis del set de datos suministrado, el cual cuenta con un total de 539 registros (filas) y 27 características (columnas), múltiples errores fueron evidenciados.

Es de relevancia mencionar que, al usar la librería Pandas para el análisis de los datos, algunas columnas son casteadas de manera errónea por la presencia de registros que tienen formato incompatible. Por lo tanto, dichos registros son resaltados, pero a su vez, la incongruencia en el tipo de dato es tomado como un error adicional. A continuación, son presentados las anomalías observadas:

3.1. Completitud

Características como *track_id*, *track_name*, *danceability*, *energy*, *key*, *loudness*, *speechiness*, *acousticness, liveness*, *tempo, time_signature* y *album_name* presentan registros faltantes (Figura 1). La Tabla 2 detalla la cantidad de errores encontrados para las diferentes características en la dimensión completitud.

```
# Verificamos la completitud de los datos.
df.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 539 entries, 0 to 538
Data columns (total 27 columns):
 #   Column                          Non-Null Count
---  ------                          --------------
 0   disc_number                     539 non-null
 1   duration_ms                     539 non-null
 2   explicit                        539 non-null
 3   track_number                    539 non-null
 4   track_popularity                539 non-null
 5   track_id                        531 non-null
 6   track_name                      532 non-null
 7   audio_features.danceability     537 non-null
 8   audio_features.energy           537 non-null
 9   audio_features.key              538 non-null
 10  audio_features.loudness         537 non-null
 11  audio_features.mode             539 non-null
 12  audio_features.speechiness      538 non-null
 13  audio_features.acousticness     538 non-null
 14  audio_features.instrumentalness 539 non-null
 15  audio_features.liveness         538 non-null
 16  audio_features.valence          539 non-null
 17  audio_features.tempo            538 non-null
 18  audio_features.id               539 non-null
 19  audio_features.time_signature   538 non-null
 20  artist_id                       539 non-null
 21  artist_name                     539 non-null
 22  artist_popularity               539 non-null
 23  album_id                        539 non-null
 24  album_name                      477 non-null
 25  album_release_date              539 non-null
 26  album_total_tracks              539 non-null
```

Figura 1. Completitud de los datos.

Tabla 2. Anomalías encontradas para la dimensión completitud.

Característica	No. de anomalías
track_id	8
track_name	7
**danceability*	2
**energy*	2
**key*	1
**loudness*	2
**speechiness*	1
**acousticness*	1
**liveness*	1
**tempo*	1
**time_signature*	1
album_name	62

Las características marcadas con un (*) presentan los siguientes caracteres “audio_features.” previo al nombre del campo.

Las columnas del tipo flotante (float) y entero (int), presentan registros con datos NaN, que hacen referencia a datos faltantes o vacíos (Figura 2A). Para las columnas del tipo carácter (str), los registros sin datos son representados igualmente por valores NaN del tipo flotante (Figura 2B).

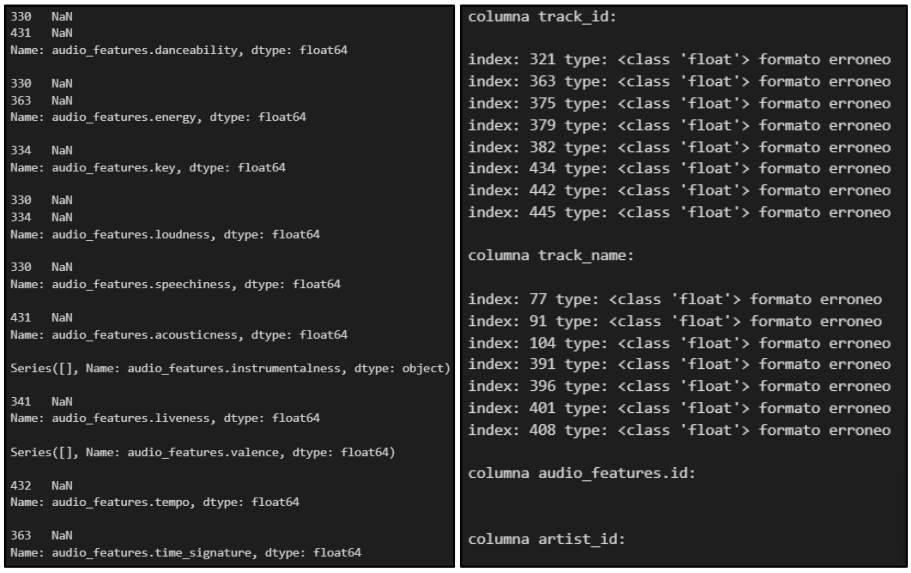


Figura 2. Registros vacíos. A) Columnas del tipo float e Int. B) Columnas del tipo str.

En la característica *album_name* es necesario tener en cuenta que, de acuerdo con la documentación de Spotify, el valor de dichos registros puede estar vacío (nulo) debido a que el álbum fue eliminado, no obstante, fue posible evidenciar que el registro existe para los álbumes “Speak Now World Tour”, y “reputation Stadium Tour Surprise Song “ en la API, por lo tanto, constituye una anomalía (Figura 3).

```
columna album_name:

index: 329 type: <class 'float'> formato erroneo. nan
index: 330 type: <class 'float'> formato erroneo. nan
index: 331 type: <class 'float'> formato erroneo. nan
index: 332 type: <class 'float'> formato erroneo. nan
index: 333 type: <class 'float'> formato erroneo. nan
index: 334 type: <class 'float'> formato erroneo. nan
index: 335 type: <class 'float'> formato erroneo. nan
index: 336 type: <class 'float'> formato erroneo. nan
index: 337 type: <class 'float'> formato erroneo. nan
```

Figura 3. Ejemplo de algunos registros vacíos en columna *album_name*.

3.2. Unicidad

El set de datos presenta 19 registros que se encuentran repetidos, el álbum “Lover” con 18 registros y un registro del álbum “Midnights (The Til Dawn Edition)” (Figura 4). Debido a que en la característica "explicit" hay un error en el formato de la casilla, el registro completo es considerado como diferente, sin embargo, al corregir el formato es posible evidenciar que la fila es un registro adicional repetido. En la Tabla 3 es posible observar la cantidad de errores de cada característica relacionados con la dimensión Unicidad.

```
# Dimensiones iniciales del set de datos.
print(f'Dimensiones iniciales: {df.shape}')
print(f'Dataset sin registros repetidos: {df.drop_duplicates().shape}. No tiene en cuenta la fila que tiene el registro con formato erroneo')
print(f'Registros repetidos: {df[df.duplicated() == True].shape}. No tiene en cuenta la fila que tiene el registro con formato erroneo')
print('')
print('Ejemplo de algunos registros repetidos:')
display(df[df.duplicated() == True][['artist_name', 'album_name','track_name']].sort_values(by= 'track_name').sample(10))

# Dado a que en la característica "explicit" hay un error en el formato de la casilla el registro es considerado como diferente, sin embargo,
# al corregir esto podemos evidenciar que es un registro adicional repetido.
df[(df['album_name'] == 'Lover') &  (df['track_name'] == 'Cruel Summer')][['explicit', 'album_name', 'track_name']].sort_values(by= 'track_name')
```

✓ 0.0s

```
Dimensiones iniciales: (539, 27)
Dataset sin registros repetidos: (521, 27). No tiene en cuenta la fila que tiene el registro con formato erroneo
Registros repetidos: (18, 27). No tiene en cuenta la fila que tiene el registro con formato erroneo

Ejemplo de algunos registros repetidos:
```

	artist_name	album_name	track_name
295	Taylor Swift	Lover	I Forgot That You Existed
306	Taylor Swift	Lover	Soon You'll Get Better (feat. The Chicks)
297	Taylor Swift	Lover	Lover
304	Taylor Swift	Lover	Death By A Thousand Cuts
303	Taylor Swift	Lover	Cornelia Street
312	Taylor Swift	Lover	Daylight
300	Taylor Swift	Lover	I Think He Knows
310	Taylor Swift	Lover	ME! (feat. Brendon Urie of Panic! At The Disco)
309	Taylor Swift	Lover	Afterglow
305	Taylor Swift	Lover	London Boy

	explicit	album_name	track_name
278	False	Lover	Cruel Summer
296	No	Lover	Cruel Summer

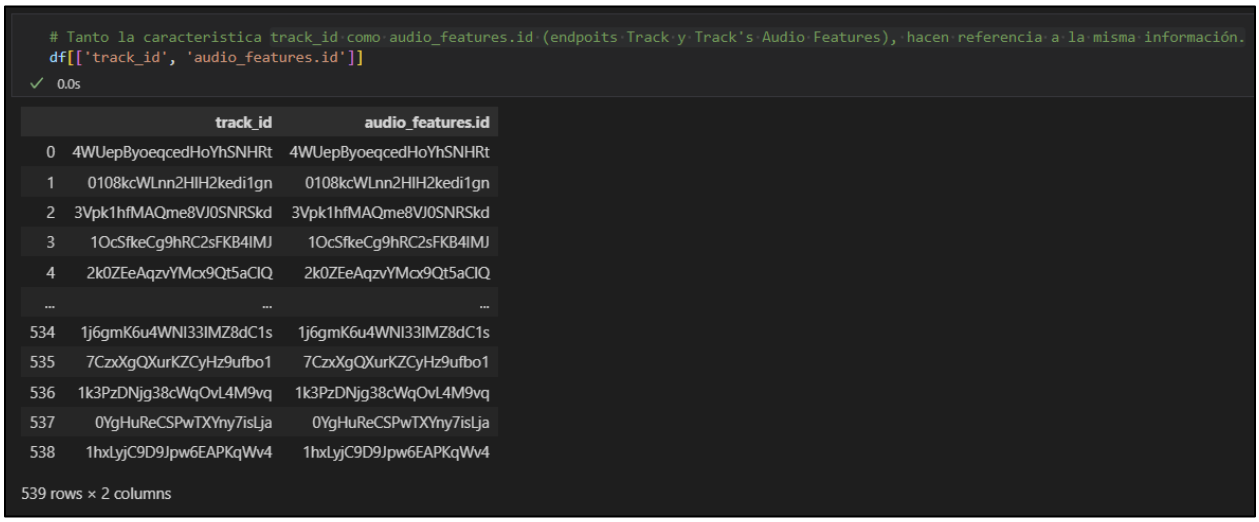
Figura 4**.** Registros repetidos en el set de datos.

Tabla 3. Anomalías encontradas para la dimensión unicidad.

Característica	No. de anomalías	Característica	No. de anomalías
disc_number	19	**instrumentalness*	19
duration_ms	19	**liveness*	19
explicit	19	**valence*	19
track_number	19	**tempo*	19
track_popularity	19	**id*	539
track_id	19	**time_signature*	19
track_name	19	*artist_id*	19
**danceability*	19	*artist_name*	19
**energy*	19	*artist_popularity*	19
**key*	19	*album_id*	19
**loudness*	19	*album_name*	19
**mode*	19	*album_release_date*	19
**speechiness*	19	*álbum_total_tracks*	19
**acousticness*	19	-	-

Las características marcadas con un (*) presentan los siguientes caracteres “audio_features.” previo al nombre del campo.

Además, las columnas *track_id* y *audio_features.id* pertenecientes a los endpoits Track y Track's Audio Features, hacen referencia a la misma información, por lo tanto, son columnas con información repetida (Figura 5). No obstante, dado que *track_id* tiene 8 casillas vacías algunos registros pueden ser tomados como diferentes.



```
# Tanto la caracteristica track_id como audio_features.id (endpoits Track y Track's Audio Features), hacen referencia a la misma información.
df[['track_id', 'audio_features.id']]
```

✓ 0.0s

	track_id	audio_features.id
0	4WUepByoeqcedHoYhSNHRt	4WUepByoeqcedHoYhSNHRt
1	0108kcWLnn2HlH2kedi1gn	0108kcWLnn2HlH2kedi1gn
2	3Vpk1hfMAQme8VJ0SNRSkd	3Vpk1hfMAQme8VJ0SNRSkd
3	1OcSfkeCg9hRC2sFKB4IMJ	1OcSfkeCg9hRC2sFKB4IMJ
4	2k0ZEeAqzvYMcx9Qt5aCIQ	2k0ZEeAqzvYMcx9Qt5aCIQ
...	...	...
534	1j6gmK6u4WNI33IMZ8dC1s	1j6gmK6u4WNI33IMZ8dC1s
535	7CzxXgQXurKZCyHz9ufbo1	7CzxXgQXurKZCyHz9ufbo1
536	1k3PzDNjg38cWqOvL4M9vq	1k3PzDNjg38cWqOvL4M9vq
537	0YgHuReCSPwTXYny7isLja	0YgHuReCSPwTXYny7isLja
538	1hxLyjC9D9Jpw6EAPKqWv4	1hxLyjC9D9Jpw6EAPKqWv4

539 rows × 2 columns

Figura 5. Características con el mismo tipo de información.

Debido a que las diferentes dimensiones se encuentran correlacionadas, la presencia de datos repetidos genera que la columna con el número de canciones por álbum no sea acorde con el número real de registros por álbum (número de canciones), como será evidenciado más adelante.

3.3. Validez

Como se mencionó, las diferentes dimensiones se encuentran relacionadas e influyen entre sí. En este caso, la validez de los datos es afectada por la completitud de los registros.

La columna *explicit,* que de acuerdo con la documentación de Spotify es del tipo booleano, presenta 5 registros de cadena de caracteres que son errores para el campo (Figura 6). Así mismo, los valores True y False, son catalogados como cadena de caracteres y precisan corrección.

```
# El formato de los datos de la columna explicit, es del tipo booleano, sin embargo, algunos de los datos son de tipo caracter.
print(df['explicit'].value_counts())

# Debido esto genera que registros que se encuentran repetidos sean tomados como diferentes.
df[(df['album_name'] == 'Lover') &  (df['track_name'] == 'Cruel Summer')][['explicit', 'album_name', 'track_name']].sort_values(by= 'track_name')
```

✓ 0.0s

```
False    480
True      54
No         4
Si         1
Name: explicit, dtype: int64
```

Figura 6. Registros de tipo cadena de caracteres en columna del tipo booleano.

Las características del tipo flotante (float) *audio_features.danceability, audio_features.energy, audio_features.key, audio_features.loudness, audio_features.speechiness, audio_features.acousticness, audio_features.liveness*, *audio_features.tempo* y *audio_features.time_signature*, presentan registros catalogados como NaN (valores no numéricos) ya que se encuentran vacíos en el set de datos (Figura 7).

```
# Caracteristicas del tipo Float con registros NaN (valores no numéricos).
for feature in df[['audio_features.danceability','audio_features.energy','audio_features.key','audio_features.loudness','audio_features.speechiness','audio_features.acousticness',
                   'audio_features.instrumentalness', 'audio_features.liveness','audio_features.valence','audio_features.tempo','audio_features.time_signature']]:
    display(df[feature][df[feature].isnull()])
```

✓ 0.0s

```
330   NaN
431   NaN
Name: audio_features.danceability, dtype: float64

330   NaN
363   NaN
Name: audio_features.energy, dtype: float64

334   NaN
Name: audio_features.key, dtype: float64

330   NaN
334   NaN
Name: audio_features.loudness, dtype: float64

330   NaN
Name: audio_features.speechiness, dtype: float64

431   NaN
Name: audio_features.acousticness, dtype: float64

Series([], Name: audio_features.instrumentalness, dtype: object)

341   NaN
Name: audio_features.liveness, dtype: float64

Series([], Name: audio_features.valence, dtype: float64)

432   NaN
Name: audio_features.tempo, dtype: float64

363   NaN
Name: audio_features.time_signature, dtype: float64
```

Figura 7**.** Registros vacíos en columnas del tipo flotante.

Por otra parte, la característica *audio_features.time_signature* que es del tipo entero (int) es catalogada como flotante ya que presenta un valor vacío (no numérico, index: 363) definido del tipo flotante (Figura 7).

Además, la característica del tipo flotante *audio_features.instrumentallness,* es definida del tipo cadena de caracteres debido a que incluye un registro (index [524]) que presenta un formato erróneo en el exponencial, por lo cual, la columna queda definida como cadena de caracteres (Figura 8).

Al mismo tiempo, es necesario tener en cuenta que esta característica presenta 235 registros con valor 0. Aunque dichos valores parecieran ser errores de los datos (sin tener en cuenta que Pandas los catalogó

como cadena de caracteres), de acuerdo con la documentación de Spotify, para la columna *audio_features.instrumentallness* no hay un rango definido, por ende, el valor numérico en si no es catalogado como erróneo (Figura 8).

```
# # Registro con error en el formato original.
print(df[df['audio_features.instrumentalness'] == '7.28x-06'].index)
print(df['audio_features.instrumentalness'][524])
# Número de registros en 0
display(df['audio_features.instrumentalness'].apply(lambda x: float(x.replace('x','e')))[df['audio_features.instrumentalness'].apply(lambda x: float(x.replace('x','e'))) == 0])
```

✓ 0.0s

```
Int64Index([524], dtype='int64')
7.28x-06

1      0.0
4      0.0
7      0.0
10     0.0
11     0.0
      ...
533    0.0
534    0.0
536    0.0
537    0.0
538    0.0
Name: audio_features.instrumentalness, Length: 235, dtype: float64
```

Figura 8**.** Errores en característica *audio_features.instrumentallness*.

De igual forma La columna *album_total_tracks* del tipo entero es catalogado como cadena de caracteres ya que 15 registros son escritos (str) (Figura 9).

```
cantidad = 0
for i, j in enumerate(df['album_total_tracks']):
    try:
        int(j)
    except:
        print(f'index: {i} type: {type(j)} {j} formato original erroneo')
        cantidad +=1

print(f'Cantidad de registros con el formato erroneo: {cantidad}')
```

✓ 0.0s

```
index: 524 type: <class 'str'> Thirteen formato original erroneo
index: 525 type: <class 'str'> Thirteen formato original erroneo
index: 526 type: <class 'str'> Thirteen formato original erroneo
index: 527 type: <class 'str'> Thirteen formato original erroneo
index: 528 type: <class 'str'> Thirteen formato original erroneo
index: 529 type: <class 'str'> Thirteen formato original erroneo
index: 530 type: <class 'str'> Thirteen formato original erroneo
index: 531 type: <class 'str'> Thirteen formato original erroneo
index: 532 type: <class 'str'> Thirteen formato original erroneo
index: 533 type: <class 'str'> Thirteen formato original erroneo
index: 534 type: <class 'str'> Thirteen formato original erroneo
index: 535 type: <class 'str'> Thirteen formato original erroneo
index: 536 type: <class 'str'> Thirteen formato original erroneo
index: 537 type: <class 'str'> Thirteen formato original erroneo
index: 538 type: <class 'str'> Thirteen formato original erroneo
Cantidad de registros con el formato erroneo: 15
```

Figura 9. Anomalías del tipo cadena de caracteres en columna del tipo entero.

Los errores relacionados con los registros de las columnas de formato cadena de caracteres están relacionados a registros faltantes del tipo NaN (Figura 10), descritos previamente en la dimensión de completitud (Figura 2A).

Finalmente, características con rango definido como *track_popularity*, *audio_features.acousticness* y *artist_popularity* presentan registros que exceden los límites permitidos (Figura 11).

Por su parte *track_popularity* presenta 7 registros por fuera de su rango (0 – 100), mientras que *audio_features.acousticness* tiene 5 datos por fuera del rango 0 - 1 , y en *artist_popularity* todos los registros exceden el límite (max: 100).

```
# Tipo de dato erroneo en caracteristicas del tipo string
for feature in df[['audio_features.id', 'track_id', 'track_name', 'audio_features.id', 'artist_id', 'artist_name', 'album_id', 'album_name', 'album_release_date']].columns:
    print('')
    print(f'columna {feature}:')

    for i, j  in enumerate(df[feature]):
            if type(j) != str:
                print(f'index: {i} type: {type(j)} formato erroneo. Valor {j}')
```

✓ 0.0s

```
columna audio_features.id:

columna track_id:
index: 321 type: <class 'float'> formato erroneo. Valor nan
index: 363 type: <class 'float'> formato erroneo. Valor nan
index: 375 type: <class 'float'> formato erroneo. Valor nan
index: 379 type: <class 'float'> formato erroneo. Valor nan
index: 382 type: <class 'float'> formato erroneo. Valor nan
index: 434 type: <class 'float'> formato erroneo. Valor nan
index: 442 type: <class 'float'> formato erroneo. Valor nan
index: 445 type: <class 'float'> formato erroneo. Valor nan

columna track_name:
index: 77 type: <class 'float'> formato erroneo. Valor nan
index: 91 type: <class 'float'> formato erroneo. Valor nan
index: 104 type: <class 'float'> formato erroneo. Valor nan
index: 391 type: <class 'float'> formato erroneo. Valor nan
index: 396 type: <class 'float'> formato erroneo. Valor nan
index: 401 type: <class 'float'> formato erroneo. Valor nan
index: 408 type: <class 'float'> formato erroneo. Valor nan

columna audio_features.id:

columna artist_id:
...
index: 443 type: <class 'float'> formato erroneo. Valor nan
index: 444 type: <class 'float'> formato erroneo. Valor nan
```

Figura 10. Campos vacíos en columnas del tipo cadena de caracteres.

```
# Registros por fuera del rango apropiado.
for feature in df[['track_popularity','audio_features.acousticness', 'artist_popularity', 'audio_features.danceability', 'audio_features.key', 'audio_features.energy',
                   'audio_features.valence', 'audio_features.time_signature']]:
    print('')
    print(f'columna {feature}:')
    for no, record in enumerate(df[feature]):
        try:
            if feature in ['audio_features.danceability',  'audio_features.energy', 'audio_features.acousticness', 'audio_features.valence']:
                if record < 0 or record > 1:
                    print(f'index: {no} valor: {record} rango erroneo')
            elif feature == 'audio_features.key':
                if record < -1 or record > 11:
                    print(f'index: {no} valor: {record} rango erroneo')
            elif feature == 'audio_features.time_signature':
                if record < 3 or record > 7:
                    print(f'index: {no} valor: {record} rango erroneo')
            else:
                if record < 0 or record > 100:
                    print(f'index: {no} valor: {record} rango erroneo')
        except:
            pass
```

✓ 0.0s

```
columna track_popularity:
index: 75 valor: -69 rango erroneo
index: 89 valor: -70 rango erroneo
index: 109 valor: -85 rango erroneo
index: 111 valor: -92 rango erroneo
index: 115 valor: -75 rango erroneo
index: 128 valor: -71 rango erroneo
index: 472 valor: 152 rango erroneo

columna audio_features.acousticness:
index: 1 valor: 5.0 rango erroneo
index: 3 valor: -0.000537 rango erroneo
index: 6 valor: -0.00354 rango erroneo
index: 527 valor: 1.5 rango erroneo
index: 535 valor: 2.0 rango erroneo

columna artist_popularity:
index: 0 valor: 120 rango erroneo
index: 1 valor: 120 rango erroneo
```

Figura 11. Registros por fuera del rango definido para las diferentes características.

A partir de los errores anteriormente expuestos, la tabla con anomalías para la dimensión validez es de la manera siguiente (Tabla 4):

Tabla 4. Anomalías encontradas para la dimensión validez.

Característica	No. de anomalías
track_popularity	7
explicit	539
track_id	8
track_name	7
**danceability*	2
**energy*	2
**key*	1
**loudness*	2
**speechiness*	1
**acousticness*	6
**instrumentalness*	539
**liveness*	1
**tempo*	1
**time_signature*	539
artist_popularity	539
album_name	46
album_total_tracks	539

Las características marcadas con un (*) presentan los siguientes caracteres “audio_features.” previo al nombre del campo.

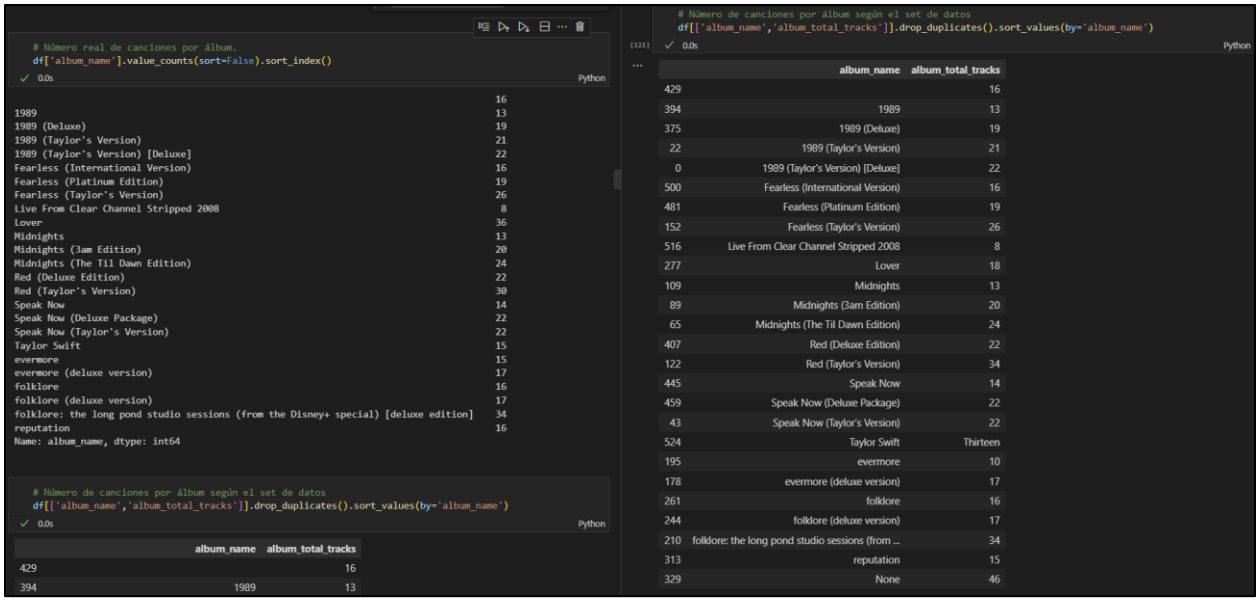
3.4. Precisión

Para la dimensión de Precisión de los datos se evidenciaron los siguientes errores (Tabla 5):

Tabla 5. Anomalías encontradas para la dimensión precisión.

Característica	No. de anomalías
Track_number	1
album_release_date	39
album_total_tracks	112

La cantidad real de canciones por álbum no coincide con el valor de la columna designada para dicho valor. Álbumes como Red (taylor’s versión), evermore , Lover, reputation, y Taylor Swift presentan una cantidad diferente de canciones con relación al valor reportado en el set de datos (Figura 12).



```
# Número real de canciones por álbum.
df['album_name'].value_counts(sort=False).sort_index()
```

```
1989                                                               16
1989 (Deluxe)                                                      13
1989 (Taylor's Version)                                            19
1989 (Taylor's Version) [Deluxe]                                   21
Fearless (International Version)                                   22
Fearless (Platinum Edition)                                        16
Fearless (Taylor's Version)                                        19
Live From Clear Channel Stripped 2008                              26
Lover                                                               8
Midnights                                                          36
Midnights (3am Edition)                                            13
Midnights (The Til Dawn Edition)                                   20
Red (Deluxe Edition)                                               24
Red (Taylor's Version)                                             22
Speak Now                                                          30
Speak Now (Deluxe Package)                                         14
Speak Now (Taylor's Version)                                       22
Taylor Swift                                                       22
evermore                                                           15
evermore (deluxe version)                                          15
folklore                                                           17
folklore (deluxe version)                                          16
folklore: the long pond studio sessions (from the Disney+ special) [deluxe edition]   17
reputation                                                         34
Name: album_name, dtype: int64
```

```
# Número de canciones por álbum según el set de datos
df[['album_name','album_total_tracks']].drop_duplicates().sort_values(by='album_name')
```

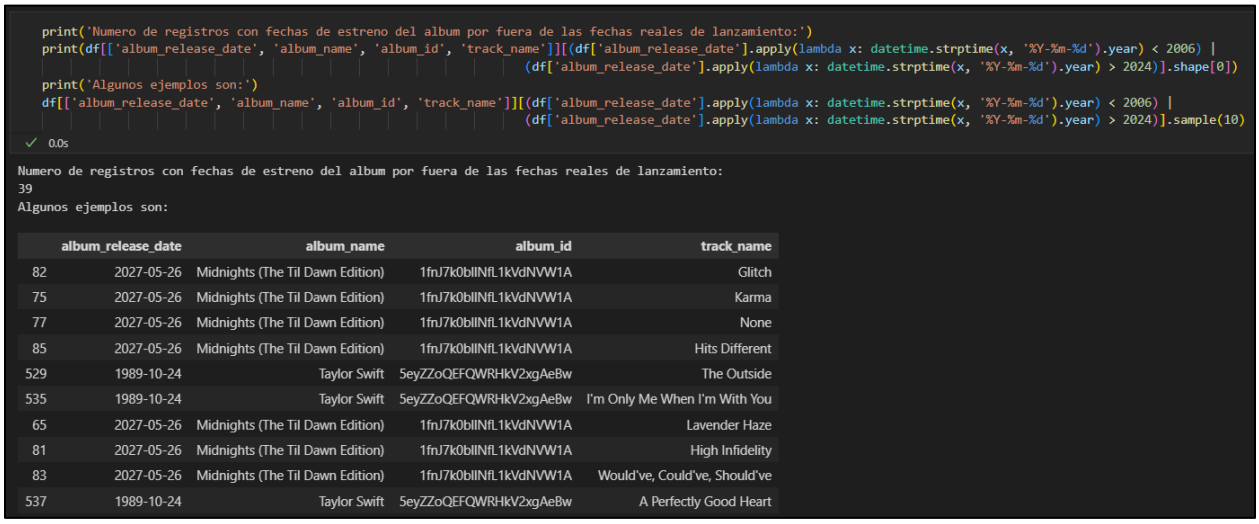
	album_name	album_total_tracks
429		16
394	1989	13

```
# Número de canciones por álbum según el set de datos
df[['album_name','album_total_tracks']].drop_duplicates().sort_values(by='album_name')
```

	album_name	album_total_tracks
429		16
394	1989	13
375	1989 (Deluxe)	19
22	1989 (Taylor's Version)	21
0	1989 (Taylor's Version) [Deluxe]	22
500	Fearless (International Version)	16
481	Fearless (Platinum Edition)	19
152	Fearless (Taylor's Version)	26
516	Live From Clear Channel Stripped 2008	8
277	Lover	18
109	Midnights	13
89	Midnights (3am Edition)	20
65	Midnights (The Til Dawn Edition)	24
407	Red (Deluxe Edition)	22
122	Red (Taylor's Version)	34
445	Speak Now	14
459	Speak Now (Deluxe Package)	22
43	Speak Now (Taylor's Version)	22
524	Taylor Swift	Thirteen
195	evermore	10
178	evermore (deluxe version)	17
261	folklore	16
244	folklore (deluxe version)	17
210	folklore: the long pond studio sessions (from ...	34
313	reputation	15
329	None	46

Figura 12**.** A) Cantidad de canciones reales por álbum. B) Cantidad de canciones por álbum reportadas en los datos.

Adicionalmente, de acuerdo con la información de la API de Spotify, los álbumes de Taylor Swift fueron estrenados desde el 2006 hasta el presente, por lo cual, la fecha reportada para el álbum “Midnights (The Til Dawn Edition)” es incorrecta ya que es superior a la actualidad. Así mismo, la fecha de estreno del álbum Taylor Swift , 1984 según el registro, es incorrecta con relación a la fecha de lanzamiento real en 2006 (Figura 13).



```
print('Numero de registros con fechas de estreno del album por fuera de las fechas reales de lanzamiento:')
print(df[['album_release_date', 'album_name', 'album_id', 'track_name']][(df['album_release_date'].apply(lambda x: datetime.strptime(x, '%Y-%m-%d').year) < 2006) |
                                                                          (df['album_release_date'].apply(lambda x: datetime.strptime(x, '%Y-%m-%d').year) > 2024)].shape[0])
print('Algunos ejemplos son:')
df[['album_release_date', 'album_name', 'album_id', 'track_name']][(df['album_release_date'].apply(lambda x: datetime.strptime(x, '%Y-%m-%d').year) < 2006) |
                                                                   (df['album_release_date'].apply(lambda x: datetime.strptime(x, '%Y-%m-%d').year) > 2024)].sample(10)
```

```
Numero de registros con fechas de estreno del album por fuera de las fechas reales de lanzamiento:
39
Algunos ejemplos son:
```

	album_release_date	album_name	album_id	track_name
82	2027-05-26	Midnights (The Til Dawn Edition)	1fnJ7k0bIlNfL1kVdNVW1A	Glitch
75	2027-05-26	Midnights (The Til Dawn Edition)	1fnJ7k0bIlNfL1kVdNVW1A	Karma
77	2027-05-26	Midnights (The Til Dawn Edition)	1fnJ7k0bIlNfL1kVdNVW1A	None
85	2027-05-26	Midnights (The Til Dawn Edition)	1fnJ7k0bIlNfL1kVdNVW1A	Hits Different
529	1989-10-24	Taylor Swift	5eyZZoQEFQWRHkV2xgAeBw	The Outside
535	1989-10-24	Taylor Swift	5eyZZoQEFQWRHkV2xgAeBw	I'm Only Me When I'm With You
65	2027-05-26	Midnights (The Til Dawn Edition)	1fnJ7k0bIlNfL1kVdNVW1A	Lavender Haze
81	2027-05-26	Midnights (The Til Dawn Edition)	1fnJ7k0bIlNfL1kVdNVW1A	High Infidelity
83	2027-05-26	Midnights (The Til Dawn Edition)	1fnJ7k0bIlNfL1kVdNVW1A	Would've, Could've, Should've
537	1989-10-24	Taylor Swift	5eyZZoQEFQWRHkV2xgAeBw	A Perfectly Good Heart

Figura 13**.** Registros de álbumes con fechas de lanzamiento erróneas.

Relacionado con el número correcto de canciones por álbum (*album_total_tracks*), un registro de canción indice 88 no es adecuado con la numeración (*track_number*) en el álbum “Midnights (The Til Dawn Edition)”, siendo un número no consecutivo (Figura 14).

```
df[['album_name', 'album_total_tracks', 'track_number']][df['album_name'] == 'Midnights (The Til Dawn Edition)'].sort_index(ascending=False).head()
```
✓ 0.0s

	album_name	album_total_tracks	track_number
88	Midnights (The Til Dawn Edition)	24	21
87	Midnights (The Til Dawn Edition)	24	23
86	Midnights (The Til Dawn Edition)	24	22
85	Midnights (The Til Dawn Edition)	24	21
84	Midnights (The Til Dawn Edition)	24	20

Figura 14**.** Registro con numero de canción erronea.

3.5. Consistencia

Para la dimensión de consistencia se utilizaron los registros que no presentan variación entre ellos en cuanto al formato y contenido, es decir, los registros que no presentan errores en cuanto a validez, unicidad y precisión (Tabla 6):

Tabla 6. Anomalías encontradas para la dimensión consistencia.

Característica	No. de anomalías	Característica	No. de anomalías
disc_number	19	**instrumentalness*	20
duration_ms	19	**liveness*	20
explicit	23	**valence*	19
track_number	19	**tempo*	20
track_popularity	26	**id*	19
track_id	26	**time_signature*	20
track_name	26	*artist_id*	19
**danceability*	21	*artist_name*	19
**energy*	21	*artist_popularity*	19
**key*	20	*album_id*	19
**loudness*	21	*album_name*	65
**mode*	20	*album_release_date*	57
**speechiness*	20	*álbum_total_tracks*	114
**acousticness*	24	-	-

Las características marcadas con un (*) presentan los siguientes caracteres “audio_features.” previo al nombre del campo.

6. Evaluación:

Es importante mencionar que el porcentaje de datos erróneos es calculado sobre el total de los registros (539) y no sobre el set de datos corregido al eliminar las filas repetidas (520) para evitar alteraciones en la información original suministrada.

Tabla 7. Evaluación de la calidad de los datos en porcentajes.

	Completitud	Unicidad	Validez	Precisión	Consistencia	Total	Calidad características
disc_number	100,0	96,5	100,0	100,0	96,5	98,6	Buena
duration_ms	100,0	96,5	100,0	100,0	96,5	98,6	Buena
explicit	100,0	96,5	0,0	100,0	95,7	78,4	Mala
track_number	100,0	96,5	100,0	100,0	96,5	98,6	Buena
track_popularity	100,0	96,5	100,0	100,0	95,2	98,3	Buena
track_id	98,5	96,5	98,5	100,0	95,2	97,7	Buena
track_name	98,7	96,5	98,7	100,0	95,2	97,8	Buena
**danceability*	99,6	96,5	99,6	100,0	96,1	98,4	Buena
**energy*	99,6	96,5	99,6	100,0	96,1	98,4	Buena
**key*	99,8	96,5	99,8	100,0	96,3	98,5	Buena
**loudness*	99,6	96,5	99,6	100,0	96,1	98,4	Buena
**mode*	100,0	96,5	100,0	100,0	96,3	98,6	Buena
**speechiness*	99,8	96,5	99,8	100,0	96,3	98,5	Buena
**acousticness*	99,8	96,5	98,9	100,0	95,5	98,1	Buena
**instrumentalness*	100,0	96,5	0,0	100,0	96,3	78,6	Mala
**liveness*	99,8	96,5	99,8	100,0	96,3	98,5	Buena
**valence*	100,0	96,5	100,0	100,0	96,5	98,6	Buena
**tempo*	99,8	96,5	99,8	100,0	96,3	98,5	Buena
**id*	98,5	0,0	98,5	100,0	96,5	78,7	Mala
**time_signature*	99,8	96,5	100,0	100,0	96,3	98,5	Buena
artist_id	100,0	96,5	100,0	100,0	96,5	98,6	Buena
artist_name	100,0	96,5	100,0	100,0	96,5	98,6	Buena
artist_popularity	100,0	96,5	0,0	100,0	96,5	78,6	Mala
album_id	100,0	96,5	100,0	100,0	96,5	98,6	Buena
album_name	88,5	96,5	91,5	100,0	87,9	92,9	Aceptable
album_release_date	100,0	96,5	100,0	92,8	89,4	95,7	Aceptable
album_total_tracks	100,0	96,5	0,0	79,2	78,8	70,9	Mala
Total	99,3	92,9	84,6	99,0	94,9	**94,1**	**Aceptable**
Calidad dimensión	Buena	Aceptable	Mala	Buena	Aceptable		

Las características marcadas con un (*) presentan los siguientes caracteres “audio_features.” Previo al nombre del campo.

7. Análisis de calidad de los datos:

En términos generales la calidad de los datos es aceptable, con un total de 94.1 sobre 100. Las diferentes dimensiones analizadas son catalogadas como buenas, aceptables y malas (40 %, 40 % y 20% del total, respectivamente), siendo validez la dimensión con la calificación más baja (84.6 %).

Desde el contexto de las características, los datos de 20 de estas son caracterizadas con calidad de datos buenos, correspondientes a un 74.1 % de la información, 2 son definidas como aceptables (7.4 %) y las características *explicit*, *audio_features.instrumentalness*, audio_features.*id*, *artist_popularity* y *album_total_tracks* son clasificadas con datos de tipo malo (18.5 %).

8. Referencias:

DAMA UK. 2018. The six primary dimensions for data quality assessment. defining data quality dimensions.