



RETO: SPOTIFY API

Data Quality Engineer - Prueba Técnica

REPORTE CALIDAD DE DATOS

Dirigido a :
Empresa R5

Encargado:
Juan José López Condori

9 de enero de 2024

Índice

| | |
|---|----------|
| 1. Introducción | 2 |
| 2. Metodología | 2 |
| 2.1. Herramientas utilizadas | 2 |
| 3. Dataset | 3 |
| 4. Análisis de calidad de datos | 4 |
| 4.1. Completeness (Compleitud). | 4 |
| 4.2. Uniqueness (Unicidad). | 7 |
| 4.3. Validity (Validez). | 8 |
| 5. Conclusiones | 8 |

Reporte de calidad de datos - Spotify API

1. Introducción

Se ha demostrado que los datos son la clave para entregar valor a los clientes y estos son una parte integral de tu estrategia comercial. Además, son un activo importante de todas las empresas.

El presente reporte de calidad de datos - Spotify API forma parte del proceso de selección para el cargo de Data Quality Engineer Junior en la empresa R5.

Esta propuesta busca explicar a detalle el proceso de indentificación de todas las anomalías de calidad de datos del dataset obtenido de la API de Spotify (dataset entregado como parte de la prueba), cabe resaltar que el reporte se basa en las seis dimensiones principales para la evaluación de la calidad de los datos [1]

2. Metodología

Para el proceso de evaluación de calidad de datos y de la identificación de anomalías se utilizó las seis dimensiones principales para la evaluación de la calidad de los datos:

- Completeness (Compleitud).
- Uniqueness (Unicidad).
- Timeliness (Actualidad).
- Validity (Validez).
- Accuracy (Exactitud).
- Consistency (Consistencia).

Considerando el dataset proporcionado se empleó solo 03 dimensiones: Completeness (Compleitud), Uniqueness (Unicidad) y Validity (Validez).

2.1. Herramientas utilizadas

Las herramientas utilizadas para la generación del reporte de calidad de datos - Spotify API, se utilizaron las siguientes herramientas/bibliotecas: Menciona las herramientas, librerías o lenguajes de programación que has utilizado (por ejemplo, Python y pandas).

- Lenguaje de programación Python
- Librería Pandas
- Librería Display

3. Dataset

El dataset proporcionado `taylor_swift_spotify.json` tenía una estructura que se detalla a continuación:

```

1 {
2   "artist_id": "06HL4z0CvFAxyc27GX",
3   "artist_name": "Taylor Swift",
4   "artist_popularity": 120,
5   "albums": [
6     {
7       "album_id": "1o59UpKw81iHROHPiSkJR0",
8       "album_name": "1989 (Taylor's Version) [Deluxe]",
9       "album_release_date": "2023-10-27",
10      "album_total_tracks": 22,
11      "tracks": [
12        {
13          "disc_number": 1,
14          "duration_ms": 212600,
15          "explicit": false,
16          "track_number": 1,
17          "audio_features": {
18            "danceability": 0.757,
19            "energy": 0.61,
20            "key": 7,
21            "loudness": -4.84,
22            "mode": 1,
23            "speechiness": 0.0327,
24            "acousticness": 0.00942,
25            "instrumentalness": 3.66e-05,
26            "liveness": 0.367,
27            "valence": 0.685,
28            "tempo": 116.998,
29            "id": "4WUepByoeqcedHoYhSNHRt",
30            "time_signature": 4
31          },
32          "track_popularity": 77,
33          "track_id": "4WUepByoeqcedHoYhSNHRt",
34          "track_name": "Welcome To New York (Taylor's Version)"
35        }, ..., ...
36      ]
37    }
38  }
39
40
```

Se indentificó que el campo “*albums*” era un campo anidado (una lista de albums) asi mismo cada uno de los albums tenía el campo “*tracks*” otro campo anidado (una lista de tracks), por lo tanto se procedió a desanidar ambos campos, exportando los siguientes archivos csv:

- dataset_albumes.csv : Con los albums desanidados.
- dataset: Con ambos campos desanidados(albums y tracks).

4. Análisis de calidad de datos

Considerando el dataset proporcionado se empleó solo 03 dimensiones: Completeness (Compleitud), Uniqueness (Unicidad) y Validity (Validez), que se detallan a continuación:

4.1. Completeness (Compleitud).

En el análisis del dataset, se identificaron registros con datos nulos en varias columnas, considerando los datos anidados:

- Album: Mediante la función `.isnull()` de pandas en un dataframe, analizamos los valores nulos de manera general como se muestra en la Figura 1

```
artist_id      0
artist_name    0
artist_popularity  0
album_id       0
album_name     2
album_release_date  0
album_total_tracks  0
tracks         0
```

Figura 1: Informacion general de datos nulos

En el campo *album_id*, se encontraron dos registros con valores nulos, tanto en la fila 16 y 20 respectivamente, como se muestra en la Figura 2

| | Index | artist_id | artist_name | album_id | album_name |
|---|-------|--------------------|--------------|------------------------|------------|
| 0 | 16 | 06HL4z0CvFAxyc27GX | Taylor Swift | 1MPAXuTVL2Ej5x0JHiSPq8 | NaN |
| 1 | 20 | 06HL4z0CvFAxyc27GX | Taylor Swift | 6fyR4wBPwLHKcRtxgd4sGh | NaN |

Figura 2: Datos nulos - campo album_name

- Tracks: Con el mismo procedimiento utilizado anteriormente, mostramos en la Figura 3, la información general de datos nulos desanidando la información del campo *tracks* :

| | |
|---------------------------------|---|
| artist_id | 0 |
| artist_name | 0 |
| artist_popularity | 0 |
| album_id | 0 |
| album_release_date | 0 |
| album_total_tracks | 0 |
| disc_number | 0 |
| duration_ms | 0 |
| explicit | 0 |
| track_number | 0 |
| track_popularity | 0 |
| track_id | 8 |
| track_name | 7 |
| audio_features_danceability | 2 |
| audio_features_energy | 2 |
| audio_features_key | 1 |
| audio_features_loudness | 2 |
| audio_features_mode | 0 |
| audio_features_speechiness | 1 |
| audio_features_acousticness | 1 |
| audio_features_instrumentalness | 0 |
| audio_features_liveness | 1 |
| audio_features_valence | 0 |
| audio_features_tempo | 1 |
| audio_features_id | 0 |
| audio_features_time_signature | 1 |

Figura 3: Información general de datos nulos

En la Figura 4 se muestran las filas con valores nulos de los campos: *album_id*, *track_id*, *track_name*, *audio_features_danceability*, *audio_features_energy*, *audio_features_key*, *audio_features_loudness*, *audio_features_speechiness*, *audio_features_acousticness*, *audio_features_liveness*, *audio_features_tempo*, *audio_features_time_signature*, se añadió el campo *Index* que indica la posición de la fila en el dataframe original.

| | Index | track_id | track_name | audio_features_danceability | audio_features_energy | audio_features_key | audio_features_loudness | audio_features_speechiness | audio_features_acousticness | audio_features_liveness | audio_features_tempo | audio_features_time_signature |
|----|-------|-------------------------|---|-----------------------------|-----------------------|--------------------|-------------------------|----------------------------|-----------------------------|-------------------------|----------------------|-------------------------------|
| 0 | 77 | 1QQii3pa5m8MEda0nbkfw | NaN | 0.661 | 0.345 | 4.0 | -14.037 | 0.1440 | 0.54500 | 0.0871 | 125.977 | 4.0 |
| 1 | 91 | 02Zkkf2zMkwRGQjZ7T4p8f | NaN | 0.638 | 0.634 | 4.0 | -6.582 | 0.0457 | 0.13300 | 0.1520 | 96.953 | 4.0 |
| 2 | 104 | 7712gjoih4QoDbXpljEk21 | NaN | 0.700 | 0.509 | 7.0 | -10.547 | 0.0789 | 0.11200 | 0.1370 | 110.947 | 4.0 |
| 3 | 321 | NaN | Gorgeous | 0.800 | 0.535 | 7.0 | -6.684 | 0.1350 | 0.07130 | 0.2130 | 92.027 | 4.0 |
| 4 | 330 | 22C0JIVhFaczZ4t9heqREN | Wildest Dreams | NaN | NaN | 8.0 | NaN | NaN | 0.06920 | 0.1060 | 139.997 | 4.0 |
| 5 | 334 | 7eGeUVeCEvEQrivmjI9Qn3 | Teardrops On My Guitar - Radio Single Remix | 0.626 | 0.427 | NaN | NaN | 0.0234 | 0.31300 | 0.1380 | 99.959 | 4.0 |
| 6 | 341 | 41T04yafZVijNq2FqvLtlId | So It Goes... | 0.574 | 0.610 | 2.0 | -7.283 | 0.0732 | 0.12200 | NaN | 74.957 | 4.0 |
| 7 | 363 | NaN | Jump Then Fall | 0.617 | NaN | 2.0 | -5.712 | 0.0274 | 0.11300 | 0.0740 | 80.007 | NaN |
| 8 | 375 | NaN | Welcome To New York | 0.789 | 0.634 | 7.0 | -4.762 | 0.0323 | 0.03480 | 0.3020 | 116.992 | 4.0 |
| 9 | 379 | NaN | All You Had To Do Was Stay | 0.605 | 0.725 | 5.0 | -5.729 | 0.0323 | 0.00201 | 0.1010 | 96.970 | 4.0 |
| 10 | 382 | NaN | Bad Blood | 0.646 | 0.794 | 7.0 | -6.104 | 0.1900 | 0.08850 | 0.2010 | 170.216 | 4.0 |
| 11 | 391 | 4FoV729rw7lhoKlMZW5K5V | NaN | 0.592 | 0.128 | NaN | -17.932 | 0.5890 | 0.82900 | 0.5270 | 78.828 | 4.0 |
| 12 | 396 | 71PmZqBXH0RUETqxpwlV0w | NaN | 0.598 | 0.786 | 2.0 | -5.572 | 0.0382 | 0.00256 | 0.1170 | 95.021 | 4.0 |
| 13 | 401 | 0TvQLMecTE8utzoNmvXRbK | NaN | 0.652 | 0.802 | 7.0 | -6.114 | 0.1810 | 0.08710 | 0.1480 | 170.157 | 4.0 |
| 14 | 408 | 7gJtmLyPTwKzhGzMBXtuXH | NaN | 0.602 | 0.896 | 1.0 | -4.267 | 0.0437 | 0.07730 | 0.0911 | 124.978 | 4.0 |
| 15 | 431 | 72GIZuUXoI4oyrS0si3Rgc | The Story Of Us - Live | NaN | 0.908 | 9.0 | -5.156 | 0.0651 | NaN | 0.8150 | 139.813 | 4.0 |
| 16 | 432 | 7mFiEij8AXPUZB7aKLbUIQ | Mean - Live/2011 | 0.429 | 0.915 | 4.0 | -4.373 | 0.0690 | 0.15400 | 0.6930 | NaN | 4.0 |
| 17 | 434 | NaN | Back To December/Apologize/You're Not Sorry - ... | 0.374 | 0.516 | 2.0 | -8.745 | 0.0294 | 0.15500 | 0.7950 | 142.057 | 4.0 |
| 18 | 442 | NaN | Enchanted - Live/2011 | 0.340 | 0.663 | 8.0 | -5.597 | 0.0331 | 0.07470 | 0.5630 | 163.678 | 4.0 |
| 19 | 445 | NaN | Mine - POP Mix | 0.696 | 0.768 | 7.0 | -3.863 | 0.0308 | 0.00461 | 0.1010 | 121.050 | 4.0 |

Figura 4: Datos nulos - campo track desanidado

En la Figura 5 se muestra la integridad de dataset según las columnas:

| | |
|---------------------------------|--------|
| artist_id | 100.00 |
| artist_name | 100.00 |
| artist_popularity | 100.00 |
| album_id | 100.00 |
| album_name | 92.59 |
| album_release_date | 100.00 |
| album_total_tracks | 100.00 |
| disc_number | 100.00 |
| duration_ms | 100.00 |
| explicit | 100.00 |
| track_number | 100.00 |
| track_popularity | 100.00 |
| track_id | 98.52 |
| track_name | 98.70 |
| audio_features_danceability | 99.63 |
| audio_features_energy | 99.63 |
| audio_features_key | 99.81 |
| audio_features_loudness | 99.63 |
| audio_features_mode | 100.00 |
| audio_features_speechiness | 99.81 |
| audio_features_acousticness | 99.81 |
| audio_features_instrumentalness | 100.00 |
| audio_features_liveness | 99.81 |
| audio_features_valence | 100.00 |
| audio_features_tempo | 99.81 |
| audio_features_id | 100.00 |
| audio_features_time_signature | 99.81 |

Figura 5: Integridad del dataset completo(%)

La presencia de datos nulos en estas columnas puede deberse a diversas razones, como errores durante la recolección de datos, registros incompletos en la fuente de origen, o incluso cambios en la estructura de la base de datos que no se han reflejado adecuadamente en la extracción de datos.

La identificación de estos datos faltantes es crucial para comprender la integridad del dataset y decidir cómo manejarlos, ya sea eliminándolos, imputando valores o aplicando estrategias específicas según el contexto del análisis.

4.2. Uniqueness (Unicidad).

En el análisis de los datos se encontró una fila duplicada, esto se identificó mediante el campo “*album.id*”, considerando que debe ser un valor único, las filas duplicadas se muestran en la Figura 6.

| | artist_id | artist_name | artist_popularity | album_id | album_name | album_release_date | album_total_tracks |
|----|--------------------|--------------|-------------------|------------------------|------------|--------------------|--------------------|
| 13 | 06HL4z0CvFAxyc27GX | Taylor Swift | 120 | 1NAmidJIEaVgA3MpcPFYGq | Lover | 2019-08-23 | 18 |
| 14 | 06HL4z0CvFAxyc27GX | Taylor Swift | 120 | 1NAmidJIEaVgA3MpcPFYGq | Lover | 2019-08-23 | 18 |

Figura 6: Datos duplicados

La Unicidad de los datos es de **96.43 %**

4.3. Validity (Validez).

En el proceso se encontró una fila con un tipo de datos que no correspondía al atributo correspondiente, en el campo “*album_total_tracks*”, donde los datos eran tipo entero se identificó un dato de tipo string, como se observa en la Figura 7

| | artist_id | artist_name | artist_popularity | album_id | album_name | album_release_date | album_total_tracks |
|----|--------------------|--------------|-------------------|------------------------|--------------|--------------------|--------------------|
| 26 | 06HL4z0CvFAxyc27GX | Taylor Swift | 120 | 5eyZZoQEFQWRHkV2xgAeBw | Taylor Swift | 1989-10-24 | Thirteen |

Figura 7: Datos con tipo de dato incorrecto

5. Conclusiones

- Se identificaron 2 filas duplicadas en el conjunto de datos, lo que indica posibles errores en la entrada de datos o inconsistencias en el proceso de almacenamiento. Se recomienda revisar y mejorar los controles de calidad en la entrada de datos para prevenir duplicados en el futuro.
- Se encontraron valores nulos en varias columnas. La presencia de valores nulos puede afectar la integridad y confiabilidad de los resultados del análisis. Se sugiere realizar una imputación de datos y establecer validaciones más estrictas en la entrada de datos para minimizar la ocurrencia de valores nulos.

Referencias

- [1] The Six Primary Dimensions For Data Quality Assessment, 2013