REPORTE DE CALIDAD DE DATOS

PRUEBA TÉCNICA GRUPO R5 - DATA QUALITY ENGINEER JUNIOR

SPOTIFY API

Anomalías identificadas en el archivo de datos *taylor_swift_spotify. json.*

Para validar las inconsistencias en la data, visualizaremos el contenido dentro de cada una de las variables.

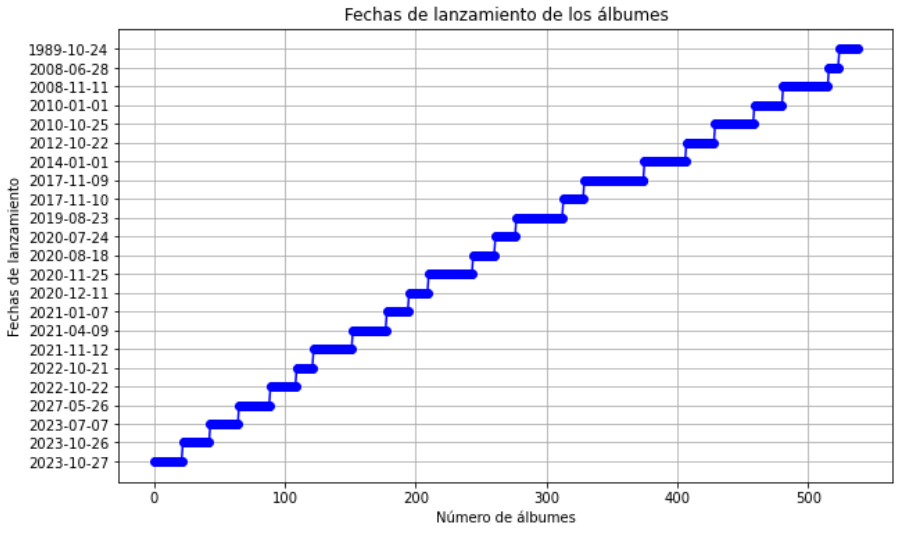
1) Iniciamos mostrando los nombres de los álbumes:

```
Nombres de los álbumes:
1989 (Taylor's Version) [Deluxe]
1989 (Taylor's Version)
Speak Now (Taylor's Version)
Midnights (The Til Dawn Edition)
Midnights (3am Edition)
Midnights
Red (Taylor's Version)
Fearless (Taylor's Version)
evermore (deluxe version)
evermore
folklore: the long pond studio sessions (from the Disney+ special) [deluxe edition]
```

Se identifican inconsistencias sobre los nombres de los álbumes, en términos de la escritura con información faltante en alguno de ellos o, que es igual a los demás.

2) Validaciones sobre la fecha de lanzamiento de los álbumes

Se genera gráfica para validaciones (el 2027 no se ordena, al parecer viene en diferente formato)



Se observan fechas de lanzamiento mayores a la fecha de hoy, lo cual no sería posible.

3) Validamos sobre los valores mínimos y máximos de la duración de los álbumes en unidades de milisegundos.

```
Duración mínima: -223093 ms
Duración máxima: 613026 ms
```

Al ser los valores de duración de la pista en ms, se presenta una inconsistencia al observar valores negativos.

4) Ahora vamos a la popularidad que tiene la pista en Spotify, contenida en la variable track_popularity

```
Popularidad mínima: -92
Popularidad máxima: 152
```

Al ser el número que ocupa en posición de popularidad, debe almacenar solo valores enteros positivos. Es decir, los valores almacenados están por fuera de los rangos permitidos.

5) Ahora, al igual que debe hacerse con la base completa, se identificarán los valores faltantes en la variable track_id

```
                album_id     album_name  ... track_id                track_name
363  1MPAXuTVL2Ej5x0JHiSPq8           None  ...     None             Jump Then Fall
375  1yGbNOtRIgdIiGHOEBaZWf  1989 (Deluxe)  ...     None         Welcome To New York
379  1yGbNOtRIgdIiGHOEBaZWf  1989 (Deluxe)  ...     None  All You Had To Do Was Stay
382  1yGbNOtRIgdIiGHOEBaZWf  1989 (Deluxe)  ...     None                  Bad Blood
```

Se deben depurar los valores faltantes para esta variable, dado que pueden generar sesgo al momento de realizar análisis.

6) Por último, validaremos la variable final, correspondiente a track_name

Se debe validar sobre los nombres para que sea consecuente la escritura de unos con otros y poder agruparlos.

Conclusiones generales:

A la base se le deben depurar los NA, posteriormente validar sobre cada variable, para así evitar problemas entre ellas, definiendo los rangos de los cuales cada una pueda moverse.

Para este proceso, deben incluirse conclusiones ligadas al contexto real, como, por ejemplo: Taylor Swift, nació en 1989, por ende, sería imposible que existan álbumes que hayan sido lanzados en esa fecha.

Entregable, prueba número 2.

Realizada por: Dallys Nicol Sinisterra Gutierrez