Informe Calidad de Datos

Grupo R5 - Direción de Analítica

Reporte de anomalías

1. Fuente de datos
2. Consideraciones y ejemplos de anomalías
3. Estadísticas (Métricas)

1. Fuente de datos

Fuente: API Spotify
Datos: Discografía Taylor Swift

La información analizada en este informe ha sido recopilada utilizando la interfaz de programación de aplicaciones (API) de Spotify, una plataforma líder en streaming de música. Específicamente, los datos examinados se centran en la artista Taylor Swift, cuya presencia en la industria musical ha sido un fenómeno notable.
La API de Spotify proporcionó acceso a una variedad de atributos detallados relacionados con las canciones, álbumes y artistas disponibles en su extenso catálogo. En este análisis, nos sumergimos en los datos asociados con Taylor Swift, explorando aspectos que van desde las características musicales de sus canciones hasta la popularidad de sus álbumes.

Definición de dato:
En el contexto de este informe, un "dato" se refiere a la información específica obtenida mediante el cruce entre una fila y una columna en el conjunto de datos. Cada celda en esta matriz de datos representa un valor singular que proporciona detalles sobre aspectos particulares de la discografía de Taylor Swift. Estos valores abarcan desde atributos detallados de canciones, álbumes y artistas hasta métricas relacionadas con la popularidad, características musicales y otra información relevante recopilada de la API de Spotify.

2. Consideraciones de Anomalías

2.01 Id de canción nulo:

Se encontraron 8 canciones con el id nulo de 575 datos.

Ejemplo:

357 NaN

2.02 Filas duplicadas:

Se encontraron 54 filas duplicadas de 575.

Ejemplo:

5 Indices de filas duplicadas:
[88, 278, 280, 282, 284]

2.03 Valores nulos:

Se encontraron 89 valores nulos de 15525 datos.

Ejemplo:

null_value_column = track_name
null_index_value = 77 NaN

2.04 Formato incorrecto en nombres de canciones según la convencion de nombramiento en inglés:

Se encontraron 89 nombres de canciones con formato incorrecto de 575 datos.

Ejemplo:

incorrect_track_names = ['willow', 'champagne problems', 'gold rush', ‘'tis the damn season']

NOTA:
Puede no ser necesariamente una anomalía. Es importante destacar que, en la industria musical, la creatividad y la expresión artística a menudo influyen en la elección de nombres de canciones, lo que puede llevar a variaciones en el formato. Este hallazgo se menciona con la precaución de que la divergencia del formato convencional puede ser intencional y parte del estilo artístico.

2.05 Caracteres mal codificados en nombres de canciones:

Se encontraron 55 nombres de canciones con caracteres mal codificados de 575 datos.

Ejemplo:

track_name_anomalies = [‘'tis the damn season', 'it’s time to go - bonus track', 'Soon You’ll Get Better (feat. The Chicks)', 'It’s Nice To Have A Friend']

2.06 Datos no booleanos en la columna ‘explicit’:

Se encontraron 6 datos no booleanos en la columna ‘explicit’ de 575 datos.

Ejemplo:

explicit_anomalies = ['Si', 'No']

2.07 Datos no numéricos en la columna ‘album_total_tracks’:

Se encontraron 15 datos no numéricos en la columna ‘album_total_tracks’ de 575 datos.

Ejemplo:

album_total_tracks_anomalies = ['Thirteen']

2.08 Formato no numérico en la columna ‘audio_features.instrumentalness’:

La columna ‘audio_features.instrumentalness’ tiene un formato no numérico.

Ejemplo:

instrumentalness_type_is_numeric = False
instrumentalness_type_anomalies = ['3.66e-05', '0', '0.0197', '5.59e-05', '0']

2.09 Datos no convertibles a numéricos en la columna ‘audio_features.instrumentalness’:

Se encontraron 1 datos no convertibles a numéricos en la columna ‘audio_features.instrumentalness’ de 575 datos.

Ejemplo:

instrumentalness_type_conver_anomalies = ['7.28x-06']

2.10 Valores fuera del rango [0,1] en la columna 'audio_features.danceability':

Se encontraron 2 valores fuera del rango [0,1] en la columna 'audio_features.danceability' de 575 datos.

Ejemplo:

danceability_anomalies = [nan]

2.11 Valores fuera del rango [0,1] en la columna 'audio_features.energy’:

Se encontraron 2 valores fuera del rango [0,1] en la columna 'audio_features.energy’ de 575 datos.

Ejemplo:

energy_anomalies = [nan]

2.12 Valores fuera del rango [0,1] en la columna 'audio_features.liveness’:

Se encontraron 1 valores fuera del rango [0,1] en la columna 'audio_features.liveness’ de 575 datos.

Ejemplo:

liveness_anomalies = [nan]

2.13 Valores fuera del rango [3,7] en la columna 'audio_features.time_signature’:

Se encontraron 1 valores fuera del rango [3,7] en la columna 'audio_features.time_signature’ de 575 datos.

Ejemplo:

time_signature_anomalies = [nan]

2.14 Valores fuera del rango [-1,11] en la columna 'audio_features.key’:

Se encontraron 1 valores fuera del rango [-1,11] en la columna 'audio_features.key’ de 575 datos.

Ejemplo:

key_anomalies = [nan]

2.15 Valores fuera del rango [-60,0] en la columna 'audio_features.loudness’:

Se encontraron 2 valores fuera del rango [-60,0] en la columna 'audio_features.loudness’ de 575 datos.

Ejemplo:

loudness_anomalies = [nan]

2.16 Valores fuera del rango [0,100] en la columna 'track_popularity’:

Se encontraron 7 valores fuera del rango [0,100] en la columna 'track_popularity’ de 575 datos.

Ejemplo:

track_popularity_anomalies = [-69, -70, -85, -92, -75, -71, 152]

2.17 Valores fuera del rango [0,100] en la columna 'artist_popularity’:

Se encontraron 575 valores fuera del rango [0,100] en la columna 'artist_popularity’ de 575 datos.

Ejemplo:

artist_popularity_anomalies = [120]

2.18 Valores fuera del rango [82000, 630000] en la columna ‘duration_ms’:

Se encontraron 5 valores fuera del rango [82000, 630000] en la columna ‘duration_ms’ de 575 datos.

Ejemplo:

duration_ms_anomalies = [-107133, -223093, 10, 1000, 3000]

Nota:

Los limites superior e inferior de 630,000 y 82,000 milisegundos se eligen basados en la duración de la canción más larga y más corta de Taylor Swift, que tienen aproximadamente 10 minutos y 1 minuto y 22 segundos respectivamente.

2.19 Valores fuera del rango [2006, 2024] en la columna ‘album_release_date’:

Se encontraron 39 valores fuera del rango [2006, 2024] en la columna ‘album_release_date’ de 575 datos.

Ejemplo:

year_anomalies = ['2027-05-26', '1989-10-24']

Nota general:

No se presentan los registros completos de los datos anómalos en este informe, pero es factible generar un archivo CSV que contenga la información completa de cada anomalía.

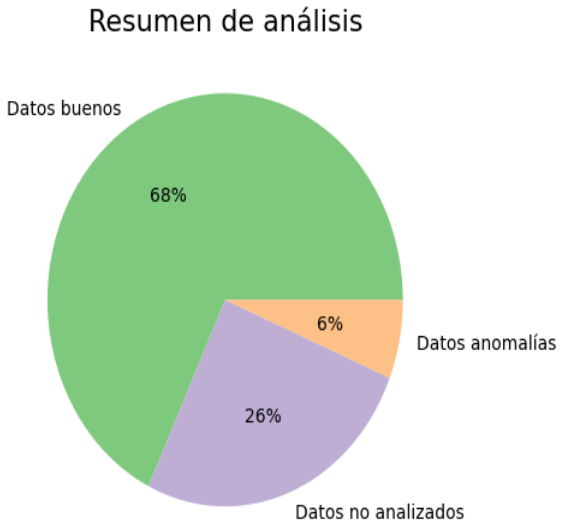
3. Estadistícas (Métricas)

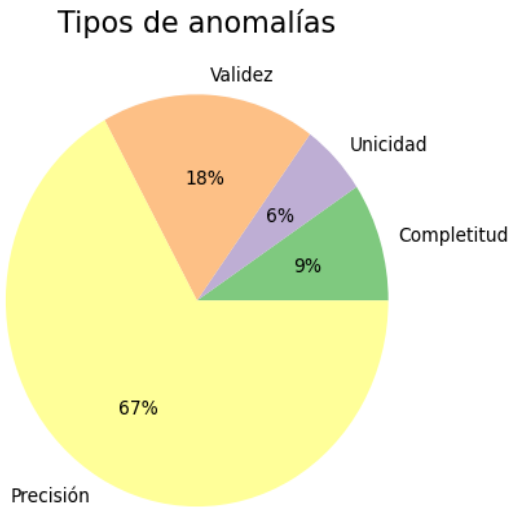
Estadísticas de los datos

Número de variables	27
Cantidad de observaciones	575
Total de datos	15525
Celdas vacías	89
Celdas vacías (%)	0.57
Filas duplicadas	54
Filas duplicadas (%)	9.39

Tipos de variables

Numérico	16
Texto	11
Fecha	0





Puntuación global

Calidad de los datos: 93.8

Puntuación del total de datos, comparada con cada una de las categorías de anomalías.

Análisis de valores únicos

Columnas de texto

explicit	4
track_id	512
track_name	331
audio_features.instrumentalness	240
audio_features.id	519
artist_id	1
artist_name	1
album_id	26
album_name	24
album_release_date	23
album_total_tracks	17

Columnas numéricas

disc_number	2
duration_ms	364
track_number	46
track_popularity	73
audio_features.danceability	267
audio_features.energy	348
audio_features.key	12
audio_features.loudness	448
audio_features.mode	2
audio_features.speechiness	292
audio_features.acousticness	401
audio_features.liveness	271
audio_features.valence	326
audio_features.tempo	450
audio_features.time_signature	3
artist_popularity	1

Columnas de fecha

Sin valores

Matriz de nulidad