INFORME

Análisis base de datos de base de datos de API de Spotify sobre Taylor Swift

Contenido

Introducción 4
Objetivos 4
Métodos 4
Variables 4
Tipo de variables 5
Estadísticas Descriptivas 7
Validación de duplicados 8
Anomalías 9
Conclusiones 11

Listado de Tablas

Tabla 1. Información de las variables 4

Tabla 2. Tipo de variables 6

Tabla 3. Estadística descriptiva de variables numéricas 7

Tabla 4.Canciones con duraciones inferiores a 3 minutos 9

Introducción

Este informe se basa en un conjunto de datos detallado obtenido a través de la API de Spotify, centrándose en la música de la artista Taylor Swift. La colección contiene 539 canciones y cubre una amplia gama de variables, desde información sobre las propias canciones hasta información específica sobre el álbum.

La calidad y precisión de estos datos son elementos clave del análisis de datos y son la base para identificar patrones y tendencias importantes que son esenciales para comprender y tomar decisiones.

Objetivos

El objetivo principal de esta revisión es detectar y caracterizar las anomalías en la calidad de los datos encontradas en el dataset. Este análisis considerará tres tipos principales de anomalías: valores faltantes, valores duplicados y discrepancias.

Métodos

Para lograr este objetivo, se utilizarán una variedad de técnicas analíticas. Se utilizará análisis de datos exploratorios para visualizar las distribuciones y comprender el comportamiento general de las variables. Además, se realizará un análisis exhaustivo de los valores faltantes, identificando patrones y cuantificando su presencia en cada variable.

Estos métodos se aplican cuidadosamente a todas las variables del conjunto de datos para identificar las anomalías más significativas y evaluar su impacto potencial en el análisis final.

Variables

El conjunto de datos contiene información sobre las siguientes variables:

Tabla 1. Información de las variables

Variable	**Descripción**	**Importancia para el análisis**
disc_number	Número de disco en el que aparece la canción.	Es importante para identificar la ubicación de la canción en el álbum.
duration_ms	Duración de la canción en milisegundos.	Es importante para identificar la duración de la canción.
Explicit	Indica si la canción contiene lenguaje o contenido explícitos.	Es importante para identificar las canciones que pueden ser inapropiadas para ciertos públicos.
track_number	Número de pista en el álbum.	Es importante para identificar la posición de la canción en el álbum.
track_popularity	Popularidad de la canción.	Es importante para identificar las canciones más populares de la artista.
track_id	ID de la canción.	Es un identificador único para la canción.
track_name	Nombre de la canción.	Es importante para identificar la canción.

Variable	Descripción	Importancia para el análisis
Danceability	Bailabilidad de la canción.	Es importante para identificar las canciones que son fáciles de bailar.
Energy	Energía de la canción.	Es importante para identificar las canciones que son animadas o enérgicas.
Key	Tono de la canción.	Es importante para identificar el tipo de música de la canción.
Loudness	Intensidad de la canción.	Es importante para identificar las canciones que son fuertes o suaves.
Mode	Modo de la canción.	Es importante para identificar el tipo de música de la canción.
Speechiness	Porcentaje de la canción que consiste en habla.	Es importante para identificar las canciones que contienen rap o spoken word.
Acousticness	Porcentaje de la canción que es acústica.	Es importante para identificar las canciones que son acústicas o instrumentales.
Instrumentalness	Porcentaje de la canción que es instrumental.	Es importante para identificar las canciones que son instrumentales.
Liveness	Porcentaje de la canción que se grabó en vivo.	Es importante para identificar las canciones que se grabaron en vivo.
Valence	Sentido de la canción.	Es importante para identificar las canciones que son positivas o negativas.
Tempo	Ritmo de la canción.	Es importante para identificar las canciones que son lentas o rápidas.
Id	ID de la canción.	Es un identificador único para la canción.
time_signature	Firma de tiempo de la canción.	Es importante para identificar el tipo de música de la canción.
artist_id	ID del artista.	Es un identificador único para el artista.
artist_name	Nombre del artista.	Es importante para identificar al artista.
artist_popularity	Popularidad del artista.	Es importante para identificar los artistas más populares.
album_id	ID del álbum.	Es un identificador único para el álbum.
album_name	Nombre del álbum.	Es importante para identificar el álbum.
album_release_date	Fecha de lanzamiento del álbum.	Es importante para identificar la fecha de lanzamiento del álbum.
album_total_tracks	Número total de pistas en el álbum.	Es importante para identificar el número total de pistas en el álbum.

Tipo de variables

Se evaluará el tipo de variable de cada una de las variables y adicionalmente, se verifica la cantidad de valores que contienen cada columna.

Existen valores faltantes en las columnas track_id, track_name, danceability, energy, loudness, speechiness, acousticness, liveness, tempo, y time_signature. La columna album_name presenta la mayor cantidad de valores faltantes, con 62 registros sin información.

Tabla 2. Tipo de variables

Columna	Valores no nulos	Tipo de datos
disc_number	539	int64
duration_ms	539	int64
explicit	539	object
track_number	539	int64
track_popularity	539	int64
track_id	531	object
track_name	532	object
danceability	537	float64
energy	537	float64
key	538	float64
loudness	537	float64
mode	539	int64
speechiness	538	float64
acousticness	538	float64
instrumentalness	539	object
liveness	538	float64
valence	539	float64
tempo	538	float64
id	539	object
time_signature	538	float64
artist_id	539	object
artist_name	539	object
artist_popularity	539	int64
album_id	539	object
album_name	477	object
album_release_date	539	object
album_total_tracks	539	object

El conjunto de datos contiene valores nulos en las siguientes columnas:

- track_id: 8 valores nulos.
- track_name: 7 valores nulos.
- danceability: 2 valores nulos.
- energy: 2 valores nulos.
- key: 1 valor nulo.
- loudness: 2 valores nulos.
- speechiness: 1 valor nulo.
- acousticness: 1 valor nulo.
- liveness: 1 valor nulo.

- tempo: 1 valor nulo.
- album_name: 62 valores nulos.

Las columnas restantes no presentan valores nulos. Es importante considerar la presencia de estos valores nulos al momento de realizar análisis o modelado sobre el conjunto de datos. Algunas opciones por considerar incluyen la eliminación de filas con valores nulos, la imputación de valores faltantes o el uso de técnicas de aprendizaje automático que sean robustas a la presencia de valores nulos. La estrategia más adecuada dependerá de la naturaleza de los datos y del objetivo del análisis.

Estadísticas Descriptivas

En relación con la tabla de descriptivas se encuentra la posible presencia de datos anómalos en las variables como:

- duration_ms: Presenta valores negativos como un valor máximo curiosamente grande.
- track_number: Se observa un valor máximo muy extremo, 46.
- track_popularity: Se evidencia presencia de valores negativos.
- Acousticness: Evidencia presencia de valores negativos y posibilidad de valores anómalos, valores superiores a 1.

Estas sospechas de anomalías de las variables se desarrollarán en la sección Anomalías. Adicionalmente se pueden evidenciar que no hay presencia de valores anómalos o atípicos para variables como: “artist_popularity”, “speechiness”, “valence”, “artist_name” y “artist_id” y demás variables no mencionadas en Tabla 3.

Tabla 3. Estadística descriptiva de variables numéricas

Variable	Conteo	Media	Desviación estándar	Mínimo	25%	50%	75%	Máximo
disc_number	539	1.032	0.175	1	1	1	1	2
duration_ms	539	236003.73	55019.87	-223093	209486.5	233626	259045.5	613026
track_number	539	11.28	7.966	1	5	10	15	46
track_popularity	539	62.918	22.499	-92	51	69	77	152
danceability	537	0.587	0.117	0.243	0.517	0.595	0.661	0.897
Energy	537	0.573	0.192	0.118	0.436	0.589	0.729	0.949
Key	538	4.587	3.246	0	2	5	7	11
loudness	537	-7.521	2.933	-17.932	-9.287	-6.942	-5.376	-1.909
mode	539	0.913	0.282	0	1	1	1	1
speechiness	538	0.058	0.073	0.023	0.031	0.038	0.056	0.912
acousticness	538	0.338	0.395	-0.004	0.036	0.168	0.664	5
liveness	538	0.163	0.142	0.034	0.097	0.115	0.162	0.931
tempo	538	122.363	30.485	68.097	96.685	119.001	143.939	208.918
time_signature	538	3.987	0.197	3	4	4	4	5

Validación de duplicados

Se encontraron 18 registros duplicados en el conjunto de datos. Estos registros duplicados se caracterizan por:

Todos los registros duplicados comparten los mismos valores en las siguientes columnas:

- track_id
- track_name
- artist_id
- artist_name
- album_id
- album_name
- album_release_date
- album_total_tracks

Las únicas columnas que difieren entre los duplicados son:

- disc_number
- duration_ms
- explicit
- track_number
- track_popularity
- danceability
- energy
- key
- loudness
- mode
- speechiness
- acousticness
- liveness
- tempo
- time_signature
- id

Algunas de las posibilidades de los valores duplicados son:

1. Posible origen de los duplicados:

   Es probable que los duplicados se deban a la combinación de los siguientes factores:

   - Presencia de canciones en diferentes ediciones de un mismo álbum.
   - Variabilidad en la duración de las canciones debido a diferentes versiones o errores de medición.
   - Inconsistencias en los metadatos de las canciones.

2. Implicaciones de los duplicados:

   Los duplicados pueden afectar negativamente a los análisis y modelos realizados sobre el conjunto de datos, ya que pueden sesgar los resultados.

   Es importante abordar la presencia de duplicados antes de realizar análisis posteriores.

3. Estrategias para manejar duplicados:
   - Eliminación de duplicados: Eliminar todas las filas duplicadas, conservando solo una instancia única de cada registro.
   - Combinación de información: Fusionar los valores de las columnas que difieren entre los duplicados, creando un registro unificado.
   - Investigación de la causa: Investigar la causa subyacente de los duplicados para determinar la mejor estrategia de manejo.

Anomalías

Analizando las diferentes variables del dataset se encuentran las siguientes anomalías:

- **Columna: duration_ms:** Se ajusta la duración en las diferentes canciones producidas se produce un intervalo de canciones usuales entre 3 y 6 min, con excepción de la canción con duración de 10 min.
  Se encuentra la presencia de 12 canciones con duración superior a 6 min. Al evaluar este grupo de canciones se encuentran diferencias entre las duraciones reales de las canciones en:
  - Dear John (Taylor's Version) (pista 47) - 4:31 minutos
  - Last Kiss (Taylor's Version) (pista 55) - 4:14 minutos
  - Dear John (pista 449) - 3:36 minutos
  - Last Kiss (pista 457) - 3:05 minutos
  - Dear John (pista 463) - 3:36 minutos
  - Last Kiss (pista 471) - 3:05 minutos

  Estas pistas tienen todas una duración inferior a 6 minutos, aunque algunas de ellas comparten el mismo nombre que versiones más largas ("All Too Well" y versiones en vivo).

  Para el caso de las canciones menores a 3 min se encuentran 33 observaciones donde 2 de ellas son inferiores a cero. Por tanto, se recomienda evaluar y verificar las siguientes canciones:

Tabla 4.Canciones con duraciones inferiores a 3 minutos

disc_number	duration_ms
18	146436
19	171818
40	146436

disc_number	duration_ms
41	171818
70	174782
72	164801
82	148781
94	174782
96	164801
106	148781
114	174782
116	164801
277	170640
278	178426
282	173386
290	171360
293	150440
295	170640
296	178426
300	173386
308	171360
311	150440
392	-107133
393	131186
408	-223093
420	10
432	1000
440	83253
472	3000
518	176316
521	173899
525	17

Adicionalmente, se encuentra una canción con una duración de alrededor de 10 min, esta corresponde la canción más extensa; una versión de ("All Too Well").

- **Columna: explicit:** Para esta variable los valores más frecuentes son: “TRUE” y “FALSE”**.** Se encuentra dos opciones como “SI” y “NO” con frecuencia 1 y 4 respectivamente, los cuales sugerimos clasificar como valores anómalos.

- **Columna: album_total_tracks:** Esta variable presenta la cantidad de pistas en el álbum. Esta se presenta de forma numérica. Por tanto, se presenta como valor anómalo: “Thirteen”. Adicionalmente, se considera importante verificar el valor

numérico “46” el cual es un valor atípico en la distribución pistas en los demás álbumes.

- **Columna: acousticness:** Presenta una distribución de valores entre la escala de 0 a 1. Sin embargo se encuentra la presencia de tres valores superiores a 1, como son: “1.5”, “2” y “5”. Estos valores se pueden evidenciar como valores anómalos.

- **track_popularity:** La variable maneja una escala de 0 a 100. Por tanto, los valores fuera de este rango se pueden considerar como valores anómalos o atípicos. Se evidencia los siguientes valores: “-69”, “-70”, “-85”, “-92”, “-75”, “-71” y “152”.

- **album_release_date:** Se encontraron valores anómalos como el año 2027 con una frecuencia de 24 observaciones y el año 1989 con 15 observaciones. Lo anterior se considera dado que para el primer álbum de Taylor Swift fue lanzado en 2006.

Posibles causas:

- **Errores de entrada de datos:** Los valores pueden haberse ingresado incorrectamente, como errores tipográficos o categorizaciones erróneas.
- **Formato inconsistente:** Los valores pueden haberse formateado de manera diferente, lo que genera inconsistencias.
- **Datos corruptos:** Es posible que algunos valores sean alterados durante la transferencia o el almacenamiento de datos.
- **Problemas con la fuente de datos externa:** Si los datos se importaron de una fuente externa, es posible que haya habido errores en los datos originales.

Conclusiones

En este informe se ha realizado un análisis de calidad de datos sobre un conjunto de datos de 539 canciones de la artista Taylor Swift. El análisis ha identificado las siguientes anomalías:

- En la columna “duration_ms” verificar las canciones con relación a su duración real. Estos valores ayudaran a generar fiabilidad a los datos.
- En la columna "explicit", se han encontrado dos valores anómalos: "SI" y "NO". Estos valores son diferentes de los valores esperados para esta columna, que son "True" o "False". La causa probable de estas anomalías es un error de entrada de datos.
- En la columna "album_total_tracks", se ha encontrado un valor anómalo: "Thirteen". Este valor es diferente de los valores esperados para esta columna, que son números enteros positivos. La causa probable de esta anomalía es un error de entrada de datos.
- En la columna "acousticness", se han encontrado tres valores anómalos: "1.5", "2" y "5". Estos valores son superiores al valor máximo esperado para esta columna, que es 1. La causa probable de estas anomalías es un error de entrada de datos o una medición incorrecta.
- En la columna "track_popularity", se han encontrado ocho valores anómalos: "-69", "-70", "-85", "-92", "-75", "-71" y "152". Estos valores están fuera del rango esperado

para esta columna, que es de 0 a 100. La causa probable de estas anomalías es un error de entrada de datos o una medición incorrecta.
- En la columna “album_release_date” se recomienda verificar los años señalados para solucionar el error de esas 39 observaciones.

Recomendaciones

Para mejorar la calidad del conjunto de datos, se recomienda realizar las siguientes acciones:

- En la columna "explicit", se recomienda eliminar los valores anómalos.
- En la columna "album_total_tracks", se recomienda verificar el valor "46" y, si es necesario, eliminarlo.
- En la columna "acousticness", se recomienda eliminar los valores anómalos.
- En la columna "track_popularity", se recomienda investigar la causa de los valores anómalos y, en función de la causa, tomar las medidas adecuadas, como eliminar los valores anómalos o reemplazarlos con valores correctos.

Impacto de las anomalías

Las anomalías encontradas en el conjunto de datos pueden afectar negativamente los análisis realizados sobre este conjunto de datos. Por ejemplo, los valores anómalos en la columna "explicit" pueden sesgar los resultados de un análisis sobre la popularidad de las canciones. Por lo tanto, es importante abordar las anomalías encontradas antes de realizar análisis posteriores.

Además de las recomendaciones específicas para cada anomalía, se recomienda realizar un análisis adicional para identificar otras posibles anomalías en el conjunto de datos. Este análisis podría incluir:

- Un análisis de la distribución de los valores en cada columna.
- Una comparación de los valores en cada columna con valores esperados o con valores de otros conjuntos de datos similares.
- Este análisis adicional ayudará a garantizar que el conjunto de datos sea de alta calidad y que pueda utilizarse para realizar análisis precisos.