PR1 Visualización de Datos

Alberto Cacho Picón

3 de Diciembre

Tras revisar numerosas posibles fuentes para el proyecto, valoré inicialmente distintos repositorios estadísticos (entre ellos el INE y varios portales con datos procedentes de Estados Unidos) que ofrecían información demográfica, económica o sociocultural. Aunque presentaban contenido interesante, ninguno se ajustaba de forma precisa al tipo de estudio que pretendía desarrollar ni a mis preferencias personales.

La búsqueda terminó siendo fructífera al acudir a *Kaggle*, donde localicé un conjunto de datos centrado en **Spotify** correspondiente al año 2025. Este dataset incluye información detallada tanto sobre canciones como sobre características musicales y metadatos de los artistas. Al tratarse de un ámbito cercano a mis gustos personales, el trabajo con este tipo de datos resulta motivador y adecuado para aplicar técnicas de análisis, visualización y modelización.

El material seleccionado destaca por su amplitud y por la variedad de variables disponibles, lo que permite examinar tendencias musicales, patrones de popularidad y posibles relaciones entre artistas y sus obras. Para disponer de una visión más completa, integré dos fuentes distintas: una orientada a las canciones y otra centrada en los artistas. Ambas se pueden consultar en *Canciones Spotify 2025* y *Metadatos de artistas*.

El hecho de que los datos correspondan a 2025 resulta especialmente relevante, ya que la música es un ámbito en constante cambio. Utilizar información reciente permite extraer conclusiones alineadas con las preferencias actuales de los usuarios y con la evolución reciente del mercado. Además, al analizar una plataforma con millones de oyentes, el estudio adquiere valor para comprender qué estilos predominan, qué artistas emergen y cómo fluctúan las tendencias.

El dataset incluye, entre otros aspectos, información sobre el género de los artistas, lo cual hace posible explorar si existen diferencias de visibilidad o popularidad según esta variable. Dado que la industria musical ha mostrado desigualdades históricas, disponer de estos datos abre la puerta a examinar si dichas brechas persisten o se han reducido.

Para obtener un único conjunto homogéneo, se llevó a cabo una fase de preparación mediante la librería `pandas`: limpieza de valores, eliminación de registros duplicados y normalización de los nombres de los artistas. Gracias a ello, la fusión de las tablas (realizada mediante el nombre del artista como clave de unión) pudo efectuarse sin inconsistencias.

El dataset resultante está formado por **5404 registros** y **28 variables**. La estructura combina **variables numéricas** (popularidad, duración, número de seguidores, edad del artista, etc.) con **variables categóricas** (género musical, tipo de artista, país, indicador de contenido explícito...). La presencia de información geográfica permite además comparar comportamientos entre distintas nacionalidades y explorar diferencias culturales en el consumo musical.

En conjunto, el dataset final constituye una base sólida y suficientemente diversa como para desarrollar un análisis exhaustivo del comportamiento musical en Spotify. Tanto el código empleado para la limpieza como el archivo final utilizado en el estudio se encuentran disponibles en mi repositorio de GitHub.

En relación con las preguntas que se abordarán mediante las técnicas de visualización, cabe señalar que todas ellas han sido definidas teniendo en cuenta las características del conjunto de datos y los objetivos expuestos previamente. Las cuestiones seleccionadas permiten explotar tanto las variables cuantitativas como las cualitativas disponibles, y se apoyan directamente en los aspectos que hacen relevante el dataset: actualidad de la información, diversidad de características musicales, presencia de metadatos asociados a artistas y disponibilidad de información geográfica.

Aunque algunas de estas preguntas han sido tratadas en proyectos similares (especialmente en estudios centrados en tendencias musicales, popularidad o análisis de géneros), la combinación concreta de variables de este dataset y la integración realizada entre canciones y artistas permite plantearlas desde una perspectiva más completa. Esto aporta valor añadido frente a visualizaciones previas que suelen centrarse únicamente en un tipo de información (p. ej., solo canciones o solo artistas). A continuación se presentan algunas de las cuestiones que pueden abordarse mediante técnicas de visualización utilizando el conjunto de datos seleccionado:

- **Popularidad y tendencias**
  - ¿Qué características musicales (como *energy*, *danceability*, *tempo* o duración) presentan mayor relación con la popularidad de las canciones?
  - ¿Qué géneros musicales muestran mayor popularidad media en 2025?
  - ¿Existe alguna asociación entre el número de seguidores de un artista y el nivel de popularidad de sus canciones?
- **Análisis de artistas**
  - ¿Cómo se distribuye la edad de los artistas presentes en el dataset y qué relación tiene con la popularidad?
  - ¿Se observan diferencias en visibilidad o éxito según el género del artista?
  - ¿Qué países concentran un mayor número de artistas influyentes dentro del catálogo analizado?

- **Características musicales**
  - ¿Qué géneros destacan por valores elevados de características acústicas como *valence*, *acousticness* o *instrumentalness*?
  - ¿Cómo varía la distribución de la duración de las canciones entre géneros o tipos de artista?
  - ¿Existen patrones diferenciados en la presencia de canciones explícitas según el género musical o el país del artista?
- **Dimensión geográfica**
  - ¿Qué países presentan mayor representación de artistas dentro del dataset?
  - ¿Se aprecian diferencias en popularidad media entre artistas según su país de origen?

A continuación, el diccionario pedido:

Diccionario de Datos

1. **track_id**: Identificador único de la canción (Spotify ID). *Tipo: Texto.*
2. **track_name**: Nombre de la canción. *Tipo: Texto.*
3. **track_number**: Número de pista dentro del álbum. *Tipo: Entero.*
4. **track_popularity**: Popularidad de la canción (0–100). *Tipo: Entero.*
5. **explicit**: Indica si la pista contiene contenido explícito. *Tipo: Booleano.*
6. **artist_name**: Nombre del artista principal. *Tipo: Texto.*
7. **artist_popularity**: Popularidad del artista (0–100). *Tipo: Entero.*
8. **artist_followers**: Número de seguidores del artista. *Tipo: Entero.*
9. **album_id**: Identificador único del álbum. *Tipo: Texto.*
10. **album_name**: Nombre del álbum. *Tipo: Texto.*
11. **album_release_date**: Fecha de lanzamiento del álbum. *Tipo: Fecha.*
12. **album_total_tracks**: Número total de pistas en el álbum. *Tipo: Entero.*
13. **album_type**: Tipo de álbum (single, album, compilation...). *Tipo: Texto.*
14. **track_duration_min**: Duración de la pista en minutos. *Tipo: Decimal.*
15. **gender**: Género del artista/persona. *Tipo: Texto.*
16. **age**: Edad del artista/persona. *Tipo: Entero.*

17. **type**: Tipo de entidad (artista, persona, etc.). *Tipo: Texto.*
18. **country**: País asociado al artista/persona. *Tipo: Texto.*
19. **city_1**: Primera ciudad relacionada. *Tipo: Texto.*
20. **district_1**: Primer distrito relacionado. *Tipo: Texto.*
21. **city_2**: Segunda ciudad relacionada. *Tipo: Texto.*
22. **district_2**: Segundo distrito relacionado. *Tipo: Texto.*
23. **city_3**: Tercera ciudad relacionada. *Tipo: Texto.*
24. **district_3**: Tercer distrito relacionado. *Tipo: Texto.*
25. **Unnamed: 0**: Columna residual del archivo. *Tipo: Entero.*
26. **index**: Índice del DataFrame exportado desde pandas. *Tipo: Entero.*