PROYECTO PODCASTS - SPOTIFY

Proyecto integrador UNICORN ACADEMY.

Agustin Felix Albornoz Bruschetti

Raquel Lorenzo Fernández

Ramyi Gacela Gossen López

Contenido

PROYECTO PODCAST - SPOTIFY

Resumen y Objetivo

Este proyecto analiza la evolución global en la publicación de episodios de podcast entre 2019 y 2024, a partir de datos obtenidos directamente de la API pública de Spotify. El objetivo es analizar tendencias, evolución y características de los podcasts a partir de los datos extraídos. Nos dimos cuenta de que, durante ese periodo, tuvo lugar un evento global que impactó profundamente a todo el mundo: la pandemia de COVID-19. Esto nos permitió, además, analizar cómo se comportaron la publicación y el consumo de podcasts bajo demanda en un contexto tan excepcional.

En total, se recopilaron más de 83.000 episodios correspondientes a 159 podcasts, procedentes de seis países (Argentina, España, Brasil, Portugal, Estados Unidos y Reino Unido) y en tres idiomas principales (español, portugués e inglés), mediante consultas a la API pública de Spotify.

Debido a las limitaciones relacionadas con la propia API (como el número máximo de consultas y paginación de resultados), se optó por focalizar el análisis en seis países representativos por idioma y volumen de producción.

Cabe señalar que no es posible realizar una comparación directa entre los idiomas, ya que únicamente se recopilaron episodios en el idioma predominante de cada país. Es decir, los datos obtenidos corresponden exclusivamente a podcasts en español para Argentina y España, en portugués para Brasil y Portugal, y en inglés para Estados Unidos y Reino Unido.

Tras el proceso de filtrado, limpieza y depuración de registros inconsistentes o incompletos, se analizaron 54.212 episodios pertenecientes a 118 podcasts. Los datos fueron estructurados y tratados en Python, y visualizados mediante paneles interactivos en Power BI.

Los detalles metodológicos, scripts y bases de datos empleados pueden consultarse en nuestro repositorio digital y en nuestra página web: https://jr3datastudio.carrd.co/

Fuente de datos

La información se obtuvo directamente de la API oficial de Spotify y fue procesada con Python, almacenando en un archivo CSV estructurado y listo para su análisis. La base de datos incluye **83.315 episodios** correspondientes a **159 podcasts** de 6 países (AR, ES, BR, PT, US, GB) y 3 idiomas (es, pt, en).

Herramientas utilizadas

- **Spotify API:** Para acceder y descargar los datos directamente desde Spotify.
- **Python:** Utilizado en la recolección, procesamiento y análisis de datos para estudiar el comportamiento de las variables clave.
- **Power BI:** Empleado para analizar los resultados y responder preguntas de negocio a través de visualizaciones claras e interactivas.

Ambas herramientas las integraremos entre sí para las siguientes etapas del proyecto.

Metodología:

1. Extracción estructurada de datos a través de la API de Spotify.

La extracción de datos se realizó mediante la API oficial de Spotify, autenticando cada petición con un acceso token. El proceso consistió en buscar podcasts por palabra clave, extraer todos sus episodios mediante paginación, debido a que la API restringe el número de episodios que pueden descargarse en una sola consulta y filtrar los resultados por país, idioma y rango de fechas (2019–2024). Los datos válidos se estructuraron en un DataFrame, incorporando variables clave como duración, fecha y editorial*(show)*.

Este **script (https://developer.spotify.com/documentation/web-api)** nos permitió construir una base de datos sólida de podcasts desde la API de Spotify, lista para su análisis descriptivo y visual en herramientas como Python (Seaborn, Matplotlib) y Power BI.

Si bien nos enfrentamos con desafíos relacionados con las limitaciones de la API, el uso de datos reales otorga un enorme valor al proyecto, tanto a nivel académico como profesional.

Ventaja	**Descripción**
Datos reales y actualizados	Se extraen directamente de la API oficial de Spotify, garantizando un reflejo fiel del estado actual de la publicación de podcasts.
Proceso repetible y adaptable	Automatizado en Python para su ejecución tantas veces como se desee; permite parametrizar fácilmente la extracción por tema, intervalo o región.
Compatibilidad con herramientas de análisis	Exportación a CSV que puede abrirse en Excel, Power BI o procesarse con scripts en Python; integración sencilla en pipelines de visualización y reporting.
Problemas y retos que encontramos	**Descripción**
Límites de la API	No se pueden pedir demasiados datos de golpe (*No se puede realizar peticiones de extracción en grandes volúmenes de datos*). Hay que esperar entre consultas para evitar bloqueos.
Hay filtro por *idioma*	Hay que usar el idioma como filtro de referencia.
Podcasts sin episodios públicos	A veces encontramos podcasts vacíos, sin episodios accesibles.
Datos faltantes o raros	Hay pocos episodios sin fecha o duración, o con errores en la información.

2. Python. Procesamiento y análisis de datos

Visualización de las columnas del dataset original, y transformaciones que fueron necesarias para el análisis

Nº	Dataset Original	Dataset Tras Limpieza
1	pais	pais
2	idioma_show	idioma_show
3	keyword	categoria
4	podcast_id	podcast_id
5	nombre_podcast	nombre_podcast
6	editorial	editorial
7	total_episodios	total_episodios
8	episodios_descargados	episodios_descargados
9	episodio_id	proporcion_episodios
10	nombre_episodio	episodio_id
11	fecha_lanzamiento	nombre_episodio
12	duracion_min	fecha_lanzamiento
13		año (nueva columna)
14		mes (nueva columna)
15		duracion_min (nueva columna)
16		categoria_duracion (nueva columna)
17		periodo_pandemia (nueva columna)

Limpieza y depuración de los datos:

- **Filtrado temporal (2019-2024):** Sólo analizamos episodios del periodo relevante, asegurando coherencia histórica y métricas comparables.Evolución temporal (gráfico de líneas) de episodios publicados entre 2019 y 2024
- **Definición del período pandémico:** Se considera como periodo de pandemia el comprendido entre el 11 de marzo de 2020 y el 5 de mayo de 2023, según la OMS.
- **Categorización de periodos:** Dividimos el análisis en tres etapas:

Pre-pandemia	Pandemia (según OMS)	Post-pandemia
Hasta 10/03/2020	11/03/2020 – 05/05/2023 (definido por OMS)	Desde 06/05/2023

- **Conversión de fechas a datetime:** Permite segmentar por año, mes y trimestre, facilitando el análisis de tendencias y la creación de series temporales.
- **Limpieza de texto:** Eliminamos símbolos y emojis para obtener nombres claros y homogéneos, mejorando la calidad de tablas y gráficos.
- **Normalización de idioma y país:** Unificamos formatos y evitamos duplicados, agilizando agrupaciones y asegurando consistencia en el análisis.
- **Traducción y capitalización de categorías:** Todas las columnas han sido traducidas al español; de este modo, se consigue uniformidad y se mejora la claridad de los datos, facilitando su lectura y posterior análisis temático.
- **Columnas derivadas:** Añadimos etiquetas útiles como periodo pandémico y categoría de duración, lo que permite comparaciones contextuales y análisis más detallados.
- **Métrica de proporción descargada:** Calculamos la relación entre episodios descargados y totales para detectar posibles vacíos de información.
- **Separación de tablas:** Organizamos los datos en tablas de episodios y de podcasts, evitando duplicados y permitiendo un modelo de análisis más limpio en Power BI.
- **Re-cálculo y orden de columnas:** Garantizamos que las columnas clave estén completas y bien organizadas, dejando el dataset listo para exportar y futuras actualizaciones.
- **Análisis descriptivo:** medias, desviaciones y boxplots por ** país, año y categoría de duración. **(... por las columnas cuantitativas que son: total_episodios, episodios_descargado y duracion de minutos. Posteriormente para un analisis del comportamiento más detallado se tomo la categoría de “periodo_pandemia”)
- **Outliers.** Detectamos y analizamos los valores atípicos presentes en el conjunto de datos. Para ello, utilizamos la gráfica Boxplots ya que nos permite visualizar de forma sencilla los registros que se alejaban del rango habitual de los datos.
- **Resumen análisis estadístico detallado:** Este informe presenta un análisis descriptivo de los podcasts de Spotify entre 2019 y 2024, basado en un CSV depurado con más de 50 000 registros y variables clave (país, idioma, fecha, duración y categoría). Se repasará el propósito del estudio, el origen y la estructura de los datos, y se mostrará cómo se procesaron con Python y se visualizaron en Power BI, manteniendo un enfoque práctico pero sin asumir conocimientos avanzados. Si deseas ver el desarrollo paso a paso, puedes consultar el notebook de Python Script_final.ipynb donde está todo el código reproducible. A lo largo del documento detallamos la evaluación y limpieza de la calidad de los datos; el análisis

exploratorio de las métricas principales y su evolución en los periodos pre-, durante y post-pandemia; el examen de correlaciones entre variables cuantitativas; y las conclusiones más reveladoras. Cerraremos con recomendaciones y futuras líneas de trabajo, cómo ajustar resultados por antigüedad, aplicar modelos de series temporales o profundizar en el análisis temático.

Informe estadístico descriptivo de podcasts de Spotify (2019 – 2024)

Dataset analizado: spotify_podcasts_limpio_2019_2024.csv (54 212 registros, 17 variables)

1 · Objetivo

El propósito de este informe es presentar los hallazgos de un análisis estadístico descriptivo de los podcasts alojados en Spotify entre 2019 y 2024. El documento está dirigido a una persona en formación en Estadística, por lo que combina explicación conceptual y resultados numéricos para facilitar la comprensión.

2 · Descripción del dataset

Variable	Tipo	Descripción
pais	categórica	País donde se descargó el episodio
idioma_show	categórica	Idioma principal del podcast
categoria	categórica	Género o temática principal
podcast_id	id	Identificador único del podcast
nombre_podcast	texto	Nombre del programa
editorial	texto	Productora o autor principal
total_episodios	numérica (entero)	Nº total de episodios publicados
episodios_descargados	numérica (entero)	Nº de episodios disponibles para descarga en el momento del scraping
proporcion_episodios	numérica (ratio)	episodios_descargados / total_episodios
episodio_id	id	Identificador único del episodio
nombre_episodio	texto	Título del episodio
fecha_lanzamiento	fecha	Fecha original de publicación
año, mes	derivados	Componentes temporales
duracion_min	numérica (continua)	Duración del episodio en minutos
categoria_duracion	categórica	Segmento de duración (muy corto ≤ 5 min, corto 5-20 min, medio 20-60 min, largo > 60 min)
periodo_pandemia	categórica ordenada	pre-pandemia (≤ feb-2020), pandemia (mar-2020 → dic-2021), post-pandemia (≥ ene-2022)

Dimensión: 54 212 episodios provenientes de 10 266 podcasts.
Distribución por periodo_pandemia: 24 507 (45 %) *pandemia* · 17 480 (32 %) *post-pandemia* · 12 225 (23 %) *pre-pandemia*.

3 · Calidad de los datos

- **Valores faltantes:** se detectaron 0,4 % de títulos de episodio vacíos (217 registros). No afectan al análisis numérico, pero se recomienda imputarlos o eliminarlos en estudios de lenguaje natural.
- **Tipos consistentes:** todas las variables cuantitativas se leen como int64 o float64. No se encontraron fechas inválidas.

4 · Exploración inicial y estadísticos globales

Métrica global	total_episodios	episodios_descargados	duracion_min
Media	1 370	1 122	32,4 min
Mediana	792	617	26,3 min
Desv. típica	1 322	1 149	31,0 min
Rango (mín–máx)	1 – 4 999	0 – 4 999	0,02 – 706 min

5. Distribución y outliers

- Las tres variables presentan colas derechas pronunciadas (asimetría positiva).
- Con la regla **IQR (Q3 + 1,5·IQR)** se identificaron **9 620 episodios atípicos** (17,7 % del total).
- Al filtrar estos valores extremos:
  - La media de total_episodios cae −38 %.
  - La media de duracion_min cae −18 %, revelando que los outliers correspondían a duraciones > 3 h y a catálogos extremadamente prolíficos.

6 · Comparativa por periodo de pandemia (sin outliers)

Periodo	n	_Total episodios_ Media · Mediana	_Episodios descargados_ Media · Mediana	_Duración (min)_ Media · Mediana
pre-pandemia	12 132	869 · 645	747 · 559	42,7 · 33,5
pandemia	20 075	1 138 · 1 000	1 013 · 890	26,8 · 17,6
post-pandemia	12 385	857 · 554	608 · 430	33,9 · 27,5

Volumen de producción: Durante la pandemia se alcanzó el catálogo más extenso (↑ 31 % respecto al total pre-pandemia).

Descargas disponibles: Siguen la misma tendencia —indican mayor trabajo de remasterización/re-host durante 2020-2021.

Duración típica: Los episodios pre-pandemia eran sensiblemente más largos (≈ 43 min). El confinamiento trajo episodios más cortos (≈ 27 min), quizá por grabaciones caseras; la post-pandemia recupera algo de longitud, pero no alcanza los valores de 2019.

7 · Correlaciones entre variables cuantitativas (sin outliers)

Método	ρ(total ↔ descargados)	ρ(total ↔ duración)	ρ(descargados ↔ duración)
Pearson	0,97	−0,46	−0,46
Spearman	0,97	−0,49	−0,49
Kendall	0,87	−0,35	−0,36

Existe una correlación casi perfecta entre el tamaño total del catálogo y los episodios realmente disponibles, dado que ambos son conteos acumulados. En cambio, la duración media guarda una correlación moderada-negativa: podcasts muy extensos suelen tener episodios más breves y viceversa.

8· Conclusiones principales

- **Fase de expansión 2020-2021**: El auge pandémico impulsó tanto la creación como la digitalización de contenido (↑ volumen y descargas).
- **Ajuste de formato**: La longitud de los episodios se acortó significativamente en 2020 y no ha vuelto a los niveles de 2019.
- **Outliers relevantes**: El 8 – 18 % de los registros inflan las medias y las varianzas; al filtrarlos entrega una foto más precisa del podcast "promedio".
- **Estructura por episodios**: Los podcasts con muchos episodios tienden a ofrecerlos casi completos (ρ≈0,97), mostrando buenas prácticas de *hosting/archiving*.

9 · Recomendaciones y próximos pasos

- **Normalizar por antigüedad**: Ajustar por años activos del podcast para comparar la productividad relativa.
- **Modelos de tendencia**: Usar descomposición STL o Prophet con la serie año-mes → episodios_nuevos para pronosticar lanzamientos.
- **Análisis de audiencia** *(si se dispone)*: Relacionar descargas con métricas de escucha para medir retorno de la inversión de episodios más largos vs. cortos.
- **Segmentación temática**: Profundizar en categorías (comedia, noticias, educación…) para detectar nichos afectados de forma diferente por la pandemia.

Con este dataset limpio y estructurado, pasamos a trabajar con Power Bi.Todos los cálculos se realizaron en Python 3.12 (pandas 2.2, seaborn 0.13, matplotlib 3.9) empleando un entorno Jupyter; se incluye notebook Script_final.ipynb con el código reproducible.

3. Dashboard Power BI:

Este proyecto final del Bootcamp de análisis de datos consistió en desarrollar una solución interactiva en Power BI para explorar la producción de podcasts en distintos países y categorías temáticas. A través del estudio de métricas como el número de episodios, la duración total y media de las publicaciones, y su evolución temporal, se llevó a cabo un análisis comparativo entre los períodos pre- y post-pandemia, omitiendo intencionadamente el tramo pandémico para aislar su impacto en la muestra. Cabe destacar que, debido a restricciones en la API de Spotify durante la descarga, los resultados reflejan el comportamiento de la muestra y no una visión global del mercado.

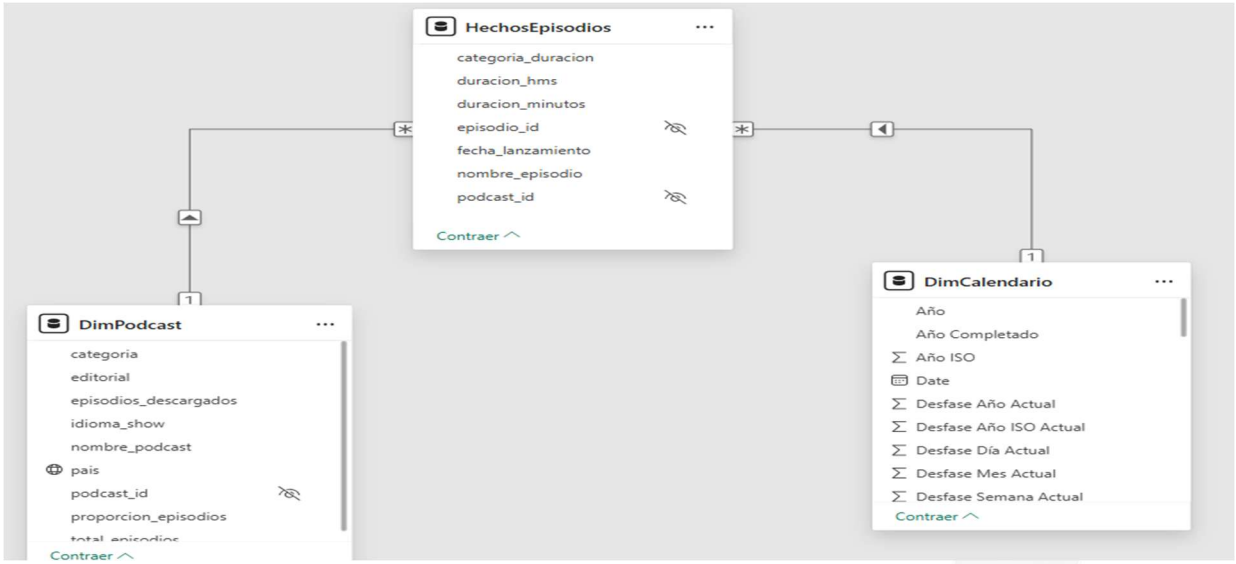
El dashboard resultante se estructura de manera secuencial y presenta un diseño claro, con filtros dinámicos por país y categoría que facilitan tanto la exploración libre de los datos como la presentación dirigida de hallazgos. Mediante el uso de modelado de datos, fórmulas DAX y técnicas de visualización y storytelling analítico, la herramienta demuestra las habilidades técnicas y críticas adquiridas a lo largo de la formación.

Preparación previa para la presentación del Dashboard

- Tras análisis y limpieza previa en Python, se importa un nuevo csv limpio a Power Bi.
- Se configura una función calendario en lenguaje M para poder incluir la categoría o columna periodo pandemia en la tabla calendario.
- Se normaliza dimensiones de la duración_min original (duracion_minutos final) dividiéndola entre 100 por que Python el decimal lo reconoce con un punto y en powerbi es una coma.
- Las clasificaciones por duración se dejaron en formato texto (categoria_duracion)
- Storytelling analítico orientado a usuarios finales.
- **Tablas creadas en Power BI**. Tabla de hechos: HechosEpisodios Contiene información a nivel de episodio individual →Clave principal: episodio_id
- **Columnas eliminadas**: columnas duplicadas o mal ubicadas como proporcion_episodios, duracion_hms. Las columnas año, mes, periodo_pandemia son eliminadas porque se encuentran en la tabla DimCalendario.
- **Dimensión: DimPodcast**. Contiene información única por podcast/show Clave principal: podcast_id
- **Dimensión: DimCalendario**
- **Tabla calendario generada en Power Query con función personalizada → Clave:** Date (relacionada con fecha_lanzamiento en HechosEpisodios)
- La categoría Es Pandemia fue añadida con este bloque de código M: Esto clasifica las fechas en 3 periodos:

Período	Rango de fechas
Pre-Pandemia	Hasta el 10 de marzo de 2020
Pandemia	Desde el 11 de marzo de 2020 al 05 de mayo de 2023
Post-Pandemia	Desde el 06 de mayo de 2023 en adelante

- **El modelo estrella** tiene las siguientes relaciones:



Resumen y estructura del dashboard

El diseño se realizó por secciones:

1. Análisis General

Esta sección proporciona un panorama general de la producción de podcasts, reuniendo datos sobre el total de episodios y minutos publicados, sus variaciones anuales y los indicadores más relevantes de crecimiento o descenso.

Su función es establecer el punto de partida para valorar cómo ha evolucionado el fenómeno del podcasting a lo largo del tiempo.

2. Evolución de Producción

Esta sección profundiza en cómo ha variado la producción de podcasts con el tiempo, examinando series temporales por país y por categoría para comparar el volumen mensual y anual. Al resaltar tendencias y momentos de mayor o menor actividad, obtenemos información clave para anticipar fluctuaciones y orientar estrategias de contenido.

3. Impacto Pre-Post Pandemia

En esta parte del informe examinamos cómo la pandemia influyó en la producción de podcasts, comparando el volumen de episodios, el total de minutos publicados y la duración media por episodio antes y después del brote pandémico. Además, desglosamos estos cambios por país y categoría temática y calculamos indicadores clave que cuantifican la variación porcentual entre ambos periodos.

Al centrar el análisis en estos periodos temporales, ganamos un contexto fundamental que enriquece el resto de métricas y permite entender mejor las dinámicas de creación durante y tras la crisis sanitaria. Este enfoque aporta profundidad y relevancia al estudio general.

4. Distribución

Aquí exploramos la longitud de los episodios, usando gráficos como histogramas y cajas para ver su distribución por país o categoría, y clasificándolos en cortos, medianos y largos. Al comparar estos grupos descubrimos distintos patrones de formato que facilitan la identificación de audiencias y estilos de contenido.

5. Outliers

Aquí buscamos los episodios atípicos, aquellos muy largos o muy cortos, o que se publican con una frecuencia inusual. Luego vemos cuánto alteran las cifras promedio y la dispersión de los datos.

4. MySQL Workbench Validación cruzada y consultas de negocio.

Aunque nuestro proyecto se apoyó sobre todo en **Python** para limpiar y preparar los datos, y en **Power BI** para analizarlos y visualizarlos, también quisimos dar un paso más: comprobar que todo lo que estábamos viendo era correcto directamente desde la base de datos.

Para eso usamos **MySQL**. Cargamos el CSV limpio, definimos bien las columnas y sus tipos de datos y, con el *Table Data Import Wizard*, dejamos todo listo para lanzar consultas SQL que nos confirmaran que los resultados coincidían al 100 % con lo procesado en Python.

Esta validación extra nos sirvió para dos cosas:

1. Asegurarnos de que Python y Power BI estaban trabajando con datos correctos.
2. Tener una forma rápida de responder preguntas de negocio sin necesidad de pasar por todo el proceso de visualización.

En esta revisión confirmamos que, después de la limpieza, teníamos **54 212 episodios** de **118 podcasts**, distribuidos en **6 países** (AR, BR, ES, GB, PT, US) y **3 idiomas** (en, es, pt).

total_episodios	podcasts_distintos	num_paises	lista_paises	num_idiomas	lista_idiomas
54212	118	6	AR,BR,ES,GB,PT,US	3	EN,ES,PT

Las consultas que hicimos fueron de todo tipo:

- Total de episodios, podcasts, países e idiomas, y sus listados.

- Evolución anual de la producción (con el pico claro de 2020).
- Distribución por categorías y duración media (global y por temática).
- Cuándo se publica más (por mes y día de la semana).
- Comparativa pre-pandemia, pandemia y post-pandemia.
- Rankings de podcasts y países líderes, y programas con mayor duración media.

Al principio intentamos cargar mi archivo CSV usando la herramienta de **Data Import Wizard** de MySQL Workbench, que es la forma más rápida y la que siempre usamos para probar. Pero con mi dataset grande (más de 54 000 filas y 17 columnas) no me estaba cargando todos los datos bien: en vez de darme los 6 países y 3 idiomas que sé que existen, solo veía 3 países. O sea, no era un problema del archivo, sino de cómo se estaba leyendo.

Como **MySQL 9.2 tiene una restricción de seguridad** que se llama *secure_file_priv*. Eso significa que MySQL solo permite cargar archivos desde una carpeta especial del servidor (en mi caso: `C:\ProgramData\MySQL\MySQL Server 9.2\Uploads\`). Si el archivo no está ahí, o si se intenta con el método "habitual", directamente no carga todo como debería.

La solución fue:

1. **Copiar mi CSV a esa carpeta segura.**
2. Usar un comando llamado `LOAD DATA INFILE` indicando claramente que los campos están separados por comas, que las cadenas van entre comillas y que las líneas acaban en `\r\n` (como en Windows).
3. Verificar con consultas (`COUNT(*), DISTINCT`) que realmente aparecían los 54 212 episodios, 118 podcasts, 6 países y 3 idiomas.

En resumen: el error no era del CSV, sino de la forma de importarlo. El Wizard sirve para bases de datos pequeñas o sencillas, pero para datasets grandes en MySQL 9.2 tuvimos que usar la carpeta segura y el comando `LOAD DATA INFILE`. Ahí sí se cargó todo perfecto.

En resumen, MySQL fue nuestro "control de calidad" y también un atajo para sacar conclusiones rápidas sin tener que abrir el dashboard.

Conclusiones

Entre 2019 y 2024, la producción de podcasts creció de forma sostenida hasta alcanzar su máximo histórico durante la pandemia de 2020. A partir de 2021, la actividad descendió y no todos los mercados recuperaron sus niveles previos: mientras Estados Unidos y Reino Unido mantuvieron un ritmo elevado, otros países redujeron su producción.

Los episodios más largos identificados como *outliers* reflejan la existencia de formatos narrativos especiales que trascienden el estándar breve, lo que evidencia la diversidad del medio.

Nuestro proceso basado en la extracción automática de datos desde la API de Spotify, limpieza en Python, validación cruzada en MySQL y visualización en Power BI, ha demostrado ser repetible, escalable y fiable, garantizando la consistencia entre las diferentes capas de análisis.

Para ampliar y enriquecer este estudio, proponemos tres líneas de mejora:

1. Incorporar métricas de consumo (audiencia, reproducciones y descargas) para conectar oferta y demanda.

2. Aplicar análisis de texto (NLP) sobre títulos y categorías para identificar intereses temáticos y patrones de contenido.

3. Automatizar todo el flujo de datos, permitiendo la actualización en tiempo real del dashboard y una visión continua y actualizada del panorama global de los podcasts.

Agradecimientos

Queremos expresar nuestro agradecimiento a los **profesores de la academia Unicorn Academy**, por su guía y exigencia académica durante todas las etapas del proyecto. Agradecemos especialmente el apoyo y la colaboración de nuestros compañeros de clase, cuya dedicación y creatividad han sido fuente constante de inspiración.

Finalmente, reconocemos el trabajo y compromiso de cada integrante del grupo: **Agustin Felix Albornoz Bruschetti, Raquel Lorenzo Fernández y Ramyi Gacela Gossen López** cuya cooperación y esfuerzo conjunto hicieron posible la realización y el éxito de este proyecto.