Proyecto 2: Preguntas de Negocio - Spotify 2023

Equipo:

Hemmy Luz Torres Ariza

Anallely Tenorio Sanchez

17 de septiembre de 2025

Laboratoria+

Ficha Técnica: Proyecto 2 Preguntas de Negocio

Objetivo

1. Ayudar al sello discográfico y al nuevo artista a tomar decisiones informadas para aumentar sus posibilidades de éxito respondiendo a las siguientes preguntas de negocio:
    - ¿Las canciones con mayor BPM (beats por minuto) tienen más reproducciones en Spotify?
    - ¿Los artistas con más canciones disponibles tienen más reproducciones?

Equipo

Hemmy Luz Torres Ariza y Anallely Tenorio Sanchez

Herramientas y tecnologías

1. Big Query
2. Looker Studio
3. Google Docs
4. Trello
5. Chat GPT
6. Recursos de la plataforma (videos youtube, Loom,etc)

Procesamiento y análisis:

A continuación se presenta el procesamiento y análisis de la información de acuerdo a las siguientes actividades:

I. Procesar y preparar la base de datos

5.1.1 Conectar/importar datos

- Cada una creó su proyecto en BigQuery y cargó las 3 tablas originales:
  - track_in_spotify
  - track_in_competition
  - track technical info

5.1.2 Identificar y manejar valores nulos

- track technical info: 95 nulos en key → reemplazados por "DESCONOCIDO".
- track_in_competition: 50 nulos en in_shazam_charts → reemplazados por 0.
- track_in_spotify: sin nulos.
- Se crearon tablas **step1** con los nulos tratados.

5.1.3 Identificar y manejar duplicados

- Revisión por track_id: no había duplicados.
- Revisión por (track_name, artist_s__name): 4 duplicados detectados en track_in_spotify.
- Criterio aplicado:
  - Mantener el ID con **mayor número de streams** (o más antiguo en caso de empate).
  - Consolidar con **máximos en playlists/charts**.
- Se crearon tablas **step1_track_in_spotify** y **step2_track_in_competition** con duplicados depurados.

5.1.4 Identificar y tratar valores atípicos en variables categóricas

- track_name y artist_s__name:
  - Se eliminaron espacios extra con TRIM.
  - Se normalizaron apóstrofes (', `, ´ → ').
  - Se limpiaron caracteres no válidos con REGEXP_REPLACE.

- Se aplicó INITCAP para estandarizar mayúsculas/minúsculas.

- Casos vacíos tras limpieza → reemplazados por "DESCONOCIDO".

- key y mode en track_tecnical_info:
  - key normalizado a mayúsculas y notación musical (A–G, #, b), otros → "DESCONOCIDO".
  - mode normalizado a "MAJOR" / "MINOR", nulos → "DESCONOCIDO".

5.1.5 Identificar y tratar valores atípicos en variables numéricas

- Se usó MIN/MAX/AVG con UNPIVOT en las 3 tablas.
- Reglas de plausibilidad:
  - bpm dentro de 40–220 → válido (65–206 en dataset).
  - Porcentajes 0–100 → válidos.
  - artist_count ≥ 1 → cumplido (1–8).
  - released_year 1930–2023 → plausible (dataset incluye canciones históricas).
  - playlists/charts ≥ 0 → cumplido.
- streams: detectado un valor no numérico → marcado y tratado en 5.1.6.

5.1.6 Verificar y cambiar tipo de datos

- Se aplicó SAFE_CAST a streams → INT64.
- El único registro no numérico quedó como NULL (documentado).
- Se creó tabla **step2_track_in_spotify** con streams en formato numérico.

5.1.7 Crear nuevas variables

- **released_date**: creada con SAFE.PARSE_DATE('%Y-%m-%d', FORMAT(...)).
  - Fechas inválidas → NULL (ninguna en este caso).
- **total_playlists**: suma de participación en playlists de las 3 plataformas:
  - in_spotify_playlists + in_apple_playlists + in_deezer_playlists.
  - Distribución: min=34, mediana=2304, max=62 623, promedio≈5669.
- Se guardaron como tablas nuevas (step5_released_date, step5_total_playlists) para mantener modularidad.

5.1.8 Unir tablas

- Utilizando el comando LEFT JOIN se generó la tabla “resumen_final_spotify” con los datos limpios de cada tabla. Asimismo se generó una tabla temporal y

auxiliar utilizando WITH para calcular el total de canciones por artista solista (583).

II. Realizar un análisis exploratorio

5.2.1 Agrupar datos según variables categóricas

- **Acción:** Se agruparon los datos por artista y por año de lanzamiento para identificar patrones de producción musical y popularidad.
- **Hallazgos:**
  - **Artistas con más canciones:** Taylor Swift lidera con 34 tracks, seguida de The Weeknd (21) y Bad Bunny (19).
  - **Canciones por año:** 2022 fue el año más prolífico con 399 canciones lanzadas.
  - **Artistas por streams:** The Weeknd y Taylor Swift superan los 14 mil millones de streams.
- **Interpretación:** Algunos artistas dominan tanto en volumen de producción como en consumo, lo cual refleja su consolidación en el mercado.

5.2.2 Visualizar variables categóricas

- **Acción:** Se construyeron gráficos de barras para mostrar las agrupaciones anteriores.
- **Hallazgos:** Los gráficos refuerzan que Taylor Swift no solo es la artista con más canciones, sino que también aparece entre los más escuchados, lo que valida su influencia.
- **Interpretación:** Visualizar por categorías ayuda a confirmar las diferencias entre artistas y años, y facilita la identificación de líderes en la industria.

5.2.3 Aplicar medidas de tendencia central

- **Acción:** Se calcularon media y mediana de streams y de playlists.
- **Hallazgos:**
  - **Streams:** Media = 514M, Mediana = 288M → la media es mucho mayor.
  - **Playlists:** Media ≈ 5.7K, Mediana ≈ 2.3K.
- **Interpretación:** La diferencia entre media y mediana confirma distribuciones sesgadas a la derecha: pocas canciones muy exitosas elevan el promedio.

5.2.4 Ver distribución

- **Acción:** Se usaron boxplots e histogramas para streams y playlists.

- **Hallazgos:**
  - La mayoría de canciones tiene menos de 500M streams, pero algunas superan los 3B.
  - En playlists, la mayoría aparece entre 1K–5K listas, pero hay casos extremos con más de 50K.
- **Interpretación:** Ambas variables presentan distribuciones asimétricas con valores atípicos (outliers) que concentran el éxito.

5.2.5 Aplicar medidas de dispersión

- **Acción:** Se calcularon desviación estándar, varianza y coeficiente de variación (CV).
- **Hallazgos:**
  - **Streams:** Desv. estándar ≈ 568M y CV ≈ 110% → gran heterogeneidad.
  - **Playlists:** Desv. estándar ≈ 8.9K y CV ≈ 157% → aún más dispersión.
- **Interpretación:** Los valores confirman que la mayoría de canciones tienen bajo rendimiento, mientras unas pocas concentran gran visibilidad.

5.2.6 Visualizar el comportamiento a lo largo del tiempo

- **Acción:** Se graficó la evolución de streams y cantidad de canciones por año.
- **Hallazgos:**
  - Streams se disparan a partir de 2020, con un pico en 2022.
  - El número de canciones lanzadas también crece desde los 2000, alcanzando un máximo en 2022 (~400).
- **Interpretación:** La pandemia y la consolidación de plataformas de streaming impulsaron tanto la producción musical como el consumo digital.

5.2.7 Calcular correlación entre variables

- **Acción:** Se calcularon correlaciones de Pearson entre streams y playlists, y entre streams y bailabilidad.
- **Hallazgos:**
  - **Streams vs Playlists:** $r = 0.78$, $r^2 = 0.61$ → relación positiva fuerte.
  - **Streams vs Bailabilidad:** $r = -0.1$, $r^2 = 0.011$ → relación negativa muy débil.
- **Interpretación:**

- Cuantas más playlists incluyen una canción, más streams acumula (playlist = visibilidad).
- La bailabilidad no explica el éxito de las canciones.

III. Aplicar la técnica de análisis

5.3.1 Responder a las preguntas de negocio

Objetivo:

Validar las preguntas de negocio planteadas mediante correlación y diagrama de dispersión.

¿Canciones con mayor BPM tienen más streams?

Análisis:

- La correlación entre BPM y streams fue **r ≈ -0.00023**, lo que representa una relación negativa prácticamente nula.
- El coeficiente de determinación fue **r² ≈ 0.000053%**, indicando que los BPM explican una fracción ínfima de la variabilidad en streams.
- Los boxplots muestran que las canciones con BPM moderados (101–120) tienden a tener **medianas de streams más altas**, pero no existe un patrón general fuerte.

Conclusión:

No se puede afirmar que un mayor BPM garantice más streams. El tempo por sí solo **no es un factor determinante del éxito** de una canción. Sin embargo, los datos sugieren que los tempos moderados (101–120 BPM) pueden asociarse con un mejor desempeño típico.

¿Aparecer en más listas implica más streams?

Análisis:

- La correlación entre playlists y streams fue de **r ≈ 0.78**, lo que indica una **relación positiva fuerte**.
- El coeficiente de determinación fue **r² ≈ 0.61**, lo cual significa que aproximadamente el **61% de la variabilidad en streams se explica por el número de playlists**.
- Los gráficos de dispersión muestran una tendencia clara ascendente, con algunos **outliers** (grandes éxitos con streams extraordinarios).

Conclusión:
Existe evidencia de que **aparecer en más playlists está fuertemente asociado con acumular más streams**. Aunque no implica causalidad directa, la magnitud del coeficiente r sugiere que la visibilidad en playlists es un factor clave en el desempeño de una canción.

IV. Resumir información en un tablero o informe

5.4.1 Representar datos a través de tablas de resumen o tarjeta de resultados

- Se crearon **scorecards** para mostrar los KPIs generales del dataset:
  - Total de canciones: **949**
  - Total de artistas: **645**
  - Promedio de streams: **514,3 M**
  - Promedio de playlists: **5,7 mil**

  Estas tarjetas permiten dar un panorama rápido y general del conjunto de datos.

5.4.2 Representar datos mediante gráficos simples

- Se incluyeron **gráficos de barras** para mostrar:
  - **Top artistas con más canciones**
  - **Top artistas por streams**
  - **Top artistas en playlists**

    Esto facilita la comparación y ranking de los artistas más relevantes.

5.4.3 Representar datos utilizando gráficos o funciones visuales avanzadas

- Se usaron **diagramas de dispersión** para responder preguntas de negocio:
  - Relación entre **BPM y streams**
  - Relación entre **playlists y streams**

    Con estos gráficos se identificaron correlaciones y se validaron hipótesis.

5.4.4 Aplicar opciones de filtro para la gestión y la interacción

- Se añadieron filtros dinámicos por:
  - **Año de lanzamiento (released_year)**
  - **Artista (artist_s__name)**

    Periodo de tiempo

    Esto permite a los usuarios explorar el dashboard según distintos criterios y profundizar en los datos.

V. Presentar resultados

Creamos una presentación colaborativa en Looker Studio con los siguientes apartados:

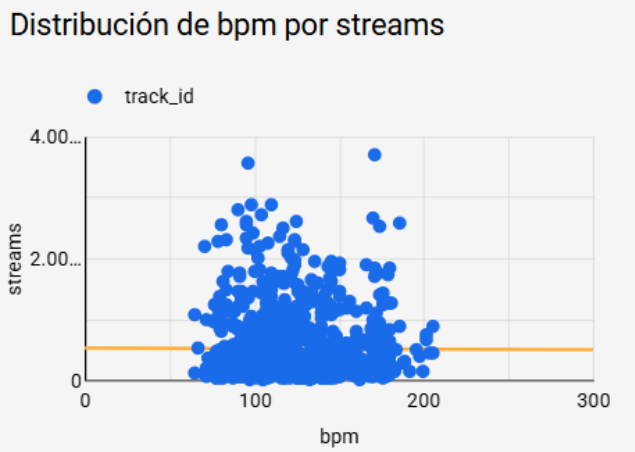
1. KPIs generales (total canciones, artistas, streams, playlists).
2. TOP artistas y evolución de canciones en el tiempo.
3. Características técnicas promedio de las canciones.
4. Pregunta de negocio 1: ¿Mayor BPM implica más streams?
5. Pregunta de negocio 2: ¿Aparecer en más playlists implica más streams?

Resultados y conclusiones:

I. Resultados

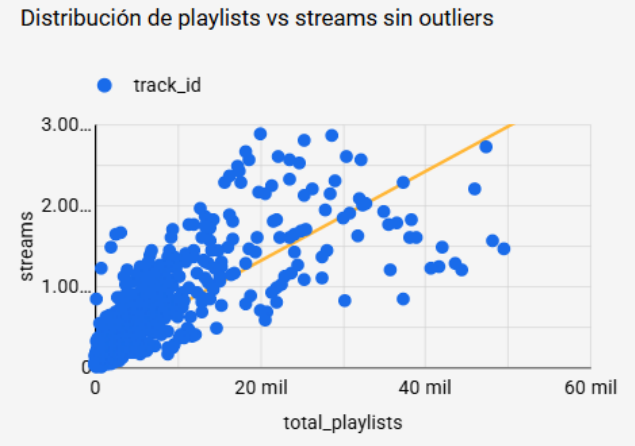
- **¿Canciones con mayor BPM tienen más streams?**

  La correlación estimada entre BPM y streams es r= -0.0002. Asimismo, se generó el sgte diagrama de dispersión:



- **¿Aparecer en más listas implica más streams?**

  La correlación estimada entre total playlists y streams es r= 0.78. Asimismo, se generó el sgte diagrama de dispersión:



II. Conclusiones

- **¿Canciones con mayor BPM tienen más streams?**

  De acuerdo con el valor de la correlación estimada entre BPM y streams (r= -0.0002), la relación entre ambas variables es negativa pero muy débil (prácticamente nula). Lo cual se evidencia en la gráfica de dispersión, ya que no se aprecia un patrón definido.

  Por lo tanto, no es posible afirmar que un mayor BPM se asocie a un incremento en los streams, por lo que esta variable, por sí sola, no resulta un predictor del éxito de una canción.

  Sin embargo, los datos sugieren que los tempos moderados (101–120 BPM) se asocian con un mejor desempeño típico (mediana más alta de streams).

- **¿Aparecer en más listas implica más streams?**

  En este caso, el valor de la correlación entre el total playlists y streams es r= 0.78, lo cual sugiere que existe una relación positiva fuerte entre ambas variables (cercano a 1). Asimismo, la gráfica de dispersión muestra una tendencia ascendente clara, con algunos outliers (éxitos extraordinarios) que refuerzan la asimetría observada en la distribución.

  En consecuencia, puede afirmarse que las canciones que aparecen en un mayor número de playlists tienden a registrar más reproducciones. Si bien la correlación no implica causalidad, la magnitud de r sugiere que la

inclusión en playlists constituye un factor clave asociado al desempeño en términos de streams.

Finalmente, los datos sugieren que el éxito de una canción está fuertemente asociado con la visibilidad en playlist, más que con sus características técnicas como el BPM.

Limitaciones / Próximos pasos:

I. Limitaciones

- Al inicio del proyecto se presentaron dificultades en el uso de la herramienta BigQuery, dado que era la primera vez que la utilizábamos. Asimismo, durante el desarrollo de las primeras actividades no contábamos con total claridad sobre la estructuración del proyecto, ya que algunas tareas se abordaban únicamente a nivel de consulta.
- Otro factor limitante fue la disponibilidad de tiempo, ya que situaciones externas, como problemas de salud y compromisos laborales, afectaron el ritmo de avance previsto.

II. Próximos pasos

- Profundizar en el uso de la herramienta, mediante práctica constante y consultas más frecuentes a los coaches, con el fin de recibir orientación y retroalimentación oportuna.
- Organizar mejor la gestión del tiempo, distribuyendo de manera equilibrada las horas de estudio y trabajo, y dejando un margen para imprevistos, con el objetivo de cumplir con el cronograma establecido.

Enlaces de interés:

1. Looker Studio: Proyecto 02 - Spotify 2023
    a. 5.2 Realizar un análisis exploratorio
    b. 5.3 Aplicar la técnica de análisis

c. Dashboard
d. Presentación - Spotify 2023

https://lookerstudio.google.com/reporting/2404a7fa-511a-4456-abbc-2d78719c743b/page/pFFXF

2. Big Query - Hemmy Torres Ariza

3. Big Query - Anallely Tenorio Sanchez