Ficha Técnica: Proyecto de Análisis de Datos

Título del Proyecto: Hipótesis

Objetivo:

Validar o refutar las hipótesis planteadas por la discográfica sobre qué hace que una canción sea más escuchada.

1. Las canciones con un mayor BPM (Beats Por Minuto) tienen más éxito en términos de cantidad de streams en Spotify.
2. Las canciones más populares en el ranking de Spotify también tienen un comportamiento similar en otras plataformas como Deezer.
3. La presencia de una canción en un mayor número de playlists se relaciona con un mayor número de streams.
4. Los artistas con un mayor número de canciones en Spotify tienen más streams.
5. Las características de la música influyen en el éxito en términos de cantidad de streams en Spotify.

Equipo:
Trabajo en dupla.

Herramientas y Tecnologías:

- BigQuery
- Power BI
- Google Docs
- Google Slide

Lenguajes

- SQL en BigQuery
- Python en Google Colab

Insumos

Dataset

Diccionario de datos

Trackinspotify

- track_id: Identificador único de la canción. Es un número entero de 7 dígitos que no se repite
- track_name: Nombre de la canción
- artist(s)_name: Nombre del artista(s) de la canción

- artist_count: Número de artistas que contribuyen a la canción.
- released_year: Año en que se lanzó la canción.
- released_month: Mes en el que se lanzó la canción.
- released_day: Día del mes en que se lanzó la canción.
- in*spotify*playlists: Número de listas de reproducción de Spotify en las que está incluida la canción
- in*spotify*charts: Presencia y ranking de la canción en las listas de Spotify
- streams: Número total de transmisiones en Spotify. Representa la cantidad de veces que la canción fue escuchada.
- track_name_limpio: Nombre de la canción sin caracteres especiales.
- artist_s__name_limpio: Nombre del artista sin caracteres especiales.
- streams_numero: Número total de transmisiones en Spotify. Representa la cantidad de veces que la canción fue escuchada. Los valores son del tipo INTEGER.
- released: fecha de lanzamiento de la canción en formato year, month, day.
- total_part_playlist: suma del número de listas de reproducción Spotify, Deezer y Apple en las que está incluida la canción.

Trackincompetition

- track_id: Identificador único de la canción. Es un número entero de 7 dígitos que no se repite
- in*apple*playlists: número de listas de reproducción de Apple Music en las que está incluida la canción
- in*apple*charts: Presencia y rango de la canción en las listas de Apple Music
- in*deezer*playlists: Número de listas de reproducción de Deezer en las que está incluida la canción
- in*deezer*charts: Presencia y rango de la canción en las listas de Deezer
- in*shazam*charts: Presencia y rango de la canción en las listas de Shazam

Track*technical*info

- track_id: Identificador único de la canción. Es un número entero de 7 dígitos que no se repite
- bpm: Pulsaciones por minuto, una medida del tiempo de la canción.
- key: Clave musical de la canción
- mode: Modo de la canción (mayor o menor)
- danceability_%: Porcentaje que indica qué tan adecuada es la canción para bailar
- valence_: Positividad del contenido musical de la canción.
- energy_: Nivel de energía percibido de la canción.
- acusticness_: Cantidad de sonido acústico en la canción.
- instrumentality_: Cantidad de contenido instrumental en la canción.
- liveness_: Presencia de elementos de actuación en vivo.
- speechiness_: Cantidad de palabras habladas en la canción.

Procesamiento y análisis

Conectar/importar datos a herramientas

- Se creó el proyecto2-hipotesis-lab y el conjunto de datos Dataset en BigQuery.
- Tablas importadas: track_in_competition, track_in_spotify, track_technical_info

Identificar y manejar valores nulos
Se identifican valores nulos a través de comandos SQL COUNT, WHERE y IS NULL.

- **track_in_competition:** 50 valores nulos en la columna **in_shazam_charts.**
- **track_in_spotify:** 0 nulos.
- **track_technical_info:** 95 valores nulos en la columna **key.**

Identificar y manejar valores duplicados
Se identifican duplicados a través de comandos SQL COUNT, GROUP BY, HAVING.

- **track_in_competition:** no hay valores duplicados**.**
- **track_in_spotify:** 4 track_name duplicadas con diferentes track_id. Para el análisis de datos no se consideró la información de estos track_name porque no hay certeza de cuáles datos son correctos, al ser 4 canciones no impacta en el resultado. La muestra total contiene 952 datos y se consideraron solo 944.
- **track_technical_info**: no hay valores duplicados.

Identificar y manejar datos fuera del alcance del análisis
Se manejan variables que no son útiles para el análisis a través de comandos SQL SELECT EXCEPT.

- **track_tecnical_info:** se excluyó la columna key por tener muchos datos nulos (95) y la columna mode por no tener información relevante para el análisis.

Identificar y manejar datos discrepantes en variables categóricas
Se identifican datos discrepantes utilizando el comandos de manejo de string, como REGEXP.

- **track_in_spotify:** Se reemplazaron por espacios vacíos los caracteres especiales de los track_name y artist_s__name , se creó una nueva columna track_name_limpio y artist_s__name_limpio.

Identificar y manejar datos discrepantes en variables numéricas

- Se identifican datos discrepantes utilizando comandos como MAX, MIN y AVG para las variables numéricas de interés para el estudio de cada base de datos.

Comprobar y cambiar tipo de dato

- **track_in_spotify:** Conversión de la variable streams de STRING a INTEGER usando comando SAFECAST, AS INT64 creando una nueva variable streams_numero, se calculan los valores MAX, MIN y AVG.

Crear nuevas variables

- **track_in_spotify:** Se creó la variable released, utilizando CONCAT.

Unir tablas

- Se crearon vistas de las tablas con los datos limpios, view_competition_limpia, view_technical_info_limpia y view_spotify_limpia
- Unión de las tablas limpias usando LEFT JOIN.
- Se creó la variable total_part_playlist usando SUM.

Construir tablas auxiliares

- Se creó tabla temporal para calcular el total de canciones por artista solista usando WITH.

Análisis exploratorio

Agrupar datos según variables categóricas

- Se importan los datos de BigQuery a Power BI.
- Se crearon tablas matrix con la cantidad de tracks por artista, cantidad de tracks por released_year y la cantidad de streams por año.

Visualizar las variables categóricas

- Se crearon gráficas de barras para la visualización de variables categóricas en Power BI.

Aplicar medidas de tendencia central y de dispersión

- Usando Matrix en Power BI se calcularon las medidas de tendencia central y dispersión para las variables bpm, streams spotify, playlist spotify, deezer, apple y total de participación playlists.

Visualizar distribución

- Utilizando Python se crearon histogramas para las variables bpm, streams_numero, total_part_playlist.

Visualizar el comportamiento de los datos a lo largo del tiempo

- Se crearon gráficos de línea en Power BI para evaluar el comportamiento de la cantidad de tracks y streams a lo largo del tiempo.

Calcular cuartiles, deciles o percentiles

- Se crearon categorías por cuartiles para las variables de características en BigQuery utilizando WITH, NTILE, IF.
- Se crearon las categorías alto y bajo para cada característica.

Calcular correlación entre variables

- En BigQuery se calculó la correlación en entre variables para cada una de las hipótesis utilizando el comando CORR.

Prueba de significancia

Planteamiento de la hipótesis:

- **Hipótesis nula (H0)**: Las características de la música NO influyen en el éxito en términos de cantidad de streams en Spotify.

- La diferencia de streams promedio en la categoría danceability son estadísticamente iguales.

  Streams promedio categoría alto = Streams promedio categoría bajo

- **Hipótesis Alternativa (H1):** Las características de la música INFLUYEN en el éxito en términos de cantidad de streams en Spotify.

- La diferencia de streams promedio en la categoría danceability son significativamente diferentes.

  Streams promedio categoría alto != Streams promedio categoría bajo

Elegir nivel de confianza (alpha)

- Para evaluar la hipótesis, seleccionamos un nivel de confianza del 95% ($\alpha = 0.05$).

Elegir estadístico de contraste adecuado y calcular el pvalor

Utilizamos dos pruebas estadísticas:

- **Test t de Student**: prueba parámetrica.
- **Test Wilcoxon (Mann-Whitney U)**: prueba no paramétrica.

Comparar el pvalor con el de alpha y concluir si aceptamos o rechazamos la H0

Característica	Wilcoxon (valor p)	Test t (valor p)	Interpretación
Danceability	0.013703012841444151	0.001996301247206946	Rechazamos la hipótesis nula: hay una diferencia significativa entre las dos categorías.
Valence	0.054718654340041556	0.13227567705145718	No podemos rechazar la hipótesis nula: no hay una diferencia significativa entre las

			dos categorías.
Liveness	0.1855317077872108 4	0.4063199701536462	No podemos rechazar la hipótesis nula: no hay una diferencia significativa entre las dos categorías.
Instrumentals	0.1029737991829157 7	0.2765756566173294	No podemos rechazar la hipótesis nula: no hay una diferencia significativa entre las dos categorías.
Acoustiness	0.4127224742352993	0.6140564152170954	No podemos rechazar la hipótesis nula: no hay una diferencia significativa entre las dos categorías.
Energy	0.0504245951548924 76	0.602990459978884	No podemos rechazar la hipótesis nula: no hay una diferencia significativa entre las dos categorías.
Speechiness	0.0001633650326623 0928	0.0002478313166170461	Rechazamos la hipótesis nula: hay una diferencia significativa entre las dos categorías.
BPM	0.8228918138922666	0.7601119588500744	No podemos rechazar la hipótesis nula: no hay una diferencia significativa entre las dos categorías.

Recordando que la hipótesis no se cumple: Hay una diferencia significativa entre el valor de streams promedio para la categoría alto y bajo, podemos observar que para danceability estar en la categoría alto si influye en el número de streams, hay más streams.

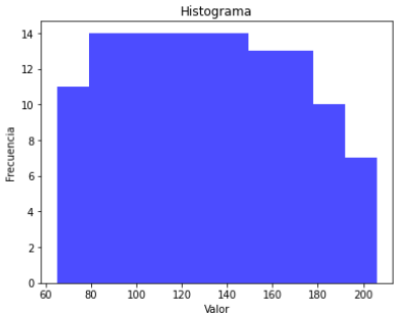
Visualizar métricas descriptivas de una variable continua

Se calcularon los valores de promedio, desviación estándar, min, max, cuartil 1, 3, 4 para cada categoría alto - bajo de cada característica.

Resultados y Conclusiones

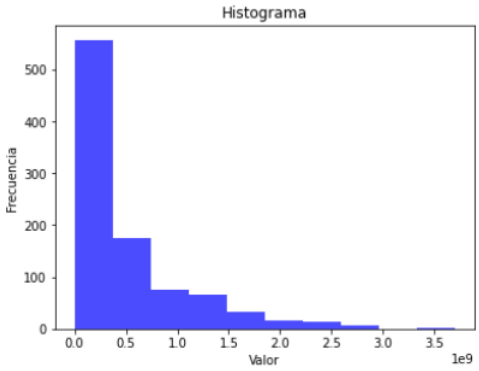
Análisis Exploratorio

BPM



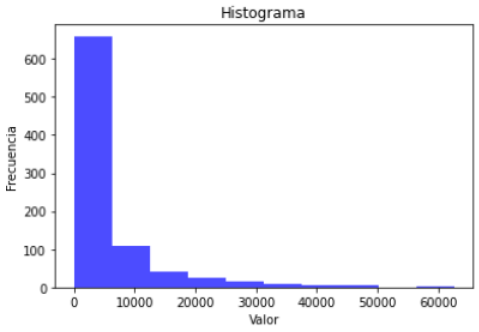
- Los valores de BPM más comunes se encuentran entre aproximadamente 85 y 145.
- Hay menos ocurrencias de valores de BPM por debajo de 80 y por encima de 180.
- La distribución parece simétrica, con una ligera tendencia hacia valores de BPM más altos.

STREAMS



- La mayoría de los valores de streams se encuentran en el rango de 0 a 500,000,000.
- Hay una frecuencia muy alta de streams en el rango de 0 a 500,000,000 mil millones, con más de 500 ocurrencias.
- Los valores de streams mayores 1,000,000,000 mil millones son mucho menos comunes.
- La distribución es asimétrica y está sesgada hacia la izquierda, con una larga cola hacia la derecha.

Total de participación playlist



- La mayor frecuencia de datos está entre 0 y 5000 mil playlists.
- La distribución de los datos es asimétrica y está sesgada a la izquierda.
- Hay muy pocos datos con valores altos de participación en playlist.

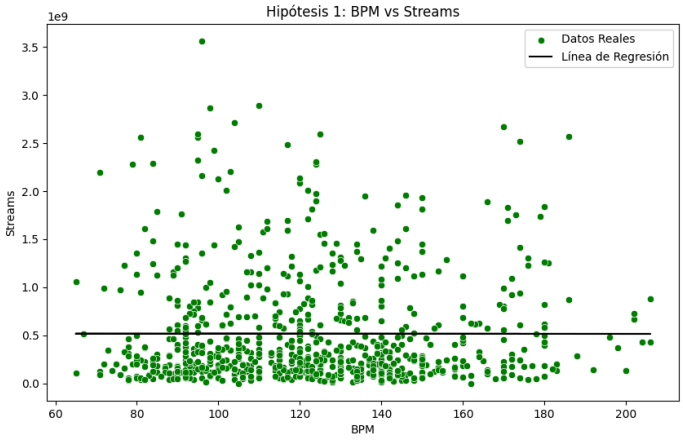
Correlación entre variables

Hipótesis 1: Las canciones con un mayor BPM (Beats Por Minuto) tienen más éxito en términos de cantidad de streams en Spotify.
Correlación
r= -0.00320018576

Interpretación: el valor de r es muy cercana a 0, lo que indica que no hay una relación lineal significativa entre las dos variables. El signo negativo indica que, si existiera alguna relación, sería una relación negativa, donde un aumento en una variable estaría asociado con una disminución en la otra. Sin embargo, dado que la magnitud es tan pequeña, esta relación negativa es prácticamente insignificante.

Regresión lineal
1.Error cuadrático medio (MSE): 3.1890603295274906e+17
2.Coeficiente de determinación (R^2): -0.0003296256420366461
3. Intercepción: 518079372.14366764
4. Coeficiente: -14448.855151167847

Interpretación:
1.El MSE mide la magnitud promedio de los errores cuadrados entre las predicciones del modelo y los valores reales. Un MSE tan grande sugiere que las predicciones están muy lejos de los valores reales, lo que indica un mal ajuste del modelo.
2.Un R^2 negativo indica que el modelo no tiene poder predictivo.
3.Intercepción: Este es el valor de streams cuando el bpm es 0. Sin embargo, como los valores de bpm no pueden ser 0, esta intersección no tiene un significado práctico directo pero es parte del cálculo del modelo.
4.Coeficiente: Este coeficiente indica que por cada incremento unitario en bpm, el "streams" disminuye en promedio en aproximadamente 14448.85 unidades. Dado que esto no parece razonable y el R^2 es negativo, esto también sugiere que el modelo no es adecuado.
Conclusión: Se refuta la hipótesis inicial de que las canciones con un mayor bpm tienen más streams.



Hipótesis 2: Las canciones más populares en el ranking de Spotify tienen comportamiento similar en otras plataformas como Deezer, Apple, Shazam.

Correlación

1. Spotify vs Deezer: r= 0.6076780201308

2. Spotify vs Apple: r= 0.55269053270411

3. Spotify vs Schazam: r= 0.6055409034998

Interpretación:

- Los valores de correlación de 0.6077 (Spotify vs Deezer), 0.5527 (Spotify vs Apple Music), y 0.6055 (Spotify vs Shazam) indican relaciones positivas moderadas entre Spotify y cada una de las otras plataformas.
- Estos valores sugieren que las tendencias de popularidad de música en Spotify tienden a reflejarse también en Deezer, Apple Music, y Shazam.
- Ninguna de estas correlaciones es extremadamente alta (cercana a 1), lo que indica que, aunque hay una relación, no es perfecta. Esto es esperable ya que cada plataforma puede tener su propia base de usuarios con preferencias ligeramente diferentes.

Regresión lineal

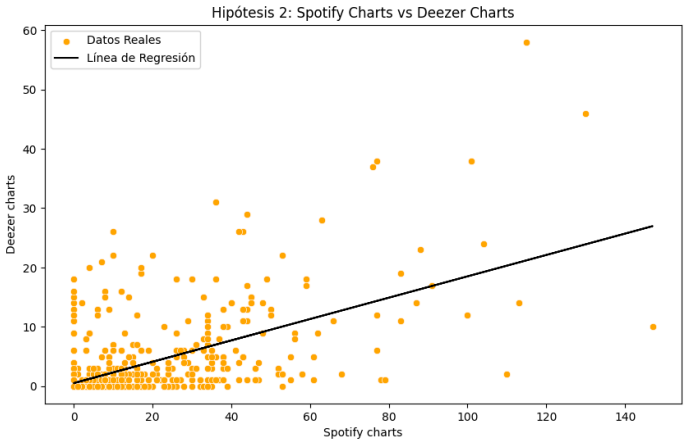
1. **Spotify vs Deezer**

Error cuadrático medio (MSE): 21.5543438226604
Coeficiente de determinación (R^2): 0.4004457933944051
Intercepción: 0.5037406320271827
Coeficiente: 0.18002131752879133

Interpretación:

El MSE de 21.55 sugiere que aunque el modelo hace un trabajo razonable prediciendo charts, hay un margen considerable de error. Sugiere que hay variabilidad en los datos que no está siendo capturada.

El R^2 de 0.40 indica que hay una relación moderada entre in_spotify_charts y in_deezer_charts, pero no es lo suficientemente fuerte como para ser el único predictor de charts.
El coeficiente positivo de 0.18 confirma que, en general, una mayor posición en spotify está asociado con una mayor posición en deezer.
Conclusión: Se valida la hipótesis inicial.

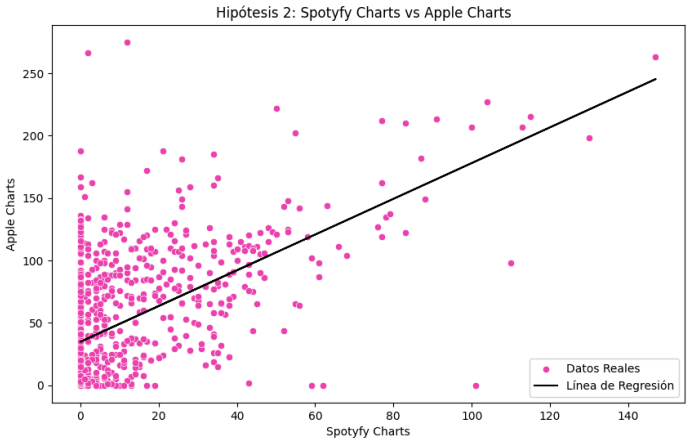


Regresión lineal

2. **Spotify vs Apple**

Error cuadrático medio (MSE): 1976.8643296982843
Coeficiente de determinación (R^2): 0.24968577987854879
Intercepción: 34.845478791842766
Coeficiente: 1.4310541598256816

Interpretación:
El MSE de 1976.86 es un valor alto lo que indica que no hay un buen ajuste del modelo a los datos.
Un R^2 de 0.2497 indica que aproximadamente el 24.97% de los datos se pueden predecir por el modelo de regresión lineal.
Intercepción cuando la variable independiente es 0, el modelo predice que la variable dependiente será aproximadamente 34.85.
Coeficiente, sugiere una relación positiva, una mayor posición en spotify está asociado con una mayor posición en deezer.
Conclusión: Se válida la hipótesis inicial, pero la relación es débil como para ser un buen predictor por sí solo.



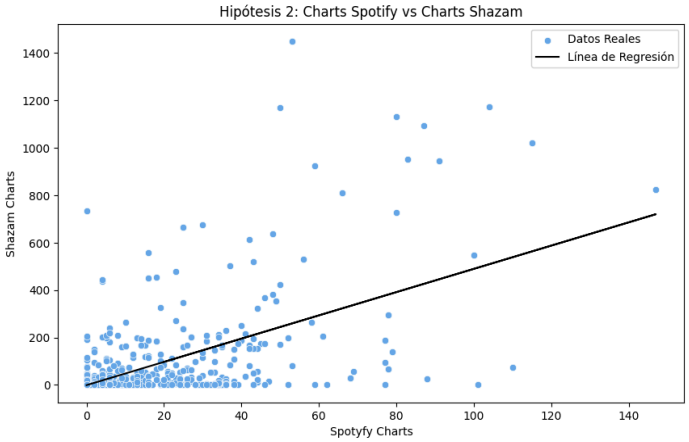
Regresión lineal

3. **Spotify vs Shazam**

Error cuadrático medio (MSE): 14735.855116718994
Coeficiente de determinación (R^2): 0.44691000277096626
Intercepción: -0.8709355811230566
Coeficiente: 4.905546289495083

Interpretación:

El MSE de 14735.86 sugiere que aunque el modelo hace un trabajo razonable prediciendo "charts in Spotify", hay un margen considerable de error.
El R^2 de 0.45 indica que hay una relación moderada, lo que sugiere que charts in Spotify es un predictor razonable del éxito en los charts de Shazam, pero hay otros factores importantes que también influyen.
El coeficiente positivo de 4.91 confirma que, en general, un mejor desempeño en los charts in spotify está asociado con un mejor desempeño en los charts in Shazam
Conclusión: Se válida la hipótesis inicial.

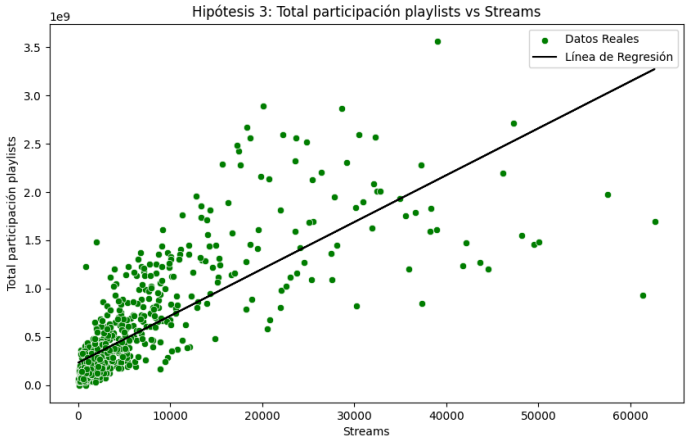


Hipótesis 3: La presencia de una canción con un mayor número de playlists se relaciona con un mayor número de streams.

Correlación
r=0.7836803010789
Interpretación:
Un valor de r indica una fuerte relación positiva. Esto sugiere que hay una relación considerablemente fuerte entre el número de playlists en las que aparece una canción y el número de streams que recibe.

Regresión lineal
Error cuadrático medio (MSE): 1.275849895855914e+17
Coeficiente de determinación (R^2): 0.5997973331266275
Intercepción: 231242214.05101973
Coeficiente: 48560.98322485796

Interpretación:
Los resultados sugieren que hay una relación positiva entre el número de playlists en las que aparece una canción y el número de streams que recibe. Aunque el r2 de 0.6 indica que esta relación explica una parte significativa de la variación en los streams, el MSE elevado muestra que hay otros factores no incluidos en el modelo que también influyen en el número de streams. Aún así, el coeficiente positivo y significativo refuerza la idea de que estar en más playlists se asocia con más streams
Conclusión: válida la hipótesis inicial.



Hipótesis 4: Los artistas con un mayor número de canciones en Spotify tienen más streams.

Correlación
r=0.80016684593280

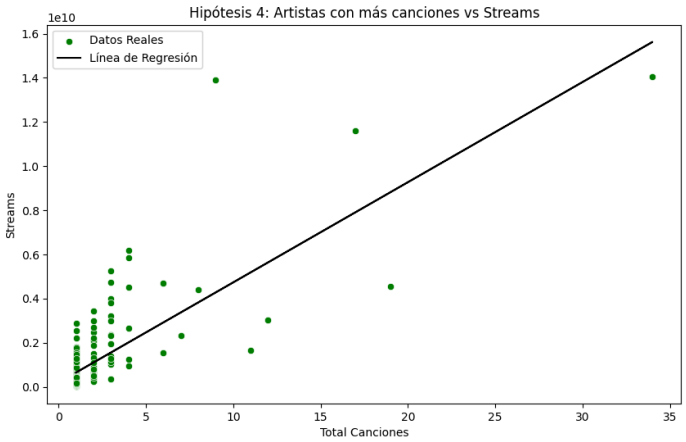
Interpretación:
El valor de r indica una correlación positiva fuerte por lo que es muy probable que los artistas con más canciones en Spotify tiendan a tener un mayor número de streams. Lo que sugiere

que la cantidad de contenido que un artista tiene en la plataforma está estrechamente vinculada con su popularidad en términos de streams.

Regresión lineal
Error cuadrático medio (MSE): 1.377236569956644e+18
Coeficiente de determinación (R^2): 0.7537531924491054
Intercepción: 194858748.11261463
Coeficiente: 453769865.76784766

Interpretación:
Los resultados sugieren una relación positiva y fuerte entre el número de canciones de un artista en Spotify y el número de streams que recibe. Aunque el MSE elevado indica que hay otros factores que también influyen en el número de streams y no están incluidos en el modelo, el r2 de 0.75 muestra que una gran parte de la variación en los streams puede ser explicada por el número de canciones. El coeficiente positivo y significativo refuerza la hipótesis de que los artistas con más canciones tienden a tener más streams.

Conclusión: Se válida la hipótesis inicial.



Hipótesis 5: Las características de la música influyen en el éxito en términos de cantidad de streams en spotify.

Correlación
Danceability: r= -0.10563589955055
Speechiness: r= -0.112773935150

Interpretación:
Los valores de r indican una relación negativa muy débil. La relación inversa es mínima, A medida que aumenta danceability y speechiness aumentan, la otra variable tiende a disminuir ligeramente.

Correlación:

Valence: r=-0.041797954869
Energy: r= -0.0257381767548
Acousticness: r=-0.00498576864
Instrumentalness:r=-0.0440399854154
Liveness:r=-0.051147025245

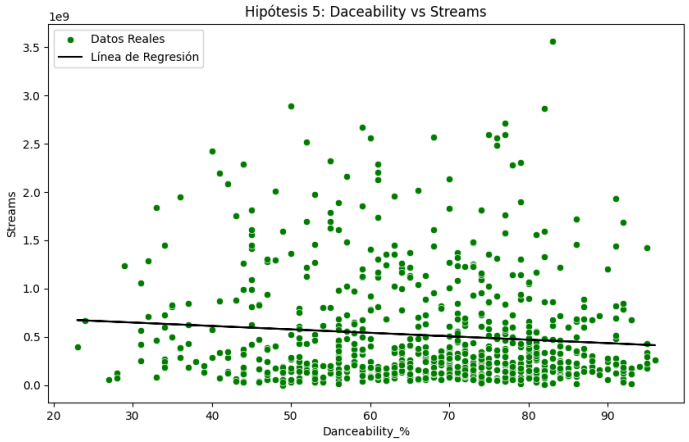
Interpretación:
Para todas las variables:

- La dirección de la relación es inversa (negativa) en todos los casos..
- La fuerza de la relación es extremadamente débil en todos los casos, con valores muy cercanos a 0, prácticamente no hay una relación lineal observable.
- No hay una relación significativa entre Valence, Energy, Acousticness, Instrument, y Liveness vs streams.

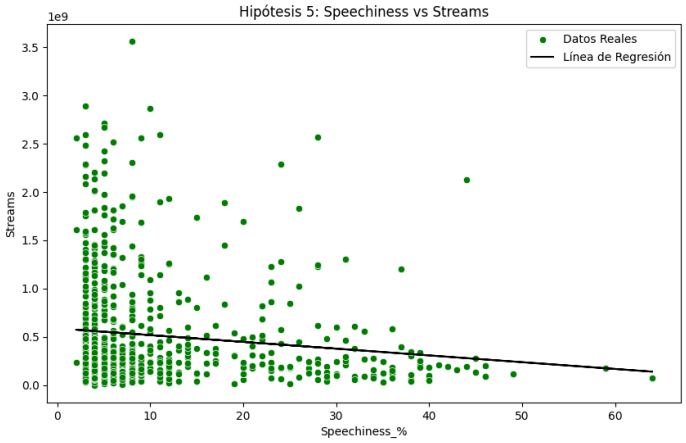
Interpretación regresión lineal para todas las características:

La hipótesis de que las características de la música influyen en el éxito en términos de cantidad de streams en Spotify no se sostiene. El modelo muestra que tienen un impacto mínimo y, de hecho, parecen tener una relación negativa con la cantidad de streams. Además, la alta magnitud del MSE y el bajo R^2 indican que el modelo no es adecuado para predecir el éxito en términos de streams basándose en las características.

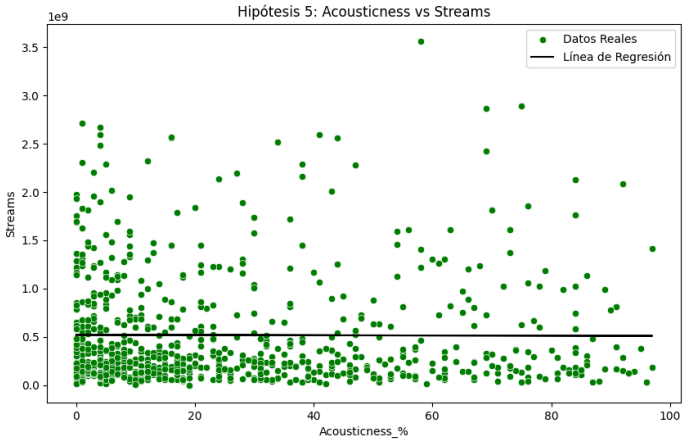
Regresión lineal Danceability:
Error cuadrático medio (MSE): 3.124828603648244e+17
Coeficiente de determinación (R^2): 0.01981828366786642
Intercepción: 754197652.8166625
Coeficiente: -3561913.4172532973



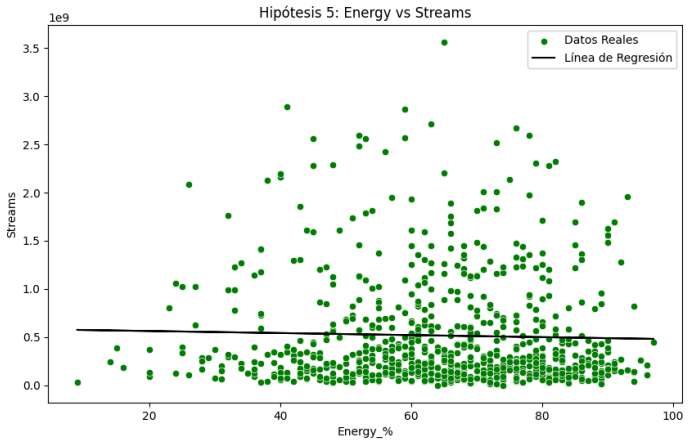
Regresión lineal Speechiness:
Error cuadrático medio (MSE): 3.1839496713280416e+17
Coeficiente de determinación (R^2): 0.0012734618744032478
Intercepción: 586839521.0596712
Coeficiente: -7004805.666278839



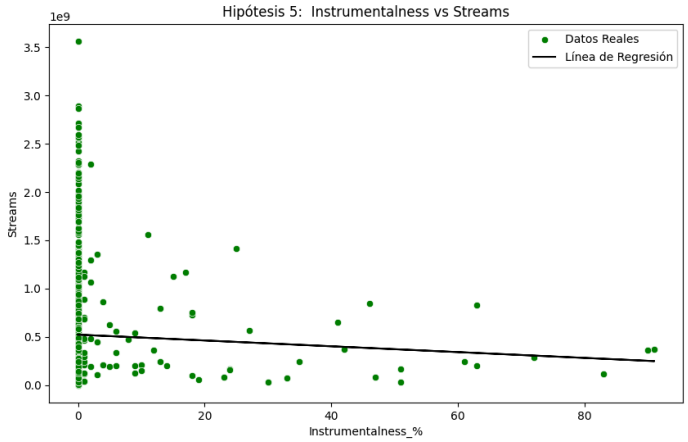
Regresión lineal Acousticness:
Error cuadrático medio (MSE): 3.1889904031590746e+17
Coeficiente de determinación (R^2): -0.000307691463716786
Intercepción: 519052870.0753346
Coeficiente: -101639.24942977371



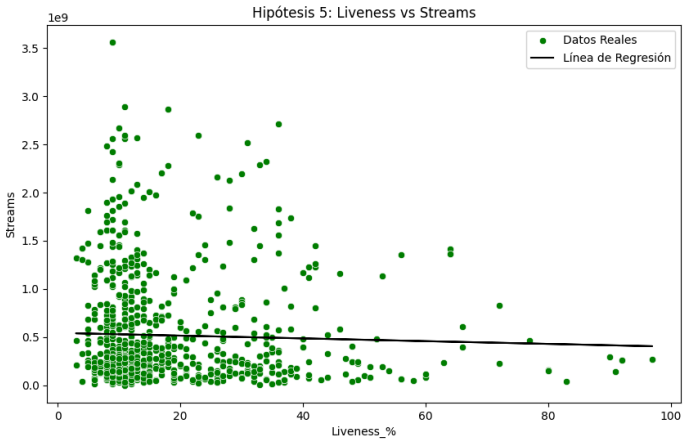
Regresión lineal Energy:
Error cuadrático medio (MSE): 3.191199432101503e+17
Coeficiente de determinación (R^2): -0.0010006093977401598
Intercepción: 584343722.8663057
Coeficiente: -1060671.6164008593



Regresión lineal Instrumentalness:
Error cuadrático medio (MSE): 3.186896882065557e+17
Coeficiente de determinación (R^2): 0.0003489945049873766
Intercepción: 521574809.3676341
Coeficiente: -3009595.266891558



Regresión lineal Liveness:
Error cuadrático medio (MSE): 3.167159734862173e+17
Coeficiente de determinación (R^2): 0.006540051127653657
Intercepción: 542243879.0230006
Coeficiente: -1424267.01610165



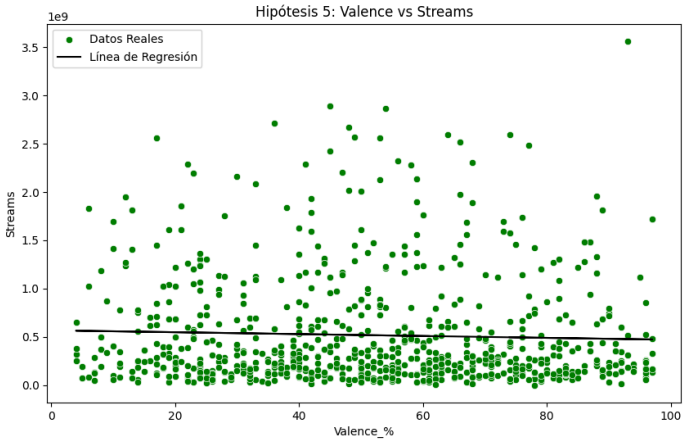
Regresión lineal Valence:

Error cuadrático medio (MSE): 3.182035009291737e+17

Coeficiente de determinación (R^2): 0.00187404416514092

Intercepción: 566138289.9384496

Coeficiente: -971945.7250246599



Recomendaciones:
Después del análisis y revisión de las hipótesis para que el lanzamiento de un artista sea exitoso y que sus canciones sean las más escuchadas se puede recomendar lo siguiente:

- Estar disponible en todas las plataformas musicales, teniendo enfoque principal en Spotify, ya que el comportamiento de está plataforma influye en las demás.
- Es relevante estar presente en un mayor número de playlists y en los rankings musicales para aumentar la visibilidad de la canción.
- Aumentar rápidamente la cantidad de canciones que se lanzan al mercado, ya que hay una correlación positiva muy fuerte entre la cantidad de canciones y los streams, es decir, a mayor número de canciones mayor número de streams.
- No hay una relación clara en que alguna de las características de la canción pueda influir en su éxito.
- Hacer campañas de marketing y colaboración con los artistas más populares del momento.

Limitaciones
Falta de información de las otras plataformas.
Desconocimiento del lenguaje de programación.