ANÁLISIS DE DATOS SPOTIFY

MEMORIA FINAL

PROYECTO II

Grado en Ciencia de Datos

2023 - 2024

Elena Navarro, Javier

Amores Giner, Pau

Ciobanu Borinschi, Luminita

Suleimanov Ismail Ogly, Rustam

etsinf

ÍNDICE

1. INTRODUCCIÓN AL PROYECTO

En la era digital, el consumo de música ha evolucionado significativamente, impulsado por plataformas de streaming que utilizan algoritmos sofisticados para recomendar contenido a los usuarios. Este proyecto tiene como objetivo profundizar en la comprensión de diversos aspectos del ecosistema musical actual a través del análisis de datos históricos y recientes. Se han definido cuatro objetivos específicos para este estudio, cada uno dirigido por un investigador principal con una línea de investigación particular.

2. OBJETIVOS ESPECÍFICOS

Descubrir los Factores Clave en las Recomendaciones de Contenido Musical

Comprender qué elementos se consideran al recomendar contenido al usuario y por qué. Desvelar el funcionamiento del sistema de recomendaciones de canciones, artistas, álbumes y géneros para así identificar y agrupar los gustos musicales de los usuarios.

Identificación de Tendencias Temporales en la Popularidad Musical

Analizar la evolución de la popularidad de los géneros musicales a lo largo del tiempo, utilizando las características de las canciones más populares. Este estudio permitirá observar y comprender las tendencias temporales en el ámbito musical.

Segmentación de Usuarios según Preferencias Regionales

Investigar las preferencias musicales de los usuarios en función de su región y el crecimiento de artistas dentro de géneros específicos. Esta segmentación ayudará a comprender mejor las dinámicas regionales en las preferencias musicales.

Predicción del éxito de la canción un artista

Realizar predicciones sobre el éxito de las canciones de los artistas basándose en datos históricos y evaluar la precisión de estas predicciones comparándolas con datos actuales. Este análisis permitirá identificar los factores que contribuyen al éxito de nuevos tracks en la industria musical.

3. DATOS EMPLEADOS

☒ *Music Recommendation System Using Spotify Data.*

Este primer fichero dispone de datos hasta 2021 y está enfocado a analizar el algoritmo de recomendación de Spotify. Está formado por distintos data sets: uno orientado a artistas, otro a canciones y otro a géneros.

Está en formato CSV. Cada fila es una canción con 19 variables entre las que destacan el nombre e identificador de la canción, y el artista que interpreta la canción, como variables categóricas. En cuanto a variables numéricas, contamos con el año de lanzamiento de la canción, una variable de cada género musical que mide de 0 a 1 la relación que tiene la canción con cada género. También disponemos de otras tablas que están relacionadas con la tabla principal mediante claves ajenas, y analizan cualidades de los artistas, géneros musicales, datos por año, etc.

https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset/input

☒ *Spotify Dataset 2023.*

Más enfocado en datos recientes, por tanto complementa muy bien el primer data set. Además tiene un enfoque parecido, ya que está compuesto por varios datasets que contienen datos de artistas, canciones y álbumes.

En este data set en formato CSV, el cual es muy similar al de "*Music Recommendation System Using Spotify Data" cada fila también es una canción/álbum o artista que actúan como variables categóricas. Cabe destacar que existen ciertas variables de interés como los followers de cada artista y hay más filas que en los data sets restantes, aparte de que es el dataset más reciente que existe de Spotify (tiene datos hasta 2023, en comparación con otros cuyos datos son hasta 2021).*

https://www.kaggle.com/datasets/tonygordonjr/spotify-dataset 2023?select=spotify_artist_data_2023.csv

☒ *Top Spotify Songs in 73 Countries:*

Por último utilizaremos este data set para encontrar el lugar de donde han sido elegidas las naciones en el top en cada uno de los países que se actualizará diariamente.

Este data set está en formato CSV y contiene las mismas variables que el mencionado anteriormente excepto algunas, como el top en el que está una canción, el álbum al que pertenece y el país en el que ha sido top 50 canciones. Esto nos permitirá realizar un análisis geográfico de las canciones de Spotify.

https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated

El cruce de datos que se realizará de los tres data sets distintos estará enfocado en recopilar la información más actual de los álbumes y artistas mediante la concatenación de BD de "*Music Recommendation System Using Spotify Dataset" y la de "Spotify Dataset 2023".* Por último, el conjunto de dos datasets concatenados lo concatenamos con uno más que es el de "*Top Spotify Songs in 73 Countries (Daily Updated)"* que es el que nos permitirá sacar la información sobre el top de canciones que han salido en la lista de las canciones más populares/mejor valoradas de cada país.

4. DESCUBRIR LOS FACTORES CLAVE EN LAS RECOMENDACIONES DE CONTENIDO MUSICAL

→ Introducción

En el mundo del streaming musical, las recomendaciones personalizadas son esenciales para mejorar la experiencia del usuario. Este estudio tiene como propósito descubrir los factores que influyen en las recomendaciones de contenido, como canciones, artistas, álbumes y géneros. Al desvelar el funcionamiento del sistema de recomendaciones, buscamos identificar patrones y agrupaciones de gustos musicales, proporcionando una visión profunda de cómo se personalizan las recomendaciones para cada usuario.

→ Metodología

Para lograr este objetivo, nos hemos centrado en las variables que Spotify utiliza para diferenciar las canciones: 'acousticness', 'danceability', 'energy', 'instrumentalness', 'liveness', 'loudness', 'speechiness' y 'valence'. Estas variables son cruciales para definir las características de las canciones. Hemos utilizado un modelo de regresión logística para evaluar la relevancia de estas características en la agrupación de canciones por género, y a partir de ello análisis en cuanto a la distribución de los géneros. Además, hemos realizado unos algoritmos simulando la recomendación de spotify para artistas y canciones y, por otro lado, hemos elaborado un PCA para obtener conclusiones sobre estas variables.

→ Resultados

El modelo de regresión trata de inferir el género de una canción basándose en las variables mencionadas. Para este análisis, se utilizó el conjunto de datos hasta 2020, seleccionando los 20 géneros más frecuentes (considerando solo el primer género listado para cada canción y su frecuencia en el dataset).

A partir de las canciones de esos géneros y sus valores en las variables, hemos entrenado el modelo de regresión logística.

Lo hemos diseñado para que asigne dos géneros (los más adecuados) a cada canción y la precisión a la primera (el primer género de los 2) es de un 38%, mientras que si tenemos en cuenta el segundo género asignado la precisión asciende al 60%.

Esto nos lleva a deducir que las variables definen bastante bien las diferencias entre canciones y géneros, ya que un 60% de acierto para 20 posibilidades distintas es bastante alto. Las variables de género obtenidas nos servirán para analizar por géneros, ya que en el dataset de canciones en el top no había datos de género y en los otros dos datasets (hasta 2020 y hasta 2023) había muchos valores faltantes. Aplicamos el modelo a todos los datasets para poder analizarlos en cuanto al género (ya que todos comparten las variables utilizadas para el modelo).

Las variables, al definir correctamente las diferencias entre canciones, también lo hacen entre artistas, lo que podría ser eficaz para recomendar artistas y canciones similares. Hemos creado un recomendador de artistas y otro de canciones que sugiere los más parecidos basándose en las diferencias mínimas entre las variables de sus canciones, simulando recomendaciones similares a las de Spotify.

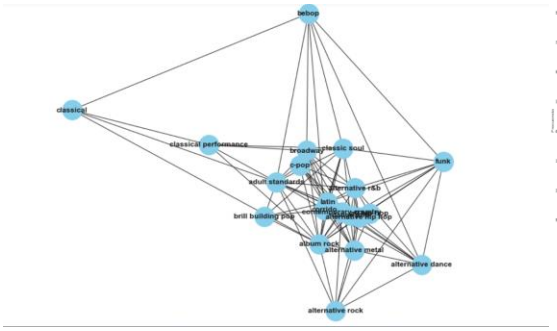
```
artistas_similares = sugerir("Plan B")
print(artistas_similares)
```

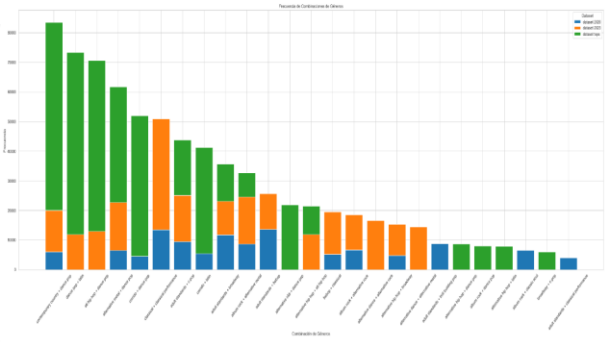
```
['Fundisha', 'Pitbull', 'Three 6 Mafia', 'Yandel', 'Sean Paul', 'J Balvin', 'Nicole Scherzinger', 'China Anne M
O', 'Jory Boy']
```

Lógicamente, sólo funciona para artistas o canciones que están dentro de nuestro dataframe . En este ejemplo, hemos solicitado para Plan B (reggeatonero latino), y nos recomienda a Pitbull, Yandel, J Balvin o Kevvo (también reggaetoneros latinos), lo cual es bastante convincente.

Por otro lado, como hemos asignado dos variables de género a cada canción (primera opción y segunda opción), hemos analizado qué géneros tienen más probabilidad de estar en una misma canción con otro género.

En primer lugar, hemos realizado un grafo, en el cual los nodos son los géneros musicales que hemos designado, y que exista una arista entre dos nodos significa que existe una canción en la que los dos están presentes (en nuestra estimación).

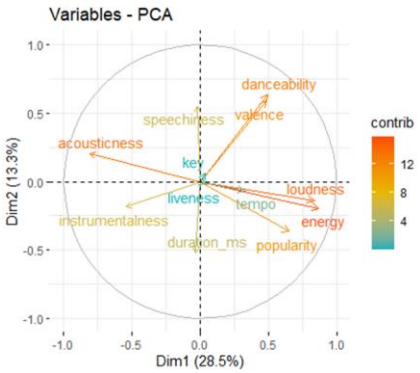


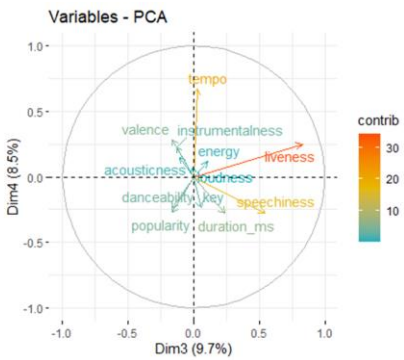


Esto nos reporta una idea de los géneros que no tienen nada que ver entre sí, y por tanto no sería intuitivo recomendar ese tipo de música si no está relacionada con la del oyente. Como podemos ver en el grafo, la mayoría de géneros sí que están relacionados, pero algunos como música clásica, funk, bebop o rock están relacionados con menos géneros.

Además, hemos sacado un gráfico de barras para saber cuales son las combinaciones de género más comunes, teniendo en cuenta las ocurrencias en las predicciones de los 3 datasets. Vemos que destacan las combinaciones country + pop, pop + latin, hip hop + dance pop, metal + pop y corrido + pop. Estos serían los géneros que se podrían recomendar en función de los que el oyente escuche, ya que estarán más estrechamente relacionados.

Por último, hemos realizado un análisis PCA sobre las variables de interés 'acousticness', 'danceability', 'energy', 'instrumentalness', 'liveness', 'loudness', 'speechiness' y 'valence', para tratar de entender las relaciones entre ellas.





De este análisis PCA obtenemos las siguientes relaciones:

Positivamente Correlacionadas:

-Energy y Tempo, Danceability y Energy, Key y Mode

Negativamente Correlacionadas:

-Acousticness y Energy, Instrumentalness y Speechiness, Liveness y Acousticness

→ Conclusiones

-Las variables 'acousticness', 'danceability', 'energy', 'instrumentalness', 'liveness', 'loudness', 'speechiness' y 'valence' están relacionadas con el género de la canción y son una forma de clasificación y aproximación a los gustos del oyente

-A partir de estas variables podemos recomendar artistas y canciones parecidas, no sólo clasificar en cuanto al género las canciones

-Existen géneros más parecidos que otros, y, por tanto, podemos recomendar también géneros parecidos según el género que escucha el oyente

-Las relaciones del PCA nos sirven para entender las relaciones entre las variables y saber cuales son similares o contrarias, y poder prever valores en las variables a partir de otras.

5. IDENTIFICACIÓN DE TENDENCIAS TEMPORALES

La música refleja las tendencias culturales y sociales de cada época. Este estudio se centra en analizar cómo ha evolucionado la popularidad de los géneros musicales a lo largo del tiempo mediante el análisis de las características de las canciones más populares. El objetivo es observar y comprender las tendencias temporales en la industria musical, identificando cambios en las preferencias del público y los factores que han influido en estos cambios.

→ Metodología

Para realizar el análisis, se utilizó la herramienta Power BI, y el trabajo se dividió en tres partes:

1. Estudio de la evolución de las características de los géneros más relevantes.

2. Análisis de la influencia de la época del año en el consumo de diferentes tipos de música.

3. Análisis de la popularidad de los géneros musicales a lo largo del tiempo.

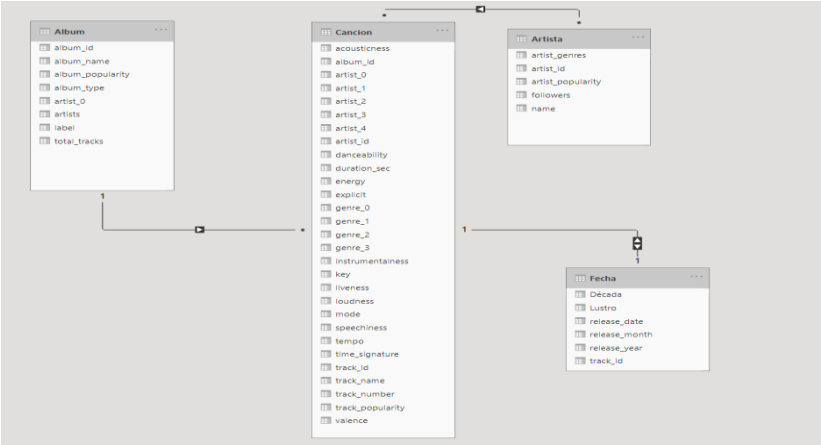
→ Preparación de Datos

Antes de comenzar el análisis, se construyó un esquema en Power Query para adaptar los datos originales y asegurar la corrección de los resultados. El modelo de datos resultante incluye:

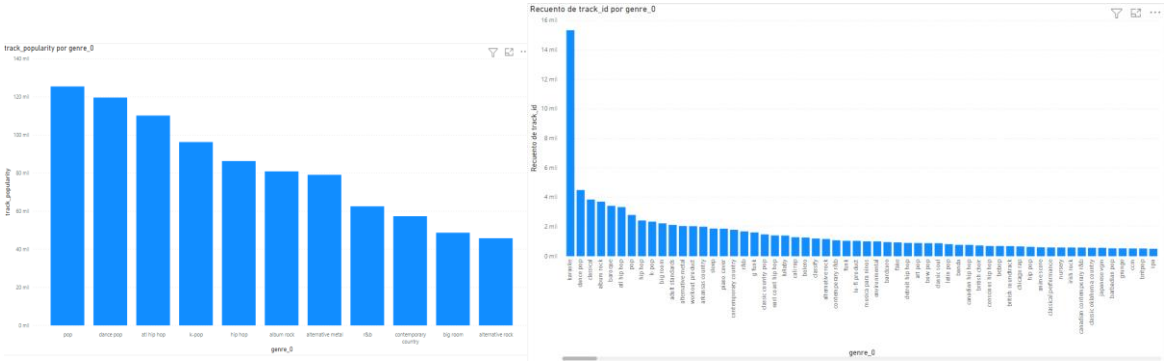
- Una tabla de hechos que representa las canciones.

- Tres dimensiones alrededor de esta tabla: álbumes, artistas y fechas.

Este modelo proporciona la estructura necesaria para realizar un análisis efectivo y obtener insights sobre la evolución de los géneros musicales.



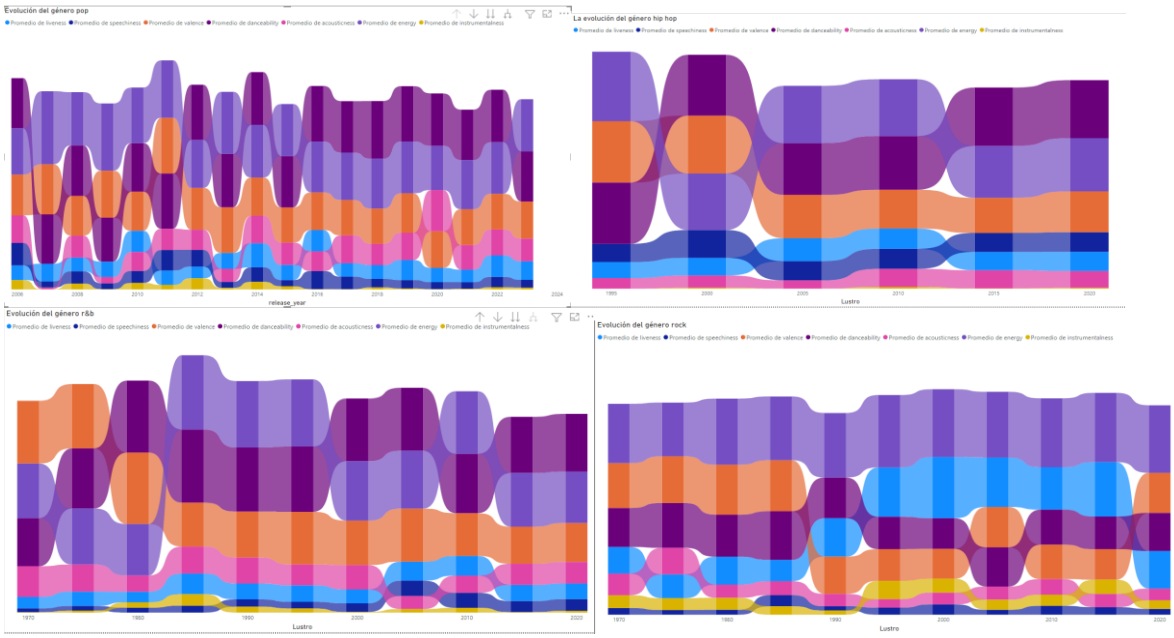
Una vez explicada la construcción del modelo vamos a determinar los géneros que vamos a estudiar. Para ello, nos centraremos en estos dos gráficos.

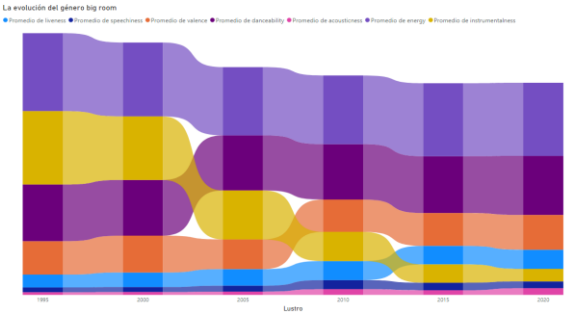


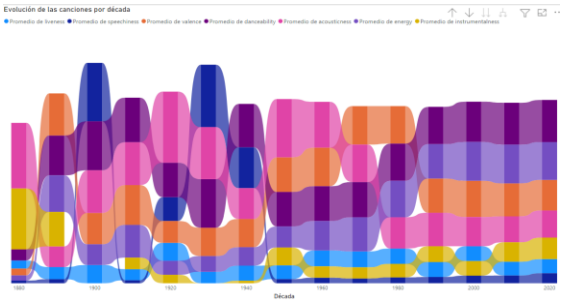
Los gráficos proporcionan información sobre el número de canciones por género en la base de datos y la media de popularidad de estas canciones. Basado en esta información, se seleccionaron los siguientes géneros para el análisis: pop, hip hop, big room, R&B y rock.

Para visualizar los datos, se consideró que los gráficos de barras eran los más adecuados para el objetivo del estudio. Estos gráficos representan el promedio de cada una de las características de las canciones a lo largo de los años, proporcionando una visión clara de cómo han evolucionado estos géneros musicales con el tiempo.

→ **Resultados**







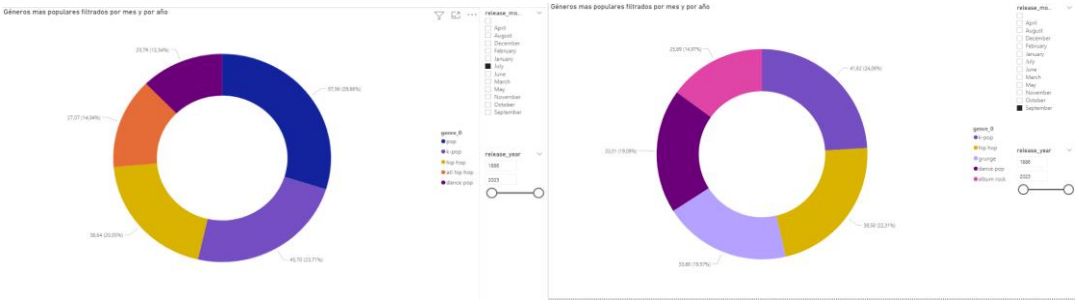
→ Conclusión 1

Como conclusiones de estos gráficos podemos extraer que el género pop, hip hop y R&B con el paso de los años tienen más canciones con un estilo "bailable" y menos canciones con un estilo "alegre". Por otro lado, el género rock muestra un aumento en la creación de canciones "en vivo" a medida que pasan los años desde su creación. También es importante destacar que el género big room experimenta un descenso considerable en la variable instrumentalness, lo que sugiere que con el tiempo este género tiene menos canciones puramente instrumentales.

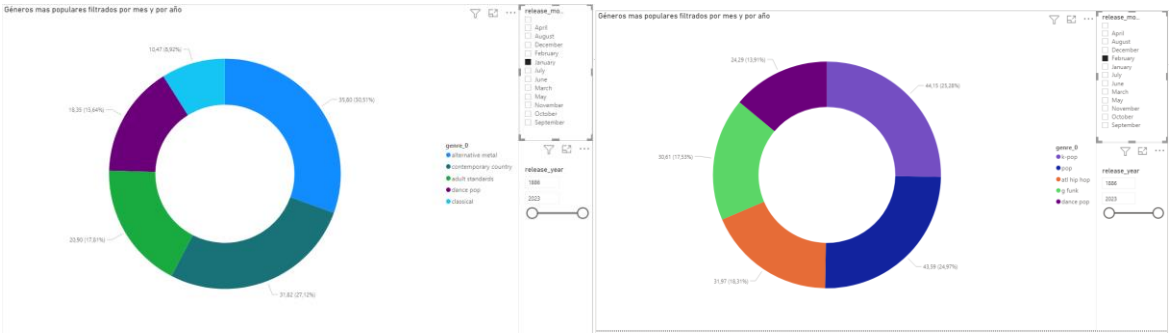
En la última gráfica, representamos la evolución de todas las canciones en conjunto. Aquí podemos observar dos picos de un estilo con una alta presencia de palabras en la canción en las décadas de 1900 y 1930. También notamos un descenso considerable en las canciones acústicas a partir de la década de 1960 y un aumento en las canciones enérgicas. A partir de la década de 1990, las canciones parecen estabilizarse siendo la mayor parte canciones "bailables".

En la segunda parte de este objetivo, analizaremos si la época del año influye en la popularidad de los géneros. Para ello, hemos creado un gráfico de anillos que muestra los géneros más populares con dos segmentadores de tiempo, que se segmentan por año de salida y por mes.

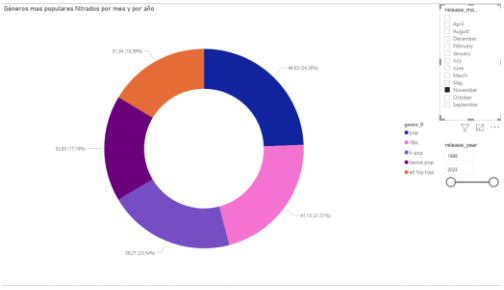
Verano



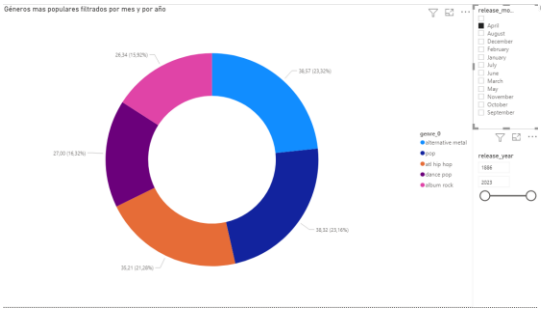
Invierno

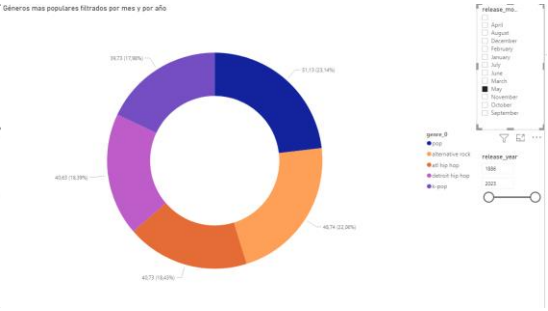


Otoño



Primavera



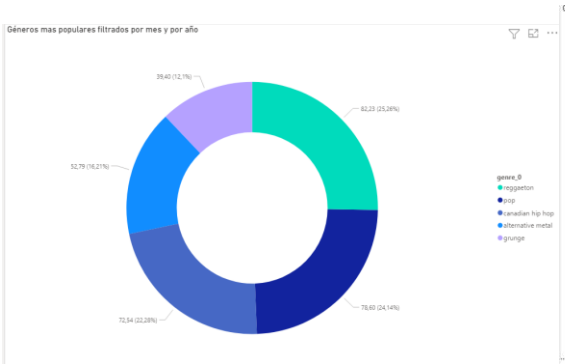


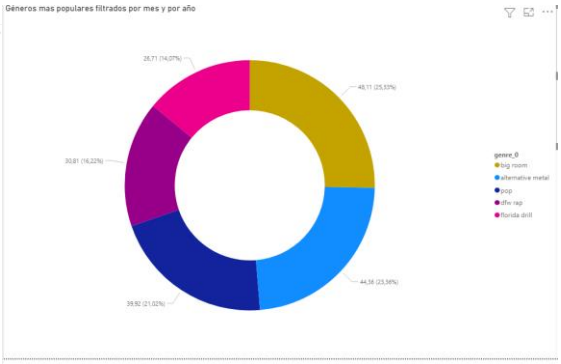
→Conclusión 2

Gracias a estos segmentadores hemos podido comprobar que en general en otoño es más popular el r&b, mientras que en primavera lo es más el metal y el rock. Además, en invierno es más popular el funk y el country, y en verano lo es el k-pop. Es importante destacar que el género pop y hip hop suele ser popular durante todo el año.

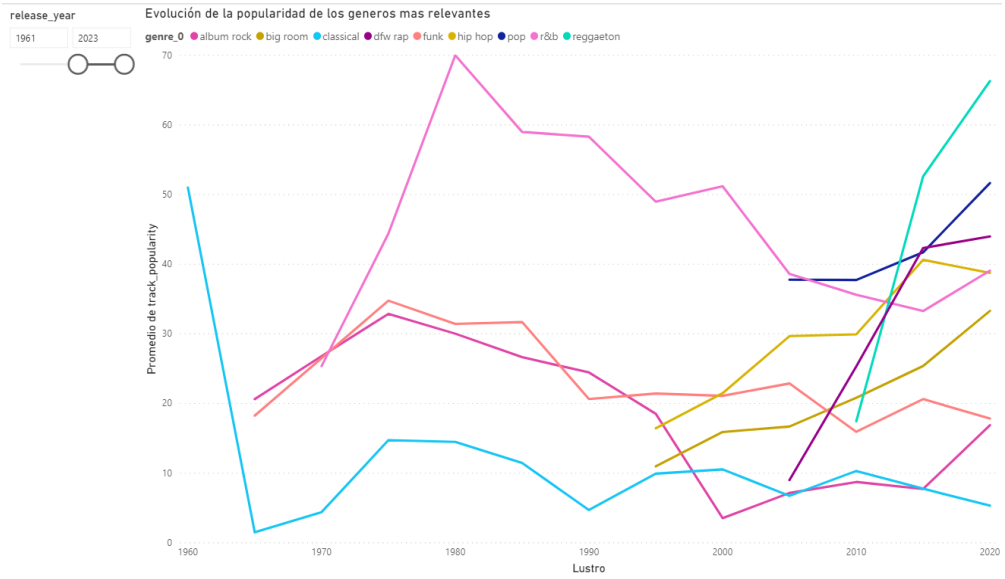
Si nos centramos únicamente en los últimos años, la situación cambia. Aparecen en el top géneros como el drill, el rap y el reggaeton y el big room. Es importante destacar que el pop y el hip hop siguen manteniéndose como unos de los géneros más populares.

A continuación, se muestran dos gráficos con ejemplos de estas explicaciones previas.





Por último, tenemos como último propósito ver la evolución de la popularidad de los géneros estudiados. Además de estos géneros estudiados vamos a implementar algunos géneros más recientes para comparar la popularidad entre estos géneros más nuevos con los más antiguos.



→ Conclusión 3

En este gráfico de líneas podemos observar el promedio de popularidad de cada género. En el eje x, cada valor está separado por un tiempo de un lustro. De aquí podemos extraer que la música clásica a partir de la década de los 60 e incluso de los 50 deja de ser el máximo exponente en cuanto a la popularidad de las canciones, siendo en la actualidad el género menos popular entre los 9 estudiados.

También se puede apreciar que el r&b alcanza un pico de popularidad promedio de 70 en la década de 1980, siendo el pico más alto del gráfico. A partir de esta década ha ido disminuyendo progresivamente hasta la actualidad. Asimismo, también podemos ver que los géneros más antiguos en este informe, que es el caso del funk y el rock, también decaen con el tiempo, siendo un poco más populares que el género de música clásica.

Por otro lado, los géneros más recientes son los que más popularidad han ganado con los últimos años, teniendo un crecimiento drástico a partir de la década de los 2000. Este es el caso del género el big room, hip hop, pop, reggaeton y el rap. Es importante destacar que el género que más popularidad ha ganado estos últimos años es el reggaeton, teniendo una popularidad media de 68 con sus canciones y, que el pop desde sus inicios, ya tenía una popularidad alta.

6. SEGMENTACIÓN DE USUARIOS SEGÚN PREFERENCIAS REGIONALES

→ Introducción

Las preferencias musicales varían significativamente según la región geográfica. Este estudio tiene como objetivo investigar cómo los usuarios de diferentes regiones del mundo prefieren distintos géneros musicales y analizar el crecimiento de artistas en estos géneros. Mediante la segmentación de usuarios por región, se busca comprender mejor las dinámicas regionales y ofrecer insights valiosos sobre cómo las preferencias locales impactan en el consumo musical.

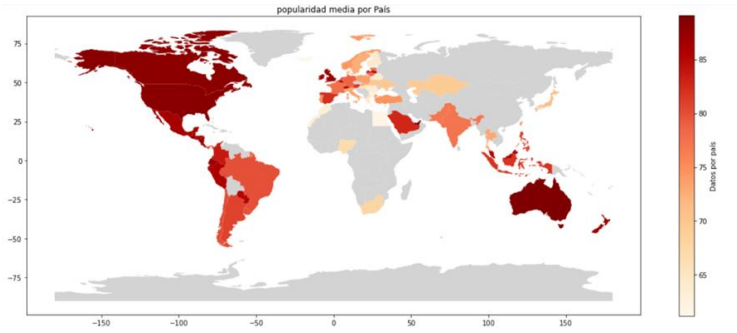
→ Metodología

Para realizar el análisis regional, se utilizó la librería `world` para visualizar la distribución de datos. La base de datos empleada contiene información sobre canciones que han aparecido en el top de Spotify en algún país. También se ha analizado el Top 50 diario por países usando R.

→ Resultados

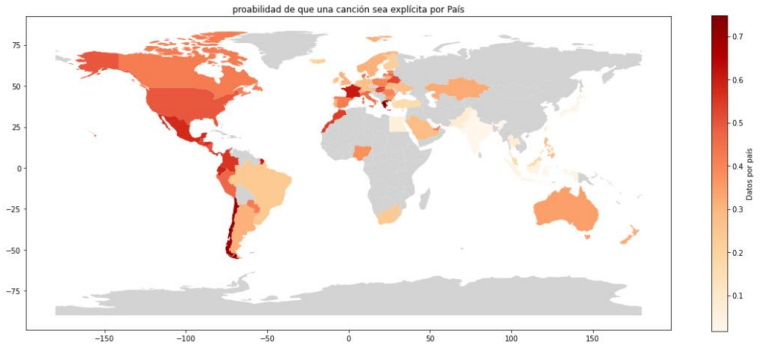
1. Popularidad Promedio de las Canciones por País:

- Se graficó la popularidad promedio de las canciones en cada país, lo que proporciona una idea de si la música escuchada en el país tiende a ser más conocida a nivel global.



2. Probabilidad de Canciones Explícitas por País:

- Se analizó la probabilidad de que una canción sea explícita en cada país, proporcionando una visión sobre el tipo de contenido musical que se prefiere en diferentes regiones.



3. Características de las Canciones por País:

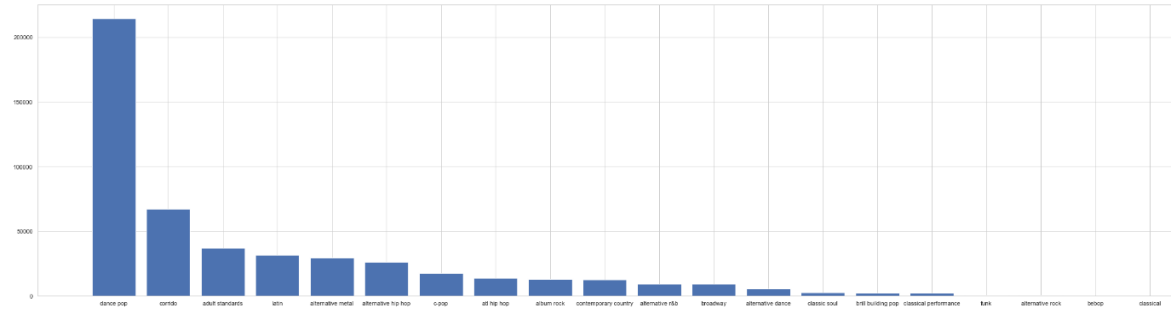
- Se realizaron gráficos para analizar diversas características de las canciones (como tempo, energía, duración, etc.) por cada país, permitiendo entender mejor las preferencias musicales regionales.

4. Análisis del Género Musical por País:

- Se predijo y graficó la frecuencia de diferentes géneros musicales en la base de datos.

- Se creó una distribución específica de géneros por país.

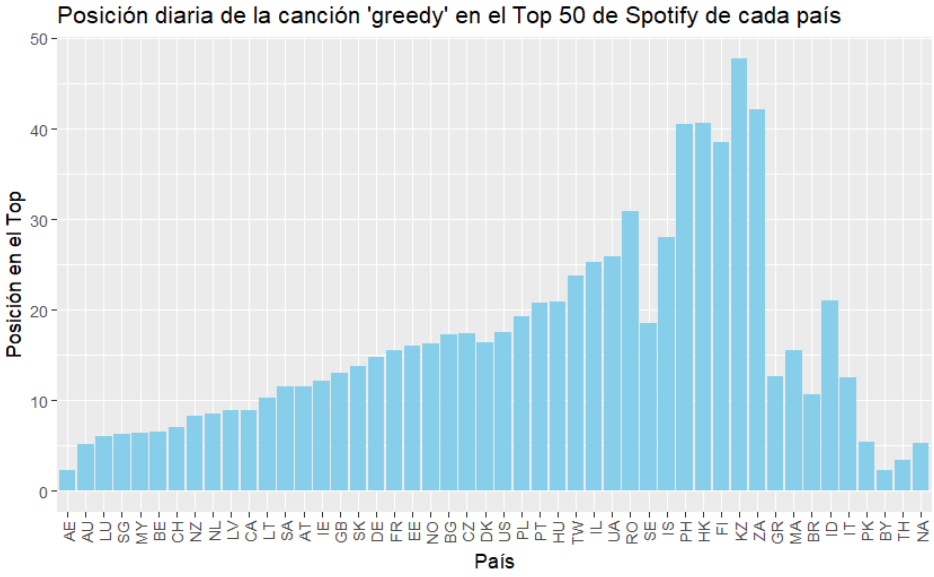
- Para mejorar la comparación entre países, se escalaron las variables dividiéndolas entre su media. Valores superiores a 1 indican más escuchas de un género en comparación con la media global.



5. Análisis Top 50 Diario

A partir del data set Top Spotify Songs in 73 Countries con fecha de 13 de Marzo de 2024, después de preparar los datos y llevar a cabo un análisis exploratorio inicial, se ha buscado información como:

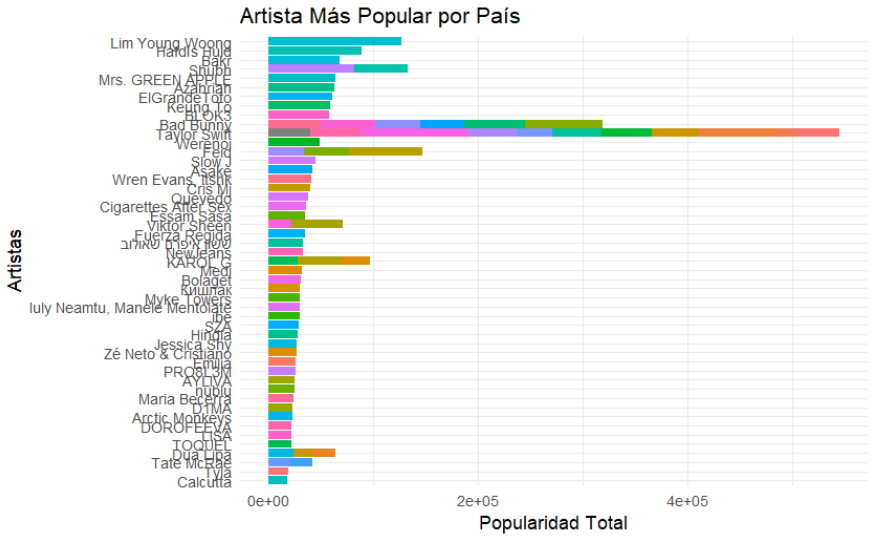
- **Canción que más aparece en todos los rankings y su posición en cada ranking:**



- **Canción que más veces aparece en el Top 1:**

##	name	num_times_top1
##	LUNA	565

- **Artista que más aparece en el top de cada país:**



→ Observaciones Destacadas

- Japón: Predominancia del Bebop.

- Bielorrusia, Corea y Kazajistán: Preferencia por el Rock.

- Finlandia, Islandia y Marruecos: Alta presencia de Música Clásica.

- Sudáfrica: Destaca el Funk.

→ Conclusión vc

El análisis de las preferencias musicales a nivel regional demuestra que existe una diversidad significativa en los gustos y consumos de música alrededor del mundo. Este estudio proporciona una visión profunda sobre cómo los usuarios de diferentes regiones prefieren géneros musicales específicos y cómo ciertos artistas se destacan en cada área geográfica. Las observaciones indican que, por ejemplo, el Bebop tiene una fuerte presencia en Japón, mientras que el Rock es dominante en Bielorrusia, Corea y Kazajistán, y la Música Clásica es popular en Finlandia, Islandia y Marruecos. Sudáfrica muestra una alta preferencia por el Funk.

La segmentación de usuarios por región ha permitido identificar no solo las preferencias por géneros musicales, sino también características particulares de las canciones que son más populares en cada país, como el tempo, la energía y la duración. Este conocimiento es valioso para la industria musical, ya que permite adaptar las estrategias de marketing y promoción a las características y preferencias específicas de cada mercado regional. Además, el análisis de las canciones más populares y los artistas más recurrentes en los rankings diarios ofrece una visión actualizada y precisa de las tendencias musicales en tiempo real.

La conclusión principal es que la variabilidad regional en las preferencias musicales es un factor crucial que debe ser considerado por los profesionales de la industria musical. Entender estas dinámicas regionales no solo facilita la promoción de música de manera más eficiente, sino que también resalta la importancia de apoyar y promocionar a los artistas locales que son más relevantes en sus respectivas áreas. Este enfoque puede conducir a una mayor aceptación y éxito en los mercados específicos, optimizando los recursos y esfuerzos en campañas de marketing musical.

7. PREDICCIÓN DEL ÉXITO DE LA CANCIÓN DE UN ARTISTA

→ **Introducción**

Predecir el éxito de una canción de un artista es un reto fundamental en la industria musical. Este propósito se enfoca en realizar pronósticos fundamentados en el análisis de datos históricos, así como en la evaluación de su precisión mediante comparaciones con datos contemporáneos. Al discernir los elementos que inciden en el éxito de una nueva composición, este estudio ofrece herramientas analíticas que pueden ser empleadas para respaldar y potenciar de manera más eficaz la difusión y promoción de la obra de un artista.

Para alcanzar este objetivo, procederemos a efectuar un análisis de regresión sobre la variable "track_popularity", la cual representa la popularidad de una canción. Nuestro propósito es desentrañar la fórmula secreta, en caso de existir, que permita identificar las combinaciones que debe emplear una persona para crear una canción viral, así como los factores que influyen en ello. Para llevar a cabo este análisis, utilizaremos el lenguaje de programación R. Como se han hecho muchas propuestas de modelos de predicción diferentes y la información tiene que ser lo más resumida posible, se explicará el modelo definitivo junto con las conclusiones que conlleva consigo mismo.

→ Metodología

Se propone el modelo definitivo que aborda los errores del modelo anterior, buscando corregir la asimetría en la variable explicada y reducir la heterocedasticidad en la medida de lo posible.

$\hat{Y} = b_0 + b_1 * danceability + b_2 * energy + b_3 * loudness + b_4 * speechiness + b_5 * acousticness + b_6 * instrumentalness + b_7 * liveness + b_8 * valence + b_9 * tempo + b_{10} * log(followers + 1) + b_{11} * album\_popularity + b_{12} * artist\_popularity + b_{13} * release\_year + b_{14} * duration\_sec$

Una vez propuesto este modelo, lo metemos en R para sacar sus características.

```
> summary(final_model)

Call:
lm(formula = log_track_popularity ~ danceability + energy + loudness +
    speechiness + acousticness + instrumentalness + liveness +
    valence + tempo + duration_sec + log(followers + 1) + album_popularity +
    artist_popularity + release_year, data = datos4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69942 -0.07788 -0.03261  0.07707  0.54827

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.482e-01  3.390e-02  25.022   <2e-16 ***
danceability        2.056e-02  1.250e-03  16.443   <2e-16 ***
energy              1.728e-02  1.264e-03  13.673   <2e-16 ***
loudness            1.730e-03  4.829e-05  35.830   <2e-16 ***
speechiness        -2.450e-02  1.219e-03 -20.092   <2e-16 ***
acousticness        1.329e-02  7.320e-04  18.150   <2e-16 ***
instrumentalness   -1.149e-02  5.702e-04 -20.155   <2e-16 ***
liveness           -1.836e-02  1.129e-03 -16.264   <2e-16 ***
valence            -1.224e-02  8.641e-04 -14.160   <2e-16 ***
tempo               6.033e-05  6.155e-06   9.801   <2e-16 ***
duration_sec        2.629e-05  1.602e-06  16.415   <2e-16 ***
log(followers + 1)  1.962e-03  1.237e-04  15.860   <2e-16 ***
album_popularity    7.577e-03  1.391e-05 544.702   <2e-16 ***
artist_popularity   7.472e-04  2.101e-05  35.568   <2e-16 ***
release_year        1.170e-03  1.675e-05  69.862   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1113 on 374317 degrees of freedom
Multiple R-squared:  0.7176,    Adjusted R-squared:  0.7176
F-statistic: 6.796e+04 on 14 and 374317 DF,  p-value: < 2.2e-16
```

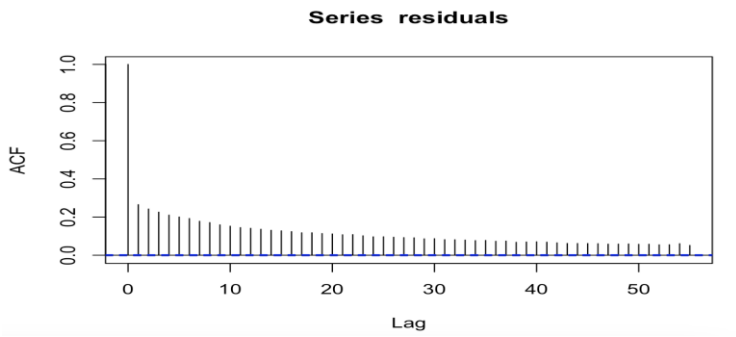
Como se puede observar, la variabilidad explicada de este nuevo modelo es bastante alta, alcanzando un 71,76%, mientras que el intervalo de residuos ha disminuido significativamente(mirar anexo para ver versión anterior del modelo). Para comprobar la eficacia del modelo se verifican las 4 propiedades que se tienen que cumplir en un modelo de regresión lineal que son:

1) ***Autocorrelación*** → para ver la existencia de la autocorrelación en este modelo, utilizaremos la prueba de Durbin-Watson. Antes de ver los resultados en R, veamos las posibles interpretaciones que tiene:

   - Si la estadística de Durbin-Watson se acerca a 2, sugiere que no hay autocorrelación de primer orden en los residuos del modelo.
   - Valores cercanos a 0 indican autocorrelación positiva.
   - Valores cercanos a 4 indican autocorrelación negativa.
   - El valor-p asociado proporciona la significancia estadística de la prueba.

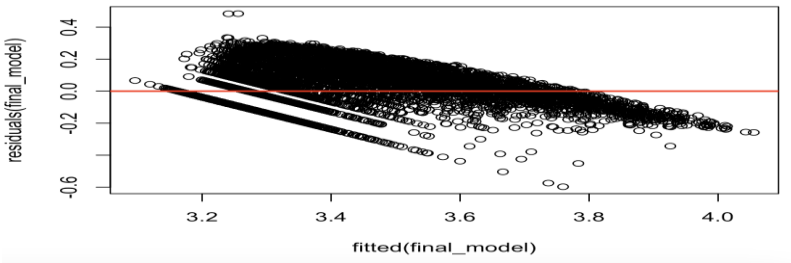
   Una vez sepamos esto, miramos en R si existe o no.

```
> durbinWatsonTest(final_model1)
 lag Autocorrelation D-W Statistic p-value
   1   -0.0006148822      2.001016   0.952
 Alternative hypothesis: rho != 0
```



Mediante el gráfico y la prueba de la Estadística de Durbin-Watson (D-W), podemos confirmar que no hay evidencia de autocorrelación en nuestro modelo definitivo. Este resultado es un punto positivo para la robustez de nuestro modelo.

2) **Heterocedasticidad** → para comprobar la existencia de la heterocedasticidad utilizaremos dos herramientas: gráfica de residuos y la prueba de Breusch-Pagan.



Con este gráfico, podemos observar una clara evidencia de la existencia de heterocedasticidad en nuestro modelo. Sin embargo, es importante destacar que es mucho menos grave que en el modelo anterior, lo cual indica una mejora en este nuevo modelo. No obstante, como medida adicional, procederemos a verificarlo mediante la prueba de Breusch-Pagan.

```
        Breusch-Pagan test

data:  final_model
BP = 5909.1, df = 14, p-value < 2.2e-16
```

Con este resultado, podemos observar una mejora significativa en el problema de heterocedasticidad que existía en nuestro modelo anterior. Esta mejora es un indicador claro de que el ajuste ha mejorado la eficiencia de la predicción del modelo, ya que hemos pasado de un valor de BP = 245176 a BP = 5909, lo cual representa una mejora significativa. Sin embargo, aún persiste el problema de heterocedasticidad, el cual será abordado de manera más efectiva en futuras iteraciones.

3) **Multicolinealidad** → tal y como se hizo anteriormente(mirar anexo), se analizará el problema de multicolinealidad mediante coeficiente VIF.

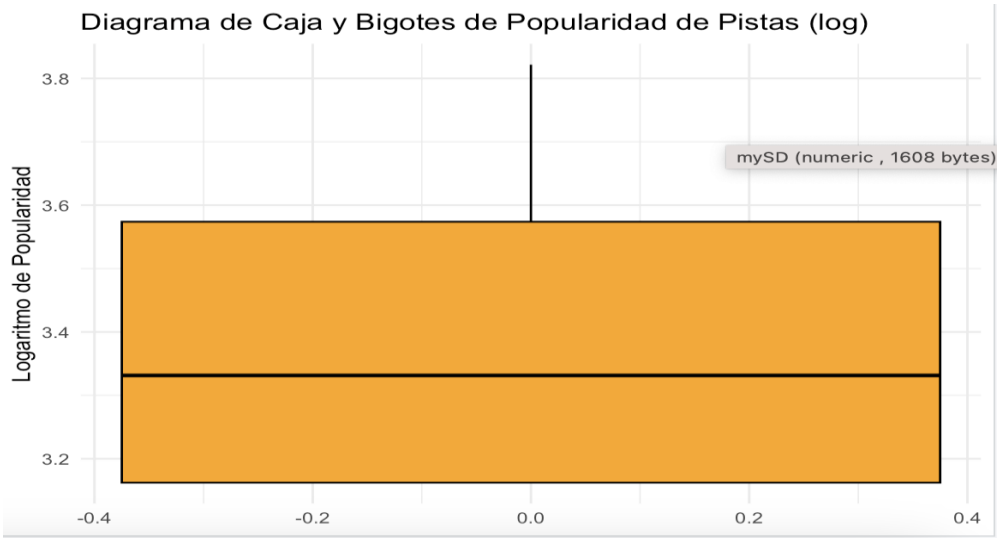
```
> print(vif_values)
      danceability             energy           loudness        speechiness       acousticness
          1.721162           3.368291           3.023008           1.192433           2.088514
  instrumentalness           liveness            valence              tempo        duration_sec
          1.413406           1.180191           1.601135           1.066633           1.054981
log(followers + 1)   album_popularity  artist_popularity       release_year
          8.332018           2.426511           9.192382           1.184695
```

Tal como se observa en este resultado, no se evidencia el problema de multicolinealidad en este modelo, lo cual sigue siendo un indicador positivo. Sin embargo, es importante destacar que los coeficientes se han incrementado y han estado cerca de rozar el límite(* → para ver).

4) **Normalidad/asimetría de la variable explicada** → Se analizará en este paso la simetría de nuestra variable explicada en el modelo. Esta evaluación se realiza para determinar la eficacia del modelo, ya que en caso de una asimetría pronunciada en la variable explicada, las predicciones pueden contener errores más significativos. Las interpretaciones son las mismas que se mencionaron anteriormente (cuanto más cercano al valor 0, mejor).

```
> print(skewness_track_popularity)
[1] 0.349777
>
```

Como se puede observar, nuestra variable explicada presenta simetría, lo cual es un indicador positivo para las predicciones de nuestro modelo definitivo y sugiere que hemos logrado resolver uno de los problemas del modelo anterior. Esta observación se respalda además con el análisis del gráfico de Box and Whisker.



→ Resultado

Como conclusión, este modelo propuesto aún presenta problemas de heterocedasticidad, aunque con una mejora significativa en comparación con el modelo anterior. Sin embargo, hemos logrado resolver el problema de la asimetría mediante las transformaciones adecuadas, pero no es suficiente. Por ende, utilizaremos el modelo de regresión PLS el cual aborda los problemas de heterocedasticidad. No se va explicar el proceso de obtención de dicho modelo ya que se obtiene de forma muy similar que en regresión lineal(lo único que se ha cambiado ha sido quitar transformaciones que no sirvieron en modelo de regresión lineal), por ende utilizaremos solo dos comando en R que son el 'summary' y 'coef' los cuales ya con una variabilidad explicada alta y resuelto el problema de heterocedasticidad nos ayudarán a sacar conclusiones.

```
pls_model <- plsr(track_popularity ~ danceability + energy + loudness + speechiness +
                    acousticness + instrumentalness + liveness + valence + tempo + followers +
                    album_popularity + artist_popularity + release_year,
                  data = datos4, scale = TRUE, validation = "CV")

summary(pls_model)
coef(pls_model)
```

```
> coef(pls_model)
, , 13 comps

                  track_popularity
danceability            0.61304900
energy                  0.25405912
loudness                0.40075109
speechiness            -0.52057364
acousticness           -0.21391130
instrumentalness       -0.28646911
liveness               -0.21960352
valence                -0.52208570
tempo                   0.06798303
followers               1.05959005
album_popularity       12.91177228
artist_popularity       0.44209768
release_year            1.19836153
```

Tras completar todos estos pasos, procedemos al paso definitivo: extraer conclusiones que ayudarán a maximizar la popularidad de una canción.

→ **Conclusiones**

- El modelo explica aproximadamente más de 70% de la variabilidad en track_popularity, lo que indica un buen ajuste.

- Las variables danceability, energy, loudness, tempo, duration_sec, followers, album_popularity, artist_popularity y release_year tienen un efecto positivo sobre track_popularity. Esto significa que cuanto más alto sea el valor de estas características, es más probable que la popularidad de la canción sea mayor.

- Por otro lado, las variables speechiness, instrumentalness,acousticness, liveness y valence tienen un efecto negativo sobre track_popularity. Esto implica que cuanto más alto sea el valor de estas características, es más probable que la popularidad de la canción sea menor.

Toda la información restante de cómo se obtuvo dicho modelo de predicción se encuentra en apartado de Anexo (en el cual puede haber una redundancia con lo escrito en este apartado)

8. COMENTARIOS FINALES

Problemas que se presentaron:

-Para hacer el análisis de distribución por países pero de los artistas, tratamos de hacer web scraping para sacar la nacionalidad de cada artista, y no lo conseguimos,ya que habían demasiados artistas como para que pudiéramos hacerlo de forma eficiente.

-Tratamos de unir los dataframes de 2020 y 2023 (concatenando), pero se perdía mucha información, por lo que decidimos estudiarlos por separado.

- Este estudio proporciona un marco útil para entender las preferencias musicales regionales, pero su impacto y utilidad pueden ampliarse significativamente mediante la integración de datos adicionales, el uso de técnicas analíticas avanzadas y la colaboración con actores clave de la industria musical. Estas mejoras aumentarían la precisión del análisis, proporcionarían insights más profundos y accionables para el desarrollo de estrategias musicales a nivel global.

- Resolver el modelo de regresión lineal que presentaba problemas de heterocedasticidad muy severos y realizar el análisis PLS para abordar el problema sin mucho detalle debido a la complejidad del análisis y muestra muy grande de datos.

- Analizar el fichero de datos de 2020 de spotify en power bi, las variables de este fichero no seguían la estructura necesaria para poder introducirse en esta herramienta de forma correcta.