AED Proyecto final

Cuál es la canción que pega

Gustavo A. Noriega C., Rubén Rico R., Jairo V. Tamayo R

2023-06-04

Proyecto final

Análisis estadístico de datos

Nombre Proyecto:¿Cuál es la canción que pega?

Tabla de Contenido

Introducción

La creación de una canción exitosa es una meta anhelada por la mayoría de los artistas, pero no siempre es fácil lograrla. En este trabajo, se propone realizar un análisis estadístico de las variables obtenidas de las dos plataformas más populares de difusión musical en línea, Spotify y YouTube, con el objetivo de identificar los atributos que influyen en la aceptación del público y el éxito comercial de una nueva canción en el mercado digital. Este estudio permitirá a los artistas enfocarse en los atributos más relevantes y optimizar así su producción, incrementando sus posibilidades de triunfar en la industria musical. Además, se pretende evaluar si es posible predecir la aceptación del público y determinar si el artista ofrece al mercado un producto apetecido.

Objetivos

1. Establecer cuáles son las variables que afectan favorablemente la aceptación del público mediante la aplicación de métodos estadísticos al conjunto de datos de las 10 mejores canciones de múltiples artistas publicadas en Spotify.
2. Evaluar si existe alguna correlación entre el número de vistas en Youtube, el número de comentarios y el número de "me gusta", y la frecuencia con la que es sugerido un tema por parte del algoritmo de la aplicación.
3. El objetivo principal de este proyecto es producir una aplicación de los métodos estadísticos multivariantes vistos a lo largo del curso utilizando uno o varios conjuntos de datos.

1. Descripción de la pregunta de investigación

El propósito de esta investigación es realizar un análisis estadístico de las variables obtenidas de dos populares plataformas de streaming de música en línea, Spotify y YouTube, para investigar los factores que influyen en la aceptación de la audiencia y el éxito comercial de nuevas canciones en el mercado digital. La pregunta principal de la investigación es: ¿Cuáles son los atributos clave que contribuyen significativamente a la aceptación por parte de la audiencia y al éxito comercial de las canciones en las plataformas digitales de música?

Los problemas de investigación identificados son los siguientes:

1. Identificar los atributos musicales y no musicales que están asociados con una mayor aceptación “likes” de la audiencia en Spotify y YouTube.
2. Determinar si existen diferencias significativas en los atributos preferidos por la audiencia entre ambas plataformas.
3. Evaluar la viabilidad de predecir la aceptación “likes” de la audiencia en función de los atributos identificados.
4. Investigar si ofrecer productos musicales con los atributos deseados influye en el éxito comercial de los artistas en la industria musical.

Al realizar un análisis estadístico de los datos obtenidos de Spotify y YouTube, esta investigación pretende proporcionar información valiosa a los artistas y productores, permitiéndoles centrarse en los atributos más relevantes y optimizar la producción de sus canciones, aumentando así sus posibilidades de éxito en la competitiva industria musical. Además, las conclusiones de este estudio pueden tener implicaciones prácticas para la toma de decisiones estratégicas en el ámbito de la promoción y distribución de música en línea.

2. Descripción de los datos

En el conjunto de datos ubicado en la URL https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube se encuentra una matriz de datos de tamaño [20,718 , 28]. Dentro de esta matriz, al menos 10 de las 28 variables son numéricas. Entre estas variables se encuentra una que registra la cantidad de reproducciones en Spotify y YouTube, así como las manifestaciones positivas o “likes”.

El conjunto de datos del enlace mencionado consta de 20,718 observaciones o instancias, y cada observación está asociada a 28 variables distintas.

Estas variables numéricas abarcan diferentes aspectos de las canciones, como el número de reproducciones o streams en Spotify y YouTube. Además, también se incluyen variables que registran manifestaciones positivas o comentarios, los cuales podrían ser indicativos del nivel de aceptación o compromiso del público con las canciones.

2.1 Descripción conceptual de los datos.

Los datos presentados en el conjunto de información corresponden a diferentes variables relacionadas con características y métricas de canciones. A continuación, se describe brevemente cada una de las variables que se utilizarán en el análisis:

Danceability: Es una medida de la aptitud de una canción para el baile, variando entre 0 y 1.

Energy: Representa la intensidad y actividad percibida en una canción, también varía entre 0 y 1.

Key: Indica la clave tonal de una canción, siendo un número del 0 al 11 que corresponde a las diferentes tonalidades musicales.

Loudness: Es una medida del volumen general de una canción en decibeles (dB).

Speechiness: Refleja la presencia de elementos hablados en una canción, con valores más altos indicando una mayor presencia de palabras habladas.

Acousticness: Indica la cantidad de elementos acústicos presentes en una canción, siendo 0 para canciones no acústicas y 1 para canciones totalmente acústicas.

Instrumentalness: Mide la probabilidad de que una canción sea instrumental, donde valores cercanos a 1 indican alta probabilidad de ser instrumental.

Liveness: Representa la percepción de si una canción fue grabada en vivo o en estudio, con valores más altos indicando mayor probabilidad de ser en vivo.

Valence: Es una medida de la positividad o negatividad de una canción, variando entre 0 y 1.

Tempo: Indica el ritmo o velocidad de una canción en beats por minuto (BPM).

Duration_ms: Es la duración de una canción en milisegundos.

Views: Representa la cantidad de visualizaciones o reproducciones que ha recibido una canción.

Likes: Indica la cantidad de “me gusta” o manifestaciones positivas que ha recibido una canción.

Comments: Muestra la cantidad de comentarios que ha generado una canción.

Stream: Representa la cantidad de veces que una canción ha sido transmitida o reproducida en una plataforma de streaming.

Danceability	Energy	Key	Loudness	Speechiness	Acousticness	Instrumentalness	Liveness
0.8389744	0.7049940	0.5454545	0.8389053	0.1836100	0.0083925	0.002330	0.6073059
0.6933333	0.7029940	0.7272727	0.8572216	0.0313278	0.0872480	0.000687	0.0322679
0.7128205	0.9229984	0.0909091	0.8971826	0.0541494	0.0426696	0.046900	0.1029934
0.7066667	0.7389947	0.1818182	0.8573276	0.0269710	0.0000140	0.509000	0.0502283
0.6800000	0.6939938	0.9090909	0.7976087	0.1773859	0.0254005	0.000000	0.0561136
0.7794872	0.8909978	1.0000000	0.8564372	0.0385892	0.0229909	0.086900	0.2876712

	Danceability	Energy	Key	Loudness	Speechiness	Acousticness	Instrumentalness	Liveness
	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.00000	Min. :0.00000	Min. :0.0000000	Min. :0.00000
	1st Qu.:0.5323	1st Qu.:0.5080	1st Qu.:0.1818	1st Qu.:0.7945	1st Qu.:0.03703	1st Qu.:0.04458	1st Qu.:0.0000000	1st Qu.:0.08067
	Median :0.6554	Median :0.6660	Median :0.4545	Median :0.8424	Median :0.05259	Median :0.19076	Median :0.0000023	Median :0.11213
	Mean :0.6370	Mean :0.6352	Mean :0.4813	Mean :0.8187	Mean :0.09895	Mean :0.29027	Mean :0.0552922	Mean :0.17933
	3rd Qu.:0.7610	3rd Qu.:0.7970	3rd Qu.:0.7273	3rd Qu.:0.8760	3rd Qu.:0.10788	3rd Qu.:0.47189	3rd Qu.:0.0004330	3rd Qu.:0.22273
	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.00000	Max. :1.00000	Max. :1.0000000	Max. :1.00000

Es importante tener en cuenta que en el conjunto de datos hay valores faltantes (NA's) en algunas de las variables. Estos valores faltantes serán tratados adecuadamente en el análisis posterior.

Con base en la exploración de los datos, se consideró que las siguientes variables no deben tenerse en cuenta en el modelo de regresión lineal múltiple: **X, Artista, Url_spotify, Track, Álbum, Tipo_álbum, Uri, Url_youtube, Title, Canal, Descripción, Licensed, official_video,** debido a que son variables categóricas, identificadores únicos y variables relacionadas con la información descriptiva. Estas variables carecen de un carácter cuantitativo directo y no proporcionan información relevante para predecir con exactitud la variable dependiente. Al excluir estas variables del análisis, podemos centrarnos en los predictores significativos que mejorarán la capacidad predictiva y la fiabilidad del modelo.

2.2 Identificación y tratamiento de datos atípicos

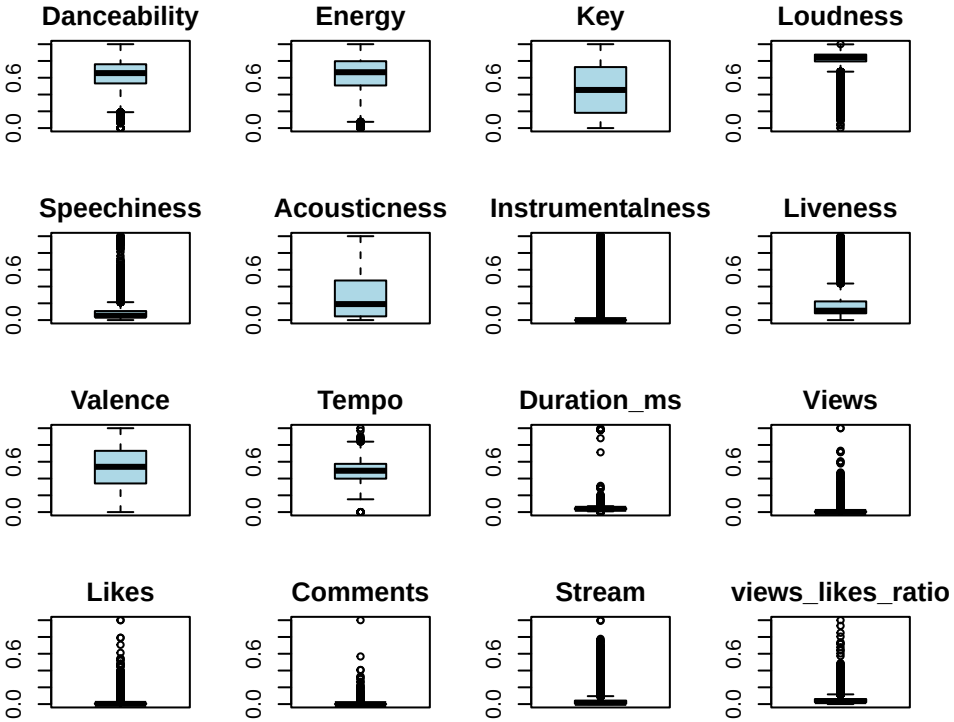
En esta parte, se identificaron los datos faltantes del conjunto de datos "datos", y se eliminaron dichas observaciones del conjunto de datos.

Adicionalmente, se calculó la relación entre las variables "Views" y "Likes" del conjunto de datos y el resultado se asignó a una nueva columna llamada "views_likes_ratio", El conjunto de datos resultante, se muestra a continuación, luego de identificar y eliminar los datos faltantes:

2.3 Análisis individual de variables.

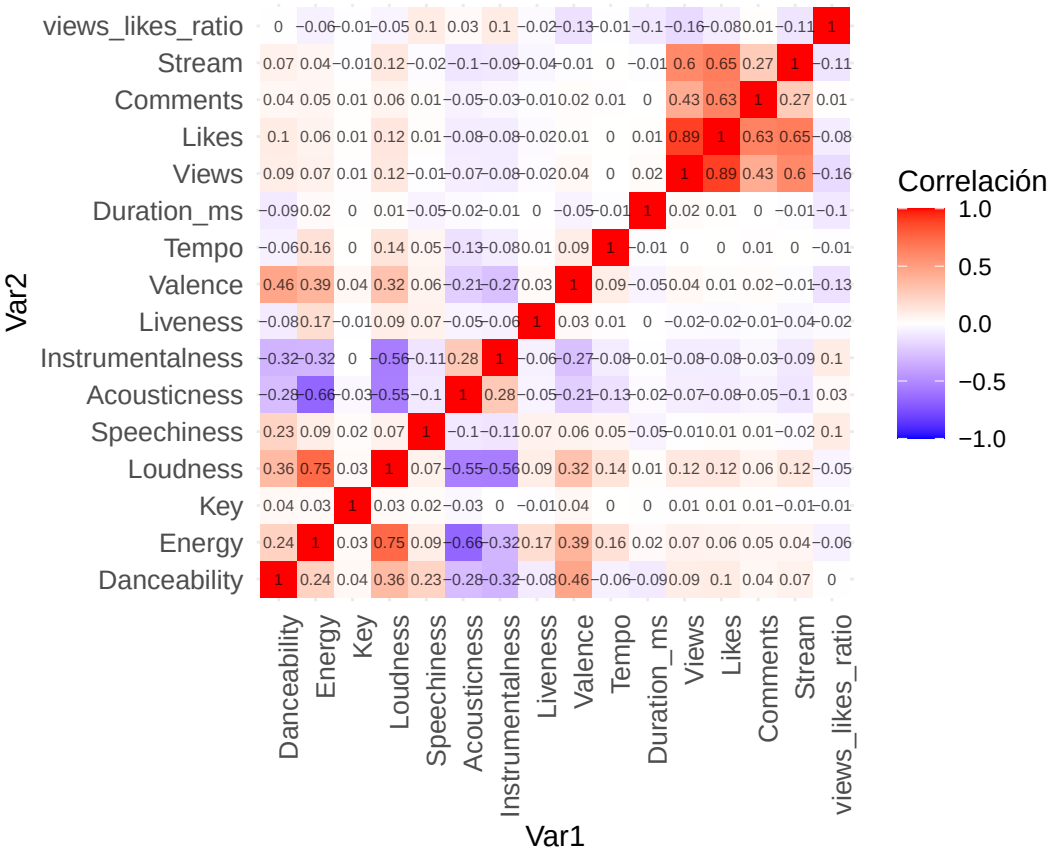
La siguiente tabla contiene los estadísticos de cada una de las variables del conjunto de datos a analizar:

A continuación se presentan los diagramas de caja o Boxplot de cada una de las variables:



2.4 Análisis conjunto de variables (correlaciones)

A continuación, se presenta la Matriz de correlaciones entre las variables del conjunto de datos.



Basado en la Matriz de correlaciones, a continuación se describen algunas de las correlaciones más relevantes, que se identificaron:

Danceability: Tiene una correlación positiva moderada con Loudness (0.36), Valence (0.46) y Energy (0.24).

Energy: Muestra una correlación positiva fuerte con Loudness (0.75).

Loudness: Tiene una correlación positiva moderada con Danceability (0.36) y Energy (0.75).

Valence: Muestra una correlación positiva moderada con Danceability (0.46) y una correlación negativa moderada con Acousticness (-0.21).

Acousticness: Tiene una correlación negativa moderada con Energy (-0.66).

Instrumentalness: Muestra una correlación negativa moderada con Acousticness (0.28).

Views: Tiene una correlación positiva fuerte con Likes (0.89).

Likes: Muestra una correlación positiva fuerte con Views (0.89) y una correlación moderada con Comments (0.63).

Comments: Tiene una correlación moderada con Views (0.43) y Likes (0.63).

Stream: Muestra una correlación moderada con Views (0.60) y Likes (0.65).

3. Presentación del análisis estadístico de los datos

Métodos y resultados.

3.1 Modelo de Regresión Lineal Múltiple.

A continuación se realiza el modelo de regresión lineal múltiple a las variables del conjunto de datos que presentaron correlación significativa:

El modelo de regresión lineal tiene las siguientes conclusiones principales:

Las variables predictoras Danceability, Views, Comments y Stream tienen un impacto significativo en la variable de respuesta Likes. Esto se evidencia por los valores p muy bajos (< 0.05) asociados a estas variables en la tabla de coeficientes.

La variable Danceability muestra una relación positiva significativa con Likes, lo que indica que a medida que aumenta la danceabilidad de una canción, es probable que aumente el número de likes.

Las variables Views, Comments y Stream también tienen una relación positiva significativa con Likes. Esto sugiere que cuanto más vistas, comentarios y transmisiones tenga una canción, es más probable que tenga más likes.

La variable Loudness no muestra una relación significativa con Likes, como se indica por el valor p elevado (0.38137). Esto implica que el nivel de sonoridad no tiene un impacto significativo en el número de likes.

El modelo tiene un coeficiente de determinación (R-cuadrado) de 0.8915, lo que significa que aproximadamente el **89.15%** de la variabilidad en la variable Likes puede ser explicada por las variables predictoras incluidas en el modelo.

En resumen, el modelo sugiere que la danceabilidad, las vistas, los comentarios y las transmisiones son factores importantes que influyen en la cantidad de likes que recibe una canción, mientras que el nivel de sonoridad no tiene un impacto significativo en los Likes.

Con el objetivo de mejorar el ajuste del modelo de regresión lineal múltiple, consideraremos la eliminación de la variable Loudness que arrojó un valor p alto (p $>$ 0.05n):

Al retirar esta variable, el modelo podría tener un mayor ajuste y mayor capacidad para explicar la variabilidad en la variable dependiente “Likes”.

En resumen, ambos modelos llegan a conclusiones similares en términos de las variables predictoras Danceability, Views, Comments y Stream, las cuales tienen un impacto significativo en la cantidad de likes. Sin embargo, la variable Loudness no muestra un efecto significativo en los likes, por lo que su inclusión en el modelo no es necesaria para explicar la variabilidad en la variable de respuesta.

A continuación se presentan los resultados de las pruebas de normalidad, homocedasticidad y correlación para la validación del modelo

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos_RL1$residuals
## D = 0.47991, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  Goldfeld-Quandt test
##
## data:  datos_RL1
## GQ = 1510.3, df1 = 9770, df2 = 9769, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

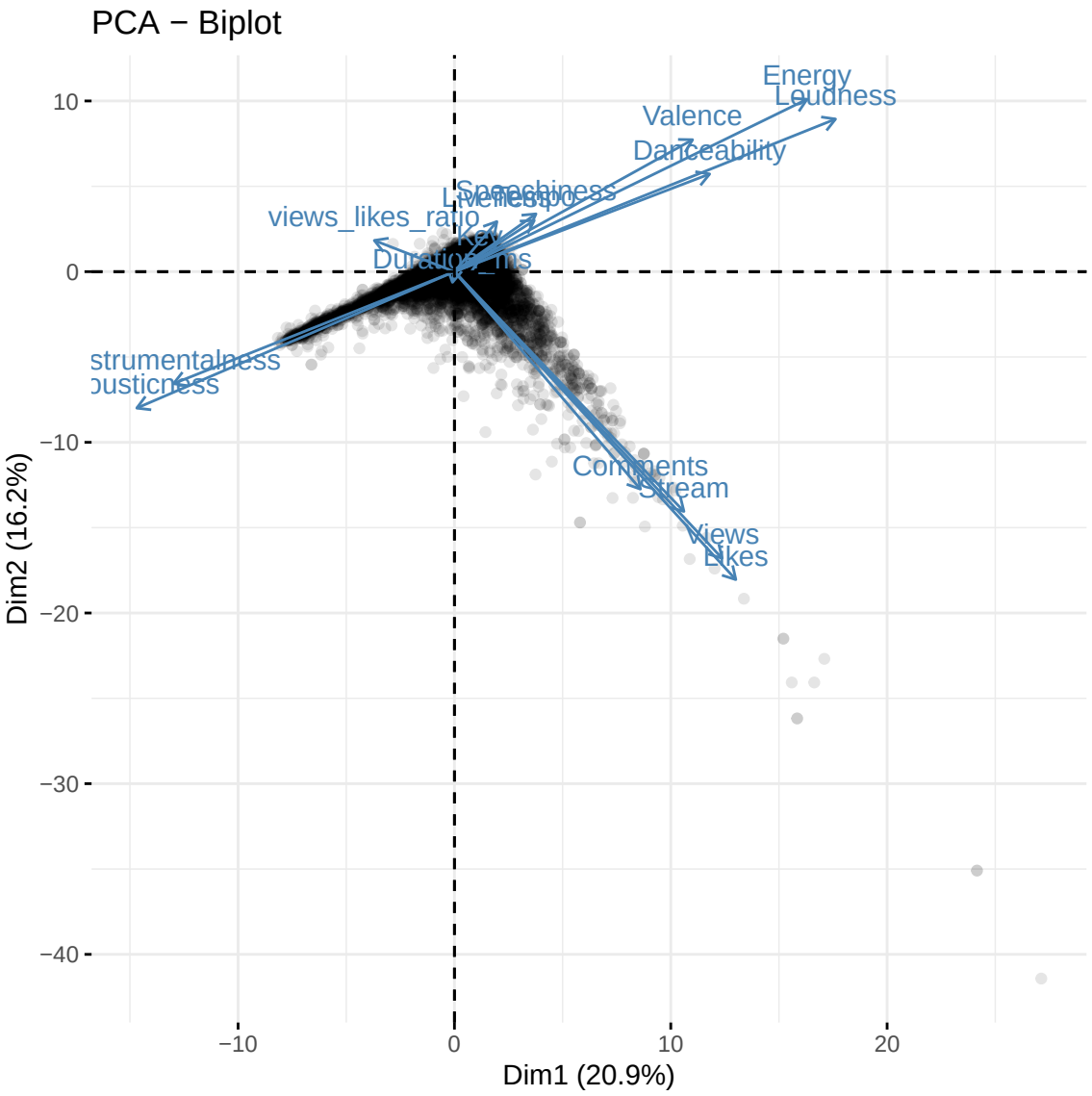
```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  datos_RL1
## LM test = 534.18, df = 1, p-value < 2.2e-16
```

3.2 Modelo PCA.

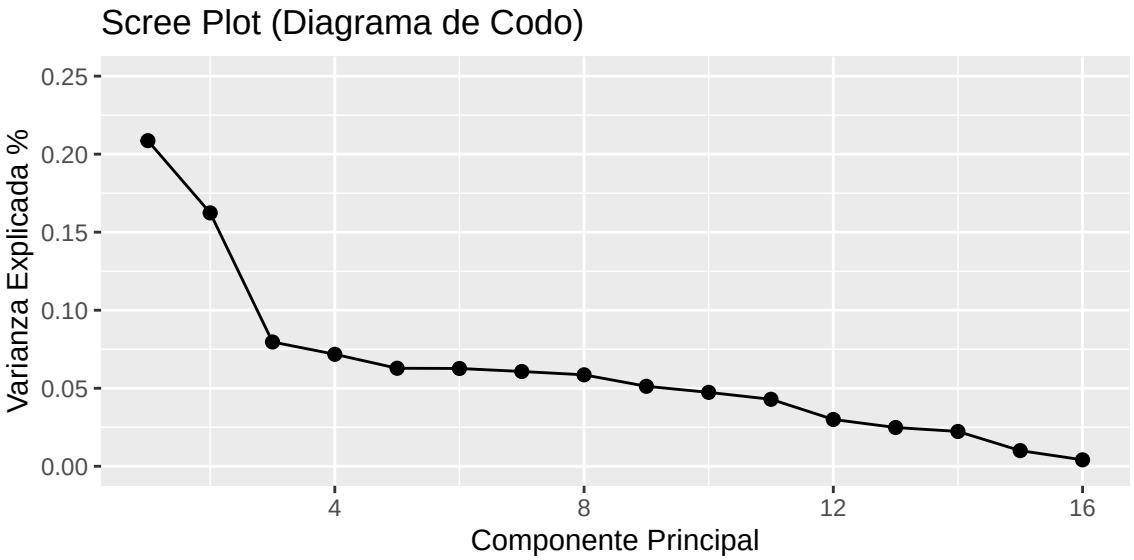
En esta etapa del análisis, se procederá a implementar el modelo PCA al conjunto de datos “datos”. El PCA, o Análisis de Componentes Principales, es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos y extraer la información más relevante.

Cuadro 1: Relación de participación de las variables en los componentes

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Danceability	0.2799606	0.1539126	0.5097511	-0.2480573	0.0717180	0.0009770	-0.0657764	0.0624970	0.1175435	-0.0359087
Energy	0.3867395	0.2716819	-0.2278834	0.1414756	-0.0295428	-0.0198800	-0.0679830	-0.1653760	0.0689910	-0.2522000
Key	0.0276038	0.0191370	0.0737596	-0.0821679	-0.5285550	-0.8066561	0.0798717	-0.1431456	-0.1189356	0.1238393
Loudness	0.4179956	0.2405056	-0.1317968	0.0509488	-0.0255146	0.0632203	-0.1220234	-0.1196550	-0.1218029	0.1660338
Speechiness	0.0894866	0.0910333	0.4133927	0.3425965	0.1866082	-0.2312357	-0.0174585	0.6149271	-0.2504042	-0.2636062
Acousticness	-0.3484034	-0.2150388	0.1650939	-0.1622501	0.1226381	-0.0276339	0.2265534	0.1413425	0.1185772	0.3891708
Instrumentalness	-0.3081307	-0.1765405	-0.0314756	0.1292779	-0.1280634	-0.0799123	0.0077761	-0.1229266	0.2848532	-0.7295579
Liveness	0.0465753	0.0789807	-0.2755659	0.3517757	0.6020618	-0.4024295	0.3475219	-0.0899936	0.1527891	0.1239857
Valence	0.2612216	0.2079440	0.2186244	-0.3268616	0.0178654	-0.0062570	0.2922205	0.0280544	0.5141329	-0.0795066
Tempo	0.0875957	0.0806888	-0.2304182	0.2994411	-0.5097463	0.2559853	0.4693629	0.4494933	0.1107633	0.1071043



El análisis del PCA reveló que las variables Danceability, Valence, Energy, Loudness, Views, Likes y Key son las más relevantes en la explicación de la variabilidad de los datos en las dimensiones 1 y 2. Estos hallazgos ayudan a identificar las dimensiones principales en los datos y permiten una reducción de la dimensionalidad al representar los datos en un espacio de menor dimensión definido por los componentes principales.



El **PC1** explica aproximadamente el 20.86% de la varianza total de los datos y muestra cargas positivas más altas en las variables Danceability, Valence, Views y Likes. Esto indica que estas variables tienen una mayor influencia en el componente principal 1.

El **PC2** explica alrededor del 16.24% de la varianza y está influenciado principalmente por las variables Energy, Loudness, y Views, todas con cargas positivas significativas.

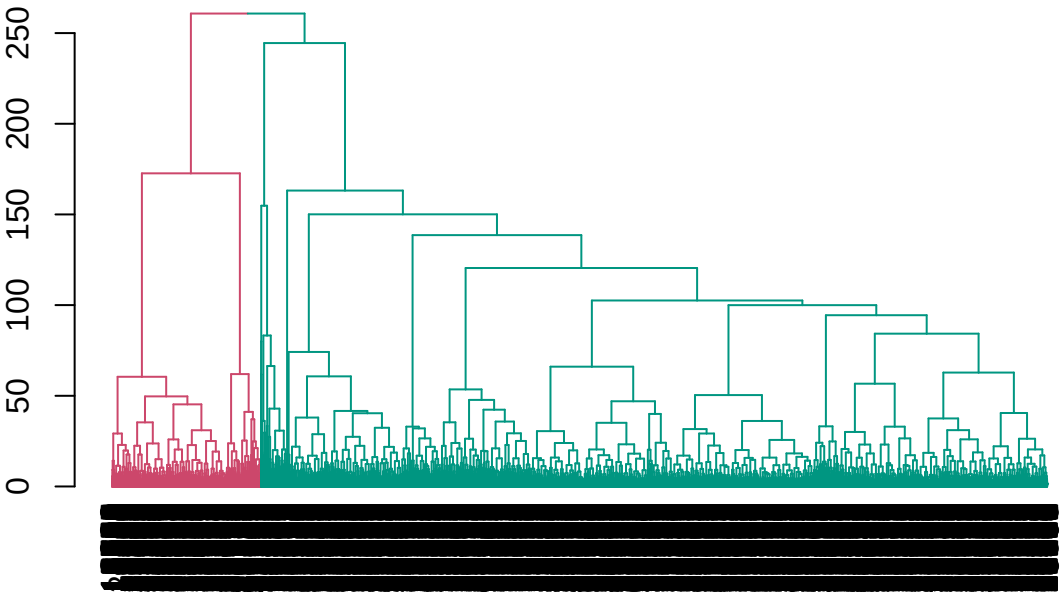
El **PC3** tiene una varianza explicada de aproximadamente el 7.96% y se ve principalmente afectado por la variable Key, con una carga positiva alta.

El **PC4** y **PC5** explican el 7.18% y 6.28% de la varianza, respectivamente. El PC4 muestra una alta carga positiva en la variable Loudness, mientras que el PC5 está influenciado por las variables Speechiness y Acousticness.

Los componentes principales restantes (**PC6** a **PC16**) explican una menor cantidad de varianza y están influenciados por diversas combinaciones de variables.

3.3 Análisis de Cluster.

El análisis de clúster (cluster) que se realizará en este capítulo, tiene como objetivo agrupar las observaciones en conjuntos homogéneos o clústeres, según la similitud de sus características. En este caso, se aplicó el análisis de clúster jerárquico utilizando diferentes métodos de enlace, como enlace completo, enlace simple, enlace promedio y enlace Ward.



Conclusiones.

- Las variables presentadas en el conjunto de datos incluyen características y métricas de canciones, como la aptitud para el baile, intensidad, clave tonal, volumen, presencia de elementos hablados, acústica, instrumentalidad, percepción de grabación en vivo, positividad, ritmo, duración, visualizaciones, "Likes", comentarios y reproducciones en streaming. Se identificaron variables categóricas y descriptivas que no aportaban información cuantitativa relevante para el modelo de regresión lineal múltiple, por lo que se excluyeron. Para evitar la distorsión generada por el desbalance de los datos dado que las variables oscilaban entre unidades y miles de millones fue necesario normalizarlas.
- El análisis de la Matriz de correlaciones reveló relaciones positivas moderadas entre Danceability y Loudness, Valence, Energy, así como entre Loudness y Danceability, Energy. Además, se encontró una correlación negativa moderada entre Valence y Acousticness. Asimismo, se identificaron correlaciones fuertes positivas entre Views, Likes, Comments y Stream. Estas correlaciones resaltan la importancia de estas variables en el análisis de los datos y proporcionan información valiosa para comprender las relaciones entre las características y métricas de las canciones estudiadas.
- El modelo de regresión lineal reveló que las variables predictoras Danceability, Views, Comments y Stream tuvieron un impacto significativo en la variable de respuesta Likes. Se encontró una relación positiva significativa entre Danceability y Likes, así como entre Views, Comments y Stream con Likes. Sin embargo, no se encontró evidencia de una relación significativa entre Loudness y Likes. El modelo tuvo un coeficiente de determinación de 0.8915, lo que indicó que aproximadamente el 89.15% de la variabilidad en Likes pudo ser explicada por las variables predictoras.
- Al realizar la comparación de los modelos de regresión lineal múltiple con y sin la variable Loudness, se concluye que las variables Danceability, Views, Comments y Stream tienen un impacto significativo en la cantidad de likes. Sin embargo, la variable Loudness no tiene un efecto significativo en los likes. Por lo tanto, se puede prescindir de la variable Loudness en el modelo, simplificando su interpretación y ajuste, mientras se mantienen las variables predictoras clave para explicar la variabilidad en los likes. Las pruebas sobre los supuestos del modelo simplificado arrojaron un p-value $<$ 2.2e-16, presenta una significancia estadística menor a 0.05, que indica una evidencia suficiente de que los datos se distribuyen de manera normal, son homocedasticos y están correlacionados.
- La exploración del Modelo PCA señaló que las variables Danceability, Valence, Energy, Loudness, Views, Likes y Key son relevantes para explicar la variabilidad de los datos. Estos resultados facilitan la identificación de las dimensiones principales y permiten reducir la dimensionalidad por los componentes principales. Igualmente, se puede apreciar la existencia de dos tendencias claramente definidas en los géneros musicales. Por un lado, hay grupos de canciones muy dinámicas, mientras que por otro lado, se destacan aquellas que predominan por su carácter acústico e instrumental.
- Hasta el **PC6** se explica aproximadamente el 64.78% de la varianza total de los datos.
- Basados en los resultados obtenidos con el modelo PCA, se realizó un análisis de Clusters para confirmar la existencia de las dos grandes tendencias en los géneros musicales, y se observó que uno de los grupos contiene un mayor número de canciones en comparación con el otro.
- Las canciones exitosas se pueden determinar por tres criterios, que sean acústicas e instrumentales, que sean muy dinámicas o por último que tengan una amplia difusión, entonces, una canción de muy buena interpretación con mucha difusión llegará a ser un éxito y por otro lado una canción muy dinámica con poca o nula difusión no lo será e inclusive podría darse que una canción de regular interpretación y medio dinamismo y con mucha difusión, llegará a ser categorizada con un éxito musical.