ENTREGA PARCIAL PROYECTO

Spotify - All Time Top 2000s

Integrantes:	Francisco Moreno
	Rafael Saavedra
	Cristóbal Yob
Profesor:	Pablo Lemus
Ayudante:	Elias Alegría
Fecha:	12 Octubre 2021

Tabla de contenido

Contexto

Spotify es una empresa de servicios multimedia que cuenta con múltiples locaciones a través del mundo, con su casa central en Suecia. En esta plataforma se pueden crear listas de reproducción con las canciones más escuchadas de cada país, según su género musical, etc. A través de la data de “Spotify - All time top 2000s mega dataset” están registradas las 2000 canciones más escuchadas a nivel mundial de manera digital en Spotify, con sus respectivos géneros, años de lanzamiento, beats por minutos, sí es bailable, etc. Se trabajará en esta data para conocer qué componentes son relevantes a la hora de crear una canción que resulte en una de las más populares de Spotify. La popularidad es importante en el algoritmo de Spotify ya que mientras más popular es una canción, tiene mayor probabilidad de aparecer en las recomendaciones realizadas por Spotify y por lo tanto alcanzar una mayor audiencia. Tras un éxito musical existen varios beneficiarios en ámbitos tanto económicos como popularidad, como lo son los artistas, representantes, disqueras, entre otros, los cuales se pueden interesar de los resultados que se obtengan.
Dado esto se Podrían Abordar las siguientes preguntas:

- ¿Qué elementos son relevantes para tener más popularidad en Spotify ?
- ¿Existe alguna correlación entre el tipo de género y la popularidad de la canción?
- ¿Cuánto afecta la energía de una canción a la popularidad alcanzada?
- ¿Cuánto afecta el ánimo de una canción a la popularidad alcanzada?
- ¿Son las canciones más bailables las más populares?
- ¿Cuánto afecta el título de una canción a la popularidad alcanzada?
- ¿Cuánto afecta el artista de una canción a la popularidad alcanzada?
- ¿Cuánto afecta el género de una canción a la popularidad alcanzada?

Si se logran responder estas preguntas, se lograría conocer las preferencias populares en el consumo de música a través de Spotify con lo que se podría establecer si las personas en general siguen un patrón musical en el transcurso del tiempo. Además, sería más fácil entender el funcionamiento de la industria y conocer cómo esta cambia y si es posible establecer ciertos patrones para lograr un éxito musical.

Base De Datos Por Utilizar

Se escogerá la base de datos Spotify - All Time Top 2000s Mega Dataset. Esta base de datos se escogió de la página kaggle.com, la cual contiene muchos datasets creados por la comunidad. Este data en específico contiene las características de las canciones más reproducidas en Spotify de todos los tiempos. Con la ayuda del algoritmo sort your music creado por Paul Lamere se podrá insertar una playlist y se entregará sus variables. A continuación, se entregarán las variables a utilizar con sus respectivas definiciones.

- **Index**: Índice de la canción en la playlist (este es índice no tiene ningún orden en específico solo es para manejar más fácil los datos).
- **Title**: Nombre de la canción (Var. independiente).
- **Artist**: Nombre del artista (Var. independiente).
- **Top Genre**: Género de la canción (Var. independiente).
- **Year**: Año de lanzamiento de la canción (Var. independiente).
- **Beats per Minute (BPM)**: El tempo de la canción (Var. independiente).

- **Energy**: La energía que produce la canción, mientras más alto más energética la canción (Var. independiente).
- **Danceability**: La facilidad de bailar la canción, mientras más alto, más fácil es bailar esta canción (Var. independiente).
- **Loudness**: La ruidosidad de la canción medido en decibeles, mientras más alta, más ruidosa (Var. independiente).
- **Valence**: El ánimo de la canción, mientras más alto, más positiva es la canción (var. independiente).
- **Length**: La duración de la canción. (Var. independiente).
- **Acoustic**: Mientras más alto, más acústica es la canción (Var. independiente).
- **Speechiness**: Mientras más alto el valor existan tendrá más palabras habladas (no cantadas), como lo puede ser una poesía o un audiobook .(Var. independiente).
- **Popularidad:** Mientras más alto el valor, más popular es en estos momentos. (var. dependiente).

Tipo de proyecto

El tipo de proyecto a utilizar seria parte de un aprendizaje no supervisado llamado clúster el cual sirve para poder agrupar datos y generar grupos que serán útiles para analizar las variables relacionadas, en donde se tendrá que analizar y asociar las variables dependientes e independientes para llegar a diversas conclusiones.
Los posibles algoritmos de este tipo pueden ser:
Centroid class Cluster, el cual funciona generando un centroide por cada k que escoja el usuario y posteriormente el algoritmo busca que los centroides sean los más cercanos a los datos para así crear un cluster.
Conectivity based cluster, el cual conecta objetos basado en la distancia.
Distribution based cluster este algoritmo se basa de que los datos se distribuyen de manera normal y se separan en clusters en base a probabilidad en vez de distancia, sufren el problema de estar overfitting usualmente.

Trabajo en R

En esta parte del trabajo se explica el proceso del trabajo realizado en R

Lectura de la Base

En este apartado fue importada la base de datos anteriormente nombrada mediante la función read.csv asignando el nombre de Sp al ser una abreviación de Spotify.

Análisis Exploratorio de Datos

Se procede a realizar un análisis exploratorio de los datos mediante un Summary (Ilustración 1 Screenshot de la consola.) con el cual podemos observar diferentes parámetros de nuestro Data Frame, es esencial revisar si nuestras variables tienen datos nulos o NA, para así corregir en el caso que existiese un error de lectura o bien imputar datos, en nuestro Data Frame no existen datos nulos ni tampoco outliers y no se logran visualizar incoherencias en estos.

Posteriormente se realizó un skim que también sirve para visualizar parámetros y nos indica si las variables están completas o tienen datos perdidos, y el valor complete rate es en todos igual a 1, por lo que se puede corroborar lo visto anteriormente en el summary, además es realizado un sum(duplicated) para ver si hay existencias de datos duplicados dando como resultado un valor de 0, por lo que no existen valores duplicados.
Al realizar un gráfico de densidad (Ilustración 2 Densidad de las variables) es posible observar que Loudness Speechiness Liveness y Acousticness, están mayormente concentrados a la izquierda del gráfico, teniendo valores cercanos a 0.
Al realizar un gráfico de tendencias (Ilustración 3 Gráfico de tendencias), podemos visualizar de manera rápida las tendencias que ha tenido cada una de las variables analizadas a través del tiempo, se puede ver que es común que aquellas canciones de mayor antigüedad tengan una mayor popularidad en promedio.
Es con todo lo anterior que podemos concluir que la calidad de los datos de nuestra base es de muy buena calidad y por ende no se necesita una imputación de datos.

Transformación de datos

Para transformar los datos fueron creados dos subconjuntos, donde en el primer subconjunto denominado Sp (Spotify) se eliminó la columna Índex que representa el índice, dato que no aporta ninguna información al trabajo, posterior a esto fueron transformados los datos de título, artista y género a factores, para así lograr una mayor facilidad en la interpretación (Ilustración 4 Screenshot 2 de la consola), a continuación se crea el segundo subconjunto llamado Spnumeric (Spotify Numeric), donde se eliminan aquellas variables no numéricas, dado que para realizar la matriz de correlaciones y otros se necesitan solo valores numéricos. este subconjunto eliminando las variables que no son numéricas(Ilustración 5 Screenshot 3 de la consola

Matriz de correlación y correlación de Pearson

En este apartado se creo una Matriz de correlación generada solamente con los datos numéricos, uno del tipo circular (Ilustración 6 Matriz de correlaciones circular) en el que se puede ver de mejor manera cuales datos son los más correlacionados, dando a entender que mientras mas grande sea el circulo, más grande la correlación y además se muestra la correlación negativa con color rojo y la positiva con un color azul para facilitar el entendimiento, pero para realizar un análisis numérico de correlaciones fue realizada una segunda matriz de correlaciones del tipo numérico (Ilustración 7 Matriz de correlación numérica), de esta es posible ver las correlaciones más grandes (en donde un valor considerado grande es >=0,5 o <=-0,5), obviando la correlación entre variables, los valores de mayor magnitud obtenidos fueron:
Loudness-Energy (correlación =0,7), Acousticness-Energy (correlación = -0,7), Valence-Danceability (correlación = 0,5) y Acousticness-Loudness (correlación=-0,5).
Gracias a los resultados anteriores es posible observar que estos pares de datos poseen una correlación alta y lo que nos indica que una mayor magnitud da pie a que una variable pueda no ser tan relevante debido a que la otra variable es muy parecida a esta.
Correlación de Pearson (Ilustración 8 Correlación Pearson importancia y selección), mediante la correlación de Pearson se puede observar que las variables más importantes para la correlación lineal son Loudness, Danceability, Year, Liveness y Speechiness, pero al analizar los pesos se observa que estos son pequeños, por lo que un modelo lineal no sería el más

adecuado para trabajar con nuestros datos, es por esto que no se procederá a trabajar con un modelo del tipo lineal.

Variables más relevantes (Selección de atributos)

En esta sección se probaron diferentes tipos de modelos para obtener que atributos serían los importantes para este trabajo.

Random Forest

Fue utilizado Random Forest (Ilustración 10 Selección de atributos Random Forest) con Sp numérico para explicar la popularidad, donde fue obtenido que las variables independientes que mejor explican el modelo son Year, Duration y Danceability (Donde pareciera que Year es un outlier debido a la separación), gracias a los resultados es posible concluir que se podrá eliminar Speechiness dado que no es una variable tan explicativa.

Reg lineal según Cp y Bic

Fue utilizada regresión linear para escoger el mejor modelo según los datos entregados por el modelo Cp (Ilustración 11 Reg lineal según Cp) en donde es obtenido un modelo con todas las variables excepto la variable Beats per minute. Utilizado el modelo BIC (Ilustración 12 Reg lineal según BICIlustración 12 Reg lineal según) se escoge un modelo que no considera las variables Valence, BPM, Duration y Acousticness.

Cluster K

Fue generado un cluster con diferentes números de k (Ilustración 13 Cluster con diferentes K gráfico) junto con la gráfica de la mejor representación del problema (Ilustración 14 Número optimo de clusters), donde fue obtenido que el numero óptimo de k son 4 , debido que es donde se tiene la mayor curvatura del codo.

Conclusión e Hipótesis

Podemos concluir que nuestra base de datos es de una excedente calidad, debido a esto no fue posible realizar una imputación de datos, además fue posible observar que se tienen cuatro variables con valores cercanos a 0, además fue posible observar que el año de estreno de la canción es una variable que impacta directamente en la Popularidad, dado que aquellas con mayor antigüedad son en promedio más Populares que las nuevas, aunque se puede observar que en el 2019 se está revirtiendo esta tendencia. Fue posible observar que solo hay 4 pares de variables con una alta correlación, las cuales de ser necesario se procederá a utilizar solo una del par para tener una menor cantidad de variables a utilizar, aunque al analizar los pesos es posible observar que no se condicen para ser trabajadas con un modelo del tipo lineal dado que son muy pequeños, posterior esto fue utilizado el proceso de clustering para conocer qué valor toma K, además fue utilizado el modelo Cp y BIC para visualizar aquellas variables más explicativas. Con los resultados obtenidos es posible pensar que se eliminara la variable Acousticness dado que posee bajos valores en OneR y Random Forest junto con Speechiness que posee valores bajos en Random Forest y Correlacion de Pearson y así tener un modelo más flexible y menos sesgado, además es posible ver que Year junto con Duration y Danceability son las tres variables que mejor explicarían el modelo.

Bonus Cpa

En esta fueron construidas las gráficas de CPA con los datos numéricos por componentes (Ilustración 15 CPA componentes principales) y por componentes acumuladas (Ilustración 16 CPA componentes principales acumulada) con estos gráficos y mediante la inclinación de codo se llega a la conclusión que se utilizara hasta el componente 2, pero a modo de validación fue probado el componente acumulativo numérico (Ilustración 17 CPA según la proporción acumulativa) en donde se obtuvo hasta el componente 7, con lo que se puede apreciar que las variables obtenidas difieren a lo conseguido en el análisis realizado en el informe.

ANEXO

```
     Index          Title             Artist           Top_Genre             Year          BPM            Energy
Min.   :   1.0   Length:1994        Length:1994        Length:1994        Min.   :1956   Min.   : 37.0   Min.   :  3.00
1st Qu.: 499.2   Class :character   Class :character   Class :character   1st Qu.:1979   1st Qu.: 99.0   1st Qu.: 42.00
Median : 997.5   Mode  :character   Mode  :character   Mode  :character   Median :1993   Median :119.0   Median : 61.00
Mean   : 997.5                                                           Mean   :1993   Mean   :120.2   Mean   : 59.68
3rd Qu.:1495.8                                                           3rd Qu.:2007   3rd Qu.:136.0   3rd Qu.: 78.00
Max.   :1994.0                                                           Max.   :2019   Max.   :206.0   Max.   :100.00
 Danceability      Loudness         Liveness        Valence         Duration       Acousticness     Speechiness
Min.   :10.00   Min.   :-27.000   Min.   : 2.00   Min.   : 3.00   Min.   :  93.0   Min.   : 0.00   Min.   : 2.000
1st Qu.:43.00   1st Qu.:-11.000   1st Qu.: 9.00   1st Qu.:29.00   1st Qu.: 212.0   1st Qu.: 3.00   1st Qu.: 3.000
Median :53.00   Median : -8.000   Median :12.00   Median :47.00   Median : 245.0   Median :18.00   Median : 4.000
Mean   :53.24   Mean   : -9.009   Mean   :19.01   Mean   :49.41   Mean   : 262.4   Mean   :28.86   Mean   : 4.995
3rd Qu.:64.00   3rd Qu.: -6.000   3rd Qu.:23.00   3rd Qu.:69.75   3rd Qu.: 289.0   3rd Qu.:50.00   3rd Qu.: 5.000
Max.   :96.00   Max.   : -2.000   Max.   :99.00   Max.   :99.00   Max.   :1412.0   Max.   :99.00   Max.   :55.000
  Popularity
Min.   : 11.00
1st Qu.: 49.25
Median : 62.00
Mean   : 59.53
3rd Qu.: 71.00
Max.   :100.00
```

Ilustración 1 Screenshot de la consola.

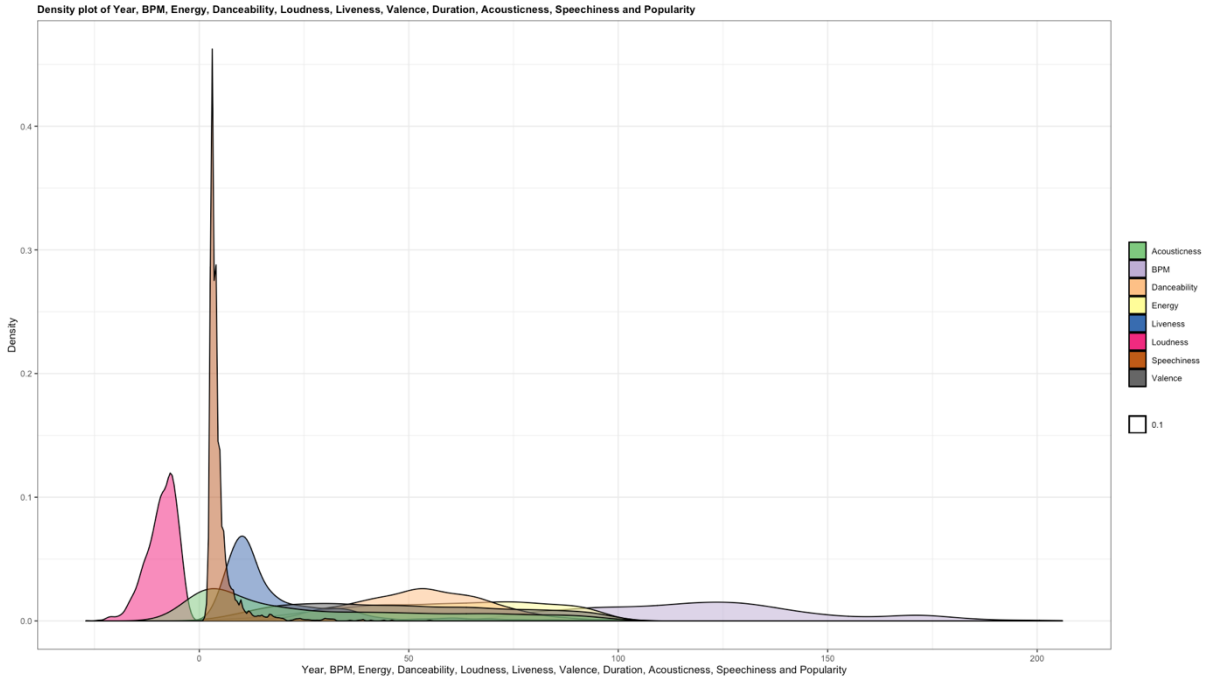

Ilustración 2 Densidad de las variables

Ilustración 3 Gráfico de tendencias

```
                Title                   Artist               Top_Genre        Year          BPM            Energy
Feeling Good      :   3   Queen              :  37   album rock      : 413   Min.   :1956   Min.   : 37.0   Min.   :  3.00
Hallelujah        :   3   The Beatles        :  36   adult standards : 123   1st Qu.:1979   1st Qu.: 99.0   1st Qu.: 42.00
One               :   3   Coldplay           :  27   dutch pop       :  88   Median :1993   Median :119.0   Median : 61.00
Always On My Mind:   2   U2                 :  26   alternative rock:  86   Mean   :1993   Mean   :120.2   Mean   : 59.68
Amsterdam         :   2   The Rolling Stones:  24   dance pop       :  83   3rd Qu.:2007   3rd Qu.:136.0   3rd Qu.: 78.00
Behind Blue Eyes  :   2   Bruce Springsteen :  23   dutch indie     :  75   Max.   :2019   Max.   :206.0   Max.   :100.00
(Other)           :1979   (Other)            :1821   (Other)         :1126
 Danceability      Loudness          Liveness        Valence         Duration       Acousticness     Speechiness
Min.   :10.00   Min.   :-27.000   Min.   : 2.00   Min.   : 3.00   Min.   :  93.0   Min.   : 0.00   Min.   : 2.000
1st Qu.:43.00   1st Qu.:-11.000   1st Qu.: 9.00   1st Qu.:29.00   1st Qu.: 212.0   1st Qu.: 3.00   1st Qu.: 3.000
Median :53.00   Median : -8.000   Median :12.00   Median :47.00   Median : 245.0   Median :18.00   Median : 4.000
Mean   :53.24   Mean   : -9.009   Mean   :19.01   Mean   :49.41   Mean   : 262.4   Mean   :28.86   Mean   : 4.995
3rd Qu.:64.00   3rd Qu.: -6.000   3rd Qu.:23.00   3rd Qu.:69.75   3rd Qu.: 289.0   3rd Qu.:50.00   3rd Qu.: 5.000
Max.   :96.00   Max.   : -2.000   Max.   :99.00   Max.   :99.00   Max.   :1412.0   Max.   :99.00   Max.   :55.000

  Popularity
Min.   : 11.00
1st Qu.: 49.25
Median : 62.00
Mean   : 59.53
3rd Qu.: 71.00
Max.   :100.00
```

Ilustración 4 Screenshot 2 de la consola

```
     Year            BPM            Energy       Danceability     Loudness        Liveness        Valence
Min.   :1956   Min.   : 37.0   Min.   :  3.00   Min.   :10.00   Min.   :-27.000   Min.   : 2.00   Min.   : 3.00
1st Qu.:1979   1st Qu.: 99.0   1st Qu.: 42.00   1st Qu.:43.00   1st Qu.:-11.000   1st Qu.: 9.00   1st Qu.:29.00
Median :1993   Median :119.0   Median : 61.00   Median :53.00   Median : -8.000   Median :12.00   Median :47.00
Mean   :1993   Mean   :120.2   Mean   : 59.68   Mean   :53.24   Mean   : -9.009   Mean   :19.01   Mean   :49.41
3rd Qu.:2007   3rd Qu.:136.0   3rd Qu.: 78.00   3rd Qu.:64.00   3rd Qu.: -6.000   3rd Qu.:23.00   3rd Qu.:69.75
Max.   :2019   Max.   :206.0   Max.   :100.00   Max.   :96.00   Max.   : -2.000   Max.   :99.00   Max.   :99.00
    Duration       Acousticness     Speechiness      Popularity
Min.   :  93.0   Min.   : 0.00   Min.   : 2.000   Min.   : 11.00
1st Qu.: 212.0   1st Qu.: 3.00   1st Qu.: 3.000   1st Qu.: 49.25
Median : 245.0   Median :18.00   Median : 4.000   Median : 62.00
Mean   : 262.4   Mean   :28.86   Mean   : 4.995   Mean   : 59.53
3rd Qu.: 289.0   3rd Qu.:50.00   3rd Qu.: 5.000   3rd Qu.: 71.00
Max.   :1412.0   Max.   :99.00   Max.   :55.000   Max.   :100.00
```

Ilustración 5 Screenshot 3 de la consola

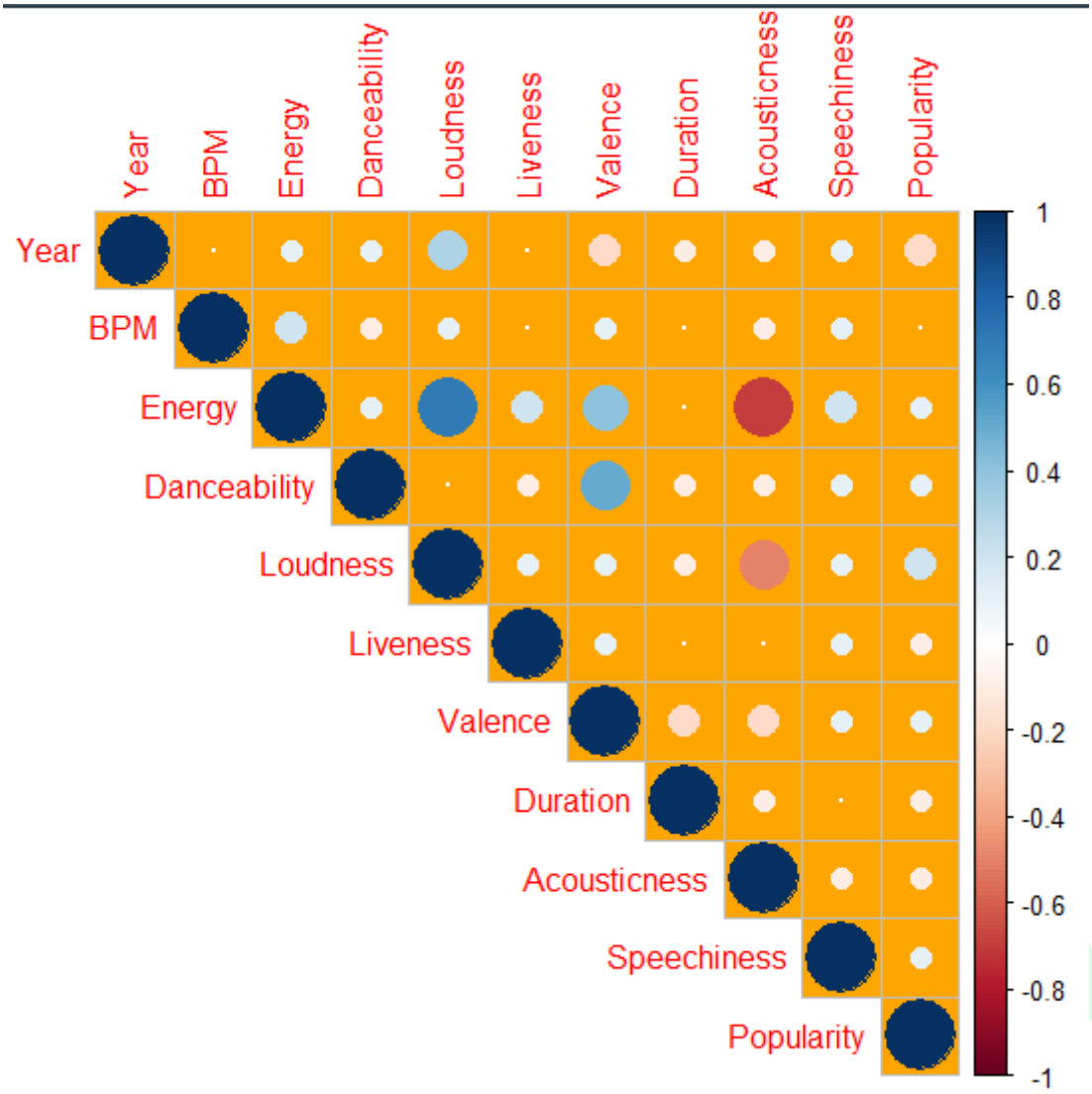

Ilustración 6 Matriz de correlaciones circular

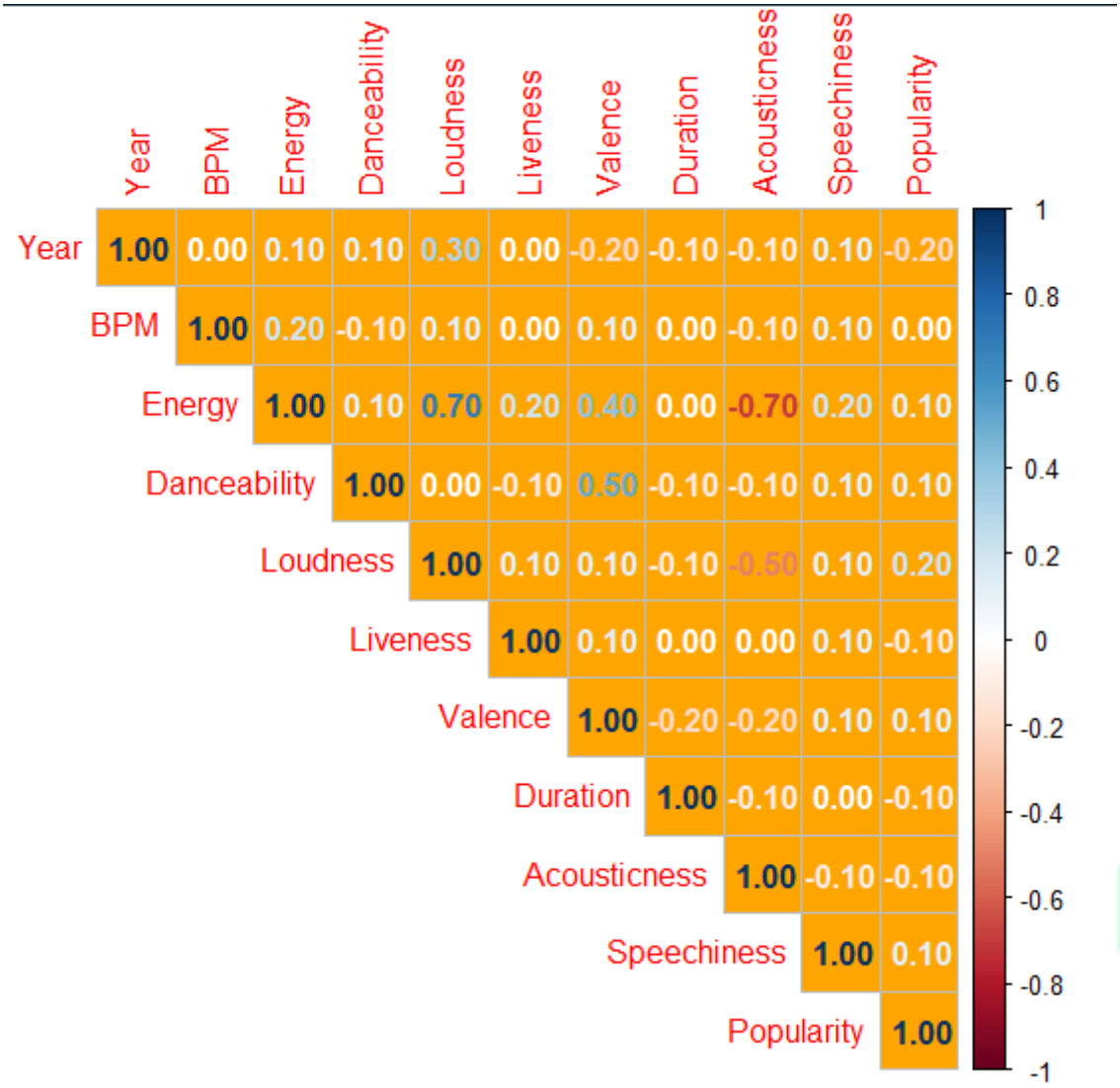

Ilustración 7 Matriz de correlación numérica

```
             attr_importance
Year              0.158962019
BPM               0.003181354
Energy            0.103392998
Danceability      0.144344281
Loudness          0.165526880
Liveness          0.111977778
Valence           0.095910821
Duration          0.065402740
Acousticness      0.087604272
Speechiness       0.111688785
> subset <- cutoff.k(weights, 5)
> f <- as.simple.formula(subset, "Popularity")
> print(f)
Popularity ~ Loudness + Year + Danceability + Liveness + Speechiness
<environment: 0x000001fbc10082f8>
```

Ilustración 8 Correlación Pearson importancia y selección

```
> weights <- oneR(Popularity~., Sp)
> print(weights)
              attr_importance
Title              0.79989970
Artist             0.45987964
Top_Genre          0.24172518
Year               0.08375125
BPM                0.01554664
Energy             0.01554664
Danceability       0.01554664
Loudness           0.05867603
Liveness           0.01554664
Valence            0.01554664
Duration           0.01554664
Acousticness       0.01554664
Speechiness        0.01554664
> subset <- cutoff.k(weights, 5)
> f <- as.simple.formula(subset, "Popularity")
> print(f)
Popularity ~ Title + Artist + Top_Genre + Year + Loudness
<environment: 0x000001fbb00fcf80>
```

Ilustración 9 Selección de atributos One r

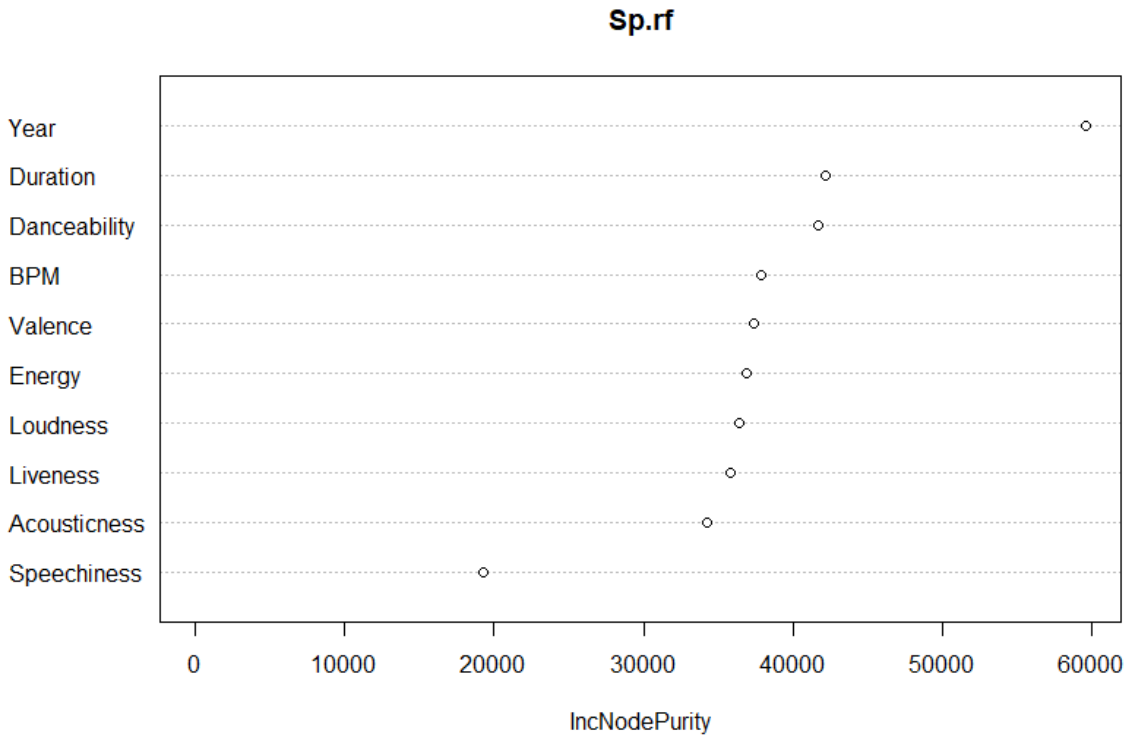

Ilustración 10 Selección de atributos Random Forest

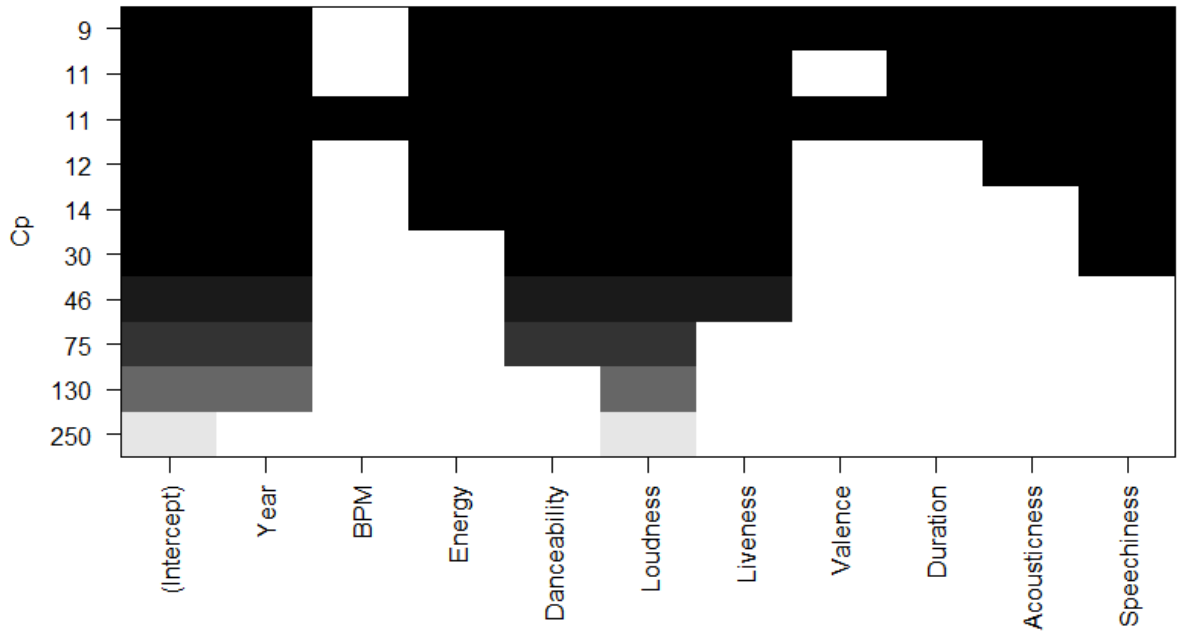

Ilustración 11 Reg lineal según Cp

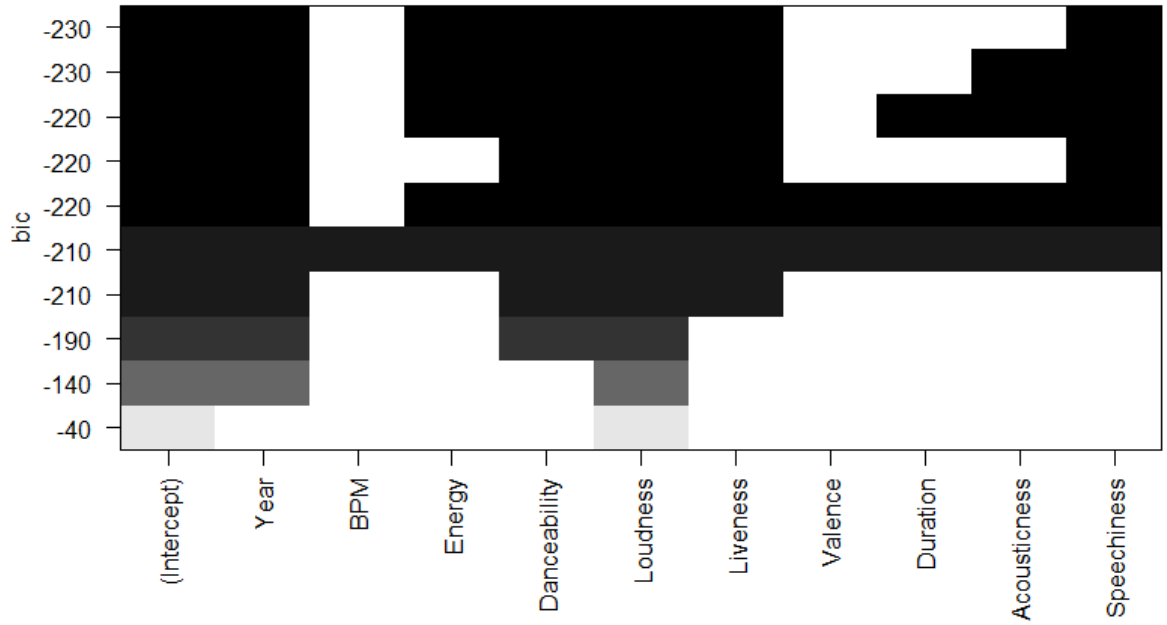

Ilustración 12 Reg lineal según BIC

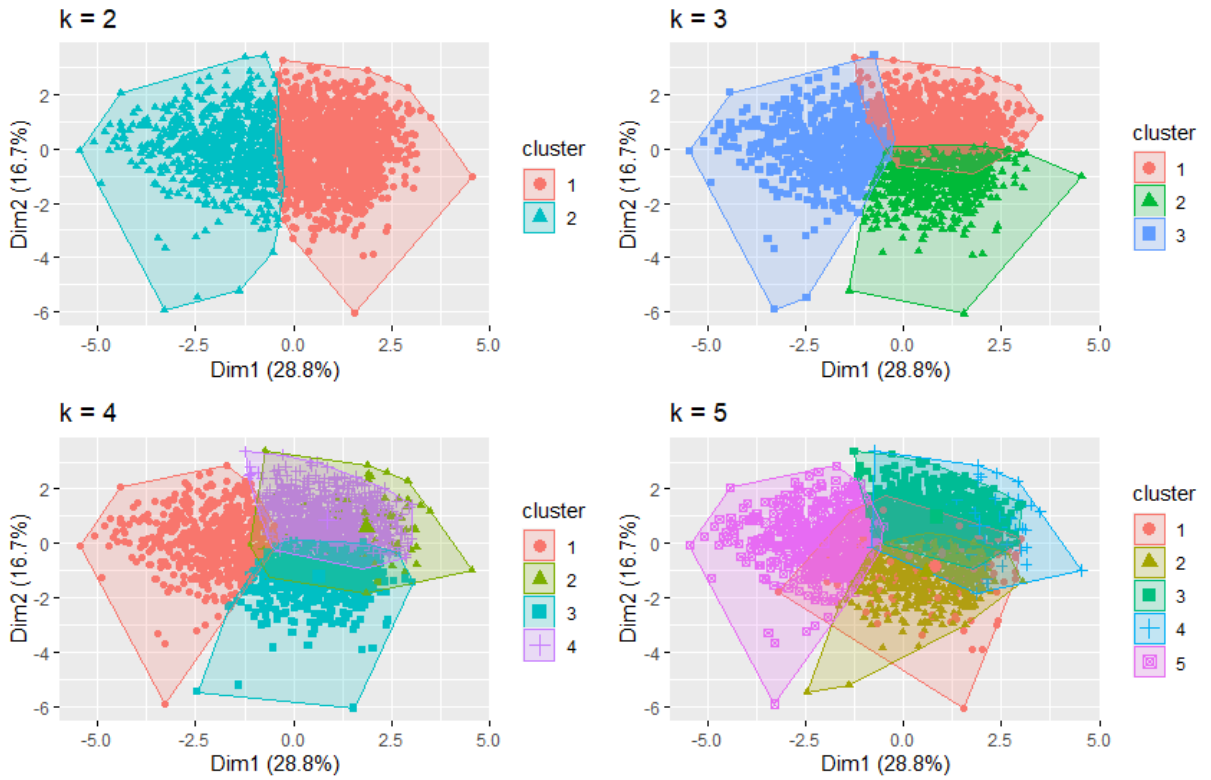

Ilustración 13 Cluster con diferentes K gráfico

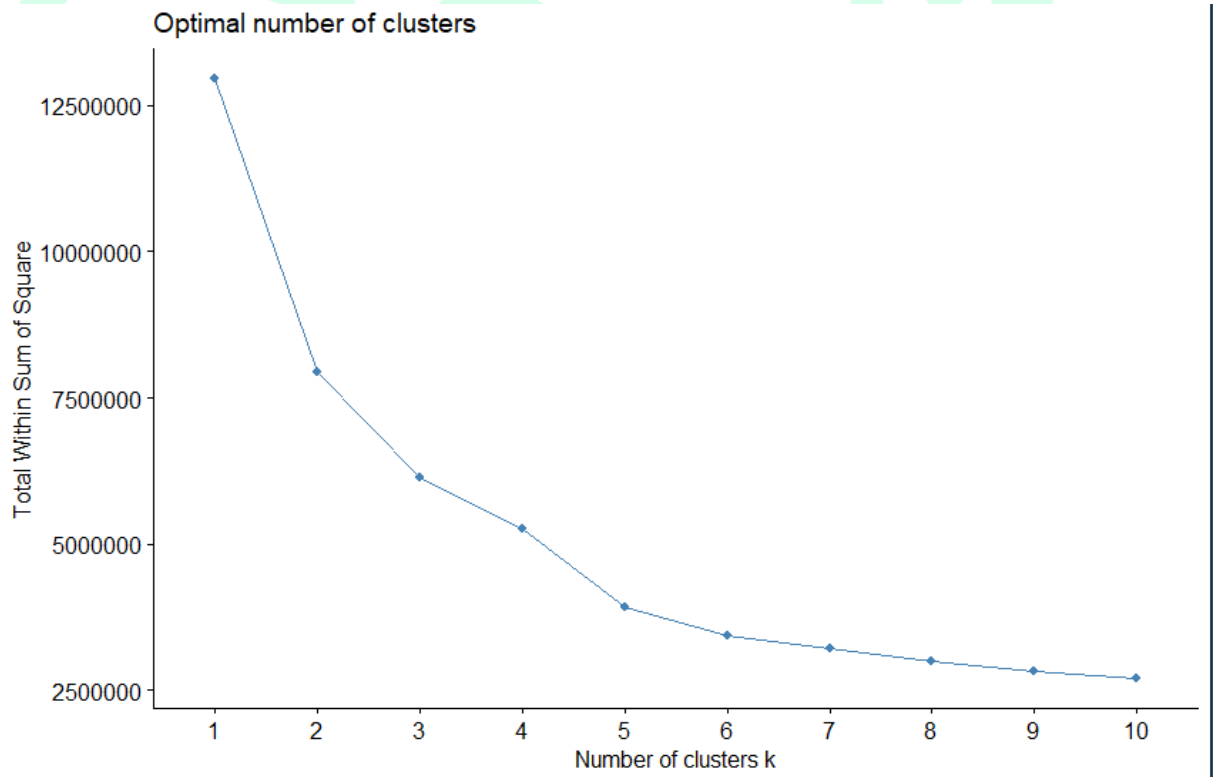

Ilustración 14 Número optimo de clusters

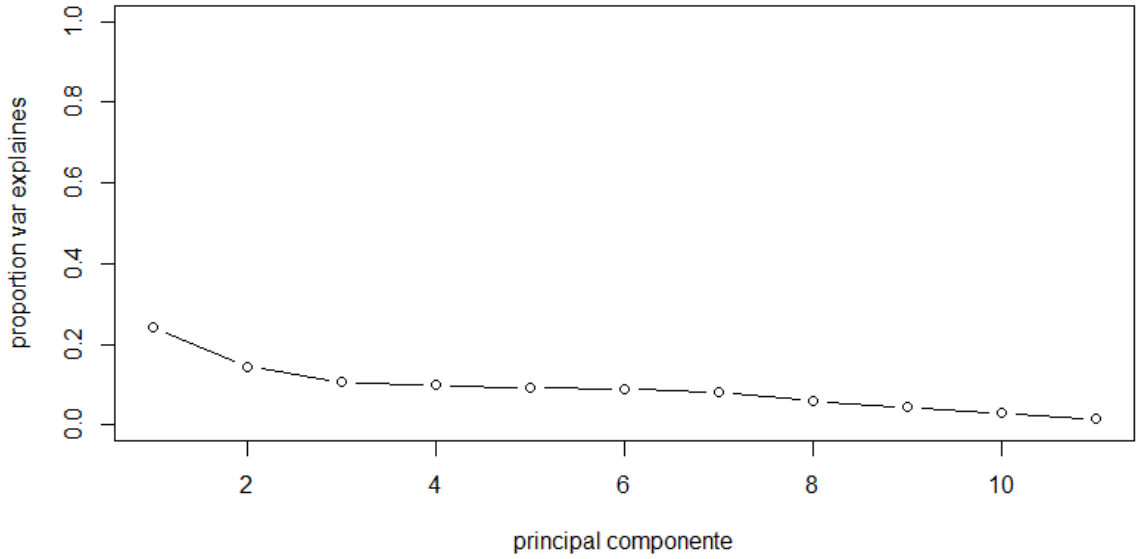

Ilustración 15 CPA componentes principales

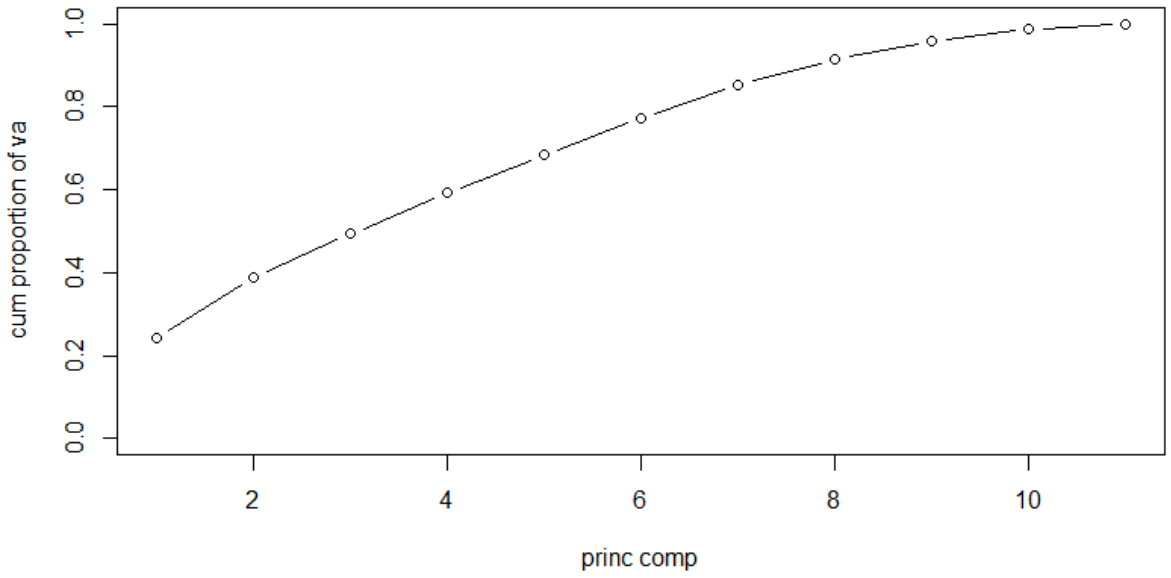

Ilustración 16 CPA componentes principales acumulada

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10
Standard deviation     1.6340 1.2606 1.0772 1.04015 1.01260 0.98302 0.94872 0.81027 0.69426 0.56350
Proportion of Variance 0.2427 0.1445 0.1055 0.09836 0.09321 0.08785 0.08182 0.05969 0.04382 0.02887
Cumulative Proportion  0.2427 0.3872 0.4927 0.59104 0.68425 0.77210 0.85392 0.91361 0.95743 0.98629
                          PC11
Standard deviation     0.38829
Proportion of Variance 0.01371
Cumulative Proportion  1.00000
```

Ilustración 17 CPA según la proporción acumulativa

Base de Datos a utilizar: [Spotify - All Time Top 2000s Mega Dataset]