



MÁSTER DE ESTADÍSTICA APLICADA PARA LA CIENCIA DE DATOS CON R SOFTWARE

¿Se pueden diferenciar los géneros
musicales a través de variables físicas y
la minería de datos?

AUTOR: Manuel López Prieto

DIRECTOR: Juan Luis López

FECHA: 26/07/2022

ENTIDAD COLABORADORA



RESUMEN

Este trabajo de investigación nace como nexo de unión de las dos grandes pasiones del autor: música y ciencia de datos.

El objetivo era saber si es posible diferenciar los géneros musicales (*rock, jazz, clásica, hip-hop...*) a través de ciertas variables físicas fácilmente obtenibles de cada grabación (*tempo, centroide espectral, intervalo de longitud de onda* entre otras).

La población de estudio consistió en 1000 grabaciones categorizadas en 10 estilos diferentes con 100 canciones por cada género. De cada una de ellas se habían extraído 9 atributos físicos a través de un software gratuito.

Los modelos estadísticos que se aplicaron han sido **k-medias** y **PAM** (Partitioning Around Medoids) respecto a estudio de partición por clúster y Análisis de Componentes Principales o **PCA** por sus siglas en inglés respecto a reducción de dimensionalidad, para concluir uniendo ambos acercamientos mediante una Agrupación Jerárquica de Componentes Principales (**HCPC**)

Finalmente se ha conseguido aislar los estilos más “añejos” (*clásica, jazz y blues*) de los más modernos (*metal, disco, rock, pop y hip-hop*).

Palabras Claves: Estadística, Clúster, K-Medias, PAM, PCA, HCPC, Música, Géneros Musicales.

AGRADECIMIENTOS

*A mis **padres**, por ser ejemplo motivante de buscar siempre la superación personal, por haberme formado en valores de respeto, amor y esfuerzo y por haberme enseñado que lo más importante que se lleva uno en la vida es la educación que recibe.*

*A mi mujer, **Silvana**, por estar siempre ahí, apoyándome, ayudándome, animándome, ~~aguantándome~~, amándome...en definitiva por ser ELLA.*

*A mi hermano, **Pitu**, por ser mi persona en el mundo y por la cantidad de música que hemos aprendido el uno del otro.*

*A mis perr@s, l@s que cruzaron el arcoíris **Wallace** y **Rosita** y la que me aguanta ahora mismo, **Frida**, por ser fieles compañer@s en el día a día y pacientes testig@s de todas mis locuras estudiantiles.*

*A **tod@s** las personas que alguna vez han compartido conmigo algún concierto, salida, disco, charla musical (en foro o en persona), porque tod@os vosotr@s sois culpables de haber aumentado los (ya sin) límites estilísticos de mi gran pasión.*

*A **Juan Luis López** y **Rosana Ferrero** en concreto y al equipo de **Máxima Formación** (con un especial recuerdo a **Alfonso Lara** por las gestiones) por haber dado forma a un máster verdaderamente apasionante, retador y a la vez entretenido. Me llevo parte de vuestro conocimiento conmigo y eso os deja a vosotros parte de mi corazón.*

ÍNDICE

RESUMEN	2
AGRADECIMIENTOS.....	3
LISTA DE FIGURAS.....	6
CAPÍTULO 1: INTRODUCCIÓN.....	7
1.1 Estado del arte	7
1.1.1 Aportación de este estudio al estado del arte	8
1.2 Finalidad y Objetivos de la Investigación	8
1.2.1 Perspectiva Personal	8
1.2.2 Perspectiva sociológica	9
1.2.3 Perspectiva musical.....	10
CAPÍTULO 2: MATERIAL Y MÉTODOS	12
2.1 Tipo de investigación.....	12
2.2 Definición de la población.....	12
2.3 Definición de las variables.....	13
2.5 Métodos usados	14
2.5.1 k-medias	14
2.5.2 PAM	16
2.5.3 PCA	18
2.5.4 HCPC.....	22
CAPÍTULO 3: ESTUDIO DE K-MEDIAS.....	24
3.1 Resultados y discusión	24
3.1.1 Análisis exploratorio de datos.....	24
3.1.2 Análisis de tendencia de la agrupación	26
3.1.3 Elección del número óptimo de grupos	27
3.1.4 Representación de los grupos.....	28
3.1.5 Importancia de las variables	29
3.1.6 Comparación de los grupos con los géneros originales	29
3.1.7 Validación	31
CAPÍTULO 4: ESTUDIO DE PAM	32
4.1 Resultados y discusión	32
4.1.1 Análisis exploratorio de datos.....	32
4.1.2 Análisis de tendencia de la agrupación	32
4.1.3 Elección del número óptimo de grupos	32
4.1.4 Comparación de los grupos con los géneros originales	32

4.1.5 Importancia de las variables	33
4.1.6 Validación	34
CAPÍTULO 5: ESTUDIO DE PCA	36
5.1 Resultados y discusión	36
5.1.1 Análisis exploratorio de datos.....	36
5.1.2 Estudio de si es adecuado el PCA en nuestro set de datos.....	36
5.1.3 Selección del número de dimensiones.....	36
5.1.4 Estudio de las variables (características).....	37
5.1.5 Estudio de los individuos (casos).....	41
CAPÍTULO 6: HCPC.....	43
CAPÍTULO 7: CONCLUSIONES	45
7.1 Conclusiones k-means.....	45
7.2 Conclusiones PAM	46
7.3 Conclusiones PCA	47
7.4 Conclusiones HCPC.....	47
7.5 Conclusiones generales.....	48
7.6 Hipotética investigación futura	49
REFERENCIAS BIBLIOGRÁFICAS	50
CAPÍTULO 8: ANEXOS	51
8.1 Código R para k-medias.....	51
8.2 Código R para PAM.....	56
8.3 Código R para PCA.....	60
8.4 Código R para HCPC.....	65
8.5 Figuras k-means.....	69
8.6 Figuras PAM	71
8.7 Figuras PCA.....	73
8.8 Figuras HCPC	78

LISTA DE FIGURAS

Figura 1. Estudio outliers (Fuente: elaboración propia)	24
Figura 2. Label vs spectral_centroid (Fuente: elaboración propia).....	25
Figura 3. Label vs beats (Fuente: elaboración propia)	26
Figura 4. Validación VAT tendencia de agrupación (Fuente: elaboración propia).....	27
Figura 5. Comparativa clúster/label y % k-means k = 10 (Fuente: elaboración propia)	28
Figura 6. Valores grupos k-means k=2 (Fuente: elaboración propia)	29
Figura 7. Comparativa clúster/label y % k-means k = 2 (Fuente: elaboración propia)	30
Figura 8. Validación k-means k = 2 (Fuente: elaboración propia).....	31
Figura 9. Comparativa clúster/label y % PAM (Fuente: elaboración propia).....	33
Figura 10. Valores grupos PAM (Fuente: elaboración propia)	34
Figura 11. Validación PAM (Fuente: elaboración propia)	35
Figura 12. Coordenadas: mapa variables 1 (Fuente: elaboración propia)	37
Figura 13. Calidad representación PCA dim1 (Fuente: elaboración propia)	38
Figura 14. Calidad representación PCA dim2 (Fuente: elaboración propia).....	39
Figura 15. Contribución a la representación PCA dim1 (Fuente: elaboración propia)	40
Figura 16. Contribución a la representación PCA dim2 (Fuente: elaboración propia)	40
Figura 17. Coordenadas: mapa de individuos 2(Fuente: elaboración propia).....	41
Figura 18. Coordenadas: mapa de individuos 3 colores (Fuente: elaboración propia)	42
Figura 19. Individuos en el mapa tras HCPC.....	43
Figura 20. Comparativa clúster/label y % HCPC (Fuente: elaboración propia).....	44
Figura 21. Validación relativa k-means (Fuente: elaboración propia)	69
Figura 22. Criterio estabilidad k-means (Fuente: elaboración propia)	70
Figura 23. Representación grupos k-means k=2 (Fuente: elaboración propia)	70
Figura 24. Centroides grupos k-means k=2 (Fuente: elaboración propia).....	70
Figura 25. Importancia variables k-means k=2 (Fuente: elaboración propia)	71
Figura 26. Número óptimo grupos PAM (Fuente: elaboración propia)	72
Figura 27. Importancia variables PAM (Fuente: elaboración propia)	73
Figura 28. Esfericidad Bartlett (Fuente: elaboración propia).....	73
Figura 29. KMO (Fuente: elaboración propia).....	74
Figura 30. Determinante matriz correlación (Fuente: elaboración propia).....	74
Figura 31. Autovalores (Fuente: elaboración propia)	74
Figura 32. Gráfico sedimentación (Fuente: elaboración propia)	75
Figura 33. Análisis paralelo Horn (Fuente: elaboración propia)	75
Figura 34. Coordenadas: mapa variables 2 (Fuente: elaboración propia)	76
Figura 35. Descripción dimensiones PCA dim1 (Fuente: elaboración propia)	76
Figura 36. Descripción dimensiones PCA dim2(Fuente: elaboración propia).....	77
Figura 37. Coordenadas: mapa de individuos 1(Fuente: elaboración propia).....	77
Figura 38. Coordenadas: mapa de individuos 10 colores (Fuente: elaboración propia)	78
Figura 39. Variables cuantitativas que mejor definen los grupos.....	78
Figura 40. Dimensiones Principales más Asociadas con los Grupos	79
Figura 41. Individuos más Representativos de Cada Grupo	80
Figura 42. Individuos de Cada Grupo más Alejados de los Otros Grupos	80

CAPÍTULO 1: INTRODUCCIÓN

1.1 Estado del arte

Queda patente en el trabajo de Sturm [1], que el reconocimiento del género musical (Musical Genre Recognition o MGR) ha sido amplio objeto de estudio: 179 tesis de máster o tesis doctorales de 2010 a 2012 (año de la publicación de este estudio).

A pesar de ello, el autor concluía que el MGR **sigue siendo un problema difícil de resolver por una máquina**.

De 2012 en adelante, gracias a las increíbles mejoras computacionales, el número de trabajos al respecto se ha multiplicado. De manera contemporánea cabe destacar a Glenn McDonald [2] y [3], responsable de la categorización por géneros para el gigante empresarial y dominador del mercado de música por streaming, Spotify.

Respecto a estudios que hayan usado alguna(s) de la(s) variables que estamos considerando aquí, podríamos destacar a Ghosal, Chakraborty, Chandra y Kumar [4], quienes, basándose en los *MFCC*, clasifican primero música con voz y música instrumental, posteriormente son capaces de aislar las señales instrumentales en función de cuatro tipos: cuerda, viento-madera, percusión y teclado y finalmente consideran la tarea de clasificar por género: *pop*, *jazz* o *bhangra* (género indio).

Mediante un acercamiento jerárquico, consiguen una **precisión** de clasificación máxima del 92.71% para voz-instrumental, 93.5% para tipo de instrumento (destacando por encima percusión y por debajo cuerdas) y un **84.6%** respecto a la diferenciación por los tres géneros.

Incluso también considerando *MFCC*, pero además usando nuestro mismo set de datos, Li & Chan [5] concluyen que los coeficientes espectrales de **Mel codifican tanto información tímbrica como tonal**, advirtiéndolo por tanto del peligro que tendría intentar clasificar por géneros canciones muy influenciadas por las tonalidades dominantes de su estilo. Calibrando los datos durante las fases de entrenamiento con transformaciones de tonalidad y tono, mejoran la clasificación de sus experimentos.

Banitalebi-Dehkordi & Banitalebi-Dehkordi [6], usando el mismo dataset y quedándose con las variables *mfcc*, *spectral_centroid*, *roll_off*, *zero_crossing_rate*, *spectral_bandwidth* y *spectral_centroid*, proponen un **nuevo método de extracción de**

características que combine las características a corto plazo y las características a largo plazo (extraídas de la combinación de las características a corto plazo).

Logran un **95.7% de precisión**, mejorando los resultados de similares estudios que se habían realizado basados en Preserving Non-Negative Matrix Factorization (TPNMF), Non-Negative Tensor Factorization (NTF), Multi Linear Principal Component Analysis (MPCA) o General Tensor Discriminate Analysis (GTDA).

Por último, pero no menos importante, puesto que la manera en la que encontré el dataset fue a través de un desafío de Kaggle [7], no quisiera terminar sin mencionar algunos de los estudios que diferentes compañeros de causa han realizado al respecto, como Elena Ignat (**PCA** de 2 y 3 dimensiones), Chadwickcasper (usando redes neuronales y algoritmos de **Deep Learning** logrando un 61% de precisión) o LC Vasconcelos (usando support-vector machines o **SVM** con un 62% de precisión).

1.1.1 Aportación de este estudio al estado del arte

Ante la muy amplia y abrumadora literatura al respecto, el presente trabajo pretende aportar al estado del arte del problema de clasificar por géneros un estudio donde se justifica científicamente cada decisión tomada, se explica detalladamente cada paso y es absolutamente **reproducible**. Usando primero **k-medias**, luego **PAM** (debo decir que no he encontrado ningún otro estudio que haga uso de este algoritmo) y posteriormente **PCA**. Concluimos con un **HCPC**, estudio que tampoco he encontrado en la bibliografía existente.

1.2 Finalidad y Objetivos de la Investigación

Seguramente el lector pueda hacerse la pregunta ¿es necesario o tiene cabida un estudio estadístico acerca de si es posible (o no) distinguir científicamente los géneros musicales?

Se puede responder en afirmativo desde diferentes perspectivas

1.2.1 Perspectiva Personal

1. La más importante para mí, pero seguramente la menos válida para el resto: el trabajo que presento nace como nexo de unión entre mis dos grandes pasiones, a saber, la música y la ciencia de datos.

Musicalmente, según me adentraba en este maravilloso mundo, he ido defendiendo cada vez más la necesidad de abolir los estilos en cuanto a los gustos. No me siento loco, sino al contrario afortunado por ser capaz de disfrutar de gente tan dispar como *Rosalía, Solange, Built to Spill, The Mystery of the Bulgarian Voices, Primal Scream, Little Simz, James Blake, Stereolab, Nilüfer Nanya, Frank Carter & The Rattlesnakes, Danny L Harle o Ritchie Hawtin*.

Lo mencionado no es azaroso ni una pedante manera de nombrar diferentes artistas (que ni siquiera entran entre mis favoritos), sino que fue mi ruta de un día de los muchos festivales a los que he asistido (concretamente, sábado 1 de junio, Primavera Sound 2019). Y pretende evidenciar que, sin sonrojo, pasé del rock a lo clásico, o de lo popero a lo electrónico no solamente sin ningún tipo de complejo, sino que todos y cada uno de los espectáculos fueron (casi) igualmente disfrutados.

Una primera motivación, por tanto, es justificar mi absoluta falta de prejuicio musical y animar a quien se autolimita a unos pocos géneros (comentarios tipo “a mí lo que me gusta es el rock”) a que simplemente se deje llevar y sea capaz de disfrutar sin etiquetas de este maravilloso arte que es la música.

Así cabe reconocer que, al comenzar el estudio, había una perversa intención por mi parte de evidenciar que, eso que se llaman estilos, eso que hace escuchar o no una canción, grupo o artista, no tiene ningún tipo de sentido, al menos en cuanto a las variables físicas que pueden definir una pieza. Quien siga adelante con la lectura, sabrá si la Estadística fue o no mi aliada en este sentido.

2. En la misma línea fue una gran motivación el leer estudios psicológicos como el de Greenberg, Kosinski, Stillwell y más [8], donde se propone una nueva clasificación. “Los géneros musicales están desfasados, pero este nuevo sistema explica por qué pueden gustarte tanto el jazz como el hip hop”. Se discuten los problemas de etiquetar la música por géneros y se introduce una nueva clasificación y estudio estadístico en función de tres variables: “**Arousal**” (el nivel de energía de la música); “**Valence**” (el espectro que va de las emociones tristes a las alegres en la música); y “**Depth**” (la cantidad de sofisticación y profundidad emocional en la música). Y según los 5 grados de personalidad (*openness, conscientiousness, extraversion, agreeableness and neuroticism*) se estudia que un sujeto podría preferir canciones con valores más altos o más bajos de cada una de aquellas tres variables.

Una nueva categorización por tanto de la música más sensitiva y quizás más válida.

1.2.2 Perspectiva sociológica

Pero no podemos dejar de lado que, finalmente, el género puede ser tanto puerta de entrada a la música como albergar internamente multitud de componentes sociológicos de adhesión a uno de ellos: cada generación ha tomado por bandera el

nuevo estilo de su época, normalmente en conflicto con el estilo (y por tanto la generación) anterior.

Lo ha hecho suyo y le ha servido para construir no solamente un bagaje cultural sino para formar una identidad a través de relaciones sociales con semejantes.

No únicamente a nivel generacional, sino a niveles más profundos como la raza, la música y algún género en concreto ha servido para fomentar los lazos de unión, denunciando en muchos casos, injusticias y convirtiéndose algunas canciones en himnos intraculturales. Pienso por ejemplo en el hip-hop para la raza negra.

Pero se cae en el peligro de que esta separación precisamente haga eso: separar (culturas, razas, edades...). En esta línea encontramos libros como el de Laurie, [9] donde se dice que “Escribir sobre la cultura popular siempre implica inclusiones y exclusiones. En los estudios sobre la música popular, estas inclusiones y exclusiones se realizan a menudo de forma tácita mediante el uso de taxonomías de género, que van desde lo más amplio (*pop*, *rock*, *jazz*) hasta lo más específico (el sonido de *Gotemburgo*, *Hi-NRG*, *Krautrock*). Desde principios de la década de 1980, el género ha pasado de ser un subconjunto de los estudios sobre la música popular a ser un marco casi omnipresente para constituir y evaluar los objetos de investigación musical”.

Además, el autor critica diferentes libros orientados a géneros musicales y su cultura subyacente “Los estudios aquí considerados tienden a excluir de antemano a los oyentes que no pertenecen a una comunidad musical fija, al igual que pasan por alto a las personas que participan en una comunidad orientada a la música sin tener ningún compromiso ontológico con los gustos musicales de sus compañeros.”

Es decir, reconociendo la importancia del estilo, aboga por la limitación que provoca tanto en el crítico como en el oyente esa categorización.

Así que otra de las motivaciones del estudio es saber si tiene razón el adolescente de la generación de turno, cuando se siente parte de un grupo y además diferente del grupo anterior. Y si muchas de las críticas a discos o canciones que nacen desde una categorización estilística tienen sentido o se debiera incidir más en otros factores.

1.2.3 Perspectiva musical

Pero el hecho es que categorizamos las canciones por géneros, porque eso facilita la búsqueda al oyente: le decimos a la aplicación musical que estamos usando (*Spotify*, *Tidal*, *Amazon Music*...) “quiero escuchar pop”, por ejemplo, y nos devolverá multitud de canciones pertenecientes a esa categoría.

Mientras escribo, hago lo anteriormente mencionado, y la app me propone cosas tan dispares como *Harry Styles*, *The Weeknd*, *Adele*, *Lorde*, *J. Balvin* o *Dani Martín*. ¿Tan parecidos son? Aún más, siguiendo con el ejemplo, el grupo pop por antonomasia

siempre ha sido *The Smiths* y el trono de rey del pop tradicionalmente se le atribuye a *Michael Jackson*. ¿Tiene sentido que ambos pertenezcan al mismo clúster?

Ahí hablamos de sentido dentro del grupo, pero otra perspectiva es dilucidar cuán alejado estaría ese grupo de otro. Por ejemplo, *Queen* se considera una banda de rock. ¿Tan alejado está su rock de otras propuestas más cercanas al pop?

Seguramente este estudio se pueda hacer usando diferentes variables. Aquí aportamos nuestro granito de arena a la causa trabajando con variables físicas como *tempo*, *spectral_centroid* o *zero_crossing_rate*.

En definitiva, todas las perspectivas anteriores parten del hecho de que etiquetar siempre ha sido una necesidad humana: la etiqueta ha servido para clasificar la realidad ya sean personas, objetos o situaciones. Y se ha tendido a catalogar el mundo para poder tener una percepción de él. Pero lo que en un principio puede ser positivo, puede acabar en dogmas y estigmatizaciones.

La música evidentemente no es ajena a ello, así que la motivación fundamental de este estudio es el saber si la física (las variables de las canciones) y la estadística apoyan o no esa categorización que necesariamente hemos creado para poder entenderla como arte.

CAPÍTULO 2: MATERIAL Y MÉTODOS

2.1 Tipo de investigación

Este estudio parte de ciertas variables físicas (descritas más abajo) extraídas de un conjunto de 1000 canciones etiquetadas por géneros y que se puede descargar en [7].

Puesto que el objetivo es saber si esas variables físicas nos pueden servir para agrupar los estilos, las técnicas usadas han sido de **Análisis de Agrupación** (o análisis clúster) y de **Análisis de Componentes Principales** (PCA). Finalmente uniremos ambos acercamientos con un breve estudio de **Agrupación Jerárquica de Componentes Principales** (HCPC).

Todo ello en **lenguaje R** junto a su entorno de desarrollo integrado o IDE más popular, **RStudio**.

Haciendo uso de multitud de librerías, pero destacando **cluster**, **factoextra**, y **FactoMineR**.

Fuente de inspiración han sido tanto los **temas 7** (Análisis de agrupación clúster) [10] y **9** (Ordenación y Reducción de la dimensionalidad) [11] del **Máster** de Estadística Aplicada Para la Ciencia de Datos Con R Software, cursado en Máxima Formación y acreditado con el Título Propio de la Universidad Nebrija de Madrid, (Ferrero) como los dos libros de Multivariate Analysis de **Alboukael Kassambara**, **Practical Guide To Cluster Analysis in R** [12] y **Practical Guide To Principal Component Methods in R** [13].

2.2 Definición de la población

La **población** de esta investigación proviene del dataset conocido como GTZAN que es el conjunto de datos público más utilizado para la evaluación en la investigación de la escucha automática para el reconocimiento de géneros musicales (MGR).

Nace del ISMIR (International Society for Music Information Retrieval) [14], una organización sin ánimo de lucro cuyo objetivo es hacer avanzar la investigación en el campo de la recuperación de información musical (MIR por sus siglas en inglés Music Information Retrieval), un campo cuyo objetivo es el desarrollo de herramientas computacionales para el procesamiento, la búsqueda, la organización y el acceso a datos relacionados con la música.

Entre otras cosas, la sociedad ISMIR fomenta el intercambio de ideas y actividades entre sus miembros, estimula la investigación y la educación en MIR, apoya y fomenta

la diversidad de miembros y disciplinas, y supervisa la organización de la conferencia anual ISMIR.

Los archivos se recopilaron en 2000-2001 a partir de diversas fuentes, como CD personales, radio y grabaciones de micrófono, con el fin de representar una amplia variedad de condiciones de grabación.

Se compone de *1000 pistas de audio* de 30 segundos de duración cada una.

Todas ellas son archivos de audio mono de 16 bits a 22050Hz en formato .wav.

Las canciones ya las había seleccionado el creador del dataset. Se puede tener acceso a ellas, pero no se aportan los nombres de las piezas ni los autores, debido a derechos de copyright.

2.3 Definición de las variables

Los audios vienen etiquetados por géneros. Hemos manipulado el set de datos para transformar esa variable en etiqueta a la que hemos llamado *label*.

Sus categorías son: *blues, classical, country, disco, hiphop, jazz, metal, pop, reggae y rock*.

El resto de las variables fueron extraídas por el autor del set de datos con la librería de Python librosa [15].

Variable	Descripción
tempo	El ritmo o la velocidad del pasaje musical
beats	Unidad rítmica en música
chroma_stft	Transformada de Fourier de Tiempo Corto (Short-time Fourier transform, STFT) está relacionada con la transformada de Fourier usada para determinar el contenido en frecuencia sinusoidal y de fase en secciones locales de una señal, así como sus cambios con respecto al tiempo.
rmse	Error cuadrático medio, en el caso de señales de audio, íntimamente ligado a la energía que en su caso sería el volumen de la señal
spectral_centroid	El centroide espectral es una medida que indica dónde está el "centro de masa" del espectro. Respecto a la música, relacionado con la percepción de "brillo" que nos puede dar la melodía y por lo tanto se usa para caracterizar el timbre musical.
spectral_bandwidth	Es el intervalo de longitudes de onda en el que una cantidad espectral radiada no es inferior a la mitad de su valor máximo

rolloff	Se refiere a la acción de un tipo específico de filtro; uno diseñado para atenuar las frecuencias por encima o por debajo de un determinado punto. Se llama roll off porque el proceso es gradual.
zero_crossing_rate	La tasa de paso por cero (ZCR) es la velocidad a la que una señal pasa de positivo a cero a negativo o de negativo a cero a positivo. Su valor se ha utilizado ampliamente tanto en el reconocimiento del habla como en la recuperación de información musical para clasificar los sonidos percusivos.
mfccn	Mel-frequency cepstral coefficients o coeficientes cepstrales de frecuencia e Mel, son coeficientes para la representación del habla basados en la percepción auditiva humana.

Tabla 1: Descripción de variables

2.5 Métodos usados

2.5.1 k-medias

El algoritmo de **k-medias** [10] crea un número de grupos k especificados por el usuario, que minimiza la suma de cuadrados de las observaciones a los centros de agrupación asignados.

Pero ¿qué es un **clúster**?

- Es un conjunto (o grupo) de observaciones similares a otras observaciones (dentro del mismo grupo) y que son distintas a las observaciones que se encuentran en otros grupos.
- Es decir, que dentro del clúster tendremos observaciones con características similares (o relacionadas)
- Y entre clústeres observaciones con características diferentes (o no relacionadas).

Concretamente, **k-medias** es un **método iterativo** que:

- Primero selecciona k puntos aleatorios como centroides iniciales.
- Entonces crea k grupos asignando cada observación a su centroide más cercano.
- Ahora calcula los centroides como el punto medio de cada clúster.
- Vuelve a reasignar cada observación a los nuevos centroides.

- Repite los dos pasos anteriores hasta que se cumpla cierto criterio de convergencia (usualmente que las observaciones no sean más reasignadas, aunque también se puede utilizar como criterio un número máximo de iteraciones. En R es 10).

Previo al estudio deberemos:

- Resumir los datos e identificar los valores ausentes (**NAs**).
- Identificar valores atípicos (**outliers**), ya que este estudio es muy sensible a ellos debido a que se basa en la media.
- **Estandarizar** los datos: la escala en la que se miden las variables y la magnitud de su varianza pueden afectar en gran medida a los resultados obtenidos por el clustering. Así que estandarizamos con la siguiente fórmula, para lograr que todas las variables tengan una distribución $N(0,1)$

$$z = \frac{x_i - \mu}{\sigma}$$

- Será a ese conjunto estandarizado al que le aplicaremos el algoritmo de k-medias.

2.5.1.1 Análisis de tendencia de agrupación

Nos preguntamos si realmente existe una estructura de agrupación en los datos. Porque si realizamos un análisis clúster sin más, obtendremos grupos (los “k” que deseemos), pero podría ser que estos no tuvieran ningún tipo de sentido.

Para ello vamos a usar un par de métodos:

- **Método estadístico: la estadística de Hopkins:** Compara la distancia de un grupo de datos aleatorios con variación similar a nuestros datos, con las distancias de los datos reales que queremos analizar (el cociente entre las primeras distancias y la suma de ambas distancias). Así que:
 - si el valor es 0.5 las distancias en el conjunto de datos aleatorio y el conjunto de datos real son similares y por tanto nuestros datos no muestran ninguna tendencia de agrupación.
 - mientras que cualquier valor que sea distinto a ese 0.5 implicará que los datos sí tienen algún tipo de agrupación.

Le indicamos por tanto en conjunto de datos o matriz y el número de puntos que queremos seleccionar de los datos. Podemos elegir tantos puntos como total de observaciones menos 1 y aplicamos la prueba para los puntos escalados.

- **Evaluación visual de tendencia (VAT):** realiza un mapa de calor con las distancias entre observaciones, reordenando las observaciones para que las que sean más similares se coloquen cercanas entre sí. En rojo se indica alta similitud (es decir, baja distancia) y en azul lo contrario. Si el conjunto de datos presenta grupos con observaciones similares, observaremos un patrón con cuadrados o rectángulos de color rojo grandes. Si no, el patrón del mapa de calor será de muchísimos cuadriláteros pequeños azules y rojos.

2.5.1.2 Elección del número óptimo de grupos

A este algoritmo hay que indicarle cuál es el número k de grupos que queremos formar. Tenemos dos maneras de saber cuál es el número óptimo:

- **Validación relativa:** aquí vamos a estudiar 30 medidas de calidad para comparar agrupaciones.
- **Según criterio de estabilidad:** aquí buscamos comparar agrupaciones realizadas sobre varios subconjuntos de datos para medir qué tan sensible es la agrupación a los cambios de la muestra. La función **clValid** de **R** nos va a calcular la proporción promedio de no superposición (APN), la distancia media (AD), la distancia media entre medias (ADM) y la figura de mérito (FOM)

2.4.1.3 Validación del algoritmo

Lo vamos a evaluar con el **ancho de silueta**: qué tan cerca está la observación de las observaciones de clústeres vecinos.

Se toma como referencia que si el valor es menor a 0.25, esa observación no está bien agrupada (no está en el grupo que debiera estar).

2.5.2 PAM

PAM o Partitioning Around Medoids, (Kaufman & Rousseeuw, 1990), es un método de partición por clúster en el cual, cada clúster está representado por uno de los objetos del clúster llamados medoides. Es menos sensible a los valores atípicos en comparación con k-means y por tanto una alternativa a este más robusta para el problema que nos ocupa (Kassambara, [12]).

Los **medoides** son objetos representativos de un conjunto de datos o de un clúster dentro de un conjunto de datos cuya suma de disimilitudes con todos los objetos del clúster es mínima.

Precisamente por esto, es menos sensible a la presencia de outliers que el algoritmo de k-medias, donde el centro del clúster se calcula mediante el uso de la media de los puntos del grupo.

Y al igual que el k-medias es un **método iterativo** (Kassambara, [12]):

- Selecciona k objetos para que se conviertan en los medoides, o en caso de que se proporcionen estos objetos utilizarlos como medoides.
- Calcula la matriz de disimilitud si no se proporcionó.
- Asigna cada objeto a su medoide más cercano.
- Para cada clúster busca si alguno de los objetos del clúster disminuye el coeficiente de disimilitud medio; si lo hace, selecciona la entidad que más disminuye este
- Si lo hace, selecciona la entidad que más disminuye este coeficiente como el medoide para este clúster.
- Si al menos un medoide ha cambiado, se vuelve al tercer punto y si no, se termina el algoritmo.

Previo al estudio, preparamos los datos igual que para k-medias.

2.5.2.1. Análisis de tendencia de agrupación

Misma manera que para k-means.

2.5.2.2 Elección del número óptimo de grupos

Para estimar el número óptimo de clústeres usaremos:

- **Método de la silueta media:** De acuerdo con Kassambara [12], la idea es calcular el algoritmo PAM utilizando diferentes valores de clústeres k. A continuación, la silueta media de los clústeres se dibuja en función del número de clústeres. La silueta media mide la calidad de una agrupación. Una silueta media alta indica una buena agrupación. El número óptimo de grupos k es el que maximiza la silueta media en un rango de valores posibles para k (Kaufman y Rousseeuw [1990]).

2.5.2.3 Validación

Misma manera que para k-means.

2.5.3 PCA

El Análisis de Componentes Principales o **PCA** (por sus siglas en inglés) tiene como principal objetivo del PCA resumir la variación (información) más importante que contiene el conjunto de datos multivariado [11] y [13]

A grandes rasgos, sus **logros** se basan en reducir la dimensión de los datos con una **pérdida de información mínima** y **simplificar** reemplazando un grupo variables correlacionadas con pocos componentes principales no correlacionados entre sí y que por lo tanto ya no dan información redundante. De esta manera aparta el ruido y hace que nos quedemos con las variables que no estén correlacionadas para poder medir correctamente el efecto de cada una sobre la respuesta.

Además, nos permite resumir la información en sencillos gráficos de dispersión.

En definitiva, es un método ideal cuando se tiene un conjunto de datos con individuos descritos por muchas variables cuantitativas relacionadas.

Si pensamos en cada variable como una dimensión, el tener más de 3 (incluso más de 2) complica en gran manera la visualización de las relaciones, ya que no tenemos visión geométrica de lo que sería un gráfico tetradimensional, por ejemplo.

Entonces, el PCA “juntará” (matemáticamente, hará una combinación lineal) las variables correlacionadas entre sí en una nueva variable, llamada componente principal, de tal manera que reduzca la dimensionalidad, idealmente a dos para poder visualizar gráficamente en un plano perdiendo la mínima información posible.

Una sucinta explicación del algoritmo podría ser la siguiente (Kassambara, [13]):

- Partimos de un conjunto de datos con muchas (más de 2) variables.
- Entonces el primer paso será encontrar el centro de los datos. La media de cada variable.
- Después buscaremos la línea que mejor se ajuste a los datos pasando por dicho punto. A esta línea la llamaremos **componente principal 1** o PC1. Como en una regresión lineal múltiple, el PC1 es una combinación lineal de las n variables del estudio, lo que significa que contiene parte de cada una de estas variables y explicará la mayor parte de la variación de los datos.
- Una vez encontramos PC1, queda por definir el segundo componente PC2 que explicará la variación restante: es la línea de mejor ajuste perpendicular (es decir, en ángulo recto) a PC1. Y también es una combinación lineal de las n variables.

- Ahora tenemos 2 componentes principales: dos variables en lugar de n , que explican las principales variaciones en las características de mis datos.
- Por último, si giramos el gráfico para que PC1 sea el eje X y PC2 el eje Y, obtenemos una representación en dos dimensiones.

La preparación de los datos es la misma que para k-medias y PAM.

2.5.3.1 Estudio de si es adecuado el PCA

Mediante tres métodos:

- **Prueba de esfericidad de Bartlett:** Esta prueba evalúa si es necesario realizar el PCA analizando si la correlación entre las variables analizadas es lo suficientemente grande como para justificar la factorización o la reducción de la matriz de correlaciones. La hipótesis nula, H_0 de esta prueba es que la matriz de correlación proviene de una población no colineal, o mejor dicho, que no existe colinealidad entre las variables (con lo que no se podría utilizar el PCA, ya que depende de la construcción de una combinación lineal de variables)
- **Adecuación muestral KMO:** Tiene el mismo objetivo que el test de Bartlett, pero se trata de saber si podemos factorizar o resumir las variables originales de forma eficiente. El punto de partida también es la matriz de correlaciones entre las variables observadas. Sabemos que las variables pueden estar más o menos correlacionadas, pero la correlación entre dos de ellas puede estar influenciada por otras. Así utilizamos la correlación parcial para medir la relación entre dos variables, eliminando el efecto del resto.
El índice KMO compara los valores de las correlaciones entre las variables y sus correlaciones parciales. Si el índice está próximo a 1 tiene sentido realizar el PCA. Si el índice es bajo, el análisis podría no ser relevante. Algunos autores han estimado que respecto al KMO general:
 - .5 a .7 son mediocres
 - .7 a .8 son buenos
 - .8 a .9 son muy buenos
 - por encima de .9 son excelentes.
- **Positividad de la matriz de correlación:** Probamos la multicolinealidad según el determinante de dicha matriz, que debe ser positivo y bajo.

2.5.3.2 Selección del número de dimensiones

Tenemos cuatro métodos principales para seleccionar el número de dimensiones:

- **Regla de Keisser-Guttman:** propone quedarse con las dimensiones con autovalores mayores a 1

- **% de varianza explicada:** propone quedarse con las dimensiones que expliquen un, por ejemplo, 70% de la varianza (el número dependerá de las necesidades de la investigación).
- **Gráfico de pantalla o sedimentación:** graficamos los valores propios vs PC y nos quedamos con el punto de inflexión: ahí donde los valores propios pasan a estabilizarse.
- **Análisis paralelo de Horn:** contrasta la variabilidad expresada en el conjunto de datos original con la obtenida de conjuntos de datos generados aleatoriamente con características similares al original. Esta opción permite controlar posibles efectos de sesgo en la elección de PC cuando el conjunto de datos es pequeño.

2.5.3.3 Estudio de las variables. Características

La **pregunta** aquí es ¿cuándo podemos decir que **dos variables son similares o distintas?**

Y, por tanto, el **objetivo:** **encontrar un número de variables sintéticas que resuman muchas variables.**

Debemos identificar las variables más importantes o relevantes para los PC seleccionados:

- *coord.* coordenadas (correlación) de las variables en cada PC. Corresponde a la correlación de la variable con cada componente.
- *cos2.* calidad de la representación para las variables en cada PC. Corresponde a calcular el cuadrado de las coordenadas (coord^2)
- *contrib.* contribución (en %) de la variable a cada PC. Se calcula así: $(\text{cos2}/\text{sum}(\text{cos2}))$

Estudiándolas una por una:

1. Coordenadas. Mapa de variables

Tendremos un gráfico con una circunferencia centrada en el origen y una serie de flechas por cada variable.

- Los ángulos (dirección de la flecha), los podemos interpretar como el signo de la correlación entre ellos.
 - si el ángulo es pequeño, la correlación es positiva
 - si el ángulo es grande (en cuadrantes opuestos), la correlación es negativa
 - si el ángulo es de 90°, no hay correlación
- Las distancias (largo de la flecha) mide la magnitud de la correlación
 - si la flecha es larga, indica una buena representación en el PC

– si la flecha es corta (origen), indica una mala representación

2. Calidad de la representación ($\cos^2$)

Toma valores entre 0 y 1:

- si $\cos^2$ está cercano a 1, el ángulo con el componente es cercano a 0 (porque $\cos$ representa coseno y recordemos que el coseno de 0° es 1) y por tanto está bien representada en ese componente
- si $\cos^2$ es cercano a 0, la variable no estaría bien representada en ese componente (de nuevo recordemos que el coseno de 90° es 0)

3. Contribución a la representación (contrib)

Para interpretar un componente, vamos a observar la variable de mayor contribución, que son las más correlacionadas con él. Las variables que se correlacionan con los primeros componentes son las más importantes en explicar la variabilidad de los datos. Las variables que no se correlacionan con ningún componente o solamente con los últimos componentes tendrán una baja contribución y por tanto podrían eliminarse del análisis.

4. Descripción de las dimensiones (dimdesc)

Nos permite **describir las dimensiones o componentes principales**, identificando las variables asociadas significativamente a cada uno. Es decir, vemos las correlaciones ordenadas y su significación.

2.5.3.4 Estudio de los individuos (*casos*)

Aquí la **pregunta** es **¿cuándo podemos decir que dos individuos son similares (o distintos) respecto a todas las variables?**

Y por tanto el **objetivo**: identificar grupos de individuos con perfiles similares o los más extremos.

Para lograr este objetivo:

- *coord.* coordenadas (proyección) de cada individuo en cada PC.
- $\cos^2$. calidad de la representación en cada PC. Corresponde a calcular el cuadrado de las coordenadas (coord^2)
- *contrib.* contribución (en %) del individuo a cada PC. Se calcula así: $(\cos^2 / \text{sum}(\cos^2))$

De nuevo los vamos a estudiar uno por uno.

1. Coordenadas. Mapa de individuos

También tendremos una circunferencia y representados los individuos por puntos. A diferencia del mapa de variables, que utilizaba las correlaciones como coordenadas, los individuos u observaciones se representan mediante sus proyecciones en el mapa. Entonces:

- las coordenadas son las proyecciones.
- La distancia al centro:
 - individuos en el centro tienen valores promedio. Es decir, que los individuos con valores promedio en todas las variables se colocarán en el centro del gráfico.
 - individuos lejos del centro destacan en alguna variable
- perfiles:
 - individuos con valores similares en las variables de estudio se agruparán en el mapa
 - individuos con perfiles distintos estarán lejos entre sí en el mapa

2. Calidad de la representación

La calidad es la distancia entre individuos, que solo puede interpretarse para individuos bien proyectado (con alto $\cos^2$).

3. Contribución a la representación

Los individuos con mayor valor de coordenada además van a contribuir más al eje.

2.5.4 HCPC

Se llama **HCPC** [13] por sus siglas en inglés, Hierarchical Clustering of Main Components (Husson, Josse, and J. 2010).

El clustering es uno de los métodos de minería de datos más importantes para descubrir conocimientos en conjuntos de datos multivariantes multivariados. El objetivo es **identificar grupos** (es decir, clústeres) **de objetos similares dentro de un conjunto de datos de interés**.

Brevemente, las dos estrategias de clustering más comunes son

_El clustering jerárquico, utilizado para identificar grupos de observaciones similares en un conjunto de datos.

_ Clustering de partición, como el algoritmo k-means, que se utiliza para dividir un conjunto de datos en varios grupos.

Pues el enfoque HCPC (Hierarchical Clustering on Principal Components) permite **combinar los tres métodos estándar** utilizados en los análisis de datos multivariantes

1. Métodos de componentes principales (PCA, CA, MCA, FAMD, MFA),
2. Agrupación jerárquica y
3. Clustering de partición, en particular el método k-means.

Es decir, usamos primero el PCA [11], que se puede utilizar como un paso de preprocesamiento antes de realizar métodos de agrupación, con el fin de eliminar el ruido de los datos.

Así que sin el ruido procedemos a agrupar ya que obtendremos clusterizaciones más estables que las obtenida mediante distancias originales.

El algoritmo es [13]

- Calcular los métodos de componentes principales: PCA, (M)CA o MFA en función de los tipos de variables en el conjunto de datos y la estructura del conjunto de datos. En este paso, puede elegir el número de dimensiones que se mantendrán en la salida especificando el argumento ncp. El valor por defecto es 5.
- Calcula el clustering jerárquico: El clustering jerárquico se realiza utilizando el criterio de Ward en los componentes principales seleccionados. El criterio de Ward se utiliza en el clustering jerárquico porque se basa en la varianza multidimensional como el análisis de componentes principales.
- Se elige el número de clústeres en función del árbol jerárquico: Se realiza una partición inicial cortando el árbol jerárquico.
- Realiza la agrupación de K-means para mejorar la partición inicial obtenida de la agrupación jerárquica. La solución de partición final, obtenida tras la consolidación con k-means, puede ser (ligeramente) diferente de la obtenida con el clustering jerárquico.

CAPÍTULO 3: ESTUDIO DE K-MEDIAS

3.1 Resultados y discusión

A lo largo del capítulo iremos presentando los resultados uniéndolos con una cierta discusión, a fin de evitar repeticiones y para lograr una estructura más legible y coherente.

Todos estos resultados y discusiones se resumirán en el apartado de Conclusiones.

3.1.1 Análisis exploratorio de datos.

Un primer problema al que nos enfrentamos es que el conjunto tiene muchísimos **outliers**.

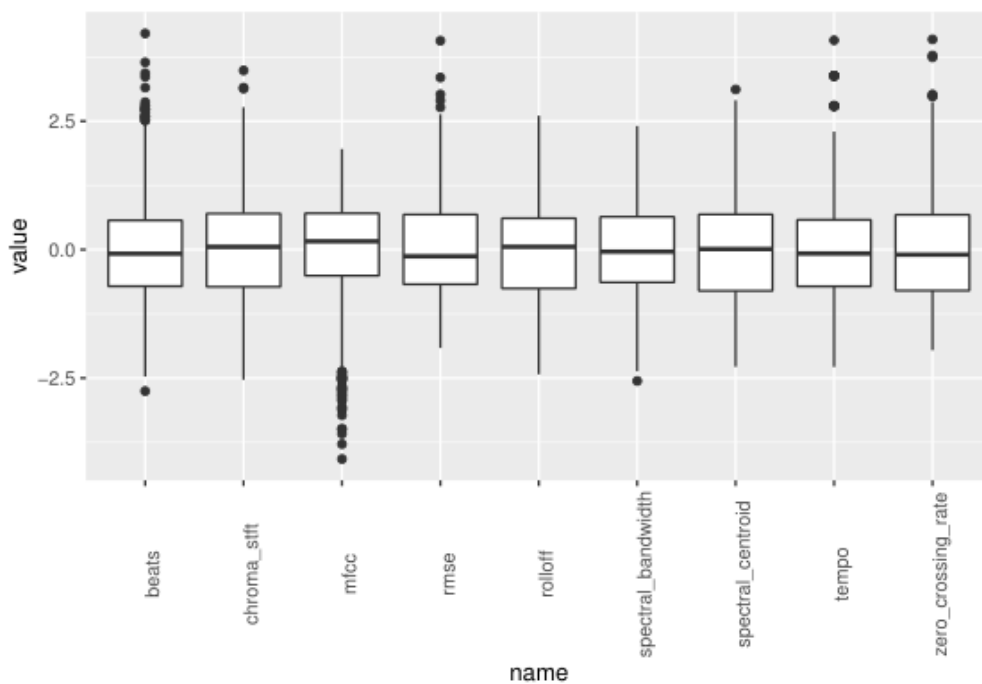


Figura 1. Estudio outliers (Fuente: elaboración propia)

Podríamos eliminarlos (seguramente la peor opción) o manipularlos (cambiar sus valores por el valor medio de la variable en concreto, por ejemplo).

Pero debido a la naturaleza del estudio **hemos decidido NO hacer nada con ellos**. Recordemos una vez más nuestro objetivo: ¿es posible separar los géneros musicales

en función de variables físicas? Y recordemos que ya aludíamos a la infinita variabilidad de cada uno de los estilos.

Por tanto, si manipuláramos los valores de las canciones más diferentes dentro del estilo en el que ha sido categorizada, ¿no estaríamos realmente manipulando el estudio para llegar en efecto a una clusterización quizás perfecta pero ficticia? ¿No estaríamos dando de lado la riqueza musical?

En cualquier caso, es por esta presencia de valores atípicos, que en la sección siguiente realizaremos un análisis PAM, algoritmo más robusto en este sentido.

Realizamos también un **análisis exploratorio de datos**, relacionando cada género o *label* con cada variable.

En este caso respecto a *spectral_centroid*:

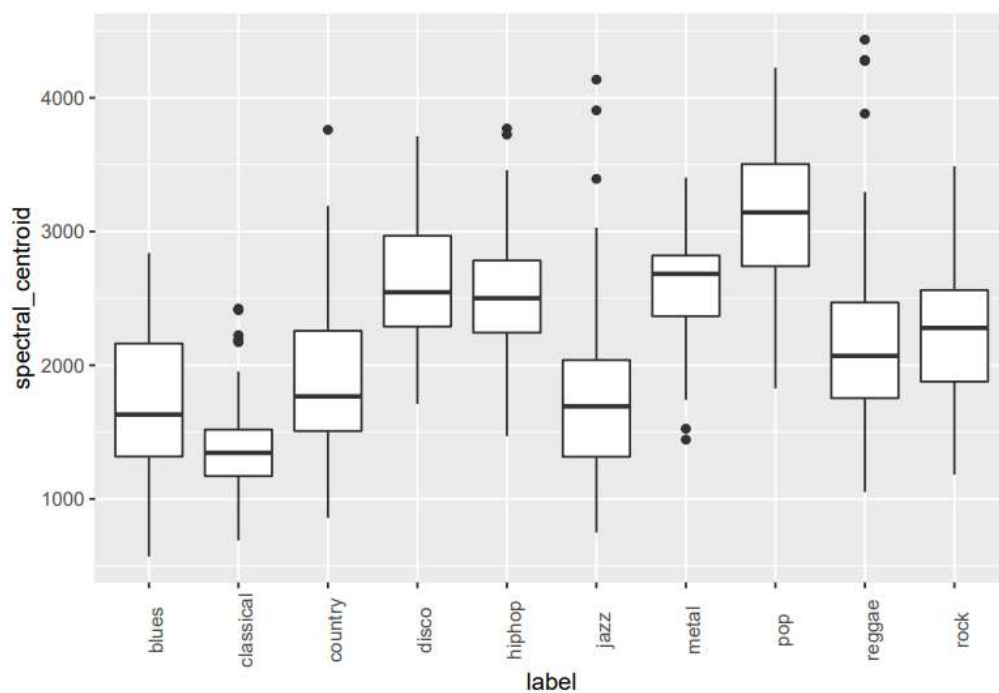


Figura 2. Label vs spectral_centroid (Fuente: elaboración propia)

No añado todas las comparativas (se pueden encontrar en los anexos, dentro del código), pero resumiremos en que, para todas las variables salvo *beats* y *tempo*, parece que *clásica* y *jazz* tienen los valores más bajos. En algunos de ellos también son bajos los valores para *blues* y *country*.

Al contrario, los valores más altos se los llevan *pop*, *metal*, *disco* y *hip-hop* tal y como se aprecia en el gráfico.

En cambio:

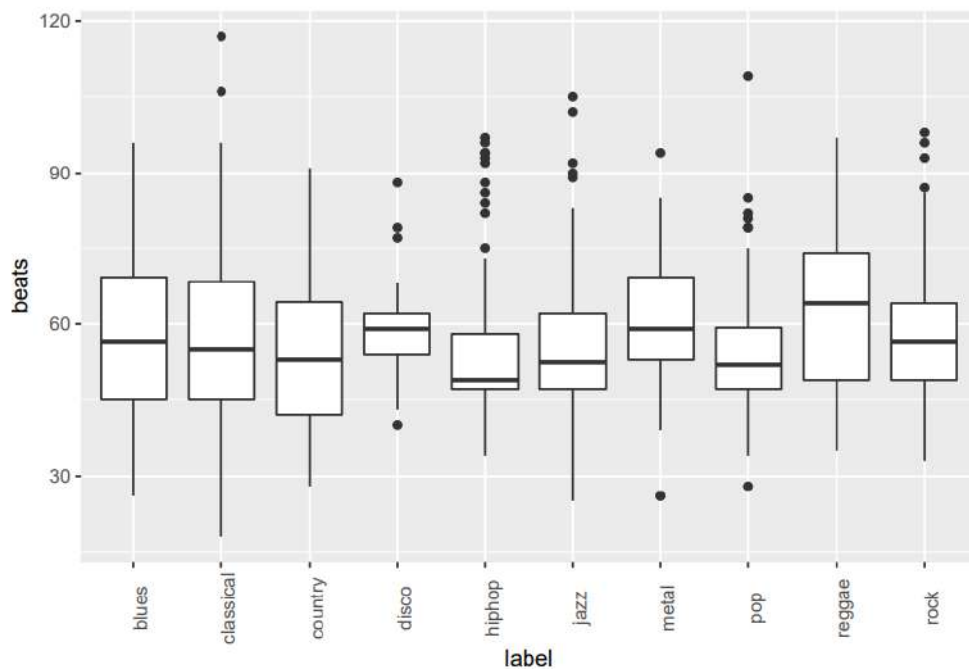


Figura 3. Label vs beats (Fuente: elaboración propia)

Para *beats* y *tempo* no parece que encontremos diferencias significativas entre los géneros, más allá de que parece que la música *disco* es muy homogénea al respecto.

Podríamos esperar por tanto de nuestro análisis clúster que, si es capaz de separar por estilos, el *jazz* y la *clásica* estén aislados del resto y juntos entre sí a través de grupos que se lleven los valores más pequeños de todas las variables salvo *beats* y *tempo*.

Y lo mismo, pero al revés para *pop*, *metal*, *disco* o *hiphop* cuyos representantes debieran caer en grupos con altos valores de esas variables.

3.1.2 Análisis de tendencia de la agrupación

- **Método estadístico: la estadística de Hopkins:**

En nuestro caso, **obtenemos un valor de 0.164696**, que es distinto de 0.5. Podemos concluir que **los datos sí muestran una tendencia de agrupación digna de analizar**.

- **Evaluación visual de tendencia (VAT):**

Este es el gráfico de nuestro set de datos:

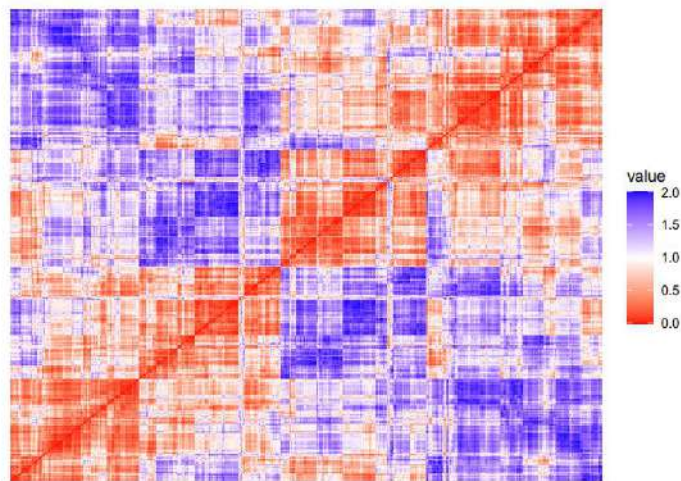


Figura 4. Validación VAT tendencia de agrupación (Fuente: elaboración propia)

Se puede apreciar que **hay tendencia de agrupación**: observamos al menos dos grandes grupos rojos a lo largo de la diagonal del cuadrado.

3.1.3 Elección del número óptimo de grupos

Respecto a los dos métodos definidos anteriormente en la explicación de k-medias (**validación relativa** o según **criterio de estabilidad**), el primero nos recomienda **2 grupos**: 9 de los 30 estadísticos estudiados determinan que ese es el número óptimo, tal y como se ve en la figura 21 de los anexos.

En cambio, tal y como se ve en la figura 22, de los 4 estadísticos usados por el segundo criterio, **dos recomienda 2 clústeres y los otros dos recomiendan 10**. Recordemos que tenemos 10 géneros, con lo que 10 sería estupendo para nuestro propósito de separar las canciones.

No obstante, lo vamos a desechar, porque tras entrenar el modelo y hacer una comparativa entre los estilos que caen en cada clúster y los de cada sujeto, llegamos a la siguiente tabla:

##		1	2	3	4	5	6	7	8	9	10
##	blues	19	0	2	20	0	25	6	1	8	19
##	classical	3	0	55	0	0	2	0	0	6	34
##	country	12	1	2	7	2	28	7	8	22	11
##	disco	4	14	1	13	4	0	42	12	10	0
##	hiphop	3	6	0	34	14	0	14	20	9	0
##	jazz	5	3	5	3	3	28	9	1	23	20
##	metal	2	3	0	16	30	0	33	12	4	0
##	pop	4	21	0	4	7	0	6	49	8	1
##	reggae	41	4	1	7	10	2	9	3	21	2
##	rock	13	3	0	15	8	7	28	2	23	1

Figura 5. Comparativa clúster/label y % k-means k = 10 (Fuente: elaboración propia)

que no se acerca para nada a nuestros objetivos: **no se aprecia que cada grupo contenga la mayor parte de cada género.**

Así que nos conformaremos con un análisis de k-medias para k = 2.

3.1.4 Representación de los grupos

Los grupos está **bien balanceados** (mismo número de sujetos en cada grupo). De hecho hay 577 en el primero y 423 en el segundo (figura 23)

Y viendo los centros de cada grupo (figura 24) concluimos que **el grupo 1 se ha llevado los valores mayores y el grupo 2 los menores de todas las variables.**

Representado gráficamente se puede apreciar mucho mejor:

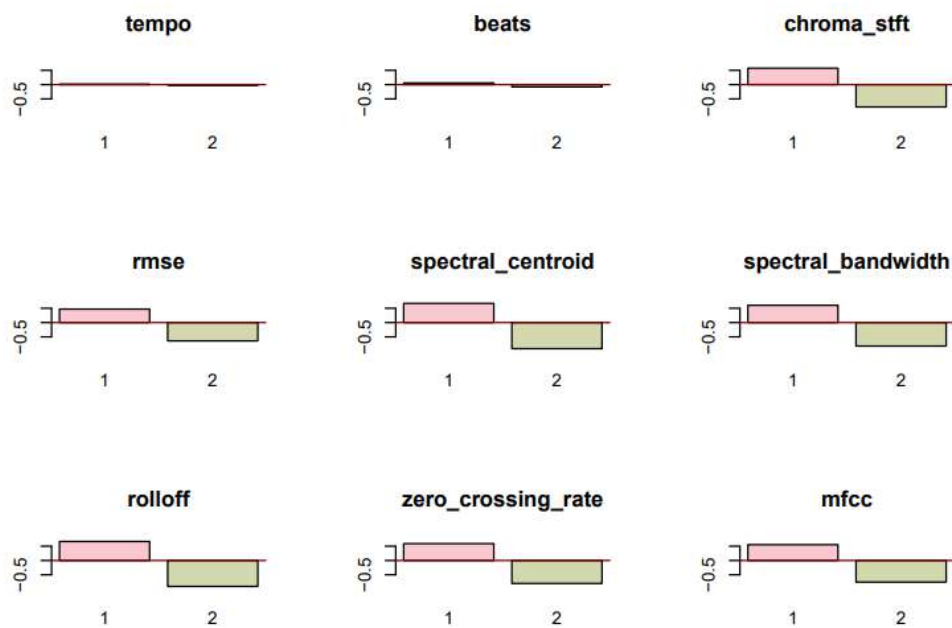


Figura 6. Valores grupos k-means k=2 (Fuente: elaboración propia)

3.1.5 Importancia de las variables

Por la figura 25 se puede apreciar que, **todas menos** *tempo*, *beats* y quizás *rmse* han sido importantes. Concretamente las que más, *spectral_centroid*, *rolloff* y *zero_crossing_rate*.

3.1.6 Comparación de los grupos con los géneros originales

Comparamos el estilo que tenía cada canción con el clúster en el que ha caído tras nuestro algoritmo de k-medias:

##			
##		1	2
##	blues	37	63
##	classical	1	99
##	country	32	68
##	disco	89	11
##	hiphop	92	8
##	jazz	20	80
##	metal	99	1
##	pop	91	9
##	reggae	49	51
##	rock	67	33

Figura 7. Comparativa clúster/label y % k-means k = 2 (Fuente: elaboración propia)

Teniendo en cuenta que cada estilo tenía 100 canciones, **podemos considerar cada número** además de como cuántos representantes de cada género hay en cada clúster, también **como el porcentaje**.

Así que el algoritmo ha sido capaz de crear un clúster que aloja **el 99% del *metal*, el 92% del *hiphop*, el 91% del *pop* y el 89% del *disco*** y otro con el **99% de *classical* y el 80% de *jazz***.

El resto de los géneros parecen repartirse entre ambos clústeres, con predominancia en el 1 de rock y en el 2 de blues y country. El reggae está casi al 50% en ambos grupos.

Recordemos que en el análisis exploratorio de datos llegábamos a conclusiones parecidas: grupo 1 con los valores más altos incluye a *disco*, *hip hop* y *metal* y grupo 2 con los valores más bajos incluye a *clásica* y *jazz*.

3.1.7 Validación

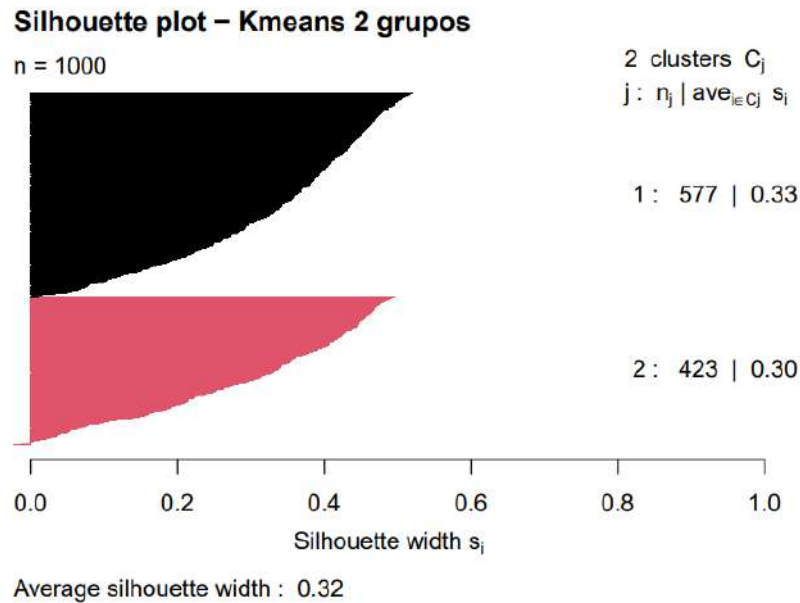


Figura 8. Validación k-means k = 2 (Fuente: elaboración propia)

Tenemos una silueta media de 0.33 para el primer clúster y de 0.30 para el segundo. Dando entonces **de manera global un 0.32**.

No es una mala agrupación, aunque debemos tener en cuenta **que hay muchas observaciones que caen por debajo del 0.25 de referencia** y que por tanto no estarían bien categorizadas.

Además, debemos tener **cuidado con las interpretaciones**, ya que, como hemos comentado, **hay bastantes outliers** y este algoritmo es especialmente sensible a estos valores atípicos.

CAPÍTULO 4: ESTUDIO DE PAM

4.1 Resultados y discusión

A lo largo del capítulo iremos presentando los resultados uniéndolos con una cierta discusión, a fin de evitar repeticiones y para lograr una estructura más legible y coherente.

Todos estos resultados y discusiones se resumirán en el apartado de Conclusiones.

4.1.1 Análisis exploratorio de datos.

Idéntico que el realizado para k-means.

4.1.2 Análisis de tendencia de la agrupación

Idéntico que el realizado para k-means.

4.1.3 Elección del número óptimo de grupos

Por el método de silueta media explicado en Material y Métodos, obtenemos que el **número óptimo de grupos es 2** tal y como se puede apreciar en la figura 26 de los anexos.

4.1.4 Comparación de los grupos con los géneros originales

Estudiamos cómo ha agrupado nuestro algoritmo, comparando el estilo de cada una de las canciones con qué género ha ido a caer a cada uno de los grupos:

	1	2
blues	50	50
classical	4	96
country	51	49
disco	98	2
hiphop	97	3
jazz	28	72
metal	100	0
pop	95	5
reggae	82	18
rock	80	20

Figura 9. Comparativa clúster/label y % PAM (Fuente: elaboración propia)

De nuevo tengamos en cuenta que había 100 canciones por género, con lo que las volumetrías de la figura 9 representa también % en cada clúster.

Así que en este caso el algoritmo ha sido capaz de crear **un clúster** que aloja el **100% del metal**, el **98% del disco**, el **97% del hiphop**, el **95% del pop**, el **82% del reggae** y el **80% del rock** y otro con el **96% de classical** y el **72% de jazz**.

Es decir, **se han separado bastante bien los géneros más modernos (clúster 1) de los más clásicos (clúster 2)**. *Country* y *Blues* están un 50-50 en cada uno de los grupos, cosa que podría tener lógica si pensamos que son dos estilos que en la actualidad se han “re-visitado” bastante.

4.1.5 Importancia de las variables

Pero seguramente lo más interesante del estudio, sea saber qué variables han contribuido más a esa partición del conjunto de datos original.

Graficamos para cada clúster:

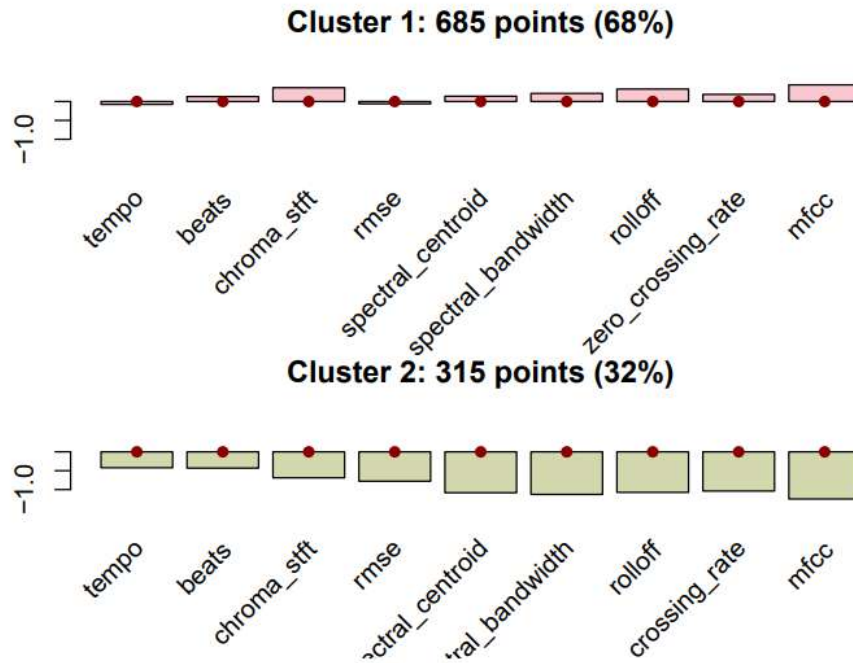


Figura 10. Valores grupos PAM (Fuente: elaboración propia)

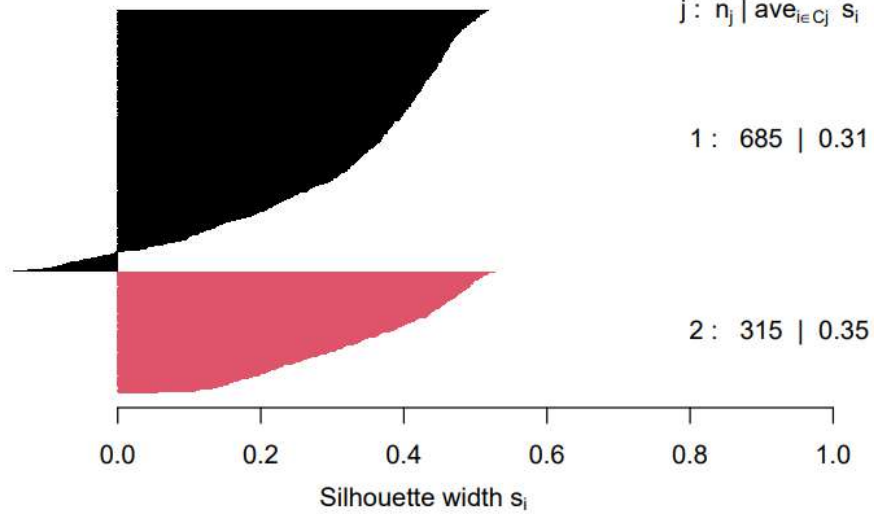
Y podemos observar que **el clúster 1 se lleva los valores más altos de cada una de las variables mientras que el 2 los más bajos.**

También mediante la figura 27 vemos que los que más aportan son, *mfcc*, *rolloff* y *spectral_bandwidth*. De nuevo *tempo* y *beats* no aportan casi nada al igual que en este caso *rmse*.

4.1.6 Validación

Silhouette plot – PAM 2 grupos

n = 1000



Average silhouette width : 0.33

Figura 11. Validación PAM (Fuente: elaboración propia)

Tenemos una silueta media de 0.31 para el primer clúster y de 0.35 para el segundo. Dando entonces **de manera global un 0.33**.

No es una mala agrupación, aunque debemos tener en cuenta que **hay muchas observaciones que caen por debajo del 0.25 de referencia** y que por tanto no estarían bien categorizadas.

CAPÍTULO 5: ESTUDIO DE PCA

5.1 Resultados y discusión

A lo largo del capítulo iremos presentando los resultados uniéndolos con una cierta discusión, a fin de evitar repeticiones y para lograr una estructura más legible y coherente.

Todos estos resultados y discusiones se resumirán en el apartado de Conclusiones.

5.1.1 Análisis exploratorio de datos.

Idéntico que el realizado para k-means y PAM.

5.1.2 Estudio de si es adecuado el PCA en nuestro set de datos

Habíamos mencionado tres métodos, y en nuestro caso,

- **Prueba de esfericidad de Bartlett:**

Obtenemos **un p-valor menor al 0.05 de referencia, podemos rechazar la hipótesis nula** y aceptar que existe colinealidad entre las variables (figura 28)

- **Adecuación muestral KMO:**

Obtenemos **0.7** por tanto, que según la clasificación explicada **podríamos considerar como bueno.** (figura 29)

- **Positividad de la matriz de correlación:**

El determinante es **positivo y bajo** como buscábamos (figura 30)

Se satisfacen de esta manera todos los supuestos del PCA, con lo que podemos seguir adelante con el estudio.

5.1.3 Selección del número de dimensiones

Tenemos cuatro métodos principales para seleccionar el número de dimensiones:

- **Regla de Keisser-Guttman:**

Según vemos en la figura 31, nos quedaríamos con **las 2 primeras dimensiones** y al límite con la tercera (los que tienen el autovalor mayor que 1)

- **% de varianza explicada:** Mirando el mismo gráfico que antes, nos quedaríamos con **las dos primeras**, que juntas suman un 76.5% de varianza explicada.
- **Gráfico de pantalla o sedimentación:** según la figura 32, pareciera que se estabiliza entre la dimensión 2 y la 3.
- **Análisis paralelo de Horn:** la figura 33 nos indica 2 dimensiones

En definitiva, **nos quedaremos con 2 componentes principales.**

5.1.4 Estudio de las variables (características)

- **Coordenada. Mapa de variables**

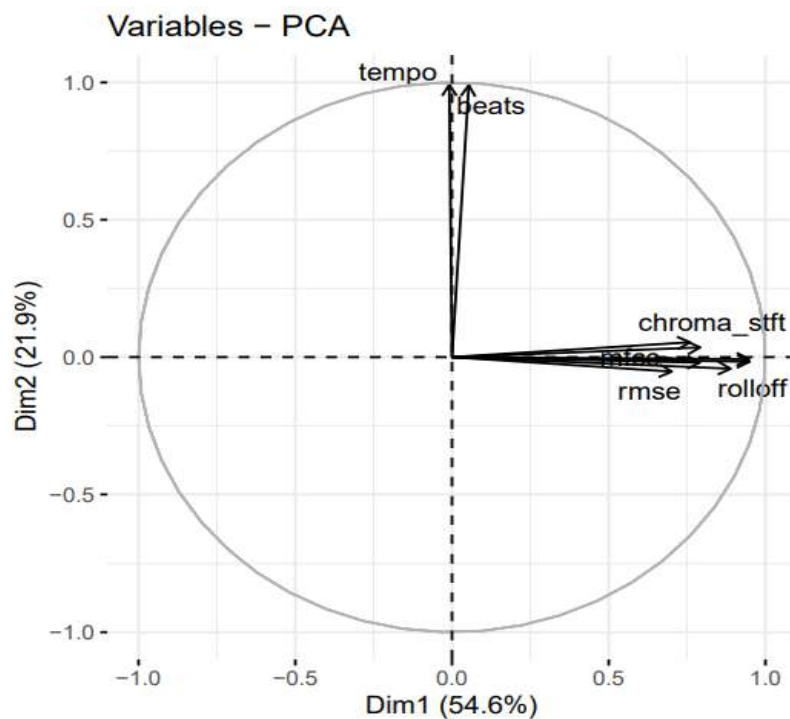


Figura 12. Coordenadas: mapa variables 1 (Fuente: elaboración propia)

Según el gráfico

- parece que **casi todas las variables están bien representadas** en el PC (flechas “largas”).
- Y que el **tempo y beats contribuyen positivamente a la segunda dimensión** además de estar muy correlacionadas (cosa lógica: el tempo se puede medir como beats por segundo) **mientras que las otras están muy correlacionadas y contribuyen positivamente a la primera.**

En la figura 34 se muestra el mismo gráfico pero además coloreado por calidad de representación.

- **Calidad de la representación (cos2)**

Para la primera dimensión:

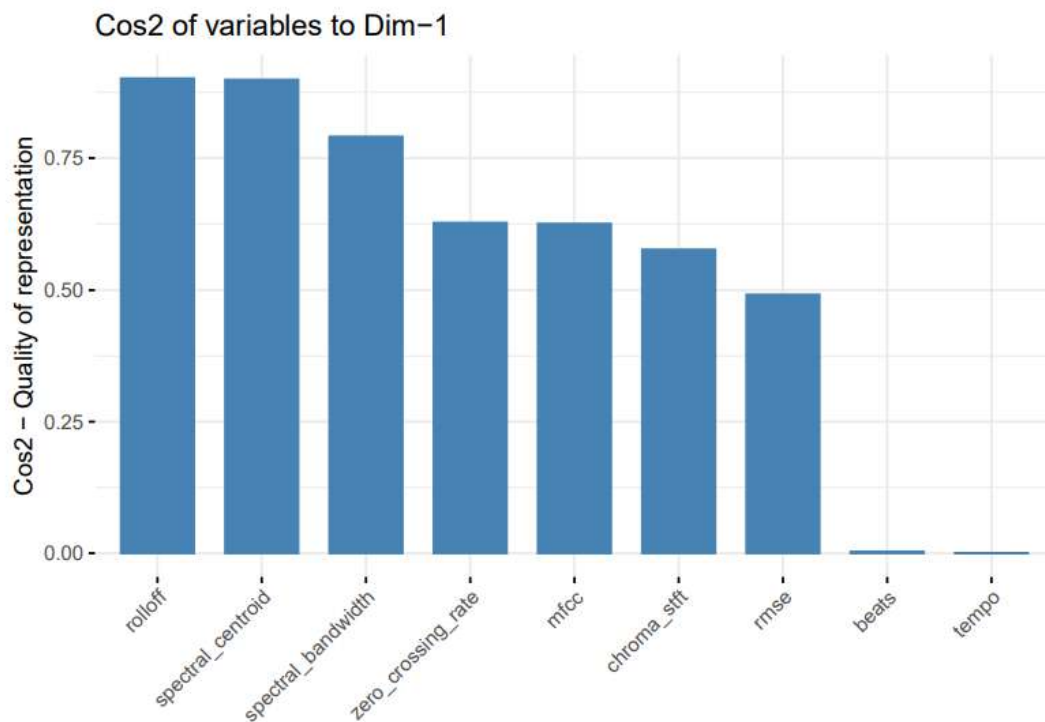


Figura 13. Calidad representación PCA dim1 (Fuente: elaboración propia)

Y para la segunda:

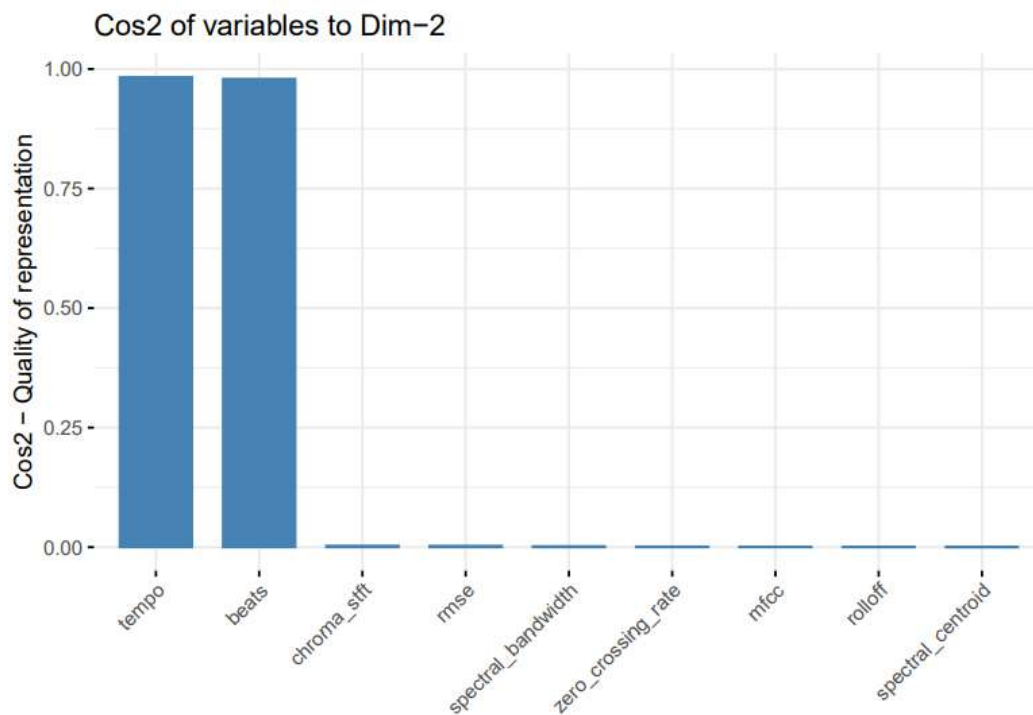


Figura 14. Calidad representación PCA dim2 (Fuente: elaboración propia)

Efectivamente *tempo* y *beats* tienen muy buena calidad de representación para la segunda dimensión y el resto de variables para la primera.

- **Contribución a la representación (contrib)**

Para el primer componente:

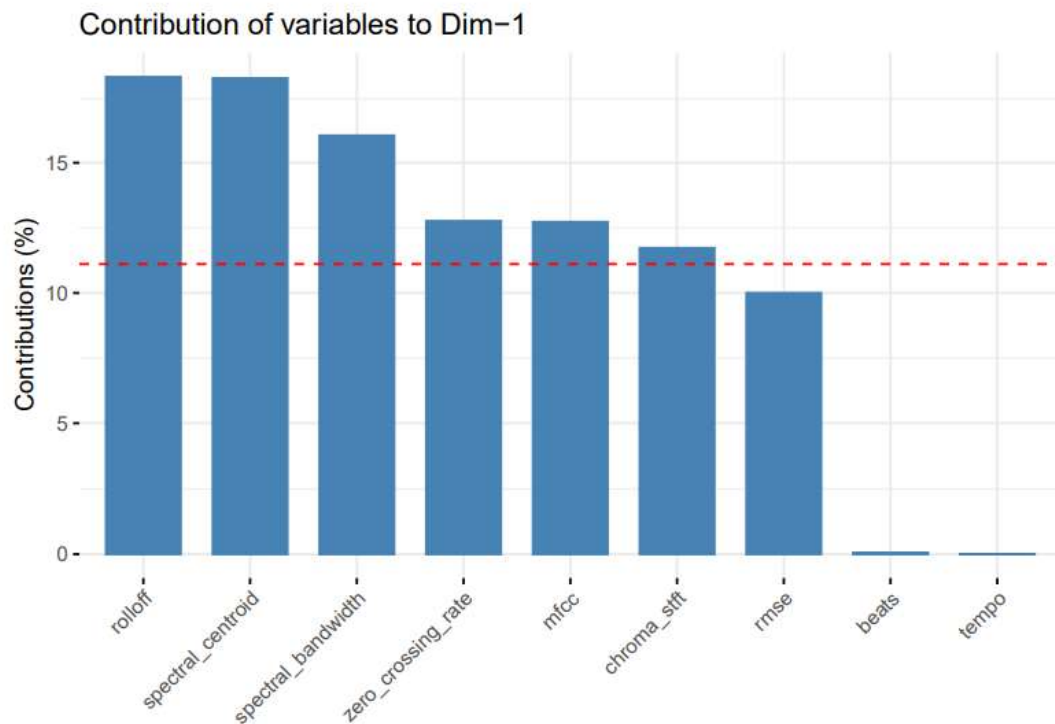


Figura 15. Contribución a la representación PCA dim1 (Fuente: elaboración propia)

Y para el segundo:

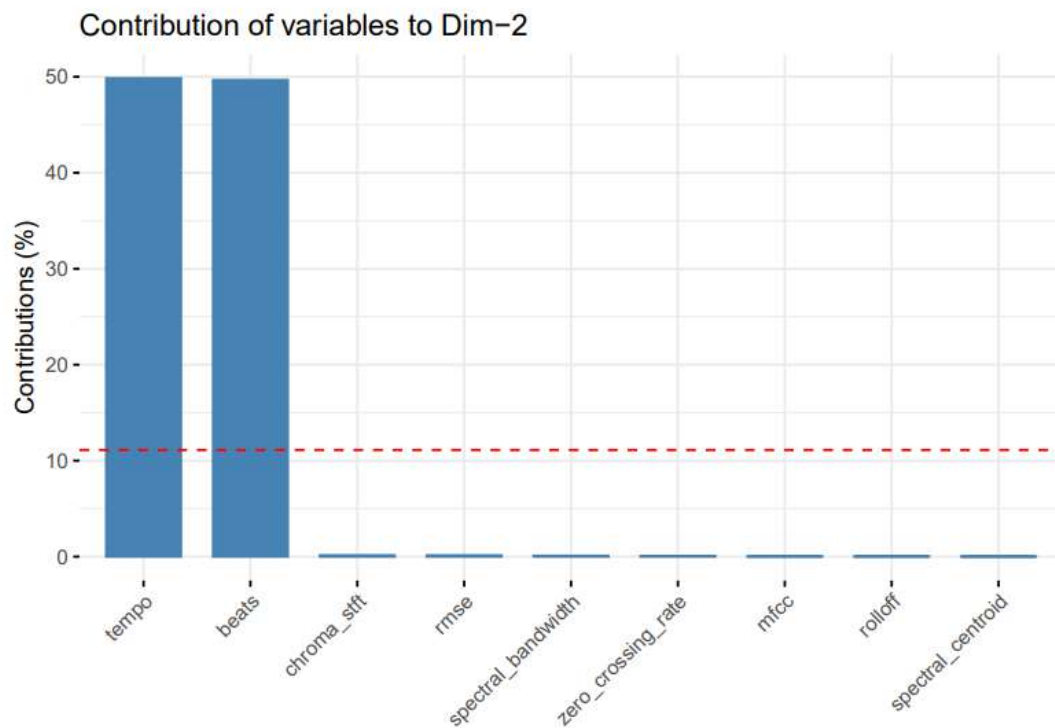


Figura 16. Contribución a la representación PCA dim2 (Fuente: elaboración propia)

Es decir, que **todas las variables están bien representadas en su dimensión correspondiente (aquellas que quedan por encima de la línea punteada roja) salvo rmse.**

- **Descripción de las dimensiones (dimdesc)**

En las figuras 35 y 36 vemos ordenadas las correlaciones por dimensión junto a sus p-valores (significación para un nivel de confianza del 95%). **Nos reafirmamos en lo comentado anteriormente.**

5.1.5 Estudio de los individuos (casos)

- **Coordenadas. Mapa de individuos**

En la figura 37 pareciera que la primera dimensión separa *clásica* y *jazz* del resto mientras que la segunda dimensión separara los propios géneros en función de sus peculiaridades.

En cualquier caso, lo podemos apreciar más claramente en este gráfico, donde le pedimos que nos pinte los 100 individuos (canciones en este caso) más importantes:

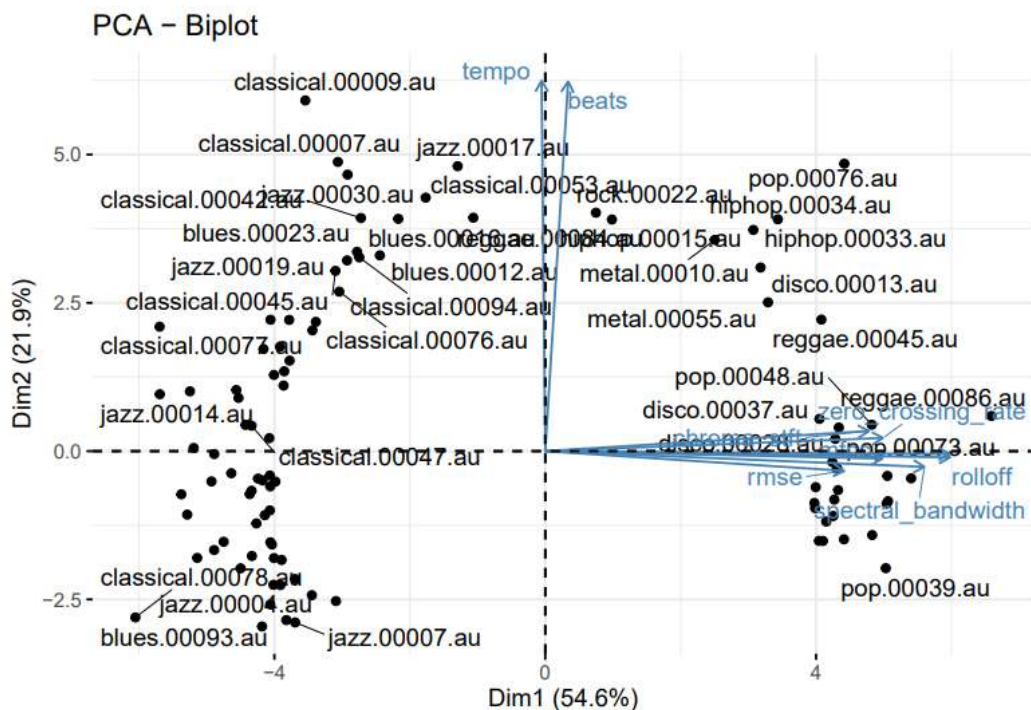


Figura 17. Coordenadas: mapa de individuos 2(Fuente: elaboración propia)

O mejor aún, en la figura 38, donde agrupamos mediante elipses los géneros y coloreamos.

Pero coloreando por cada estilo no visualizamos del todo bien cómo ha separado la primera dimensión. Es por ello que creamos **otro gráfico a tres colores**, donde lo que se pretende es usar mismo color para los estilos que quedan a la **izquierda** de la primera dimensión (**azulados**), otro color para los que quedan a la **derecha** (**amarillos**) y otro color para los que están **a caballo** entre ambos (**naranjas**):

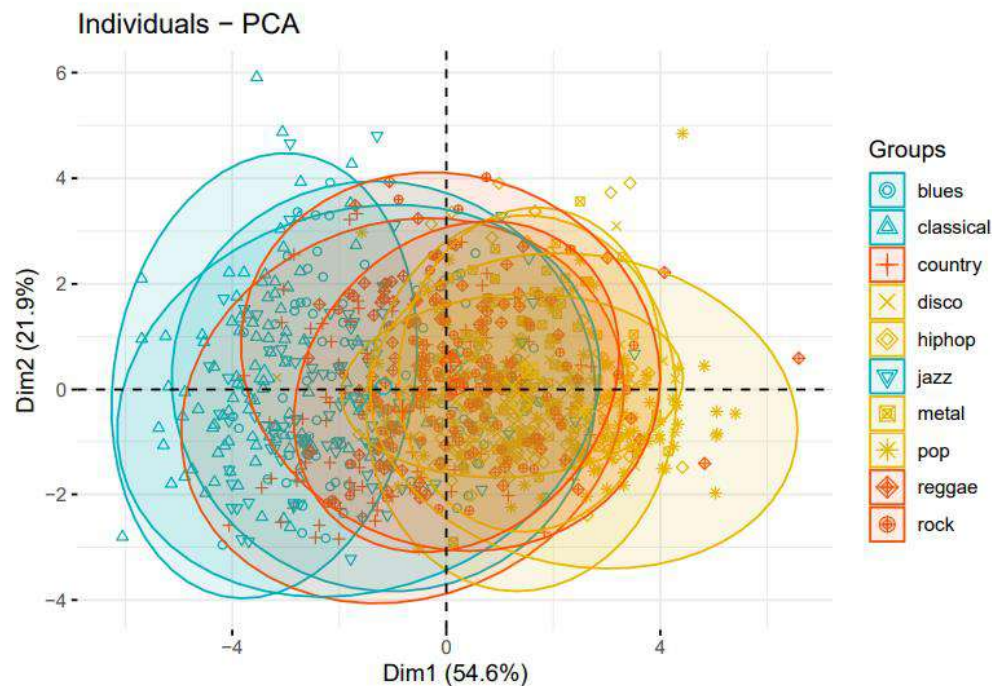


Figura 18. Coordenadas: mapa de individuos 3 colores (Fuente: elaboración propia)

Aquí podemos apreciar que la primera dimensión separa *blues, clásica y jazz* de *disco, hiphop, metal y pop*. *Country, reggae y rock* parecen a caballo entre ambas.

Además, esta dimensión explica un 54.6% de la variabilidad.

Mientras que la segunda dimensión separa los propios géneros en función de sus peculiaridades. Es decir, dentro de un mismo estilo, nos está separando aquellas canciones que tienen mayor número de *beats* (y por tanto *tempo* más acelerado) de las que menos (y por tanto más lentas).

CAPÍTULO 6: HCPC

Tras realizar el estudio de agrupación jerárquica de los componentes principales, llegamos a esta clasificación

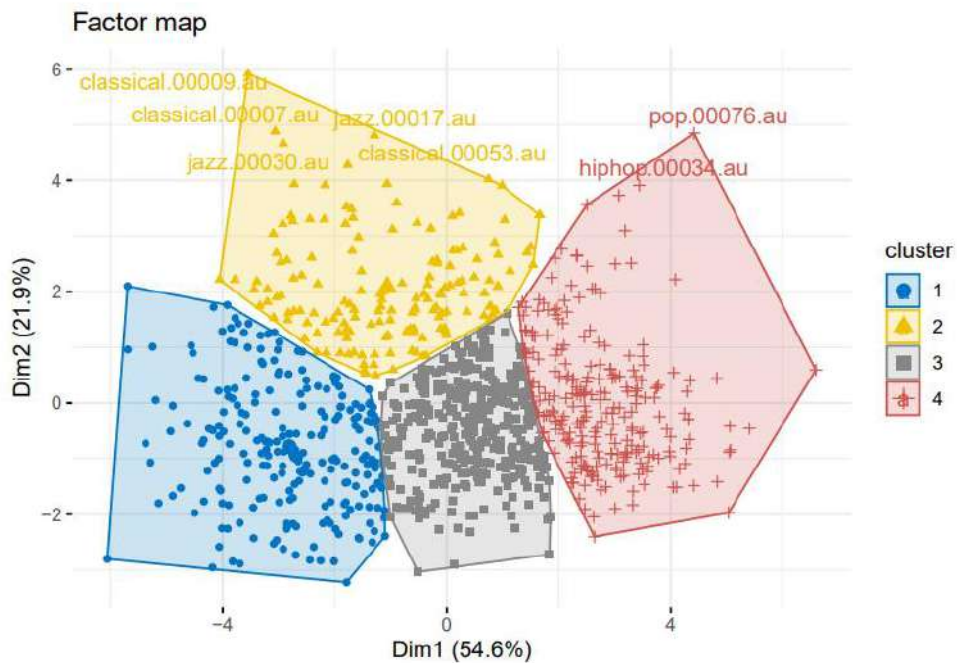


Figura 19. Individuos en el mapa tras HCPC

Donde, según el gráfico 40, obtenemos:

- _un clúster 1 con valores muy bajos/negativos para las dimensiones 1 y 2 originales.
- _un clúster 2 con valores bajos/negativos de la dimensión 1 y altos/positivos de la dimensión 2.
- _un clúster 3 con valores bajos/ positivos de la dimensión 1 y bajos /negativos de la dimensión 2.
- _un clúster 4 con valores altos/positivos para las dimensiones 1 y 2.

En el gráfico 39 vemos que las variables que más contribuyeron a cada dimensión son prácticamente las mismas que la del estudio de k-means.

En el gráfico 38 vemos los individuos más representativos de cada grupo, e interesante es también el gráfico 42 donde se enumeran los individuos de cada grupo que están más alejados del resto de grupos. Es decir, son canciones que solamente podrían estar en el clúster en el que están.

Se comenta más al respecto, en Conclusiones.

Pero vayamos a lo realmente interesante, que es saber cómo ha clasificado el algoritmo:

```
##
##      1  2  3  4
## blues 33 29 36  2
## classical 76 21  3  0
## country 36 18 36 10
## disco  2  2 56 40
## hiphop  1  9 55 35
## jazz  60 16 17  7
## metal  0  9 42 49
## pop    1  5 18 76
## reggae 11 38 38 13
## rock  14 13 60 13
```

Figura 20. Comparativa clúster/label y % HCPC (Fuente: elaboración propia)

CAPÍTULO 7: CONCLUSIONES

7.1 Conclusiones k-means

Se ha concluido que:

- I. Hay tendencia de agrupación, con lo que es válido realizar este tipo de estudio.
- II. El número óptimo de grupos es 2.
- III. Esos dos grupos separan *pop*, *metal*, *disco*, *hiphop* y en menor medida *rock* de *clásica*, *jazz* y en menor medida *blues* y *country*. *Reggae* esta casi igualmente representado en los dos grupos.
- IV. Las variables que más contribuyen a esta separación son, por orden, *spectral_centroid*, *rolloff*, *zero_crossing_rate*, *chroma_sftf*.
- V. Los sujetos con mayores valores de las anteriores van al grupo 1 mientras que los que tienen menores valores caen en el 2.
- VI. *Beats* y *tempo* no contribuyen a separar.
- VII. Lo anterior cuadra con el análisis exploratorio de datos realizado previamente.
- VIII. La validación del estudio de clusterización, mediante el gráfico de silueta no es ni bueno ni malo: una silueta media de 0.32.
- IX. Pero hay que tener cuidado con las interpretaciones realizadas, porque el conjunto de datos tiene muchos outliers y este algoritmo es muy sensible a la presencia de estos valores atípicos.
- X. En cualquier caso, por la naturaleza del estudio, se decide no hacer nada con esos valores, puesto que se estima que trabajarlos sería adulterar la propia variabilidad que tiene la música.
- XI. Y por ello se realiza el estudio de la sección siguiente.

Ahora bien, ¿qué me dicen los puntos IV y V? Es decir, ¿tiene sentido que esas variables separen por grupos y que *clásica* y *jazz* tengan menores valores que *metal*, *disco* y *hip hop*?

- La variable que más contribuyó a separar fue *spectral_centroid*. Según [16] donde se explican en detalle las variables estudiadas, un valor más alto del SC corresponde a una mayor energía de la señal que concentra las frecuencias más altas. Y musicalmente, la frecuencia se refiere al número de ciclos por segundo de la onda sonora.
Por tanto, sí es lógico que tengamos frecuencias más bajas en una orquesta, ya sea de *clásica* o de *jazz*, que en una banda de *metal*, donde la distorsión de las

guitarras y la propia batería estará provocando una mayor acumulación de frecuencias altas y por tanto mayor valor de SC.

- En la misma línea con el rolloff. Si lo hemos definido como un filtro diseñado para atenuar las frecuencias por encima o por debajo de un punto determinado, es lógico que el valor de ese filtro sea menor para *clásica* y *jazz*, estilos que no tienen estridentes cambios de frecuencia que en *metal* donde si nos podemos encontrar con grandes picos de ella.
- Y lo mismo para el *zero_crossing_rate*, definido como la velocidad a la que una señal cambia de positivo a cero a negativo o de negativo a cero a positivo.
- Realmente, viendo las definiciones, ya podemos intuir que son variables muy correlacionadas entre sí. De hecho, hemos visto que eran de las que más aportaban a la primera dimensión del PCA.
- Y tiene sentido también que *beats* y *tempo* no contribuyan a separar. Al final acaban definiendo lo que sería la “velocidad” de una canción. Y cada estilo tiene tanto sus “lentas” como sus “rápidas”.
Se podría estar tentado a pensar que, por ejemplo, el *jazz* y la *clásica* debieran tener menos *beats*.
Quien así lo crea, se haría un gran favor al escuchar el *Chasin’ The Trane* que el magnífico saxofonista (y posiblemente el mayor ídolo de quien escribe) *John Coltrane*, grabó en directo en el Village Vanguard en 1961 o *El Vuelo del Moscardón*, de *Rimsky-Korsakov*.

7.2 Conclusiones PAM

Se ha concluido que:

- I. Hay tendencia de agrupación, con lo que es válido realizar este tipo de estudio.
- II. El número óptimo de grupos es 2.
- III. Esos dos grupos separan *pop*, *metal*, *disco*, *hiphop*, *reggae* y *rock* de *clásica* y *jazz*, quedando casi igualmente representados en ambos grupos *blues* y *country*.
- IV. Las variables que más contribuyen a la separación son por orden sobre todo *mfcc* y luego *rolloff*, *spectral_bandwidth* y *spectral_centroid*.
- V. La validación del estudio de clusterización, mediante el gráfico de silueta no es ni bueno ni malo: una silueta media de 0.33.

En este caso, la variable que más ha separado ha sido el *mfcc* (las otras ya las hemos discutido anteriormente)

- Los MFCC (coeficientes cepstrales en las frecuencias de Mel), al final por representar texturas musicales de corta duración (en este set de datos teníamos nada menos que 20 para cada audio de 30 segundos que hemos reducido a una sola variable quedándonos con su media) codifican muy fielmente la información tímbrica y tonal, como vimos en [13]. Es decir, que en la misma línea que lo comentado para las otras variables, no es descabellado obtener valores altos para *metal* o *disco* y bajos para *clásica* y *jazz*.

7.3 Conclusiones PCA

Se ha concluido que:

- I. Es adecuado realizar un estudio PCA.
- II. Que el número óptimo de dimensiones es dos.
- III. Que a la primera dimensión contribuyen positivamente todas las variables salvo *rmse*, *beats* y *tempo*.
- IV. Que a la segunda dimensión contribuyen positivamente *beats* y *tempo*.
- V. Que la primera dimensión separa *disco*, *hiphop*, *metal* y *pop* de *blues*, *clásica* y *jazz* (mayores valores para los primeros). Y *country*, *reggae* y *rock* quedan a caballo entre ambas. Además, esta dimensión explica algo más de la mitad de la variabilidad.
- VI. Que la segunda dimensión lo que separa es los propios géneros en función de *tempo* y *beats*: las más rápidas o con más *beats* arriba del plano y las más lentas abajo.

Tal y como comentábamos en las conclusiones del k-means, aquí lo que ha ocurrido es que todas las variables de onda, que estaban muy correlacionadas, han aportado a generar la primera dimensión, mientras que las de velocidad (*tempo* y *beats*), también correlacionadas entre sí, han generado la segunda.

Una variable que no ha aportado absolutamente nada en ninguno de los estudios realizados ha sido el *rmse*. Y es bastante lógico, ya que lo relacionábamos en la definición con el volumen de la señal. Pero además hacíamos hincapié en que las canciones se habían recuperado de distintas fuentes, precisamente para evitar diferenciaciones por tipo o calidad de grabación.

7.4 Conclusiones HCPC

Los resultados del HCPC han sido muy interesantes, ya que han venido a afinar aún más la clusterización que habíamos obtenido por métodos anteriores: recordemos que pudimos separar por las variables de onda pero no por las de tiempo y aquí parece que, eliminando el ruido, cada uno de los dos grupos obtenidos se ha separado en otros dos, según tengan mayor o menor ritmo.

El primer grupo se lleva gran parte de jazz y clásica. El mayor representante del grupo 2 es el reggae y al 3 ya se van uniendo el *rock* el *disco*, el *hip hop* y el *metal*, para terminar con un grupo 4 muy pop y bastante metal.

El *blues* predomina más en la suma de los grupos 1 y 2, emparentándose así más con el *jazz* y la *clásica* (normal), al contrario que el *rock*, que predomina en los 3 y 4 emparentándose así más con *metal* o *pop* (también lógico).

Reggae y *Country* siguen estando casi perfectamente a caballo entre ambos.

Es una categorización muy bonita, porque parecería incluso un árbol genealógico de la música.

Es interesante también ver (figura 42) 1 los individuos que están más alejados del resto de grupos (en 1 *clásica*, en 2 *clásica* y *jazz*, en 3 *metal*, *rock*, *blues* y *pop* y en 4 mucho *pop* y algo de *reggae* y *hip-hop*) que comparado con lo anterior nos habla a las claras de la dificultad de realizar este clúster por estilos.

En definitiva, una separación por estilos muy parecida a la que ya teníamos (más antiguos vs más modernos) pero bastante más fina.

7.5 Conclusiones generales

1. En líneas generales, todos los algoritmos han sido capaces de separar con mayor o menor éxito los dos géneros más tradicionales (*clásica* y *jazz*) de los más novedosos (*pop*, *metal*, *hiphop* o *disco*).

2. ¿En qué manera responde todo lo visto a las preguntas que hacía en la introducción? ¿Tienen sentido los estilos?

A priori no parece descabellado que esos estilos más añejos se hayan podido categorizar en grupos diferentes que los más modernos, simplemente pensando en el tipo de música: mientras que los primeros son eminentemente instrumentales y suelen destacar vientos, piano y ciertas cuerdas tipo violines, violonchelos etc, los segundos suelen ser muy monolíticos en cuanto a formación de voz, bajo, guitarra/s y batería.

Por tanto, siendo las variables predominantes propiedades de onda, es lógico que unos que tienen voz y donde es típico que la formación esté tocando a la vez se diferencien

de otros donde aparte de la ausencia de cantantes, típicamente se basan en solos con sus correspondientes cambios de frecuencia.

3. Pero fijémonos en algo muy curioso: ningún método ha sido capaz de aislar dos géneros tan aparentemente antagónicos como el pop y el metal.

Pienso ahora en una divertida crónica [17] del miércoles 6 de Julio de 2022, festival MadCool. Coincidían Metallica y Carly Rae Jepsen. El reportero narra que grupos de fans de la segunda se acercaron a puestas heavys para animarlos a que no fueran estrechos de miras y los acompañaran al concierto de la canadiense.

Terminaron todos juntos, cantando primero “Nothing Else Matters” y luego “Call Me Maybe”.

Solamente puedo definir eso como mágico. Como música de verdad: arte que une.

Y me encantaría que los protagonistas de la anécdota supieran, que es que realmente, a la Estadística también le ha dado igual que fuerais fans de unos o de otra: con todas las cosas que creéis que os diferencian, los algoritmos os han ido a llevar al mismo clúster.

4. Por tanto, no hemos encontrado evidencias estadísticamente significativas, que justifiquen las adhesiones a un estilo, pero aún más importante, contraposiciones a otro género. Nuestro estudio le dice al adolescente de esta generación, que las variables físicas estudiadas no explican que se enamore de un estilo en lucha con el género que escuchaba el adolescente que alguna vez fue su padre.

Nos mantenemos por tanto en la idea inicial: la música se disfruta, independientemente del género al que se le haya querido categorizar. Y posiblemente cuanto más se extienda esta idea, más rica será la música del futuro, hecha por intérpretes que no teman reconocer que disfrutaban igualmente bailando una cumbia que marcándose un air guitar al ritmo de cualquier canción heavy. Porque en este caso no tendrán problemas en quedarse con **lo mejor de cada cultura** y crear arte **nuevo y excitante**: esas apasionantes cosas que **solamente pueden nacer del mestizaje**.

7.6 Hipotética investigación futura

Uno de los grandes límites de este estudio es que ha partido de una categorización de géneros muy simple: 10 estilos. Y ya hemos visto [11] que en Spotify hay más de 3000. Tomando como ejemplo el pop, podríamos encontrar una diversidad tan enorme como adult-pop, art-pop, bubblegum pop, c-pop, dance-pop, dream-pop, power pop, progressive pop, synth-pop y un largo etcétera. Quizás este estudio sería mucho más rico (y sí, mucho más complicado) habiendo partido de más canciones y con una categorización más granular para cada canción.

REFERENCIAS BIBLIOGRÁFICAS

- [1] Sturm, B (2012). *A Survey Of Evaluation in Music Genre Recognition*, Aalborg University
- [2] Glenn McDonald. Géneros de Spotify https://verne.elpais.com/verne/2020/12/02/articulo/1606909760_903985.html
- [3] Glenn McDonald. Entrevista <https://www.mondosonoro.com/blog-musica/glenn-mcdonald-spotify/>
- [4] Ghosal, A, Chakraborty, R, Chandra, B & Kumar, S (2012). *Music Classification based on MFCC Variants and Amplitude Variation Pattern: A Hierarchical Approach*
- [5] Li, T, & Chan, A (2011). *Genre classification and the invariance of MFCC features to Key and Tempo*
- [6] Banitalebi-Dehkordi, M & Banitalebi-Dehkordi, A. *Music Genre Classification Using Spectral Analysis and Sparse Representation Of The Signals*
- [7] Desafío Kaggle <https://www.kaggle.com/insiyeh/musicfeatures>
- [8] Greenberg, D, Kosinski, M, Stillwell, D & Monteiro, B (mayo, 2016). *The Song Is You: Preferences for Musical Attribute Dimensions Reflect Personality*.
- [9] Laurie, T (2014). *Music Genre as Method*
- [10] Ferrero, R. (2022). *Análisis de Agrupación Clúster*.
- [11] Ferrero, R. (2022). *Ordenación y Reducción de la Dimensionalidad*.
- [12] Kassambara, A. (2017). *Practical Guide To Cluster Analysis in R*.
- [13] Kassambara, A. (2017). *Practical Guide To Principal Component Methods in R*.
- [14] ISMIR <https://ismir.net/>
- [15] Tutorial librería para extraer variables <https://librosa.org/doc/latest/index.html>
- [16] Explicación variables <https://www.sciencedirect.com/topics/engineering/spectral-centroid#:~:text=2.-,Spectral%20centroid,being%20concentrated%20within%20higher%20frequencies.>
- [17] Crónica MadCool <https://jenesaipop.com/2022/07/07/437234/metallica-carly-mad-cool/>

CAPÍTULO 8: ANEXOS

8.1 Código R para k-medias

```
## -----  
## librerías  
library(tidyverse)  
library(tidyr)  
library(ggplot2)  
library(clustertend)  
library(pheatmap)  
library(factoextra)  
library(cluster)  
library(NbClust)  
library(clValid)  
library(GGally)  
library(flexclust)  
library(FeatureImpCluster)  
library(ggrepel)  
library(cluster)  
library(pheatmap)  
library(rstatix)  
  
## -----  
## Cargamos los datos  
music <- read.csv("data.csv", sep=",")  
head(music)  
  
## -----
```

```
## Manipulación de datos
```

```
##Nos interesa que el nombre de cada fila sea realmente el nombre del corte musical
```

```
music <- music %>% remove_rownames %>% column_to_rownames(var="filename")
```

```
## La columna label es un factor:
```

```
music$label <- as.factor(music$label)
```

```
## Agrupamos los mfcc (media)
```

```
music$mfcc <- (music$mfcc1+music$mfcc2+music$mfcc3+music$mfcc4+music$mfcc5+music$mfcc6+  
              music$mfcc7+music$mfcc8+music$mfcc9+music$mfcc10+music$mfcc11+  
              music$mfcc12+music$mfcc13+music$mfcc14+music$mfcc15+music$mfcc16+  
              music$mfcc17+music$mfcc18+music$mfcc19+music$mfcc20)/20
```

```
## dataframe final
```

```
music_ok <- music %>%
```

```
  select(label,tempo,beats,chroma_stft,rmse,spectral_centroid,spectral_bandwidth,rolloff,  
         zero_crossing_rate, mfcc)
```

```
## -----
```

```
## Estandarización de los datos
```

```
music_std <- scale(music_ok[,-1])
```

```
## -----
```

```
## Estudio de outliers
```

```
long <- pivot_longer(as.data.frame(music_std) ,cols=1:9)
```

```
ggplot(long, aes (x=name, y=value)) +  
  geom_boxplot() +  
  theme(axis.text.x = element_text(angle=90))
```

```
## -----
```

```
## Análisis exploratorio de datos
```

```
# Gráficas de label en función de varias features
```

```
music_ok %>%
```

```
  group_by(label) %>%
```

```
  get_summary_stats(rmse,  
                    type = "mean_sd")
```

```
ggboxplot( x = "label", y = "beats",
```

```
  data = music_ok,
```

```
  add = c("mean"),
```

```
  add.params= list (color="red"))
```

```
ggboxplot( x = "label", y = "spectral_centroid",
```

```
  data = music_ok,
```

```
  add = c("mean"),
```

```
  add.params= list (color="red"))
```

```
ggboxplot( x = "label", y = "rolloff",
```

```
  data = music_ok,
```

```
  add = c("mean"),
```

```
  add.params= list (color="red"))
```

```
ggboxplot( x = "label", y = "zero_crossing_rate",
```

```
  data = music_ok,
```

```
  add = c("mean"),
```

```
  add.params= list (color="red"))
```

```
## -----
```

```
## Validez de análisis clúster
```

```
## Estadística de Hopkins
```

```
set.seed(123)
```

```
hopkins(music_std, n = nrow(music_std) - 1)
```

```
## Evaluación visual de tendencia (VAT)
```

```
music_dist <- dist(music_std)
```

```
fviz_dist(music_dist, lab_size = .1, show_labels = FALSE )
```

```
## -----
```

```
## Elección del número óptimo de grupos
```

```
#Validación relativa:
```

```
  NbClust(music_std, method = "kmeans", min.nc = 2, max.nc = 10)
```

```
#Según criterio de estabilidad
```

```
stab <- clValid(scale(music_ok[, -1]),
```

```
               nClust = 2:10, maxitems = 1000, clMethods = "kmeans",
```

```
               validation = "stability")
```

```
optimalScores(stab)
```

```
## -----
```

```
## Creación de los algoritmos
```

```
music_km2 <- kmeans(music_std, centers = 2)
```

```
music_km10 <- kmeans(music_std, centers = 10)
```

```
## -----
```

```
## Representación de los grupos para 10
```

```
## -----
```

```
## Tabla de clasificación para 10
```

```
table(music_ok$label, music_km10$cluster)
```

```
## -----
```

```
## Representación de los grupos para 2
```

```
## Tamaño
```

```
music_km2$size
```

```
## Centroides
```

```
music_km2$centers
```

```
## lo mismo pero con las variables sin escalar
```

```
aggregate(music_ok[,-1], list(music_km2$cluster), mean)
```

```
aggregate(music_ok[,-1], list(music_km2$cluster), sd)
```

```
## gráfico de lo anterior para cada grupo
```

```
set.seed(123)
```

```
par(mar=c(1,1,1,1))
```

```
cl_kcca <- as.kcca(music_km2, music_std)
```

```
barplot(cl_kcca)
```

```
## gráfico de lo anterior para cada variable:
```

```
barplot(cl_kcca, bycluster = FALSE)
```

```
## -----
```

```
## Importancia de las variables
```

```
importance <- FeatureImpCluster(cl_kcca, as.data.table(music_std))
```

```
plot(importance)
```

```
## -----
```

```
## Tabla de clasificación
```

```
table(music_ok$label, music_km2$cluster)
```

```
## -----  
  
## Validación de silueta  
  
library(cluster)  
  
sk <- silhouette(music_km2$cluster, music_dist) # gráfico de silueta  
  
plot(sk, main="Silhouette plot - Kmeans 2 grupos",  
      cex.names=0.8, col=1:2, nmax=100)
```

8.2 Código R para PAM

```
## -----  
  
## librerías  
  
library(tidyverse)  
  
library(tidyr)  
  
library(ggplot2)  
  
library(clustertend)  
  
library(pheatmap)  
  
library(factoextra)  
  
library(cluster)  
  
library(NbClust)  
  
library(clValid)  
  
library(GGally)  
  
library(flexclust)  
  
library(FeatureImpCluster)  
  
library(ggrepel)  
  
library(cluster)  
  
library(pheatmap)
```

```
library(rstatix)
```

```
## -----
```

```
## Cargamos los datos
```

```
music <- read.csv("data.csv", sep=",")
```

```
head(music)
```

```
## -----
```

```
## Manipulación de datos
```

```
## Nos interesa que el nombre de cada fila sea realmente el nombre del corte musical
```

```
music <- music %>% remove_rownames %>% column_to_rownames(var="filename")
```

```
## La columna label es un factor:
```

```
music$label <- as.factor(music$label)
```

```
## Agrupamos los mfcc (media)
```

```
music$mfcc <- (music$mfcc1+music$mfcc2+music$mfcc3+music$mfcc4+music$mfcc5+music$mfcc6+  
              music$mfcc7+music$mfcc8+music$mfcc9+music$mfcc10+music$mfcc11+  
              music$mfcc12+music$mfcc13+music$mfcc14+music$mfcc15+music$mfcc16+  
              music$mfcc17+music$mfcc18+music$mfcc19+music$mfcc20)/20
```

```
## dataframe final
```

```
music_ok <- music %>%
```

```
  select(label,tempo,beats,chroma_stft,rmse,spectral_centroid,spectral_bandwidth,rolloff,  
         zero_crossing_rate, mfcc)
```

```
## -----
```

```
## Estandarización de los datos
```

```
music_std <- scale(music_ok[,-1])
```

```
## -----
```

```
## Estudio de outliers
```

```
long <- pivot_longer(as.data.frame(music_std) ,cols=1:9)
```

```
ggplot(long, aes (x=name, y=value)) +  
  geom_boxplot() +  
  theme(axis.text.x = element_text(angle=90))
```

```
## -----
```

```
## Validez de análisis clúster
```

```
## Estadística de Hopkins
```

```
set.seed(123)
```

```
hopkins(music_std, n = nrow(music_std) - 1)
```

```
## Evaluación visual de tendencia (VAT)
```

```
music_dist <- dist(music_std)
```

```
fviz_dist(music_dist, lab_size =.1, show_labels = FALSE )
```

```
## PAM
```

```
## -----
```

```
## Número de grupos
```

```
## Método silueta
```

```
fviz_nbclust(music_std, pam, method = "silhouette") +  
  geom_vline(xintercept = 2, linetype = 2) +  
  labs(subtitle = "Silhouette Method")
```

```
## Método Elbow
```

```
fviz_nbclust(music_std, pam, method = "wss") +  
  geom_vline(xintercept = 2, linetype = 2) +  
  labs(subtitle = "Elbow Method")
```

```

## Método gap_stat
fviz_nbclust(music_std, pam, method = "gap_stat") +
  labs(subtitle = "Gap statistic Method")

## -----

## Algoritmo PAM

music_pam2 <- pam(music_std, 2)
print(music_pam2)

## -----

## Tabla comparativa
table(music_ok$label, music_pam2$cluster)

## -----

## Importancia de las variables
set.seed(123)
par(mar=c(1,1,1,1))
cl_kcca <- as.kcca(music_pam2, music_std)
barplot(cl_kcca)

## incluso para cada variable
barplot(cl_kcca, bycluster = FALSE)
importance <- FeatureImpCluster(cl_kcca, as.data.table(music_std))
plot(importance)

## -----

## Validación
library(cluster)
sk <- silhouette(music_pam2$cluster, music_dist) # gráfico de silueta
plot(sk, main="Silhouette plot - PAM 2 grupos",

```

```
cex.names=0.8, col=1:2, nmax=100)
```

8.3 Código R para PCA

```
## -----
```

```
## librerías
```

```
library(tidyverse)
```

```
library(tidyr)
```

```
library(clustertend)
```

```
library(pheatmap)
```

```
library(factoextra)
```

```
library(cluster)
```

```
library(NbClust)
```

```
library(clValid)
```

```
library(GGally)
```

```
library(flexclust)
```

```
library(FeatureImpCluster)
```

```
library(ggrepel)
```

```
library(cluster)
```

```
library(pheatmap)
```

```
library(rstatix)
```

```
library(tidyr)
```

```
library(ggplot2)
```

```
library(ggrepel)
```

```
library(cluster)
```

```
library(rstatix)
```

```
library(psych)
```

```
library(FactoMineR)
```

```
library(ggplot2)
```

```

## -----

## Cargamos los datos

music <- read.csv("data.csv", sep=",")

head(music)

## -----

## Manipulación de datos

## Nos interesa que el nombre de cada fila sea realmente el nombre del corte musical
music <- music %>% remove_rownames %>% column_to_rownames(var="filename")

## La columna label es un factor:
music$label <- as.factor(music$label)

## Agrupamos los mfcc (media)
music$mfcc <- (music$mfcc1+music$mfcc2+music$mfcc3+music$mfcc4+music$mfcc5+music$mfcc6+
              music$mfcc7+music$mfcc8+music$mfcc9+music$mfcc10+music$mfcc11+
              music$mfcc12+music$mfcc13+music$mfcc14+music$mfcc15+music$mfcc16+
              music$mfcc17+music$mfcc18+music$mfcc19+music$mfcc20)/20

## dataframe final
music_ok <- music %>%
  select(label,tempo,beats,chroma_stft,rmse,spectral_centroid,spectral_bandwidth,rolloff,
         zero_crossing_rate, mfcc)

## -----

## Estandarización de los datos

music_std <- scale(music_ok[,-1])

## -----

## Estudio de outliers

```

```
long <- pivot_longer(as.data.frame(music_std) ,cols=1:9)

ggplot(long, aes (x=name, y=value)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle=90))
```

```
## -----
```

```
## Validez de análisis clúster
```

```
## Evaluación visual de tendencia (VAT)
```

```
music_dist <- dist(music_std)
fviz_dist(music_dist, lab_size =.1, show_labels = FALSE )
```

```
## -----
```

```
## Dataset definitivo (quitamos el level)
```

```
music_ok2 <- music_ok[,-1]
summary(music_ok2)
music_std <- scale(music_ok[,-1])
```

```
## -----
```

```
## ¿Es adecuado hacer un PCA?
```

```
## Según la prueba de esfericidad de Bartlett:
```

```
  cortest.bartlett(music_ok2, n = 98)
```

```
## Según la adecuación muestral KMO:
```

```
KMO(cor(music_ok2))
```

```
## Positividad de la matriz de correlación
```

```
det(cor(music_ok2))
```

```

## -----

##Algoritmo PCa
res_pca <- PCA(music_ok2, graph=FALSE)
summary(res_pca)

## -----

## Selección del número de dimensiones

# Regla de Keisser-Guttman
get_eigenvalue(res_pca)

# % de varianza explicada**
fviz_screplot(res_pca)

# Gráfico de pantalla o sedimentación**:
fviz_screplot(res_pca)

# Análisis paralelo de Horn**
fa.parallel(music_ok2, fa = "pc")

## -----

## Estudio de las variables (características)
( var <- get_pca_var(res_pca) )

# Coordinada. Mapa de variables
fviz_pca_var(res_pca, repel=TRUE)

# grafico
fviz_pca_var(res_pca, col.var = "cos2",
  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
  repel = TRUE # Avoid text overlapping
)

```

```

# Calidad de la representación (cos2)
fviz_cos2(res_pca, choice = "var", axes = 1)
fviz_cos2(res_pca, choice = "var", axes = 2)

# Contribución a la representación (contrib)**
fviz_contrib(res_pca, choice = "var", axes = 1)
fviz_contrib(res_pca, choice = "var", axes = 2)

# Descripción de las dimensiones (dimdesc)
res_desc <- dimdesc(res_pca, axes = c(1,2), proba = 0.05)
res_desc$Dim.1
res_desc$Dim.2

## -----

## Estudio de los individuos (casos)
( var <- get_pca_ind(res_pca) )

# Coordinadas. Mapa de individuos
fviz_pca_ind(res_pca, repel=TRUE)

# gráfico
fviz_pca_ind(res_pca,
  geom.ind = "point", # show points only (nbut not "text")
  col.ind = music_ok$label, # color by groups
  palette = c("#00AFBB", "#00AFBB", "#FC4E07", "#E7B800", "#E7B800", "#00AFBB",
    "#E7B800", "#E7B800", "#FC4E07", "#FC4E07"),
  addEllipses = TRUE, # Concentration ellipses
  legend.title = "Groups"
)

```

```

# Calidad de la representación (cos2)

fviz_cos2(res_pca, choice = "ind", axes = 1,top = 50)
fviz_cos2(res_pca, choice = "ind", axes = 2,top = 50)
head(var$cos2, 50)

#Contribución a la representación (contrib)

fviz_contrib(res_pca, choice = "ind", axes = 1,top = 50)
fviz_contrib(res_pca, choice = "ind", axes = 2 ,top = 50)

## -----

## los individuos más importantes

#las 9 primeras variables/ 100 sujetos más importantes
fviz_pca_biplot(res_pca, repel = T,
                select.var = list(contrib = 9),
                select.ind = list(contrib = 100),
                ggtheme = theme_minimal())

```

8.4 Código R para HCPC

```

## -----

## librerías

library(tidyverse)
library(tidyr)
library(clustertend)
library(pheatmap)
library(factoextra)
library(cluster)
library(NbClust)
library(clValid)

```

```
library(GGally)
library(flexclust)
library(FeatureImpCluster)
library(ggrepel)
library(cluster)
library(pheatmap)
library(rstatix)
library(tidyr)
library(ggplot2)
library(ggrepel)
library(cluster)
library(rstatix)
library(psych)
library(FactoMineR)
library(ggplot2)
```

```
## -----
```

```
## Cargamos los datos
```

```
music <- read.csv("data.csv", sep=",")
```

```
head(music)
```

```
## -----
```

```
## Manipulación de datos
```

```
## Nos interesa que el nombre de cada fila sea realmente el nombre del corte musical
```

```
music <- music %>% remove_rownames %>% column_to_rownames(var="filename")
```

```
## La columna label es un factor:
```

```
music$label <- as.factor(music$label)
```

```
## Agrupamos los mfcc (media)
```

```
music$mfcc <- (music$mfcc1+music$mfcc2+music$mfcc3+music$mfcc4+music$mfcc5+music$mfcc6+
              music$mfcc7+music$mfcc8+music$mfcc9+music$mfcc10+music$mfcc11+
              music$mfcc12+music$mfcc13+music$mfcc14+music$mfcc15+music$mfcc16+
              music$mfcc17+music$mfcc18+music$mfcc19+music$mfcc20)/20
```

```
## dataframe final
```

```
music_ok <- music %>%
```

```
  select(label,tempo,beats,chroma_stft,rmse,spectral_centroid,spectral_bandwidth,rolloff,
         zero_crossing_rate, mfcc)
```

```
## -----
```

```
## Estandarización de los datos
```

```
music_std <- scale(music_ok[, -1])
```

```
## -----
```

```
## Estudio de outliers
```

```
long <- pivot_longer(as.data.frame(music_std) ,cols=1:9)
```

```
ggplot(long, aes (x=name, y=value)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle=90))
```

```
## -----
```

```
## Validez de análisis clúster
```

```
## Evaluación visual de tendencia (VAT)
```

```
music_dist <- dist(music_std)
```

```
fviz_dist(music_dist, lab_size =.1, show_labels = FALSE )
```

```
## -----
```

```
## Dataset definitivo (quitamos el level)
```

```
music_ok2 <- music_ok[, -1]
```

```
summary(music_ok2)
```

```
music_std <- scale(music_ok[, -1])
```

```
## -----
```

```
## ¿Es adecuado hacer un PCA?
```

```
## Según la prueba de esfericidad de Bartlett:
```

```
cortest.bartlett(music_ok2, n = 98)
```

```
## Según la adecuación muestral KMO:
```

```
KMO(cor(music_ok2))
```

```
# Positividad de la matriz de correlación
```

```
det(cor(music_ok2))
```

```
## -----
```

```
## Algoritmo HCPC
```

```
music.pca <- PCA(music_ok2, ncp = 2, graph = FALSE)
```

```
music.hcpc <- HCPC(music.pca, graph = FALSE)
```

```
## -----
```

```
## Variables cuantitativas que describen mejor cada grupo**
```

```
music.hcpc$desc.var$quanti.var
```

```
## -----
```

```
## Descripción de cada grupo por las variables cuantitativas**
```

```
music.hcpc$desc.var$quanti
```

```
## -----
```

```
## Dimensiones principales que están más asociadas con los grupos**
```

```
music.hcpc$desc.axes
```

```
## -----
##Individuos representativos de cada grupo**
music.hcpc$desc.ind$para

## -----
##Individuos más lejanos a otros grupos de cada grupo**
music.hcpc$desc.ind$dist

## -----
##Tabla de clasificación del algoritmo
tabla_a <- music.hcpc$data.clust
table(music_ok$label, tabla_a$clust)
```

8.5 Figuras k-means

- Among all indices:
- 9 proposed 2 as the best number of clusters
- 6 proposed 3 as the best number of clusters
- 2 proposed 4 as the best number of clusters
- 1 proposed 6 as the best number of clusters
- 1 proposed 8 as the best number of clusters
- 3 proposed 9 as the best number of clusters
- 2 proposed 10 as the best number of clusters

***** Conclusion *****

- According to the majority rule, the best number of clusters is 2

Figura 21. Validación relativa k-means (Fuente: elaboración propia)

```
##          Score Method Clusters
## APN 0.04910638 kmeans          2
## AD  2.44611508 kmeans          10
## ADM 0.18456464 kmeans          2
## FOM 0.64179929 kmeans          10
```

Figura 22. Criterio estabilidad k-means (Fuente: elaboración propia)

```
## [1] 577 423
```

Figura 23. Representación grupos k-means k=2 (Fuente: elaboración propia)

```
##          tempo          beats chroma_stft          rmse spectral_centroid
## 1  0.01597047  0.05937615   0.5733982  0.4706273         0.6703875
## 2 -0.02178478 -0.08099300  -0.7821530 -0.6419668        -0.9144529
## spectral_bandwidth  rolloff zero_crossing_rate          mfcc
## 1          0.6031350  0.6653019          0.5893273  0.5534347
## 2          -0.8227161 -0.9075158          -0.8038815 -0.7549216
```

Figura 24. Centroides grupos k-means k=2 (Fuente: elaboración propia)

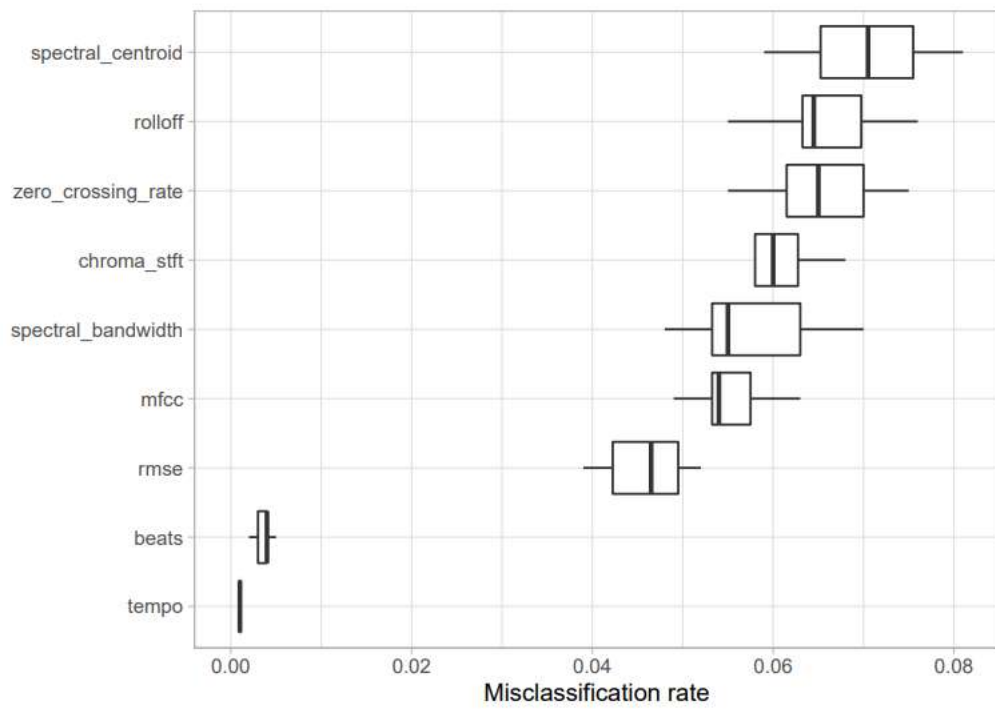


Figura 25. Importancia variables k-means k=2 (Fuente: elaboración propia)

8.6 Figuras PAM

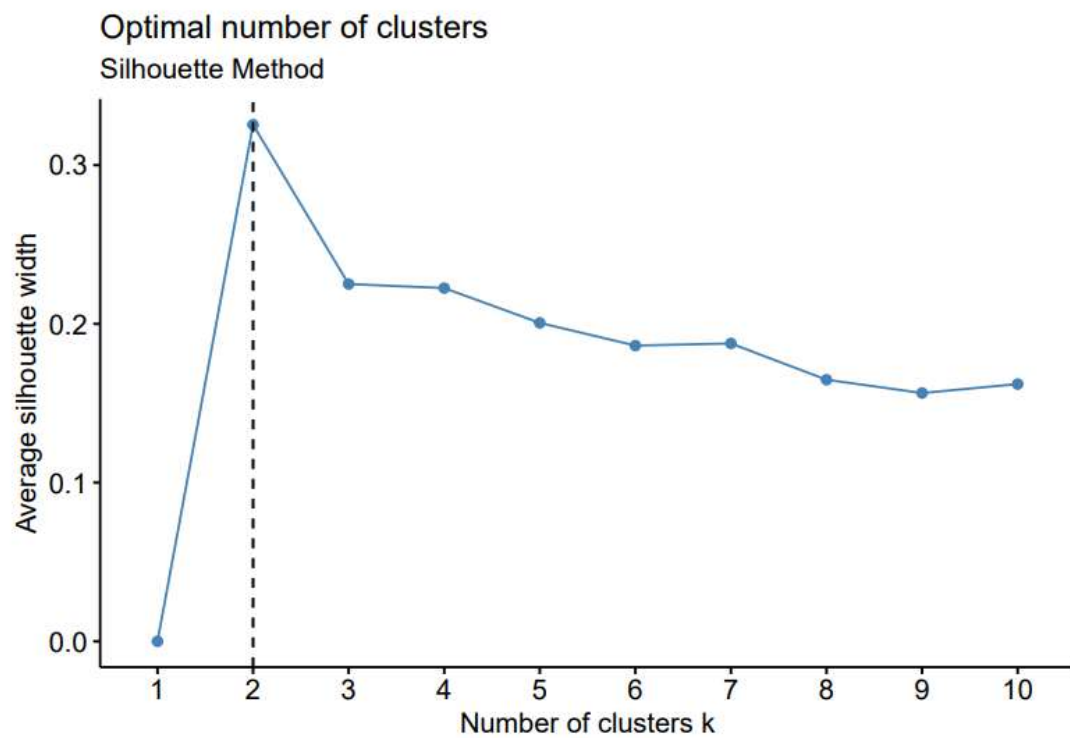


Figura 26. Número óptimo grupos PAM (Fuente: elaboración propia)

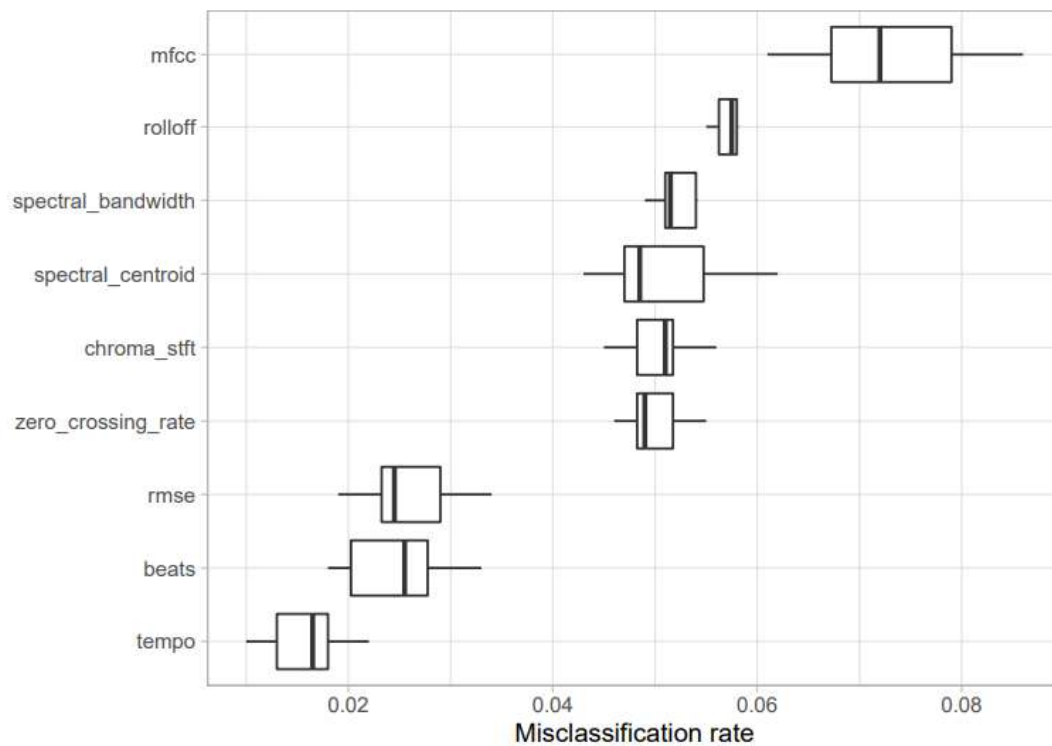


Figura 27. Importancia variables PAM (Fuente: elaboración propia)

8.7 Figuras PCA

```
## $chisq
## [1] 14331.61
##
## $p.value
## [1] 0
##
## $df
## [1] 36
```

Figura 28. Esfericidad Bartlett (Fuente: elaboración propia)

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor(music_ok2))
## Overall MSA = 0.7
## MSA for each item =
##          tempo          beats          chroma_stft          rmse
##          0.49          0.49          0.80          0.71
## spectral_centroid spectral_bandwidth          rolloff zero_crossing_rate
##          0.70          0.82          0.80          0.64
##          mfcc
##          0.68
```

Figura 29. KMO (Fuente: elaboración propia)

```
## [1] 5.567109e-07
```

Figura 30. Determinante matriz correlación (Fuente: elaboración propia)

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1 4.914290050      54.60322278      54.60322
## Dim.2 1.971914607      21.91016230      76.51339
## Dim.3 1.054600215      11.71778016      88.23117
## Dim.4 0.549618571       6.10687301      94.33804
## Dim.5 0.323427269       3.59363632      97.93167
## Dim.6 0.135518319       1.50575910      99.43743
## Dim.7 0.030380863       0.33756514      99.77500
## Dim.8 0.015424599       0.17138443      99.94638
## Dim.9 0.004825507       0.05361675     100.00000
```

Figura 31. Autovalores (Fuente: elaboración propia)

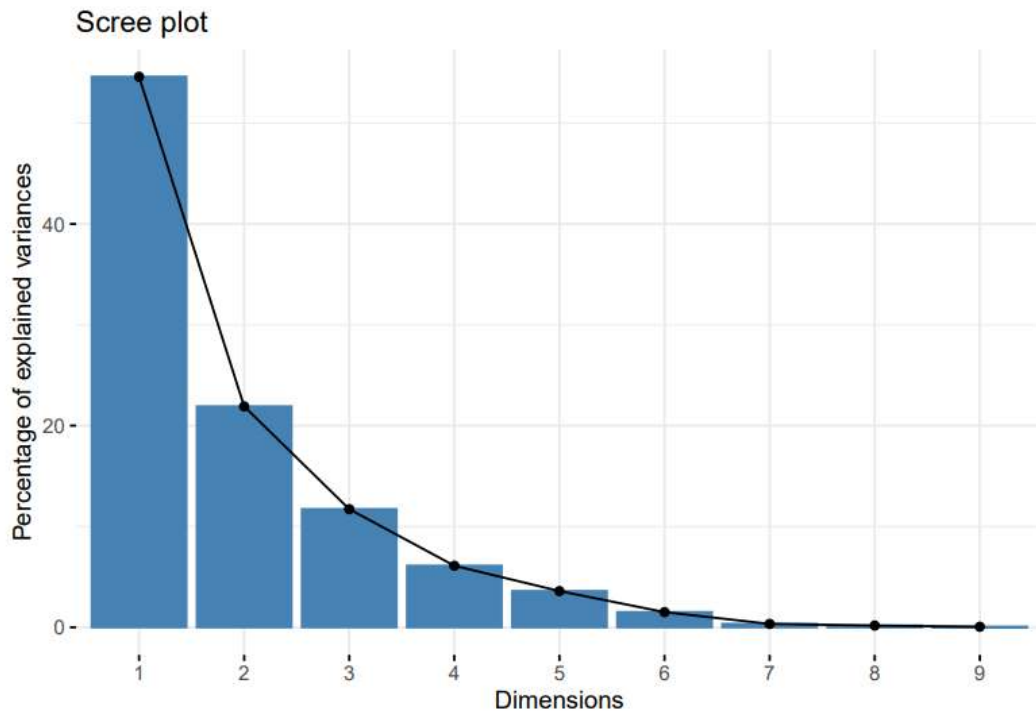
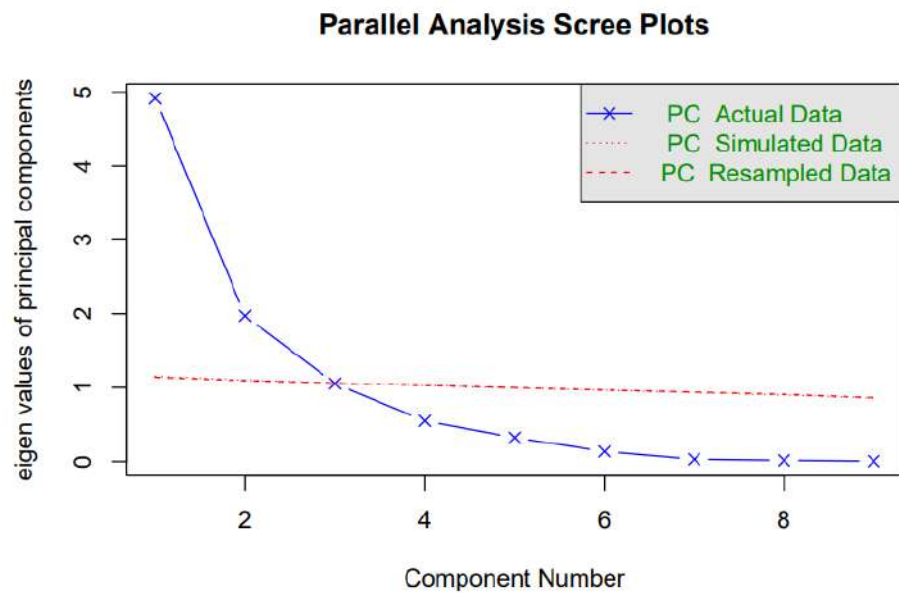


Figura 32. Gráfico sedimentación (Fuente: elaboración propia)



Parallel analysis suggests that the number of factors = NA and the number of components = 2

Figura 33. Análisis paralelo Horn (Fuente: elaboración propia)

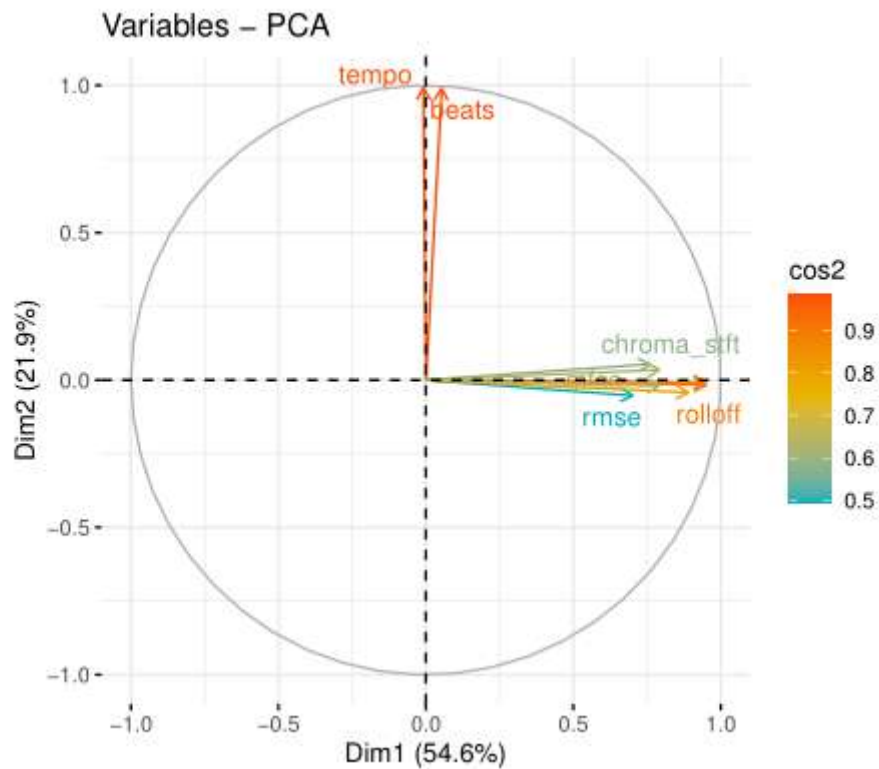


Figura 34. Coordenadas: mapa variables 2 (Fuente: elaboración propia)

```
## $quanti
##
## correlation      p.value
## rolloff          0.9493654 0.000000e+00
## spectral_centroid 0.9481403 0.000000e+00
## spectral_bandwidth 0.8892098 0.000000e+00
## zero_crossing_rate 0.7919521 4.749806e-216
## mfcc              0.7908016 5.410600e-215
## chroma_stft       0.7591260 2.741636e-188
## rmse              0.7011027 9.463907e-149
##
## attr(,"class")
## [1] "condes" "list"
```

Figura 35. Descripción dimensiones PCA dim1 (Fuente: elaboración propia)

```
## $quanti
##      correlation p.value
## tempo    0.9914678      0
## beats    0.9896747      0
##
## attr(,"class")
## [1] "condes" "list"
```

Figura 36. Descripción dimensiones PCA dim2(Fuente: elaboración propia)

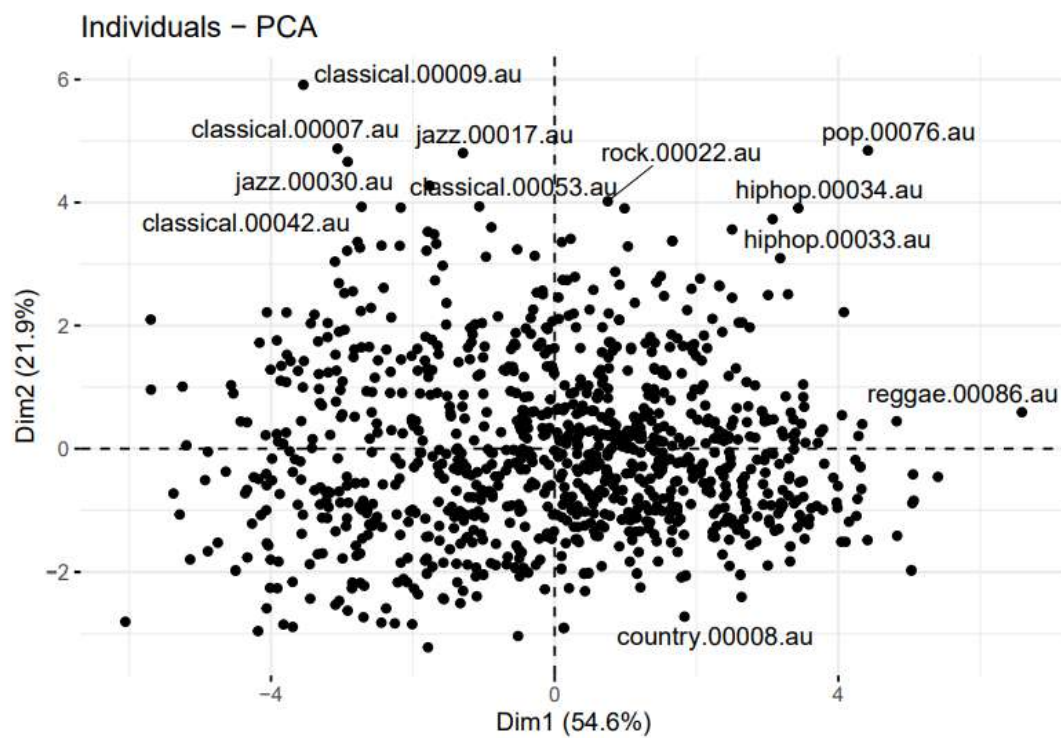


Figura 37. Coordenadas: mapa de individuos 1(Fuente: elaboración propia)

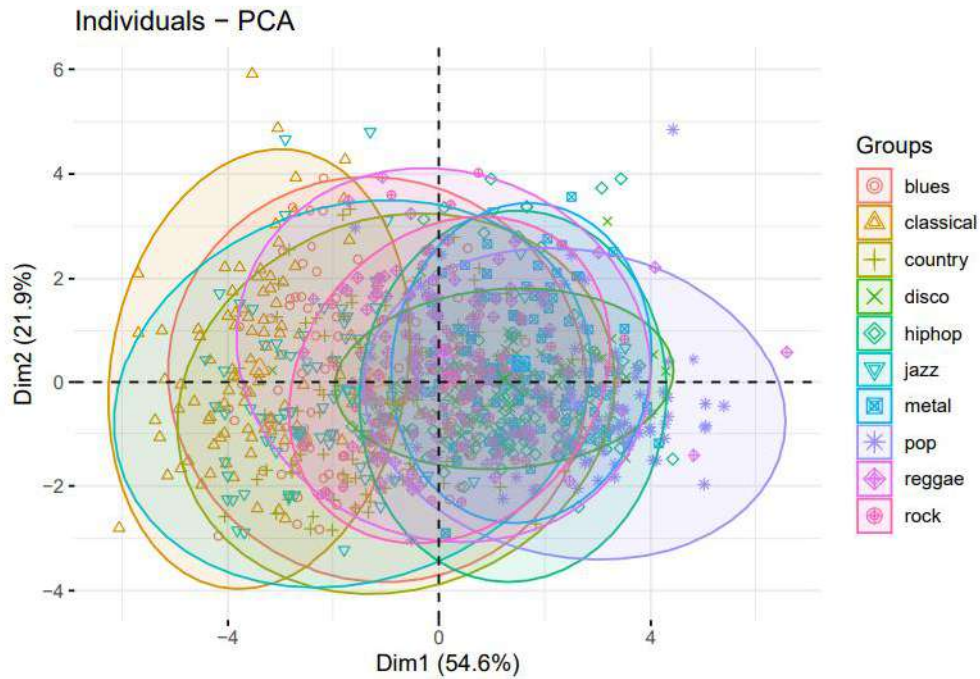


Figura 38. Coordenadas: mapa de individuos 10 colores (Fuente: elaboración propia)

8.8 Figuras HCPC

##		Eta2	P-value
##	spectral_centroid	0.7349932	1.308158e-286
##	rolloff	0.7247774	1.969458e-278
##	spectral_bandwidth	0.6297880	2.452460e-214
##	zero_crossing_rate	0.5378926	2.037528e-166
##	mfcc	0.4920766	5.436986e-146
##	chroma_stft	0.4833990	2.485037e-142
##	beats	0.4693142	1.609414e-136
##	tempo	0.4631469	5.044112e-134
##	rmse	0.3805556	4.068707e-103

Figura 39. Variables cuantitativas que mejor definen los grupos

```

##
## Link between the cluster variable and the quantitative variables
## =====
##              Eta2          P-value
## Dim.1 0.8017735  0.000000e+00
## Dim.2 0.4650048  8.992642e-135
##
## Description of each cluster by quantitative variables
## =====
## $'1'
##              v.test Mean in category Overall mean sd in category Overall sd
## Dim.2 -9.758981      -0.7844625 1.012412e-14      1.115584  1.404249
## Dim.1 -22.143893     -2.8100114 3.987366e-15      1.106620  2.216820
##              p.value
## Dim.2 1.688440e-22
## Dim.1 1.194598e-108
##
## $'2'
##              v.test Mean in category Overall mean sd in category Overall sd
## Dim.2 20.495706      2.08643 1.012412e-14      0.9603464  1.404249
## Dim.1 -6.609099     -1.06211 3.987366e-15      1.3296966  2.216820
##              p.value
## Dim.2 2.351503e-93
## Dim.1 3.866661e-11
##
## $'3'
##              v.test Mean in category Overall mean sd in category Overall sd
## Dim.1 4.604078      0.4296226 3.987366e-15      0.7672187  2.216820
## Dim.2 -7.738728     -0.4574332 1.012412e-14      0.8548886  1.404249
##              p.value
## Dim.1 4.142978e-06
## Dim.2 1.004162e-14
##
## $'4'
##              v.test Mean in category Overall mean sd in category Overall sd
## Dim.1 22.29024      2.744435 3.987366e-15      0.8857044  2.21682
##              p.value
## Dim.1 4.594742e-110

```

Figura 40. Dimensiones Principales más Asociadas con los Grupos

```

## Cluster: 1
## classical.00002.au      blues.00021.au      jazz.00040.au      jazz.00046.au
##      0.1248626      0.1357739      0.1638671      0.2187678
##      jazz.00039.au
##      0.2216342
## -----
## Cluster: 2
## reggae.00066.au blues.00002.au reggae.00005.au jazz.00054.au reggae.00097.au
##      0.05831445      0.08737625      0.11665903      0.18093763      0.24711110
## -----
## Cluster: 3
## disco.00039.au country.00011.au rock.00026.au blues.00065.au
##      0.1280006      0.1303881      0.1311801      0.1436623
## disco.00032.au
##      0.1735327
## -----
## Cluster: 4
## pop.00064.au metal.00053.au disco.00004.au hiphop.00077.au disco.00001.au
##      0.1380789      0.1440340      0.1837292      0.2639146      0.2825110

```

Figura 41. Individuos más Representativos de Cada Grupo

```

## Cluster: 1
## classical.00078.au classical.00055.au classical.00084.au classical.00063.au
##      6.894161      5.633543      5.361308      5.273740
##      blues.00093.au
##      5.244953
## -----
## Cluster: 2
## classical.00009.au classical.00007.au jazz.00017.au jazz.00030.au
##      6.737017      5.666549      5.536561      5.447313
## classical.00053.au
##      5.165062
## -----
## Cluster: 3
## metal.00066.au metal.00045.au blues.00071.au rock.00024.au pop.00049.au
##      3.623665      3.623665      3.353900      3.312055      3.213999
## -----
## Cluster: 4
## reggae.00086.au pop.00076.au pop.00051.au hiphop.00034.au pop.00039.au
##      6.247098      6.134194      4.974243      4.852670      4.842656

```

Figura 42. Individuos de Cada Grupo más Alejados de los Otros Grupos