¿Qué música escuchamos?

Grupo N°5: Flavia Felicioni, Adrian Marino, Claudio Collado

Resumen

Spotify es una aplicación multiplataforma empleada para la reproducción de música vía streaming. Cuenta con 345 millones de usuarios activos y 155 millones de usuarios suscritos, lo cual lleva a la generación de una enorme cantidad de datos con posibilidad de ser explotados para la obtención de conocimiento.
En particular spotifycharts provee información diaria y semanal de los artistas que forman parte del top 200 a nivel mundial así como también a nivel país. Tomando esto como base junto con datos obtenidos de la API de Spotify forman el conjunto de datos utilizados en este trabajo práctico

1. Introducción

El presente trabajo práctico tiene como finalidad principal la aplicación de conceptos y metodologías de análisis vistos en la primera parte de la materia *Data Mining* perteneciente a la *Maestría en explotación de datos y descubrimiento de conocimiento.*
Para su elaboración se toma como referencia la aplicación del proceso **KDD (Knowledge Discovery in Database)** ya que corresponde a la metodología de descubrimiento de conocimiento estudiado durante el cursado de la materia.

2. Preguntas Analizadas

A continuación se listan las preguntas utilizadas como guía para el desarrollo de este Trabajo Práctico:

- **Pregunta N°1:**
    a) *Aquellos temas que ingresaron al TOP 3 durante el periodo 2018 - 2020 ¿Cuántas semanas permanecieron* **en** *cada posición?*
    b) *Para aquellos temas que ingresaron al TOP 1 durante el periodo 2018-2020 ¿Su permanencia en esta posición es únicamente en cantidad de semanas consecutivas o existen temas que logran recuperar esta posición luego de su caída a posiciones inferiores?*
- **Pregunta N°2:**
    a) *Para los temas que ingresaron al TOP 10 durante el periodo 2018-2020 ¿Cuáles son los atributos (features) que más incidencia tienen para su permanencia en este TOP?*
    b) *De los atributos identificados anteriormente ¿Cómo es su comportamiento para aquellos temas que permanecen pocas y muchas semanas en este TOP?*
- **Pregunta N°3:** *Para aquellos temas que ingresaron al TOP 1 durante el periodo 2018-2020 y en aquellas semanas donde se producen cambios de temas ¿Como es el comportamiento de los atributos en estos cambios? ¿El tema que ingresa tiene valores de atributos similares al tema que desplaza?*
- **Pregunta N°4:** *Considerando el momento desde que un tema es publicado y llega al TOP 10 ¿Cuántas semanas transcurren? ¿Esto sucede rápidamente o existen temas publicados mucho tiempo atrás que igualmente logran ingresar? ¿Cuáles son los atributos de las canciones que llegaron más rápidamente al TOP 10?*

- **Pregunta N°5:** *Para aquellos temas que ingresaron al TOP 1 durante el periodo 2018-2020 y caen de esta posición ¿Cuántas semanas pasan hasta que llegan al TOP 5 , TOP 10 y no vuelven de él?*

3. Datos disponibles

El conjunto de datos entregados fue un backup de MongoDB en formato JSON. En total fueron tres colecciones de MongoDB que almacenan los siguientes conjuntos de datos:

- **Artist:** Incluye información de los artistas que obtuvieron alguna posición en el ranking durante el período analizado.
- **Artist_audio_features:** Guarda metadatos de las canciones que obtuvieron alguna posición en el ranking durante el período analizado.
- **Charts:** Tabla de rankeo de canciones (TOP 200) durante un periodo determinado

4. Análisis Inicial

Como punto de partida se ejecutaron las siguientes actividades en forma iterativa con el objetivo final de obtener el dataset en condiciones de ser utilizado para responder las preguntas guía indicadas anteriormente:

a) Análisis exploratorio de las colecciones en forma individual
b) Integración de las colecciones según un criterio específico

4.1. Análisis exploratorio de las colecciones en forma individual

Inicialmente, se realizaron tareas para simplificar las colecciones, renombrando campos y colecciones para tener datos más coherentes. Finalmente, nos quedamos con aquellos campos que contaban con información relevante para nuestro análisis:

- **Charts**: La colección **charts** la nombramos como **track_weekly_top_x**: Segmentamos la colección en top 1, 50, 100 y 200. Esto nos permitió realizar análisis manejando una menor cantidad de datos ya que dadas las limitaciones de hardware que teníamos, en algunos casos no contábamos con la memoria suficiente para cargar el dataset completo. Por otro lado eliminamos repetidos ya que todos los documentos estaban duplicados (Ej: Filtrando por la posición 1 teníamos 318 documentos en el TOP 1 cuando deberían ser aproximadamente 157). También encontramos varios temas para los cuales no se pudo encontrar su par en **track_features** teniendo que ser excluidos del análisis.
- **Artist Audio Features:** La colección **artist_audio_features** la renombramos a **track_features** y quitamos varios campos repetidos, en especial a nivel artista.

4.2. Integración de las colecciones según un criterio específico

Para realizar el análisis exploratorio necesario para contestar las preguntas planteadas, era necesario integrar las colecciones **track_weekly_top_x** y **track_features**.
El primer desafío fue pensar cuál era la clave de asociación entre ambas colecciones: En la colección **track_features** encontramos que había varias versiones de un mismo track en álbumes con distinto id, siendo estas re-ediciones del mismo álbum. Por ejemplo, encontramos un tema de Eminem que estaba en 83 álbumes con distinta clave, pero en realidad eran reediciones del mismo álbum. Finalmente, encontramos que la “clave única” a nivel tema era <artist, track, album_id>.

Luego, al analizar la colección **track_weekly_top_x** solo teníamos dos formas de identificar un track:

- <artist_name + track_name>
- <track_url>

La opción de usar la url de cada track era la más razonable, ya que esta identifica unívocamente cada track en **track_features**. Esta fue nuestra primera opción, pero luego de hacer el join encontramos que teníamos una pérdida de información del 37% para TOP 200, dado que para varios temas en **track_weekly_top_x** no encontramos el track correspondiente en **track_features**. Entendemos que hay una inconsistencia con alguno de los dos campos: track_features.url o track_weekly_top_x.track_url.
Por esto mismo optamos por la primera opción: usar la clave <artist_name + track_name>. Sabemos que en principio no es la mejor opción, pero dada la pérdida de información observada anteriormente optamos por esta alternativa.
Teniendo la clave de join se nos presentó otro desafío: Esta clave no es única en **track_features**. Dado esto, realizamos una agrupación por esta clave y realizamos la mediana para los campos numéricos y tomamos el valor de los categóricos (que era el mismo en todos los campos). Finalmente hicimos join segmentado por cantidad de posición como ya aclaramos anteriormente por un tema de restricciones de hardware, especialmente memoria RAM, quedando la siguientes colecciones listas para su análisis:

- track_weekly_top_10 join track_features
- track_weekly_top_50 join track_features
- track_weekly_top_100 join track_features
- track_weekly_top_200 join track_features

Con esta integración sólo tenemos una pérdida de información entre 6 y 9% lo cual es considerablemente menor al 37% de la primera opción analizada.

5. Resultados obtenidos como respuestas a las preguntas planteadas

Obtenidas las 4 colecciones definitivas estas fueron importadas a R para realizar los análisis específicos que permitan obtener las respuestas a las preguntas guía planteadas. A continuación se observan estas respuestas.

5.1. Pregunta N°1

a) ***Aquellos temas que ingresaron al TOP 3 durante el periodo 2018 - 2020 ¿Cuántas semanas permanecieron en cada posición?***

En el Gráfico N°1 se observa la cantidad de semanas que permaneció cada tema en una posición perteneciente al TOP 3

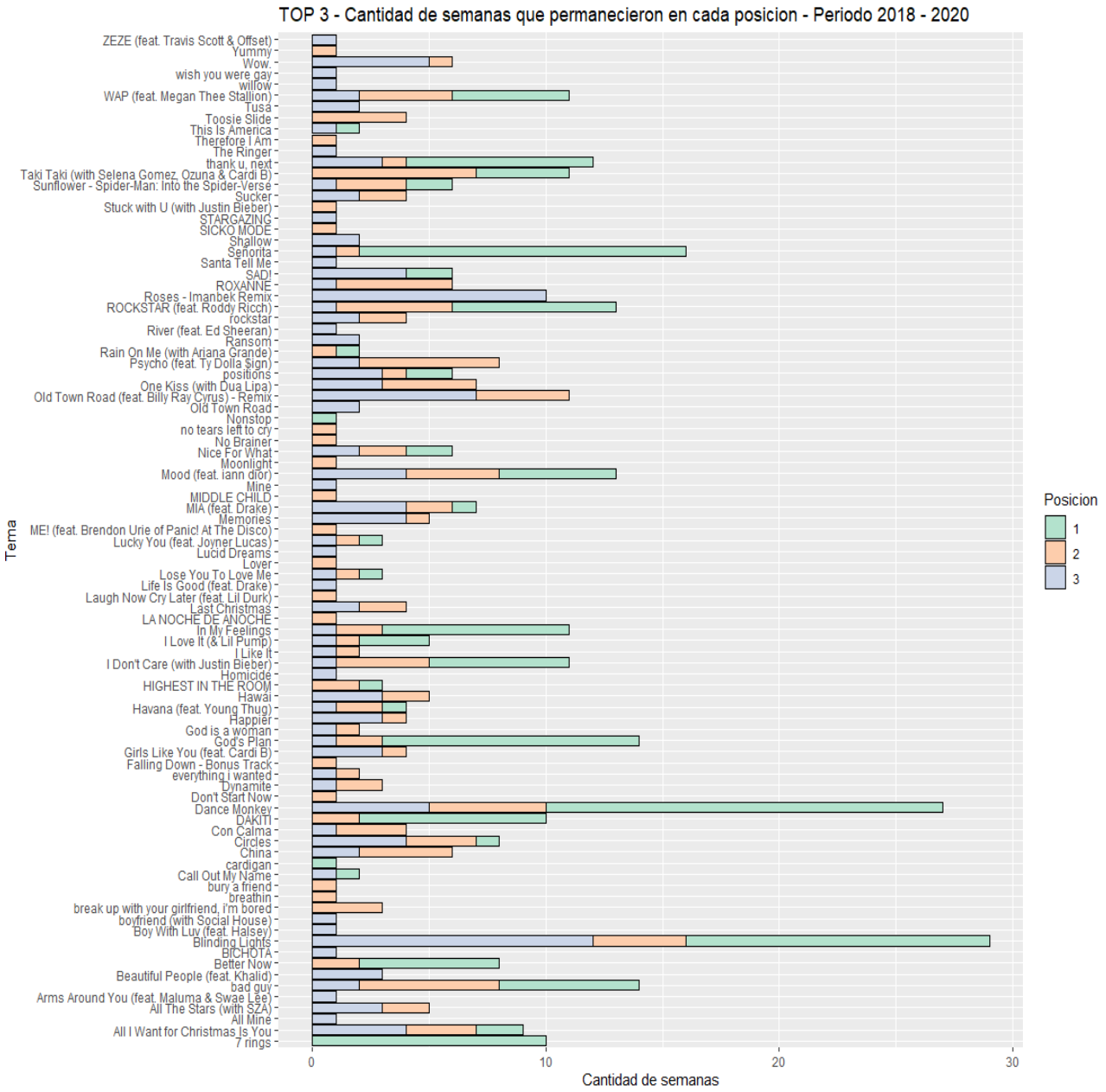

Gráfico N°1

Del Gráfico N°1 se obtienen las siguientes conclusiones principales:

- 91 temas formaron parte de este periodo en el TOP 3
- El periodo fue dominado por el grupo de temas compuesto por ***Blinding Lights*** y ***Dance Monkey*** ya que obtuvieron una considerable cantidad de semanas en cada posición del TOP 3.
- Luego del grupo de temas dominantes es interesante considerar el grupo de temas constituido por ***Señorita***, ***God's Plan***, ***Bad Guy***, ***Mood***, ***Rockstar***, ***Thank u next***, ***WAP***, ***Taki Taki***, ***In my feelings*** y ***I don't care*** ya que ocuparon las 3 posiciones y en total estuvieron más de 10 semanas.
  Una mención especial corresponde al tema ***7 Rings*** que logró permanecer 10 semanas en TOP 1 y ninguna semana en las posiciones 2 y 3.
- La permanencia de una considerable cantidad de semanas en el TOP 1 no se relaciona con una permanencia similar en las posiciones 2 y 3. En este sentido se observa un comportamiento disímil ya que algunos temas tienen una cantidad semejante de

semanas en cada posición y otros un comportamiento sesgado a una posición específica.

- Se observa una cantidad importante de temas que llegan y permanecen 1 sola semana, siendo la gran mayoría pertenecientes a las posiciones 2 y 3. Una mención especial corresponde a los temas ***Nonstop*** y ***Cardigan*** que permanecieron 1 semana en TOP 1 y ninguna semana en las posiciones 2 y 3

b) Para aquellos temas que ingresaron al TOP 1 durante el periodo 2018-2020 ¿Su permanencia en esta posición es únicamente en cantidad de semanas consecutivas o existen temas que logran recuperar esta posición luego de su caída a posiciones inferiores?

El Gráfico N°2 contiene la evolución semanal de cada tema en el TOP 1: Cada punto rojo corresponde a un único tema en una semana específica, por lo tanto puntos en una misma línea horizontal corresponden a un mismo tema que estuvo en TOP 1 en diferentes semanas (consecutivas o no). Los saltos de línea en forma vertical indica el cambio de TOP 1 entre diferentes temas.

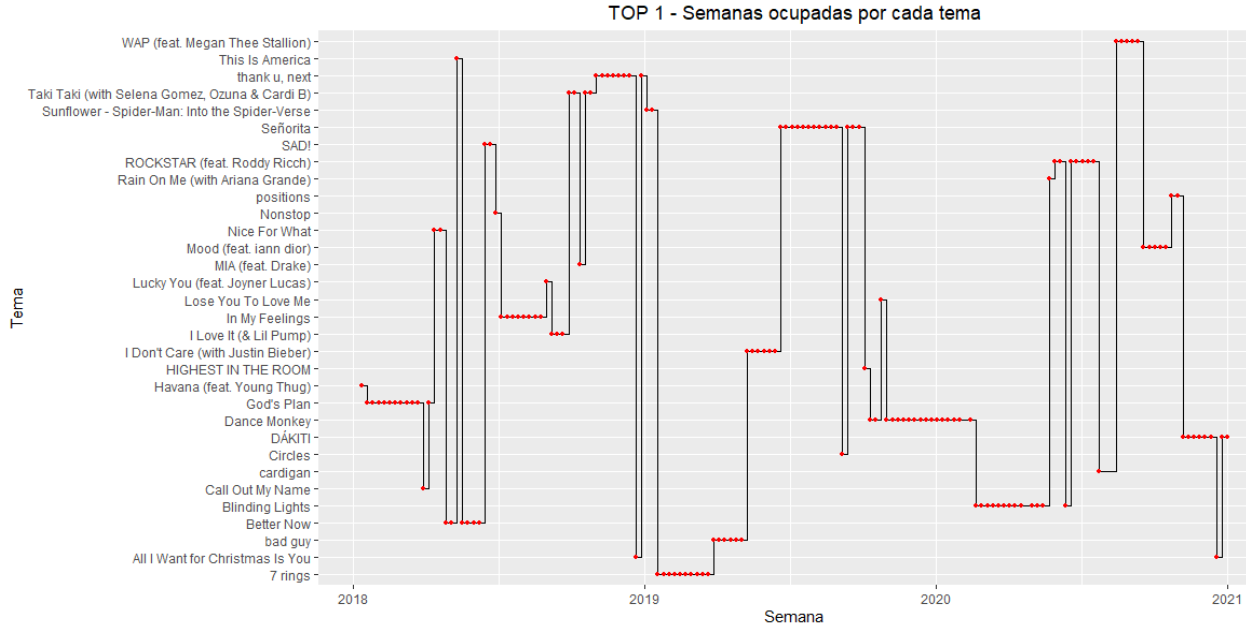

Gráfico N°2

Del análisis del Gráfico N°2 se obtienen las siguientes conclusiones principales:

- 32 temas formaron parte de este periodo en el TOP 1.
- De los 32 temas los siguientes 9 temas lograron recuperar el TOP 1 luego de perderlo: ***God's Plan, Better now, Taki Taki, Thank u next, Señorita, Dance monkey, Blinding lights, Rockstar*** y ***Dakiti***.
- En 8 de los 9 temas que perdieron el TOP 1 y lograron recuperarlo la pérdida de esta posición corresponde a 1 semana, lo cual evidencia que esta fue generada por un tema que logró un éxito puntual que lo llevó a TOP 1 esa semana y luego no pudo mantener la posición.
  Siguiendo con este análisis es de destacar el tema ***Blinding lights*** que logró recuperar el TOP 1 luego de transcurridas 3 semanas desde la pérdida.
- De los temas que lograron recuperar el TOP 1 la permanencia en esta posición luego de la recuperación es en su mayoría de 1 a 3 semanas, destacándose el tema ***Dance monkey*** que se mantuvo durante 11 semanas.

- En la semana de navidad del 2018 y 2020 el tema ***All I want for christmas is you*** llega al TOP 1, lo cual está relacionado con la festividad de ese periodo. Es importante tener esto en cuenta ya que permite saber con anticipación que el tema que se encuentre en el TOP 1 previo a la semana de navidad es probable que caiga de posición debido a esto.

5.2. Pregunta N°2

a) Para los temas que ingresaron al TOP 10 durante el periodo 2018-2020 ¿Cuáles son los atributos (features) que más incidencia tienen para su permanencia en este TOP?
b) De los atributos identificados anteriormente ¿Cómo es su comportamiento para aquellos temas que permanecen pocas y muchas semanas en este TOP?

Para contestar estas preguntas nos enfocamos en los siguientes criterios:

- **Criterio N°1**: Determinar cuáles son las características a tener en cuenta para que un artista permanezca en el TOP 1.
- **Criterio N°2**: Determinar cuáles son las características a tener en cuenta para mantenerse el mayor tiempo posible en el TOP 1 (medido en semanas).

Para definir qué características son más importantes bajo los criterios anteriormente definidos, utilizamos la función **importance** del algoritmo **random forest**. Esta nos permite ordenar las variables por orden de importancia. Luego, tomamos las 4 variables más importantes para realizar este análisis.

5.2.1 Criterio N°1

El Gráfico N°3 muestra los resultados del algoritmo Random Forest para las clases:
- TOP 1
- TOP 2 al TOP 10

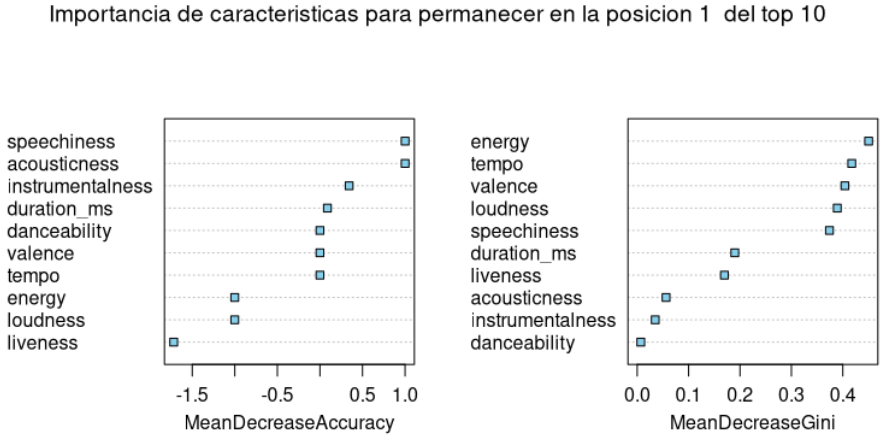

Gráfico N°3

De esta manera obtenemos el orden de importancia de las variables para separar estos dos grupos. Tenemos dos métricas distintas:

- **Mean decrease accuracy**: Mide el grado de precisión de la clasificación que aporta cada característica.

- **Mean decrease Gini**: Mide el grado de información que aporta cada característica a los nodos del árbol de decisión.

Para este análisis elegimos la métricas **Mean decrease Gini.**
En el Gráfico N°4 observamos el grado de cada característica en función de la posición en el TOP 10

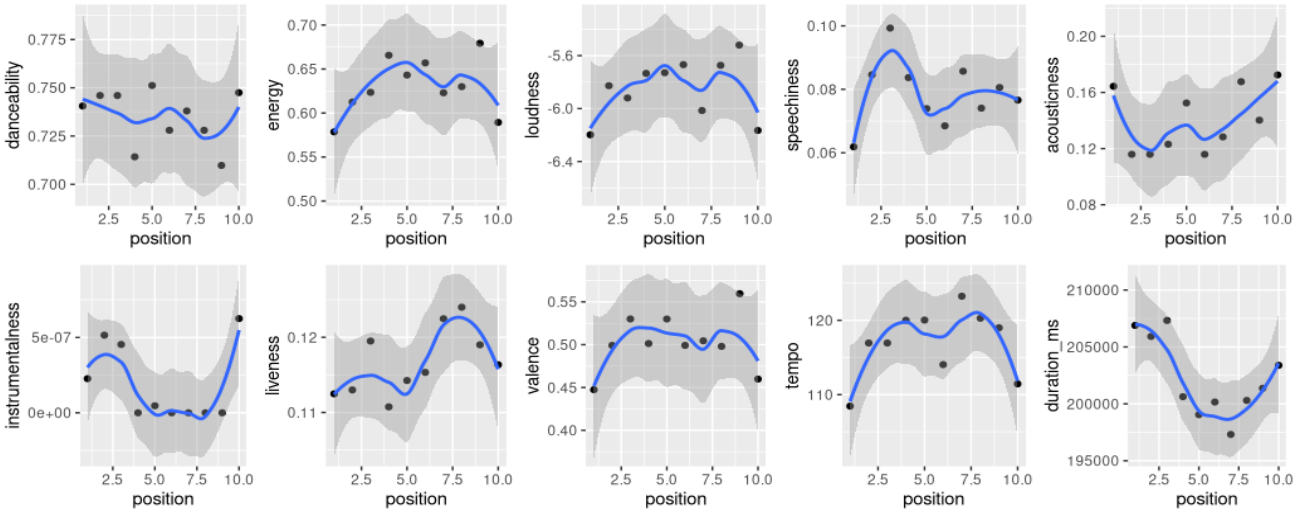

Gráfico N°4

Del Gráfico N°4 analizamos las siguientes cuatro características más importantes:

Energía:
- Se puede apreciar que la posición 1 es la de menor energía en el TOP 10, indicando que son temas musicales tranquilos, de poca intensidad.
- La posición 5 tiene la mayor energía siendo temas rápidos, fuertes y ruidosos.
- Para entrar al TOP 10 un tema debería tener una energía moderada ya que está muy cercana a 0.5

Tempo:
- En concordancia con la energía los temas en la posición 1 son los de menor tempo, es decir aquellos con menor velocidad o ritmo.
- Nuevamente una coincidencia con energía para las posiciones 5, 7 y 8 donde se tenemos temas con más velocidad y energía.

Valencia (Positividad):
- Es una medida de la positividad que transmite el tema, felicidad, alegría, euforia.
- Otra vez tenemos una coincidencia en valores con las dos características anteriores.
- Temas en TOP 1 y TOP 10 tienen un cierto equilibrio en cuanto a positividad a diferencia de las demás posiciones que en promedio tiene mayor positividad

Loudness (Volumen):
- Otra variable muy relacionada con las anteriores. Tenemos niveles de volumen o fuerza bajos en el TOP 1, teniendo que ver con una baja energía y velocidad. Lo mismo sucede con el TOP 10.
- Para valores intermedio de posiciones el volumen/fuerza aumentan acompañando la energía y velocidad que ya registramos en esas posiciones.

Conclusiones:
- Como conclusión primero podemos apreciar que las variables a simple vista tienen un cierto nivel de correlación.

- Además también se puede apreciar que tenemos aproximadamente 3 niveles que concuerda en algún punto en todas la variables. Posición 1, posiciones intermedias y posición 10.
- En cada uno de estos niveles los valores en cada variable acompañan, es decir: si hay mucha energía también hay mucha positividad, velocidad y volumen. Lo mismo sucede para valores bajos.
- Los temas más votados son aquellos que tienen energía y positividad medios, una velocidad media/alta (110 es el tempo de una velocidad considerable [1]) y un volumen o fuerza baja.
- Los temas en la posición 10 siguen el mismo patrón que la posición 1 pero con niveles ligeramente más altos.
- Las posiciones intermedias siempre tienen la mayor energía, positividad, velocidad y volumen.

5.2.2 Criterio N°2

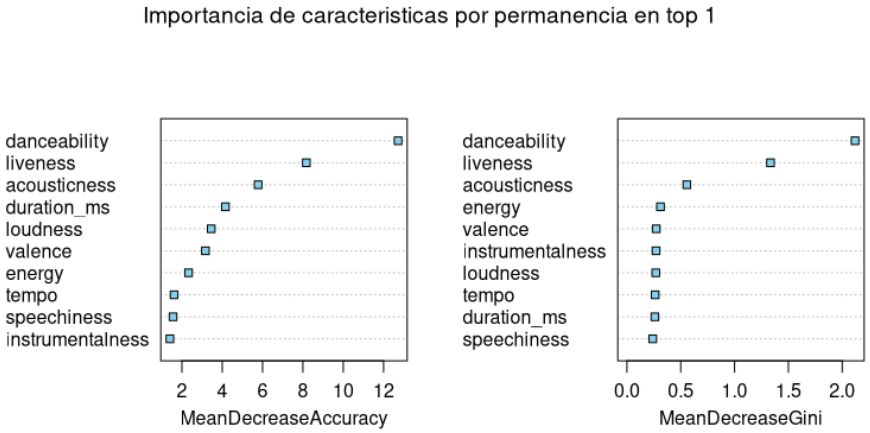

Gráfico N°5

Para este criterio vamos a usar la misma métrica que el criterio anterior: **Mean decrease Gini**. En el Gráfico N°6 observamos el grado de cada característica en función de la permanencia en semanas en el top 1

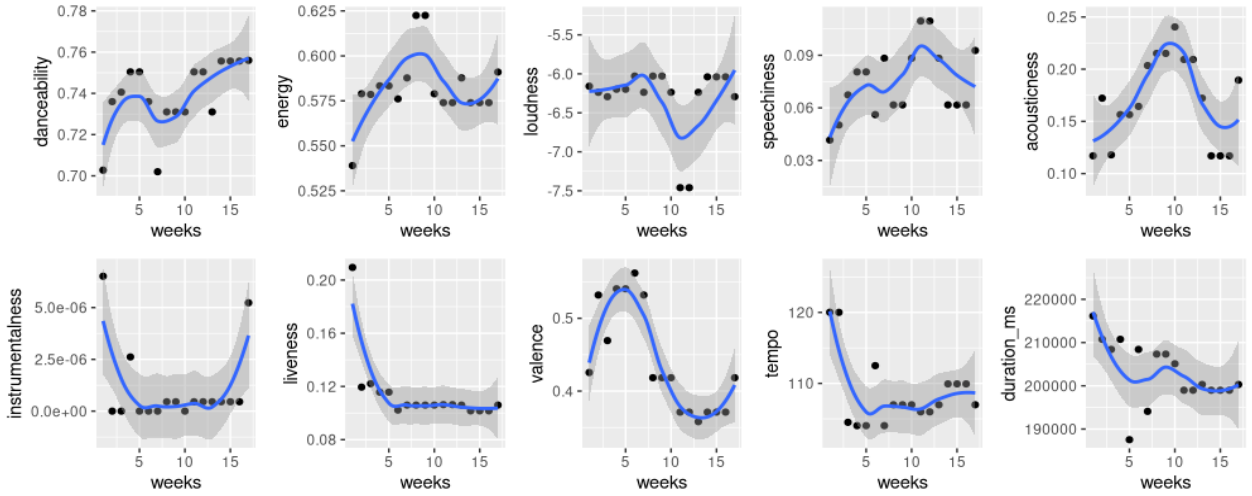

Gráfico N°6

[1] https://youtu.be/dwVP07gsqOM

Del Gráfico N°6 analizamos las cuatro características más importantes:

Danceability:
- Mide cuán bailable es un tema.
- Esta variable tiene un rango entre 0 y 1, por lo cual vemos que el top 1 tiene un alto componente de temas bailables (> 0.7).
- Los temas que más tiempo permanecen son los más bailables.

Liveness:
- Mide el grado de presencia de una audiencia en la grabación de un tema.
- Se aprecia que ningún tema en el top 1 tiene audiencia, tenemos valores muy bajos, por debajo de 0.2. Los valores > 0.8 describen temas temas en vivo con audiencia.
- En definitiva, sólo tenemos temas grabados en un estudio.

Acousticness:
- Define cuan acústica es una pista.
- En general el TOP 1 tiene un componente muy baja acústica lo que concuerda con que son todos temas muy bailables (> 0.7).
- En el rango medio de permanencia toma más fuerza lo acústico pero estamos en valores muy bajos.

Energía:
- Los temas de mayor permanencia son aquellos con niveles medio de energía, como sucedía con el criterio 1.
- Los niveles de energía no tienen una gran varianza todas las permanencias tiene niveles más o menos parecidos, entre 0.55 y 0.6.
- Se determina más energía en las permanencias medias pero la variación es baja.

Observaciones:

- En definitiva si queremos componer temas que permanezcan el mayor tiempo posible en el TOP 1, vamos a querer que tengan
    - Temas para bailar (es una tendencia muy marcada).
    - Niveles medios de energía
    - Poca o ninguna componente acústica.
    - Temas grabados en estudio (es una tendencia muy marcada).
    - Y una energía equilibrada. Es decir, ni heavy metal y baladas.

5.3. Pregunta N°3

Para aquellos temas que ingresaron al TOP 1 durante el periodo 2018-2020 y en aquellas semanas donde se producen cambios de temas ¿Como es el comportamiento de los atributos en estos cambios? ¿El tema que ingresa tiene valores de atributos similares al tema que desplaza?

El Gráfico N°7 contiene la evolución semanal de los 8 atributos para aquellos temas del TOP 1: Cada punto rojo corresponde a un único tema en una semana específica, por lo tanto puntos en una misma línea horizontal corresponden a un mismo tema que estuvo en TOP 1 en diferentes semanas (consecutivas o no). Los saltos de línea en forma vertical indican el cambio de tema entre diferentes semanas.

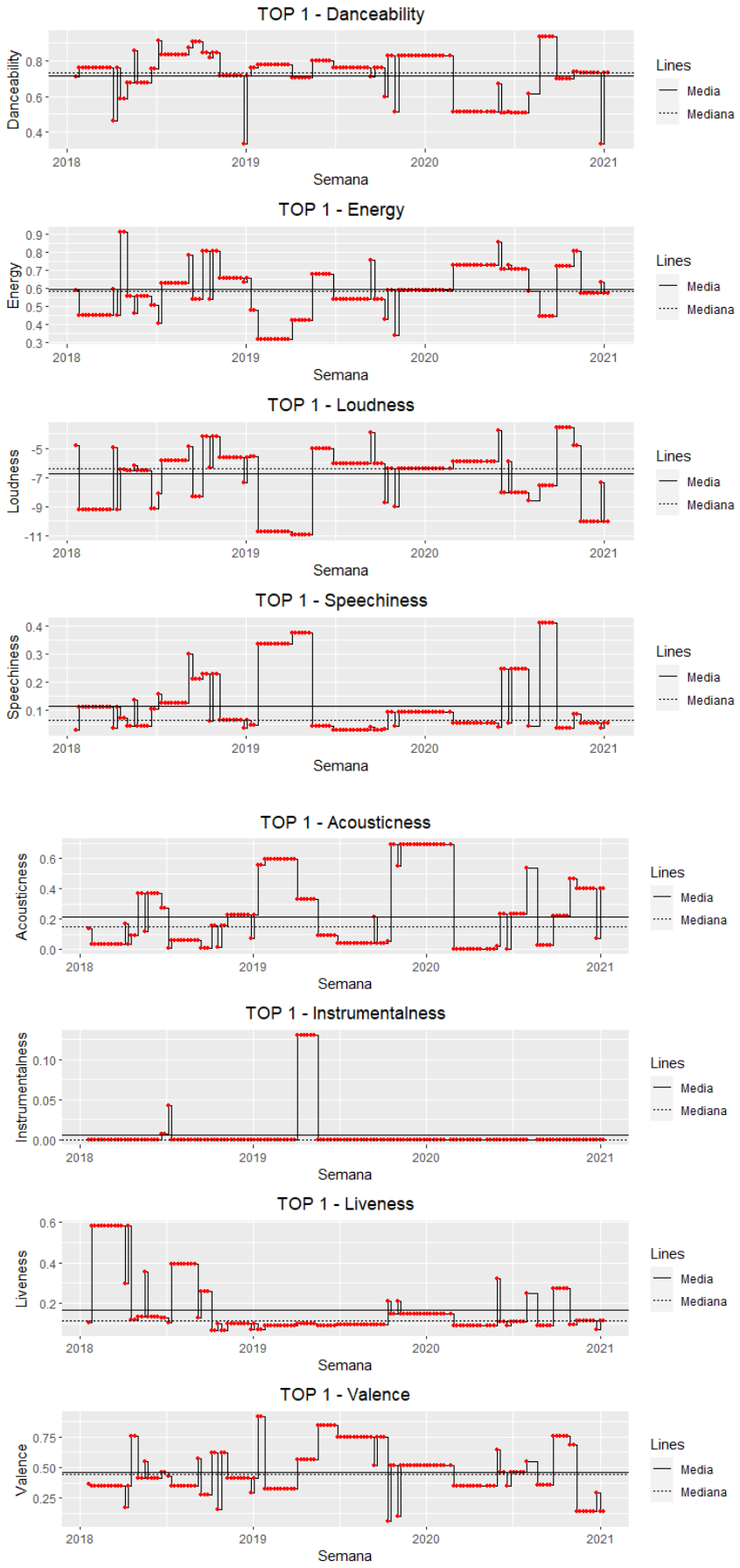
TOP 1 - Danceability
Danceability
0.8
0.6
0.4
2018
2019
2020
2021
Semana
Lines
Media
Mediana
TOP 1 - Energy
Energy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
2018
2019
2020
2021
Semana
Lines
Media
Mediana
TOP 1 - Loudness
Loudness
-5
-7
-9
-11
2018
2019
2020
2021
Semana
Lines
Media
Mediana
TOP 1 - Speechiness
Speechiness
0.4
0.3
0.2
0.1
2018
2019
2020
2021
Semana
Lines
Media
Mediana
TOP 1 - Acousticness
Acousticness
0.6
0.4
0.2
0.0
2018
2019
2020
2021
Semana
Lines
Media
Mediana
TOP 1 - Instrumentalness
Instrumentalness
0.10
0.05
0.00
2018
2019
2020
2021
Semana
Lines
Media
Mediana
TOP 1 - Liveness
Liveness
0.6
0.4
0.2
2018
2019
2020
2021
Semana
Lines
Media
Mediana
TOP 1 - Valence
Valence
0.75
0.50
0.25
2018
2019
2020
2021
Semana
Lines
Media
Mediana

Gráfico N°7

Del análisis del Gráfico N°7 se obtienen las siguientes conclusiones principales:

- Cuando en una semana ingresa un nuevo tema los valores de los atributos presentan un comportamiento diverso según el atributo que se esté analizando: En algunos casos se observa un comportamiento ascendente o descendente en valores similares al tema que reemplaza y en otros casos se observan grandes saltos ascendentes / descendentes.
- Analizando todos los atributos se identifican las siguientes particularidades:
    - Danceability: Los saltos pronunciados a valores inferiores que se observan coinciden con el mismo comportamiento identificado previamente relacionados a las semanas de navidad del 2018 y 2020 cuando el tema ***All I want for christmas is you*** llega al TOP 1.
    - Intrumentalness: La mayoría de los valores se encuentran cercanos a cero, salvo 2 temas donde suben levemente y 1 tema donde la suba es considerable.
- El resto de los atributos presentan comportamientos dispares entre diferentes temas observándose comportamientos ascendentes / descendentes moderados.

5.4. Pregunta N°4

Considerando el momento desde que un tema es lanzado y llega al TOP 10 ¿Cuántas semanas transcurren? ¿Esto sucede rápidamente o existen temas lanzados mucho tiempo atrás que igualmente logran ingresar? ¿Cuáles son los atributos de las canciones que llegaron más rápidamente al TOP 10?

Para responder esta pregunta se usaron la fecha de lanzamiento del álbum "album_release_date" y la fecha de la semana en la cual cada canción ingresa por primera vez al TOP 10 (mínima "week_start" para cada canción).
Un primer filtrado que se realizó es que la fecha de lanzamiento fuera posterior al 01/01/2018 porque sólo disponemos de los charts a partir de esa fecha. Como ejemplo de canciones descartadas con este filtrado, la canción más antigua que se encontró fue ***Jingle Bell Rock*** de ***Bobby Helms*** que ingresó al TOP 10 el 21/12/2018, cercano a la navidad y el álbum al que pertenece se lanzó el 2/12/1957, es decir posiblemente la canción haya ingresado en alguna navidad de años previos al 2018.
Sin embargo, se puede destacar que tras realizar este primer filtrado se conservan un 94% de las canciones incluidas en el TOP 10. Es decir, la mayor parte de las canciones escuchadas entre 2018-2021 fueron lanzadas en ese mismo lapso de tiempo.
Tras el filtrado se realizó un análisis exploratorio de la relación entre las fechas mencionadas (mínima "week-start" y "album_release_date") que se muestra en el Gráfico N° 8.

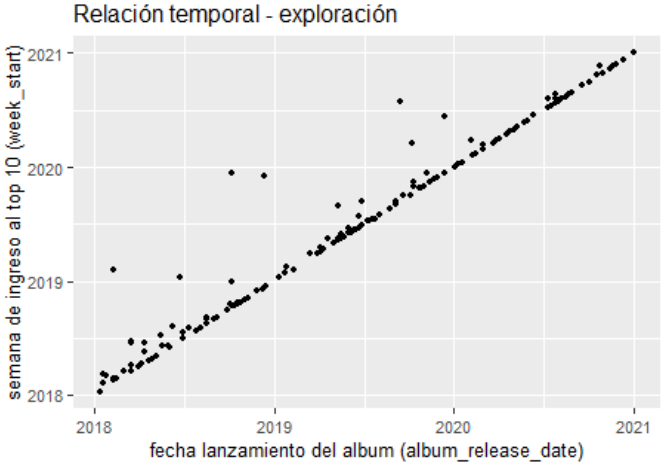

Gráfico N°8

La tendencia claramente lineal que se observa en el Gráfico N°8 (con una muy alta correlación) permite confirmar la hipótesis de que en gran medida las canciones ingresan al TOP 10 en fechas muy cercanas a las del lanzamiento del álbum.
En lo que sigue se cuantifican estas relaciones a partir del cálculo de la diferencia temporal entre ambas fechas, variable a la cual se denomina delta_t y que se expresa discretizada en unidad de "weeks". Por ej., si el valor de delta_t es igual a "0 weeks" indica que la canción ingresó por primera vez al TOP 10 en la misma semana de lanzamiento del álbum.
En el gráfico N°9 (derecha-superior) se presenta el boxplot del delta_t en el cual se visualiza la aparición de muchos outliers y un boxplot muy angosto que tiene los siguientes parámetros estadísticos: bigote inferior Q1 vale 0, la mediana Q3 es igual a 0 y el valor del bigote superior Q5 resulta igual a 2. Asimismo se calculó que la moda resulta igual a 0 coincidiendo con la mediana, mientras que la media resulta igual a 3 (esta diferencia respecto de la mediana es esperable por la gran presencia de outliers).
En la parte inferior del Gráfico N° 9 se aprecia el histograma identificando con color rojo a los valores que se encuentran por debajo del bigote superior calculado (resultado de aplicar la técnica de preprocesamiento de outliers mediante el rango intercuartil) y en azul los valores outliers que podrían ser removidos. Este histograma presenta un marcado sesgo a la derecha.

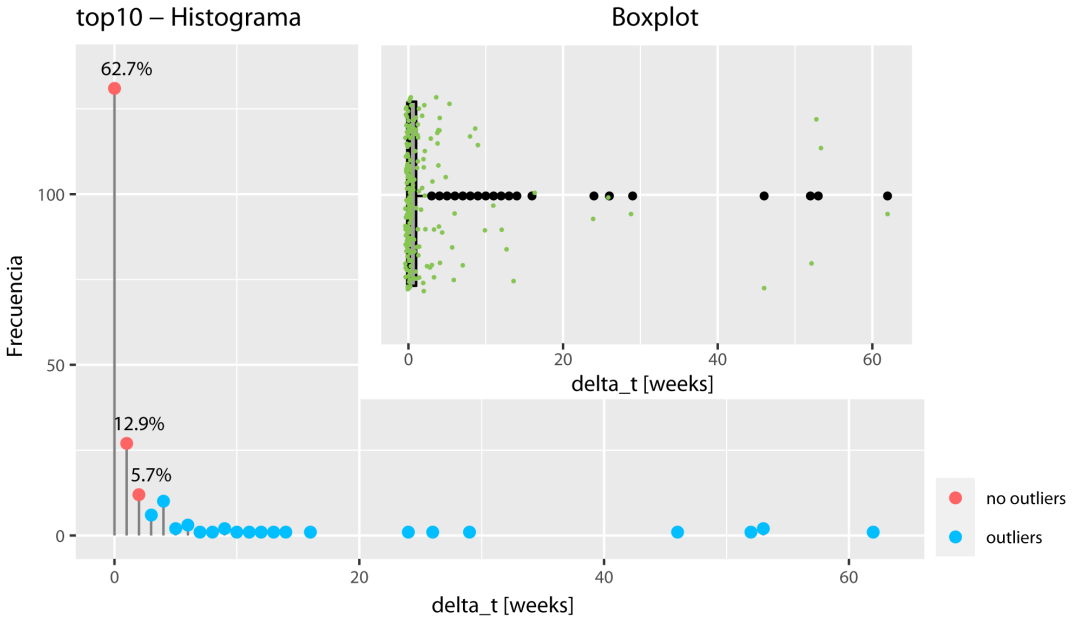

Gráfico N° 9

A partir de este análisis conjunto se puede concluir que en el TOP 10 2018 - 2021 se observa una marcada tendencia a escuchar canciones nuevas (en una proporción del 94%).
Además, las canciones en su mayoría han ingresado rápidamente respecto de su lanzamiento. Esto se puede concluir a partir de las observaciones cuantitativas derivadas del Gráfico N°9:

- El 62.7% de las canciones ingresan en la misma semana que salió el disco. Si se incluye también la semana posterior se llega a un 75.6%. Hasta cuatro semanas después de su fecha de lanzamiento rondan al 89%.
- Si bien hay canciones más antiguas que llegan al TOP, las cantidades son marginales. Por ej., sólo un 7% de canciones llegan luego de transcurridas al menos 10 semanas desde su lanzamiento.

Luego se encontró que el 59% de las canciones del TOP 10 son el único hit de su álbum mientras que el 15% llegaron al TOP de a pares. Del 26% restante se determinó la lista de álbumes que incluyen al menos tres hits. Esto se muestra en el Gráfico N° 10 donde además se

discrimina por la cantidad de semanas que tardaron en llegar al TOP 10. Se puede concluir que las canciones de los álbumes con más hits ingresan al TOP rápidamente.

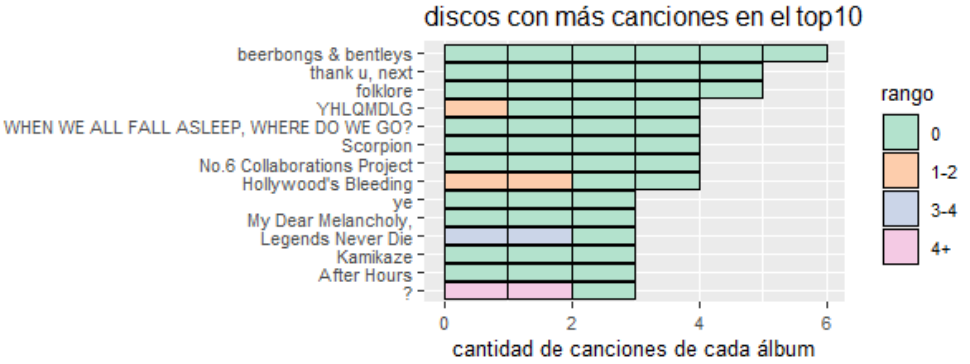

Gráfico N°10

A partir de la información obtenida nos interesa observar la relación de los atributos de las canciones que debutaron en el TOP 10 en menos de 5 semanas respecto del lanzamiento, discriminando por "cantidad de hits por disco". Esto se presenta en el Gráfico N°11. Para realizar este análisis se descartaron las features "instrumentalness" y "liveness" en concordancia con lo analizado exhaustivamente en la pregunta 2.

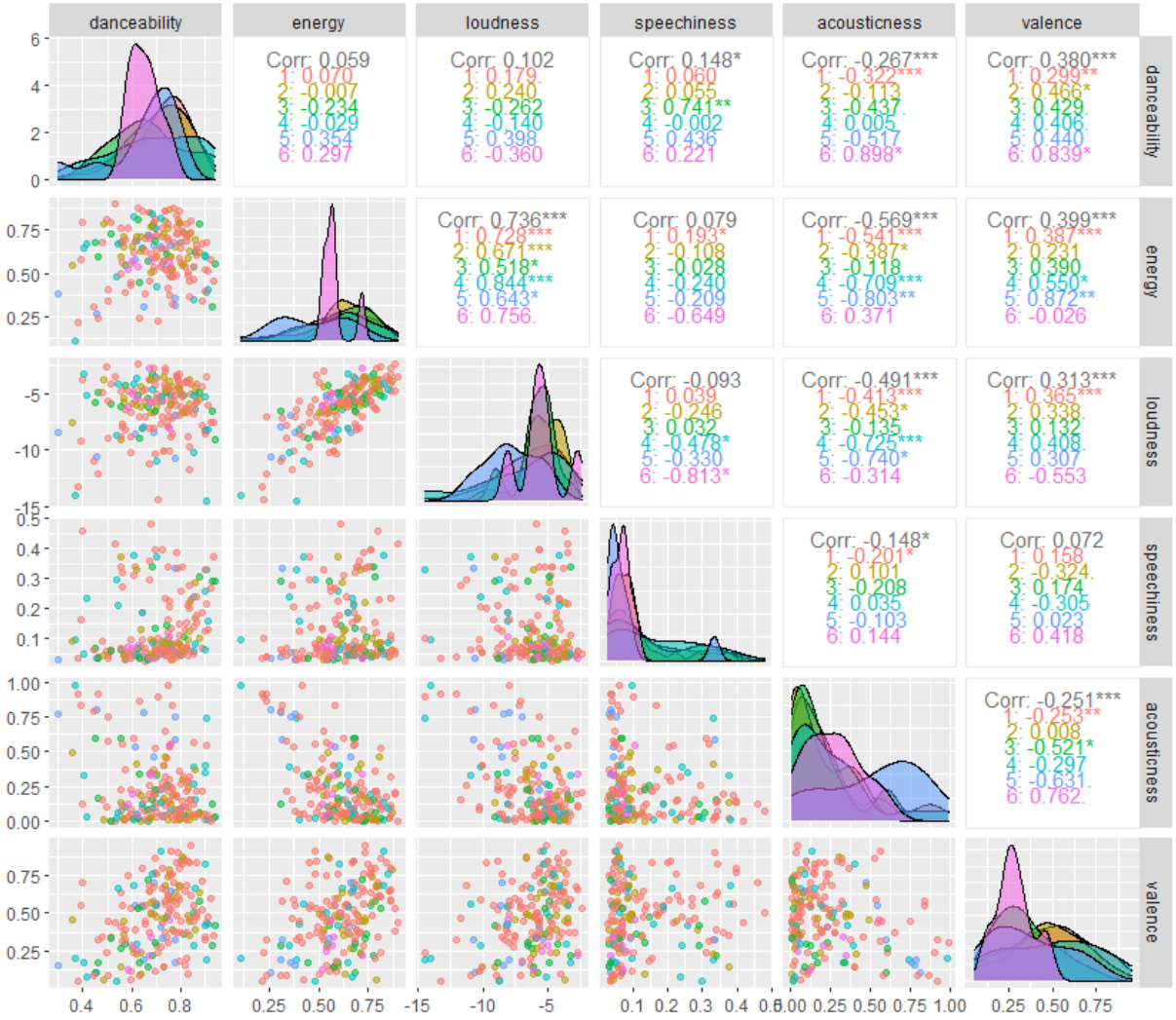

Gráfico N° 11: Dispersión y correlación segmentado por cantidad de hits del álbum.

A partir del Gráfico N° 11 de pares se pueden destacar las siguiente observaciones:

- Alta correlación positiva entre "energy" y "loudness" en todos los grupos analizados. El valor máximo se alcanza para canciones de álbumes que tienen 4 hits.
- Alta correlación negativa entre "energy" y "acousticness" principalmente entre canciones de álbumes de 4 y 5 hits (a excepción de canciones del álbum de 6 hits).
- La mayor dispersión de features se presenta en canciones de álbumes con 1 y 2 hits, manteniendo sus valores dentro de los márgenes ya analizados en pregunta 2.
- El único álbum que tiene 6 hits presenta una esperable variación más concentrada en todas sus features. Esta concentración se hace evidente al compararlas con la variación de features en los otros grupos que concuerda con lo analizado en pregunta 2. Se aprecia alta correlación entre “danceability” y “valence” para estos 6 hits.
- La energía de las canciones de los álbumes de 5 hits se concentra en torno a dos valores, el más alto coincidiendo con los álbumes de al menos 4 hits (que concentran valores mayoritariamente por encima de 0.75).

A partir de lo observado, usando el coeficiente de correlación de Pearson junto al análisis gráfico se podría eliminar redundancia en el dataset dada la alta correlación entre “energy” y “loudness” (asumiendo un umbral de 0.7). Esta eliminación podría extenderse a otras features si se analiza correlación segmentada de acuerdo a cantidad de hits por álbum.

5.5. Pregunta N°5

Para aquellos temas que ingresaron al TOP 1 durante el periodo 2018-2020 y caen de esta posición ¿Cuántas semanas pasan hasta que llegan al TOP 5 , TOP 10 y no vuelven de él?

Para contestar esta pregunta definimos la siguiente métrica:

Permanencia de descenso:

- Se refiere a la cantidad de semanas que un tema permanece en el TOP 10 luego de caer de la posición 1.
- En algún sentido nos da una medida de la velocidad de descenso de posiciones de un tema medida en semanas.

En el Gráfico N°12 observamos un histograma de esta nueva variable:

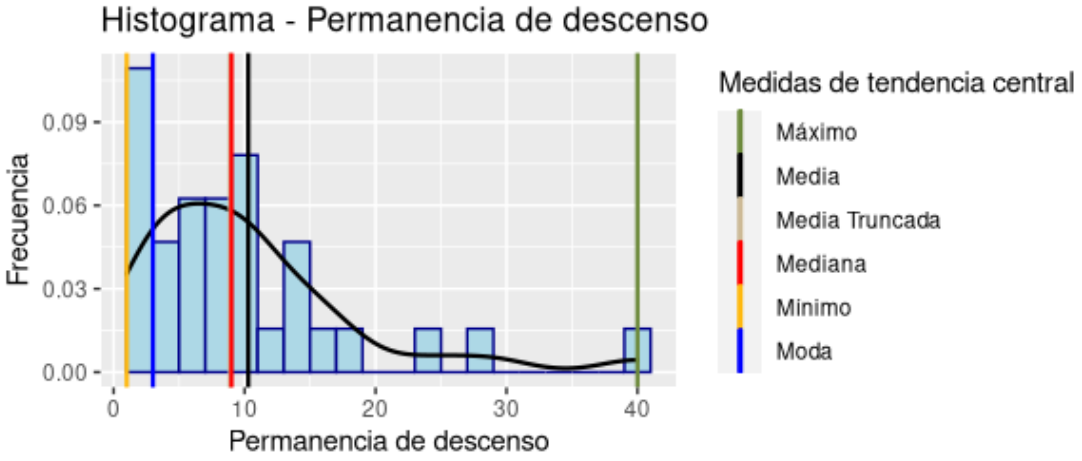

Gráfico N°12

Del análisis del Gráfico N°12 se obtienen las siguientes observaciones:

- La tendencia es una permanencia de 10 semanas en el TOP 10.
- Hay muy pocos temas que permanecen hasta un máximo de 40 semanas, lo cual es esperado.
- La distribución tiene un cierto sesgo hacia la izquierda, lo que indica que la mayoría de los temas permanecen pocas semanas (menos de 2).

Visto esto, en el Gráfico N°13 observamos la permanencia de descenso del TOP 10 segmentada por rango de posiciones (hasta caer fuera del TOP 10):

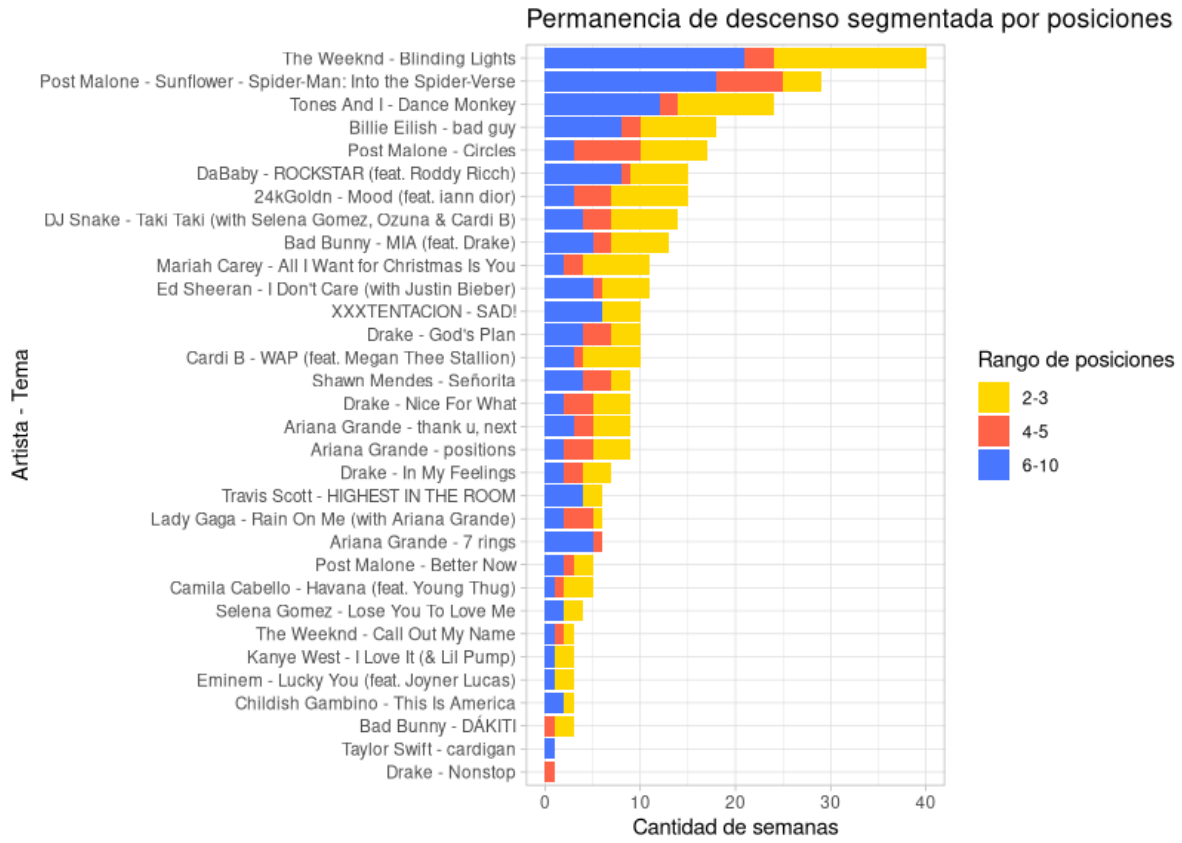

Gráfico N°13

Del análisis del Gráfico N°13 se obtienen las siguientes observaciones:

- Se aprecia que inicialmente ***The Weekend - Blinding Lights*** es el tema con más permanencia con 40 semanas en el TOP 10. Cabe destacar que tiene muy poca permanencia entre las posiciones 4 y 5, es decir que hay diferencia con otros temas donde si el descenso es más uniforme como el caso de ***Drake - God's Plan*** o ***Ariana Grande - Positions***. De esta manera los temas con menor permanencia parecen ser más uniformes en el descenso de posiciones en comparación.
- La artista ***Ariana Grande*** tiene temas en general con una permanencia cercana.
- Otros artistas como ***Drake*** y ***Post Malone*** tienen mayor variabilidad de permanencia teniendo temas con mucha permanencia y otros con muy poca como el caso de ***Nonstop*** de ***Drake***.
- Luego hay artistas que tiene mucho permanente pero un solo tema en el TOP 10 para todo el periodo como ***Billie Eilish, Tons And I, Da Baby, 24kGoldn, DJ Snake, Bad Bunny, Mariah Carey***, etc.. Todos ellos sólo tienen un tema en el TOP 10 en todo el periodo (2018-2020) lo que indica el nivel de dificultad para mantener la permanencia de descenso.