TRABAJO PRÁCTICO ENTREGABLE 1
Procesamiento de datos y bases de datos noSQL

Grupo 4 (martes): López Malizia, Álvaro; Padula, Eliana; Rossi, Fabiana A.

RESUMEN

En el presente trabajo nos propusimos integrar los conocimientos relacionados con el preprocesamiento de datos, integración, reducción de dimensionalidad, construcción de variables y gestión de datos, vistos en clase, mediante una Base de Datos NO-SQL orientada a documentos (MongoDB) y el IDE R-Studio del lenguaje de programación R.

Se realizó un análisis exploratorio de información relativa a *rankings* semanales de Spotify de los 200 temas más escuchados a nivel mundial. Se complementó esta información con el estudio de las características de dichos temas, para entender las relaciones existentes entre algunas variables del *dataset*. Además, se integraron datos provenientes de otras fuentes de información que proveyeron el precio de las acciones de Spotify en el período de tiempo analizado. Por último, se utilizaron herramientas gráficas para generar visualizaciones del contenido de nuestro dataset y las relaciones intrínsecas de los datos para facilitar su interpretabilidad.

A partir de nuestro análisis concluimos que la posición en el *chart* de distintos temas se correlaciona positivamente con la edad del tema al momento de ingresar al *chart*, pero no con otras características de audio de los temas. La cantidad de mercados de distribución de un tema influye en su popularidad, pero existen temas con comportamientos anómalos. Por otro lado, un estudio multivariado indicó que los temas presentan cierta estacionalidad en sus características de audio. El precio de las acciones de Spotify no está influenciado por el total de *streams*, pero sí por la declaración de pandemia COVID.

INTRODUCCIÓN

El volumen de datos que se generan continuamente ha ido creciendo de modo exponencial durante los últimos años, por lo que existe una necesidad urgente de utilizar herramientas computacionales que asistan en la extracción de nueva información, en el proceso de descubrimiento de la información en datasets (KDD). El objeto de KDD es la revelación e interpretación de nuevos patrones en la información, que son imperceptibles en un primer momento debido a su gran volumen.

En este trabajo práctico procesamos y limpiamos datos provenientes de los *rankings* semanales de Spotify de los 200 temas más escuchados a nivel mundial entre 2018 y 2021 y sus características asociadas. Integramos y redujimos la dimensionalidad de nuestro dataset, creamos nuevas variables, y aplicamos diversas herramientas de visualización para facilitar la interpretación de las relaciones entre las variables del mismo.

METODOLOGÍA

Los datos fueron generados con un script R utilizado para descargar la información relativa a los *charts* (*rankings*) semanales de los 200 temas más escuchados a nivel mundial, desde Spotify Charts. Se tomaron datos en el intervalo de tiempo comprendido entre el 12 de Enero de 2018 hasta el 15 de Enero de 2021.

Para cada posición en los *charts* se obtuvieron datos relativos a los temas (*tracks*) interpretadas por los distintos artistas, desde la API de Spotify. Una lista completa de las variables obtenidas y sus características puede encontrarse en el siguiente link.

Las bases de datos generadas fueron importadas al procesador de bases de datos orientado a documentos, MongoDB, y se utilizó la interfaz gráfica de usuario Robo3T para procesar los datos. Luego de verificar la existencia de duplicados en las colecciones importadas, se eliminaron dichos registros. Se integraron los datos en una única base de datos, y se promediaron los datos de las variables numéricas correspondientes a las características de audio, que presentaban el mismo autor y el mismo nombre del tema. Se decidió descartar aquellas variables no pertinentes a nuestro análisis (como direcciones de url o imágenes). También se descartaron aquellas variables con poca variabilidad (no informativas), y algunas fuertemente correlacionadas. Particularmente, para el análisis de la edad de cada canción, se

eliminaron aquellos registros cuya granularidad en la fecha de emisión del álbum era distinta a "día" (representaban un 4,77% de los registros). Luego de revisar la integridad de los datos en la nueva base de datos generada, se eliminaron aquellos registros que presentaban un porcentaje de datos faltantes mayor al 50% (consideramos que el número de registros con estas características era minoritario y el dataset completo era una muestra representativa). Se transformó el tipo de dato en cada atributo a aquél más apropiado en cada caso. Se realizaron *queries* pertinentes a cada pregunta de nuestro trabajo en MongoDB-Robo3T, y se importaron las tablas resultantes como *dataframes* a R, utilizando la librería mongolite. Por otro lado, se obtuvo información del precio de las acciones de Spotify Technology desde Yahoo Finance. Los valores se colectaron semanalmente en el período comprendido entre el 2 de Abril de 2018 y el 1 de Febrero de 2021, y se evaluó el precio de cierre por acción de la compañía.

El *dataframe* ***charts*** presentaba los siguientes atributos:

- **Position** (posición en el *chart*: 1 a 200),
- **Track_Name** (nombre del tema),
- **Artist** (nombre del artista),
- **Streams** (número de reproducciones),
- **week_start/week_end** (comienzo y fin, respectivamente, de la semana para un dado *chart*)

El *datafarme* ***audiofeatures*** presentaba los siguientes atributos:

- **artist_name** (nombre del artista intérprete),
- **track_name** (nombre del tema),
- **danceability** (descripción de cuán apropiada es un tema para bailar: valores entre 0-1),
- **energy** (representa una percepción de intensidad y actividad: valores entre 0 y 1),
- **loudness** (volumen medido en decibeles dB),
- **instrumentalness** (predice cuántas vocales hay en un tema: valores entre 0 y 1),
- **liveness** (detecta la presencia de una audiencia),
- **valence** (medida de positividad de un tema: toma valores entre 0 y 1),
- **tempo** (tempo promedio del tema medido en beats por minuto BPM),
- **time_signature** (medida que representa número de beats por bar de medición),
- **duration_ms** (duración en milisegundos),
- **speechiness** (explica el número de palabras en un tema),
- **acousticness** (acústica),
- **album_release_date** (fecha de emisión del álbum)[1] ,
- **album_release_date_precision** (granularidad en fecha de emisión de álbum: día/mes/año)
- **album_type** (tipo de álbum: single, album o compilado),
- **market** (código de país ISO 3166-1 alpha-2)[2],
- **track_number** (número del tema en el álbum).

Además, se crearon los siguientes atributos***:***

- **Edad del tema** (Diferencia entre la fecha de emisión del tema y la aparición en el *chart*, para cada una de las apariciones),
- **Permanencia ininterrumpida** (Se calculó para cada tema el número de semanas que permanecieron en el *chart* de manera ininterrumpida, para cada una de las instancias que ingresaron al *chart*),
- **Popularidad** (Cantidad total de apariciones de un tema en el *chart* a lo largo del tiempo),
- **Índice de éxito del artista** (Se calculó utilizando datos escalados entre 1 y 100)

$$índice\ de\ éxito\ =\ \frac{MPI \times 0{,}5\ PM \times ST}{PMA}$$

donde los términos son:

MPI= máxima permanencia ininterrumpida con cualquiera de los temas de cada artista (semanas)
PM= posición media de todas los temas de cada artista
ST= *streams* totales por artista
PMA= posición mínima alcanzada

[1] *En aquellos casos donde había más de un dato para el mismo Artista - tema - tipo de álbum, se tomó la menor fecha.*
[2] *Cuando existía un número diferente de países para el mismo Artista - tema se utilizó la máxima cantidad de mercados disponibles.*

La detección de *outliers* en distribuciones univariadas se realizó considerando valores superiores a 1,5 veces la distancia intercuartil; se consideraron *outliers* extremos aquellos con valores superiores a 3 veces la distancia intercuartil. En el caso de distribuciones multivariadas, la determinación de *outliers* se realizó utilizando la distancia de Mahalanobis o la técnica de *Isolation Forest*, indicado en cada caso.

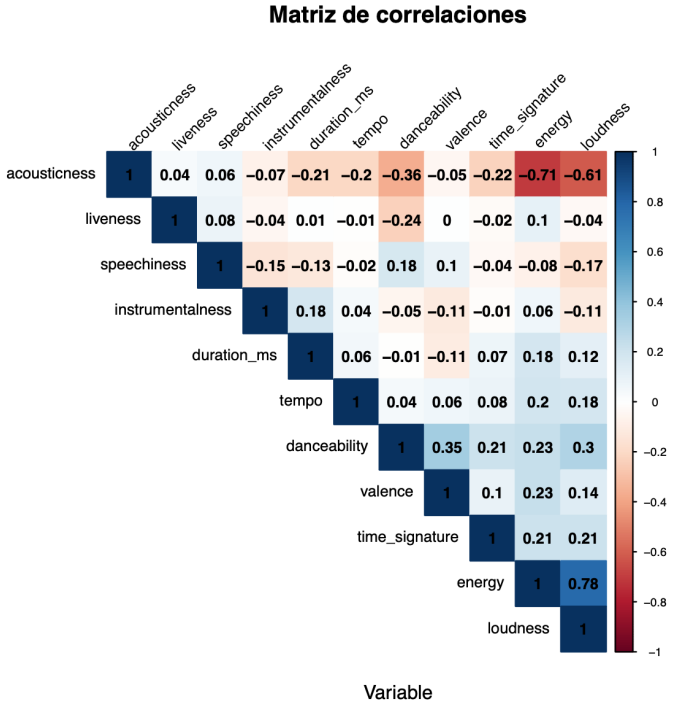
Para el agrupamiento de datos por semestres, se consideró que el inicio de los semestres consecutivos era el 01 de Enero de 2018.

Se consideraron las fechas 30 de Enero de 2020 y 11 de Marzo de 2020 como las fechas en las cuales se declaró a COVID-19 como una emergencia sanitaria y una pandemia, respectivamente, según la Organización Mundial de la Salud[3].

RESULTADOS

Descripción de las variables utilizadas en el trabajo práctico

Se construyó una matriz de correlaciones de los atributos numéricos analizados para cada tema, y se calculó su varianza a partir de los datos normalizados (**Figura 1**).



Atributos	Varianza
duration_ms	0.0006
loudness	0.0044
time_signature	0.0068
tempo	0.0135
danceability	0.0286
speechiness	0.0287
liveness	0.0416
instrumentalness	0.0450
energy	0.0532
valence	0.0561
acousticness	0.0980

Figura 1- *Izquierda: Matriz de correlaciones entre las variables numéricas correspondientes a características de audio de los distintos temas en el chart. El valor dentro de cada casillero indica el coeficiente de correlación, y el color del mismo indica el sentido de la correlación (positiva en color azul y negativa en color rojo). Derecha: Análisis de "Low Variance Filter": varianza de las características de audio de los temas del dataset, ordenadas de manera creciente.*

Los resultados de la **Figura 1** indican que existen atributos fuertemente correlacionados, como *energy* y *loudness*, y otros negativamente correlacionados como *acousticness* y *energy*. De la primera asociación, podemos deducir que el aporte de información derivada del análisis de la variable *loudness* será prescindible. Esta deducción se apoya además en el análisis de la varianza de cada atributo: *loudness* presenta muy poca variabilidad en la dispersión de sus valores, por lo que no aportará información substancial. Los atributos *duration_ms* y *time_signature* también se encuentran entre los tres menos variables, por lo que también se eliminarán del análisis posterior.

Hipótesis 1 : Existen atributos que se relacionan con la posición o permanencia ininterrumpida de cada tema en el *chart*

Con el objeto de analizar si existen características o atributos del audio de los temas que pudieran relacionarse con la posición de las mismas en el *chart*, se realizó un análisis del valor

[3] *"Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)". World Health Organization (WHO). 30 January 2020. Archived from the original on 31 January 2020. Retrieved 30 January 2020, y "WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020". World Health Organization (WHO). 11 March 2020. Archived from the original on 11 March 2020. Retrieved 12 March 2020.*

promedio de los valores normalizados de los distintos atributos en los temas que se encontraban en los primeros y últimos 10 temas del *chart*. Los resultados se representaron con un *radarchart* que se muestra en la **Figura 2**.

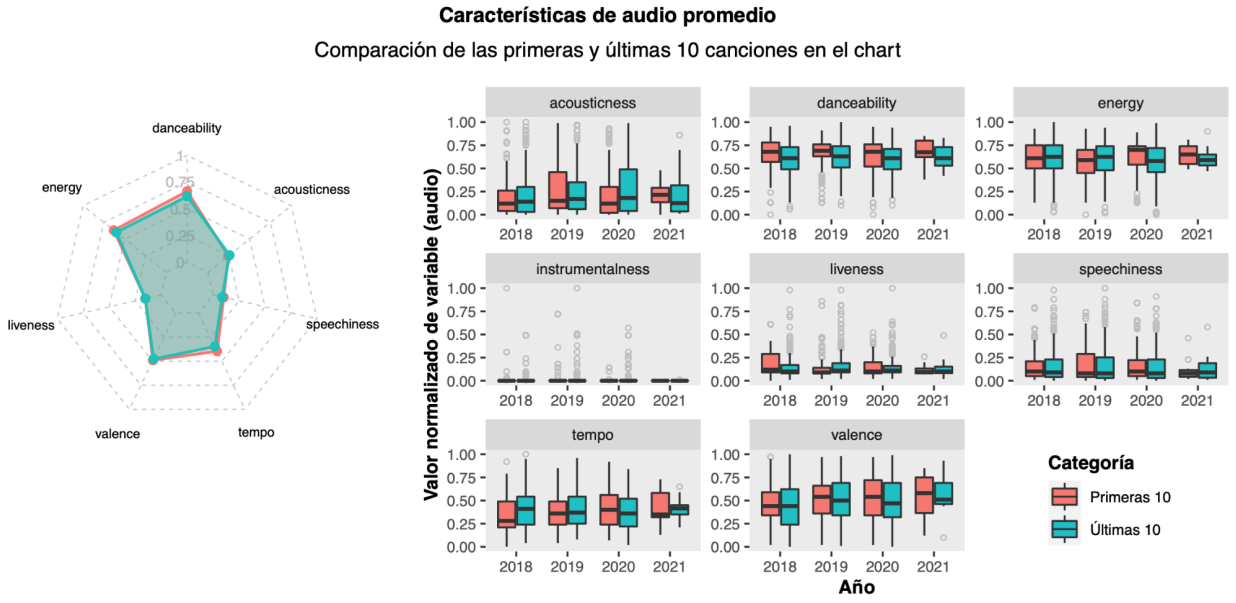


Figura 2- *Características de audio promedio de los primeros y últimos 10 temas del chart. Izquierda: Radarchart con las características de audio promedio. Derecha: Cada atributo está representado en un gráfico, donde los boxplot representan los valores de los atributos para los temas en cada categoría, agrupando los datos por año.*

El polígono formado por la unión de los valores medios de los atributos de los temas de cada categoría presentó un grado solapamiento elevado, por lo que se decidió realizar un análisis de la distribución de cada variable (normalizada) agrupando los datos por año, para poder además evaluar su dispersión y la presencia de *outliers* (**Figura 2**). Los resultados indicaron que la distribución de los valores de cada variable en cada categoría es semejante, y que existen variables como *acousticness, danceability, liveness* y *speechiness* que presentan *outliers*. Por otro lado, la variable *instrumentalness*, si bien mostró que tenía una varianza no menor en el análisis de *Low Variance Filter* (tabla embebida en **Figura 1**), es una variable que tiene la mayor parte de los valores agrupados en pocos valores (ver), y por ende será eliminada de los próximos análisis.

A continuación se creó el atributo *edad* para estudiar la relación entre la edad del tema en cada semana del *chart* y la posición asociada de dicho tema (**Figura 3**).

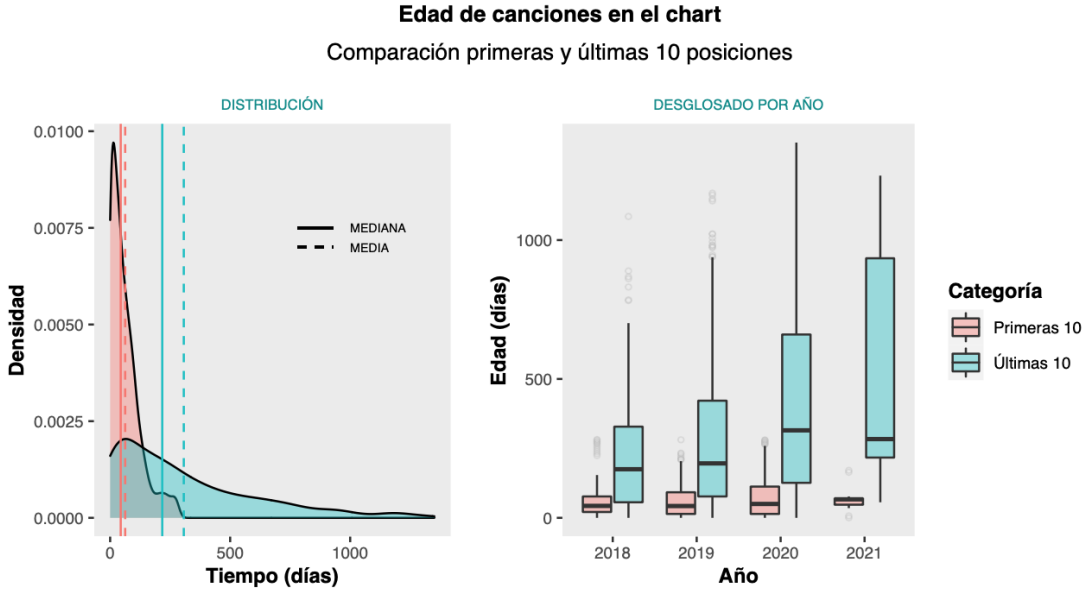


Figura 3- *Edad de temas en el chart, comparación entre las primeras y últimas 10 posiciones del chart. Izquierda: El gráfico de densidad representa la distribución de edades de los temas en las categorías analizadas. Se indica la mediana con una línea contínua, y la media con una línea punteada. Derecha: Edad de los temas, desglosado por año.*

Los resultados indicaron que los temas que se encuentran en las primeras 10 posiciones tienen una distribución de edad centrada en valores de menor magnitud respecto de los temas en las últimas 10 posiciones. Esta tendencia se respeta al agrupar los datos en cada uno de los años analizados.

A continuación, y con el objeto de estudiar la relación entre la duración de la permanencia ininterrumpida en el *chart* de cada tema y distintas variables de audio de los mismos, se creó un nuevo atributo llamado *máxima permanencia*. El mismo se construyó calculando el número máximo de semanas consecutivas que cada tema se encontraba en el *chart*. La distribución de dicha variable se presenta en la **Figura 4**.

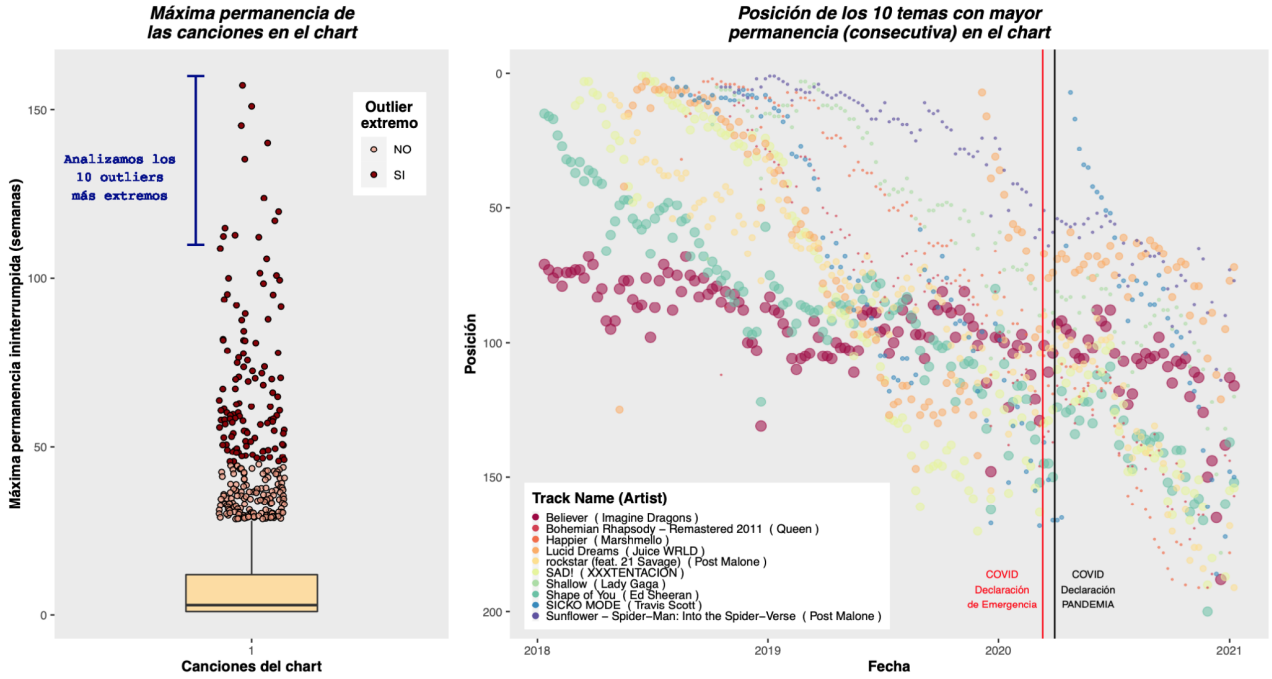


Figura 4- *Determinación de la máxima permanencia ininterrumpida de los temas en el chart (medida en semanas). Izquierda: Boxplot que indica la distribución de la máxima permanencia de los temas en el chart. Se indican outliers y outliers extremos con distintos colores. Derecha: Scatter plot que muestra la posición de los 10 temas con mayor permanencia consecutiva en el chart a lo largo del tiempo. El color de los puntos indica temas distintos, el tamaño del punto tiene una correlación directa con la permanencia en el chart. Se indican con líneas verticales la declaración de estado de emergencia y de pandemia debido a COVID-19, según la organización mundial de la salud.*

Como indicado anteriormente, se consideró *outlier* a todo aquél valor que era mayor que 1,5 veces la distancia intercuartil de la distribución de valores, y *outlier extremo* a aquellos valores que se encontraban a una distancia mayor a 3 veces el rango intercuartil. Se analizaron los 10 *outliers* más extremos en el tiempo, y se observó su posición a lo largo del tiempo. Los resultados indican que la mayoría de los temas presentan el mismo comportamiento: su historia en el *chart* comienza en posiciones más bajas (mejor *ranking*) y luego desciende lentamente. La pendiente está correlacionada con su tiempo de permanencia en el *chart*. Existen temas con comportamiento diferencial, como el tema “Lucid Dreams” (Juice WRLD) o “Sicko mode” (Travis Scott).

Por último, quisimos determinar si había alguna característica del audio distintiva de los temas que se encontraban en esas 10 posiciones de mayor permanencia, en comparación con el resto de los temas del *chart*. Para ello se construyeron los gráficos de la **Figura 5**.

Si bien se observó que el estadístico de distribución central de cada variable en ambos grupos es diferente, la dispersión de cada distribución se superpone, por lo que podría pensarse que no existen diferencias en los atributos de cada categoría.

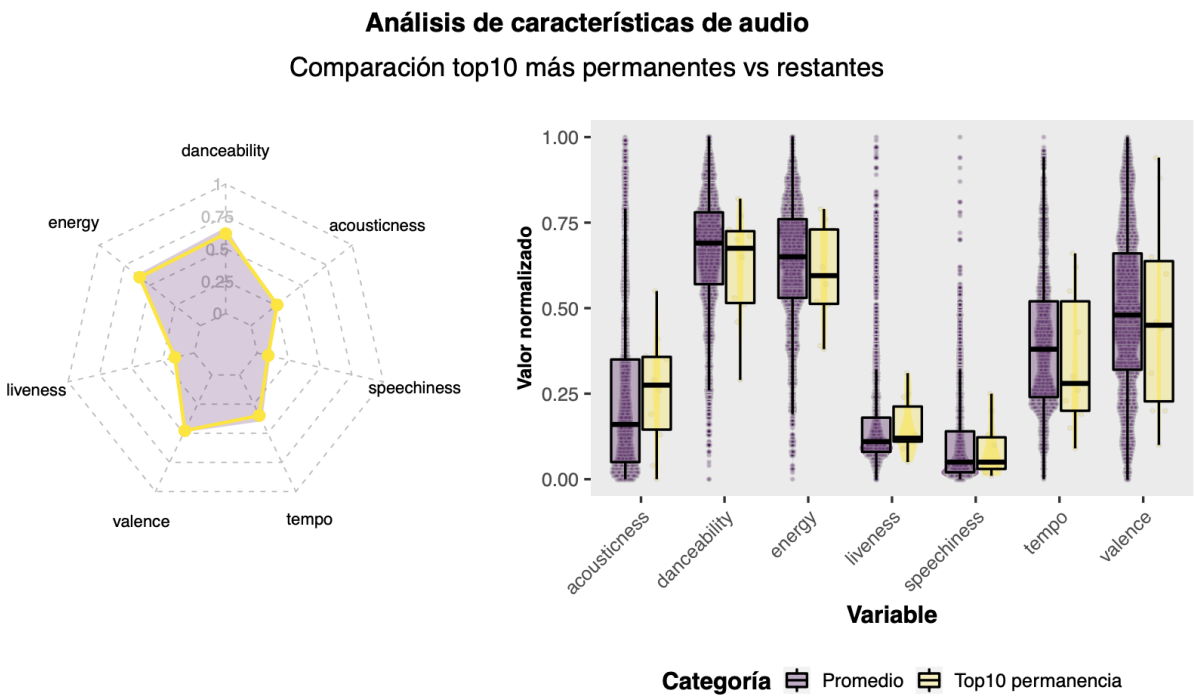


Figura 5- *Comparación de las características de audio de los 10 temas más permanentes en el chart versus los restantes. Izquierda: El gráfico de radarchart compara las características de audio promedio de las categorías analizadas. Derecha: Superposición de boxplot y violin-plot de los valores de los atributos analizados en las categorías: 10 temas más permanentes (Top10 permanencia) y los temas restantes (Promedio).*

Hipótesis 2: El éxito de un artista presenta un pico, luego de lo cual disminuye

A fin de caracterizar el éxito de los artistas en nuestra base de datos, creamos un *índice de éxito* que contempla la *máxima permanencia* en el *chart* de alguna de sus temas, la *posición media* y *mínima* de sus temas, y la cantidad de *streams* totales que suman sus temas en un rango de tiempo dado (**Figura 6**). Dicho índice fue analizado por semestres, siendo el 01-01-2018 el comienzo del primer semestre.

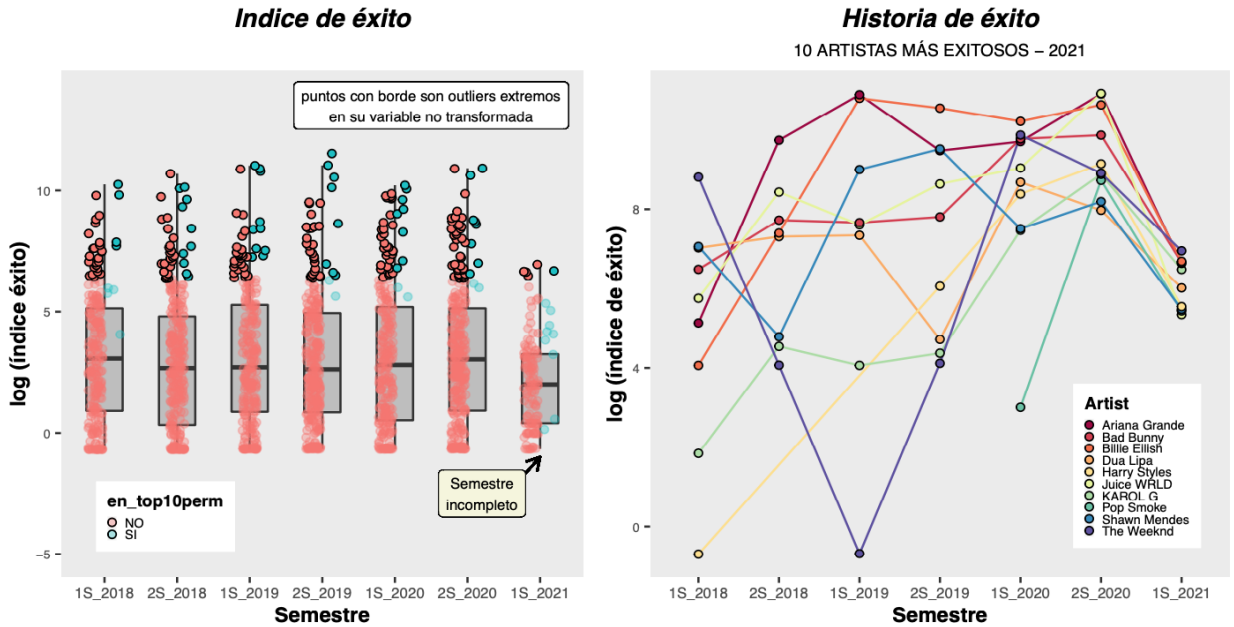


Figura 6- *Se analizó el índice de éxito para los artistas que entraron al chart en el período analizado. Izquierda: Los boxplot muestra el log(índice de éxito) de los artistas del chart, agrupando por semestre. El color de los puntos diferencia si estos artistas tienen (verde) o no (naranja) un tema que se encuentre en la lista de 10 temas más permanentes (en_top10perm). El borde de los puntos indica si esos valores son outliers extremos en la distribución de éxito [NO log(éxito)] para cada semestre. Derecha: Historia del éxito de los artistas más exitosos del año 2021.*

El análisis de la distribución del índice de éxito por semestre indicó que el mismo parece ser constante en el tiempo, salvo para el primer semestre 2021 (esto último se debe a que la

definición de índice de éxito contempla suma de *streams* totales, y los datos correspondientes a este semestre son incompletos y por ende debería desestimarse). Los *outliers* extremos de la distribución del índice de éxito muestra que muchos de los artistas muy exitosos tuvieron además temas en el *subset* de 10 temas con mayor permanencia en el *chart*. No obstante, hay muchos otros artistas que no tuvieron temas dentro de esta categoría y aún así son también considerados como *outliers extremos* de la distribución de índice de éxito, es decir, muy exitosos.

Al analizar la historia de éxito de los 10 artistas más exitosos en 2021, se puede observar que existen distintos patrones en las curvas que definen el comportamiento, siendo "Ariana Grande" y "Billie Ellish" artistas que tuvieron un éxito constante en un período mayor a dos años, y otros como "Pop Smoke" o "The Weeknd", que presentaron un pico de éxito más recientemente, a partir del año 2020.

Hipótesis 3: Existen atributos relacionados con el tema que influyen sobre su popularidad (cantidad de veces que dicho tema aparecerá en el *chart*).

Se definió la variable popularidad como la cantidad de veces que un tema aparece en el *chart*. A fin de estudiar la relación entre dicha variable y atributos de los temas, se analizó el número de países en los que se distribuía el tema (mercados disponibles) (**Figura 7**), el número de temas adicionales con los que cuenta el artista (**Figura 8**), y la posición del tema dentro de los distintos temas (**Figura 9**).

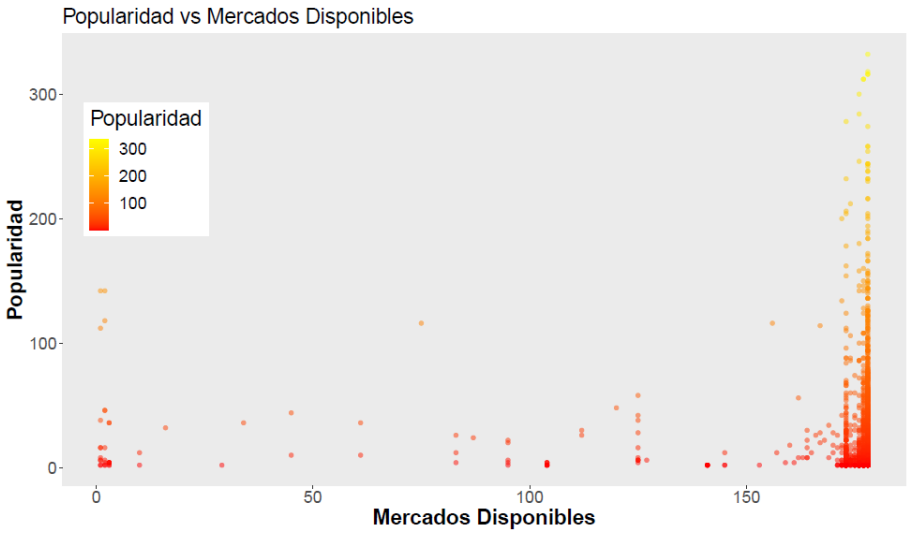


Figura 7- *Análisis de la relación entre los atributos* ***popularidad*** *del tema y la máxima cantidad de* ***mercados disponibles*** *en los que dicho tema se distribuyó.*

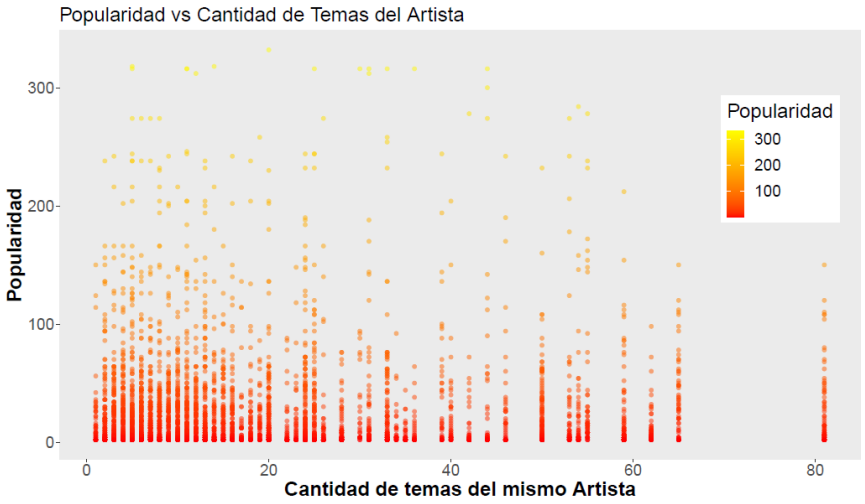


Figura 8- *Relación entre los atributos* ***popularidad*** *del tema, y* ***producción*** *de su artista intérprete (****cantidad de temas del mismo artista****).*

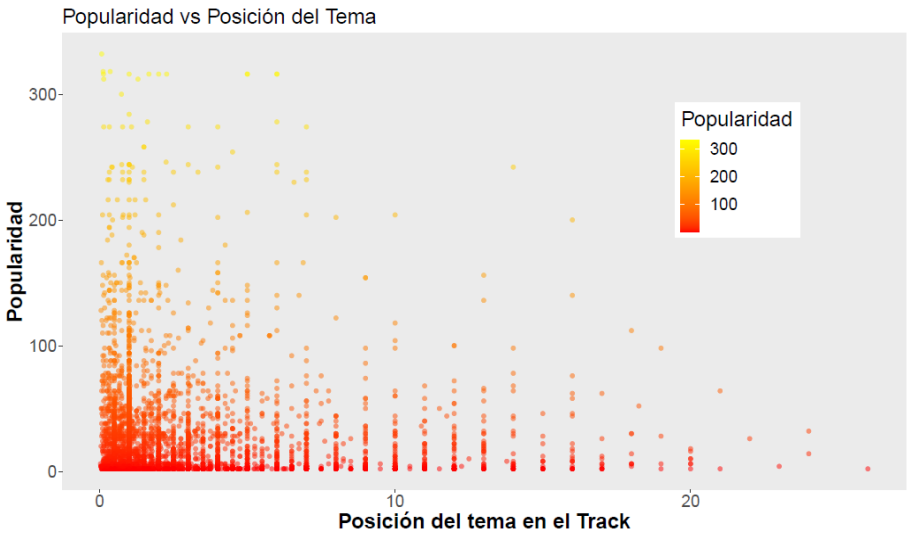


Figura 9- *Relación entre los atributos* ***popularidad*** *y* ***posición media*** *del tema.*

De acuerdo a las **Figuras 7, 8,** y **9**, la popularidad de un tema parece correlacionar positivamente con una mayor cantidad de mercados de distribución y una menor posición media en el *chart*. En cambio, la distribución en cuanto a la cantidad de temas adicionales del mismo artista y la popularidad de sus temas no es tan clara.

En base a estas observaciones decidimos profundizar nuestro análisis sobre el peso de los mercados disponibles en la popularidad del tema.

Países en los que se distribuye cada tema

Naturalmente, esperábamos que a mayor cantidad de mercados disponibles, mayor fuese la popularidad de un tema. Particularmente, nos interesaba conocer la presencia de países que tuvieran un peso preponderante en forma individual para poder posicionar un tema en el *chart*. Nuestros resultados indicaron que la tendencia general es la esperada. Los temas de mayor popularidad se concentran en la región de mayor cantidad de mercados disponibles (margen inferior derecho (**Figura 7**).

Detección de instancias más alejadas de la distribución utilizando la distancia de Mahalanobis

Realizamos un análisis multivariado de Mahalanobis para detectar la presencia de valores alejados de la distribución. Los 10 elementos de mayor distancia se indican en la **Tabla 1.**

N	TEMA	POPULARIDAD	Nº MÁXIMO DE PAÍSES DISPONIBLES	DISTANCIA Mahalanobis
1	Thinking out Loud	142	1	8,323545
2	Don't Stop Me Now - Remastered 2011	142	2	8,281196
3	Body (feat. brando)	112	1	8,094511
4	Another One Bites The Dust - Remastered 2011	118	2	8,092124
5	Mirrors	2	1	7,816448
6	Notificação Preferida - Ao Vivo	2	1	7,816448
7	Call You Mine (feat. Bebe Rexha)	6	1	7,810246
8	LOYAL (feat. Drake)	6	1	7,810246
9	Wonderful Christmastime - Edited Version / Remastered 2011	8	1	7,807611

10	Takeaway (feat. Lennon Stella)	38	1	7,805449

Tabla 1- *Análisis de Mahalanobis. Se muestran los 10 temas con mayor distancia de Mahalanobis.*

Dentro de esta lista, se destaca la existencia de dos grupos diferentes.

- **Grupo 1:** Temas que tienen **pocos países** en los cuales estuvieron disponibles, pero sin embargo presentan un **elevado número de apariciones en el *chart* (alta popularidad)**, marcados en celeste.
- **Grupo 2:** Temas que tuvieron un bajo índice de popularidad y una baja cantidad de mercados disponibles, marcados en amarillo.

Entre los elementos más alejados de la distribución, no aparecieron temas con una gran cantidad de mercados disponibles y una baja popularidad. Estos podrían considerarse como un tercer grupo de datos que se desvían de la distribución. No obstante, al ordenar los datos de forma ascendente para la popularidad y descendente para la disponibilidad de mercados, encontramos que la cantidad de datos que aparecen con estas características es numerosa y no se encuentran entre los datos más alejados según el análisis de distancias de Mahalanobis. Algunos ejemplos se muestran en la **Tabla 2**.

N	TEMA	POPULARIDAD	NÚMERO MÁXIMO DE PAÍSES DISPONIBLES	DISTANCIA Mahalanobis
1	...Ready For It?	2	178	0.5837743
2	1.5	2	178	0.5837743
3	1000 Doves	2	178	0.5837743
4	1000 Nights (feat. Meek Mill & A Boogie Wit da Hoodie)	2	178	0.5837743
5	100k Cash	2	178	0.5837743
6	2012	2	178	0.5837743

Tabla 2- *Detalle de registros con comportamiento inesperado, pero que no presentan valores extremos de distancia de Mahalanobis.*

Luego centramos nuestra atención sobre la identidad sobre los elementos del **primer grupo.** Los mismos se destacan en la **Figura 7**, ya que se agrupaban en el margen superior derecho, y se indican a continuación.

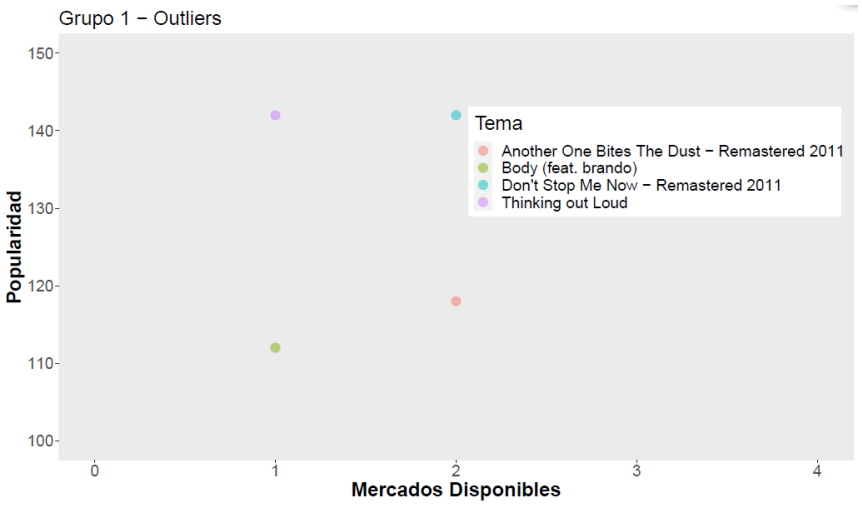


Figura 10- *Análisis de la relación entre los atributos **popularidad** del tema y la máxima cantidad de **mercados disponibles** en los que dicho tema se distribuyó, para los elementos del **grupo 1**.*

A partir de estos resultados nos surgieron las siguientes inquietudes: ¿cuál es la identidad de los países entre los cuales se distribuyeron estos temas?, ¿por qué la combinación de estos **países-temas** tuvo tanto peso en el *chart* global?, ¿cómo fue la permanencia en el *chart* de estos temas a lo largo del tiempo?

Los países en los cuales se distribuyeron los temas del grupo 1 se indican en la **Tabla 3.**

TEMA	PAÍSES DISPONIBLES
Another One Bites The Dust – Remastered 2011	Japón, Canadá y Estados Unidos
Body (feat. Brando)	Canadá
Don't Stop Me Now – Remastered 2011	Japón, Canadá y Estados Unidos
Thinking out Loud	España

Tabla 3- *Identidad de los países donde se distribuyen los temas del grupo 1.*

De acuerdo a la cantidad de población de Japón y Estados Unidos, es razonable que la combinación de ambos influya de forma preponderante en el *chart*. Sin embargo, es llamativo que temas sólo distribuidos en España o Canadá se encuentren entre los temas más populares, dado que ambos países tienen un menor peso demográfico.

Al analizar la permanencia y posición en el *chart* de esos temas a lo largo del tiempo, se observaron dos comportamientos diferentes. Por un lado, uno de los temas se mantiene constante en posiciones más bajas dentro del *chart* (*Thinking out Loud*), mientras que los otros tres restantes ingresan en posiciones altas y tienen un descenso sostenido a lo largo del tiempo (**Figura 11**).

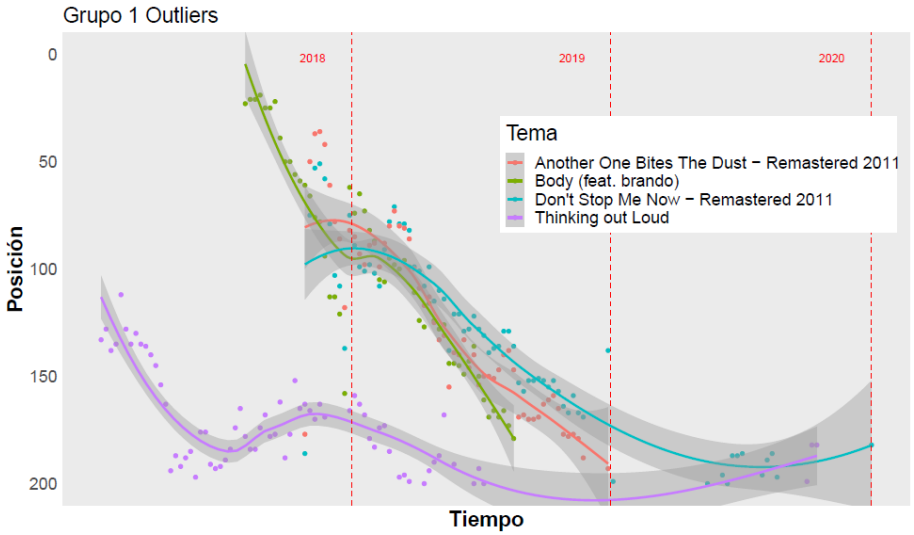


Figura 11- *Descripción de la posición de los temas del grupo 1 en el período analizado.*

Adicionalmente, se comparó el comportamiento de estos 4 temas, con aquellos de menor distancia de Mahalanobis. Comparativamente, estos últimos evidencian menores apariciones en el chart, distribuidas en un período de tiempo más acotado que los temas del Grupo 1 (**Figura 12**).

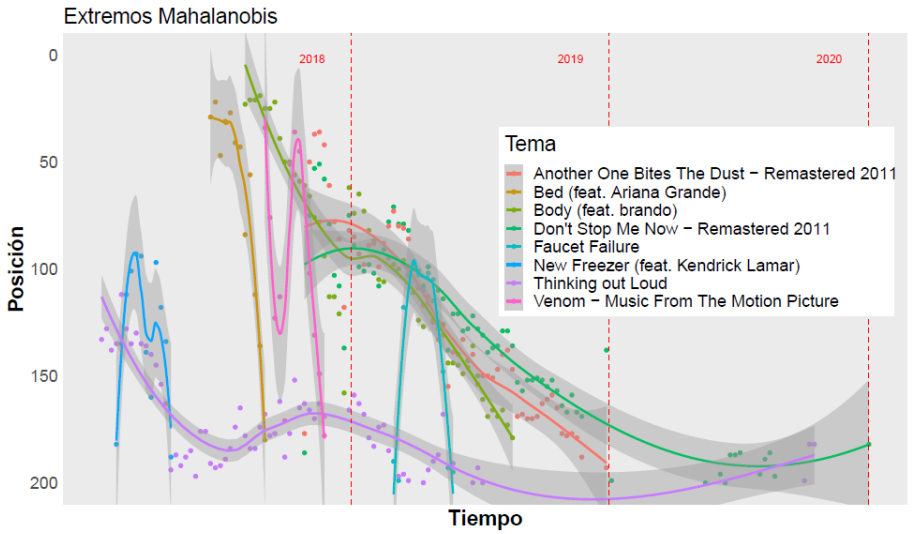


Figura 12- *Comparación de la posición de los 4 temas con mayor y menor distancia de Mahalanobis, a lo largo del tiempo.*

Hipótesis 4: Existe estacionalidad en los atributos de los temas que ingresan al chart

Como se pudo observar previamente, existe una diferencia en el comportamiento de los atributos a través del tiempo, donde puntualmente se observa un cambio pos-comienzo de la pandemia (COVID). Consecuentemente, nos preguntamos si existe estacionalidad dentro del año en los atributos de los temas observados.

Para realizar dicho análisis y obtener una cantidad suficiente de observaciones, se consideró que los atributos para cada semana del año se podría medir como:

$$Feature_t = Feature_{ti}\frac{Streams_{ti}}{Streams_t}$$

donde t indica la semana en evaluación (*week_start_date*) e i es el tema en esa semana por el mínimo nivel de granularidad considerado en la base de datos (artista-tema).

El ejercicio se realizó con los primeros top 10 de cada semana y se realizó un análisis gráfico de caras de Chernoff (**Figura 13**).

Figura 13- *Caracterización de atributos multivariados utilizando la técnica de caras de Chernoff en los primeros 10 temas del chart.*

Se observa que tanto considerando o no los datos colectados durante la pandemia, la música más escuchada (top 10) en el mes de Diciembre se diferencia notablemente de la música escuchada en los meses restantes.

Se realizó luego el mismo ejercicio, pero con las 200 posiciones del *chart* (**Figura 14**).

Figura 14- *Caracterización de atributos multivariados utilizando la técnica de caras de Chernoff en los 200 temas del chart.*

Nuestros resultados demostraron que al incorporar una mayor cantidad de datos (donde cada tema puede permanecer a más de un mes), las caras parecerían tener rasgos más similares entre sí. Aún en ese caso, la conclusión arribada es que incluso considerando los *charts* desde COVID, Diciembre se muestra como un mes atípico en los datos.

Outliers

A continuación nos preguntamos si era posible que los temas considerados en Diciembre estuvieran enmascarados como *outliers*.

Para responder dicha pregunta, se observaron los *outliers* considerando el tema y sus *features* o atributos. A diferencia de los *outliers* que se buscaban en otras secciones, en este caso el objetivo era preguntarse si había temas significativamente diferentes según sus *features*.

Para realizar dicho ejercicio, se realizó la técnica de *Isolation Forest. S*e consideraron 26 tracks como *outliers*; ya que se corresponde con aquellos casos cuyos valores de iforest score superan el percentil 99. (outliers considerados: iforest score > 0.56502)[4].

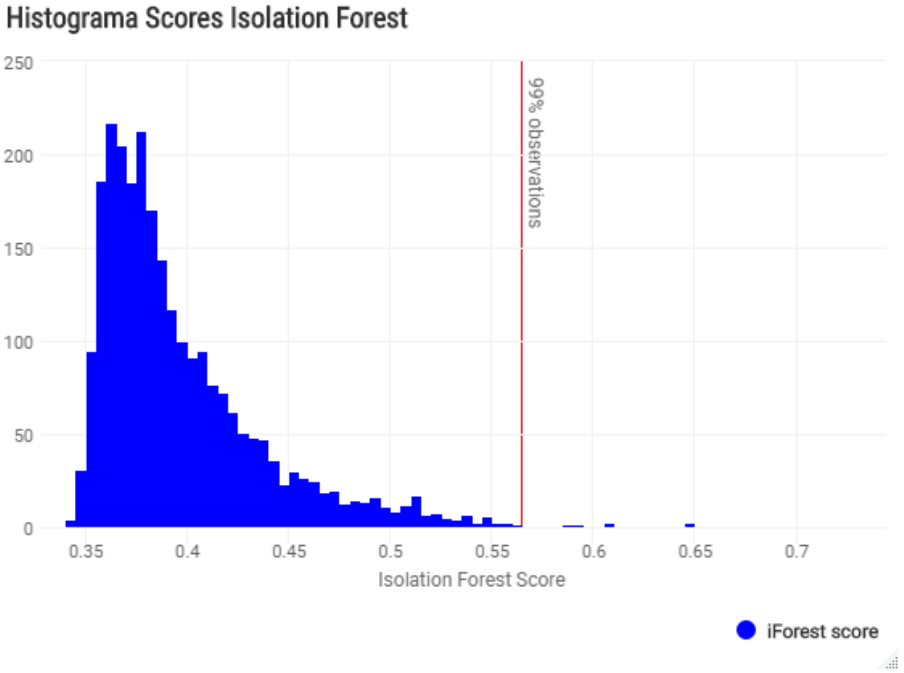


Figura 15- *Histograma de la distribución de los scores arrojados por la técnica de Isolation Forest.*

Se realizó nuevamente el ejercicio de las caras de Chernoff, pero excluyendo los *outliers* provistos por *Isolation Forest*. El resultado se muestra en la **Figura 16** e indicó que Diciembre presentó un comportamiento atípico respecto de los otros meses.

[4] Outliers_Estacional.xlsx

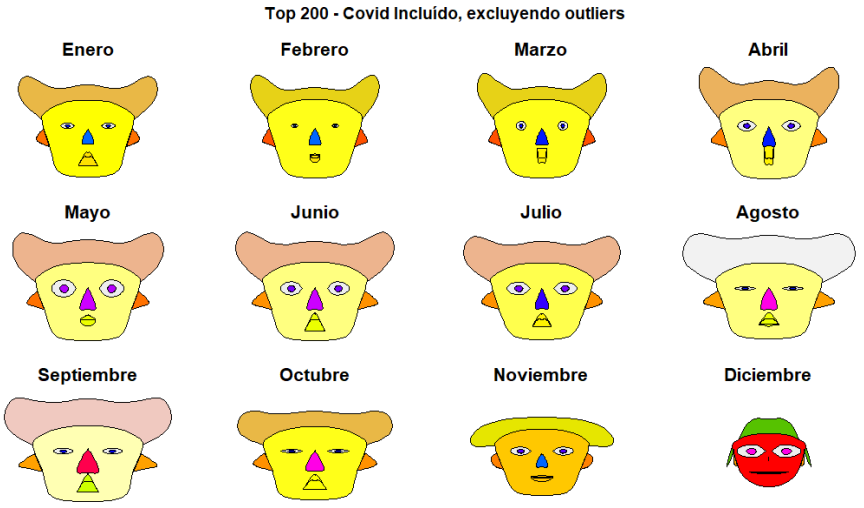


Figura 16- *Caracterización de atributos multivariados utilizando la técnica de caras de Chernoff en los 200 temas del chart, excluyendo los outliers determinados por las técnica de Isolation Forest.*

A continuación se tomaron los atributos de mayor varianza (*acousticness, energy* y *valence*) y se observó su comportamiento a través de los meses (**Figura 17**).

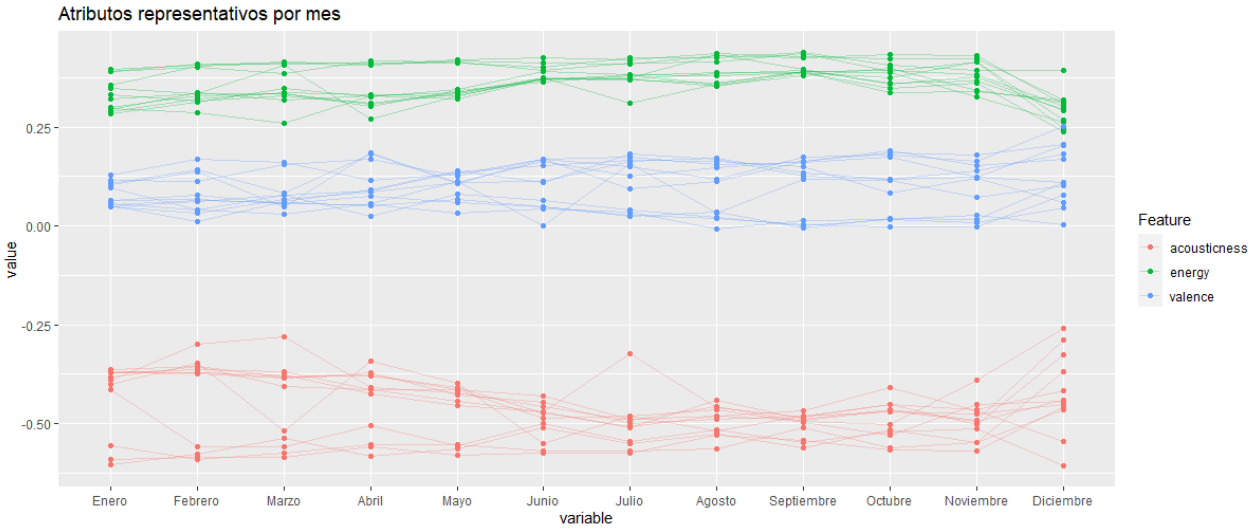


Figura 17- *Caracterización de atributos acousticness, energy y valence a lo largo de los meses del año.*

Nuestros resultados indicaron que el nivel de alegría (*valence*) de la música en diciembre se dispersa y eleva, al igual que el atributo *acousticness,* mientras que la energía (*energy*) baja, y consecuentemente por una correlación positiva, también lo hará el nivel medido en decibeles (*loudness*).

Hipótesis 5: El precio de las acciones de Spotify se correlaciona con el número de streams totales

A continuación evaluamos el comportamiento del precio de cierre de las acciones de Spotify, y la comparamos con el número de *streams* totales por semana (**Figura 17**). Consideramos que esta última variable podría reflejar indirectamente el número de usuarios de esta plataforma, a nivel mundial.

Los resultados indican que en el rango de fechas estudiadas, el número de *streams* totales presenta una tendencia global positiva. Si se evalúa la suma semanal de *streams* puede observarse que la misma muestra un comportamiento irregular, y que siempre aumenta en las semanas que anteceden el fin de año.

Si se analiza el precio de cierre (semanal) de las acciones de Spotify, no parece haber una correlación entre el costo de las acciones y la cantidad de *streams* totales. No obstante, se

observa un aumento claro en el precio unitario luego de la declaración de la pandemia de COVID19.

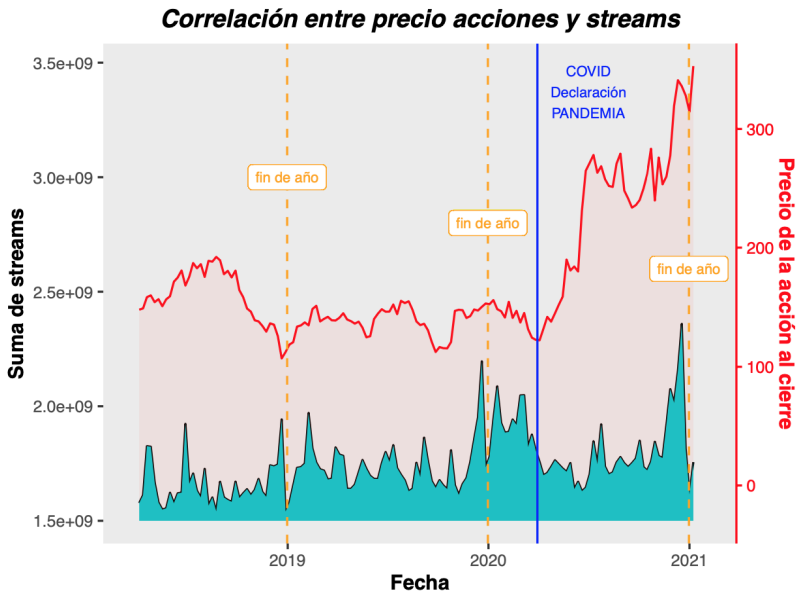


Figura 17- *Análisis secuencial del precio unitario de las acciones de Spotify (precio de cierre) en el período comprendido entre 06 de Abril de 2018 y 08 de Enero de 2021. La curva negra con área color verde se corresponde con el valor de la suma de streams semanales, y la curva roja con área rosada se corresponde con el valor del precio de cierre de la acción de Spotify. Las líneas verticales punteadas color naranja indican el fin de cada año, y la línea vertical azul indica la declaración de COVID como una pandemia.*

DISCUSIÓN

En el presente trabajo se realizó un análisis exploratorio de información relativa a *rankings* semanales de Spotify de los 200 temas más escuchados a nivel mundial, y se integró información de otras fuentes de datos.

Una de las hipótesis de este trabajo era que **existían atributos que se relacionaban con la posición o permanencia ininterrumpida de cada tema en el *chart*.** De los distintos atributos analizados, concluimos que la edad de un tema se correlaciona positivamente con su posición en el *chart*. Los otros atributos analizados no mostraban una clara relación con la posición o permanencia ininterrumpida en el *chart*. Para aquellos temas con mayor permanencia ininterrumpida en el *chart*, se vió que la mayoría presentaba un comportamiento similar caracterizado por un descenso paulatino en su posición a lo largo del tiempo y no influenciado por la declaración de la pandemia COVID. Por el contrario, el tema "Lucid Dreams" (Juice WRLD) presentó un comportamiento diferente, con un ascenso marcado en las posiciones del *chart* hacia fines de 2019. El mismo podría deberse a la repentina muerte de su intérprete en diciembre de dicho año (noticia). De un modo similar, el tema "Sicko mode" (Travis Scott) mostró un ascenso en el *ranking* de los 200 temas más escuchados de Spotify luego de la declaración de la pandemia COVID-19. Este comportamiento se correlaciona con la incorporación del rappero en los videojuegos "Fortnite", que se llevó a cabo con una serie de conciertos virtuales que comenzaron el 23 de Abril de 2020 y el cual fue reproducido casi 46 millones de veces.

Por otro lado, creamos un **índice de éxito** para describir cómo variaba el comportamiento de un artista dado a lo largo del tiempo. Entre los artistas más exitosos, existían algunos con un valor de índice de éxito sostenido en el tiempo, y otros que mostraban un pico de éxito.

Dado que no encontramos una relación clara entre la posición de cada tema y sus características de audio, nos propusimos investigar si **existían otros atributos del tema que tenían influencia sobre su popularidad (definida como la cantidad de veces que dicho tema aparecerá en el *chart*).** Nuestros resultados indicaron que la cantidad de mercados de distribución de un tema influye en su popularidad, pero que existen temas con comportamientos anómalos (instancias más alejadas de la distribución, según la distancia de Mahalanobis). Entre éstos se encontraban temas con alta popularidad y un número acotado de mercados de distribución disponibles. En este punto, es de esperar que exista un sesgo en cuanto a la penetrancia de la aplicación Spotify en los distintos mercados, dada por un número de usuarios

diferente en cada caso, que no estamos considerando en nuestro análisis. Llamativamente, ninguno de los temas más alejados de la distribución fue registrado dentro de los 10 principales temas de mayor permanencia ininterrumpida. Esto nos indica que si bien el atributo *permanencia ininterrumpida* puede ser informativa para el éxito o popularidad de un artista, el mismo quizás deba complementarse con información del atributo *popularidad* (que no contempla interrupciones temporales de presencia en el *chart*), lo cual le conferirá mayor robustez.

A continuación quisimos determinar si **existía estacionalidad en los atributos de los temas que ingresan al *chart*.** A tal fin, se realizó un análisis multivariado de caras de Chertoff, utilizando los datos de temas en las primeras 10 o 200 posiciones del *chart* de cada semana. En ambos casos se observó que los mismos presentan cierta estacionalidad, siendo Diciembre un mes distintivo, incluso cuando se contempla en el análisis el período post declaración de pandemia debido a COVID. Al analizar la estacionalidad de los atributos de audio de los temas con mayor varianza (*energy, valence, acousticness)*, se observó que dichas variables presentan un comportamiento que acompaña la tendencia del análisis de las caras de Chertoff.

Por último decidimos estudiar si **el precio de las acciones de Spotify correlacionaba con el número de *streams* totales.** Si bien el precio de cierre de las acciones de Spotify no mostró una relación con la cantidad de *streams* totales en el período analizado, se observó un claro aumento en el precio unitario luego de la declaración de la pandemia de COVID19. Este comportamiento fue compartido por múltiples empresas en el área de la tecnología (nota).

En conclusión, en este trabajo práctico se procesó un gran volumen de datos provenientes de los *rankings* semanales de Spotify. Luego de limpiar, integrar, reducir la dimensionalidad del *dataset* y crear nuevas variables, aplicamos herramientas de visualización diversas y se revelaron nuevos patrones en la información, que eran imperceptibles en un primer momento.