Descripción de Atributos en cada colección

-Artista:

Atributos: "_id" "Artist"

Dimensión dataset: 676 2

Haciendo un *unique* se encuentra que este dataset no tiene valores repetidos y haciendo un *is.na* se encuentra que no hay missings.

-Charts

Atributos: "_id" "Position" "Track_Name" "Artist" "Streams" "URL" "week_start" "week_end"

Dimensión dataset: 63600 8

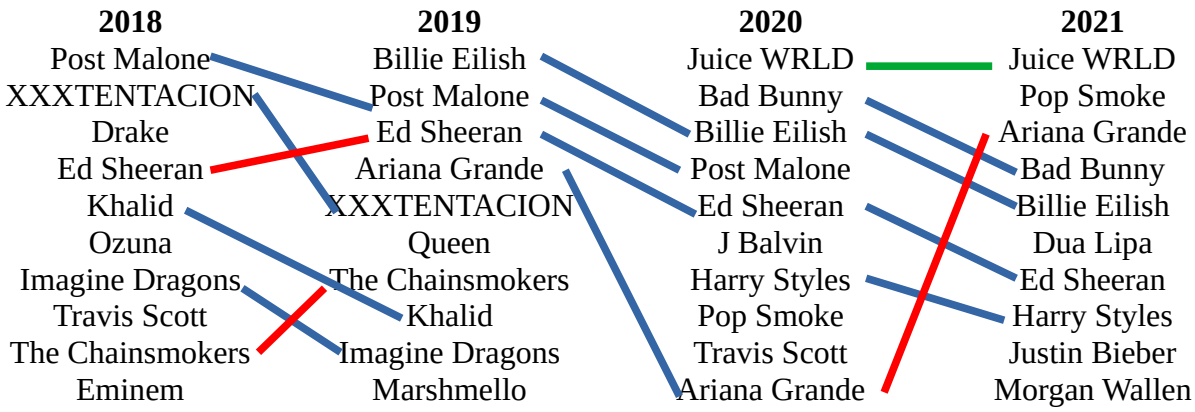
Haciendo un *is.na* se encuentra que no hay missings, pero haciendo una exploración se encuentra que hay filas de datos que si bien tienen distinto identificador, tienen los mismos datos. Unificando estas filas repetidas, el tamaño del nuevo dataset es:

Dimensión dataset: 31404 8

Cantidad de datos por año en dataset charts:

2018	2019	2020	2021
20400	21200	21200	800

- Artistas con mas tracks en las listas por año (primeros 10 puestos):



El color y pendiente de las lineas da idea de que tan rapido ascienden o descienden en los rankins.

Los artistas que mas meten canciones en el ranking 200 son Post Malone, Billie Eilish, Ed Sheeran y Ariadna Grande (por ahora), que por tres años seguidos han metido varias canciones en este ranking.

- Streams

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3589018	5151455	6536358	8647701	9846342	71467874

El 50% de las reproducciones estan por debajo del orden de las ~6,53 millones de reproducciones y el maximo de reproducciones es del orden de ~71,47 millones de reproducciones. Del histograma se ve que las demas medidas de tendencia central no son representativas. El boxplot de Streams total y por año tiene muchos outliers por tanto no es una tecnica adecuada para analizar outliers.

Falta analisis outliers con otras tecnicas.

-Artistas con mas reproducciones

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3747900	5097888	6162743	6832971	7799853	21176356

El 50% de los artistas tienen en promedio por debajo de ~6,16 millones de reproducciones con un maximo de ~21,18 millones. Del histograma se ve que las demas medidas de tendencia central no son representativas. El boxplot tiene muchos outliers por tanto no es una tecnica adecuada para analizar outliers

Cuando se consideran las reproducciones medias y maximas, el ranking es el siguiente:

Mean	**Max**
Tones And I	Ariana Grande
SAINt JHN	Drake
Wham!	Shawn Mendes
Internet Money	Olivia Rodrigo
Jawsh 685	Ed Sheeran
Powfu	Mariah Carey
Mariah Carey	Bad Bunny
24kGoldn	The Weeknd
Bobby Helms	Tones And I
Joel Corry	Billie Eilish

Considerando reproducciones medias parecen meterse mas artistas antiguos en las listas, los artistas nuevos son los que meten hits muy reproducidos. Esto se puede considerar al decidir si analizar popularidad por reproducciones medias o maximas para cada artista.

- Canciones Top

En total durante los años del analisis 39 canciones han ocupado el primer lugar.

Si se discrimina por año se ve que las canciones top cambian año a año, se podria discriminar mas fino para ver como cambian temporada a temporada pero creo que seria mas ruidoso puesto que no se saca tanta informacion debido a que son artistas que pegan un hit y no necesariamente que tengan una popularidad constante o consolidada. Lo que si se puede hacer es usar estas canciones para cruzar con los datos de features y observar porque son exitosas.

2018		**2019**		**2020**		**2021**	
Havana (feat. Young Thug)	Camila Cabello	*Sunflower - Spider-Man: Into the Spider-Verse*	Post Malone	*Dance Monkey*	Tones And I	*DÁKITI*	Bad Bunny
God's Plan	Drake	*7 rings*	Ariana Grande	*The Box*	Roddy Ricch	*drivers license*	Olivia Rodrigo
Call Out My Name	The Weeknd	*bad guy*	Billie Eilish	*Blinding Lights*	The Weeknd		
Nice For What	Drake	*I Don't Care (with Justin Bieber)*	Ed Sheeran	*THE SCOTTS*	THE SCOTTS		
Better Now	Post Malone	*Señorita*	Shawn Mendes	*Rain On Me (with Ariana Grande)*	Lady Gaga		
This Is America	Childish Gambino	*Circles*	Post Malone	*ROCKSTAR (feat. Roddy Ricch)*	DaBaby		
SAD!	XXXTENTACION	*HIGHEST IN THE ROOM*	Travis Scott	*cardigan*	Taylor Swift		
Nonstop	Drake	*Dance Monkey*	Tones And I	*Savage Love (Laxed - Siren Beat)*	Jawsh 685		
In My Feelings	Drake	*Lose You To Love Me*	Selena Gomez	*WAP (feat. Megan Thee Stallion)*	Cardi B		
Lucky You (feat. Joyner Lucas)	Eminem			*Mood (feat. iann dior)*	24kGoldn		

-Artistas con mas tracks top en las listas por año (primeros 5 puestos):

La diferencia con los que meten mas tracks en general,es que estos pueden estar aqui solo porque pegaron un hit. Se pueden cruzar estas listas con las de tracks en general para descartar los que son populares solo por un hit en un solo año y no vuelven a aparecer.

2018		**2019**		**2020**		**2021**	
Drake	4	Post Malone	2	24kGoldn	1	Bad Bunny	1
Ariana Grande	1	Ariana Grande	1	Ariana Grande	1	Olivia Rodrigo	1
Bad Bunny	1	Billie Eilish	1	Bad Bunny	1		
Camila Cabello	1	Ed Sheeran	1	Cardi B	1		
Childish Gambino	1	Selena Gomez	1	DaBaby	1		

- Audio Features

Atributos: "_id" "acousticness" "album_id" "album_images" "album_name" "album_release_date" "album_release_date_precision" "album_release_year" "album_type" "analysis_url" "artist_id" "artist_name" "artists" "available_markets" "danceability" "disc_number" "duration_ms" "energy" "explicit" "external_urls_spotify" "instrumentalness" "is_local" "key"

"key_mode" "key_name" "liveness" "loudness" "mode" "mode_name"
"speechiness" "tempo" "time_signature" "track_href" "track_id"
"track_name" "track_number" "track_preview_url" "track_uri" "type" "valence"

Dimensión dataset: 182049 40

Hay missings:

track_preview_url 77280
album_release_year 415

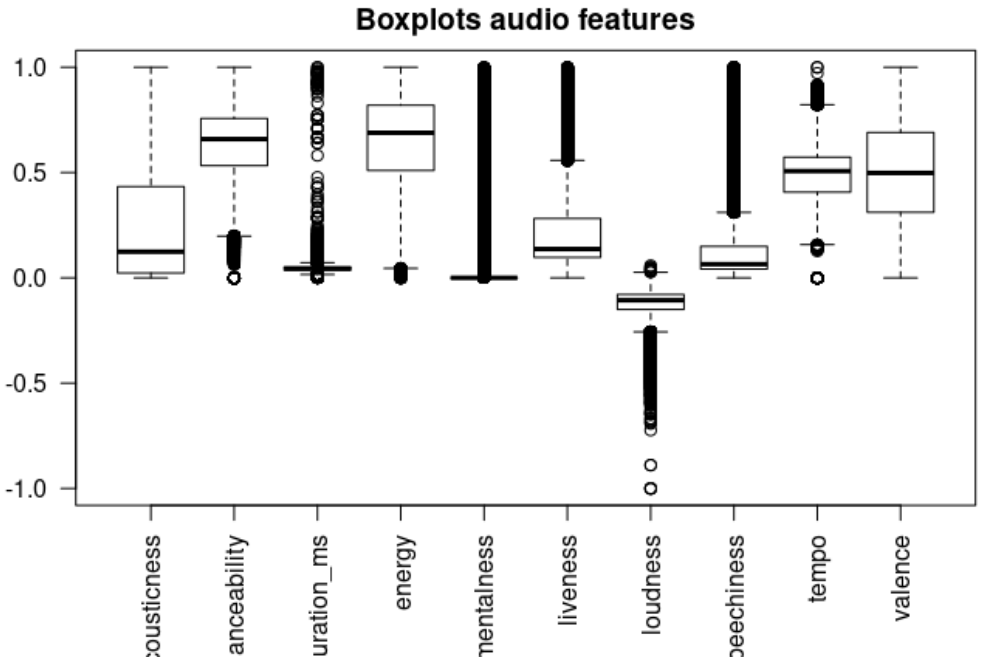
Como no son atributos relevantes podemos no considerarlos.

Haciendo un *unique* el dataset queda de dimensiones:

156621 40

Lo que quiere decir que hay filas 25428 repetidas. Se hace el analisis unificando estas filas.

Los boxplots normalizados de las audio features (para visualizarlos juntos) quedan asi:



para las variables:

"acousticness" "danceability" "duration_ms" "energy" "instrumentalness"
"liveness" "loudness" "speechiness" "tempo" "valence"

Para la mayoria de features no sirve este analisis. Los valores de instrumentalness son muy pequeños la mayoria.

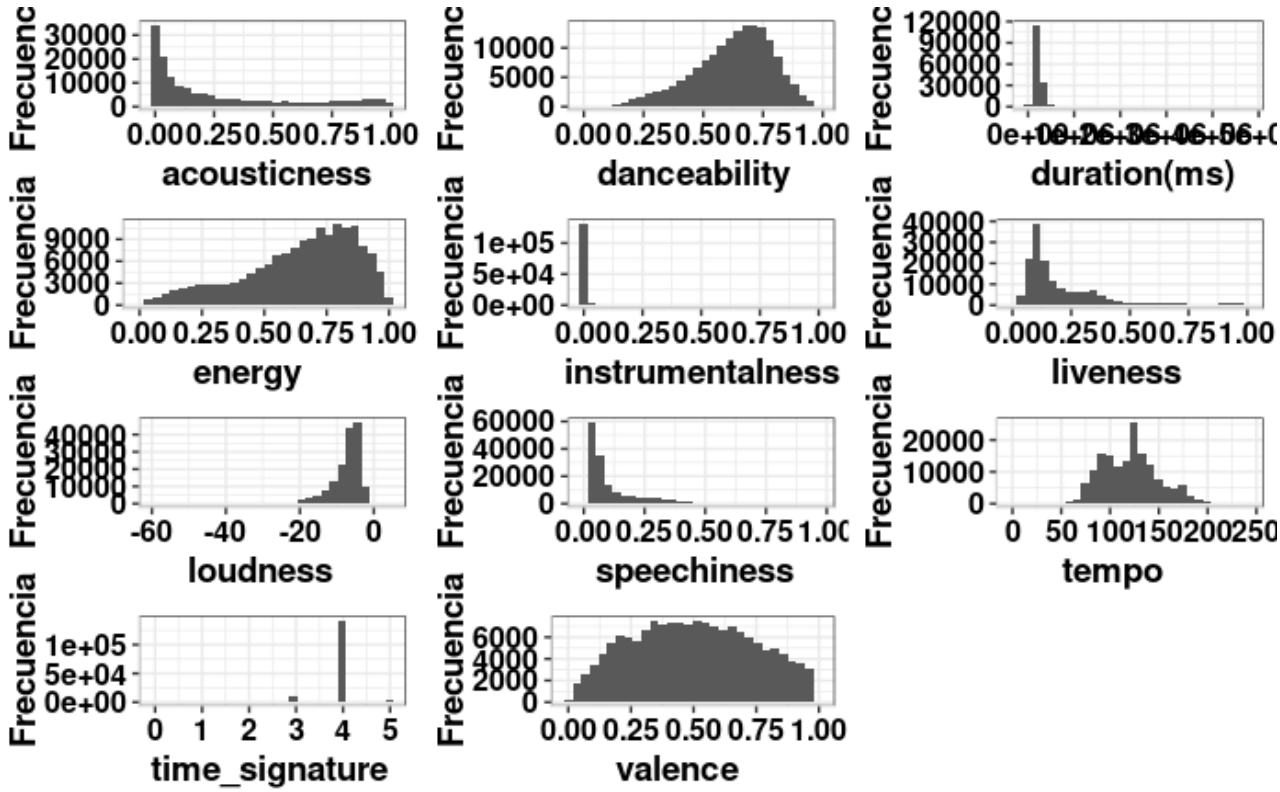
Teniendo la informacion de cuales artistas son los mas populares por numero de tracks en los listados y por tracks numero uno, podemos reducir nuestro analisis de features que hacen que una cancion o artista sea top a estas canciones y artistas.

Antes hay que cruzar la lista de artistas con los del archivo de audio features para que solo esten artistas del archivo de artistas. Al cruzar verificamos que esta todo ok.

Los resultados pueden cambiar al sacar outliers, pero la estructura de analisis que se acaba de describir se puede aplicar igual.

-Distribuciones

Histogramas de las variables audio features:



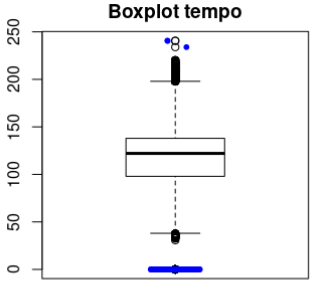
Observamos de la variable time_signature que se puede deducir directamente que la gran mayoria de las canciones en los listados estan compuestas en compases simples 4 tiempos. Para las demas analizamos los boxplots y z_scores para encontrar outliers.

- Outliers

- Z-score

Analisis preliminar con metodos de z-score y z-score modificado para los audio features.
Analizamos con z-score las distribuciones mas simetricas, que vendrian siendo tempo y valence.

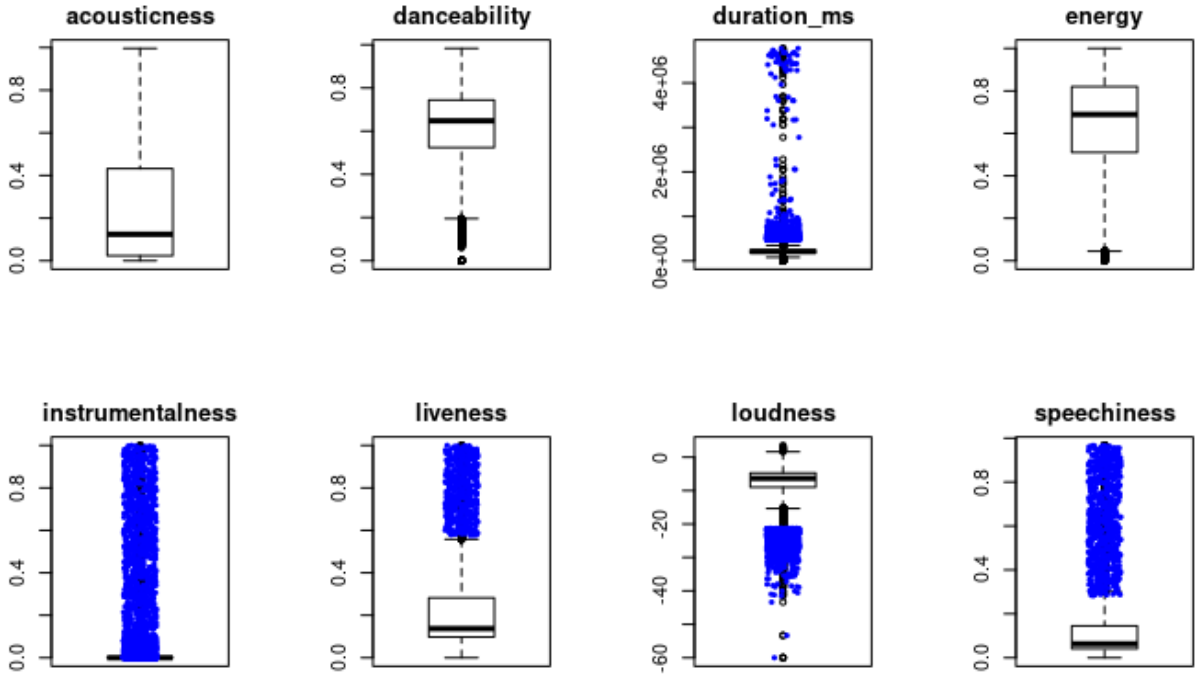
Para valence no se detectan outliers con este metodo, para tempo:



Los datos que tienen tempo 0 es porque seguramente los llenaron con ceros por no tener datos. Estos datos se pueden impugnar mas adelante.

-Z-score modificado

Para las otras variables de audio features:



Como se puede observar el metodo marca muchos puntos como outliers, esto es porque las distribuciones son muy asimetricas, por tanto debemos explorar otros metodos.

- Outliers variables dummy

Calculamos el z-score para t*rackxartist* (cantidad de tracks que cada artista tiene metidas en las listas, tambien se discrimina por año), *streamsxartist* (reproducciones por artista, estan calculadas, reproducciones medias, donde se meten algunos artistas viejos y reproducciones maximas, donde priman los artistas nuevos), *top_tracks* (canciones en el top, se discriminan tambien por año, esta es variable categorica), *NumberOnexartist* (artistas con mas canciones en el primer lugar, se discriminan por tambien por año).

Falta crear streamsxtrack, para determinar las canciones con mas reproducciones, este tendria que tener resultados parecidos a con top_track que se hace con las posiciones.

Falta agrupar las features por los artistas top en listas (con mas canciones o con mas canciones numero uno) y por las canciones top en las listas. No se como se calcularian outliers aqui, o si mas bien aqui ya se entraria a modelar.

Y ahi voy…

despues de esto metemos lo de Marcos.