Entrega Pre-TP1: Data mining en Música: Preparación de los Datos

Integrantes: Paola Grajeda, Adrián Paredes

29 de septiembre de 2019

Dataset *audio_features*

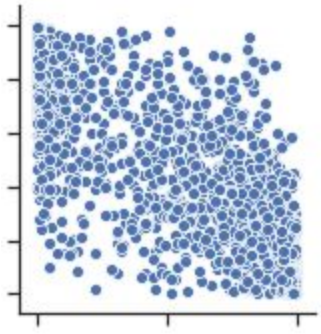
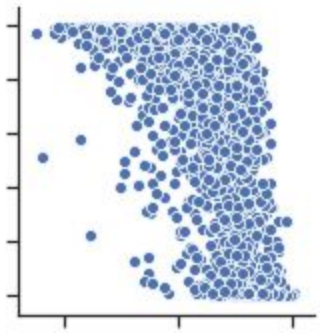
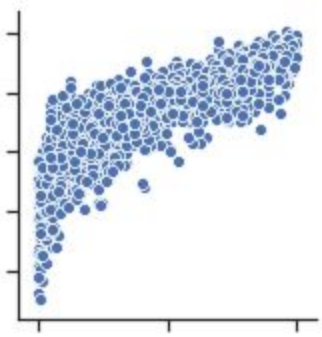
El dataset no tiene valores faltantes.

Para graficar la *scatter matrix* se omiten los siguientes features no numéricos:

- analysis_url
- track_href
- type
- uri
- key (category: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)
- mode (category: 0, 1)
- time_signature (category: 1, 3, 4, 5)

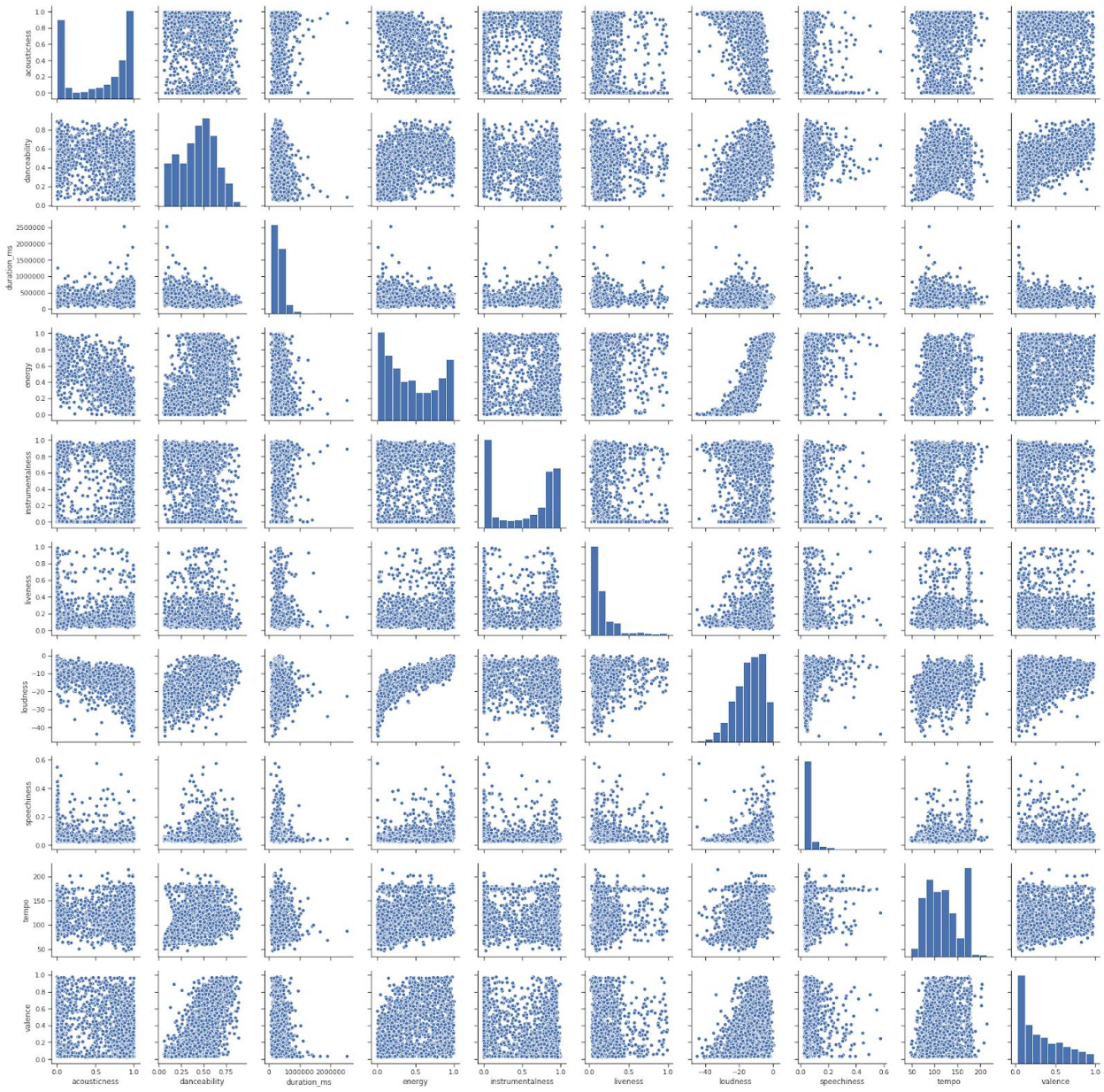
El análisis de la *scatter matrix* y la *correlation matrix* nos muestra que hay correlación entre **energy** y **acousticness** y entre **loudness** y **energy**.

Si somos un poco más laxos con el *cutoff*, podríamos considerar que **acousticness** y **loudness** también están correlacionadas.

		
energy y acousticness (-0.833)	acousticness y loudness (-0.716542)	loudness y energy 0.866154

El feature candidato para reducir la dimensionalidad podría ser **energy**.

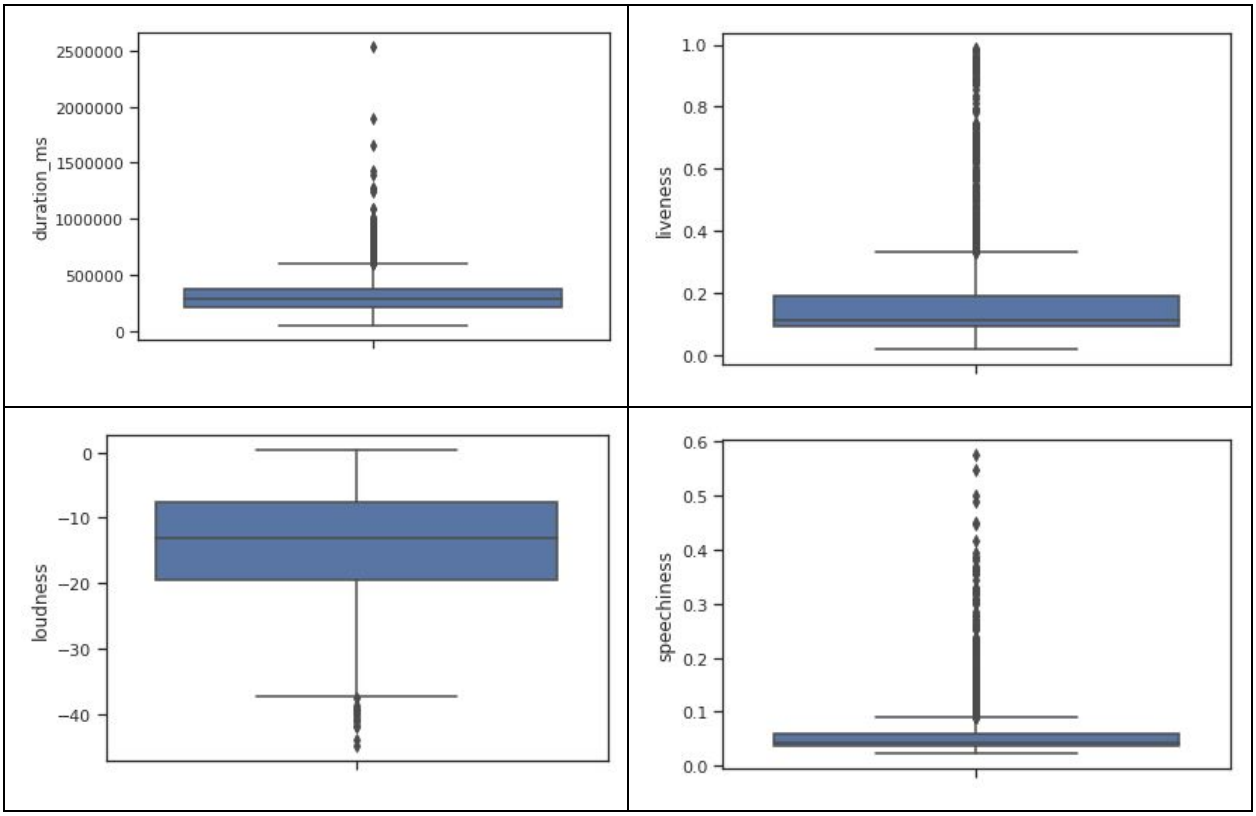
Al estandarizar los valores, obtenemos exactamente los mismos resultados:



Scatter matrix estandarizada

Graficando distintos boxplots, se detecta que las siguientes variables tienen outliers:

- duration_ms
- liveness
- loudness
- speechiness



Boxplots de features con outliers

Se considera outlier extremo al valor fuera del rango [Q1-3*RI ; Q3+3*RI], siendo Q1 y Q3 el primer y tercer quartil y RI el rango intercuartil.

Se calcula la cantidad de outliers extremos por variable:

variable	value
speechiness	154
liveness	122
duration_ms	31

Speechiness y **liveness** tienen entre un 5,5% y 6,9% de sus datos considerados como outliers extremos.

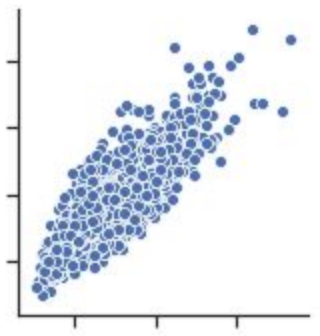
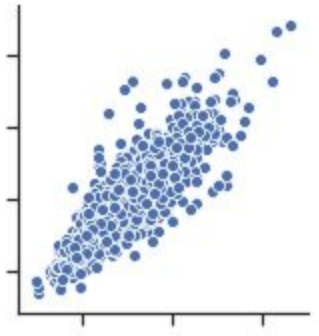
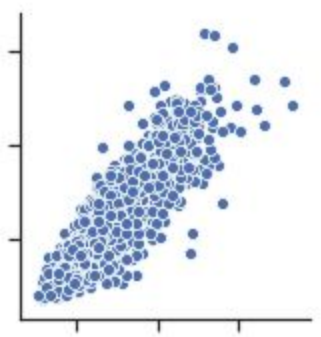
Dataset *audio_analysis*

Se resumen las 12 variables de timbre y las 12 de pitches tomando la media y el desvío estándar de todas las canciones, obteniendo así un dataset de 48 columnas.

A pesar de que en *audio_features* tenemos 2206 canciones, sólo contamos con 2205 datasets de timbres y 2205 datasets de pitches, con lo cual probablemente nos falte la información de análisis de una canción.

Se encuentra que las siguientes variables están correlacionadas:

- **timbre_std_3** con **timbre_std_10**
- **timbre_std_7** con **timbre_std_9**
- **timbre_std_5** con **timbre_std_10**

		
timbre_std_3 y timbre_std_10 (0.845)	timbre_std_7 y timbre_std_9 (0.868)	timbre_std_5 y timbre_std_10 (0.871)

Los features que se pueden omitir del análisis son:

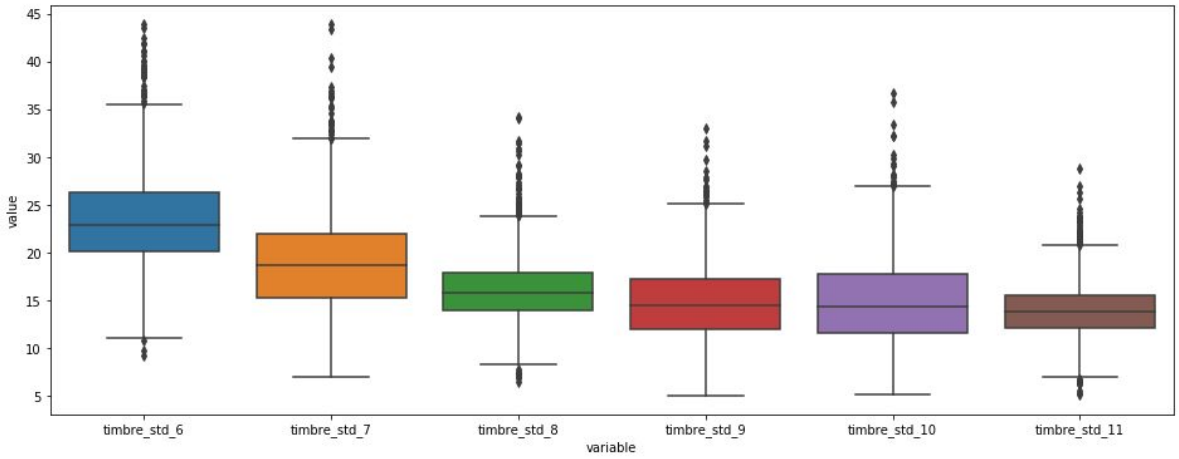
- Timbre_std_10
- Timbre_std_7

Se calcula la *scatter matrix* completa, pero no se incluye en el presente informe debido a las dimensiones de la misma (48x48).

Al estandarizar los valores, obtenemos exactamente los mismos resultados.

A modo de prueba en el armado del dataset, se reemplaza la media y el desvío por la mediana y la MAD respectivamente. La intención es detectar si se obtiene mayor cantidad de correlaciones. Sin embargo, las correlaciones obtenidas son las mismas. Se conserva el código, ya que podrá ser utilizado en la 2da parte del TP para evaluar si se consigue mejor clusterización que con las primeras medidas de resumen.

Volviendo al análisis original, se detectan outliers en las 48 variables. En los boxplot se puede apreciar una simetría a pesar de los outliers. En el gráfico sólo se muestran 6 features, pero el patrón se repite en los 42 restantes.



Boxplots de 6 de las variables resumen construidas con las desviaciones estándar de los timbres 6, 7, 8, 9, 10 y 11.

Se encuentran outliers extremos en 27 features; **pitch_std_0** y **pitch_std_1** tienen 17 y 8 respectivamente. El resto de los features tienen menos de 8.

# outliers	variables
1	9
2	4
3	3
4	2
5	1
6	2
7	4
8	1
17	1

Tabla que indica la frecuencia de cada cantidad de outliers.
Por ejemplo: 1 variable tiene 17 outliers y 4 variables tienen 2 outliers.

En la segunda parte del TP se podrá sustituir dichos valores con N/A y aplicar algún método de imputación de faltantes, o se deberá utilizar algún algoritmo que sea robusto a valores extremos. Sin embargo, puede resultar más relevante el trabajo de limpieza en el dataset de *audio_features* donde existen variables con un 5% y 6% de outliers extremos.