Emmanuel Cardozo
Lucas Didone

Trabajo práctico integrador.
Análisis de datos.

1. Introducción y motivación.

Dataset escogido: *Datos de distintas canciones en* **Spotify**.

El dataset presenta diferentes características de canciones disponibles en Spotify, y el objetivo es a partir de las mismas poder predecir si un nuevo tema será del agrado de la persona que tiene la playlist activa.

Elegimos este dataset ya que nos pareció interesante el tema planteado, además técnicamente es desafiante ya que el conjunto de datos presenta una gran cantidad de variables categóricas, lo que añade una cuota extra de dificultad según nuestro criterio.

2. Análisis exploratorio inicial.

- **Visualización de las primeras filas del conjunto de datos.**

	acousticness	danceability	duration	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence	label
0	0.713	0.514	100125	0.521	0.816000	8	0.1120	-14.835	0	0.0444	119.879	4	0.143	1
1	0.192	0.714	207019	0.614	0.000000	4	0.2630	-6.935	1	0.0319	123.969	4	0.582	1
2	0.333	0.630	216200	0.455	0.000004	5	0.1270	-9.290	1	0.0292	139.931	4	0.199	1
3	0.601	0.810	136413	0.221	0.210000	5	0.1840	-11.005	1	0.0429	109.960	4	0.798	1
4	0.883	0.465	181440	0.459	0.000173	6	0.0692	-8.137	0	0.0351	90.807	4	0.288	1
5	0.524	0.633	244360	0.401	0.000000	4	0.1230	-12.549	1	0.0439	134.978	4	0.523	1
6	0.597	0.507	183573	0.795	0.000000	9	0.2960	-6.966	1	0.0607	165.540	4	0.900	0
7	0.452	0.825	259102	0.435	0.609000	1	0.0953	-9.582	1	0.0568	119.038	4	0.243	1
8	0.748	0.420	366179	0.324	0.839000	9	0.0723	-14.700	0	0.0556	183.020	3	0.330	1
9	0.913	0.292	197613	0.246	0.088300	0	0.2090	-9.758	1	0.0330	140.316	4	0.249	1

- **Resumen de 5 números (sobre variables numéricas).**

	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	valence
min	0.000001	0.10700	0.00925	0.000000	0.02400	-29.60100	0.02340	0.0332
25%	0.037150	0.48000	0.42325	0.000000	0.09455	-10.17350	0.03590	0.2970
50%	0.244500	0.60600	0.63150	0.000010	0.12900	-7.27000	0.04875	0.4830
75%	0.678500	0.71575	0.80475	0.002245	0.26475	-5.09775	0.11300	0.6845
max	0.994000	0.98600	0.99500	0.967000	0.97900	-0.53300	0.72100	0.9750

- **Descripción de las variables.**

Se identifican los tipos de datos de las variables. Además, se clasifican en variables de entrada y de salida.

Variable	Tipo de dato	Descripción	¿Es informativa para un problema de clasificación?	Entrada o salida
acousticness	Numérico (ordinal).	Indica el grado de acusticidad de una canción.	Sí	Entrada
danceability	Numérico (ordinal).	Indica que tan bailable es una canción.	Sí	Entrada
energy	Numérico (ordinal).	Indica el nivel de energía de una canción.	Sí	Entrada
instrumentalness	Numérico (ordinal).	Indica qué tan instrumental es una canción.	Sí	Entrada
key	Categórico.	Indica la tonalidad en la que se encuentra la canción.	Sí (One Hot Enconding)	Entrada
liveness	Numérico (ordinal).	Indica qué tan "en vivo" es la canción.	Sí	Entrada
loudness	Numérico (ordinal).	Indica el volumen general de una pista en decibelios (dB)	Sí	Entrada
mode	Categórico.	Indica la modalidad (mayor o menor) de una canción. Contenido melódico.	Sí (One Hot Enconding).	Entrada
speechiness	Numérico (ordinal).	Indica la proporción de discurso que	Sí	Entrada

		hay en la canción.		
tempo	Categórico.	El tempo general estimado de una pista en pulsaciones por minuto (BPM).	Sí (One Hot Enconding)	Entrada
time_signature	Categórico.	El compás es una convención de notación para especificar cuántos tiempos hay en cada compás.	Sí (One Hot Enconding)	Entrada
valence	Numérico (ordinal)	Describe la positividad musical que transmite una canción.	Sí	Entrada
label	Categórico.	Indica si la canción es del gusto de la persona que tiene esta playlist activa	No (ya que es la variable de salida)	Salida

Variables de entrada: Análisis por tipo.

Variables numéricas:

- **Skewness (asimetría):**

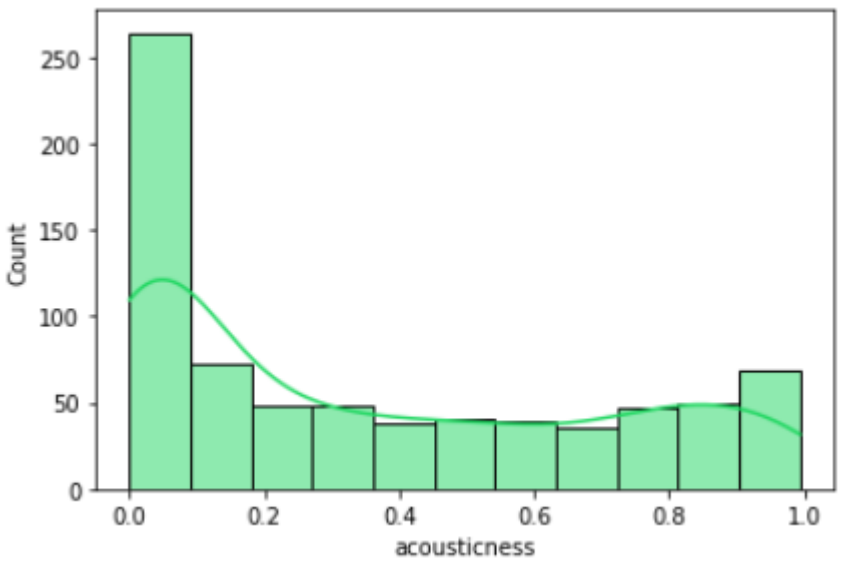
VARIABLE	SKEW
ACOUSTICNESS	0.534804
DANCEABILITY	-0.311981
ENERGY	-0.458765
INSTRUMENTALNESS	2.488166
LIVENESS	2.156240
LOUDNESS	-1.693115
SPEECHINESS	2.040370
VALENCE	0.104812

- **Kurt (curtosis)**

VARIABLE	SKEW
ACOUSTICNESS	-1.210296
DANCEABILITY	-0.296706
ENERGY	-0.758962
INSTRUMENTALNESS	4.518012
LIVENESS	4.842701
LOUDNESS	3.232917
SPEECHINESS	4.149336
VALENCE	-0.911007

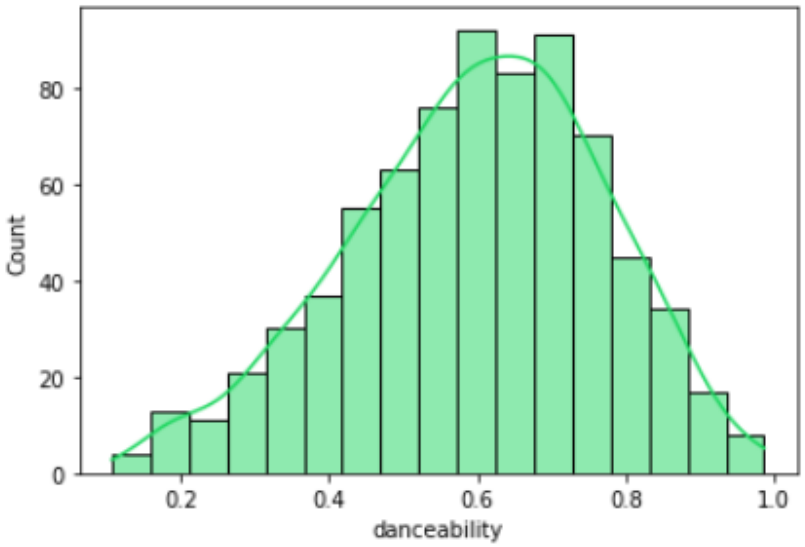
1. acousticness

mean	0.357394
std	0.338405
min	0.000001
25%	0.037150
50%	0.244500
75%	0.678500
max	0.994000



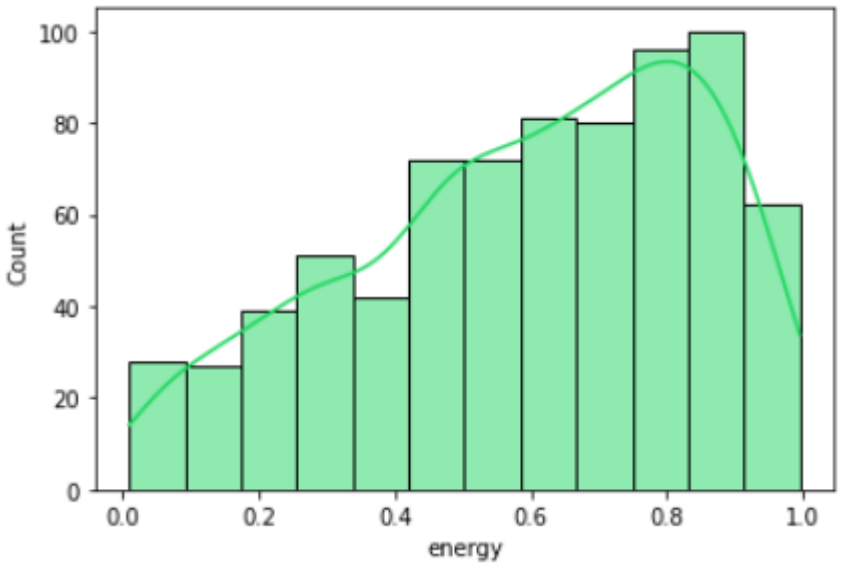
2. danceability

danceability	
mean	0.596439
std	0.172036
min	0.107000
25%	0.480000
50%	0.606000
75%	0.715750
max	0.986000



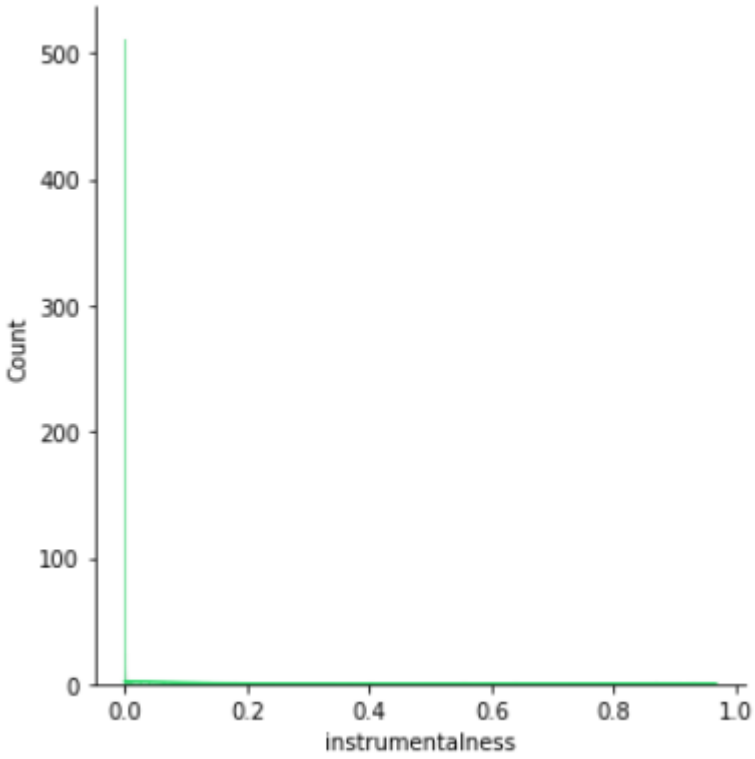
3. energy

mean	0.594188
std	0.253301
min	0.009250
25%	0.423250
50%	0.631500
75%	0.804750
max	0.995000



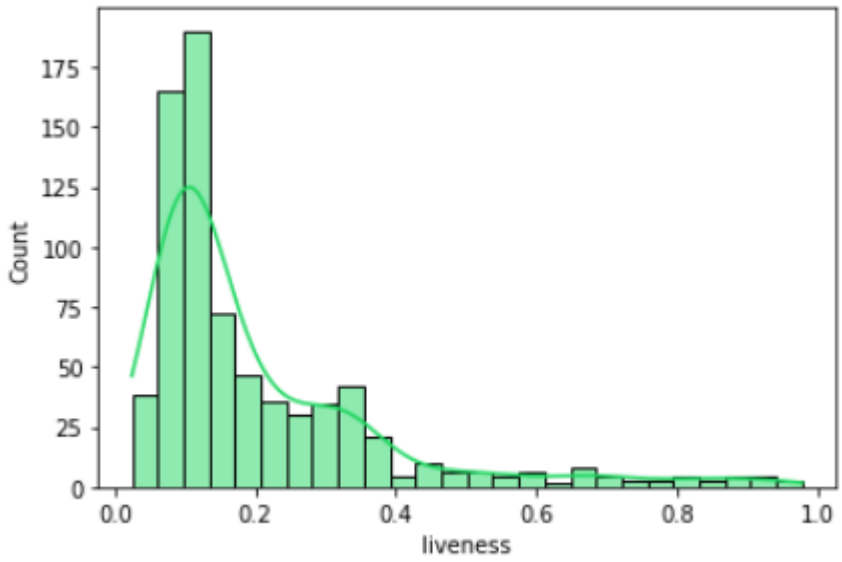
4. instrumentalness

mean	0.100245
std	0.259921
min	0.000000
25%	0.000000
50%	0.000010
75%	0.002245
max	0.967000



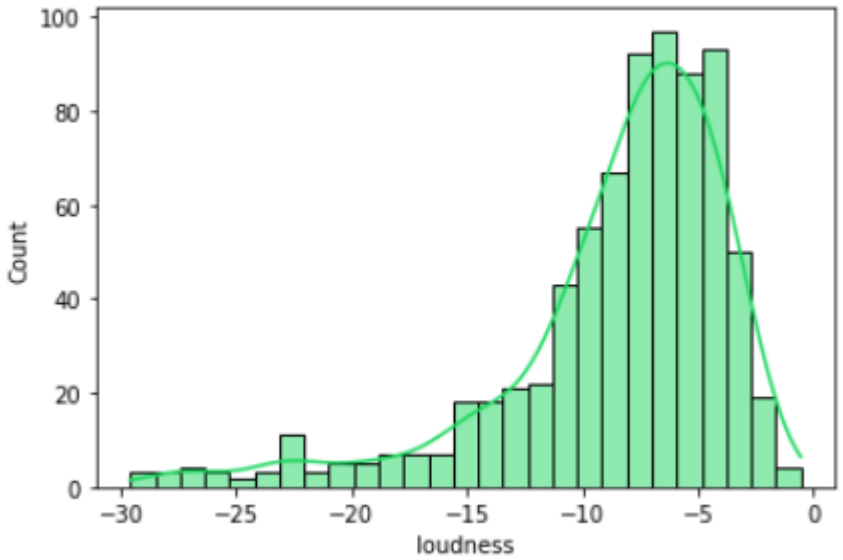
5. liveness

mean	0.203376
std	0.177609
min	0.024000
25%	0.094550
50%	0.129000
75%	0.264750
max	0.979000



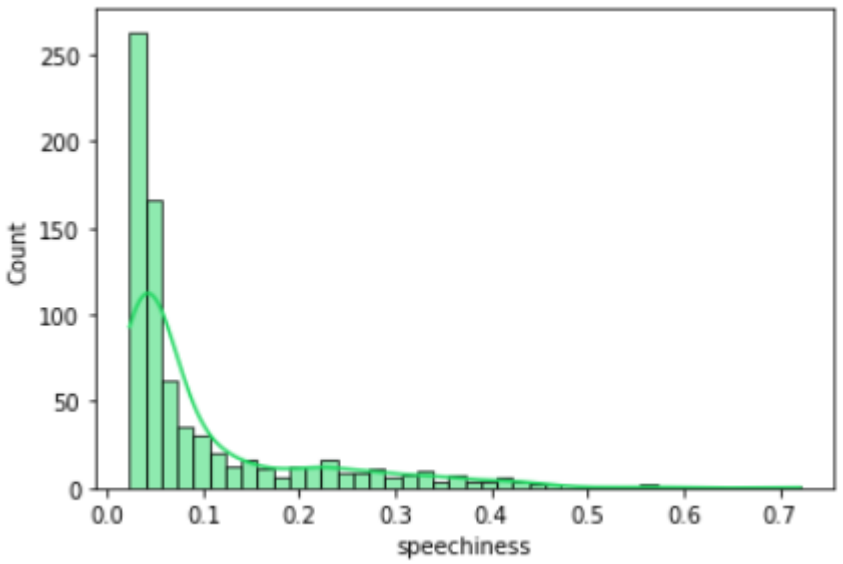
6. loudness

mean	-8.509339
std	5.039488
min	-29.601000
25%	-10.173500
50%	-7.270000
75%	-5.097750
max	-0.533000



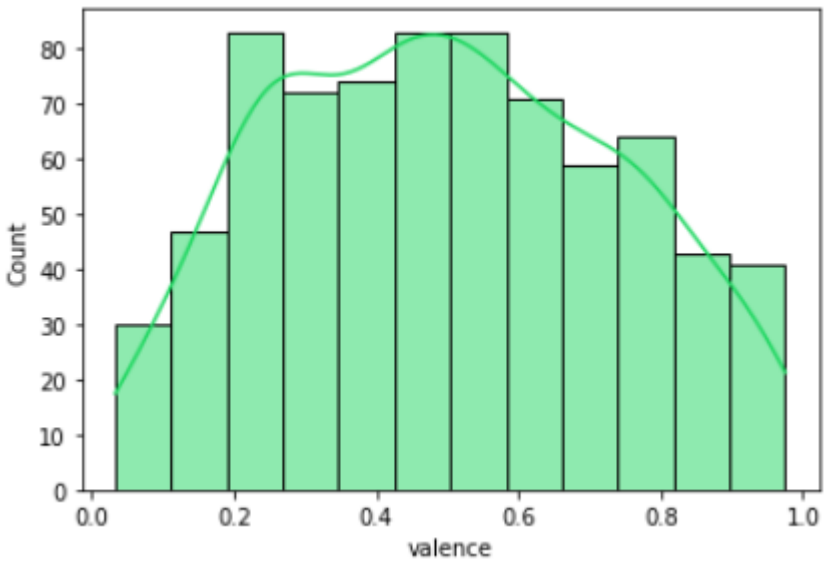
7. speechiness

mean	0.098966
std	0.104715
min	0.023400
25%	0.035900
50%	0.048750
75%	0.113000
max	0.721000



8. valence

mean	0.497321
std	0.239615
min	0.033200
25%	0.297000
50%	0.483000
75%	0.684500
max	0.975000



Conclusiones

A excepción de las columnas "acousticness", la aproximación de las features del dataset de Spotify a una variable normal es correcta (en el caso de "acousticness", se podría aproximar a 2 normales).

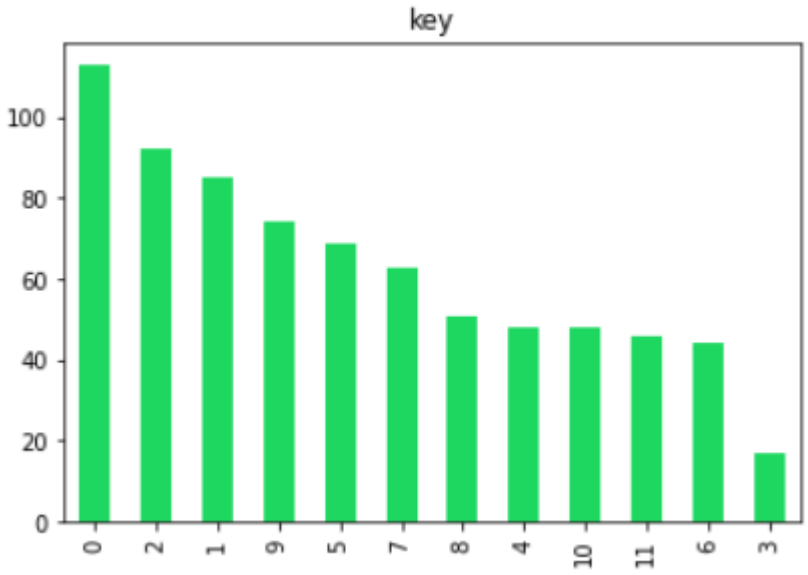
En cuanto a la curtosis de las features, las columnas duration,instrumentalness, liveness, loudness, speechiness y time_signature presentan distribuciones leptocúrticas (mayor concentración alrededor de la media), mientras que las restantes features presentan una distribución platicúrtica.

Con respecto a la oblicuidad de las features o "skewness", se consideran bastante simétricas: ‘acousticness’, danceability, energy, key, tempo y valence, mientras que las restantes features, se consideran considerablemente asimétricas.

Variables categóricas:

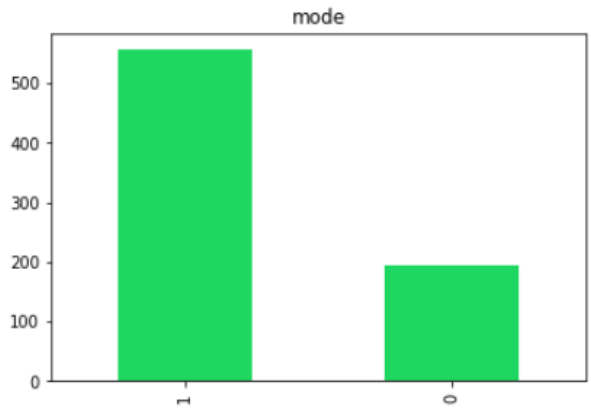
1. **key:** frecuencia de cada categoría

0	0.022667
1	0.058667
2	0.061333
3	0.064000
4	0.064000
5	0.068000
6	0.084000
7	0.092000
8	0.098667
9	0.113333
10	0.122667
11	0.150667



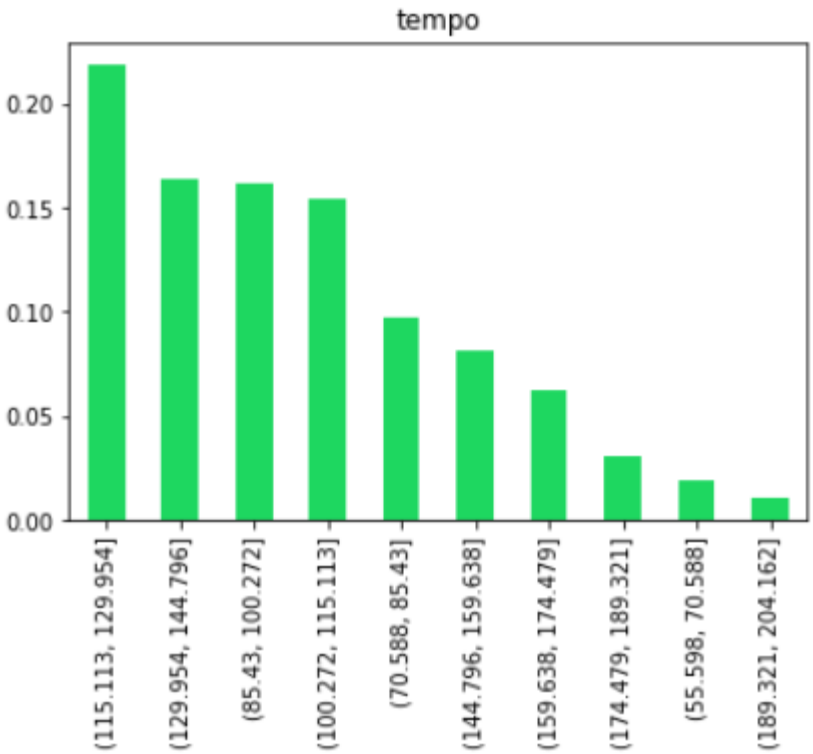
2. **mode:** frecuencia de cada categoría

0	0.258667
1	0.741333



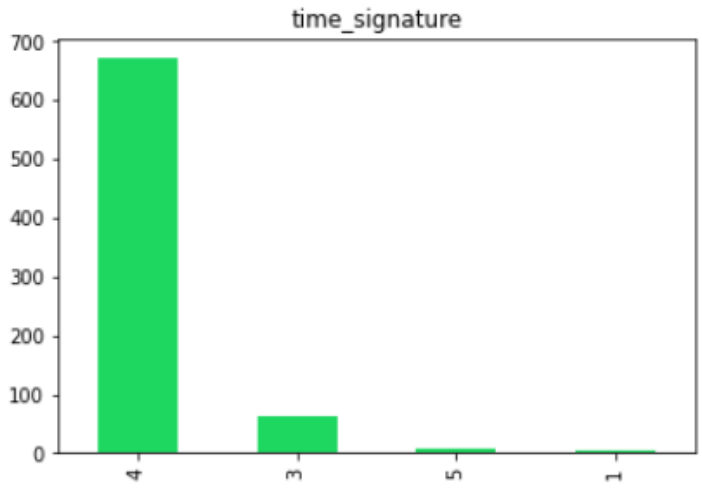
3. **Tempo:** frecuencia de cada categoría (agrupados en 10 bins).

0	0.010667
1	0.018667
2	0.030667
3	0.062667
4	0.081333
5	0.097333
6	0.154667
7	0.161333
8	0.164000
9	0.218667



4. **time_signature:** frecuencia de cada categoría

0	0.008000
1	0.012000
2	0.085333
3	0.894667



Conclusiones

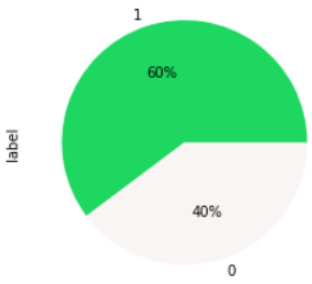
En lo referente a las features con variables categóricas, las features "time_signature" (tipo de compás) y "mode" (modos de un tono: mayor o menor), son las de menor número de variables categóricas. La feature "key" (tono), es la de mayor cantidad número de variables. Además, para poder analizar la variable tempo, fue necesario agrupar sus datos en grupos (se determinó que la cantidad de los mismos sea de 10).

No se evidencian variables compuestas

Análisis de la variable de salida

Frecuencia relativa de cada clase:

label	
0	0.397333
1	0.602667



La variable de salida "label" (like o no de una canción) se encuentra levemente desbalanceada en favor de la clase 1 (es decir la clase que indica que al usuario le gustó la canción).

Análisis de datos faltantes

acousticness	0
danceability	0
duration	0
energy	0
instrumentalness	0
key	0
liveness	0
loudness	0
mode	0
speechiness	0
tempo	0
time_signature	0
valence	0
label	0

No se evidencia la presencia de valores faltantes o nulos.