Práctica de Laboratorio:
Análisis Exploratorio de Datos - Data Wrangling

Docente: Ana Maria CuadrosValdivia

ALVAREZ ASTETE JHEEREMY MANUEL

Para realizar el Análisis Exploratorio de datos, lo primero que deberíamos hacer es intentar responder a las siguientes preguntas (data wrangling):

Paso 1: Analiza el comportamiento de tus datos.

- Un registro es una entidad, describa que representa un registro

  Cada fila del dataset representa una **canción** individual y contiene:

  - El artista (`Singer`), título (Song), género (`Genre`) y década (`Decade`).
  - Análisis emocional (`Valence`, `Arousal`, `Vader_Score`, `Sentiment`).
  - Características textuales (`Total_Word_Count`, `Average_Word_Length`, etc.).
  - Clasificación emocional (`Cluster_Label`).

- ¿Cuántos registros hay?
  - ¿Son demasiado pocos?

```
df.shape
```

```
7]:  (4583, 18)
```

  - ¿Son muchos y no tenemos Capacidad (CPU+RAM) suficiente para procesarlo?

  4583 registros es un tamaño **moderado**.
  Es **suficiente para realizar análisis exploratorios, visualizaciones e incluso entrenar modelos de clasificación supervisada simple**, especialmente si el conjunto de datos está balanceado en sus clases.
  - ¿Hay datos duplicados? NO

- ¿Qué datos son discretos y cuáles continuos?

Columna	Tipo de dato	Tipo lógico (discreto / continuo / categórico)

`Singer`	object	Categórica (nombre del artista)
`Song`	object	Categórica (título de la canción)
`Genre`	object	Categórica (género musical)
`Vader_Score`	object	Categórica / Compleja (es un diccionario o string JSON, requiere procesamiento)
`Valence`	float64	Continua (valor emocional entre 0 y 9)
`Arousal`	float64	Continua (nivel de activación emocional entre 0 y 9)
`Decade`	object	Categórica (década en que salió la canción)
`Sentiment`	object	Categórica (etiqueta emocional: feliz, triste, etc.)
`Total_Word_Count`	int64	Discreta (cantidad total de palabras)
`Unique_Word_Count`	int64	Discreta (cantidad de palabras únicas)
`Fraction_Unique_Words`	float64	Continua (proporción entre 0 y 1)
`Numer_Unique_Word_Lengths`	int64	Discreta (cantidad de longitudes únicas de palabras)
`Min_Word_Length`	int64	Discreta (longitud mínima de palabra en letras)
`Max_Word_Length`	int64	Discreta (longitud máxima de palabra en letras)
`Average_Word_Length`	float64	Continua (promedio de longitud de palabras)
`Sum_All_Word_Lengths`	int64	Discreta (suma total de caracteres de todas las palabras)
`Cluster_Label`	object	Categórica (etiqueta de clúster emocional)

- Muchas veces sirve obtener el tipo de datos: texto, int, double, float
  - ¿Cuáles son los tipos de datos de cada columna?

```
Unnamed: 0                  int64
Singer                     object
Song                       object
Genre                      object
Vader_Score                object
Valence                   float64
Arousal                   float64
Decade                     object
Sentiment                  object
Total_Word_Count            int64
Unique_Word_Count           int64
Fraction_Unique_Words     float64
Numer_Unique_Word_Lengths   int64
Min_Word_Length             int64
Max_Word_Length             int64
Average_Word_Length       float64
Sum_All_Word_Lengths        int64
Cluster_Label              object
dtype: object
```

  - ¿Entre qué rangos están los datos de cada columna?, valores únicos, min, max

```
Unnamed: 0                  4583
Singer                      1368
Song                        4507
Genre                          5
Vader_Score                 4570
Valence                     4513
Arousal                     4452
Decade                        10
Sentiment                      4
Total_Word_Count             831
Unique_Word_Count            388
Fraction_Unique_Words       3780
Numer_Unique_Word_Lengths     13
Min_Word_Length                3
Max_Word_Length               19
Average_Word_Length         4156
Sum_All_Word_Lengths        1942
Cluster_Label                  4
dtype: int64
```

  - ¿Todos los datos están en su formato adecuado?
    SI
  - Los datos tienen diferentes unidades de medida?
    No. Todas las puntuaciones (`Valence`, `Arousal`) están en escalas **estandarizadas del 1 al 9**, y los demás no tienes un rango o una medida, solo es por cantidad.
  - Cuáles son los datos categóricos, ¿hay necesidad de convertirlos en numéricos? No

Categóricos:

- `Genre`, `Sentiment`, `Cluster_Label`, `Decade`

- ¿Qué representa un registro?
  - Describe qué representa cada fila.
    - El artista (`Singer`), título (Song), género (`Genre`) y década (`Decade`).
    - Análisis emocional (`Valence`, `Arousal`, `Vader_Score`, `Sentiment`).
    - Características textuales (`Total_Word_Count`, `Average_Word_Length`, etc.).
    - Clasificación emocional (`Cluster_Label`).
  - Si es una data etiquetada, como interpretas la información de las clases?
  - ¿Hay niveles de granularidad de los datos? Por ejemplo, datos a nivel país, región, ciudad. Años, meses, días, horas, minutos, etc.
    Sí: `Decade`, `Genre`, `Artist` permiten agrupar por **tiempo, estilo o autor**.

- ¿Están todas las filas completas o tenemos campos con valores nulos?
  Con respecto a la fecha, si hay, pero para el uso no es necesario
  - En caso que haya demasiados nulos: ¿Queda el resto de información inútil?. Se debe agregar o combinar sus datos.
    No
  - Si se agregan datos debe comprobar que siguen el mismo comportamiento. Por ejemplo, tiene la misma media, mediana, etc.

- ¿Siguen alguna distribución?
  Usa describe() y analiza los valores.

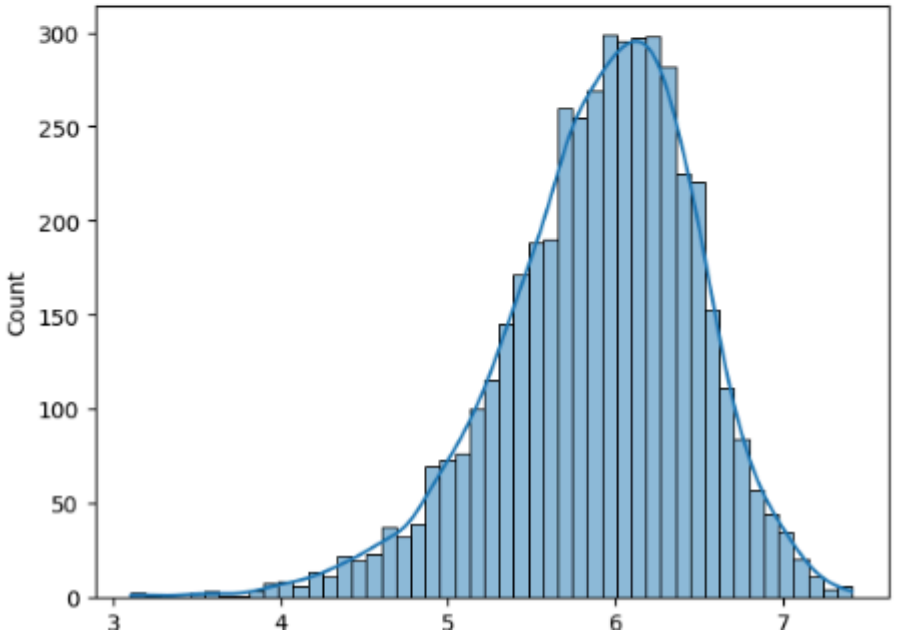
Valence	Arousal	Total_Word_Count	Unique_Word_Count	Fraction_Unique_Words	Numer_Unique_Word_Lengths	Min_Word_Length	Max_Word_Length	Average_Word_Length	Sum_All_Word_Lengths
583.000000	4583.000000	4583.000000	4583.000000	4583.000000	4583.00000	4583.000000	4583.00000	4583.000000	4583.000000
5.900379	4.286748	315.121318	120.089243	0.411622	9.70216	1.007419	10.33297	3.688579	1156.172813
0.580384	0.387666	216.950667	78.188572	0.126308	1.42271	0.088328	2.43877	0.327569	803.203633
3.107586	3.195156	52.000000	5.000000	0.024725	4.00000	1.000000	5.00000	2.743902	220.000000
5.560333	4.010104	168.000000	70.000000	0.321633	9.00000	1.000000	9.00000	3.473557	621.000000
5.961364	4.266053	250.000000	96.000000	0.402477	10.00000	1.000000	10.00000	3.659148	920.000000
6.300595	4.521461	394.500000	141.000000	0.494169	11.00000	1.000000	11.00000	3.862868	1431.000000
7.412500	6.422523	2828.000000	1027.000000	0.872727	16.00000	3.000000	119.00000	6.991803	10794.000000

[16]:

	Valence	Arousal
count	4583.000000	4583.000000
mean	5.900379	4.286748
std	0.580384	0.387666
min	3.107586	3.195156
25%	5.560333	4.010104
50%	5.961364	4.266053
75%	6.300595	4.521461
max	7.412500	6.422523

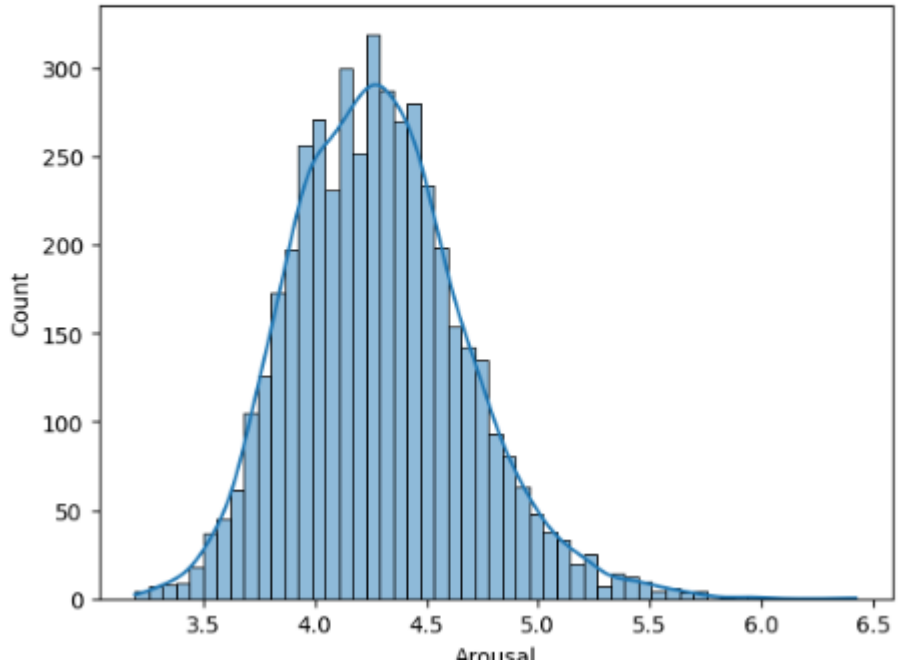
[17]:
```
import seaborn as sns
sns.histplot(df['Valence'], kde=True)
```

[17]:
```
<Axes: xlabel='Valence', ylabel='Count'>
```



```
import seaborn as sns
sns.histplot(df['Arousal'], kde=True)
```

```
<Axes: xlabel='Arousal', ylabel='Count'>
```



- Usa medidas estadísticas:
  - Medidas de tendencia central: media aritmética, geométrica, armónica, mediana, moda, desviación estándar.
  - Correlación y covarianza: permite entender la relación entre dos variables aleatorias.

Sí, puedes usar todas esas variables, pero en el análisis estadístico y de correlación solo tienen sentido usar las variables numéricas (como Valence, Arousal, Total_Word_Count, y las métricas extraídas de Vader_Score).

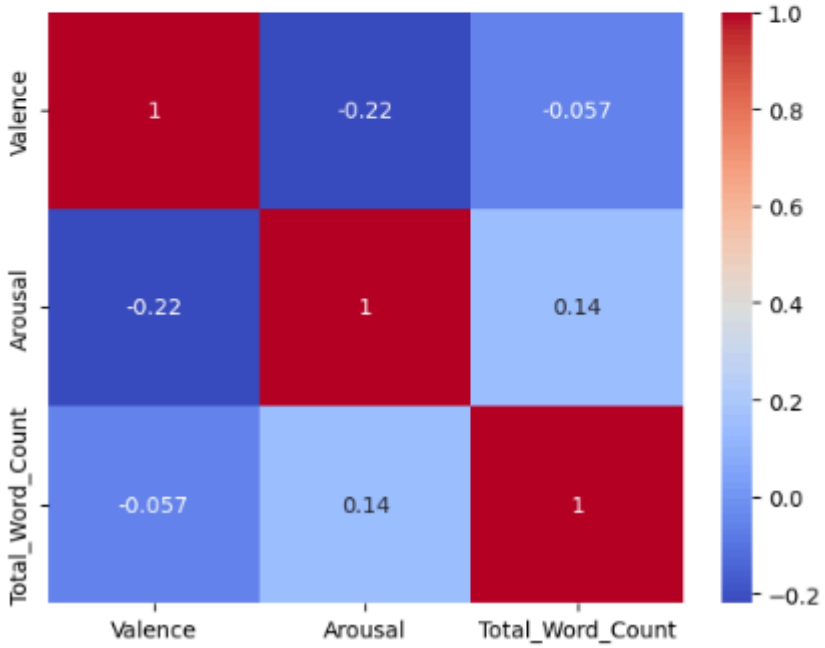
- ¿Hay correlación entre features (características)?

```
corr_matrix = df[['Valence', 'Arousal', 'Total_Word_Count']].corr()
print(corr_matrix)
```

```
                  Valence   Arousal  Total_Word_Count
Valence          1.000000 -0.219290         -0.056895
Arousal         -0.219290  1.000000          0.142734
Total_Word_Count -0.056895  0.142734          1.000000
```

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.heatmap(df[['Valence', 'Arousal', 'Total_Word_Count']].corr(), annot=True, cmap='cool
plt.show()
```




```
cov_matrix = df[['Valence', 'Arousal', 'Total_Word_Count']].cov()
print(cov_matrix)
```

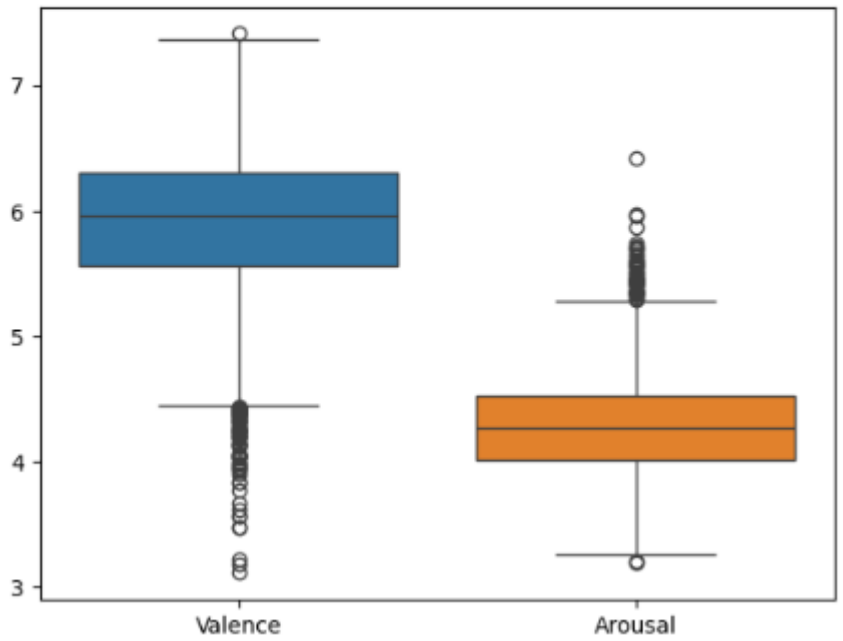
```
                   Valence    Arousal  Total_Word_Count
Valence           0.336846  -0.049339         -7.163961
Arousal          -0.049339   0.150285         12.004538
Total_Word_Count -7.163961  12.004538      47067.592001
```

Paso 2. Análisis de outliers

- ¿Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y “contaminan” ó desvían las distribuciones)

```
[30]: sns.boxplot(data=df[['Valence', 'Arousal']])
```

```
[30]: <Axes: >
```



- ¿Podemos eliminarlos? ¿Es importante conservarlos?
- son errores de carga o son reales?

Sí, si son reales.
Porque las emociones extremas pueden ser justamente lo que hace especial a ciertas canciones o estilos.

Paso 3: Visualización

- Las variables que podemos representar son:
  - Variables categóricas: Gráfico de barras y circular
  - Variables numéricas: Una variable: histogramas, dos variables: boxplot

Gráfico de barras: comparar cantidades de una variable.

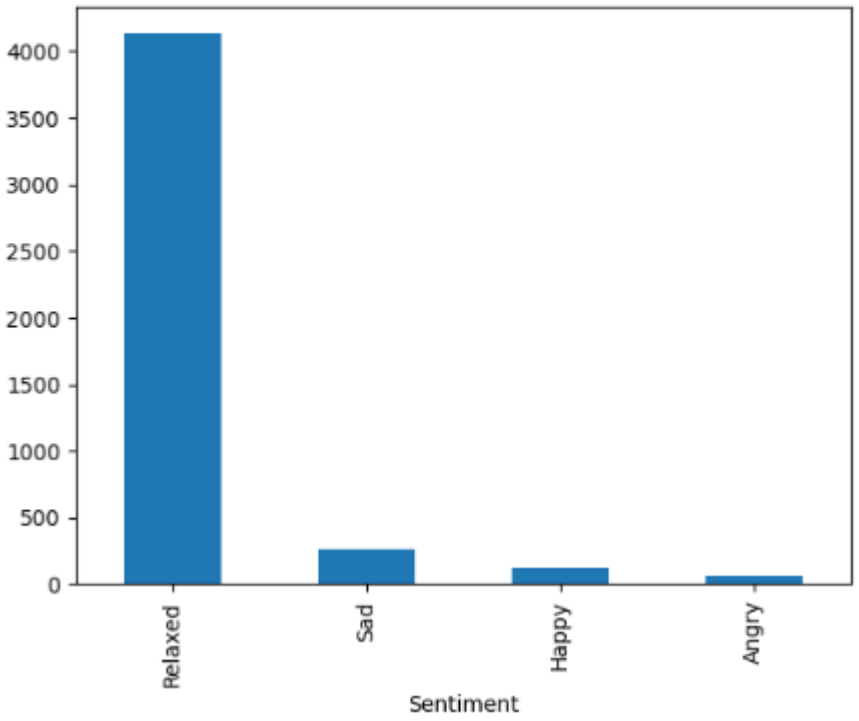
Gráfico circular: para representar porcentajes y proporciones.

Boxplot: representa los datos numéricos a través de sus cuartiles pudiendo representar los outliers.

Scatterplot: muestra el grado de relación entre dos variables.

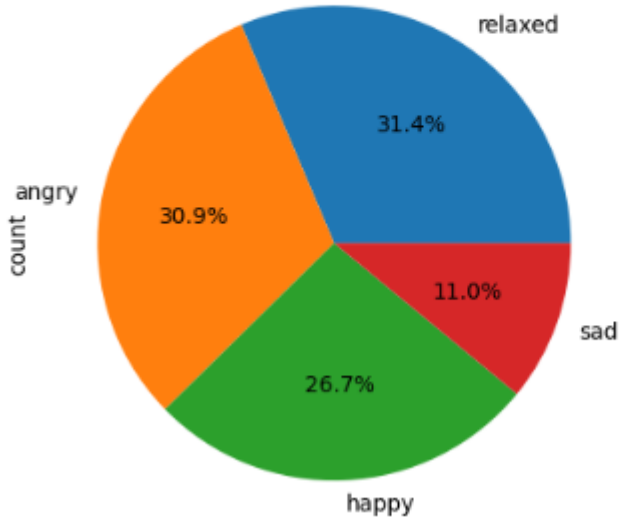
```
df['Sentiment'].value_counts().plot(kind='bar')
```

```
<Axes: xlabel='Sentiment'>
```



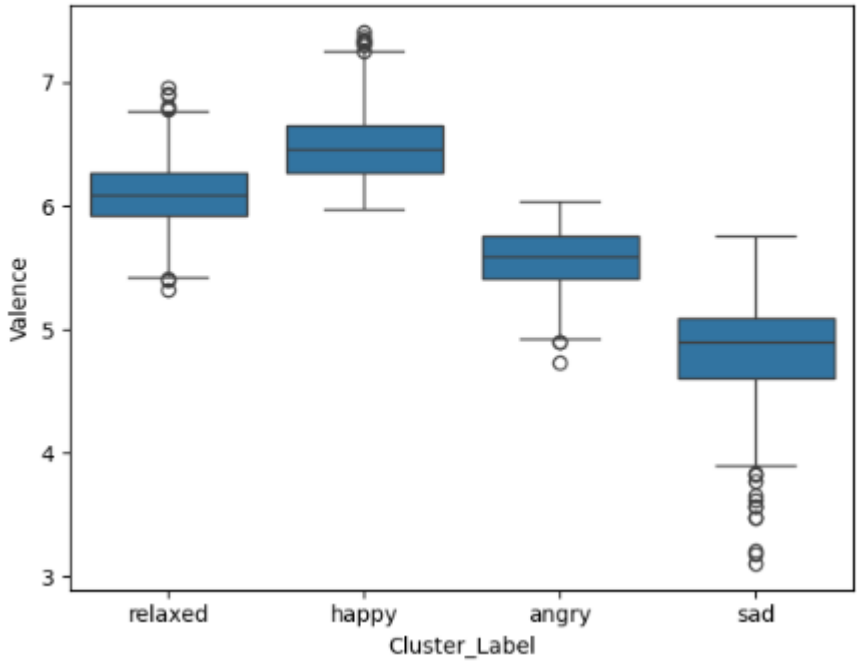
```
df['Cluster_Label'].value_counts().plot(kind='pie', autopct='%1.1f%%')
```

```
<Axes: ylabel='count'>
```



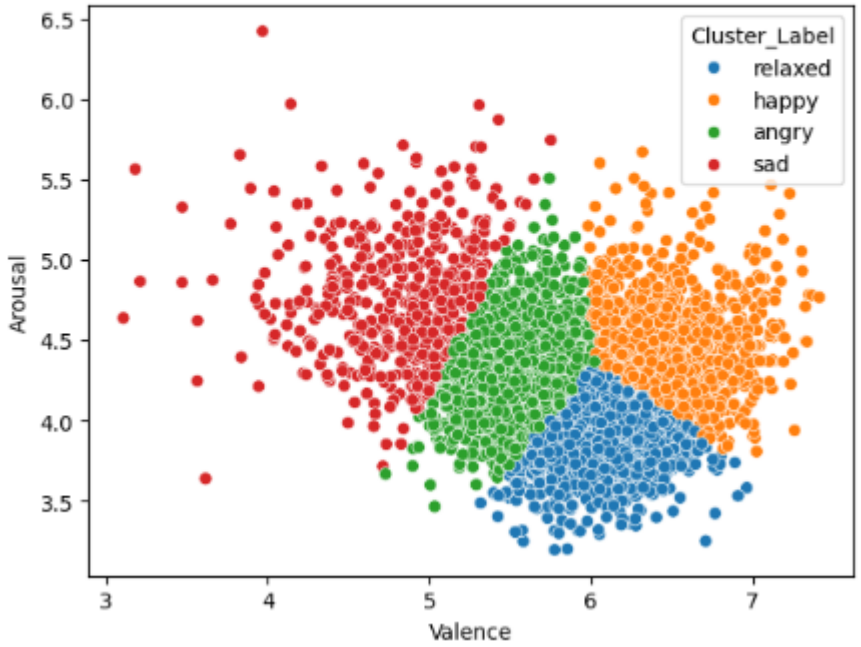

```
sns.boxplot(x='Cluster_Label', y='Valence', data=df)
```

```
<Axes: xlabel='Cluster_Label', ylabel='Valence'>
```



```
sns.scatterplot(x='Valence', y='Arousal', hue='Cluster_Label', data=df)
```

```
<Axes: xlabel='Valence', ylabel='Arousal'>
```



Paso 4. Encuentra un problema potencial en tus datos.

- Si es un problema de tipo supervisado:
  - ¿Cuál es la columna de “salida”? ¿binaria, multiclase?

`Sentiment` o `Cluster_Label` multiclase

  - ¿Está balanceado el conjunto salida?

```
df['Sentiment'].value_counts(normalize=True)
```

```
Sentiment
Relaxed    0.900720
Sad        0.056295
Happy      0.027711
Angry      0.015274
Name: proportion, dtype: float64
```

```
df['Cluster_Label'].value_counts(normalize=True
```

```
Cluster_Label
relaxed    0.314205
angry      0.309186
happy      0.267074
sad        0.109535
Name: proportion, dtype: float64
```

- ¿Cuáles parecen ser features importantes? ¿Cuáles podemos descartar?

Features importantes:

- **Valence y Arousal:** Son las medidas emocionales básicas que reflejan la carga afectiva y la activación de la canción.

- **Vader_Score (especialmente `compound`):** Complementa el análisis emocional, con un puntaje compuesto que resume sentimiento general.

- **Total_Word_Count, Unique_Word_Count, Average_Word_Length:** Características textuales que reflejan complejidad y variedad en la letra.

- **Genre y Decade:** Variables categóricas que pueden influir en el sentimiento o tipo de cluster.

- **Cluster_Label y Sentiment:** Variables de salida o para análisis supervisado.

Features que podrían descartarse o usar con precaución:

- **Singer y Song:** Muy específicos, pueden crear sobreajuste o ruido si usas un modelo general.

- **Min_Word_Length, Max_Word_Length, Sum_All_Word_Lengths, Numer_Unique_Word_Lengths:** Algunas pueden ser redundantes o menos relevantes si ya usas promedios y conteos.

- **Columnas con muchos valores nulos o inconsistentes:** Ejemplo, si `Decade` tiene muchos valores como `nan0`, es mejor limpiarlos o imputarlos.

    - ¿Estamos ante un problema dependiente del tiempo? Es decir un TimeSeries.**N**o es un problema Time Series tradicional, porque cada fila representa una canción independiente, sin un orden temporal que se analice secuencialmente.
    - Si fuera un problema de Visión Artificial: ¿Tenemos suficientes muestras de cada clase y variedad, para poder hacer generalizar un modelo de Machine Learning? No

- La distribución, tendencia de las variables varía en el tiempo? NO
- ¿Hay algún problema notable con la calidad de los datos?

Valores nulos o incorrectos: como `nan0` en `Decade`.

Outliers: valores extremos en Valence, Arousal o conteos de palabras.

Desbalance en la variable objetivo: especialmente en `Sentiment`.

Duplicados o entradas repetidas: verificar y limpiar.

Datos textuales con letras muy cortas o vacías: que pueden sesgar métricas.

- ¿Existe alguna relación sorprendente entre las variables?

- **Baja correlación entre Valence y Vader_compound:** Puede indicar que diferentes métodos de análisis emocional no capturan lo mismo.

- **Posible correlación entre género y emociones:** Algunos géneros pueden tender a ciertas emociones.

- **Relación entre longitud de letras y emociones:** canciones más largas podrían tener sentimientos más complejos o variados.

Conclusión
¿Qué podemos aprender de este análisis?

Datos de entrada y salida claros:
El dataset contiene múltiples características relevantes para el análisis emocional de canciones, como medidas de valencia, activación (arousal), puntajes de sentimiento y características textuales. Las variables de salida para clasificación son multiclase (`Sentiment` y `Cluster_Label`).

Balance de clases:
`Sentiment` está muy desbalanceado, dominado por la clase "Relaxed", lo que puede dificultar entrenar modelos supervisados efectivos con esta variable. En cambio, `Cluster_Label` presenta una distribución mucho más balanceada, por lo que es una mejor opción como variable objetivo para tareas de clasificación.

Características importantes:
Las variables emocionales (`Valence`, `Arousal`), junto con características textuales y categóricas (`Genre`, `Decade`), son las más informativas. Variables muy específicas como el nombre del artista o canción pueden no aportar valor predictivo y pueden descartarse.

No es un problema de series temporales tradicional:
Aunque hay un campo `Decade` para agrupar canciones por época, no hay un orden temporal secuencial que influya en la predicción directa. Sin embargo, analizar tendencias por décadas puede aportar insights históricos.

Calidad de los datos:
Se detectaron valores nulos, posibles outliers y registros duplicados, que deben ser tratados para asegurar un análisis y modelado robusto. Además, hay que cuidar la consistencia de datos categóricos (por ejemplo, corregir valores erróneos en `Decade`).

Relaciones y correlaciones:
Se observa baja correlación entre algunas medidas emocionales (como `Valence` y el puntaje `compound` de Vader), lo que indica que diferentes métodos aportan perspectivas complementarias del sentimiento.

Implicaciones para el modelado:

- Para un modelo supervisado, `Cluster_Label` es la mejor variable objetivo por balance y representatividad.

- Es recomendable realizar un preprocesamiento exhaustivo: manejo de nulos, codificación de variables categóricas y tratamiento de outliers.

- Posibles diferencias en emociones por género o década podrían mejorar la interpretación y selección de características.