Informe de Fundamentos de la Ciencias de Datos

“COVERS DE UNA DÉCADA ICÓNICA DEL SIGLO PASADO (Amazon Music)”

Integrantes:

Videla, Braian.

Perna, Ignacio.

Introducción

El objetivo principal de este análisis es realizar un análisis exhaustivo del conjunto de datos, caracterizando sus atributos, valores atípicos y valores nulos, así como la realización de una limpieza de datos.

(FALTA)

…

En el contexto del análisis de datos, usaremos la biblioteca pandas, una herramienta fundamental en Python que permite la manipulación y análisis eficiente de datos. Pandas tiene diferentes funcionalidades que nos ayudarán en dicho análisis exhaustivo.

Lectura y Escritura de Datos :

- `pandas` permite importar datos desde diversas fuentes, como archivos CSV, Excel, SQL, JSON, y más, mediante funciones como `read_csv()`, `read_excel()`, y `read_sql()`.
- También facilita la exportación de datos a estos formatos mediante funciones como `to_csv()`, `to_excel()`, y `to_sql()`.

Manipulación de datos :

- Filtrado y Selección : Posibilidad de seleccionar filas y columnas utilizando etiquetas o condiciones lógicas.
- Indexación : Permite establecer índices personalizados para facilitar la selección y alineación de datos.
- Agrupación : La función `groupby()`permite agrupar datos y realizar operaciones agregadas, como sumas y promedios, sobre grupos específicos.

Manejo de Datos Faltantes :

- `pandas`Ofrece métodos para identificar y manejar valores nulos, como `isnull()`, `dropna()`, y `fillna()`, que permiten limpiar el conjunto de datos de manera eficiente.

Análisis estadístico :

- Métodos integrados para calcular estadísticas descriptivas (media, mediana, desviación estándar) y realizar análisis exploratorio básico.

Visualización :

- Aunque `pandas`no se enfoca en la visualización, se integra bien con bibliotecas como `matplotlib`y `seaborn`para crear gráficos a partir de los datos, facilitando la exploración visual.

Para empezar, podemos ver un pequeño pantallazo de lo que se tratará nuestro dataset. Con una función de pandas “pd.read_cesv(“Covers.csv”)” podremos observar los primeros y últimos 5 registros con sus respectivas columnas.

Observamos entonces, que tenemos 980 registros con 17 variables a analizar.

Tipo de Datos

Con el objetivo de adentrarnos en el análisis exhaustivo de los datos, necesitamos entender cuáles son las características estructurales de nuestro dataset.

Del método info de pandas, sabemos que por lo menos a primera vista, no tenemos valores nulos en nuestro conjunto, y que tenemos un 11,76% de datos que son Cualitativos, Categóricos y Nominales porque no tienen un orden definido: Track y Artist.

Vemos que la variable Duration está catalogada como lo que podría ser un dato cualitativo, pero la analizaremos más adelante. Por ahora lo tomaremos como en realidad debería ser: un dato cuantitativo.

Luego, tenemos que el 29,41% son datos Cuantitativos, Numéricos y Discretos que representan cantidades o medidas que pueden contarse de manera específica y no toman valores intermedios: Time Signature, Key, Mode, Popularity y Year.

Y por último, tenemos un 58,82 % de datos Cuantitativos, Numéricos y Continuos que representan cantidades o medidas que no pueden contarse fácilmente ya que pueden tomar valores como fracciones o decimales: Danceability, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence y Tempo.

Descripción de las variables

Track: el título de la canción.

Artist: el intérprete o grupo que grabó la canción.

Duration: la duración de la canción, medida en minutos y segundos.

Time Signature: la métrica musical de la canción, indica el número de pulsaciones por compás.

Danceability: una medida de qué tan adecuada es una pista para bailar, basada en el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general.

Energy: una medusa de intensidad y actividad en la canción, donde los valores más altos indican una pista más enérgica.

Key: la tonalidad musical en la que está compuesta la canción, representada por un número entero.

Loudness: el volumen promedio de la canción, medido en decibelios(dB).

Mode: la modalidad de la pista, indica si la canción está en tono mayor o menor.

Speechiness: una medida de la presencia de palabras habladas en una pista, valores más altos indican cualidades más parecidas al habla.

Acousticness: una medida de la calidad acústica de la pista, valores más altos indican una probabilidad de ser acústica.

Instrumentaleness: una medida que indica presencia de voces, valores más altos representan pistas más instrumentales.

Liveness: una medida de la probabilidad de que la pista se haya interpretado en vivo, valores más altos indican más ruido de audiencia.

Valence: una medida de la positividad musical de la pista, valores más altos indican música más positiva o alegre.

Tempo: la velocidad o ritmo de la pista, medida en pulsaciones por minuto(BPM).

Popularity: una puntuación que refleja la popularidad de la pista, generalmente basada en los recuentos de transmisiones y otras métricas.

Year: el año en el que se lanzó la canción.

Planteamiento inicial de hipótesis

La medida que indica el número de pulsaciones por compás debería estar directamente relacionada con el ritmo de la pista medida en pulsaciones por minuto que a su vez quizás está relacionado con la intensidad de la canción. Suponemos que a mayor pulsaciones por compás, mayor deberían ser las pulsaciones por minuto, y mayor actividad: TIME SIGNATURE - TEMPO - ENERGY.

Si una canción es muy adecuada para bailarla, tenderá a ser más alegre: DANCEABILITY - VALENCE.

Una canción con volumen muy elevado, suponemos que debería ser más enérgica: LOUDNESS - ENERGY.

Si una pista tiene muchas probabilidades de haber sido interpretadas en vivo, entonces es muy probable que tenga un volumen elevado: LIVENESS - LOUDNESS.

A su vez, si una canción tiene popularidad alta, es muy probable que se haya interpretado en vivo al tener mayor ruido de audiencia o viceversa: POPULARITY - LIVENESS.

Las canciones más adecuadas para bailar podrían ser más populares que otras teniendo en cuenta la época: DANCEABILITY - POPULARITY.

Una pista muy instrumental debería tener valores directamente opuestos a la cantidad de palabras habladas y viceversa: INSTRUMENTALNESS - SPEECHINESS.

Primera descripción estadística de las variables

Usando el método “describe()” de Pandas, podemos observar en nuestras variables, detalles estadísticos como:

Count : la cantidad de valores no nulos en la columna. En este caso, todos tienen 980 valores.

Mean (Media) : el promedio de los valores en la columna.

Std (Desviación estándar) : mide la dispersión de los datos alrededor de la media. Una mayor desviación indica mayor variabilidad en los datos.

Min (Valor mínimo) : el valor más bajo en la columna.

25% (Primer cuartil) : el valor por debajo del cual se encuentra el 25% de los datos, también conocido como el percentil 25.

50% (Mediana o segundo cuartil) : el valor central que separa la mitad superior e inferior de los datos.

75% (Tercer cuartil) : el valor por debajo del cual se encuentra el 75% de los datos, también conocido como el percentil 75.

Max (Valor máximo) : el valor más alto en la columna.

En dicho método podremos observar que las variables Track y Artist, al tener datos cualitativos nominales, la librería Pandas las elimina del análisis descriptivo porque se encarga de tomar datos estadísticos, lo que es imposible con este tipo de datos.

Time Signature: Sus valores se encuentran en un rango de 1 a 5 con un desvío estándar cercano a cero, por lo que podríamos estar observando una distribución normal porque los datos se encuentran cerca de la media.

Danceability: Sus valores se encuentran en un rango de 0 a 1, con un desvío estándar cercano a cero, obteniendo la misma conclusión anterior.

Energy: Sus valores se encuentran en un rango de 0 a 1, con un desvío estándar cercano a cero. Vemos también que el valor mínimo de ésta variable es de una magnitud 116 veces menor que el promedio, indicando que tenemos por lo menos un outlier.

KEY: Sus valores están en un rango de 0 a 12, con un desvío estándar medianamente elevado, lo que nos indica que los valores de esta variable están muy dispersos alrededor de la media, es decir, que existe una amplia variabilidad en los tonos musicales en el conjunto de canciones analizadas.

Loudness: Como es una medida en decibelios, su rango es de -100 a 0. En este caso, vemos que el 75% de los datos se encuentran más cercanos al límite superior, por lo que tenemos la mayoría de canciones con un nivel de ruido elevado.

Mode: Al ser una variable discreta, no tiene sentido analizar promedios, desvíos o cuartiles. Sus posibles valores son 0 o 1 indicando si la canción está en tono mayor o menor.

Speechiness: El rango se mueve entre 0 y 1. Notamos que la media de 0,059923 indica que la presencia de palabras habladas en una pista por lo general es baja. El primer y tercer cuartil refuerzan esta idea con datos similares(0.031 y 0.038). Con desvío estándar cercano a 0 podemos suponer distribución normal. Notamos también la presencia de algún outlier al tener un valor máximo 12 veces más grande que la media.

Acousticness: El rango se mueve entre 0 y 1. La media de 0.33 indica que, en promedio, hay una presencia acústica moderada en las pistas. Se puede ver una concentración significativa de pistas con niveles bajos de elasticidad. Y teniendo un mínimo con valor 15 mil veces menor que la media, habrá, outliers por el lado de la pista poco acústica.

Instrumentales: El rango va de 0 a 1. Vemos que la mayor cantidad de calores se concentran en valores muy bajos y en su mayoría son ceros o cercanos al mismo. La disparidad entre el valor máximo y el 75% nos indica la presencia de outliers.

Liveness: El rango va de 0 a 1. Tendremos la mayoría de los valores cercanos a 0, por lo que habrá poco ruido de audiencia, es decir, menor probabilidad de que se haya interpretado en vivo.

Valence: El rango está entre 0 y 1. Podemos ver por el valor de la media y la relativa paridad entre cuartiles, que parece ser una distribución normal. Aunque tenemos un claro outlier por el lado de una pista musical completamente triste con valor mínimo 62 mil veces menor que la media.

Tempo: Su rango se encuentra entre 50 y 216 pulsaciones por minuto. Una desviación estándar elevada nos indica que nuestras pistas varían bastante en ésta característica.

Popularity: Su rango se ubica entre 0 y 100 puntos de popularidad. Vemos que en ésta década la popularidad variaba bastante, la mayoría tiende a ser medianamente reconocida, pero con algunos casos atípicos de gran popularidad y de baja popularidad.

Year: El rango de los años está entre 1970 y 1979.

Limpieza del conjunto de datos

Registros repetidos

Queremos trabajar con datos limpios y claros para analizar nuestro conjunto, por lo que deberíamos buscar registros repetidos, para no incluirlos dentro de nuestro trabajo.

En primer lugar tomamos como clave de los registros la dupla Artista-Canción y las agrupamos con la función Group By de Pandas para ver si conduce a algún problema de repetición de registros.

Analizando los resultados obtuvimos 980 filas, que es la cantidad inicial de filas que teníamos previamente, por lo que podemos concluir que NO tenemos registros repetidos en nuestro Dataset.

Valores nulos

Para seguir puliendo, vamos a buscar valores nulos para saber cómo tratarlos próximamente: eliminarlos, transformarlos o analizar la razón de la nulidad de los mismos.

Ejecutando una sumatoria de valores con el método isnull() de Pandas que marca la presencia de valores “NULL” observamos que esta suma de nulos en todos los registros nos devuelve 0, por lo que podemos afirmar la ausencia de valores nulos en todo el conjunto.

Ejecutamos también una sumatoria de valores con el método isna() de Pandas que marca la presencia de valores “NaN” (Not a Number”) y en todas las variables obtuvimos una suma de 0, por lo que no tenemos valores NaN en nuestro conjunto.

Para terminar de asegurarnos, como anteriormente cuando describimos estadísticamente las variables no había ninguna de las cuantitativas que presentara

valores extraños que pueda considerarse nulos, vamos a agrupar los valores solamente por cualitativas y encontrar posibles nulos o errores de carga.

Sabiendo que no tenemos valores NULL o NaN, como la librería Pandas no nos muestra todos los datos, los imprimimos en un for para comprobar que no tenemos uno o más datos cualitativos “Unknown” o cualquier nombre extraño. Dicha impresión estará dada por el artista y la cantidad de canciones que éste tiene. Gracias al método items() que convierte la serie devuelta por el método Value_counts() en una vista de pares (índice, valor).

Luego comprobamos los datos de los Artistas “10cc”,”GQ” y “M” solo por tener nombres extraños y asegurarnos de que no sean muestras basura y resultaron tener datos válidos por lo que no habría necesidad de eliminar ni transformar ninguna variable.

Pudimos comprobarlo seleccionando la columna “Artist” del dataset y definiendo una lista de los nombres de artistas que queremos buscar en el dataset. Usando la función isin() solo mostrará los valores que le pasemos, en este caso, la lista de artistas.

Conversión de tipos

Al principio habíamos dicho que la variable DURATION la íbamos a tratar como un dato cuantitativo porque es la medición de la duración de una canción. Así que, para poder incluirla en el análisis y tratar con ella, vamos a convertir en un dato numérico aplicando un reemplazo de “:” (razón por la que Pandas la tomaba como un String) por un “.” y castear a un tipo float.

Para completar el análisis descriptivo anterior le aplicamos el método describe().

Es una variable que en esta (MUESTRA O ESPACIO MUESTRAL)???? tendrá un rango de 0 a 30 minutos por canción. Con un promedio de 3.6 minutos y un máximo de 26 minutos.

Por otro lado, notamos que la variable LOUDNESS que mide el volumen promedio de la canción, tiene valores negativos y más adelante en el análisis de

correlación vamos a ver que todas las demás variables incluidas tienen un rango de 0 a 1. Como siempre es conveniente trabajar con rangos iguales en ambos lados, vamos a invertir el rango LOUDNESS que actualmente tiene un rango teórico de -100 a 0, a un intervalo entre 0 y 1.

Para poder hacerlo, tendremos que aplicarle a la columna LOUDNESS una función de valor absoluto donde convertimos todo su rango negativo en uno positivo entre 0 y 1 dividiéndolos por 100.

#ACLARACIÓN: Siempre tener en cuenta trabajar con copias al realizar cambios o transformaciones por si en algún momento requiera trabajar con el Dataset original.

Matriz de Correlación

Llegamos al punto donde ya tenemos todos los datos básicos y necesarios de cada variable particular, por lo que ahora queremos estudiar el comportamiento bivariado de todas las variables cuantitativas de nuestro conjunto. Para eso, vamos a utilizar la **Matriz de Correlación**, pero antes debemos quitar de nuestro DataFrame los datos **cualitativos** porque no existe una relación numérica entre dos Strings.

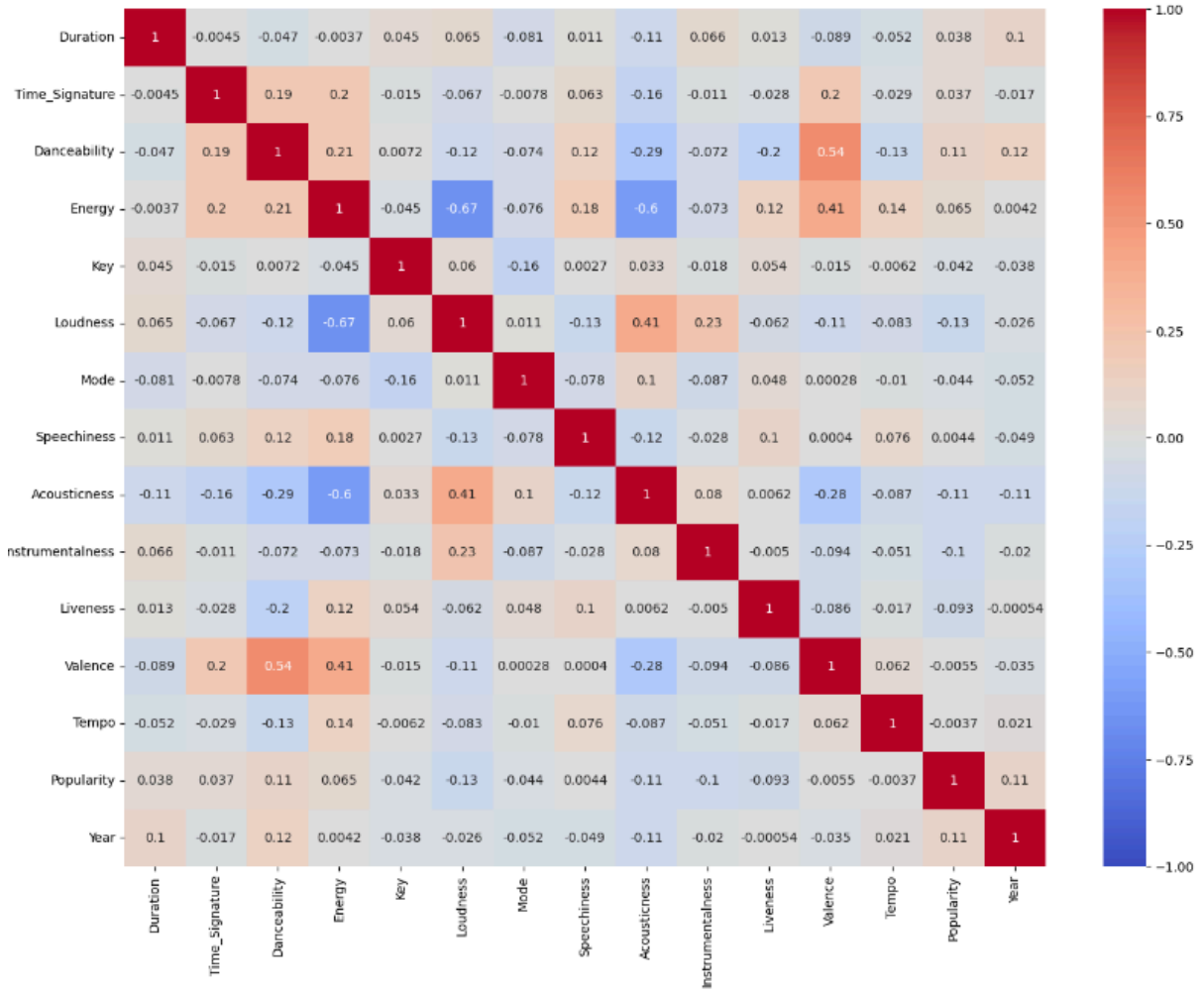
Para esto eliminamos las columnas Track y Artist y luego lanzamos la matriz de correlación entre todas las variables con el método Corr() de Pandas .

Al ver la matriz de correlaciones, podemos notar que en la diagonal se encuentra solo “1”, esto se debe a que una variable se relaciona completamente con sí misma.

	Duration	Time_Signature	Danceability	Energy	Key	Loudness	Mode
Duration	1.000000	-0.004463	-0.047164	-0.003672	0.045059	0.064686	-0.080998
Time_Signature	-0.004463	1.000000	0.187562	0.204889	-0.015184	-0.067192	-0.007765
Danceability	-0.047164	0.187562	1.000000	0.214106	0.007224	-0.117018	-0.074487
Energy	-0.003672	0.204889	0.214106	1.000000	-0.045086	-0.670980	-0.076433
Key	0.045059	-0.015184	0.007224	-0.045086	1.000000	0.059878	-0.164355
Loudness	0.064686	-0.067192	-0.117018	-0.670980	0.059878	1.000000	0.011323
Mode	-0.080998	-0.007765	-0.074487	-0.076433	-0.164355	0.011323	1.000000
Speechiness	0.010918	0.062508	0.122851	0.183647	0.002664	-0.132885	-0.077697
Acousticness	-0.106135	-0.162942	-0.289435	-0.602552	0.032761	0.406708	0.104428
Instrumentalness	0.066365	-0.010567	-0.072341	-0.072815	-0.017979	0.232382	-0.086634
Liveness	0.012645	-0.027574	-0.203165	0.124811	0.054373	-0.062452	0.047852
Valence	-0.089477	0.196694	0.543124	0.407548	-0.015310	-0.107765	0.000277
Tempo	-0.051571	-0.028967	-0.126015	0.136926	-0.006183	-0.082913	-0.010323
Popularity	0.038248	0.037160	0.110817	0.065086	-0.042028	-0.129457	-0.043979
Year	0.101437	-0.017081	0.120210	0.004188	-0.038449	-0.025810	-0.051700

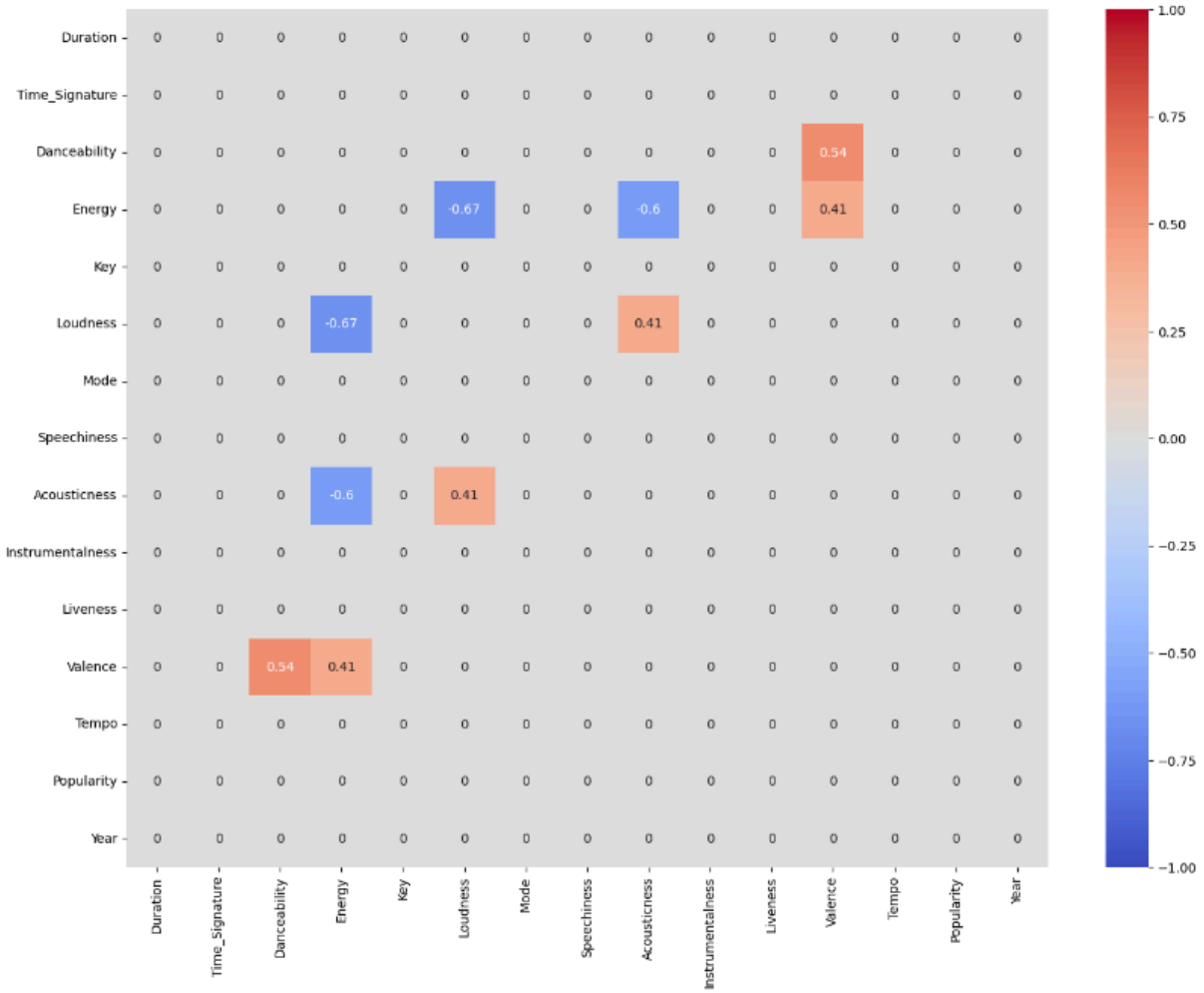
Para una mejora en cuanto a la visualización podemos representar esta matriz gráficamente y nos ayudará a identificar las correlaciones elevadas o bajas, asignando colores a cada correlación. Esto lo haremos con la característica de **Mapa de Calor** o **Heatmap** de la librería “**Seaborn**” montada sobre “**Matplotlib**”.

Este Heatmap nos pide una cota mínima, una cota máxima y un valor central, donde tonalidades de color ROJO representan valores cercanos a 1 y tonalidades de color AZUL representan valores cercanos a -1.



Vamos a pasar en limpio esta matriz quitándole información difusa como la correlación entre mismas variables. Y establecer una condición para mostrar solo aquellos coeficientes que su valor sea digno de estudio. En este caso, basándonos en el estándar aceptado actualmente, no tenemos ni un solo coeficiente que llegue al valor absoluto de 0.7, por lo que vamos a establecer un límite en 0.4 y estudiar el comportamiento entre aquellas variables que lo cumplan. También removemos los valores de la diagonal(utilizando la librería Numpy) ya que no aportan información para la correlación bivariada.

Igualmente, se estudiará más adelante el motivo de la poca o casi nula relación entre la mayoría de las variables.



Luego de analizar este último gráfico decidimos estudiar la correlación de las siguientes variables:

DANCEABILITY - VALENCE (coeficiente= 0.54)

ENERGY - LOUDNESS (coeficiente= 0.67)

ENERGY - ACOUSTICNESS (coeficiente= 0.6)

ENERGY - VALENCE (coeficiente= 0.41)

LOUDNESS - ACOUSTICNESS (coeficiente= 0.41)

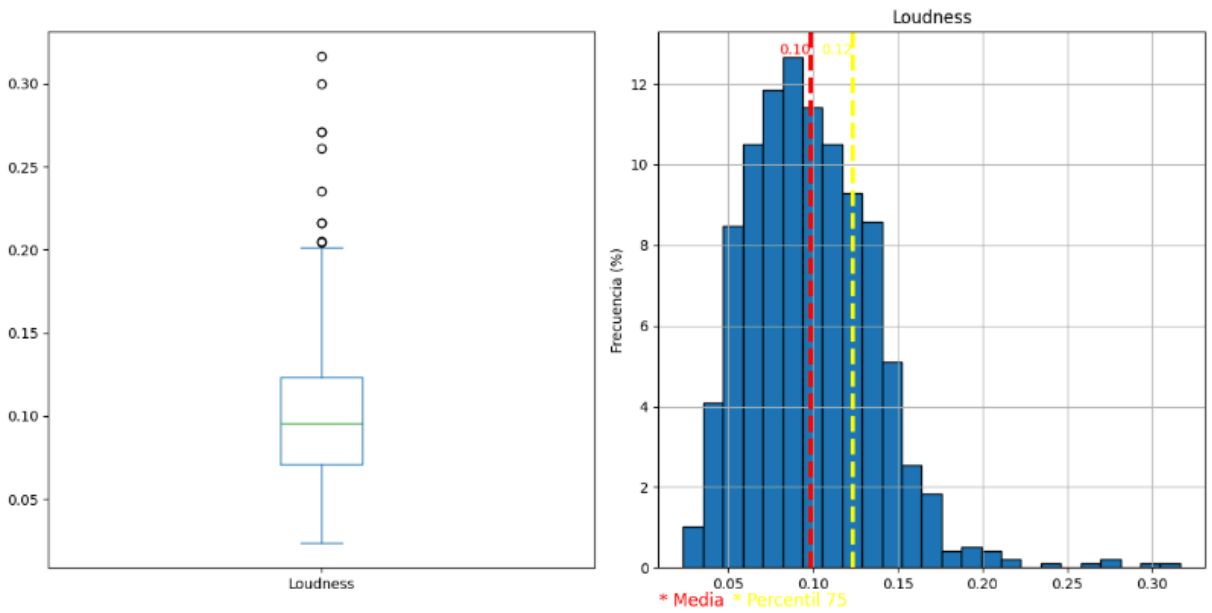
Estudio de correlación de cada variable

Antes de hacer el **análisis bivariado**, vamos a observar en profundidad individualmente cada variable para comprenderlas bien antes de compararlas con otra.

Una buena y correcta forma de estudiar cada variable individualmente es con un **BoxPlot** o un **gráfico de barras.** Ambos tipos de gráficos son muy valiosos en análisis exploratorio, ya que permiten obtener una visión clara y rápida de cómo se comporta una variable, facilitando la toma de decisiones en etapas posteriores del análisis. En este caso, usaremos ambas.

Loudness

Podemos observar gracias a la **descripción estadística de las variables** que vamos a tener que aplicar alguna especie de transformación de escalado a ésta variable porque es la única dentro de nuestras tuplas de correlación mayor a 0,4 que no se mueve en un rango entre 0 y 1.



Procedemos a imprimir el **coeficiente de asimetría** y la **Kurtosis,** dos medidas estadísticas que proporcionan información adicional sobre la distribución de la variable.

Coeficiente de asimetría: 0.94

Kurtosis: 2.67

Observamos gráficamente que la distribución de **Loudness** con un **coeficiente de asimetría** positivo tiene un sesgo a derecha, donde podemos ver que hay muchos valores extremos repartidos por derecha de la media pero con poca frecuencia de valores altos. Con una **Kurtosis** elevada sabemos que ésta variable tiene una cola pesada significando la presencia de valores atípicos.

kurtosis < 3 cola ligera

kurtosis > 3 cola pesada

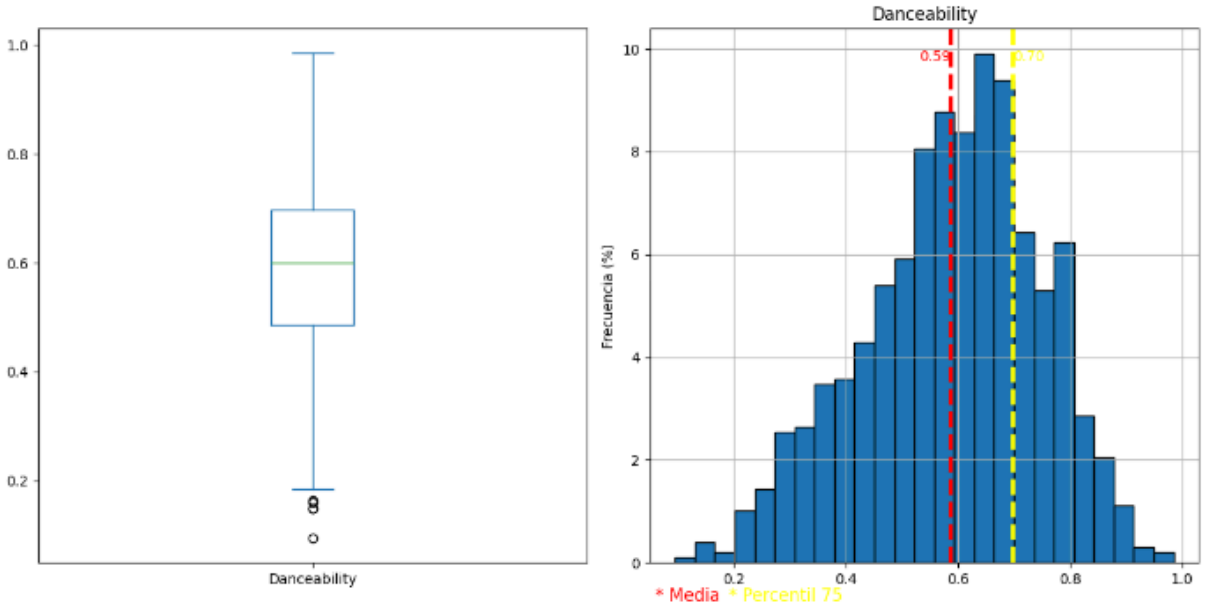
kurtosis = 3 distribución normal

coeficiente de asimetría < 0 asimétrica hacia la izquierda

coeficiente de asimetría > 0 asimétrica hacia la derecha

coeficiente de asimetría = 0 distribución normal

Danceability



Coeficiente de asimetría: -0.34

Kurtosis: -0.29

Lo que vemos tanto en el boxplot como en el histograma de la variable Dancebility es que habíamos supuesto en la descripción estadística de las variables, se trata de una distribución normal con un ligero sesgo a la izquierda

```
Coeficiente de asimetría: -0.34

Kurtosis: -0.29
```