NOMBRE	MATERIA	CI	NUMERO DE PREGUNTA
Joel Modesto Anara Michua	Inteligencia Artificial inf-354	10911944	2

2. Seleccióne un datatset tabular de al menos 1000 columnas, 14 filas. Si elige imágenes igualmente puede convertir la imagen en datos tabulares de NxM. De esta selección indique cual es la clase o si no tiene.

RESPUESTA: Se selecciono el siguiente dataset llamado '*Conjunto de datos de un millón de canciones + Spotify*' de donde tomamos 14 columnas y 1100 filas entre los cuales la clase denota la columna **GENERO** ya que es la columna dependiente de todas las demas este data set puede encontrarlo en el siguiente enlace Million Song Dataset + Spotify + Last.fm.

a. Sin el uso de librerías en Python programe el percentil y cuartil de cada columna. Que distribución se puede aplicar en su caso normal, Bernoulli, gaussiana, poisson, otros. Indique la razón de su uso graficando con matplotlib.

CODIGO DE LOS CUARTILES Y PERCENTILES

```
import pandas as pd
import matplotlib.pyplot as plt
#indice i=(p/100)*n donde n es el numero de observaciones
def percentil(p,n,v):
     i=(p/100)*(n-1)
     if(i%1!=0):
          i=int(i)+1 #ENVES DE LA INTERPOLACION
     else:
         i=int(i)

     percentiles=i
     print("n: ",n,"p: ",p)
     if(n%2==0 or p!=50 ):
      perce=(v[percentiles]+v[percentiles-1])/2
     else:
      perce=v[percentiles]
     return perce
def cuartiles(p,n,v):
   if(p==1):
     return percentil(25,n,v)
   if(p==2):
     return percentil(50,n,v)
   if(p==3):
      return percentil(75,n,v)


print("------------ARCHIVO------")
archivo=pd.read_csv("DatasetEx.csv")
tempo=archivo["tempo"].tolist()
anio=archivo["year"].tolist()
danceabilidad=archivo["danceability"].tolist()
modo=archivo["mode"].tolist()
```

```
tempo=sorted(tempo)
anio=sorted(anio)
danceabilidad=sorted(danceabilidad)
print("------PERCENTILES------")
print("PERCENTIL 12 DE TEMPO: ",percentil(12,len(tempo),tempo))
print("PERCENTIL 50 DE AÑO: ",percentil(50,len(anio),anio))
print("PERCENTIL 80 DE DANCEABILIDAD:
",percentil(80,len(danceabilidad),danceabilidad))
print("---CUARTILES DE TEMPO----")
print("CUARTIL 1: ",cuartiles(1,len(tempo),tempo))
print("CUARTIL 2: ",cuartiles(2,len(tempo),tempo))
print("CUARTIL 3: ",cuartiles(3,len(tempo),tempo))
print("---CUARTILES DE AÑO----")
print("CUARTIL 1: ",cuartiles(1,len(anio),anio))
print("CUARTIL 2: ",cuartiles(2,len(anio),anio))
print("CUARTIL 3: ",cuartiles(3,len(anio),anio))
print("---CUARTILES DE DANCEABILIDAD----")
print("CUARTIL 1: ",cuartiles(1,len(danceabilidad),danceabilidad))
print("CUARTIL 2: ",cuartiles(2,len(danceabilidad),danceabilidad))
print("CUARTIL 3: ",cuartiles(3,len(danceabilidad),danceabilidad))
```

Corrida:

```
------PERCENTILES------
n:  1100 p:  12
PERCENTIL 12 DE TEMPO:  91.46199999999999
n:  1100 p:  50
PERCENTIL 50 DE AÑO:  2006.0
n:  1100 p:  80
PERCENTIL 80 DE DANCEABILIDAD:  0.6445000000000001
---CUARTILES DE TEMPO----
n:  1100 p:  25
CUARTIL 1:  103.9015
n:  1100 p:  50
CUARTIL 2:  123.3625
n:  1100 p:  75
CUARTIL 3:  144.72699999999998
---CUARTILES DE AÑO----
n:  1100 p:  25
CUARTIL 1:  2001.0
n:  1100 p:  50
CUARTIL 2:  2006.0
n:  1100 p:  75
CUARTIL 3:  2009.0
---CUARTILES DE DANCEABILIDAD----
n:  1100 p:  25
CUARTIL 2:  2006.0
n:  1100 p:  75
CUARTIL 3:  2009.0
---CUARTILES DE DANCEABILIDAD----
n:  1100 p:  25
---CUARTILES DE DANCEABILIDAD----
```

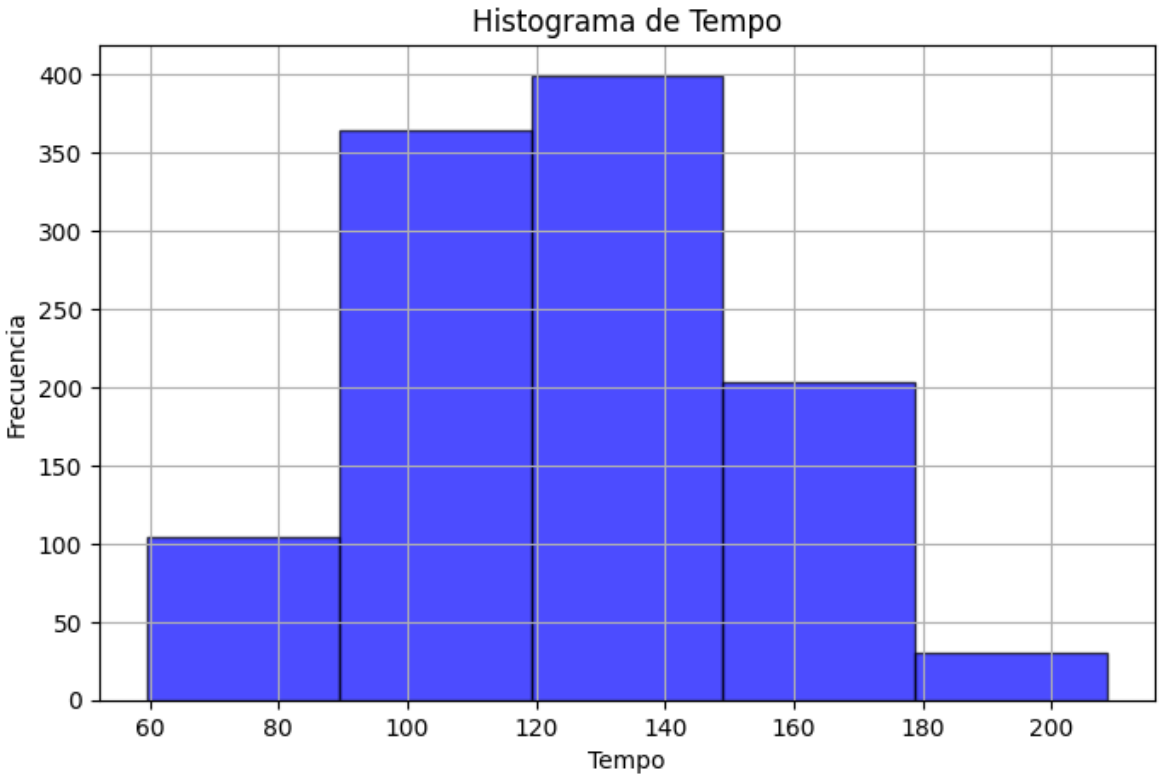
```
---CUARTILES DE DANCEABILIDAD----
n:  1100 p:  25
n:  1100 p:  25
CUARTIL 1:  0.41
n:  1100 p:  50
CUARTIL 2:  0.515
n:  1100 p:  75
n:  1100 p:  50
CUARTIL 2:  0.515
n:  1100 p:  75
CUARTIL 3:  0.622
```

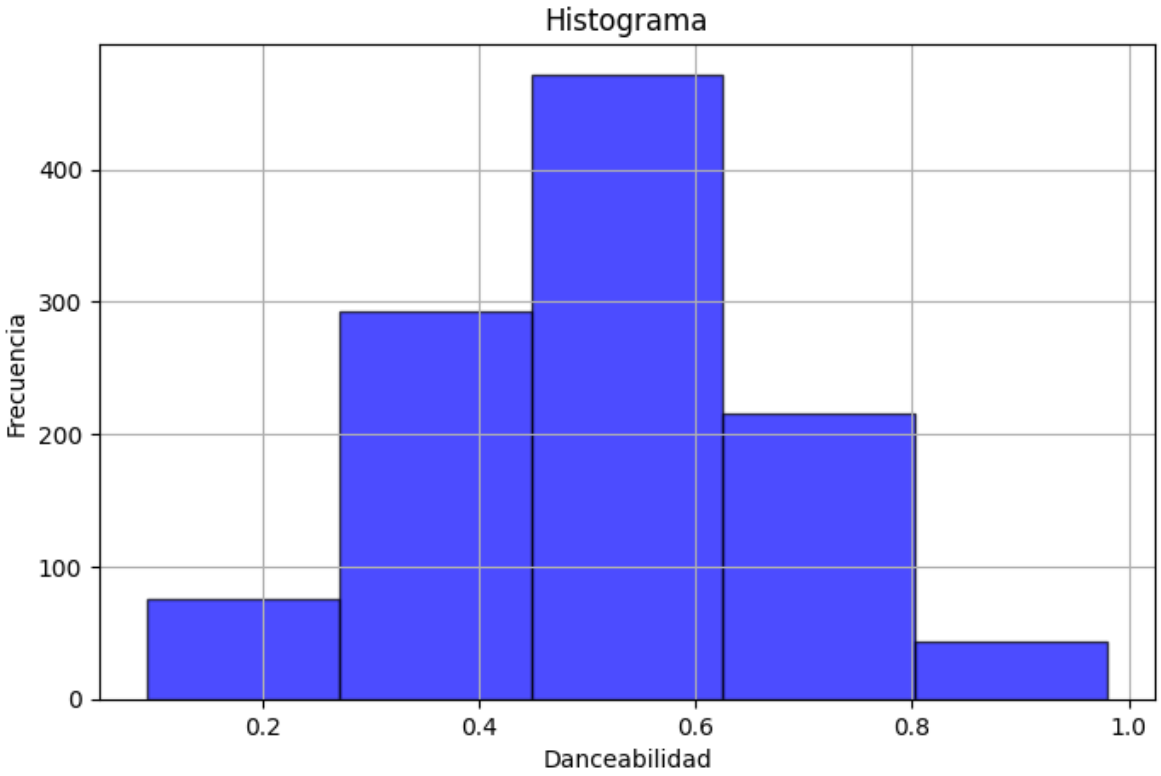
CODIGO DE LAS GRAFICAS

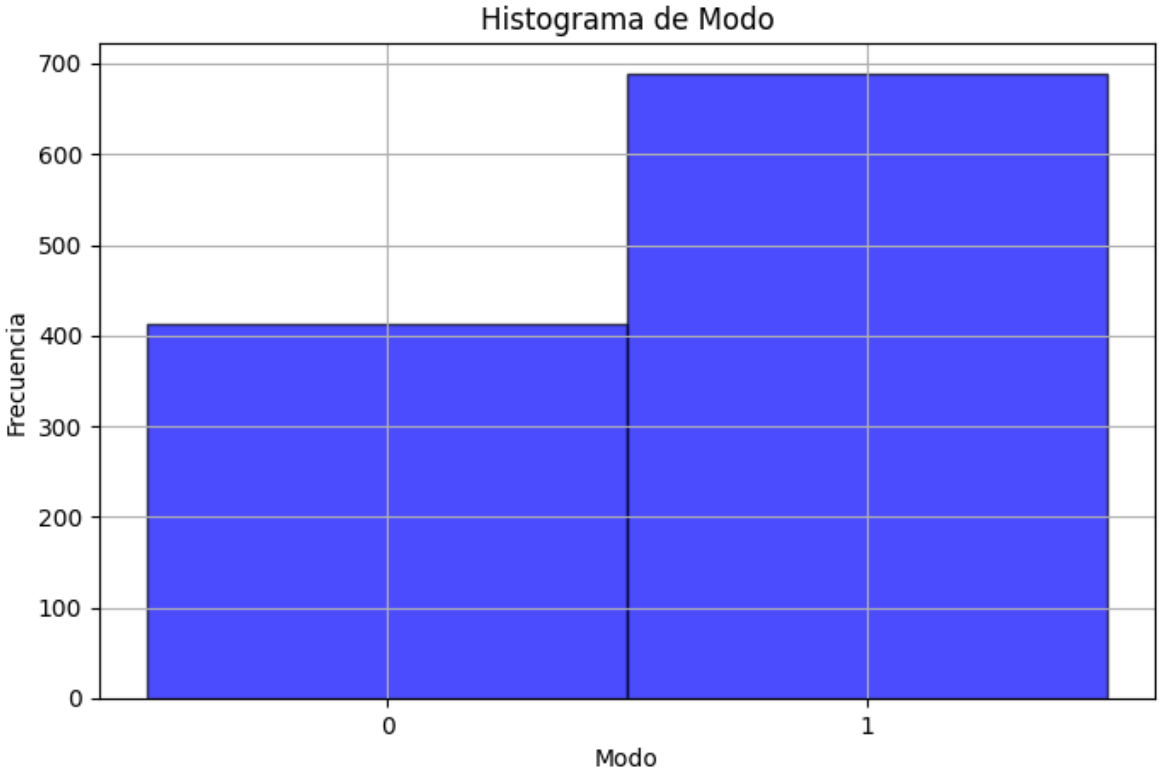
```
    print("---grafico de histograma de tempo------")
  plt.figure(figsize=(8, 5))
  plt.hist(tempo, bins=5, alpha=0.7, color='blue', edgecolor='black')
  plt.title('Histograma de Tempo')
```

```
plt.xlabel('Tempo')
plt.ylabel('Frecuencia')
plt.grid()
plt.show()
print("---grafico de histograma de Danceabilidad---")
plt.figure(figsize=(8, 5))
plt.hist(danceabilidad, bins=5, alpha=0.7, color='blue', edgecolor='black')
plt.title('Histograma')
plt.xlabel('Danceabilidad')
plt.ylabel('Frecuencia')
plt.grid()
plt.show()
print("---grafico de histograma de Modo---")

plt.figure(figsize=(8, 5))
plt.hist(modo, bins=[-0.5, 0.5, 1.5], alpha=0.7, color='blue', edgecolor='black')
plt.title('Histograma de Modo')
plt.xlabel('Modo')
plt.ylabel('Frecuencia')
plt.xticks([0, 1])  # Ajusta las etiquetas del eje x
plt.grid()
plt.show()
```

GRAFICAS







EXPLICACION

Histograma de Tempo

La distribución Gamma podría ser una opción adecuada para describir el **histograma de Tempo** que presentamos, sobre todo considerando su flexibilidad por el echo de:

1. Asimetría
2. Datos que nunca son negativos
3. Ajustable a diferentes formas

Histograma de Danceabilidad

El conjunto de datos que mide la **danceabilidad** en una escala de 0 a 1, formando una campana asimétrica con más valores a la izquierda que a la derecha (es decir, con un sesgo positivo), podría seguir una distribución Beta con las siguientes caracteristicas:

1. Escala de 0 a 1
2. Asimetria por el sesgo positivo
3. Ajustable a diferentes formas

Histograma de Modo

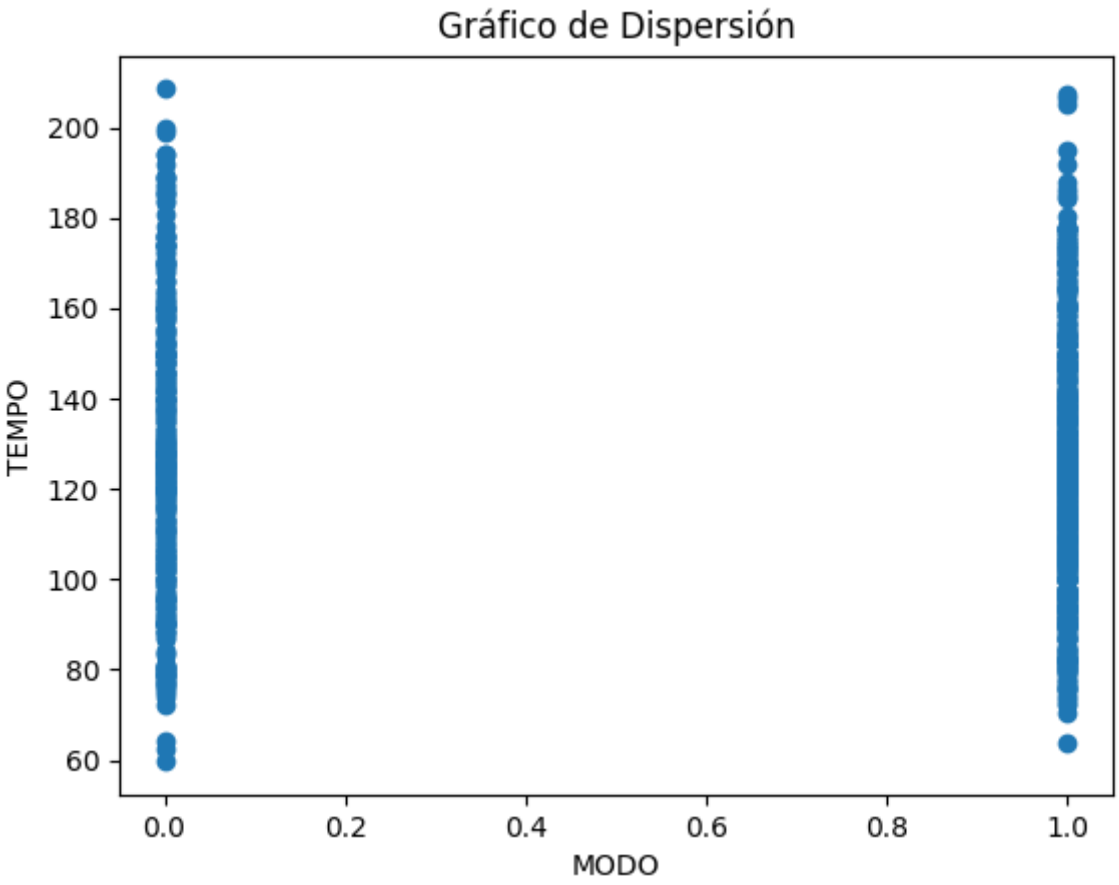
Dado que el histograma de **modo** solo se tiene valores (0 y 1), la distribución que se ajusta mejor a los datos es una distribución binomial por la siguiente caracteristica dado que existe una proporcion de p que es la probabilidad de exito que seria en este caso la cantidad de 1 y una proporcion de q que seria de fracaso que esta representada por ceros y n el numero total en el conjunto de datos.

b. De al menos tres columnas seleccionadas por usted indique que datos son relevantes de estas, grafique la misma (puede ser dispersión o mapa de calor, otros), indique al menos 4 características por columna seleccionada.

CODIGO

```
# ---- INCISO B ----
plt.scatter(auxmodo,auxdanceabilidad)

# Añadir títulos y etiquetas
plt.title("Gráfico de Dispersión")
plt.xlabel("MODO")
plt.ylabel("DANCEABILIDAD")

# Mostrar el gráfico
plt.show()


plt.scatter(auxmodo,auxtempo)

# Añadir títulos y etiquetas
plt.title("Gráfico de Dispersión")
plt.xlabel("MODO")
plt.ylabel("TEMPO")

# Mostrar el gráfico
plt.show()

plt.scatter(auxtempo,auxdanceabilidad)
```

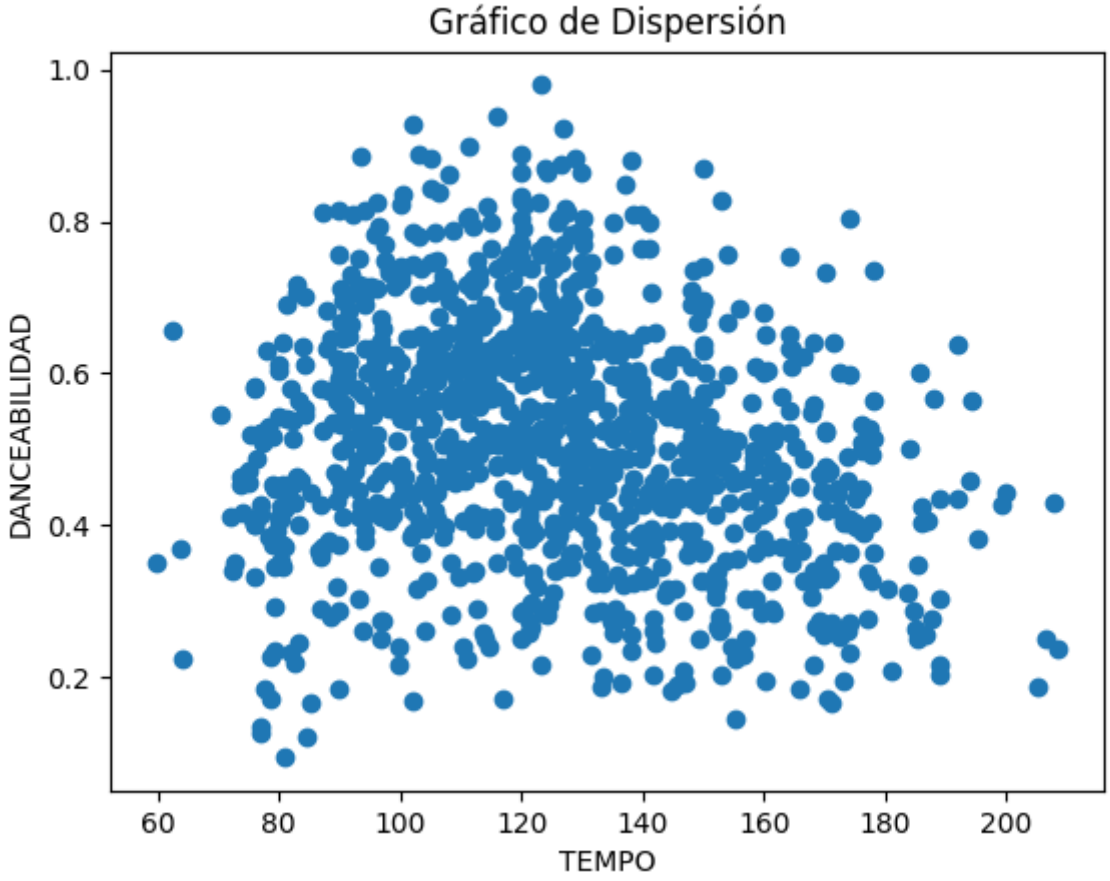
```
# Añadir títulos y etiquetas
plt.title("Gráfico de Dispersión")
plt.xlabel("TEMPO")
plt.ylabel("DANCEABILIDAD")

# Mostrar el gráfico
plt.show()
```



Caracteristicas:

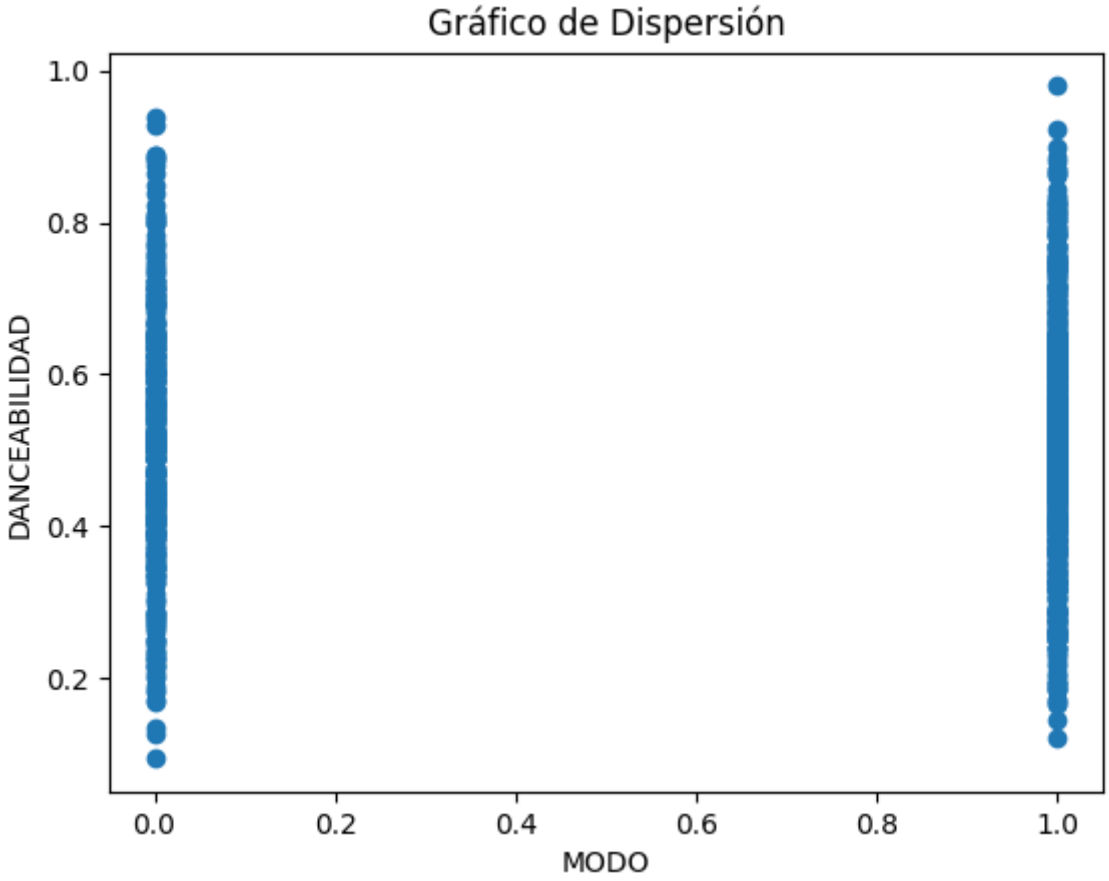
1. en la grafica observamos que hay mas datos del modo mayor que menor segun el dataset
2. observamos que tanto de tonalidad mayor o tonalidad menor son usados en distintos tempos variados para distintos generos en especifico
3. observamos que el tempo y el modo menor tiene un mayor valor que el tempo del modo mayor.
4. graficamente observamos que indistintamente del tiempo la ritmica de la cancion afectara para definir mas adelante el genero, en este caso no define el tempo y la tonalidad segun la danceabilidad

podremos elegir eso.



CARACTERISTICAS:

1. observamos que la danceabilidad en funcion al tiempo puede ser muy variado en especial hay pocas canciones que tienen un tempo alto en este caso 200 y son danzables
2. Observamos que las que mas se danzan estan en un tempo de 120 bpm aproximadamente
3. Observamos segun la grafica que la mayor cantidad de canciones tienen una danceabilidad de 0.5 aproximadamente y con distintos tiempos lo cual vemos que el tempo no es definitivo para ver si una musica es bailable o no
4. vemos un dato atipico en el tempo 80 ya que es la cancion que menos danceabilidad tiene probablemente sea una balada o una cancion que su parte ritmica no sea muy movida y variada.



CARACTERISTICAS

1. observamos que existe una cancion que es la que menos se danza y esta pertenece al modo menor y observamos que su danceabilidad es menor a los del dato mayor del modo mayor
2. vemos que el modo no es un factor que determina si una cancion es danzable o no sino que existen otros factores como el tempo que con la danceabilidad si nos daban un indicio de ciertos tiempos son danzables
3. en este caso observamos que el modo mayor posee un grado mayor de4 danceabilidad segun el dataset

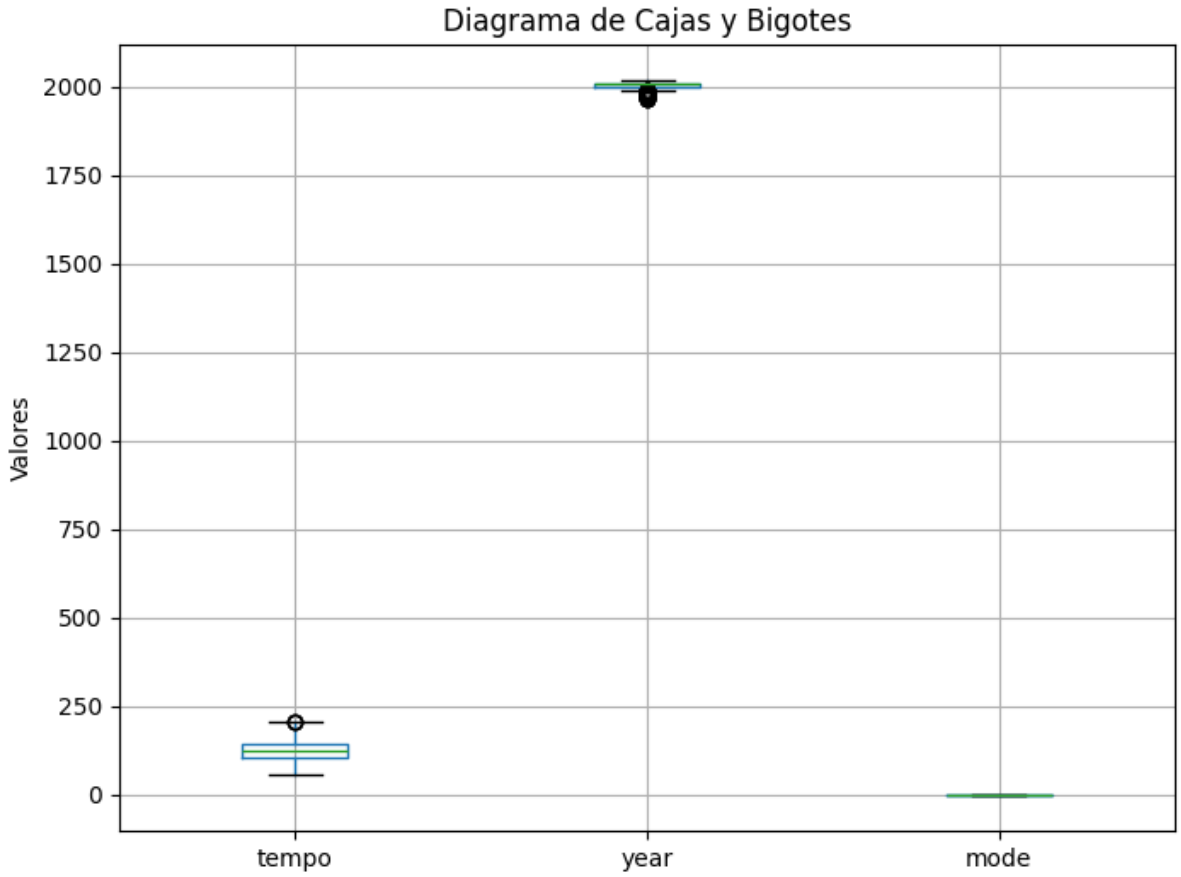
c. Obteniendo la media, mediana, moda con el uso de librerías, grafique un diagrama de cajas-bigote de al menos 3 columnas. Explique el resultado. CODIGO:

```
    #INCISO C


# Media
media_col1 = archivo["tempo"].mean()
media_col2 = archivo["year"].mean()
media_col3 = archivo["mode"].mean()

# Mediana
mediana_col1 = archivo["tempo"].median()
mediana_col2 = archivo["year"].median()
mediana_col3 = archivo["mode"].median()
```

```
# Moda

print("MEDIAS")
print("media1: ",media_col1,"media2: ",media_col2,"media3: ",media_col3)

print("MEDIANAS")
print("meidana 1",mediana_col1,"meidana 2",mediana_col2,"meidana 3",mediana_col3)

moda_col1 = archivo["tempo"].mode()[0]
moda_col2 = archivo["year"].mode()[0]
moda_col3 = archivo["mode"].mode()[0]

print("Moda Tempo: ", moda_col1)
print("Moda año: ", moda_col2)
print("Moda Modo: ", moda_col3)


# Gráfico de Cajas y Bigotes (Boxplot)
plt.figure(figsize=(8, 6))
archivo.boxplot(column=["tempo", "year", "mode"])
plt.title('Diagrama de Cajas y Bigotes')
plt.ylabel('Valores')
plt.show()
```

CORRIDA:

```
MEDIAS
media1:  125.07505363636365 media2:  2003.9972727272727 media3:  0.6254545454545455
MEDIANAS
meidana 1 123.3625 meidana 2 2006.0 meidana 3 1.0
Moda Tempo:  100.001
Moda año:  2007
Moda Modo:  1
```



EXPLICACION

Este gráfico de cajas y bigotes representa tres variables: **tempo**,**year** y **mode**. A continuación, ofrezco una explicación que integra los resultados estadísticos:

Tempo: Media: 125.08 Mediana: 123.36 Moda: 100.00 La mayoría de los valores del tempo se encuentran entre 100 y 140, lo que sugiere que la mayoría de las canciones tienen un tempo en ese rango. La media (125.08) y la mediana (123.36) están bastante cerca, indicando que los datos no están muy sesgados. Sin embargo, hay un valor muy alto (alrededor de 250) que se considera un "outlier" o valor atípico, sugiriendo que existe alguna canción que se destaca en cuanto a su tempo.

Year: Media: 2004.00 Mediana: 2006.00 Moda: 2007.00 La variable de año está muy concentrada alrededor del 2000, con una media de 2004.00 y una mediana de 2006.00. Esto implica que la mayoría de las canciones en este conjunto de datos son de ese periodo. Los bigotes son cortos, lo que indica que no hay mucha variación en los datos. Si bien hay algunos valores atípicos, estos están relativamente cerca del año 2000. La moda (2007) sugiere que este año es el más frecuente en el conjunto de datos.

Mode: Media: 0.63 Mediana: 1.00 Moda: 1.00 La variable de mode parece ser binaria (solo toma valores de 0 y 1). La media (0.63) y la mediana (1.00) indican que hay una ligera predominancia de valores 1, lo que significa que la mayoría de las canciones probablemente están categorizadas en la opción "mayor" si estamos hablando de música. La moda (1.00) refuerza esta observación, ya que este es el valor más frecuente en el conjunto de datos.