** ML EDA 4 Spotify Data (Module 2)**

Questions:

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import requests
import warnings
warnings.filterwarnings('ignore')

df = pd.read_csv('spotify.csv')
df
```

Out[1]:

	Artist	Track Name	Popularity	Duration (ms)	Track ID
0	Drake	Rich Baby Daddy (feat. Sexyy Red & SZA)	92	319191	1yeB8MUNeLo9Ek1UEpsyz6
1	Drake	One Dance	91	173986	1zi7xx7UVEFkmKfv06H8x0
2	Drake	IDGAF (feat. Yeat)	90	260111	2YSzYUF3jWqb9YP9VXmpjE
3	Drake	First Person Shooter (feat. J. Cole)	88	247444	7aqfrAY2p9BUSiupwk3svU
4	Drake	Jimmy Cooks (feat. 21 Savage)	88	218364	3F5CgOj3wFlRv51JsHbxhe
...	...	...	...	...	...
435	French Montana	Splash Brothers	44	221863	3fBsEOnzwtlkpS0LxXAZhN
436	Fat Joe	All The Way Up (feat. Infared)	64	191900	7EzwtgfW7khBrpvaNPtMoT
437	A$AP Ferg	Work REMIX (feat. A$AP Rocky, French Montana, ...	69	283693	7xVLFuuYdAvcTfcP3IG3dS
438	Diddy	Another One Of Me (feat. 21 Savage)	65	220408	4hGmQboiou09EwhcTWa0H6
439	Rick Ross	Stay Schemin	68	267720	0nq6sfr8z1R5KJ4XUk396e

440 rows × 5 columns

In [2]:
```
df.head()
```

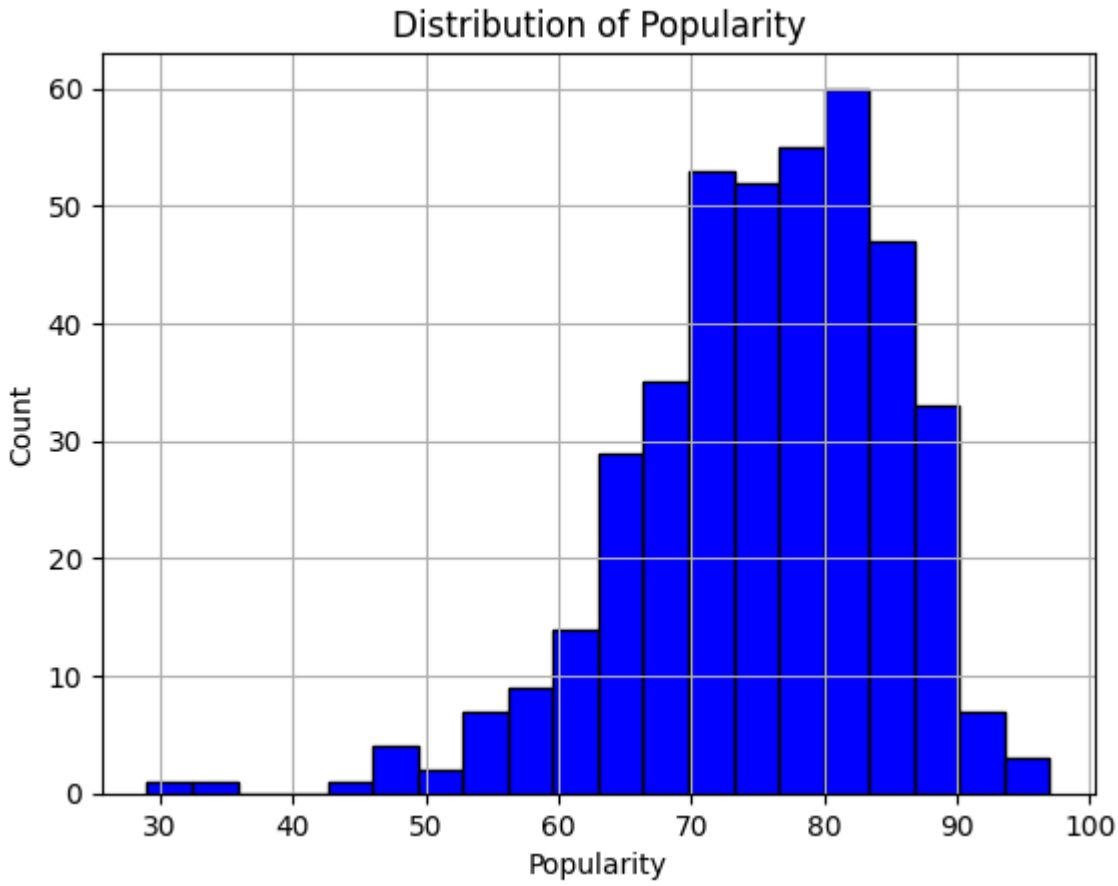
Out[2]:

	Artist	Track Name	Popularity	Duration (ms)	Track ID
0	Drake	Rich Baby Daddy (feat. Sexyy Red & SZA)	92	319191	1yeB8MUNeLo9Ek1UEpsyz6
1	Drake	One Dance	91	173986	1zi7xx7UVEFkmKfv06H8x0
2	Drake	IDGAF (feat. Yeat)	90	260111	2YSzYUF3jWqb9YP9VXmpjE
3	Drake	First Person Shooter (feat. J. Cole)	88	247444	7aqfrAY2p9BUSiupwk3svU
4	Drake	Jimmy Cooks (feat. 21 Savage)	88	218364	3F5CgOj3wFlRv51JsHbxhe

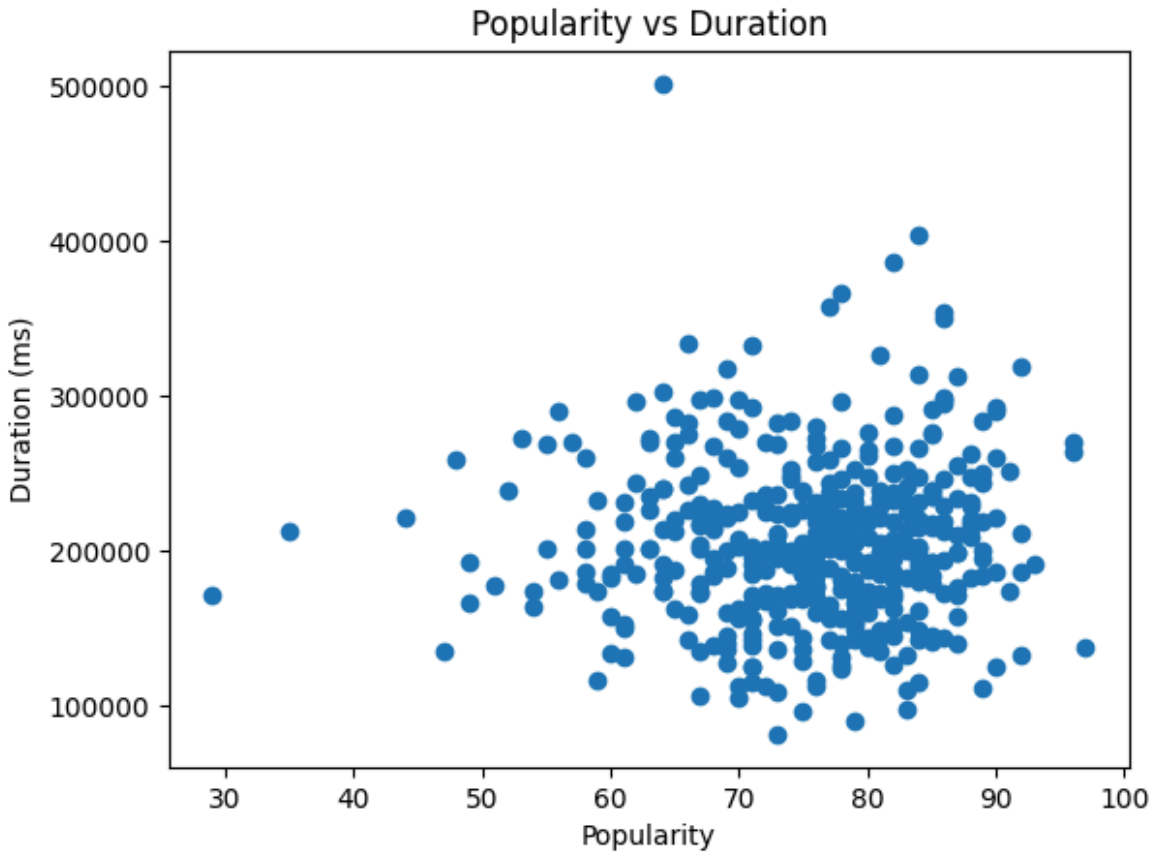
In [3]:
```
# Q1 Read the dataframe, check null value if present then do the needful, check
# Check for null values and handle them if present
if df.isnull().sum().any():
    df = df.dropna()

# Check for duplicate rows and remove them if present
if df.duplicated().any():
    df = df.drop_duplicates()

# Display the cleaned dataframe info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 413 entries, 0 to 438
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Artist         413 non-null    object
 1   Track Name     413 non-null    object
 2   Popularity     413 non-null    int64 
 3   Duration (ms)  413 non-null    int64 
 4   Track ID       413 non-null    object
dtypes: int64(2), object(3)
memory usage: 19.4+ KB
```

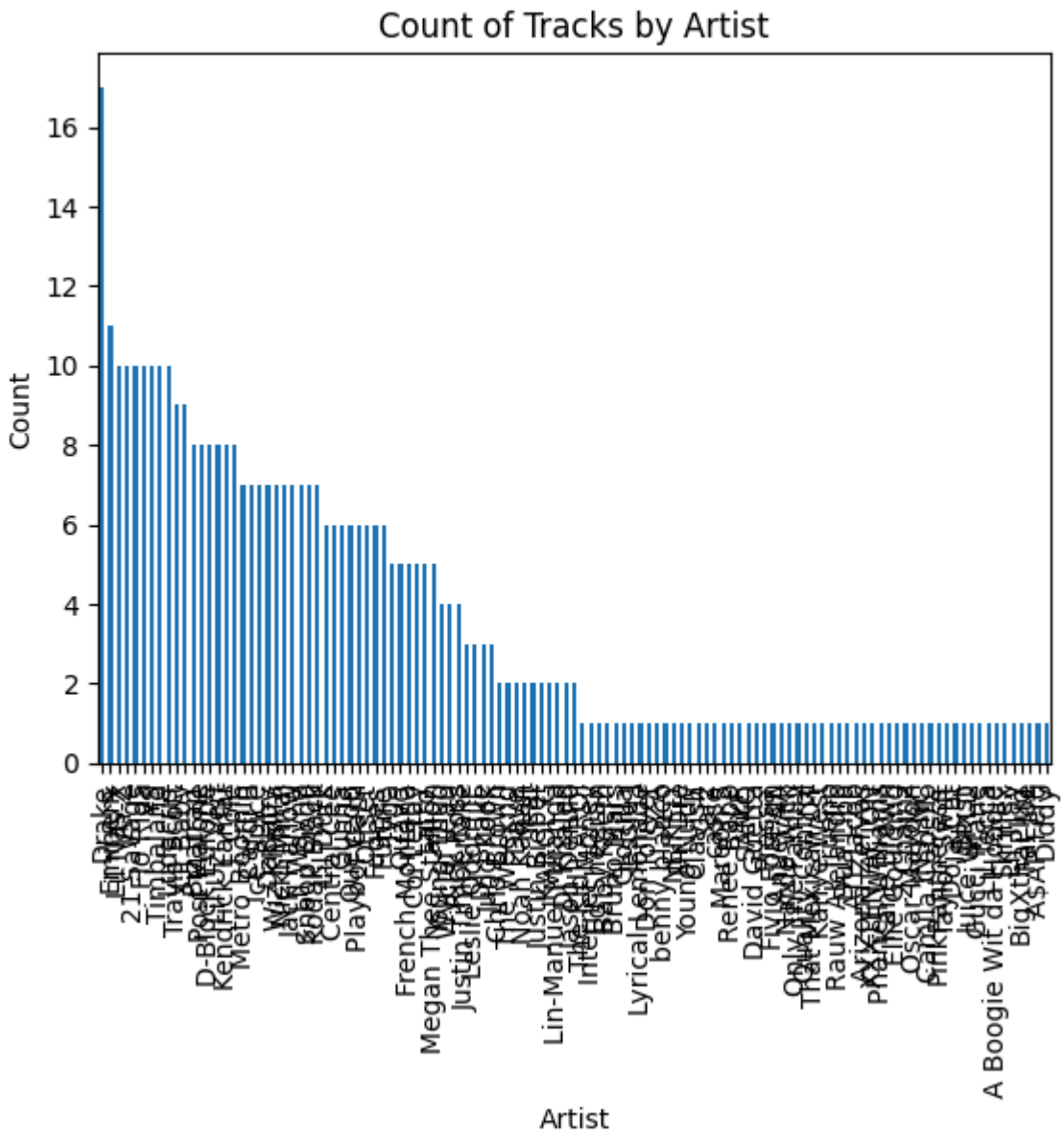
In [ ]:
```
# Q2 What is the distribution of popularity among the tracks in the dataset? Vis

df['Popularity'].hist(bins=20, color='blue', edgecolor='black')
plt.xlabel('Popularity')
plt.ylabel('Count')
plt.title('Distribution of Popularity')
plt.show()
```



In [5]:
```
# Q3 Is there any relationship between the popularity and the duration of tracks

plt.scatter(df['Popularity'], df['Duration (ms)'])
plt.xlabel('Popularity')
plt.ylabel('Duration (ms)')
plt.title('Popularity vs Duration')
plt.show()
```



In [8]:
```
# Q4 Which artist has the highest number of tracks in the dataset? Display the c

artist_counts = df['Artist'].value_counts()
artist_counts.plot(kind='bar')
plt.xlabel('Artist')
plt.ylabel('Count')
plt.title('Count of Tracks by Artist')
plt.show()
```



In [9]:
```
# Q5 What are the top 5 least popular tracks in the dataset? Provide the artist

least_popular_tracks = df.nsmallest(5, 'Popularity')
least_popular_tracks[['Artist', 'Track Name']]
```

Out[9]:

	Artist	Track Name
207	Pressa	Attachments (feat. Coi Leray)
231	Justin Bieber	Intentions
413	French Montana	Splash Brothers
225	Lil Baby	On Me - Remix
407	Wyclef Jean	911 (feat. Mary J. Blige)

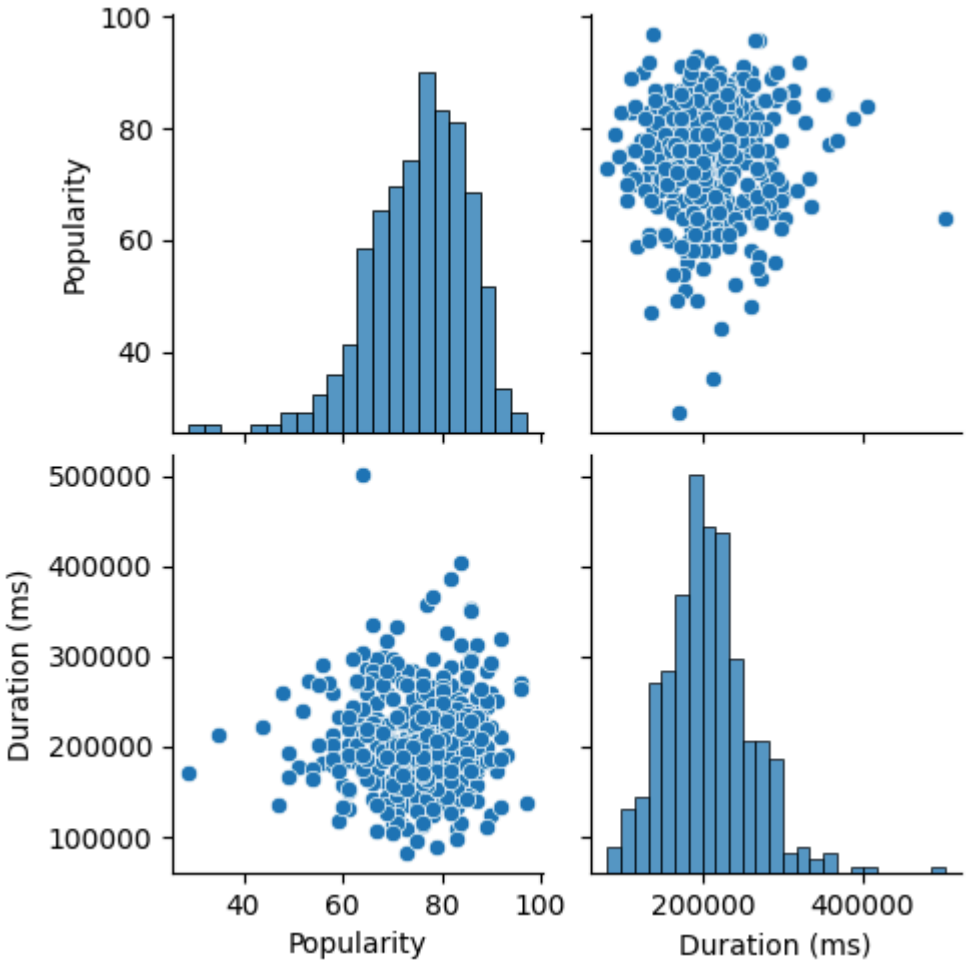
In [10]:
```
# Q6 Among the top 5 most popular artists, which artist has the highest popularity

top_artists = df.nlargest(5, 'Popularity')
average_popularity = top_artists.groupby('Artist')['Popularity'].mean()
average_popularity
```

Out[10]:
```
Artist
21 Savage       96.0
Drake           92.0
Jack Harlow     97.0
Travis Scott    93.0
¥$              96.0
Name: Popularity, dtype: float64
```

In [11]:
```
# Q7 For the top 5 most popular artists, what are their most popular tracks? Lis

top_artists = df.nlargest(5, 'Popularity')
top_tracks = top_artists.groupby('Artist')['Track Name'].first()
top_tracks
```

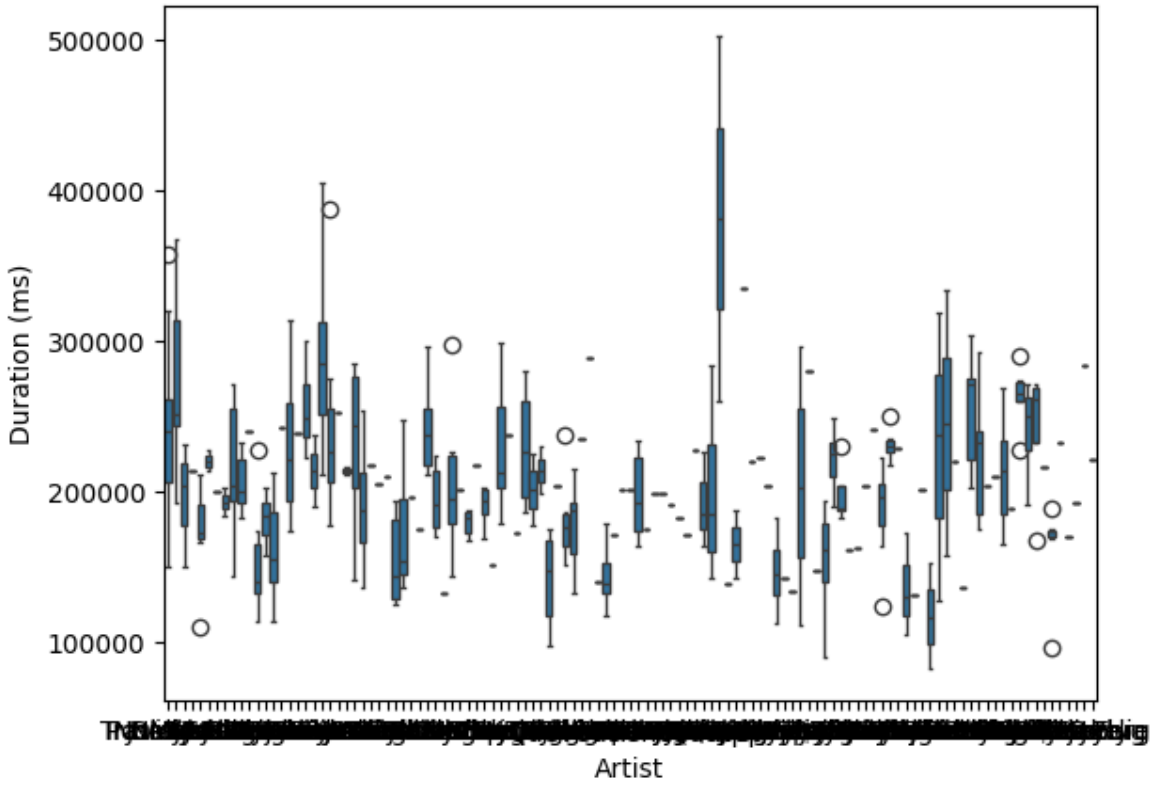
Out[11]:
```
Artist
21 Savage                                      redrum
Drake           Rich Baby Daddy (feat. Sexyy Red & SZA)
Jack Harlow                                Lovin On Me
Travis Scott                FE!N (feat. Playboi Carti)
¥$                                          CARNIVAL
Name: Track Name, dtype: object
```

In [12]:
```
# Q8 Visualize relationships between multiple numerical variables simultaneously

sns.pairplot(df[['Popularity', 'Duration (ms)']])
plt.show()
```

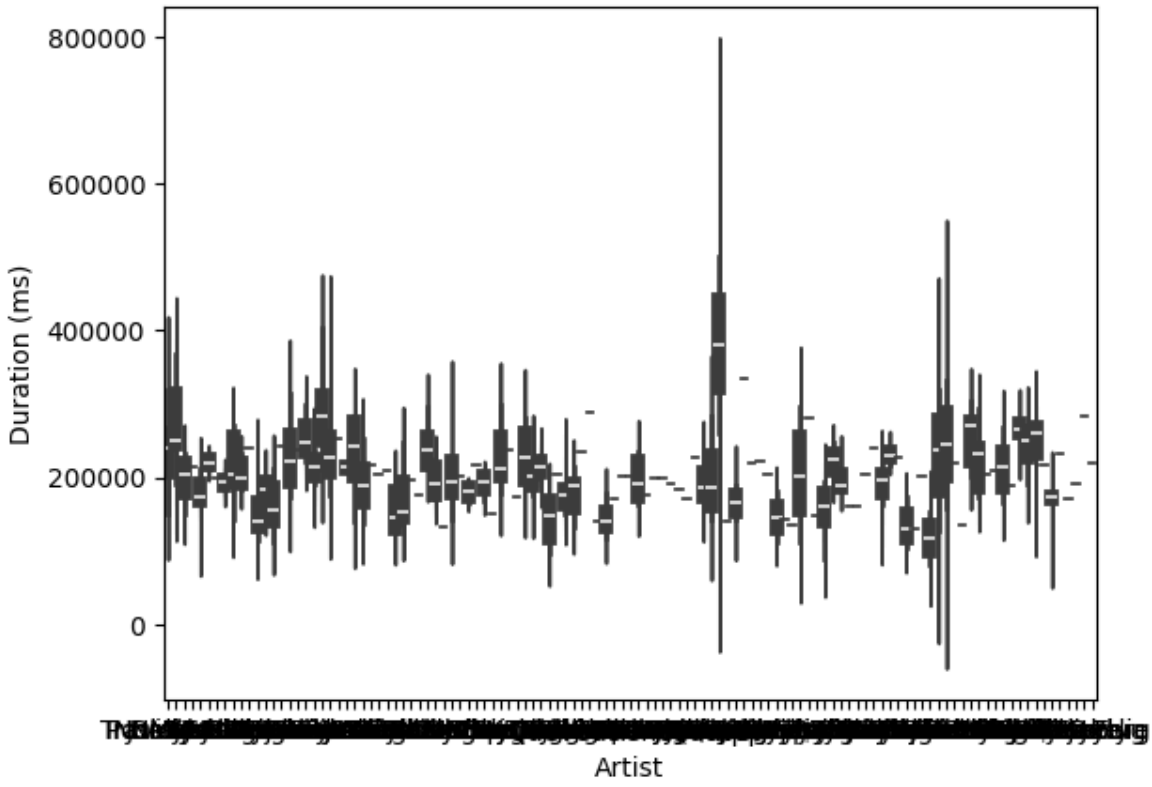


In [13]:
```
# Q9 Does the duration of tracks vary significantly across different artists? Ex
```

```
sns.boxplot(x='Artist', y='Duration (ms)', data=df)
plt.show()
```



In [14]:
```
#using violin plot
sns.violinplot(x='Artist', y='Duration (ms)', data=df)
plt.show()
```



In [ ]: