In [30]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```python
df=pd.read_excel('rolling_stones_spotify.xlsx')
```

In [5]:
```python
df.head()
```

Out[5]:

	Unnamed: 0	name	album	release_date	track_number	id	uri	acousticness	danceability	energy	instrumentalness	liveness	loudness
0	0	Concert Intro Music - Live	Licked Live In NYC	2022-06-10	1	2IEkywLJ4ykbhi1yRQvmsT	spotify:track:2IEkywLJ4ykbhi1yRQvmsT	0.0824	0.463	0.993	0.996000	0.932	-12.913
1	1	Street Fighting Man - Live	Licked Live In NYC	2022-06-10	2	6GVgVJBKkGJoRfarYRvGTU	spotify:track:6GVgVJBKkGJoRfarYRvGTU	0.4370	0.326	0.965	0.233000	0.961	-4.803
2	2	Start Me Up - Live	Licked Live In NYC	2022-06-10	3	1Lu761pZ0dBTGpzxaQoZNW	spotify:track:1Lu761pZ0dBTGpzxaQoZNW	0.4160	0.386	0.969	0.400000	0.956	-4.936
3	3	If You Can't Rock Me - Live	Licked Live In NYC	2022-06-10	4	1agTQzOTUnGNggyckEqiDH	spotify:track:1agTQzOTUnGNggyckEqiDH	0.5670	0.369	0.985	0.000107	0.895	-5.535
4	4	Donâ€™t Stop - Live	Licked Live In NYC	2022-06-10	5	7piGJR8YndQBQWVXv6KtQw	spotify:track:7piGJR8YndQBQWVXv6KtQw	0.4000	0.303	0.969	0.055900	0.966	-5.098

In [6]:
```python
df = df.drop('Unnamed: 0', axis=1)
```

In [7]:
```python
df.head()
```

Out[7]:

	name	album	release_date	track_number	id	uri	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechine
0	Concert Intro Music - Live	Licked Live In NYC	2022-06-10	1	2IEkywLJ4ykbhi1yRQvmsT	spotify:track:2IEkywLJ4ykbhi1yRQvmsT	0.0824	0.463	0.993	0.996000	0.932	-12.913	0.11
1	Street Fighting Man - Live	Licked Live In NYC	2022-06-10	2	6GVgVJBKkGJoRfarYRvGTU	spotify:track:6GVgVJBKkGJoRfarYRvGTU	0.4370	0.326	0.965	0.233000	0.961	-4.803	0.07
2	Start Me Up - Live	Licked Live In NYC	2022-06-10	3	1Lu761pZ0dBTGpzxaQoZNW	spotify:track:1Lu761pZ0dBTGpzxaQoZNW	0.4160	0.386	0.969	0.400000	0.956	-4.936	0.11
3	If You Can't Rock Me - Live	Licked Live In NYC	2022-06-10	4	1agTQzOTUnGNggyckEqiDH	spotify:track:1agTQzOTUnGNggyckEqiDH	0.5670	0.369	0.985	0.000107	0.895	-5.535	0.19
4	Donâ€™t Stop - Live	Licked Live In NYC	2022-06-10	5	7piGJR8YndQBQWVXv6KtQw	spotify:track:7piGJR8YndQBQWVXv6KtQw	0.4000	0.303	0.969	0.055900	0.966	-5.098	0.09

In [8]:
```python
#checking for duplicates
duplicates = df.duplicated().sum()
```

In [9]:
```python
duplicates
```

Out[9]:
```
0
```

In [10]:
```python
# check for missing values
df.isna().sum()
```

Out[10]:
```
name                0
album               0
release_date        0
track_number        0
id                  0
uri                 0
acousticness        0
danceability        0
energy              0
instrumentalness    0
liveness            0
loudness            0
speechiness         0
tempo               0
valence             0
popularity          0
duration_ms         0
dtype: int64
```

In [11]:
```python
# checking for outliers
df.describe()
```

Out[11]:

	track_number	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	duration_ms
count	1610.000000	1610.000000	1610.000000	1610.000000	1610.000000	1610.00000	1610.000000	1610.000000	1610.000000	1610.000000	1610.000000	1610.000000
mean	8.613665	0.250475	0.468860	0.792352	0.164170	0.49173	-6.971615	0.069512	126.082033	0.582165	20.788199	257736.488199
std	6.560220	0.227397	0.141775	0.179886	0.276249	0.34910	2.994003	0.051631	29.233483	0.231253	12.426859	108333.474920
min	1.000000	0.000009	0.104000	0.141000	0.000000	0.02190	-24.408000	0.023200	46.525000	0.000000	0.000000	21000.000000
25%	4.000000	0.058350	0.362250	0.674000	0.000219	0.15300	-8.982500	0.036500	107.390750	0.404250	13.000000	190613.000000
50%	7.000000	0.183000	0.458000	0.848500	0.013750	0.37950	-6.523000	0.051200	124.404500	0.583000	20.000000	243093.000000
75%	11.000000	0.403750	0.578000	0.945000	0.179000	0.89375	-4.608750	0.086600	142.355750	0.778000	27.000000	295319.750000
max	47.000000	0.994000	0.887000	0.999000	0.996000	0.99800	-1.014000	0.624000	216.304000	0.974000	80.000000	981866.000000

In [26]:
```python
songs = df.groupby('album')['name'].count()
```

In [27]:
```python
songs
```

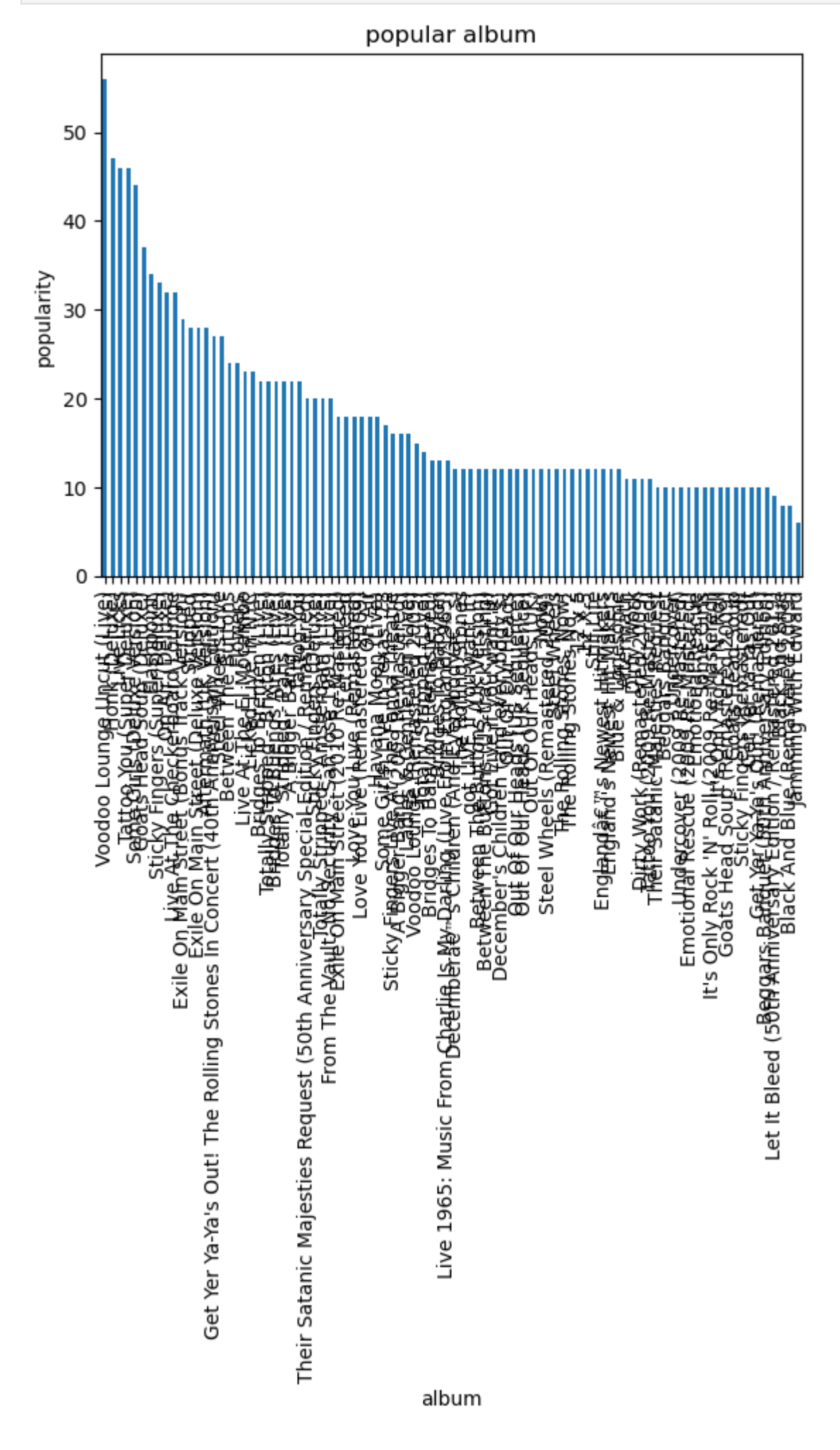
Out[27]:
```
album
12 X 5                              12
12 x 5                              12
A Bigger Bang (2009 Re-Mastered)    16
A Bigger Bang (Live)                22
Aftermath                           11
                                    ..
Undercover                          10
Undercover (2009 Re-Mastered)       10
Voodoo Lounge (Remastered 2009)     15
Voodoo Lounge Uncut (Live)          56
got LIVE if you want it!            12
Name: name, Length: 90, dtype: int64
```

In [35]:
```python
song_counts = df['album'].value_counts()
```

In [36]:
```python
song_counts
```

Out[36]:
```
Voodoo Lounge Uncut (Live)                                 56
Honk (Deluxe)                                              47
Live Licks                                                 46
Tattoo You (Super Deluxe)                                  46
Some Girls (Deluxe Version)                                44
                                                           ..
Beggars Banquet (50th Anniversary Edition)                 10
Let It Bleed (50th Anniversary Edition / Remastered 2019)   9
Black And Blue                                              8
Black And Blue (Remastered 2009)                            8
Jamming With Edward                                         6
Name: album, Length: 90, dtype: int64
```

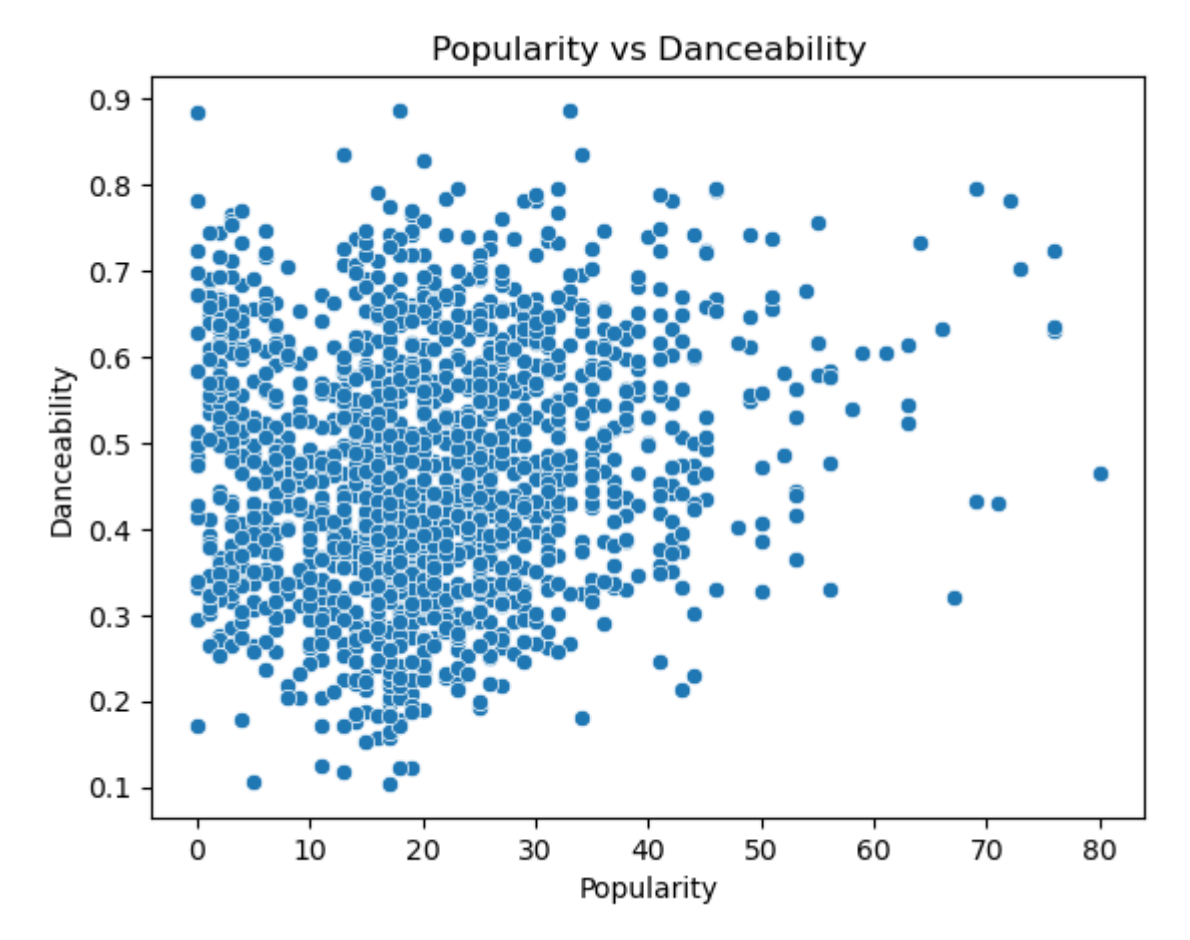
In [47]:
```python
# Plot horizontal bar chart
song_counts.plot(kind='bar')
plt.xlabel('album')
plt.ylabel('popularity')
plt.title('popular album')
plt.show()
```



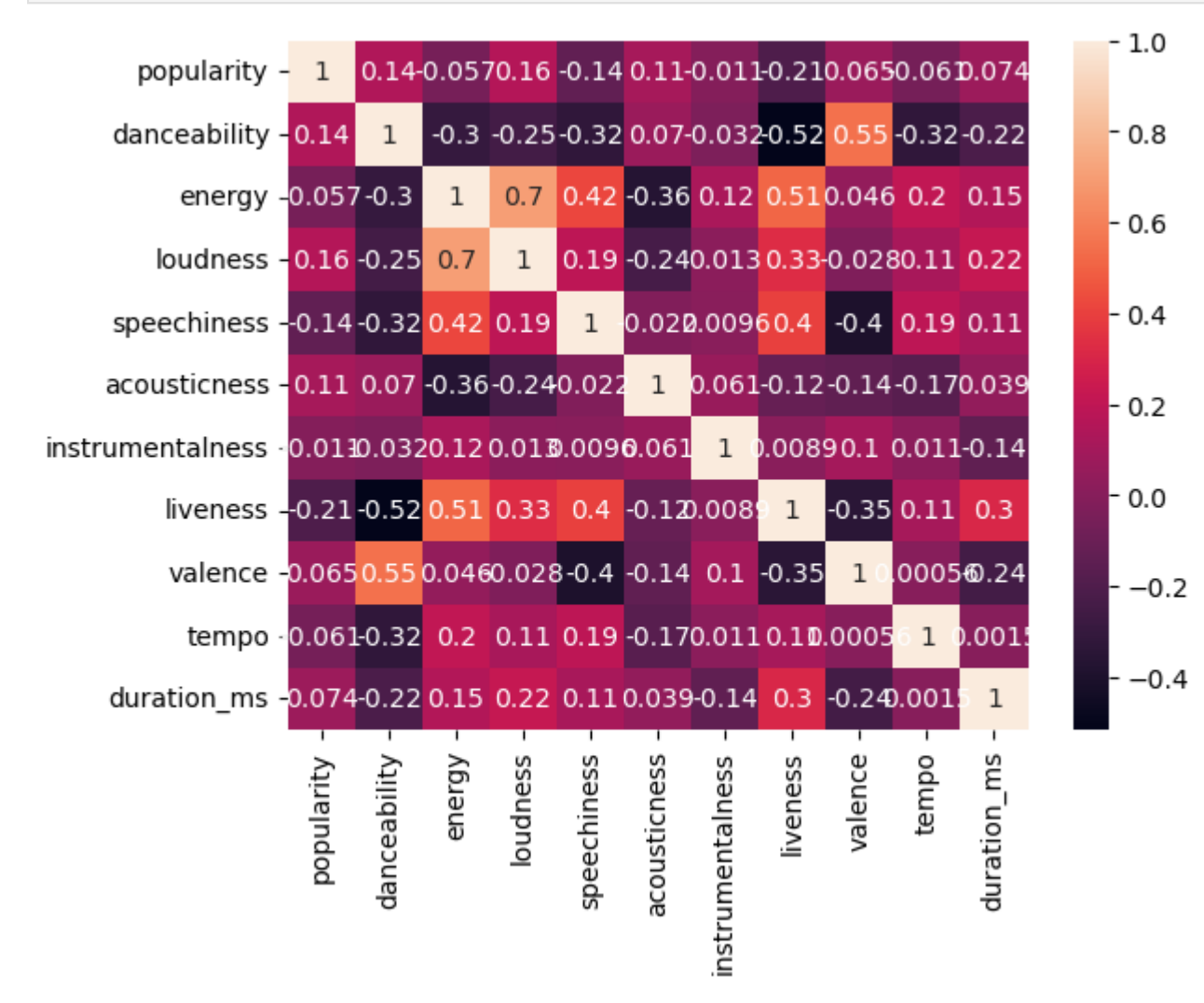
from the above graph Voodoo Lounge Uncut (Live) and Honk (Deluxe) are two popular albums

In [48]:
```python
# Exploratory data analysis
sns.scatterplot(data=df, x='popularity', y='danceability')
plt.xlabel('Popularity')
plt.ylabel('Danceability')
plt.title('Popularity vs Danceability')
```

Out[48]:
```
Text(0.5, 1.0, 'Popularity vs Danceability')
```



In [49]:
```python
corr = df[['popularity', 'danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo','duration_ms'
sns.heatmap(corr, annot=True)
plt.show()
```

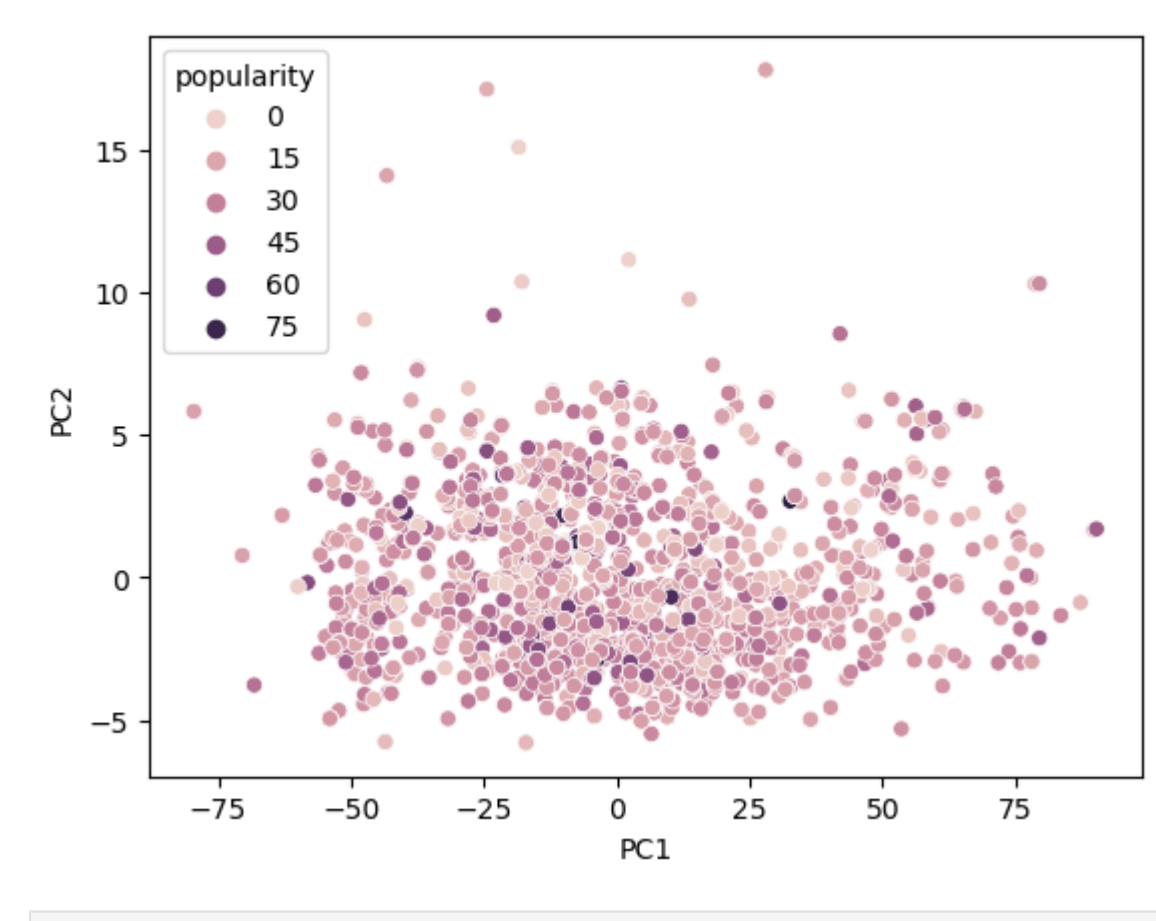


In [50]:
```python
from sklearn.decomposition import PCA
```

In [53]:
```python
pca = PCA(n_components=2)
features_transformed = pca.fit_transform(df[['danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 't
```

In [55]:
```python
sns.scatterplot(x=features_transformed[:, 0], y=features_transformed[:, 1], hue=df['popularity'])
plt.xlabel('PC1')
plt.ylabel('PC2')
```

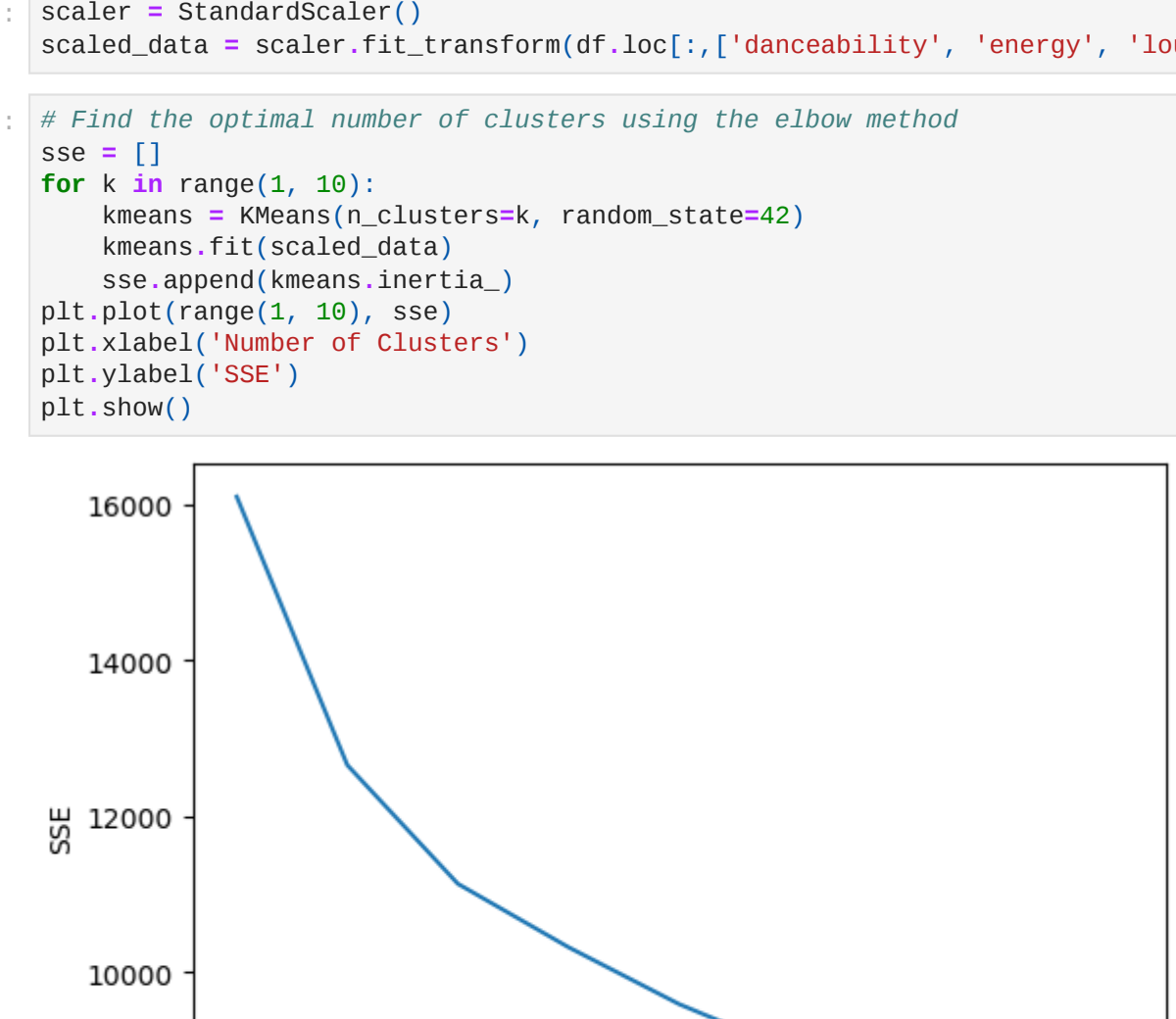
Out[55]:
```
Text(0, 0.5, 'PC2')
```



In [56]:
```python
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

In [62]:
```python
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df.loc[:,['danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 't
```

In [63]:
```python
# Find the optimal number of clusters using the elbow method
sse = []
for k in range(1, 10):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_data)
    sse.append(kmeans.inertia_)
plt.plot(range(1, 10), sse)
plt.xlabel('Number of Clusters')
plt.ylabel('SSE')
plt.show()
```



In [64]:
```python
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(scaled_data)
```

Out[64]:
```
KMeans(n_clusters=4, random_state=42)
```

In [65]:
```python
clusters = pd.DataFrame(df, columns=df.columns)
clusters['cluster'] = kmeans.labels_
```

In [68]:
```python
clusters.head()
```

Out[68]:

	name	album	release_date	track_number	id	uri	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechine
0	Concert Intro Music - Live	Licked Live In NYC	2022-06-10	1	2IEkywLJ4ykbhi1yRQvmsT	spotify:track:2IEkywLJ4ykbhi1yRQvmsT	0.0824	0.463	0.993	0.996000	0.932	-12.913	0.11
1	Street Fighting Man - Live	Licked Live In NYC	2022-06-10	2	6GVgVJBKkGJoRfarYRvGTU	spotify:track:6GVgVJBKkGJoRfarYRvGTU	0.4370	0.326	0.965	0.233000	0.961	-4.803	0.07
2	Start Me Up - Live	Licked Live In NYC	2022-06-10	3	1Lu761pZ0dBTGpzxaQoZNW	spotify:track:1Lu761pZ0dBTGpzxaQoZNW	0.4160	0.386	0.969	0.400000	0.956	-4.936	0.11
3	If You Can't Rock Me - Live	Licked Live In NYC	2022-06-10	4	1agTQzOTUnGNggyckEqiDH	spotify:track:1agTQzOTUnGNggyckEqiDH	0.5670	0.369	0.985	0.000107	0.895	-5.535	0.19
4	Donâ€™t Stop - Live	Licked Live In NYC	2022-06-10	5	7piGJR8YndQBQWVXv6KtQw	spotify:track:7piGJR8YndQBQWVXv6KtQw	0.4000	0.303	0.969	0.055900	0.966	-5.098	0.09

In [69]:
```python
#to print the features of cluster
for i in range(kmeans.n_clusters):
    print(f'Cluster {i}:')
    print(clusters[clusters['cluster'] == i].describe())
```

```
Cluster 0:
       track_number  acousticness  danceability      energy  instrumentalness  \
count    508.000000    508.000000    508.000000  508.000000        508.000000
mean       7.746063      0.157733      0.553057    0.809270          0.056819
std        6.526874      0.145471      0.120879    0.122161          0.096405
min        1.000000      0.000114      0.181000    0.421000          0.000000
25%        3.000000      0.033050      0.468750    0.709750          0.000180
50%        7.000000      0.115500      0.563000    0.838000          0.008440
75%       10.000000      0.242000      0.633250    0.909000          0.069100
max       46.000000      0.686000      0.887000    0.981000          0.498000

         liveness    loudness  speechiness       tempo     valence  \
count  508.000000  508.000000   508.000000  508.000000  508.000000
mean     0.294923   -6.708183     0.050291  123.944476    0.762242
std      0.245778    2.618641     0.022621   24.887473    0.141705
min      0.021900  -13.189000     0.023200   70.003000    0.365000
25%      0.103250   -8.964250     0.035400  108.084500    0.656500
50%      0.218500   -6.616500     0.043100  121.753000    0.783500
75%      0.379500   -4.496750     0.057950  137.955000    0.869000
max      0.994000   -1.382000     0.147000  201.878000    0.974000

       popularity    duration_ms  cluster
count  508.000000     508.000000    508.0
mean    22.828740  229499.226378      0.0
std     13.766731   65402.911109      0.0
min      0.000000   76546.000000      0.0
25%     15.000000  187160.000000      0.0
50%     22.000000  227026.500000      0.0
75%     30.000000  271732.750000      0.0
max     80.000000  436000.000000      0.0
Cluster 1:
       track_number  acousticness  danceability      energy  instrumentalness  \
count    532.000000    532.000000    532.000000  532.000000        532.000000
mean      10.422932      0.193690      0.358784    0.924855          0.110515
std        7.324319      0.182676      0.095137    0.075542          0.209983
min        1.000000      0.000009      0.104000    0.559000          0.000000
25%        5.000000      0.046150      0.292750    0.903750          0.000268
50%        9.000000      0.139000      0.358000    0.955000          0.010195
75%       15.000000      0.303250      0.426250    0.975000          0.096475
max       47.000000      0.920000      0.614000    0.997000          0.991000

         liveness    loudness  speechiness       tempo     valence  \
count  532.000000  532.000000   532.000000  532.000000  532.000000
mean     0.844962   -5.316331     0.108948  137.156096    0.439310
std      0.201420    1.740774     0.065344   28.818755    0.172192
min      0.044000  -13.770000     0.029300   46.525000    0.015900
25%      0.794750   -6.406250     0.065825  117.807500    0.330000
50%      0.938500   -5.086500     0.093550  137.296500    0.429000
75%      0.969000   -4.131250     0.130250  154.457500    0.562500
max      0.998000   -1.014000     0.624000  216.304000    0.902000

       popularity    duration_ms  cluster
count  532.000000     532.000000    532.0
mean    17.772556  314265.546992      1.0
std      7.827168  134874.335712      0.0
min      0.000000   32946.000000      1.0
25%     13.000000  230989.500000      1.0
50%     18.000000  280640.000000      1.0
75%     22.000000  360852.750000      1.0
max     50.000000  981866.000000      1.0
Cluster 2:
       track_number  acousticness  danceability      energy  instrumentalness  \
count    197.000000    197.000000    197.000000  197.000000        197.000000
mean       7.898477      0.248485      0.507010    0.832046          0.752010
std        6.638620      0.223449      0.125947    0.122338          0.147871
min        1.000000      0.000055      0.107000    0.516000          0.380000
25%        3.000000      0.039800      0.428000    0.773000          0.640000
50%        7.000000      0.194000      0.515000    0.840000          0.770000
75%       10.000000      0.422000      0.602000    0.933000          0.876000
max       37.000000      0.824000      0.750000    0.999000          0.996000

         liveness    loudness  speechiness       tempo     valence  \
count  197.000000  197.000000   197.000000  197.000000  197.000000
mean     0.420160   -6.771477     0.055232  125.841305    0.722951
std      0.305442    2.736743     0.031562   27.309186    0.204017
min      0.028100  -13.197000     0.026200   55.581000    0.016500
25%      0.152000   -8.749000     0.034000  109.572000    0.612000
50%      0.325000   -6.879000     0.042300  124.917000    0.771000
75%      0.666000   -4.585000     0.061700  142.270000    0.885000
max      0.990000   -1.568000     0.188000  189.824000    0.969000

       popularity    duration_ms  cluster
count  197.000000     197.000000    197.0
mean    21.167513  201904.786802      2.0
std     12.568480   64814.467300      0.0
min      0.000000   48640.000000      2.0
25%     14.000000  148560.000000      2.0
50%     20.000000  203333.000000      2.0
75%     29.000000  244853.000000      2.0
max     63.000000  493373.000000      2.0
Cluster 3:
       track_number  acousticness  danceability      energy  instrumentalness  \
count    373.000000    373.000000    373.000000  373.000000        373.000000
mean       7.592493      0.458825      0.491038    0.559362          0.076433
std        4.611247      0.246959      0.132189    0.151085          0.179945
min        1.000000      0.001290      0.172000    0.141000          0.000000
25%        4.000000      0.229000      0.382000    0.458000          0.000041
50%        7.000000      0.516000      0.477000    0.559000          0.002080
75%       10.000000      0.666000      0.604000    0.665000          0.041500
max       33.000000      0.994000      0.835000    0.918000          0.915000

         liveness    loudness  speechiness       tempo     valence  \
count  373.000000  373.000000   373.000000  373.000000  373.000000
mean     0.293761   -9.796981     0.046987  113.325735    0.466308
std      0.253745    3.019243     0.029459   30.478500    0.205577
min      0.026100  -24.408000     0.024400   57.772000    0.000000
25%      0.116000  -11.375000     0.030700   90.184000    0.296000
50%      0.192000   -9.603000     0.039400  110.198000    0.472000
75%      0.355000   -7.774000     0.049800  130.531000    0.589000
max      0.978000   -2.890000     0.242000  205.687000    0.963000

       popularity    duration_ms  cluster
count  373.000000     373.000000    373.0
mean    22.109920  245055.294906      3.0
std     14.779608   96340.998868      0.0
min      0.000000   21000.000000      3.0
25%     11.000000  176253.000000      3.0
50%     21.000000  227533.000000      3.0
75%     31.000000  297386.000000      3.0
max     76.000000  673266.000000      3.0
```

In [ ]: