In [1]:
```
import glob
import pandas as pd

extension = 'csv'
all_filenames = sorted([i for i in glob.glob('datasets/spotify_artists/*.{}'.format(exte
# combine all files in the list
combined_csv = pd.concat([pd.read_csv(f) for f in all_filenames])
combined_csv.drop(combined_csv.columns[combined_csv.columns.str.contains('unnamed',case
# display
print ("Number of Columns before data cleansing: %d"%len(combined_csv.columns))
print ("Number of rows before cleansing: %d"%combined_csv.title.count())
combined_csv.describe()
display(combined_csv)
```

```
Number of Columns after data cleansing: 3
Number of rows after cleansing: 2653
```

	title	all_artists	popularity
0	Rowboat	Johnny Cash	26
1	Better Days - Single Edit	Bruce Springsteen	38
2	Contigo	Enrique Iglesias	38
3	Technologic - Vitalic Remix	Daft Punk	20
4	A cause	Céline Dion	25
...	...	...	...
92	More Than Just Friends	Mariah Carey	25
93	Make a Mistake	Brad Paisley	23
94	Night Train	James Brown	45
95	Grown so Ugly	The Black Keys	34
96	Christmas Lullaby	Mannheim Steamroller	8

2653 rows × 3 columns

In [2]:
```
combined_csv.rename(columns={'title':'title_spoty'}, inplace=True)
```

In [3]:
```
#drop repeated (title, artist)
combined_csv.drop_duplicates(subset = ['title_spoty','all_artists'], keep = 'first', inp
# display
print ("Number of Columns after data cleansing: %d"%len(combined_csv.columns))
print ("Number of rows after cleansing: %d"%combined_csv.title_spoty.count())
combined_csv.describe()
```

```
Number of Columns after data cleansing: 3
Number of rows after cleansing: 2443
```

Out[3]:

	popularity
count	2443.000000
mean	32.145722
std	19.296451
min	0.000000
25%	19.000000
50%	32.000000

75%	46.000000
max	86.000000

In [4]:
```
combined_csv['index'] = range(1, len(combined_csv) + 1)
```

In [5]:
```
combined_csv.describe()
```

Out[5]:

	popularity	index
count	2443.000000	2443.000000
mean	32.145722	1222.000000
std	19.296451	705.377677
min	0.000000	1.000000
25%	19.000000	611.500000
50%	32.000000	1222.000000
75%	46.000000	1832.500000
max	86.000000	2443.000000

In [6]:
```
# export to csv
combined_csv.to_csv("spotify_combined.csv", index=False, encoding='utf-8-sig')
```