Spotify Data Analysis

This is complete code for Spotify data analysis. Dataset is available on [www.Kaggle.com (http://www.Kaggle.com)](http://www.Kaggle.com)

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
#import track dataset
df_tracks = pd.read_csv("tracks.csv")
df_tracks.head(5)
```

Out[2]:

	id	name	popularity	duration_ms	explicit	artists	id_artists	release_date	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
0	35iwgR4jXetI318WEWsa1Q	Carve	6	126903	0	['Uli']	['45tIt06XoI0Iio4LBEVpls']	1922-02-22	0.645	0.4450	0	-13.338	1	0.4510	0.674	0.7440	0.151	0.127	104.851	3
1	021ht4sdgPcrDgSk7JTbKY	Capítulo 2.16 - Banquero Anarquista	0	98200	0	['Fernando Pessoa']	['14jtPCOoNZwquk5wd9DxrY']	1922-06-01	0.695	0.2630	0	-22.136	1	0.9570	0.797	0.0000	0.148	0.655	102.009	1
2	07A5yehtSnoedViJAZkNnc	Vivo para Quererte - Remasterizado	0	181640	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	1922-03-21	0.434	0.1770	1	-21.180	1	0.0512	0.994	0.0218	0.212	0.457	130.418	5
3	08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	176907	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	1922-03-21	0.321	0.0946	7	-27.961	1	0.0504	0.995	0.9180	0.104	0.397	169.980	3
4	08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	163080	0	['Dick Haymes']	['3BiJGZsyX9sJchTqcSA7Su']	1922	0.402	0.1580	3	-16.900	0	0.0390	0.989	0.1300	0.311	0.196	103.220	4

Top 5 Least Popular Songs

In [3]:
```
df_tracks.sort_values('popularity', ascending = True).head(5)
```

Out[3]:

	id	name	popularity	duration_ms	explicit	artists	id_artists	release_date	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
546130	181rTRhCcggZPwP2TUcVqm	Newspaper Reports On Abner, 20 February 1935	0	896575	0	['Norris Goff', 'Chester Lauck', 'Carlton Bric...	['3WCwCPDMpGzrt0Qz6quumy', '7vk8UqABg0Sga78GI3...	1935-02-20	0.595	0.262	8	-17.746	1	0.9320	0.993	0.007510	0.0991	0.320	79.849	4
546222	0yOCz3V5KMm8l1T8EFc60i	恋は水の上で	0	188440	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949	0.418	0.388	0	-8.580	1	0.0358	0.925	0.000014	0.1050	0.439	94.549	4
546221	0y48Hhwe52099UqYjegRCO	私の誕生日	0	173467	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949	0.642	0.178	5	-11.700	1	0.0501	0.993	0.000943	0.0928	0.715	119.013	4
546220	0xCmgtf9ka07hkZg3D6PaV	エル・チョクロ (EL CHOCLO)	0	205280	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949	0.695	0.467	0	-12.236	0	0.0422	0.827	0.000000	0.0861	0.756	125.941	4
546219	0tBXS3VuCPX7KWUFH2nros	恋は不思議なもの	0	185733	0	['Hibari Misora']	['1m5pMY5blqJwdxJ7vqQtuN']	1949	0.389	0.388	2	-8.221	1	0.0351	0.869	0.000000	0.0924	0.372	72.800	4

Most Popular Songs whose popularity is greater than 90

In [4]:
```
df_tracks.query('popularity > 90', inplace = False).sort_values('popularity', ascending = False).head(5)
```

Out[4]:

	id	name	popularity	duration_ms	explicit	artists	id_artists	release_date	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
93802	4iJyoBOLtHqaGxP12qzhQI	Peaches (feat. Daniel Caesar & Giveon)	100	198082	1	['Justin Bieber', 'Daniel Caesar', 'Giveon']	['1uNFoZAHBGtllmzznpCI3s', '20wkVLutqVOYrc0kxF...	2021-03-19	0.677	0.696	0	-6.181	1	0.1190	0.3210	0.000000	0.420	0.464	90.030	4
93803	7lPN2DXiMsVn7XUKtOW1CS	drivers license	99	242014	1	['Olivia Rodrigo']	['1McMsnEElThX1knmY4oliG']	2021-01-08	0.585	0.436	10	-8.761	1	0.0601	0.7210	0.000013	0.105	0.132	143.874	4
93804	3Ofmpyhv5UAQ70mENzB277	Astronaut In The Ocean	98	132780	0	['Masked Wolf']	['1uU7g3DNSbsu0QjSEqZtEd']	2021-01-06	0.778	0.695	4	-6.865	0	0.0913	0.1750	0.000000	0.150	0.472	149.996	4
92810	5QO79kh1waicV47BqGRL3g	Save Your Tears	97	215627	1	['The Weeknd']	['1Xyo4u8uXC1ZmMpatF05PJ']	2020-03-20	0.680	0.826	0	-5.487	1	0.0309	0.0212	0.000012	0.543	0.644	118.051	4
92811	6tDDoYIxWvMLTdKpjFkc1B	telepatía	97	160191	0	['Kali Uchis']	['1U1el3k54VvEUzo3ybLPIM']	2020-12-04	0.653	0.524	11	-9.016	0	0.0502	0.1120	0.000000	0.203	0.553	83.970	4

Changing dataframe index to relase_date

In [5]:
```
df_tracks.set_index('release_date', inplace = True)
df_tracks.index = pd.to_datetime(df_tracks.index)
df_tracks.head(5)
```

Out[5]:

	id	name	popularity	duration_ms	explicit	artists	id_artists	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
release_date																			
1922-02-22	35iwgR4jXetI318WEWsa1Q	Carve	6	126903	0	['Uli']	['45tIt06XoI0Iio4LBEVpls']	0.645	0.4450	0	-13.338	1	0.4510	0.674	0.7440	0.151	0.127	104.851	3
1922-06-01	021ht4sdgPcrDgSk7JTbKY	Capítulo 2.16 - Banquero Anarquista	0	98200	0	['Fernando Pessoa']	['14jtPCOoNZwquk5wd9DxrY']	0.695	0.2630	0	-22.136	1	0.9570	0.797	0.0000	0.148	0.655	102.009	1
1922-03-21	07A5yehtSnoedViJAZkNnc	Vivo para Quererte - Remasterizado	0	181640	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.434	0.1770	1	-21.180	1	0.0512	0.994	0.0218	0.212	0.457	130.418	5
1922-03-21	08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	176907	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.321	0.0946	7	-27.961	1	0.0504	0.995	0.9180	0.104	0.397	169.980	3
1922-01-01	08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	163080	0	['Dick Haymes']	['3BiJGZsyX9sJchTqcSA7Su']	0.402	0.1580	3	-16.900	0	0.0390	0.989	0.1300	0.311	0.196	103.220	4

Converting duraiton of songs into seconds

In [6]:
```
df_tracks['duration'] = df_tracks['duration_ms'].apply(lambda x: round(x/1000))
df_tracks.drop('duration_ms', inplace = True, axis = 1)
```

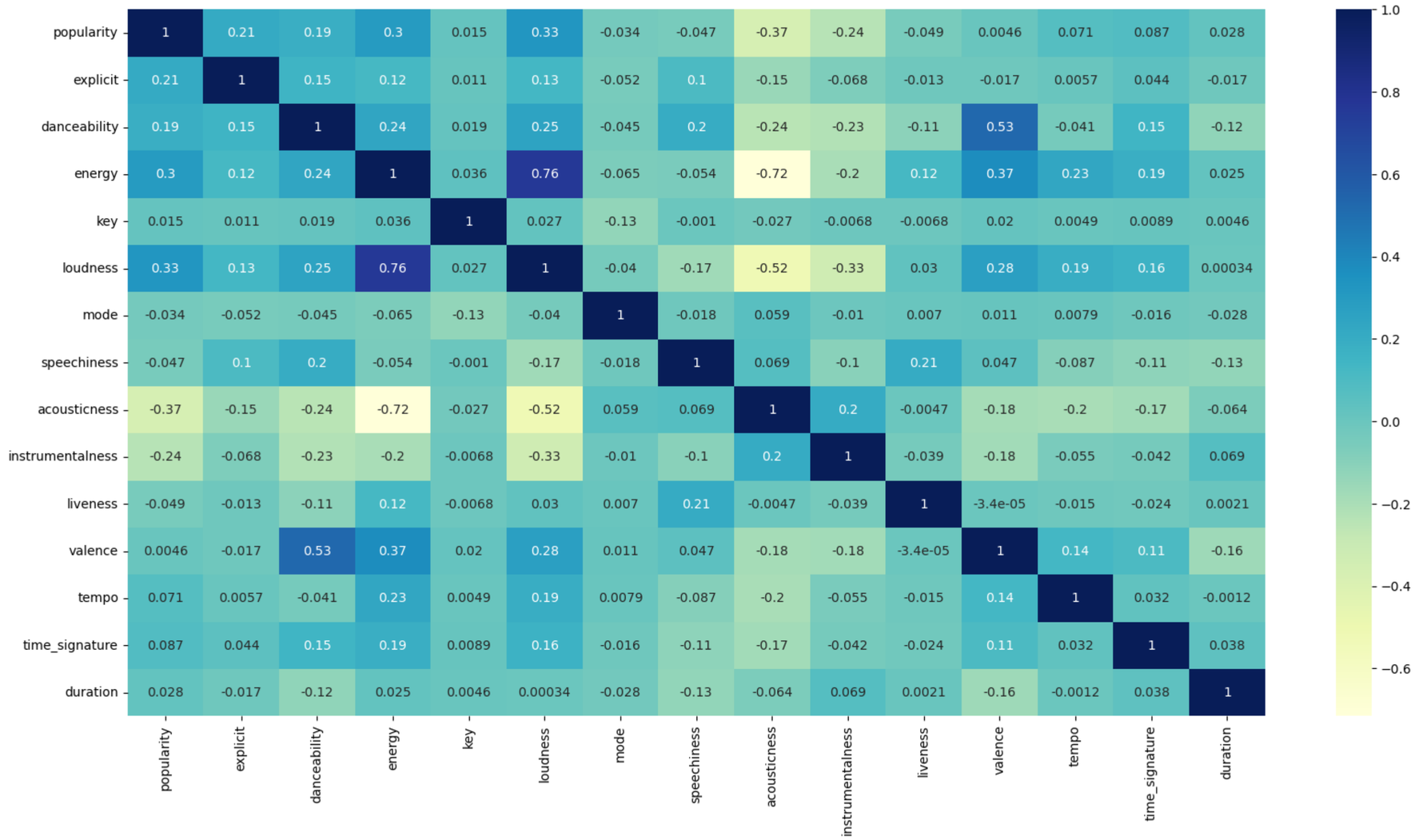
In [7]:
```
df_tracks.head(5)
```

Out[7]:

release_date	id	name	popularity	explicit	artists	id_artists	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	duration
1922-02-22	35iwgR4jXetI318WEWsa1Q	Carve	6	0	['Uli']	['45tIt06XoI0Iio4LBEVpls']	0.645	0.4450	0	-13.338	1	0.4510	0.674	0.7440	0.151	0.127	104.851	3	127
1922-06-01	021ht4sdgPcrDgSk7JTbKY	Capítulo 2.16 - Banquero Anarquista	0	0	['Fernando Pessoa']	['14jtPCOoNZwquk5wd9DxrY']	0.695	0.2630	0	-22.136	1	0.9570	0.797	0.0000	0.148	0.655	102.009	1	98
1922-03-21	07A5yehtSnoedViJAZkNnc	Vivo para Quererte - Remasterizado	0	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.434	0.1770	1	-21.180	1	0.0512	0.994	0.0218	0.212	0.457	130.418	5	182
1922-03-21	08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.321	0.0946	7	-27.961	1	0.0504	0.995	0.9180	0.104	0.397	169.980	3	177
1922-01-01	08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	0	['Dick Haymes']	['3BiJGZsyX9sJchTqcSA7Su']	0.402	0.1580	3	-16.900	0	0.0390	0.989	0.1300	0.311	0.196	103.220	4	163

Correlation heatmap between variables

In [8]:
```
plt.figure(figsize = (20,10))
df_tracks_corr = sns.heatmap(df_tracks.corr(), cmap = 'YlGnBu', annot = True)
```



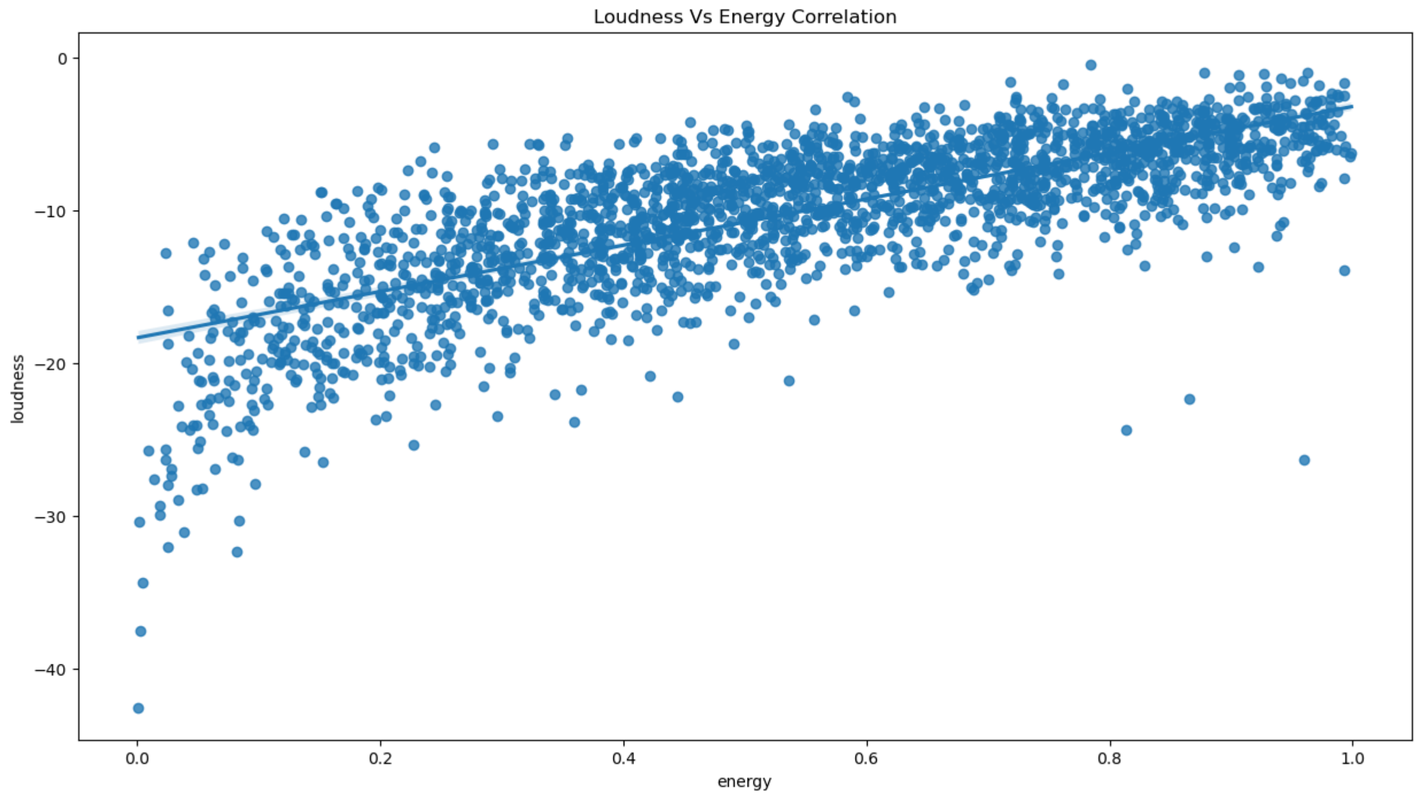
Let's create 02 regression plots to show correlation between different variables for that I would take 0.4% sample data from original data

In [9]:
```
sample_df = df_tracks.sample(frac = 0.004)
print(len(sample_df))
```

```
2347
```

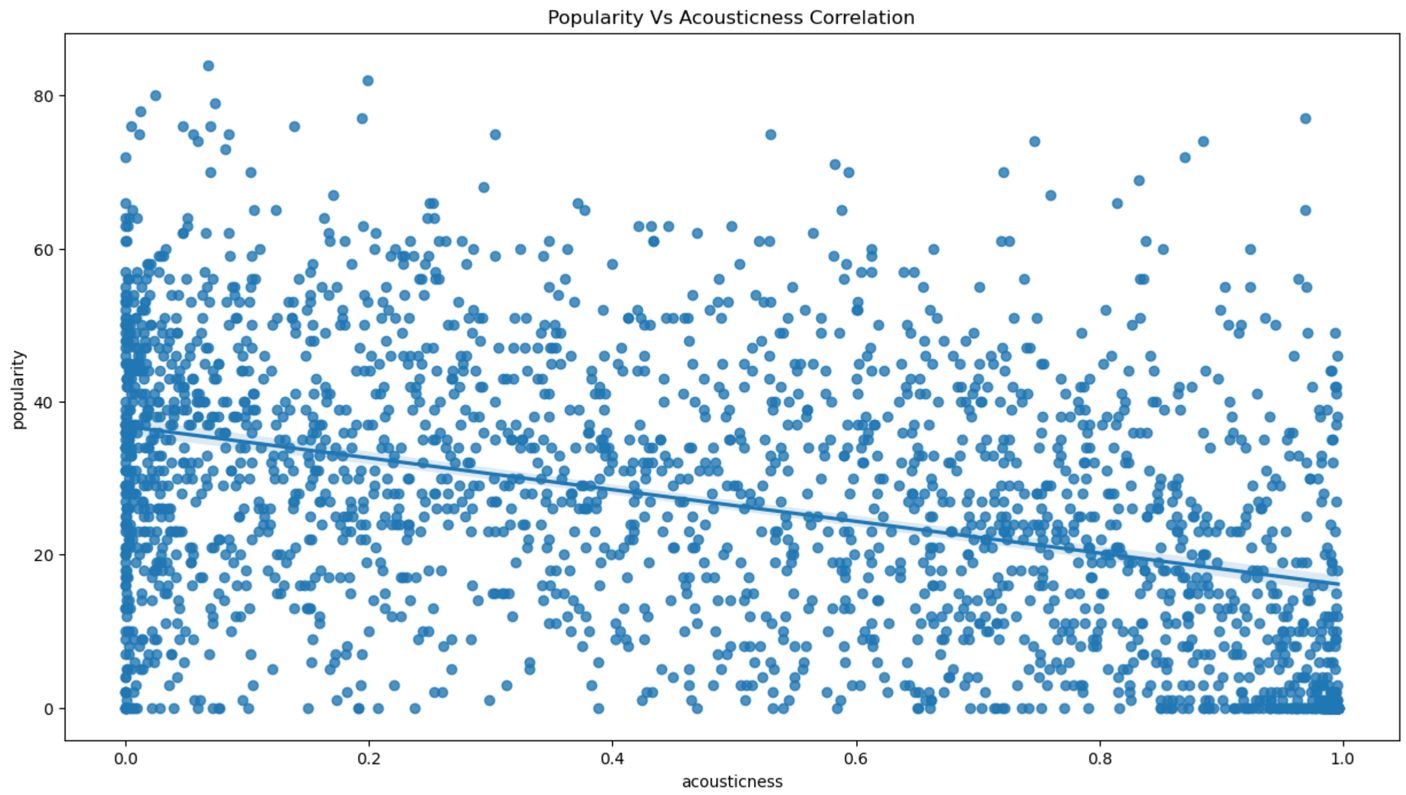
In [10]:
```
plt.figure(figsize = (15,8))
sns.regplot(data= sample_df, y='loudness', x='energy').set(title='Loudness Vs Energy Correlation')
```

Out[10]:
```
[Text(0.5, 1.0, 'Loudness Vs Energy Correlation')]
```



In [11]:
```
plt.figure(figsize = (15,8))
sns.regplot(data = sample_df, y='popularity', x='acousticness').set(title = 'Popularity Vs Acousticness Correlation')
```

Out[11]:
```
[Text(0.5, 1.0, 'Popularity Vs Acousticness Correlation')]
```



Create New Column "year" from "release_date"

In [12]:
```
df_tracks['date'] = df_tracks.index.get_level_values('release_date')
df_tracks.date = pd.to_datetime(df_tracks.date)
years = df_tracks.date.dt.year
```

In [13]:
```
df_tracks.head(5)
```

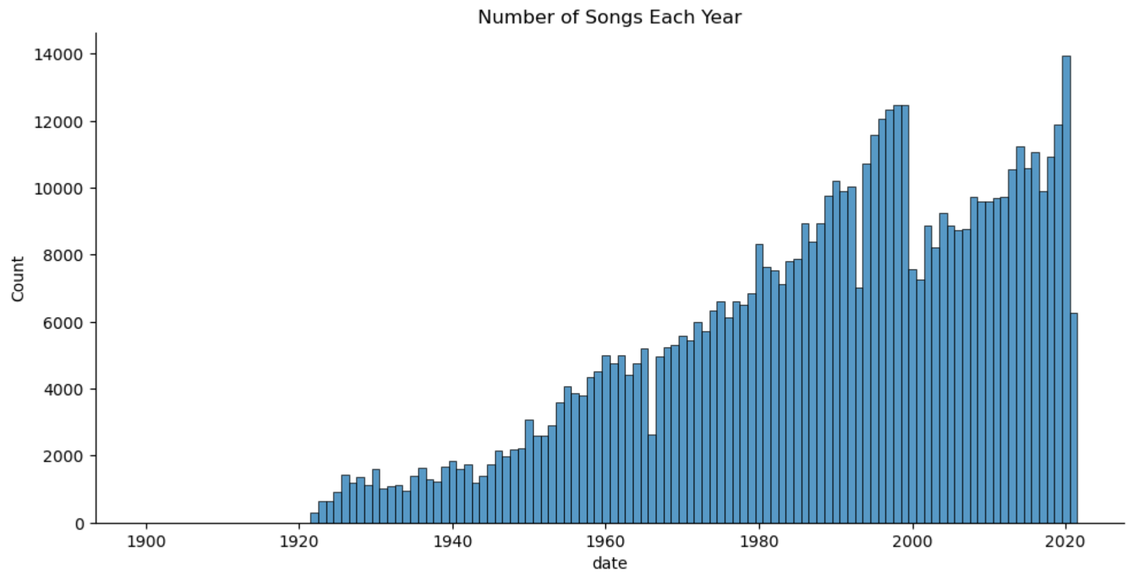
Out[13]:

release_date	id	name	popularity	explicit	artists	id_artists	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	duration	date
1922-02-22	35iwgR4jXetI318WEWsa1Q	Carve	6	0	['Uli']	['45tIt06XoI0Iio4LBEVpls']	0.645	0.4450	0	-13.338	1	0.4510	0.674	0.7440	0.151	0.127	104.851	3	127	1922-02-22
1922-06-01	021ht4sdgPcrDgSk7JTbKY	Capítulo 2.16 - Banquero Anarquista	0	0	['Fernando Pessoa']	['14jtPCOoNZwquk5wd9DxrY']	0.695	0.2630	0	-22.136	1	0.9570	0.797	0.0000	0.148	0.655	102.009	1	98	1922-06-01
1922-03-21	07A5yehtSnoedViJAZkNnc	Vivo para Quererte - Remasterizado	0	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.434	0.1770	1	-21.180	1	0.0512	0.994	0.0218	0.212	0.457	130.418	5	182	1922-03-21
1922-03-21	08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	0	['Ignacio Corsini']	['5LiOoJbxVSAMkBS2fUm3X2']	0.321	0.0946	7	-27.961	1	0.0504	0.995	0.9180	0.104	0.397	169.980	3	177	1922-03-21
1922-01-01	08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	0	['Dick Haymes']	['3BiJGZsyX9sJchTqcSA7Su']	0.402	0.1580	3	-16.900	0	0.0390	0.989	0.1300	0.311	0.196	103.220	4	163	1922-01-01

Total No.of Songs each year since 1922 using Distribution plot

In [14]:
```
sns.displot(years, discrete = True, aspect= 2, height = 5, kind='hist').set(title='Number of Songs Each Year')
```

Out[14]:
```
<seaborn.axisgrid.FacetGrid at 0x28f5de5bd90>
```



Duration of Songs over the years using barplot

In [15]:
```
plt.figure(figsize = (18,7))
total_duration = df_tracks.duration
sns.barplot(x=years, y=total_duration, data=df_tracks, errwidth=False).set(title='year Vs Duration')
plt.xticks(rotation = 90)
```

Out[15]:
```
(array([  0,   1,   2,   3,   4,   5,   6,   7,   8,   9,  10,  11,  12,
         13,  14,  15,  16,  17,  18,  19,  20,  21,  22,  23,  24,  25,
         26,  27,  28,  29,  30,  31,  32,  33,  34,  35,  36,  37,  38,
         39,  40,  41,  42,  43,  44,  45,  46,  47,  48,  49,  50,  51,
         52,  53,  54,  55,  56,  57,  58,  59,  60,  61,  62,  63,  64,
         65,  66,  67,  68,  69,  70,  71,  72,  73,  74,  75,  76,  77,
         78,  79,  80,  81,  82,  83,  84,  85,  86,  87,  88,  89,  90,
         91,  92,  93,  94,  95,  96,  97,  98,  99, 100]),
 [Text(0, 0, '1900'),
  Text(1, 0, '1922'),
  Text(2, 0, '1923'),
  Text(3, 0, '1924'),
  Text(4, 0, '1925'),
  Text(5, 0, '1926'),
  Text(6, 0, '1927'),
  Text(7, 0, '1928'),
  Text(8, 0, '1929'),
  Text(9, 0, '1930'),
  Text(10, 0, '1931'),
```

Import Feature Dataset for analyzing Genre

In [16]:
```
df_genre = pd.read_csv('SpotifyFeatures.csv')
df_genre.head(5)
```

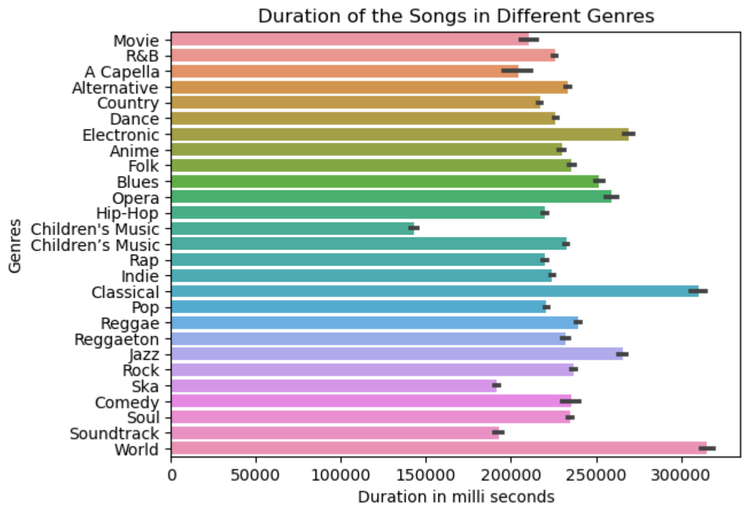
Out[16]:

	genre	artist_name	track_name	track_id	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
0	Movie	Henri Salvador	C'est beau de faire un Show	0BRjO6ga9RKCKjfDqeFgWV	0	0.611	0.389	99373	0.910	0.000	C#	0.3460	-1.828	Major	0.0525	166.969	4/4	0.814
1	Movie	Martin & les fées	Perdu d'avance (par Gad Elmaleh)	0BjC1NfoEOOusryehmNudP	1	0.246	0.590	137373	0.737	0.000	F#	0.1510	-5.559	Minor	0.0868	174.003	4/4	0.816
2	Movie	Joseph Williams	Don't Let Me Be Lonely Tonight	0CoSDzoNIKCRs124s9uTVy	3	0.952	0.663	170267	0.131	0.000	C	0.1030	-13.879	Minor	0.0362	99.488	5/4	0.368
3	Movie	Henri Salvador	Dis-moi Monsieur Gordon Cooper	0Gc6TVm52BwZD07Ki6tIvf	0	0.703	0.240	152427	0.326	0.000	C#	0.0985	-12.178	Major	0.0395	171.758	4/4	0.227
4	Movie	Fabien Nataf	Ouverture	0IusIXpMROHdEPvSl1fTQK	4	0.950	0.331	82625	0.225	0.123	F	0.2020	-21.150	Major	0.0456	140.576	4/4	0.390

Duration fo Songs in different genre using barplot

In [17]:
```
plt.title("Duration of the Songs in Different Genres")
sns.color_palette('rocket', as_cmap = True)
sns.barplot(x = "duration_ms", y ='genre', data = df_genre)
plt.xlabel('Duration in milli seconds')
plt.ylabel('Genres')
```

Out[17]: Text(0, 0.5, 'Genres')



Top 5 Genre by popularity

In [18]:
```
sns.set_style(style='darkgrid')
plt.figure(figsize = (10,5))
famous = df_genre.sort_values('popularity', ascending = False).head(10)
sns.barplot( y='popularity',x='genre', data = famous).set(title='Top 5 Genre by Popularity')
```

Out[18]: [Text(0.5, 1.0, 'Top 5 Genre by Popularity')]