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

```
In [2]: df_tracks = pd.read_csv("C:\Data analytics\Spotify Dataset\tracks.csv")
        df_tracks.head()
```

```
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  2Wo0qPVsbXbSMMT  Carve  13  126893  0  [ZAP]  [3N6WVuEKe4ukfqa7MQ]  2017-10-22  0.645  0.4402  0  -13.338  1  0.0512  0.644  0.5118  0  0.1715  0.7266  113.379  4
1  3ZUWwgPVqQPgZTMC  Captain 3-D - Underwater Remix  0  90350  0  [YqPFCDVbKQwqeSwBeDWrY]  2022-05-01  0.695  0.3630  0  -22.136  1  0.0879  0.757  0.0060  0  0.1115  0.5180  90.000  4
2  3TVXqKX5oeeUWCSNNa  The Less Exploited Remix  0  101640  0  [Dynamic Concept]  [3ZcZ5u1ZeVSM6BZRcAnZnC]  2022-03-21  0.404  0.1775  1  -21.180  1  0.0512  0.694  0.0218  0  0.1115  0.2180  90.000  4
3  3IPHqqNqLYbAaBbbbb  El Poseeedor - Remixxxxxxx  0  27667  0  [Dynamic Concept]  [3ZcZ5u1ZeVSM6BZRcAnZnC]  2022-03-21  0.531  0.0946  7  -27.961  1  0.0504  0.985  0.0045  0.0180  0  0.0970  0.0400  0  4
4  3Ow4SSqJWCAXSaaaqm  Lady of the Evening  0  163000  0  [Dita von Teese]  [3MLDZuJ5yJ6xUFy3xC5kTa]  1922  0.4430  0.1680  3  -18.9930  1  0.0380  0.969  0.1080  0  0.1080  0.3110  90.000  4
```

```
In [3]: df.info()
```

```
Out[3]:
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 586672 entries, 0 to 586671
        Data columns (total 21 columns):
        #   Column            Non-Null Count  Dtype
        ---  ------            --------------  -----
        0   id                586672 non-null  object
        1   name              586601 non-null  object
        2   popularity        586672 non-null  int64
        3   duration_ms       586672 non-null  int64
        4   explicit          586672 non-null  int64
        5   artists           586672 non-null  object
        6   id_artists        586672 non-null  object
        7   release_date      586672 non-null  object
        8   danceability      586672 non-null  float64
        9   energy            586672 non-null  float64
        10  key               586672 non-null  int64
        11  loudness          586672 non-null  float64
        12  mode              586672 non-null  int64
        13  speechiness       586672 non-null  float64
        14  acousticness      586672 non-null  float64
        15  instrumentalness  586672 non-null  float64
        16  liveness          586672 non-null  float64
        17  valence           586672 non-null  float64
        18  tempo             586672 non-null  float64
        19  time_signature    586672 non-null  int64
        20  release_date      586672 non-null  object
        dtypes: float64(11), int64(7), object(3)
        memory usage: 89.5+ MB
```

```
In [4]: sorted_df = df_tracks.sort_values("popularity", ascending= True).head(10)
        sorted_df
```

```
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
0  546130  10x Pitbull ft.Lil Pitbull Fatboy Slim  0  366576  0  ['Pitbull', 'Fatboy Slim', 'Lil Pitbull']  ['0BTJeCYFWGcnVQ1dqunvy', '7nKKaXgRBgQ7bUG...']  1900-02-20  0.595  0.262  8  -17.748  1  0.6320  0.663  0.007510  0  0.0970  0.0400  0.007510  4
1  546222  2pOLCKbWnbbbTBbFzBG  Black Chick  0  186640  0  ['Black Chick']  ['1m0pMFTWSqAJ7mEpQxA']  1949  0.418  0.398  0  -9.930  1  0.0384  0.925  0.000014  0  0.0970  0.0400  0.000014  4
2  546223  5u4Ww4eEEEbcnUgMQ7  Bridal  0  250200  0  ['Bridal']  ['1m0pMFTWSqAJ7mEpQxA']  1949  0.695  0.407  0  -12.236  0  0.0422  0.027  0.000000  0  0.0970  0.0400  0.000000  4
3  546224  3WCbqMNdTNcZgBPBAP  CHOCLO  0  205000  0  ['Bridal']  ['1m0pMFTWSqAJ7mEpQxA']  1949  0.695  0.407  0  -12.236  0  0.0422  0.027  0.000000  0  0.0970  0.0400  0.000000  4
4  546226  0BXSSUAUTPCNnKdFUnuu  ZLUTBUB  0  195733  0  ['ZLUTBUB']  ['1v7Ldf7Ldf7Ldf7Ldf7Ld']  1949  0.390  0.390  2  -11.5  1  0.0391  0.969  0.000000  0  0.0970  0.0400  0.000000  4
5  546228  5yKxQTCNcbHhhYnTHYNB  SHAFTA  0  160427  0  ['Shaft']  ['1m0pMFTWSqAJ7mEpQxA']  1949  0.631  0.246  5  -11.993  1  0.0356  0.951  0.000000  0  0.0970  0.0400  0.000000  4
6  546227  0ropDwCWrSaqQzPQ4qQ  Shaft  0  176707  0  ['Shaft', 'Isaac Hayes']  ['2Wb4QBeeeeeeeeeeeeeW', 'ezRKNVVqqqqqqqqqqqqq']  1949-04-10  0.633  0.317  7  -13.947  1  0.0430  0.492  0.000000  0  0.0970  0.0400  0.000000  4
7  546230  0ec3DUeeUCzEdeeeeeeeu  Shaft  0  161547  0  ['Shaft']  ['1m0pMFTWSqAJ7mEpQxA']  1949  0.529  0.646  0  -9.462  0  0.0418  0.794  0.000000  0  0.0970  0.0400  0.000000  4
8  546231  6ecZPLLuuXcTTzeeQ4q  Sunset (Sounds From The Movie Shaft)  0  177692  0  ['Shaft', 'Isaac Hayes']  ['2Wb4QBeeeeeeeeeeeeeW', '3Ftw4KTgAJ3qPQ7Q7Q...']  1949-09-19  0.66  0.321  0  -13.429  0  0.0330  0.830  0.000000  0  0.0970  0.0400  0.000000  4
9  546234  3Mwwm6cZQYWWKDNyAUT  Sunset (Theme From Shaft) (Instrumental)  0  175976  0  ['Isaac Hayes']  ['2Wb4QBeeeeeeeeeeeeeW']  1949-09-06  0.646  0.341  0  -12.177  1  0.0670  0.466  0.000000  0  0.0970  0.0400  0.000000  4
```

```
In [5]: df_tracks.describe().T.round(2)
```

```
Out[5]:
        count  mean  std  min  25%  50%  75%  max
popularity  586672.00  27.57  18.54  0.00  13.00  27.00  39.00  100.00
duration_ms  586672.00  230911.18  126326.30  30.00  174066.75  214653.50000000000  263867.00000000000  5621218.000000000
explicit  586672.00  0.04  0.19  0.00  0.00  0.00  0.00  1.00
danceability  586672.00  0.58  0.18  0.0000  0.4530  0.5730000000000000  0.6880000000000000  0.9810000000000000
energy  586672.00  0.6158  0.2515  0.0  0.4430  0.6490000000000000  0.8430000000000000  1.0000000000000000
key  586672.00  5.22  3.53  0.0  2.0000  5.0000000000000000  8.0000000000000000  11.0000000000000000
loudness  586672.00  -10.3567  5.0853  -60.0  -12.0610  -8.6330000000000000  -6.4820000000000000  5.3760000000000000
mode  586672.00  0.6597  0.4741  0.0  1.0000  1.0000000000000000  1.0000000000000000  1.0000000000000000
speechiness  586672.00  0.0842  0.1050  0.0000  0.0356  0.0489000000000000  0.0817000000000000  0.9650000000000000
acousticness  586672.00  0.4993  0.3997  0.0000  0.0.0000  0.4200000000000000  0.9960000000000000  0.9960000000000000
instrumentalness  586672.00  0.1341  0.3588  0.0  0.0000  0.0000014  0.0000911  1.000
liveness  586672.00  0.2136  0.1905  0.0  0.0969  0.1390000000000000  0.2760000000000000  1.000
valence  586672.00  0.4192  0.2576  0.0  0.2040  0.4040000000000000  0.6460000000000000  0.9960000000000000
tempo  586672.00  118.4947  29.7918  0.00  96.0000  117.9440000000000  136.3210000000000  249.983000000000
time_signature  586672.00  3.8732  0.4732  0.0  4.0000  4.0000000000000000  4.0000000000000000  5.000
```

```
In [6]: most_popular = df_tracks.query("popularity > 90", inplace = False).sort_values("popularity", ascending = False)
        most_popular[:10]
```

```
Out[6]:
        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  62KKD  4UaWOLhHpaGFzYpJhQ  Ghostbusters  100  196692  0  ['Ghostbusters', 'Ray Parker Jr.']  ['1AP6XZWGWGnqCSC5v...', '2GdxKLuzqRTDnzBuF...']  2021-10-19  0.677  0.696  0  -8.491  1  0.1140  0.32100  0.000000  0  0.0970  0.0400  0.000000  4
1  62M0D  7iPNZDHAHcVYJAACWnCS  Ghostbusters  98  242514  0  ['Ghostbusters']  ['2AKnMnkEf7FGJ4mmYmaoE']  2021-01-08  0.585  0.438  10  -8.761  1  0.0601  0.72100  0.000013  0  0.0970  0.0400  0.000013  4
2  62N0A  1CXHpoAXcAZTfGKNdNJ7  Ain't Nobody in The Jungle  98  132760  0  ['Ain't Nobody in The Jungle']  ['2Ezq1ZAJ5CBDwuDQ3ZgUKE']  2021-01-08  0.774  0.696  4  -8.965  0  0.0613  0.17600  0.000000  0  0.0970  0.0400  0.000000  4
3  62QTM  5Ozwm7mVTBqDPLfq  Ain't Nobody (Live from Soul Train)  97  233627  0  ['Brian Auger']  ['1Ysy4uAfcCZ1Dwe8TSDFZ']  2020-03-20  0.680  0.826  0  -5.407  1  0.0589  0.53120  0.000012  0  0.0970  0.0400  0.000012  4
4  62R1J  6d5Ovx4mdbc7Ynqy1vcdB  Ain't Nobody (Live)  97  185000  0  ['Chaka Khan']  ['2Ezq1ZAJ5CBDwuDQ3ZgUKE']  2022-02-18  0.653  0.526  11  -9.018  0  0.0502  0.11200  0.000000  0  0.0970  0.0400  0.000000  4
5  62S0Q  3SqNHEDSZ5PdYGDVb9Ok  Winding Wire  96  200640  0  ['Fleur']  ['1Ed4Ed7eXfMCQ4qcTQPP']  2020-03-20  0.514  0.730  1  -9.934  1  0.0398  0.00146  0.000095  0  0.0970  0.0400  0.000095  4
6  62W0G  7kHXbcTbWxCDnHGGdh  Leave The Door Open  96  242596  0  ['Bruno Mars', 'Anderson .Paak', 'Silk Sonic']  ['0TulbJFKXYNZfCUuWvNX', '5EHNbCVnKDULNdbGU...']  2021-05-05  0.588  0.616  5  -7.964  1  0.0324  0.16030  0.000000  0  0.0970  0.0400  0.000000  4
7  62X0H  8ESMDdGdvSFDdMFVN  Bad Business  95  194600  0  ['Filter']  ['2kqBQ4mPFUbcwGQPnbD']  2020-08-10  0.756  0.620  8  -7.079  0  0.2520  0.41400  0.000000  0  0.0970  0.0400  0.000000  4
8  62aqW  6GuwFSdDdDwbHMMMB7  Sisters  94  210667  0  ['Irving Berlin']  ['2kqBQ4mPFUbcwGQPnbD']  2022-02-07  0.748  0.463  2  -8.453  1  0.0429  0.00242  0.128000  0  0.0970  0.0400  0.127100  4
9  62b0B  2RJ8CBDNCQNy3MbMBMBQ  Heartbreak Anniversary  94  205640  0  ['Giveon']  ['2MuB8kTxAfMCdCqnpcTQP']  2020-03-27  0.448  0.465  0  -8.904  1  0.0791  0.52400  0.000001  0  0.0970  0.0400  0.000001  4
```

```
In [7]: df_tracks.set_index("release_date", inplace=True)
        df_tracks.index = pd.to_datetime(df_tracks.index)
        df_tracks.head()
```

```
Out[7]:
        id  name  popularity  duration_ms  explicit  artists  id_artists  release_date  danceability  energy  key  loudness  mode  speechiness  acousticness  instrumentalness  liveness  valence  tempo  time_signature
release_date
1922-03-22  3Ow4SqJWCAXSaaaqm  Lady of the Evening  0  163000  0  ['Dita von Teese']  ['3MLDZuJ5yJ6xUFy3xC5kTa']  1922  0.4430  0.1680  3  -18.9930  1  0.0380  0.969  0.1080  0  0.1080  0.3110  90.000  4
1922-09-01  3TVXqKX5oeeUWCSNNa  Captain 3-D - Underwater Remix  0  90350  0  ['Dynamic Concept']  ['3ZcZ5u1ZeVSM6BZRcAnZnC']  2022-05-01  0.695  0.3630  0  -22.136  1  0.0879  0.757  0.0060  0  0.1115  0.5180  90.000  4
1922-09-01  3TVXqKX5oeeUWCSNNa  The Less Exploited Remix  0  101640  0  ['Dynamic Concept']  ['3ZcZ5u1ZeVSM6BZRcAnZnC']  2022-03-21  0.404  0.1775  1  -21.180  1  0.0512  0.694  0.0218  0  0.1115  0.2180  90.000  4
1922-09-01  3IPHqqNqLYbAaBbbbb  El Poseeedor - Remixxxxxxx  0  27667  0  ['Dynamic Concept']  ['3ZcZ5u1ZeVSM6BZRcAnZnC']  2022-03-21  0.531  0.0946  7  -27.961  1  0.0504  0.985  0.0045  0.0180  0  0.0970  0.0400  0  4
1922-09-01  3Ow4SqJWCAXSaaaqm  Lady of the Evening  0  163000  0  ['Dita von Teese']  ['3MLDZuJ5yJ6xUFy3xC5kTa']  1922  0.4430  0.1680  3  -18.9930  1  0.0380  0.969  0.1080  0  0.1080  0.3110  90.000  4
```

```
In [8]: df.shape[0]*df.shape[1].round(2)
```

```
Out[8]:
        artists  [ 'Doctor Beaufort']
        name  [ 'Ain't Nobody (Live)']
        dtype: object
        [ 'Ain't Nobody (Live)']
```

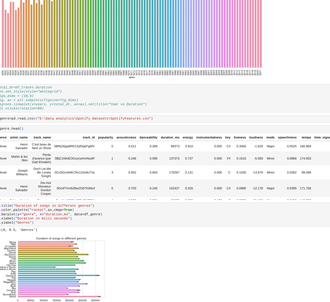
```
In [9]: df_tracks.duration_sec = df_tracks.duration_ms/1000
        df_tracks.drop("duration_ms", axis=1,inplace = True, errors=)
```

```
In [10]: df_tracks.duration.head()
```

```
Out[10]:
        release_date
        1922-03-22  163.000
        1922-09-01  27.667
        1922-09-01  90.350
        1922-09-01  101.640
        1922-09-01  163.000
        Name: duration, dtype: int64
```

```
In [11]: corr_df = df_tracks.drop("key","mode","explicit", axis=1).corr(method="pearson")
         plt.figure(figsize=(14,6))
         heatmap = sns.heatmap(corr_df,annot=True, fmt='.1g', vmax=1, vmin=-1, center= 0, cmap="inferno", linewidths=1, linecolor="black")
         heatmap.set_title('Correlation HeatMap Between Variable')
         heatmap.set_xticklabels(heatmap.get_xticklabels(), rotation=90)
```

```
Out[11]: [Text(0.5, 0, 'popularity'),
          Text(1.5, 0, 'duration_ms'),
          Text(2.5, 0, 'danceability'),
          Text(3.5, 0, 'energy'),
          Text(4.5, 0, 'loudness'),
          Text(5.5, 0, 'speechiness'),
          Text(6.5, 0, 'acousticness'),
          Text(7.5, 0, 'instrumentalness'),
          Text(8.5, 0, 'liveness'),
          Text(9.5, 0, 'valence'),
          Text(10.5, 0, 'tempo'),
          Text(11.5, 0, 'time_signature'),
          Text(12.5, 0, 'duration')]
```




```
In [12]: sample_df = df_tracks.sample(int(0.004*len(df_tracks)))
```

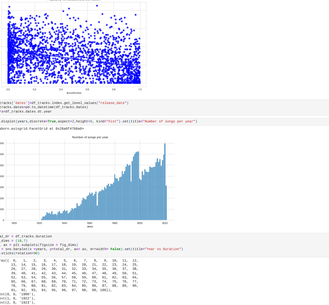
```
In [13]: print(len(sample_df))
```

```
Out[13]:
         2348
```

Progression map between loudness and energy

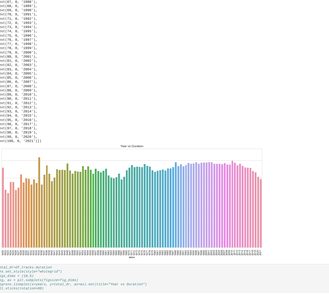
```
In [14]: plt.figure(figsize=(8,6))
         fig_dims = (8,6)
         sns.regplot(data= sample_df, y="loudness", x="energy", color='r').set(title = 'Loudness vs Energy Correlation')
```

```
Out[14]: [Text(0.5, 0, 'Loudness vs Energy Correlation')]
```




```
In [15]: plt.figure(figsize=(8,6))
         fig_dims = (8,6)
         sns.regplot(data= sample_df, y="popularity", x="acousticness", color="b").set(title = "Popularity vs Acoustiness Correlation")
```

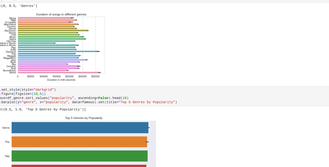
```
Out[16]: [Text(0.5, 0, 'Popularity vs Acoustiness Correlation')]
```




```
In [17]: df_tracks['hours'].head(10)
         df_tracks['hours'].tail(10)
         sorted_df_tracks = df_tracks.sort_values('hours')
         sorted_df_tracks.dropna.df_copy()
```

```
In [18]: sns.displot(years,bins=range(1900,2023,5),height=5, kind="bin", ).set(title="Number of songs per year")
         random.seed(pd.Period(freq='A-DEC', year=1900))
```

```
Out[18]:
         <seaborn.axisgrid.FacetGrid at 0x28b4f47b9a0>
```




```
In [19]: total_df = df_tracks.duration
         fig_dims = (18,7)
         fig = plt.subplots(figsize = fig_dims)
         fig = sns.boxplot(x= years, y=total_df, ax= ax, orient='v', false).set(title="Year vs Duration")
         plt.xticks(rotation=90)
```

```
Out[19]: [Text(0, 0, '1922'), Text(1, 0, '1923'), Text(2, 0, '1924'), Text(3, 0, '1925'), Text(4, 0, '1926'), Text(5, 0, '1927'), Text(6, 0, '1928'), Text(7, 0, '1929'), Text(8, 0, '1930'), Text(9, 0, '1931'), Text(10, 0, '1932'), Text(11, 0, '1933'), Text(12, 0, '1934'), Text(13, 0, '1935'), Text(14, 0, '1936'), Text(15, 0, '1937'), Text(16, 0, '1938'), Text(17, 0, '1939'), Text(18, 0, '1940'), Text(19, 0, '1941'), Text(20, 0, '1942'), Text(21, 0, '1943'), Text(22, 0, '1944'), Text(23, 0, '1945'), Text(24, 0, '1946'), Text(25, 0, '1947'), Text(26, 0, '1948'), Text(27, 0, '1949'), Text(28, 0, '1950'), Text(29, 0, '1951'), Text(30, 0, '1952'), Text(31, 0, '1953'), Text(32, 0, '1954'), Text(33, 0, '1955'), Text(34, 0, '1956'), Text(35, 0, '1957'), Text(36, 0, '1958'), Text(37, 0, '1959'), Text(38, 0, '1960'), Text(39, 0, '1961'), Text(40, 0, '1962'), Text(41, 0, '1963'), Text(42, 0, '1964'), Text(43, 0, '1965'), Text(44, 0, '1966'), Text(45, 0, '1967'), Text(46, 0, '1968'), Text(47, 0, '1969'), Text(48, 0, '1970'), Text(49, 0, '1971'), Text(50, 0, '1972'), Text(51, 0, '1973'), Text(52, 0, '1974'), Text(53, 0, '1975'), Text(54, 0, '1976'), Text(55, 0, '1977'), Text(56, 0, '1978'), Text(57, 0, '1979'), Text(58, 0, '1980'), Text(59, 0, '1981'), Text(60, 0, '1982'), Text(61, 0, '1983'), Text(62, 0, '1984'), Text(63, 0, '1985'), Text(64, 0, '1986'), Text(65, 0, '1987'), Text(66, 0, '1988'), Text(67, 0, '1989'), Text(68, 0, '1990'), Text(69, 0, '1991'), Text(70, 0, '1992'), Text(71, 0, '1993'), Text(72, 0, '1994'), Text(73, 0, '1995'), Text(74, 0, '1996'), Text(75, 0, '1997'), Text(76, 0, '1998'), Text(77, 0, '1999'), Text(78, 0, '2000'), Text(79, 0, '2001'), Text(80, 0, '2002'), Text(81, 0, '2003'), Text(82, 0, '2004'), Text(83, 0, '2005'), Text(84, 0, '2006'), Text(85, 0, '2007'), Text(86, 0, '2008'), Text(87, 0, '2009'), Text(88, 0, '2010'), Text(89, 0, '2011'), Text(90, 0, '2012'), Text(91, 0, '2013'), Text(92, 0, '2014'), Text(93, 0, '2015'), Text(94, 0, '2016'), Text(95, 0, '2017'), Text(96, 0, '2018'), Text(97, 0, '2019'), Text(98, 0, '2020'), Text(99, 0, '2021')]
```



```
In [20]: # total_df=df_tracks.duration
         # fig_dims = (18,7)
         # fig= plt.subplots(figsize=fig_dims)
         # fig= sns.boxplot(x= years, y=total_df, ax= ax, orient='v').set(title="Year vs Duration")
         # plt.xticks(rotation=90)
         # plt.savefig("C:\Data analytics\Spotify Dataset\Spotifyfeatures.png")
```

```
In [21]: df_genre.head()
```

```
Out[21]:
genre  artist_name  track_id  popularity  acousticness  danceability  duration_ms  energy  instrumentalness  key  liveness  mode  speechiness  tempo  time_signature  valence
0  Movie  Henri Salvador  C'est bien ma chance (aznavour)  0  0.611  0.389  99373  0.910  0.000  C#  0.346  1  0.959  Major  0.0525  166.969  4/4  0.814
1  Movie  Martin & Eliot  Waltz - Piano Version  0  0.246  0.590  137373  0.737  0.000  F#  0.151  0.159  Minor  0.0868  174.003  4/4  0.816
2  Movie  Joseph Kosma  Le temps des cerises (1945 / Remastered)  3  0.952  0.663  170267  0.131  0.000  C  0.103  0.268  Major  0.0362  89.488  5/4  0.368
3  Movie  Henri Salvador  Dans mes rêves (Remastered 2014)  0  0.703  0.240  192427  0.326  0.000  C#  0.098  -12.179  Major  0.0395  171.758  4/4  0.227
```

```
In [22]: plt.title("Division of songs in different genres")
         sns.pie(df["genre"], ax,image/free)
         sns.barplot(x="genre", x="duration_ms", data=df_genre)
         plt.xlabel("Genre")
         plt.ylabel("Duration")
         plt.show()
```

```
Out[22]: Text(0.5, 0, 'Genre')
```



```
In [23]: sns.set_style(style="darkgrid")
         plt.figure(figsize=(8,6))
         famous_df_genre.sort_values("popularity", ascending=False).head(10)
         sns.barplot("genre", x="popularity", data=famous).set(title="Top 5 Genres by Popularity")
```

```
Out[23]: [Text(0.5, 0, 'Top 5 Genres by Popularity')]
```