ROLL NO:5840

DATA VISUALIZATION PROJECT

DATASET: Top Hits Of Spotify From 2009 To 2019

FUNCTIONS USED FOR ANALYZING DATA:

1.dataset.shape
2.dataset.columns
3.dataset.info()
4.dataset.isnull()
5.dataset.isnull().sum()
6.dataset.describe()

PLOTS USING SEABORN LIBRARY :

1.Line Plot
2.Bar Plot
3.Kde Plot
4.Joint Plot
5.Violin Plot
6.Box Plot
7.Strip Plot
8.Swarm Plot

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2000 to 2019')
print(df)
```

```
              artist                                                song  \
0       Fatman Scoop                                        Be Faithful
1            Madonna  4 Minutes (feat. Justin Timberlake & Timbaland)
2     Alexandra Burke                                         Hallelujah
3     Black Eyed Peas                                  I Gotta Feeling
4           Flo Rida                                      Right Round
...              ...                                               ...
1101       Sam Smith           Dancing With A Stranger (with Normani)
1102     Post Malone                                          Circles
1103         t.A.T.u.                          All The Things She Said
1104      Will Smith                                           Switch
1105             NSG                                          Options

      duration_ms  explicit  year  popularity  danceability  energy  key  \
0          164506      True  2009          49         0.649   0.713    7
1          189693     False  2009          71         0.753   0.931    2
2          217826     False  2009          63         0.177   0.425    2
3          289133     False  2009          80         0.743   0.766    0
4          204640     False  2009          74         0.720   0.672    7
...           ...       ...   ...         ...           ...     ...  ...
1101       171029     False  2019          75         0.741   0.520    8
1102       215280     False  2019          85         0.695   0.762    0
1103       214440      True  2020          39         0.527   0.834    5
1104       197666     False  2020          34         0.873   0.900    5
1105       240081      True  2020          57         0.836   0.621    1

      loudness  mode  speechiness  acousticness  instrumentalness  liveness  \
0       -6.488     1       0.2950      0.000787          0.000000    0.3180
1       -4.922     1       0.0652      0.009940          0.006960    0.2340
2       -6.211     0       0.0291      0.654000          0.000000    0.1950
3       -6.375     1       0.0265      0.087300          0.000000    0.5090
4       -6.852     1       0.0551      0.009000          0.000000    0.2320
...        ...   ...          ...           ...               ...       ...
1101    -7.513     1       0.0656      0.450000          0.000002    0.2220
1102    -3.497     1       0.0395      0.192000          0.002440    0.0863
1103    -5.767     0       0.0474      0.041100          0.005990    0.1050
1104    -4.325     0       0.1800      0.010700          0.000502    0.5560
1105    -4.684     0       0.0894      0.389000          0.000092    0.1040

      valence    tempo                        genre
0      0.6290  101.129                      hip hop
1      0.7670  113.029                          pop
2      0.0942  182.571                          pop
3      0.6100  127.960                 hip hop, pop
4      0.7050  124.986                 hip hop, pop
...       ...      ...                          ...
1101   0.3470  102.998                          pop
1102   0.5530  120.042                      hip hop
1103   0.3810  179.920                          pop
1104   0.4780  102.516                 hip hop, pop
1105   0.7620  101.993  World/Traditional, hip hop

[1106 rows x 18 columns]
```

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
df.shape
```

```
(1106, 18)
```

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
df.columns
```

```
Index(['artist', 'song', 'duration_ms', 'explicit', 'year', '
       'danceability', 'energy', 'key', 'loudness', 'mode', '
       'acousticness', 'instrumentalness', 'liveness', 'valen
       'genre'],
      dtype='object')
```

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1106 entries, 0 to 1105
Data columns (total 18 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   artist            1106 non-null   object 
 1   song              1106 non-null   object 
 2   duration_ms       1106 non-null   int64  
 3   explicit          1106 non-null   bool   
 4   year              1106 non-null   int64  
 5   popularity        1106 non-null   int64  
 6   danceability      1106 non-null   float64
 7   energy            1106 non-null   float64
 8   key               1106 non-null   int64  
 9   loudness          1106 non-null   float64
 10  mode              1106 non-null   int64  
 11  speechiness       1106 non-null   float64
 12  acousticness      1106 non-null   float64
 13  instrumentalness  1106 non-null   float64
 14  liveness          1106 non-null   float64
 15  valence           1106 non-null   float64
 16  tempo             1106 non-null   float64
 17  genre             1106 non-null   object 
dtypes: bool(1), float64(9), int64(5), object(3)
memory usage: 148.1+ KB
```

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
df.isnull()
```

	artist	song	duration_ms	explicit	year	popularity	danceability	energy	key	loudness
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
1101	False	False	False	False	False	False	False	False	False	False
1102	False	False	False	False	False	False	False	False	False	False
1103	False	False	False	False	False	False	False	False	False	False
1104	False	False	False	False	False	False	False	False	False	False
1105	False	False	False	False	False	False	False	False	False	False

1106 rows × 18 columns

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
df.isnull().sum()
```

```
artist              0
song                0
duration_ms         0
explicit            0
year                0
popularity          0
danceability        0
energy              0
key                 0
loudness            0
mode                0
speechiness         0
acousticness        0
instrumentalness    0
liveness            0
valence             0
tempo               0
genre               0
dtype: int64
```

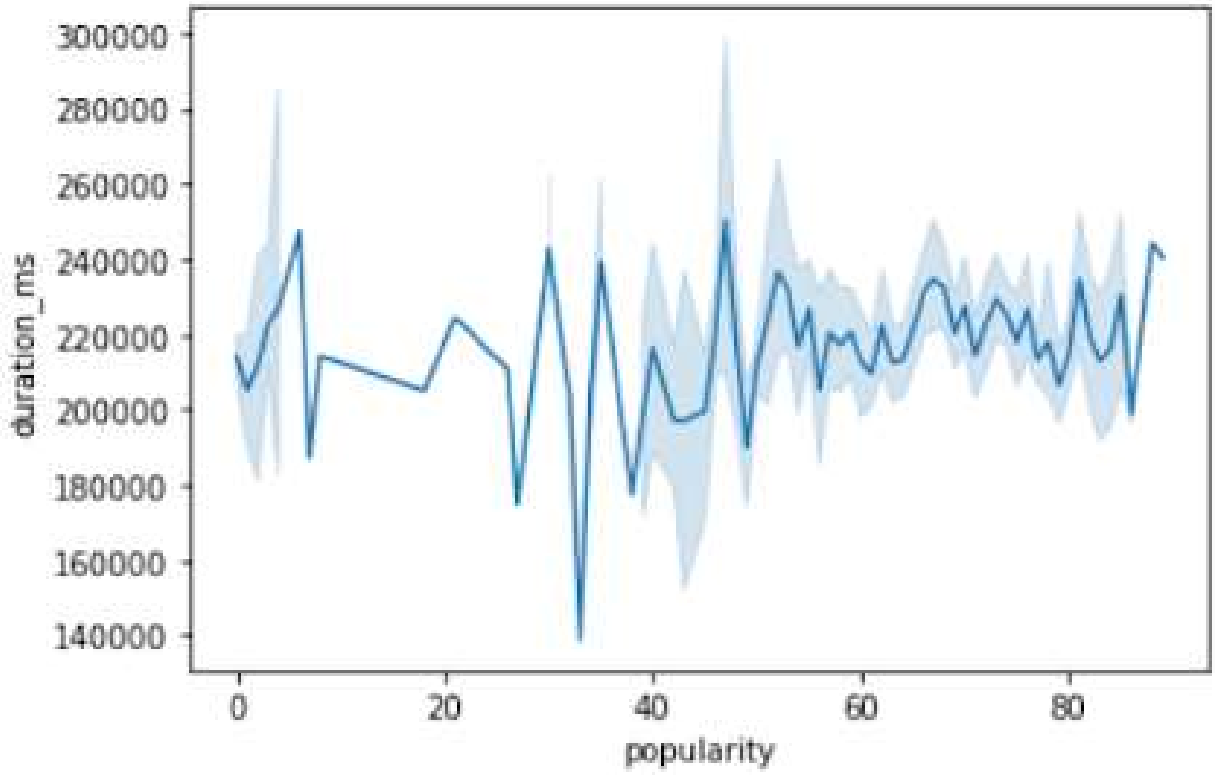
```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
df.describe()
```

	duration_ms	year	popularity	danceability	energy	key
count	1106.000000	1106.000000	1106.000000	1106.000000	1106.000000	1106.000000
mean	219611.028029	2014.051537	59.550633	0.665689	0.714273	5.422242
std	35978.227356	3.124572	24.943405	0.133177	0.150741	3.686813
min	113000.000000	2009.000000	0.000000	0.177000	0.054900	0.000000
25%	198053.750000	2011.000000	57.000000	0.585000	0.618250	2.000000
50%	216013.000000	2014.000000	68.000000	0.674000	0.732000	6.000000
75%	237210.000000	2017.000000	75.000000	0.753000	0.830000	9.000000
max	484146.000000	2020.000000	89.000000	0.964000	0.985000	11.000000

1.LINE PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.lineplot(x='popularity',y='duration_ms',data=df)
```

```
<AxesSubplot:xlabel='popularity', ylabel='duration_ms'>
```

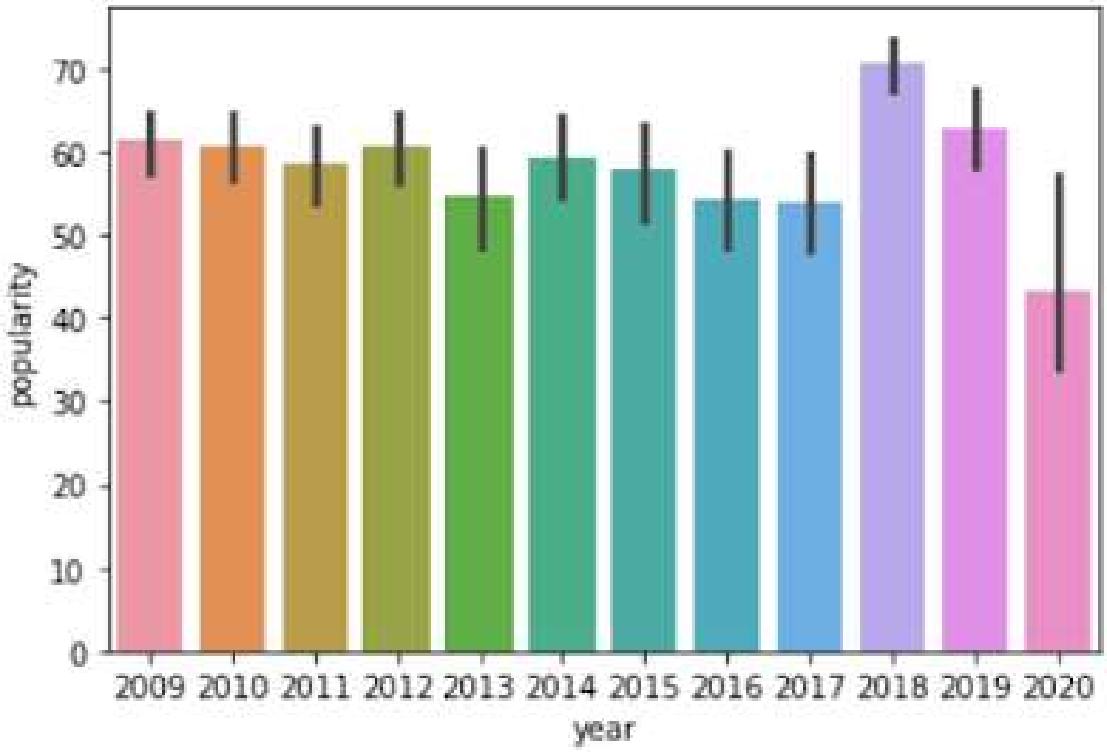


In the above line plot , the x-axis represents the popularity of the songs whereas the y-axis represents the duration of song played in units of milliseconds. The lowest played duration of songs is 113000 ms in between the range of popularity 20 - 40 whereas the highest played duration of songs is 484146 in between the range of popularity 40 - 60.

2.BAR PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.barplot(x='year',y='popularity',data=df)
```

```
<AxesSubplot:xlabel='year', ylabel='popularity'>
```

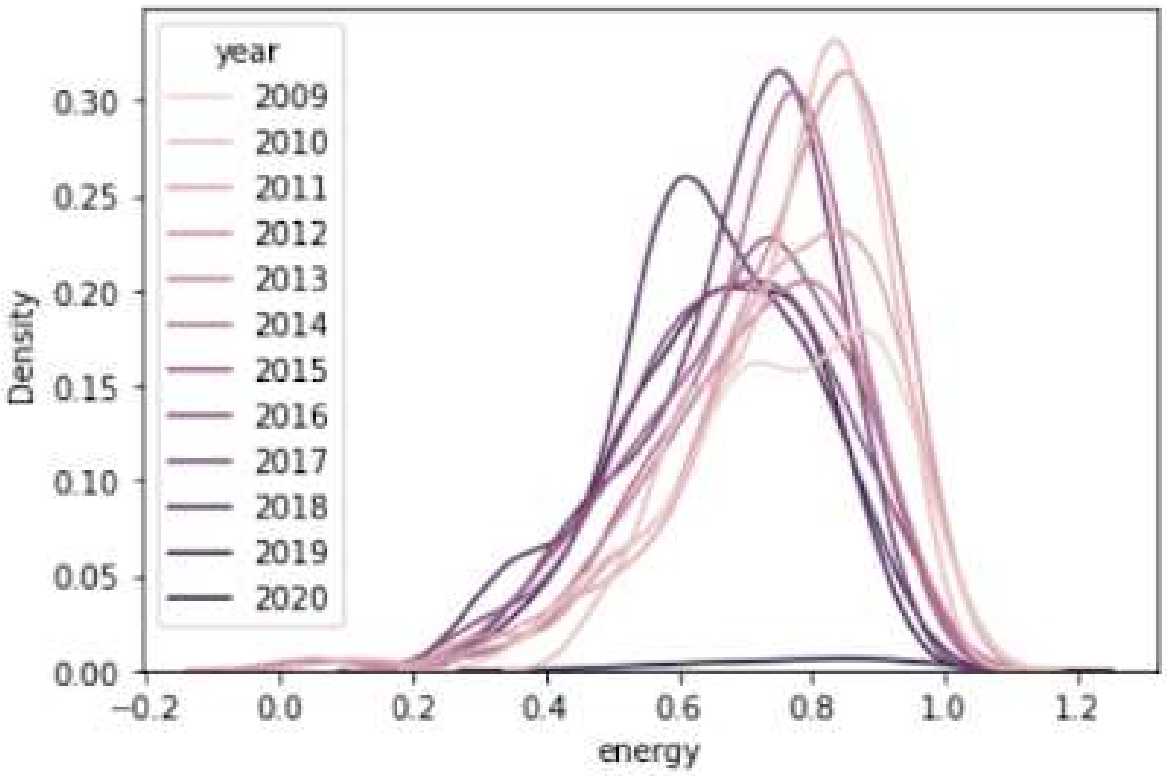


In the above bar plot , the x-axis represents the years in which the songs were released whereas the y-axis represents the popularity of the songs. The popularity of songs is highest in the year 2018 whereas the popularity of the songs is lowest in the year 2020.

3.KDE PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.kdeplot(x='energy',hue='year',data=df)
```

```
<AxesSubplot:xlabel='energy', ylabel='Density'>
```

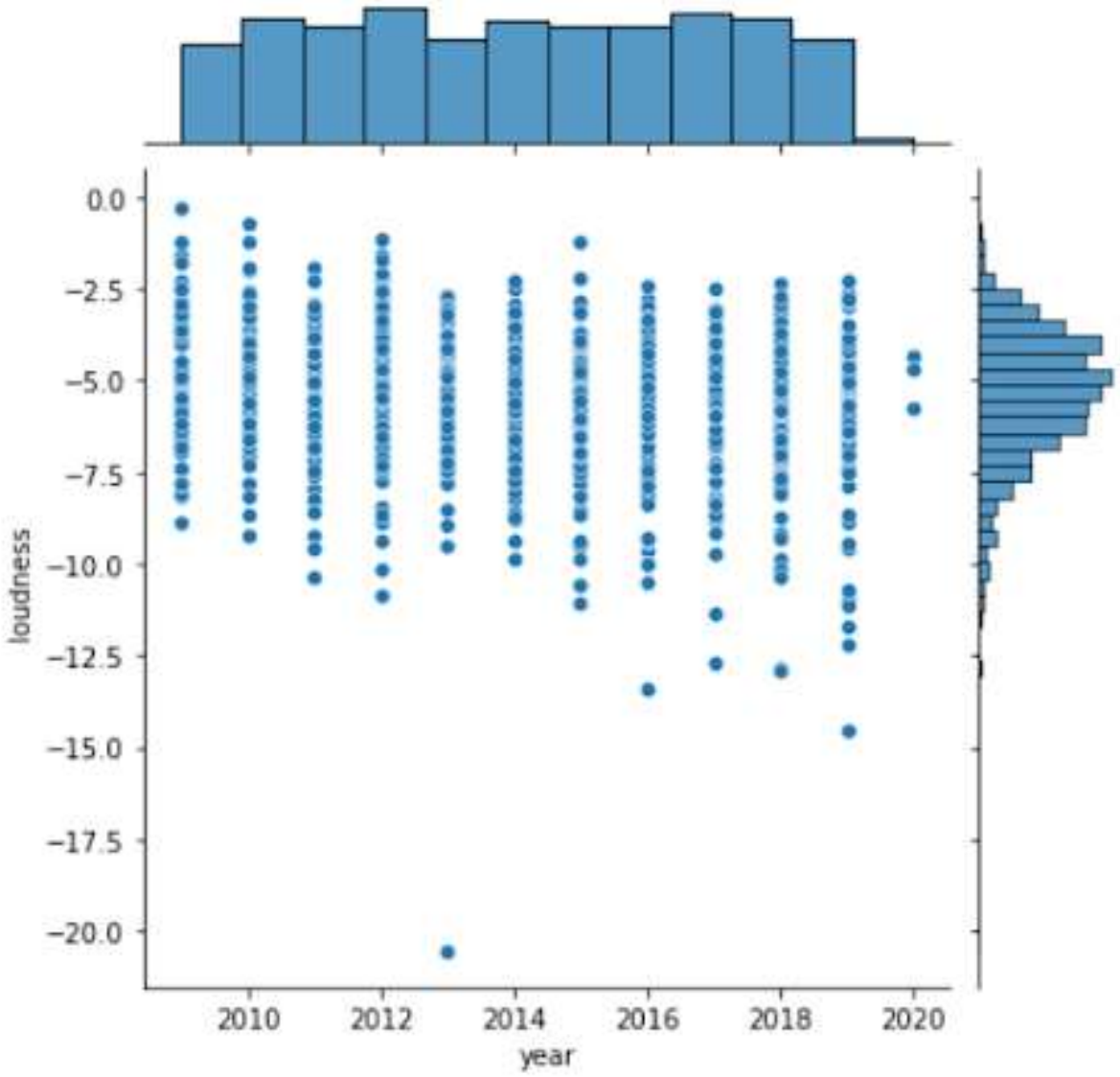


In the above kde plot , the x-axis represents the energy level of the songs whereas the y-axis represents the density of count . The lowest energy of songs is 0.054900 in the year 2022 whereas the highest energy of songs is 0.985000 in the year 2010.

4.JOINT PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.jointplot(x='year',y='loudness',data=df,kind='scatter')
```

```
<seaborn.axisgrid.JointGrid at 0x2a3b9fce280>
```

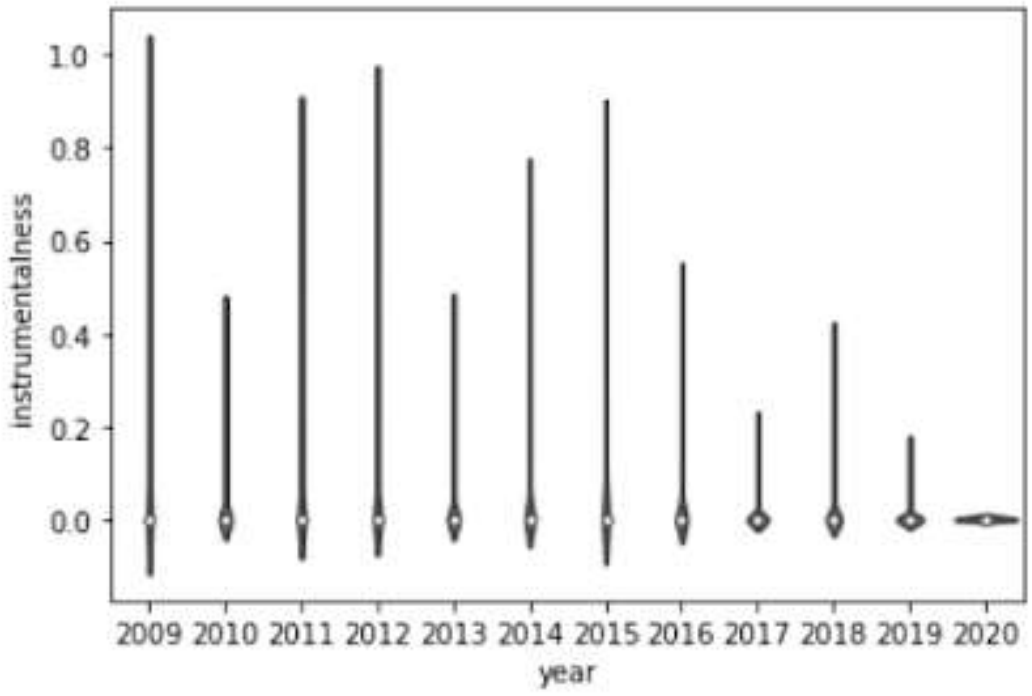


In the above joint plot , the x-axis represents the years in which the songs were released whereas the y-axis represents the measure of loudness of the songs. The loudness of songs is highest in the year 2009 whereas the loudness of the songs is lowest in the year 2020.

5.VIOLIN PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.violinplot(x='year',y='instrumentalness',data=df,linewidth=1.9)
```

```
<AxesSubplot:xlabel='year', ylabel='instrumentalness'>
```

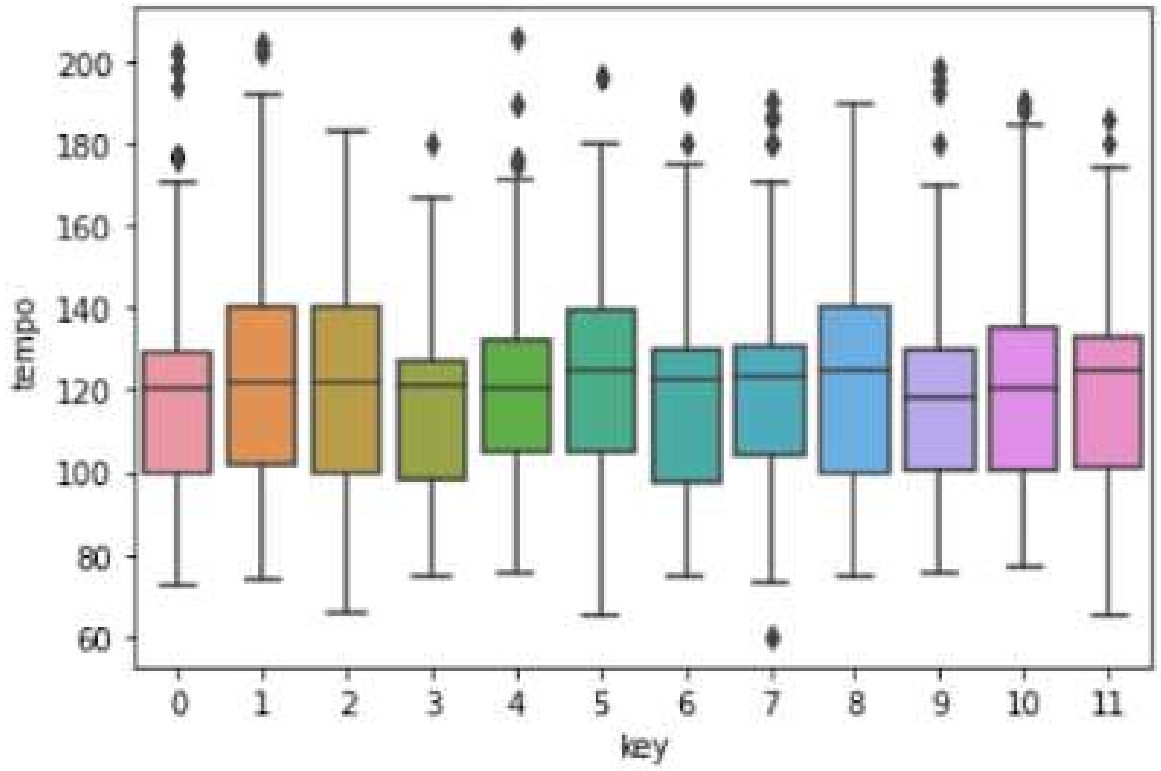


In the above violin plot , the x-axis represents the years in which the songs were released whereas the y-axis represents the instrumentalness of the songs. The instrumentalness of songs is highest in the year 2009 whereas the instrumentalness of the songs is lowest in the year 2020.

6.BOX PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.boxplot(x='key',y='tempo',data=df)
```

```
<AxesSubplot:xlabel='key', ylabel='tempo'>
```

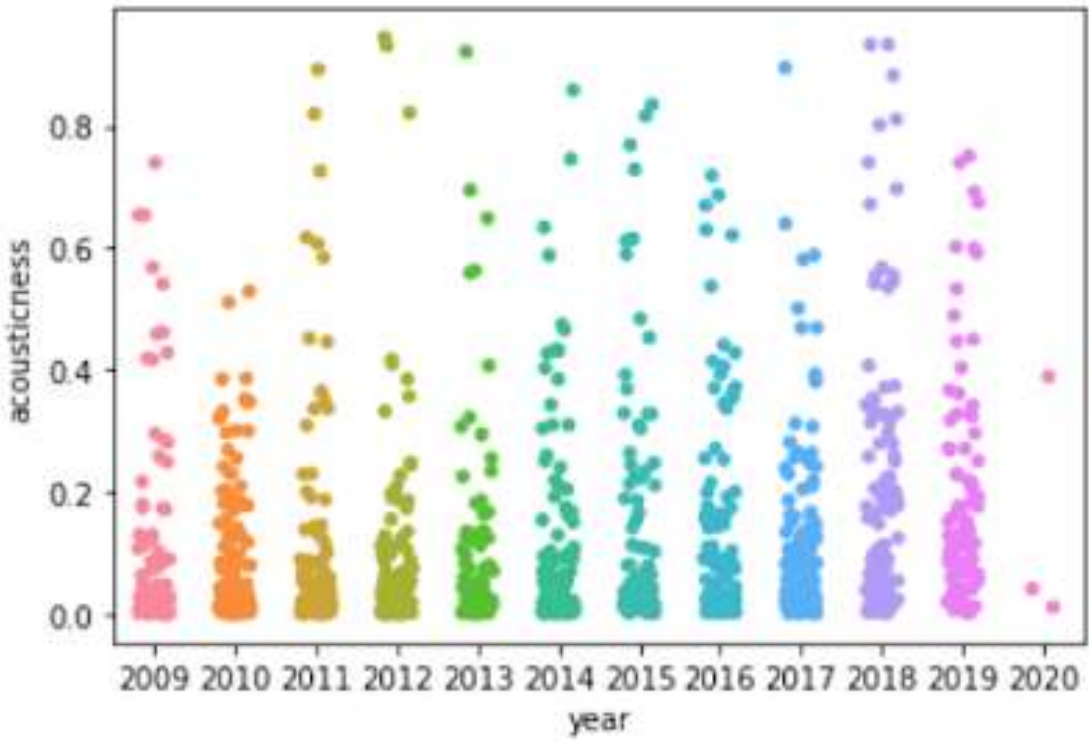


In the above box plot , the x-axis represents the key of the songs whereas the y-axis represents the tempo of the songs.The tempo is lowest is 60.019000 when the key is 7 whereas the tempo is highest is 205.570000 when the key is 4 .

7.STRIP PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.stripplot(x='year',y='acousticness',data=df,jitter='0.2')
```

```
<AxesSubplot:xlabel='year', ylabel='acousticness'>
```

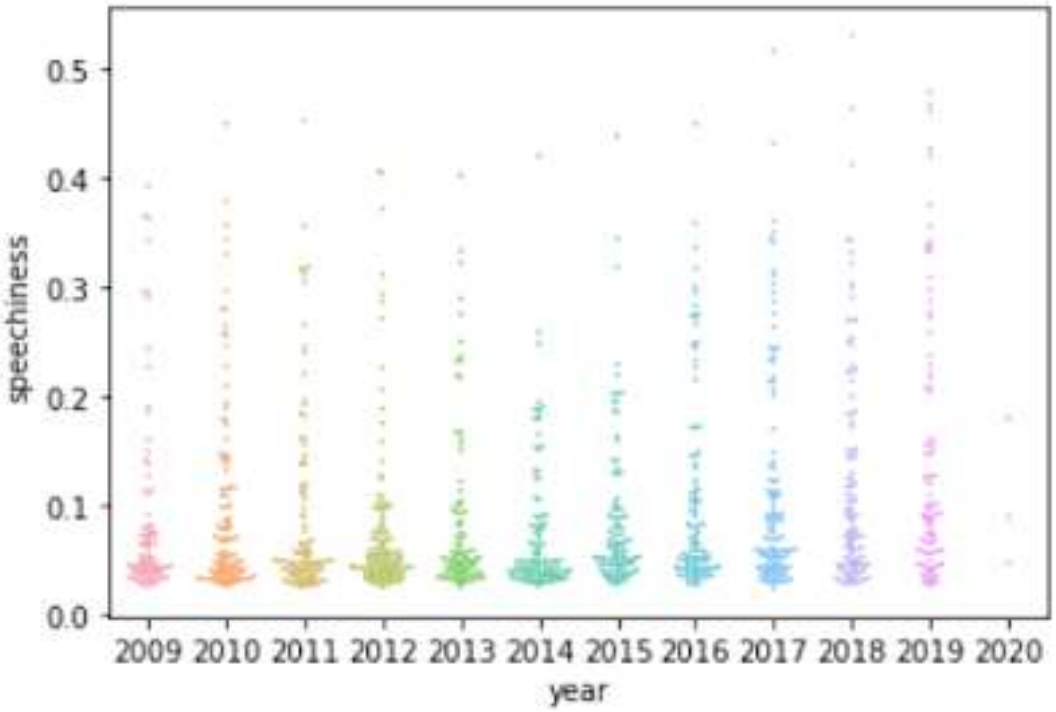


In the above strip plot , the x-axis represents the years in which the songs were released whereas the y-axis represents the acousticness of the songs.The acousticness of songs is highest in the year 2018 whereas the acousticness of the songs is lowest in the year 2020.

8.SWARM PLOT:

```
import seaborn as sns
df=sns.load_dataset('top hits of spotify from 2009 to 2019')
sns.swarmplot(x ='year', y = 'speechiness', data = df, size =1.5)
```

```
<AxesSubplot:xlabel='year', ylabel='speechiness'>
```



In the above swarm plot , the x-axis represents the years in which the songs were released whereas the y-axis represents the speechiness of the songs.The speechiness of songs is mostly found in the year 2014 whereas the speechiness of the songs is least found in the year 2020.