CA 1

CSE353

EDA Project

Submitted By : Royal Chaudhary
Reg No. : 12016265
Section: K20CH

Introduction to dataset.

This dataset is about The unpopular songs on Spotify.

- **Basic** Structure

```
In [13]: df = pd.read_csv("./unpopular_songs.csv")
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10877 entries, 0 to 10876
Data columns (total 17 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   danceability      10877 non-null  float64
 1   energy            10877 non-null  float64
 2   key               10877 non-null  int64
 3   loudness          10877 non-null  float64
 4   mode              10877 non-null  int64
 5   speechiness       10877 non-null  float64
 6   acousticness      10877 non-null  float64
 7   instrumentalness  10877 non-null  float64
 8   liveness          10877 non-null  float64
 9   valence           10877 non-null  float64
 10  tempo             10877 non-null  float64
 11  duration_ms       10877 non-null  int64
 12  explicit          10877 non-null  bool
 13  popularity        10877 non-null  int64
 14  track_name        10877 non-null  object
 15  track_artist      10877 non-null  object
 16  track_id          10877 non-null  object
dtypes: bool(1), float64(9), int64(4), object(3)
memory usage: 1.3+ MB
```

This dataset contains the audio characteristics of over 10,000 unpopular songs and 16 columns that are as follows:

1 danceability
2 energy
3 loudness
4 mode
5 speechiness
6 acousticness

7 instrumentalness
8 liveness
9 valence
10 tempo
11 . duration ms
12 explicit
13 popularity
14 track name
15 track artist
16 track id

As for the null values, we have zero null values so we can easily work on this dataset without worrying about missing or improper values.

```
In [15]: df.isnull().sum()

Out[15]: danceability        0
         energy              0
         key                 0
         loudness            0
         mode                0
         speechiness         0
         acousticness        0
         instrumentalness    0
         liveness            0
         valence             0
         tempo               0
         duration_ms         0
         explicit            0
         popularity          0
         track_name          0
         track_artist        0
         track_id            0
         dtype: int64
```

Why this dataset

The dataset I choose belongs to the music domain. I choose this dataset because With so many songs so readily available to them, music lovers tend to be very discriminating and have rather short attention spans, which means picking up on similar patterns, lyrics or recurring themes between two or more of your songs, could easily turn them off, regardless of how subtle the similarities are.

I, personally am a music lover, so I listen to dozens of songs on daily basis and have found in past 4- 5 years that there still do exist so many great songs which are unheard by a large portion of human population. So, to find what categorises songs as popular or unpopular I chose this dataset out of my own curiosity.

For instance here's how top 5 rows of this dataset would look like:

```
In [16]: df.head()
Out[16]:
```

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	explicit	popularity	track
0	0.530	0.770	4	-6.633	0	0.0389	0.284	0.501000	0.744	0.623	120.144	225696	False	2	No F
1	0.565	0.730	1	-6.063	1	0.0730	0.365	0.000000	0.237	0.511	130.026	158093	False	2	W
2	0.427	0.546	4	-8.727	1	0.0849	0.539	0.015200	0.368	0.435	78.345	167262	False	2	
3	0.421	0.531	7	-5.516	1	0.0262	0.706	0.000208	0.110	0.383	85.080	236832	False	2	Afte
4	0.537	0.804	8	-7.378	0	0.1570	0.379	0.000489	0.323	0.543	139.950	239400	False	2	

What are the insights we will be finding here

We'd mainly analyse the data by finding patterns or by clusterizing it to identify the different types of unpopular songs.

We'd be comparing all sorts of data based on loudness, popularity, danceability, tempo etc. with the help of charts and graphs to analyse the factors behind the unpopularity of the songs on the dataset.

Tools Used

- Pandas
- NumPy
- Seaborn
- matplotlib