Top 10,000 Songs on Spotify

1960-Now

Ali Bakhshesh

Analysis

1.Data Exploration

This dataset has recorded 10000 Songs on spotify science 1960 until now.

The dataset consists of following attributes:

1. **Index**
2. **Track URI**
3. **Track Name**
4. **Artist URI**
5. **Artist Name**
6. **Album URI**
7. **Album Name**
8. **Album Artist URI**
9. **Album Artist Name**
10. **Album Release Date**
11. **Album Image URL**
12. **Disk Number**
13. **Track Number**
14. **Track Preview URL**
15. **ISRC**
16. **Added By**
17. **Added At**
18. **Artist Genres**
19. **Label**
20. **Copyright**
21. **Time Duration:** The length of time a pitch, or tone, is sounded.
22. **Explicit:** One that has curse words or language or art that is generally deemed sexual, violent, or offensive in nature.
23. **Popularity:** This is the main feature in our analysis and in most of the tests we study about popularity of various types of music.
24. **Danceability:** Measured using a mixture of song features such as beat strength, tempo stability, and overall tempo.
25. **Energy:** The energy concept in music is usually used in relation to sound power, the amplitude of sound, etc.
26. **Loudness:** A way to measure audio levels based on the way humans perceive sound.
27. **Mode:** the term mode or *modus* is used in a number of distinct senses, depending on context.
28. **Speechiness:** If the speechiness of a song is above 0.66, it is probably made of spoken words, a score between 0.33 and 0.66 is a song that may contain both music and words, and a score below 0.33 means the song does not have any speech.
29. **Acousticness:** This value describes how acoustic a song is. A score of 1.0 means the song is most likely to be an acoustic one.

30. **Instrumentalness:** This value represents the amount of vocals in the song. The closer it is to 1.0, the more instrumental the song is.
31. **Liveness:** This value describes the probability that the song was recorded with a live audience.
32. **Valance:** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g., happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g., sad, depressed, angry).
33. **Tempo:** In musical terminology, tempo also known as Beats per minute, is the speed or pace of a given piece.
34. **Time Signature:** Indicate how many beats are in each measure of a piece of music, as well as which note value is counted as a beat.

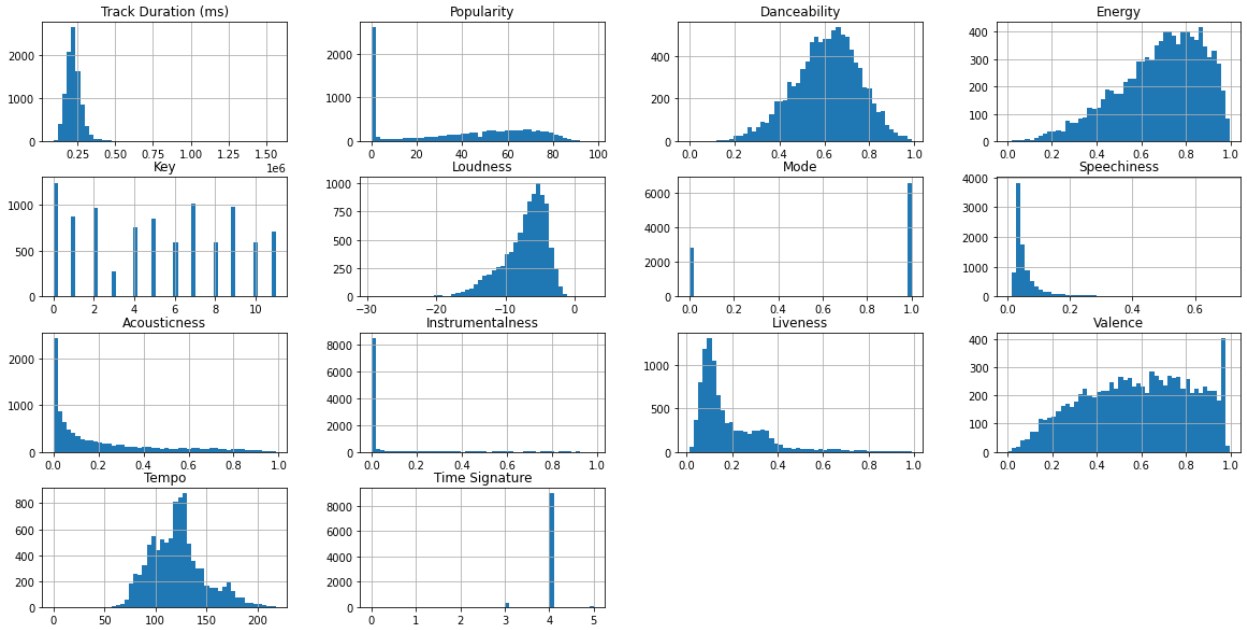
The first 20 attributes are not explained because I didn't use them in analysis and doing statistical tests.

As it was mentioned before first 20 columns are not so useful so they were eliminated so we have a general view of the remaining data in the blow.

	Track Duration (ms)	Popularity	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Time Signature
0	216270	0	0.617	0.872	8.0	-12.305	1.0	0.0480	0.015800	0.112000	0.4080	0.504	111.458	4.0
1	237120	64	0.825	0.743	2.0	-5.995	1.0	0.1490	0.014200	0.000021	0.2370	0.800	127.045	4.0
2	312533	56	0.677	0.665	7.0	-5.171	1.0	0.0305	0.560000	0.000001	0.3380	0.706	74.981	4.0
3	233400	42	0.683	0.728	9.0	-8.920	1.0	0.2590	0.568000	0.000051	0.0384	0.833	75.311	4.0
4	448720	0	0.319	0.627	0.0	-9.611	1.0	0.0687	0.675000	0.000073	0.2890	0.497	85.818	4.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9994	165800	79	0.623	0.727	11.0	-5.570	0.0	0.0562	0.184000	0.000020	0.3090	0.400	125.975	4.0
9995	176640	17	0.720	0.841	9.0	-6.373	1.0	0.0340	0.000354	0.011200	0.3380	0.767	130.978	4.0
9996	227619	62	0.719	0.806	9.0	-6.802	0.0	0.0389	0.000132	0.088900	0.3610	0.626	123.037	4.0
9997	153442	87	0.534	0.855	1.0	-4.923	0.0	0.1830	0.060700	0.000263	0.3460	0.420	122.060	4.0
9998	166266	69	0.744	0.620	5.0	-7.930	1.0	0.2460	0.214000	0.001160	0.1030	0.711	128.103	4.0

	Track Duration (ms)	Popularity	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Time Signature
count	9.446000e+03	9446.000000	9446.000000	9446.000000	9446.000000	9446.000000	9446.00000	9446.000000	9446.000000	9446.000000	9446.000000	9446.000000	9446.000000	9446.000000
mean	2.255781e+05	38.538747	0.608168	0.685176	5.176371	-7.197891	0.69765	0.065562	0.204485	0.026908	0.185229	0.583383	121.562274	3.961571
std	5.413862e+04	29.627115	0.145697	0.189766	3.583477	3.248467	0.45930	0.061830	0.245704	0.116918	0.147560	0.238876	26.315202	0.251001
min	9.122600e+04	0.000000	0.000000	0.000020	0.000000	-29.368000	0.00000	0.000000	0.000003	0.000000	0.012000	0.000000	0.000000	0.000000
25%	1.933232e+05	0.000000	0.515000	0.564000	2.000000	-8.903000	0.00000	0.033200	0.018225	0.000000	0.089200	0.396000	102.602250	4.000000
50%	2.204465e+05	44.000000	0.617000	0.713000	5.000000	-6.452000	1.00000	0.043100	0.092600	0.000005	0.128000	0.595000	120.505000	4.000000
75%	2.512498e+05	65.000000	0.710000	0.836000	8.000000	-4.862000	1.00000	0.067900	0.308750	0.000487	0.245000	0.780000	134.479500	4.000000
max	1.561133e+06	98.000000	0.988000	0.997000	11.000000	2.769000	1.00000	0.711000	0.987000	0.985000	0.989000	0.995000	217.913000	5.000000

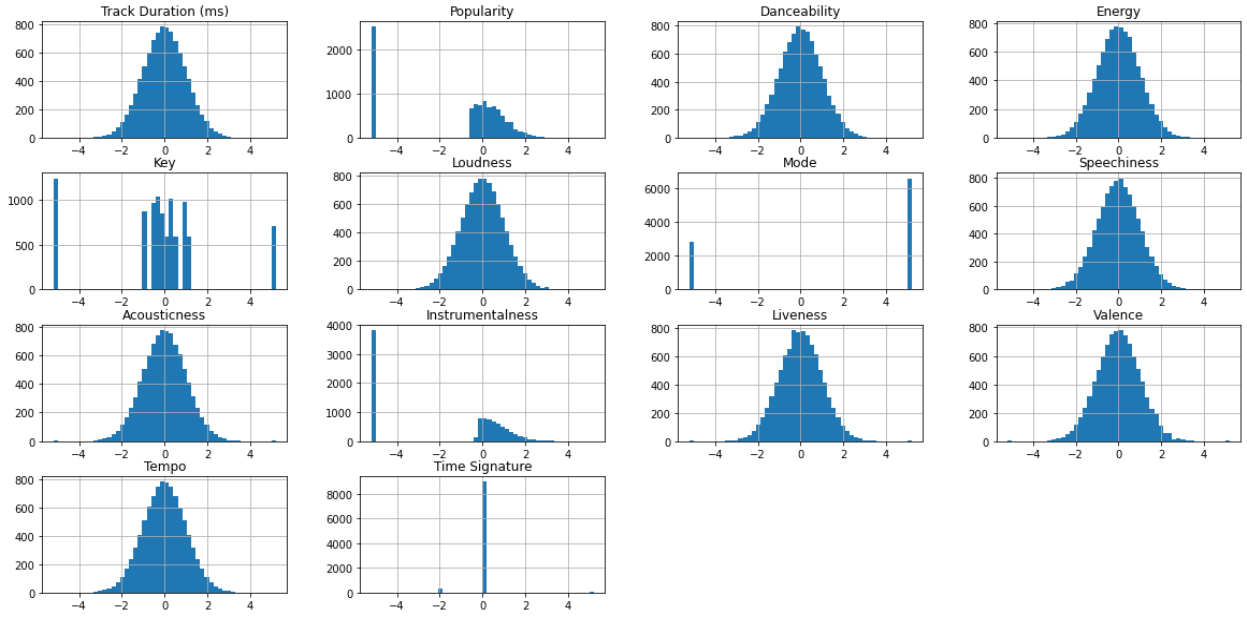
And here there is a histogram which shows the distribution of our numerical data.



2. Data preprocessing:

The data preprocessing and cleaning include following steps:

- **Omitting null values**: All the rows with null values could be eliminated or we could fill them with some proper values, but we did the first one because the number null values was not so huge compared to rest of the data.
- **Removing some columns**: As I said before some of the attributes of data specially String attributes were not useful in our analysis, so we dropped them.
- **Normalizing the data**: Many statistical tests require data to be normally distributed so we did that in the next step. Here is the new histogram of distribution of numerical data.



3. Statistical Tests and Analysis

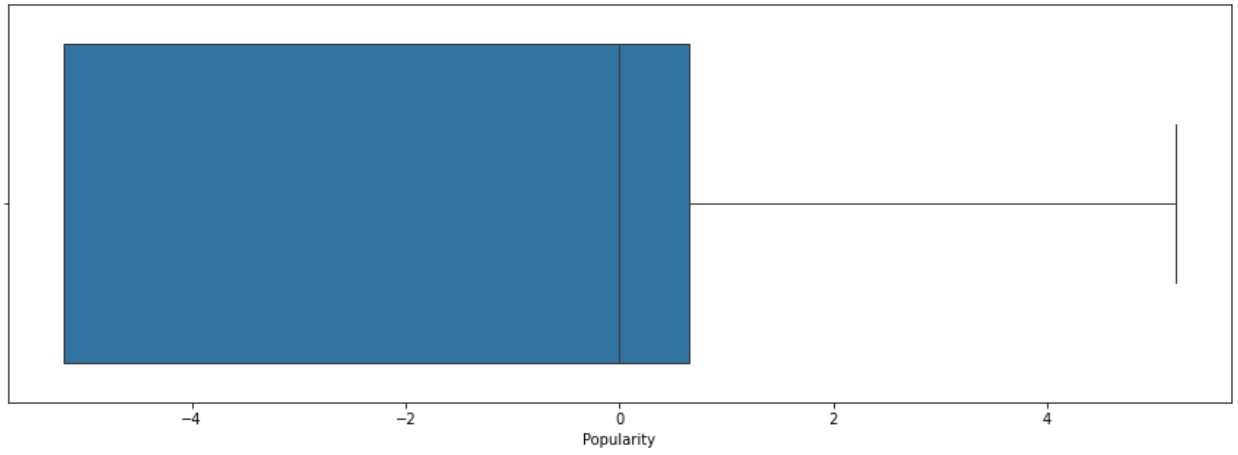
Question1: Music in which tempo range are more likely to be listened by people?

To answer this question, we divided data into two categories. High and Low tempo. Then we compared the average popularity in each group using *Two Sample t-test*.

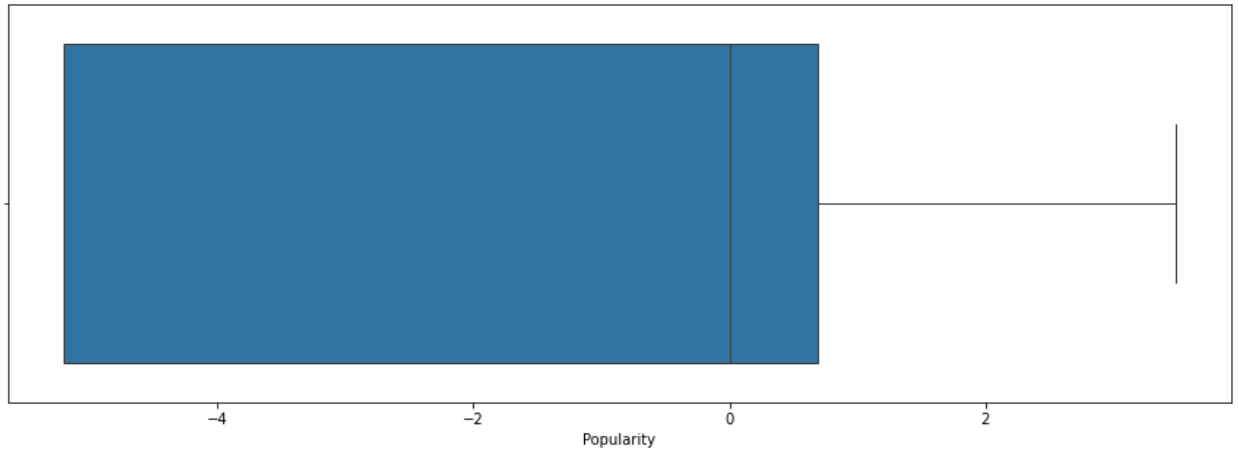
The null hypothesis for this test is that the average popularity for each category is equal it means that we assume that tempo has no effect on popularity of songs.

NOTE: α is considered equal to 0.05 in all tests.

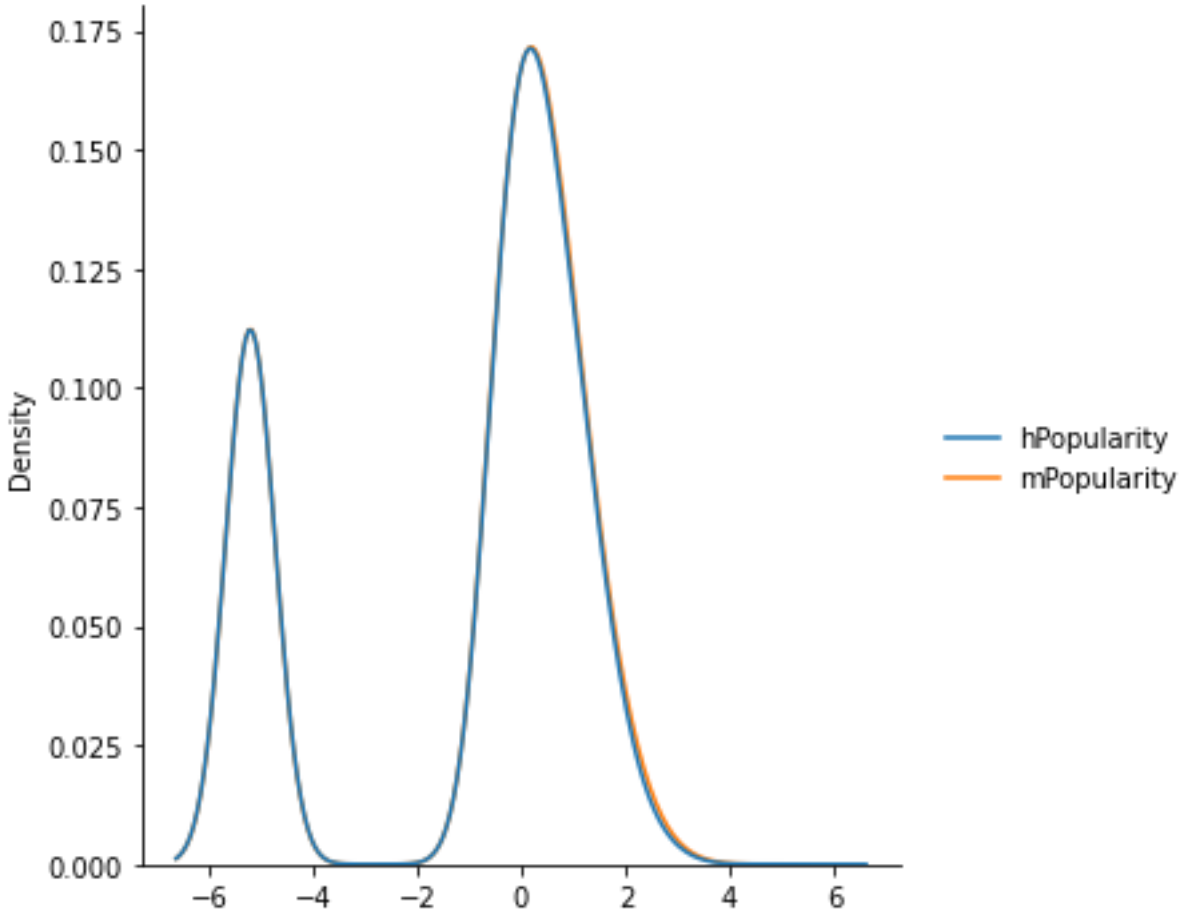
Our ***P_VALUE*** in this test is approximately equal to 0.43. So, we can say that the null hypothesis has been accepted. It means that high or low tempo has no effect on popularity.



Popularity of low tempo songs box plot



Popularity of high tempo songs box plot



High and low tempo songs distribution plot

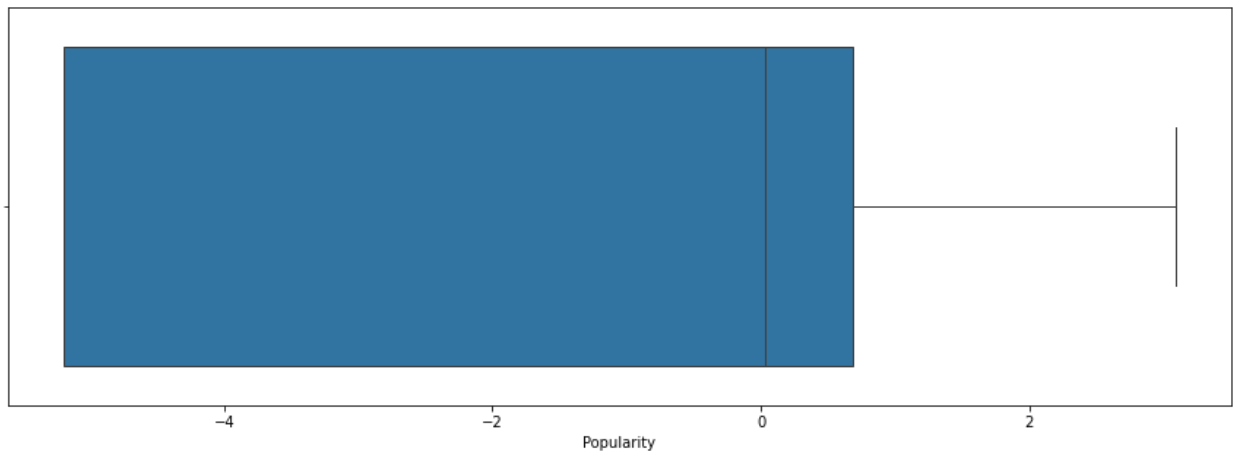
Question2: Do people prefer to listen to long music or short music?

In this question we have again divided data into two categories Long and Short song and we want to see whether time duration of songs has effect on popularity or not?

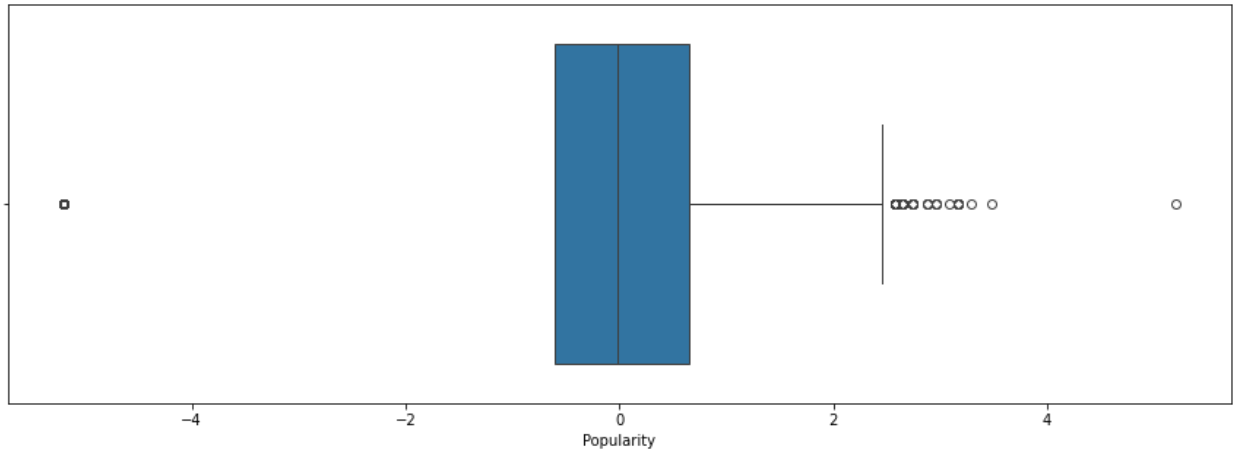
The null hypothesis is that average popularity of two categories is equal. We will again use *two sample t-test* to check if our null hypothesis is right or not.

The ***P_VALUE*** in this test is almost equal to 10^-5. So, it seems that one of the two categories is more popular than other one.

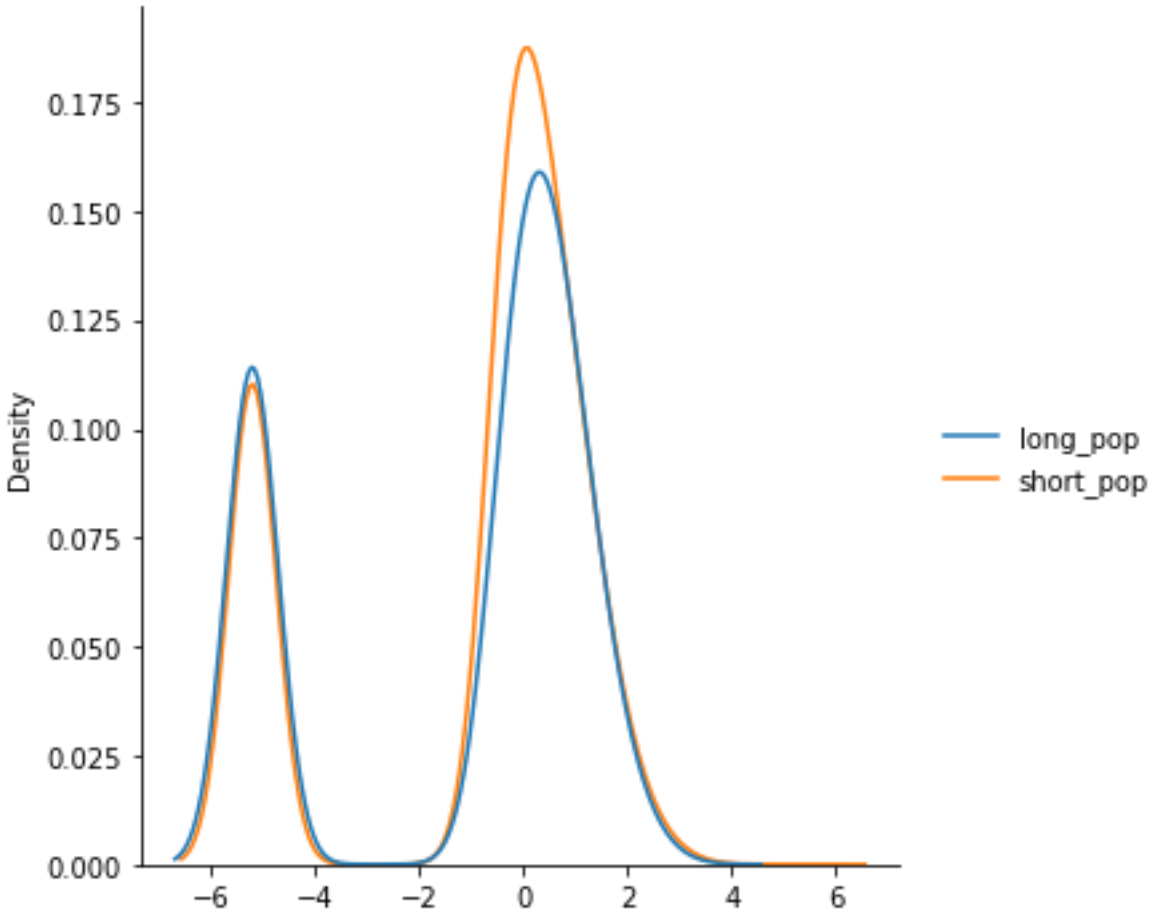
But by looking at plots we cannot see significant difference.



Popularity of long songs box plot



Popularity of short songs box plot



Popularity of long and short songs distribution plot

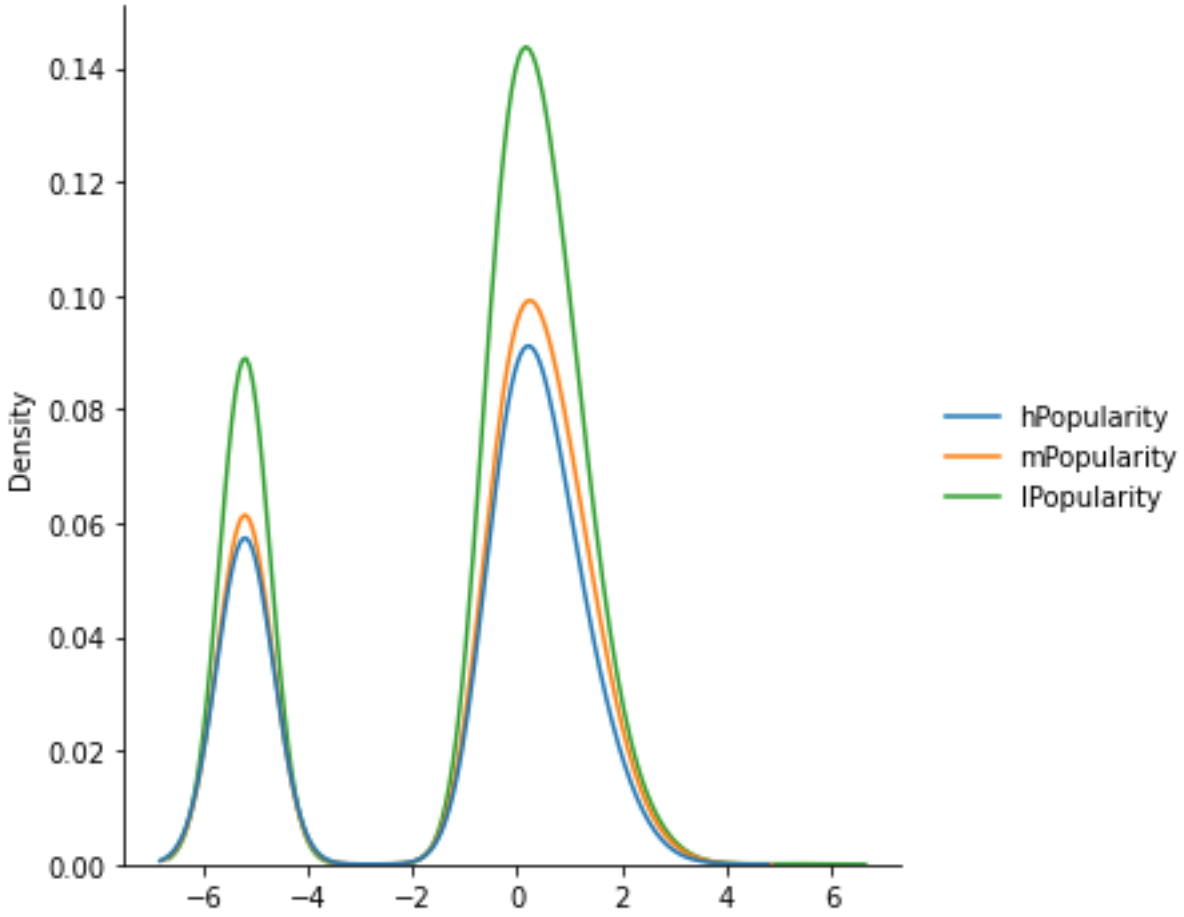
According to these plots, the popularity of short songs is a little more than long songs, but the difference is not remarkable.

Question3: Does being energetic affect popularity of songs?

Unlike previous questions in this question, we have three categories, Low energy, Moderate energy. and High energy. Like other tests the null hypothesis is that the average popularity of all categories is equal. It means that being energetic does not affect popularity.

In this test we have three sample so, we use *ANOVA* test.

The ***P_VALUE*** for this test is about 0.04 and it means that our null hypothesis has been rejected but it's probable that the statistical test is invalid because by looking at plots we cannot see a significant difference in distribution of popularity of each category.



Popularity of high, moderate and low energy songs

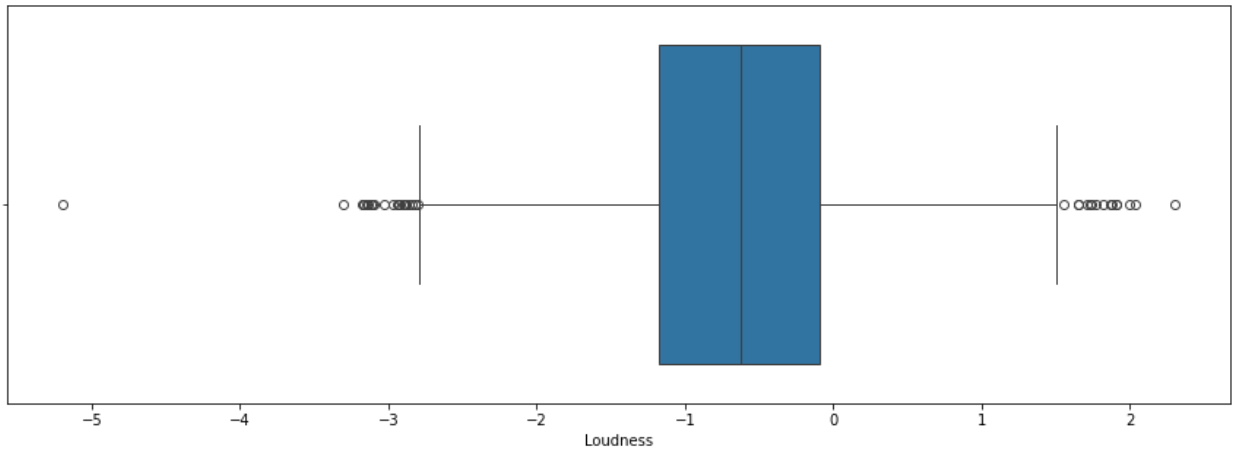
Question4: How does loudness relate to energy?

Against other tests in this test we won't talk about popularity but we will analyze How loudness effects the amount of energy. According to the previous test in this test we can indirectly find the relation between being loudness and being popular.

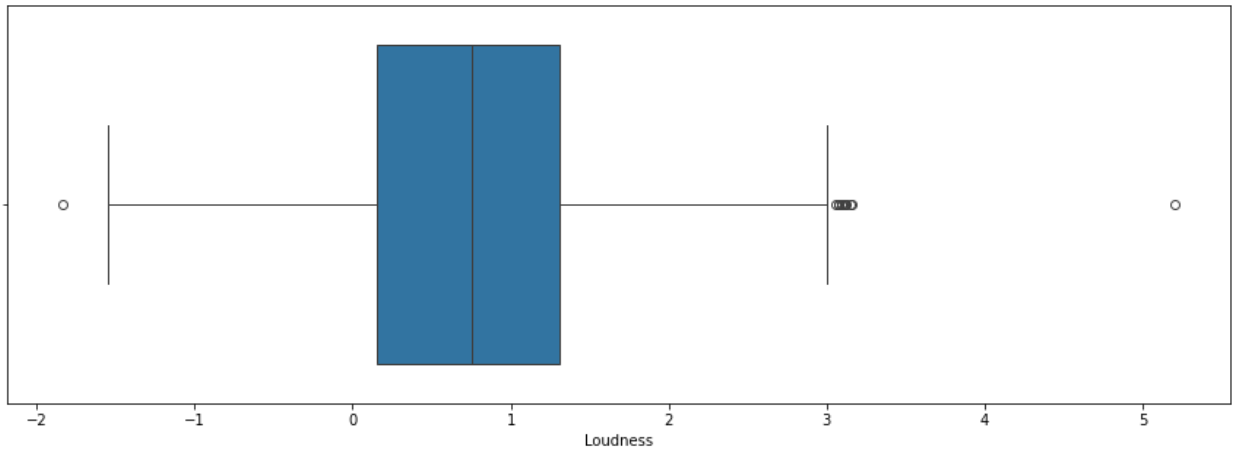
To answer this question, we use three categories in the last test and check if the average amount of loudness in all categories is equal or not.

Like other tests the null hypothesis in this test is that the average of all categories is equal which means that there is no relation between loudness and energy of songs.

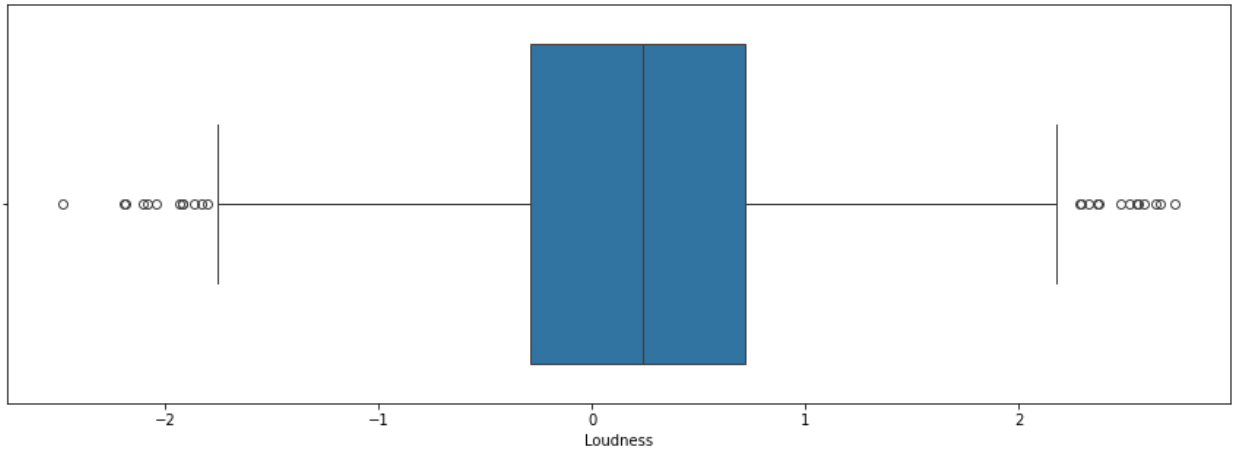
P_VALUE in this test has been acquired very small value (approximately 0) so our null hypothesis is rejected, and we expect significant difference in the average amount of loudness of each category. By looking at the plots we see that this difference is obvious.



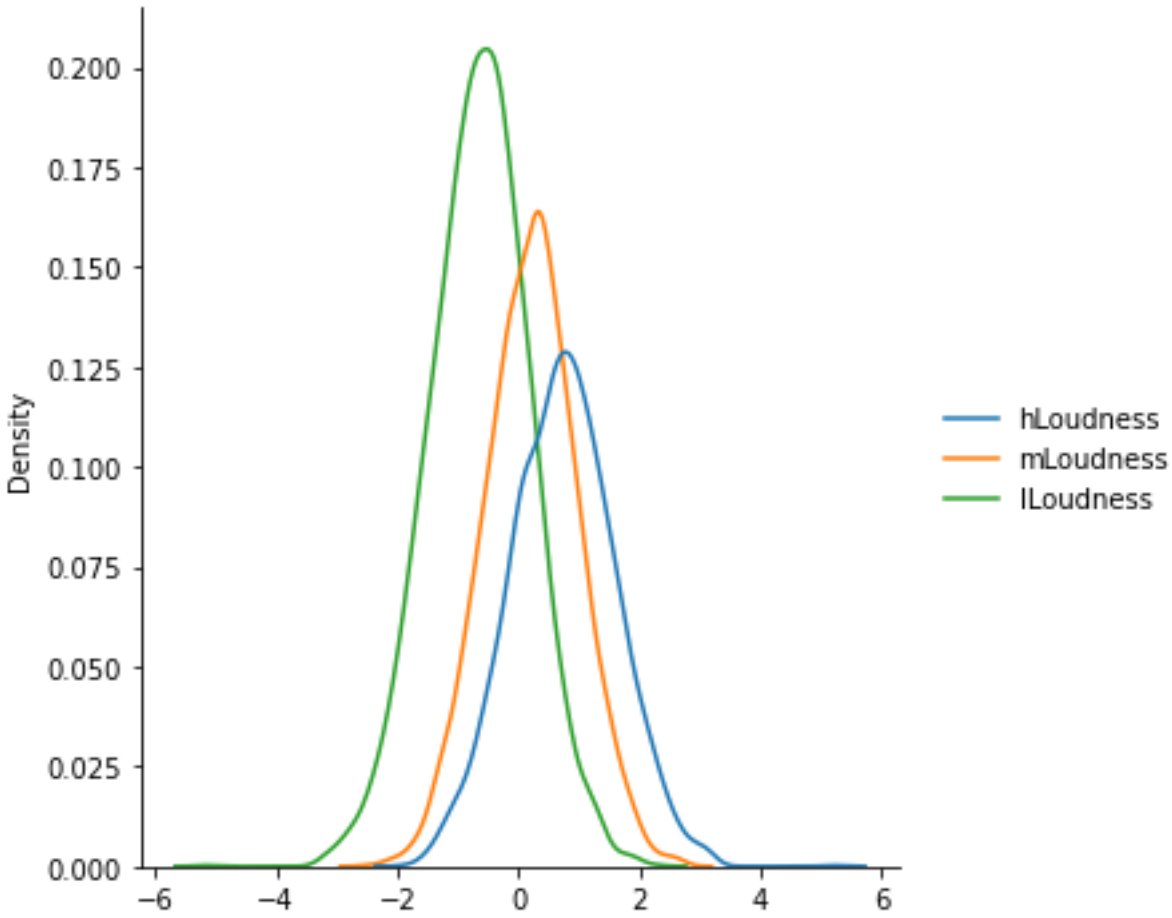
loudness of low energy songs box plot



loudness of high energy songs box plot



loudness of moderate energy songs box plot



loudness of high, low and moderate energy songs distribution plot

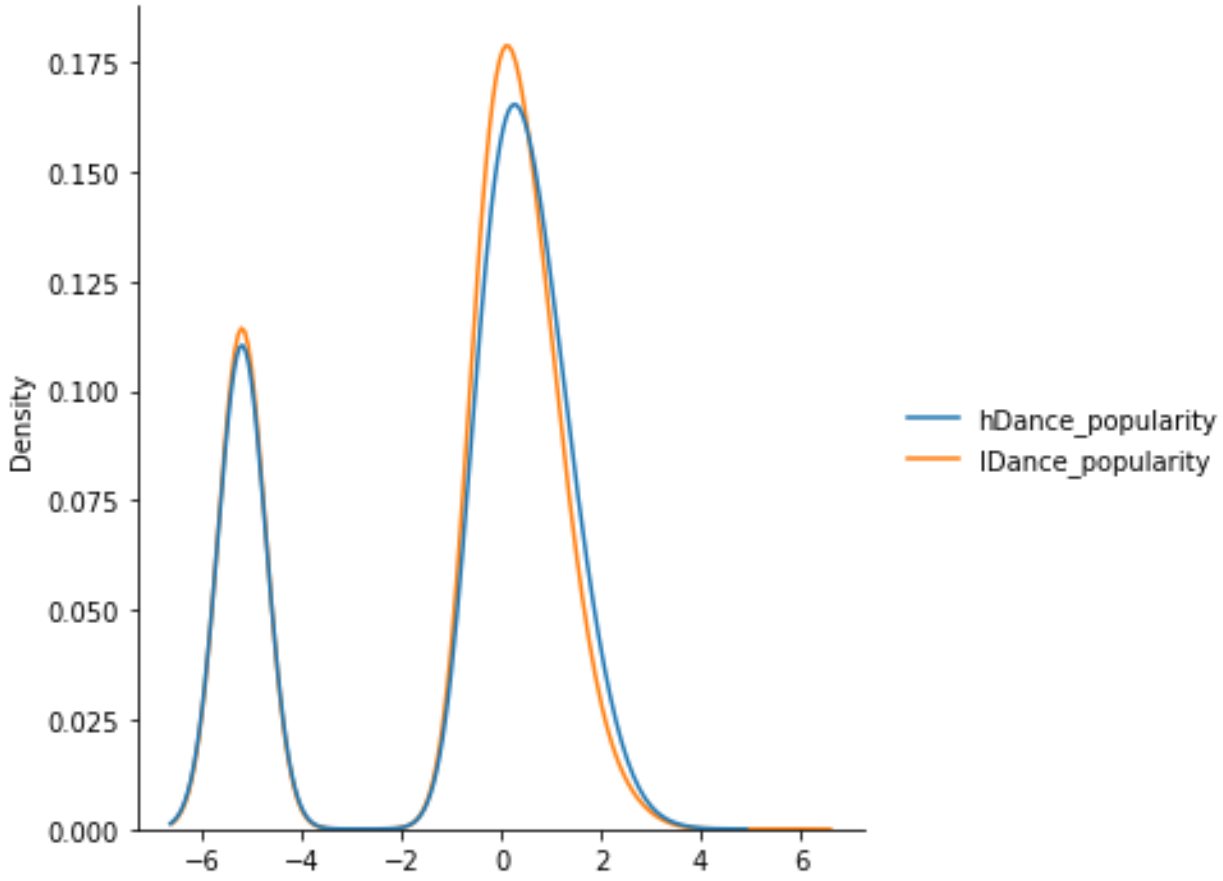
Question5: Is danceability related to popularity?

Here we have two categories; Danceable and Non danceable songs and again we want to check if danceability has any effect on popularity.

As always, we assume NOT.

The acquired ***P_VALUE,*** 0.03 rejects our null hypothesis.

By looking at plots it is clear that danceable songs are more popular among people but the difference is not significant.



4.Result

According to the done tests we couldn't find reliable relation between any feature of songs and their popularity although some tests told us that there are some differences in different type of categories but after visualizing, we saw that there is no significant difference.

The following correlation matrix can admit our conclusion.