**MUSICAL INFERENCES**

STAT 500

Fall 2019

Vineeth Suhas Challagali

Venkata Nihaal Aakula

Penn State Great Valley

School of Graduate Professional Studies

The Pennsylvania State University

**ABSTRACT**

Music is an art form and cultural activity whose medium is sound organized in time [1]. There has been an immense interest that is being developed towards the music by the people all over the world. Internet has revolutionized the way industries work and music industry is no different here. There are so many websites and applications that are available to the people which provides music of different tastes. Spotify is one of the music platforms which has got a huge attention as it provides huge collection of songs for the audience of different types. Spotify compiles a list of songs that are streamed most often over the course of a year. This project is all about exploring the top 100 songs of 2018 compiled by Spotify using statistics and draw the meaningful insights from that and concluding with the lessons learnt and future prospects of the project.

**INTRODUCTION**

**CONTEXT**

**PROBLEM DESCRIPTION**

How different characteristics like loudness, valence, energy and others can affect the popularity of the song?

**DATA**

The data is extracted from the online data science platform, Kaggle [2]. The data set is all about the songs that are streamed most often over the course of the year. This data is compiled by Spotify for the year of 2018. Spotify provides a web API to get different audio features of the songs. All the features mentioned in the dataset are calculated by Spotify.

The data consists of hundred songs with thirteen different features alongside the Spotify URI, Name and Artist of the song. The total data occupies around 12 kilobytes of memory.

The following are the columns of the dataset:

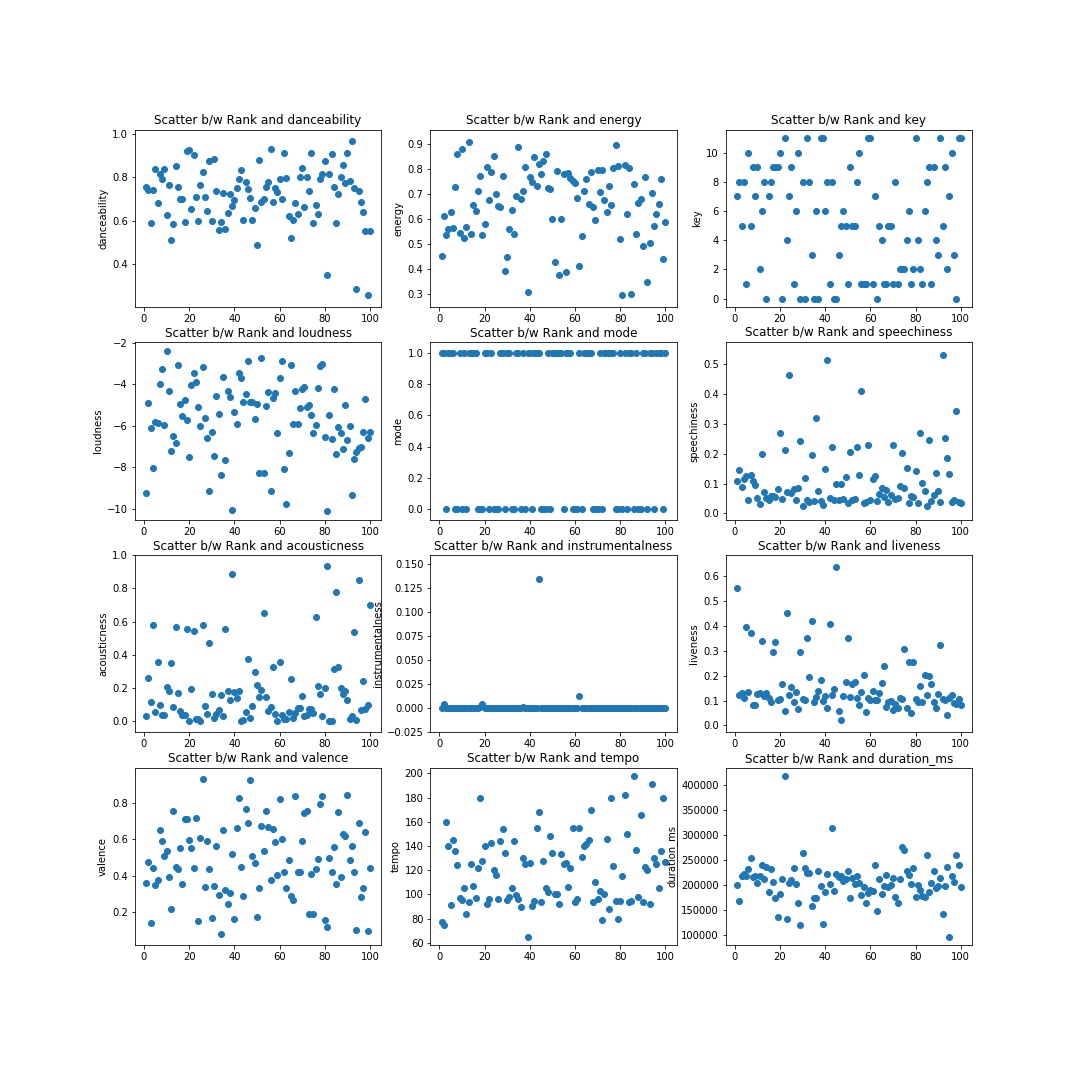
|  |  |
| --- | --- |
| **Column** | **Description** |
| Id | Spotify URI of the song |
| Name | Name of the song |
| Artist | Artist(s) of the song |
| Danceability | Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. |
| Energy | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. |
| Key | The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. |
| Loudness | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. |
| Mode | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| Speechiness | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| Acousticness | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| Instrumentalness | Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| Liveness | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| Valence | 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). |
| Tempo | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| Duration | The duration of the track in milliseconds. |
| TimeSignature | An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). |

The sampling method used to collect this dataset can be considered as systematic sampling. As it followed the pattern of sorting the songs in the order of their frequency of streaming. As this dataset provides the most popular songs in a year, it would be easier to analyze the characteristics which are important for a song to be popular and also the songs which are from a particular baseline. But It would be disadvantageous to apply various statistical methods for these kinds of samples and also there could be other factors which could affect the song popularity like the familiarity of the artist.

**Exploratory Data Analysis**

**Descriptive statistics**

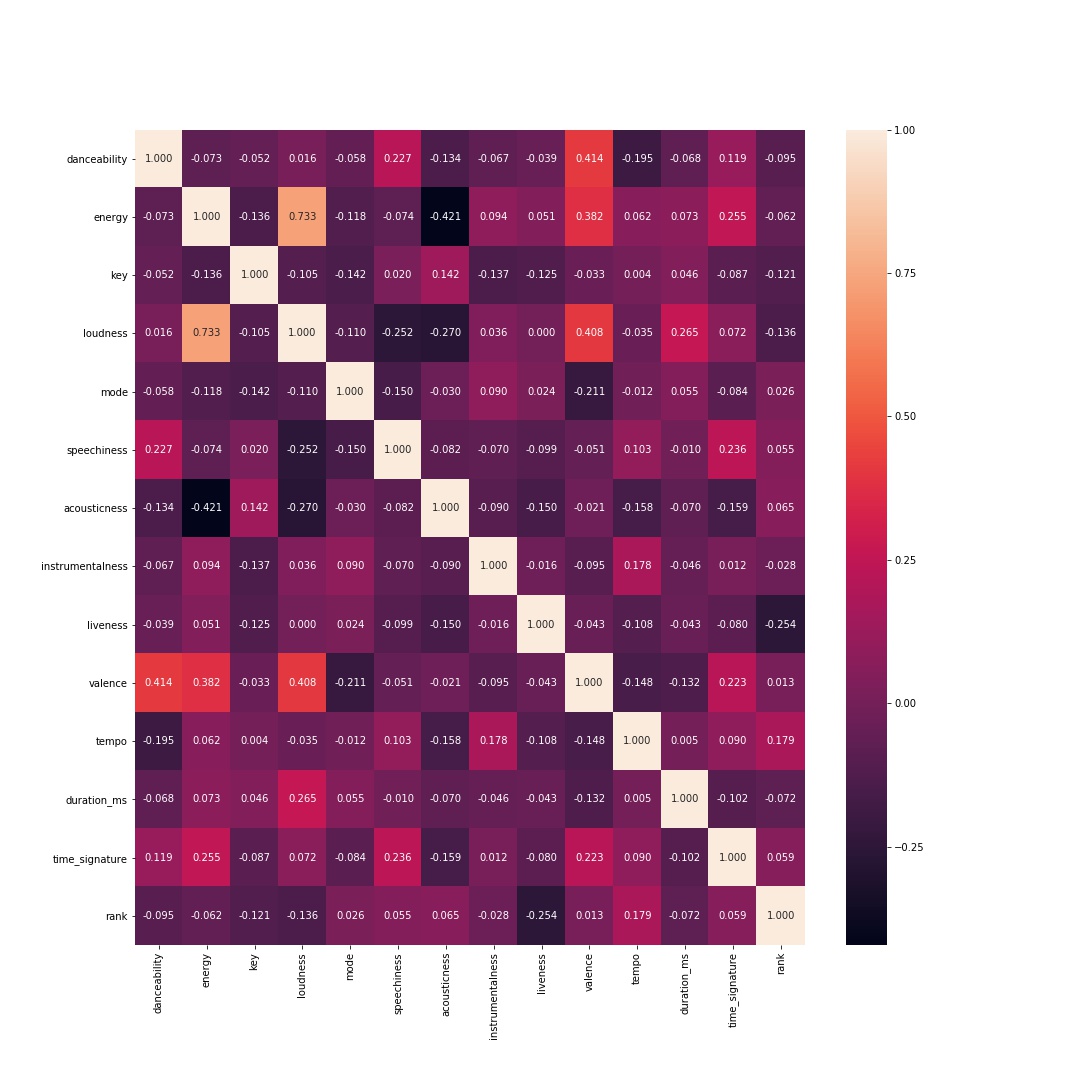
* Data preprocessing:
  + As we can see that the data consists of an Id field which is actually the URI but not the rank of the song. Adding an additional rank column for the sake of analysis.
  + Also, fortunately there are no missing values in the dataset which makes sense as these are the set of top 100 songs which might not miss any important characters.
  + Below is the correlation between rank and all the other characteristics for the better understanding of the behavior of characters as the rank goes up.



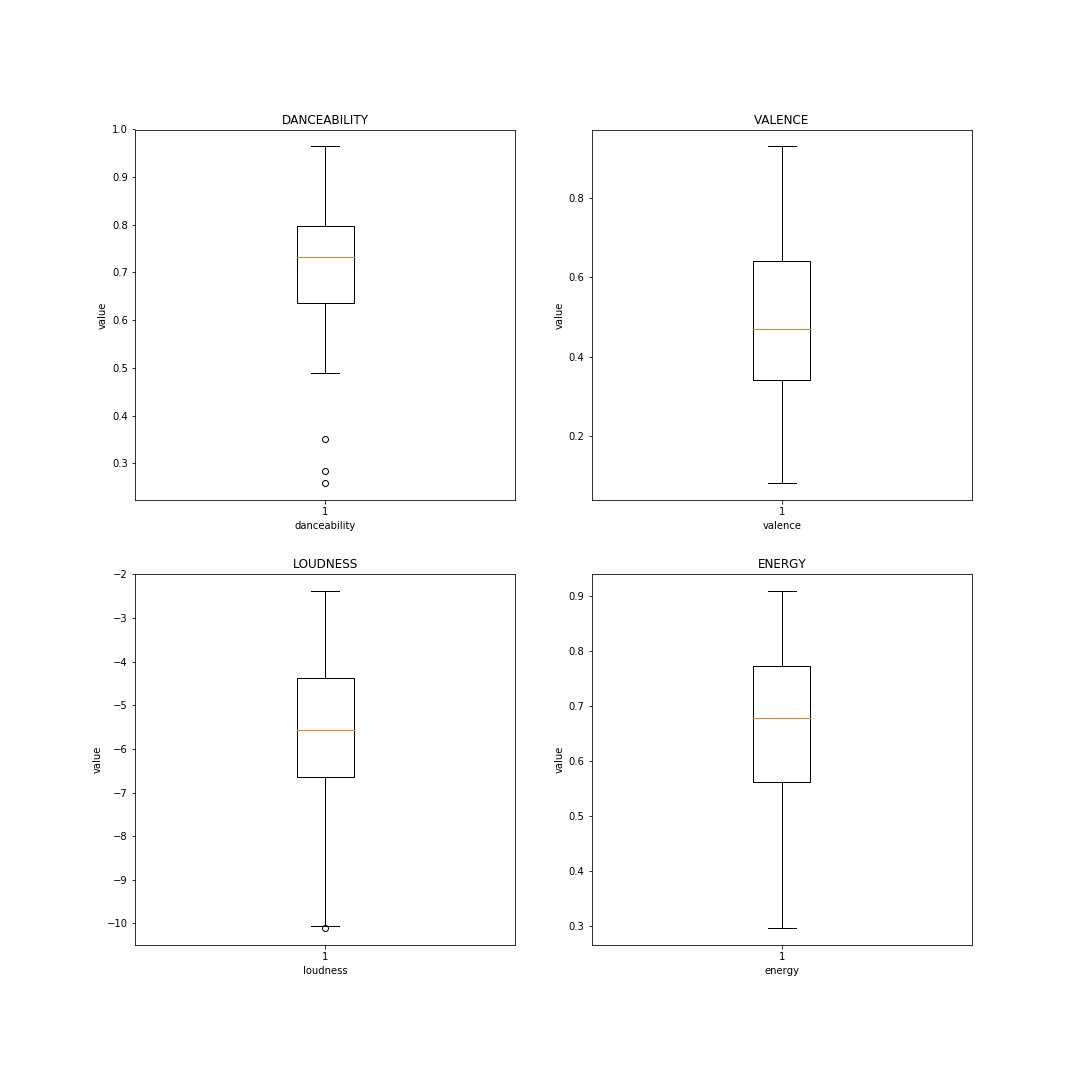
* Characteristics of a data set
  + The basic description of the dataset including central measures, ranges and standard deviations are shown below

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **danceability** | **energy** | **key** | **loudness** | **mode** | **speechiness** | **acousticness** | **instrumentalness** | **liveness** | **valence** | **tempo** | **duration\_ms** | **time\_signature** |
| **count** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **mean** | 0.716 | 0.659 | 5.33 | -5.677 | 0.59 | 0.115 | 0.195 | 0.00158 | 0.158 | 0.4844 | 119.90 | 205206.78 | 3.98 |
| **std** | 0.131 | 0.145 | 3.676 | 1.777 | 0.494 | 0.104 | 0.22094 | 0.0134 | 0.1116 | 0.2061 | 28.79 | 40007.89 | 0.200 |
| **min** | 0.258 | 0.296 | 0.0 | -10.109 | 0.0 | 0.0232 | 0.000282 | 0.0 | 0.0215 | 0.0796 | 64.934 | 95467.0 | 3.0 |
| **25%** | 0.635 | 0.562 | 1.75 | -6.6505 | 0.0 | 0.0453 | 0.040 | 0.0 | 0.0946 | 0.3410 | 95.73075 | 184680.0 | 4.0 |
| **50%** | 0.733 | 0.678 | 5.0 | -5.566 | 1.0 | 0.074 | 0.109 | 0.0 | 0.1185 | 0.4705 | 120.116 | 205047.5 | 4.0 |
| **75%** | 0.798 | 0.772 | 8.25 | -4.36375 | 1.0 | 0.137 | 0.24775 | 3.0875E-05 | 0.17075 | 0.6415 | 140.02275 | 221493.25 | 4.0 |
| **max** | 0.964 | 0.909 | 11.0 | -2.384 | 1.0 | 0.53 | 0.934 | 0.134 | 0.636 | 0.9310 | 198.075 | 417920.0 | 5.0 |

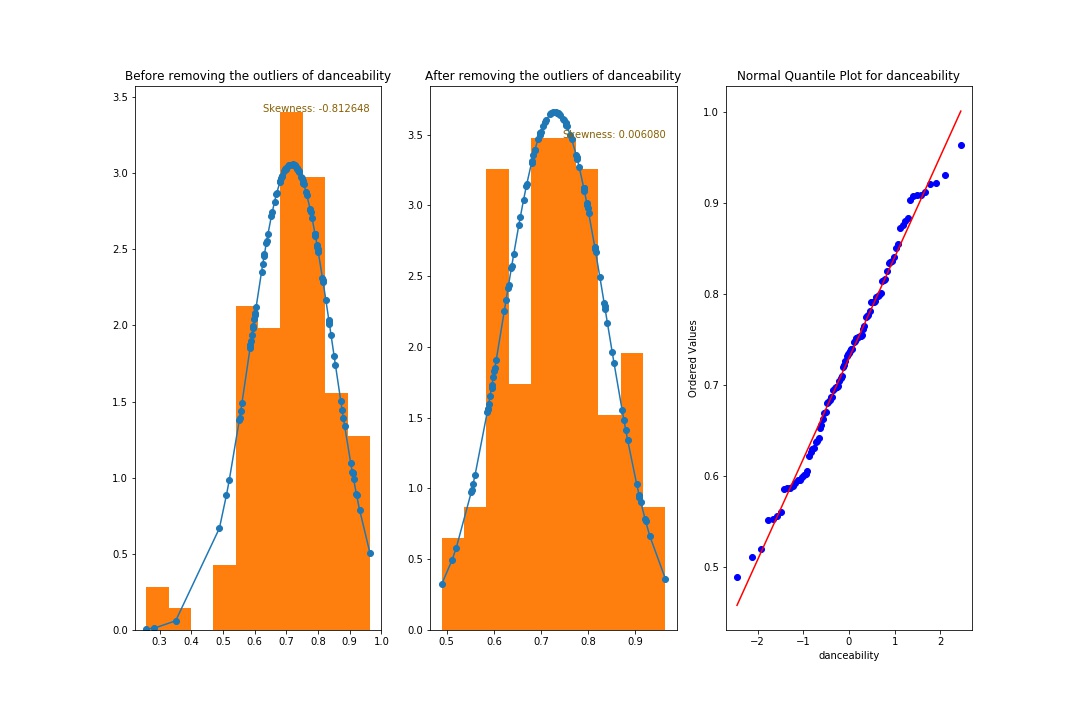
* + It would be difficult to understand the actual relationships from the above descriptive statistics and the correlations. So, it’s better to dig into the correlation between different characters which could give us some meaningful information.
  + Below is the heat map consisting of correlation coefficients with the light color defining more correlation and dark color defining the least correlation.

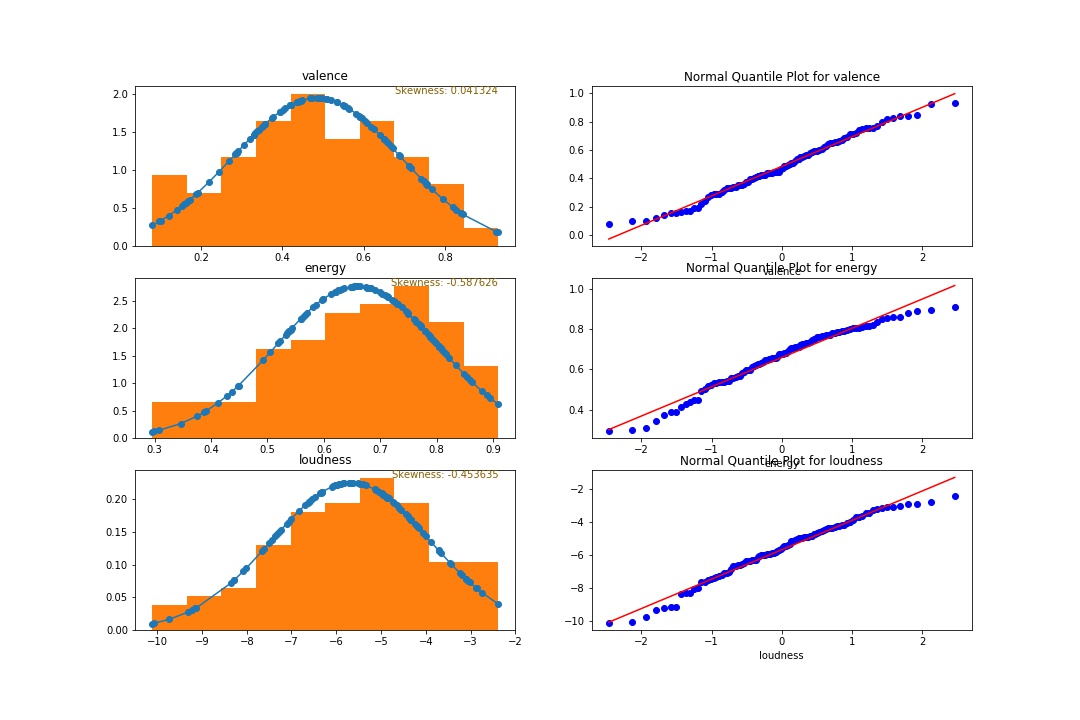


* + From the above heat map, we can see that there is a highest correlation between Loudness and Energy. Also, there are few characters like danceability and valence which are significantly correlated which can be useful in determining the popularity of the song. We are considering only these correlated characters for our further analysis.
  + The below boxplots for each of the selected characters can help us understand the central tendency and the spread of the data along with the outliers which can significantly affect our analysis.



* + We can clearly see that there exists the outliers for the danceability and all others looks good. The histograms can help us understand the distributions of each of the characters.
  + The below diagram represents the histograms and normal quantile plots respectively for each of the character. We can observe that there is a clear pattern of normality that exists for each of the characters and this can help us apply few of the inferencial methods on them to understand the population paramters.





**Inferential statistics**

**Claim**: In the recent times, the songs are majorly composed of Energy

**Hypothesis Test**:

* STEP-1: Symbolic form of the claim

To validate the claim, we are conducting a hypothesis test to see if the songs have mean\_energy > 0.5

* STEP - 2 & 3: Symbolic form of alternative claim and

mean\_energy = 0.5;

H0: mean\_energy = 0.5

H1: mean\_energy > 0.5

* STEP - 4: Significance Level

Considering the significance level of alpha = 0.05

* ALL THE DIFFERENT PARAMETERS:

n = 100 # Size of the sample

popmean = 0.5 # Population mean

sample = songs['energy'] # Extract the energies of the sample songs

mean\_energy\_of\_sample = np.mean(sample) # Mean of the sample

sd\_energy\_of\_sample = np.std(sample) # standard deviation of the sample

* STEP - 5: Identify the test statistic

Because the claim is made about the population mean m, the sample statistic

most relevant to this test is the sample mean x, and we use the t distribution.

* Requirements:

1. Either the data is normally distributed or n > 30 [According to sample, energy is normal and sample size is > 30]

2. Assuming the sample to be the simple random [If we consider the population to be of popular songs and this is a sample of top 100 songs]

3. The sample is independent

* STEP - 6: Calculate the P-value

print(stats.ttest\_1samp(songs['energy'], 0.5))

* STEP - 7: Make a decision

The above statement results in the sample statistic and pvalue as follow:

Ttest\_1sampResult(statistic=10.964576604381161, pvalue=8.713371408794922e-19)

As the P-value is less than alpha, we can reject the null hypothesis.

**Conclusion**:

THERE IS SUFFICIENT EVIDENCE TO SUPPORT THE CLAIM THAT

THE SONG SHOULD BE MAJORLY COMPOSED OF ENERGY TO BE MORE POPULAR.

**Presentation of the results - show usefulness and novelty of the results**

**Reasons why the results might not be correct**

**Description of ways in which the study could be improved.**

**REFERENCES**

[1] <https://en.wikipedia.org/wiki/Music>

[2] <https://www.kaggle.com/nadintamer/top-spotify-tracks-of-2018>