Report For
Musica Recommendation Engine

Data Science
Assignment#3

Muhammad Anique 20L-2171

Hammad Zahoor 20L-0971

Kaashaan Ali Karim 20L-2073

Introduction:

The dataset under examination is sourced from Spotify, a leading music streaming platform known for its vast and diverse collection of songs. Spotify offers an ideal source of data for our analysis and the development of a music recommendation system. In this report, we delve into the exploration of this dataset to gain a comprehensive understanding of its components and properties, with the ultimate goal of creating a recommendation system that provides users with personalized music suggestions.

Problem Statement:

The objective of this report is to design and implement a music recommendation system that leverages the wealth of information contained within the Spotify dataset. Music recommendation systems have become increasingly important as music libraries continue to grow, making it challenging for users to discover new music that aligns with their preferences. Our aim is to address this challenge by developing a system that considers a variety of musical attributes, such as acousticness, danceability, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, and valence, to recommend songs that cater to individual tastes and moods.

Variables:

Artists:

This categorical variable denotes the artists or musical acts responsible for each song. It provides information about the creators of the music.

Acousticness:

A numeric feature that quantifies the degree to which a song is acoustic or non-electronic. Values close to 1.0 suggest a more acoustic sound, while values close to 0.0 indicate a less acoustic, more electronic sound.

Danceability:

A numeric feature that measures a song's suitability for dancing. Higher values imply greater danceability.

Energy:

This numeric attribute reflects the intensity and activity level of a song. Higher energy values represent more energetic music.

ID:

A unique identifier associated with each song in the dataset.

Instrumentalness:

A numeric feature indicating the extent to which a song is instrumental. Values near 1.0 signify instrumental music, while values near 0.0 imply the presence of lyrics or vocals.

Key:

A categorical variable representing the musical key of each song. The key provides information about the fundamental pitch of the music.

Liveness:
A numeric feature that characterizes the presence of a live audience during a song's recording. Higher values suggest live recordings.

Loudness:
A numeric feature that quantifies the loudness of a song. More positive values indicate louder music.

Mode:
This categorical variable denotes the musical mode of each song (major or minor), which influences its emotional tone.

Name:
The title of each song, providing information about its content or subject matter.

Speechiness:
A numeric attribute that indicates the presence of spoken words or lyrics in a song. Higher values suggest a higher proportion of speech.

Tempo:
A numeric feature representing the tempo or beats per minute (BPM) of each song. It determines the song's speed or pace.

Valence:
A numeric feature that reflects the positivity or happiness of a song. Higher valence values suggest a more positive emotional tone.

Univariate Analysis

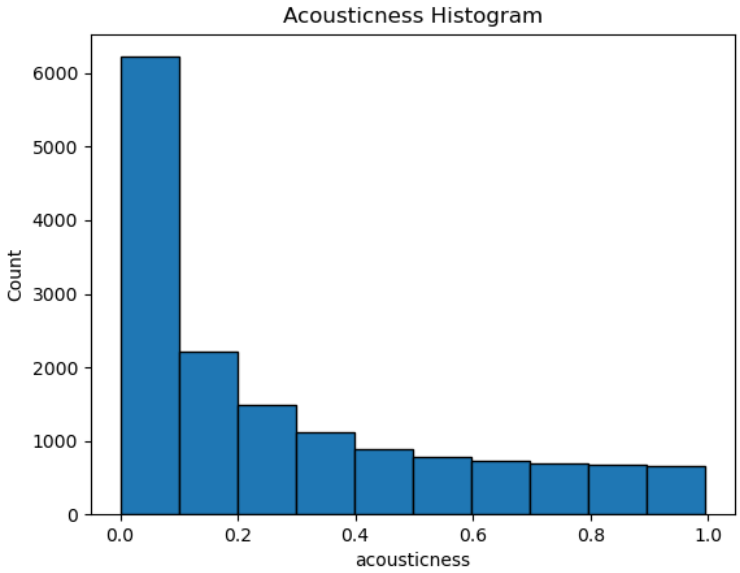
The following section presents the univariate analysis of the song dataset, focusing on individual variables to gain insights into their distributions and characteristics

	acousticness	danceability	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	valence
count	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000
mean	0.276240	0.622192	0.632248	0.046076	5.254669	0.182581	-7.155861	0.623845	0.099286	120.306547	0.492853
std	0.287896	0.166186	0.207826	0.178125	3.558730	0.145066	4.152240	0.484435	0.102389	30.102370	0.238314
min	0.000000	0.000000	0.000020	0.000000	0.000000	0.000000	-54.376000	0.000000	0.000000	0.000000	0.000000
25%	0.035250	0.520000	0.499000	0.000000	2.000000	0.096050	-8.289500	0.000000	0.036700	96.915000	0.310000
50%	0.163000	0.637000	0.654000	0.000001	5.000000	0.123000	-6.290000	1.000000	0.054300	119.966000	0.488000
75%	0.456000	0.743000	0.792000	0.000216	8.000000	0.225000	-4.821000	1.000000	0.113000	140.024500	0.680000
max	0.996000	0.980000	1.000000	1.000000	11.000000	0.989000	1.342000	1.000000	0.902000	220.099000	0.990000

Histograms:

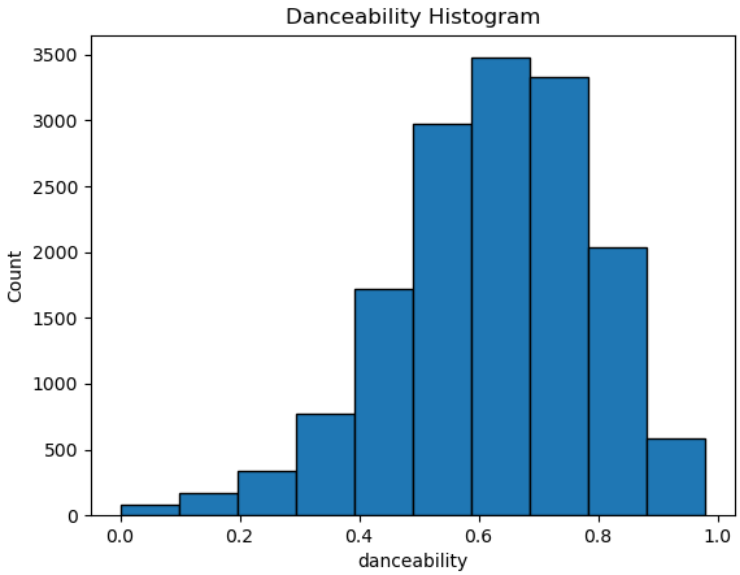
Acousticness:

Distribution is right skewed which implies that the majority of songs have less acousticness.



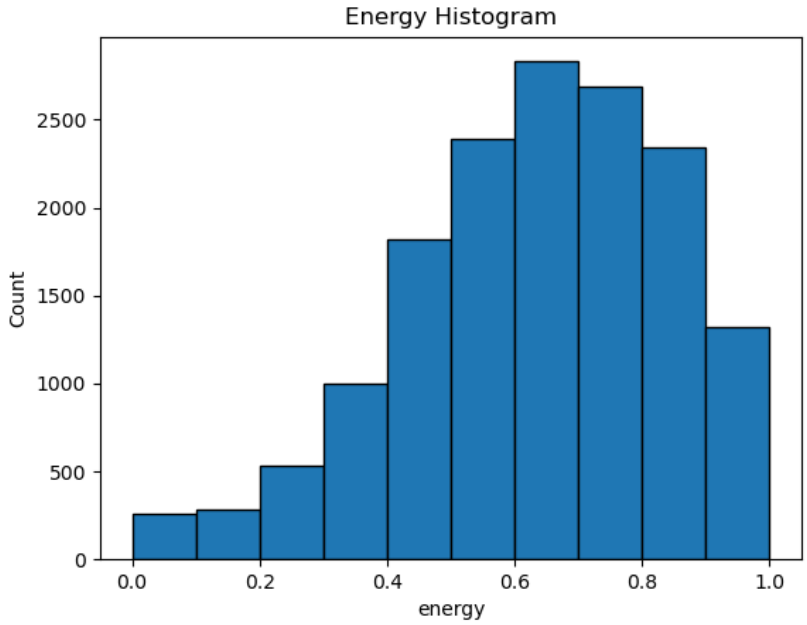
Danceability:

Distribution is left skewed which implies that the majority of songs have a high level of danceability.



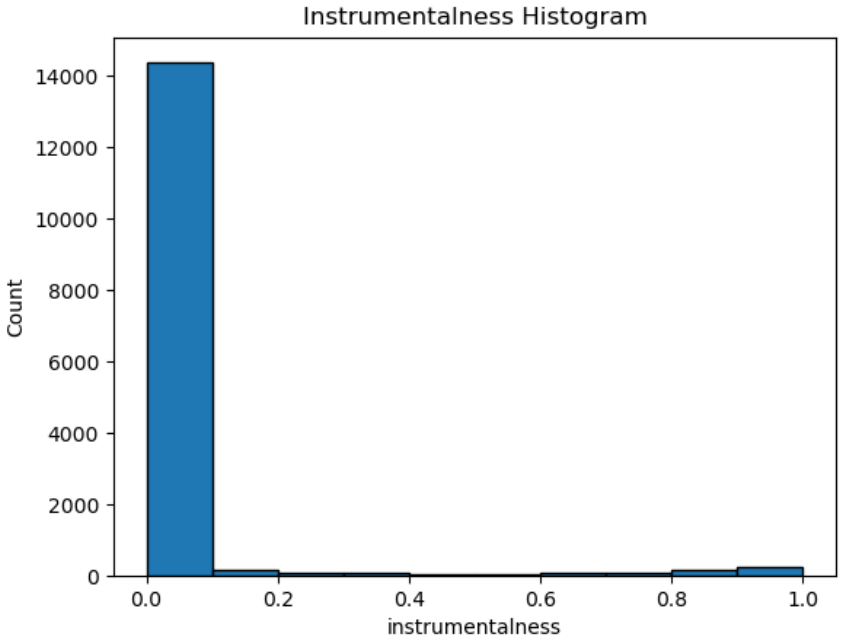
Energy:

Distribution is left skewed which implies that the majority of songs have a high level of energy.



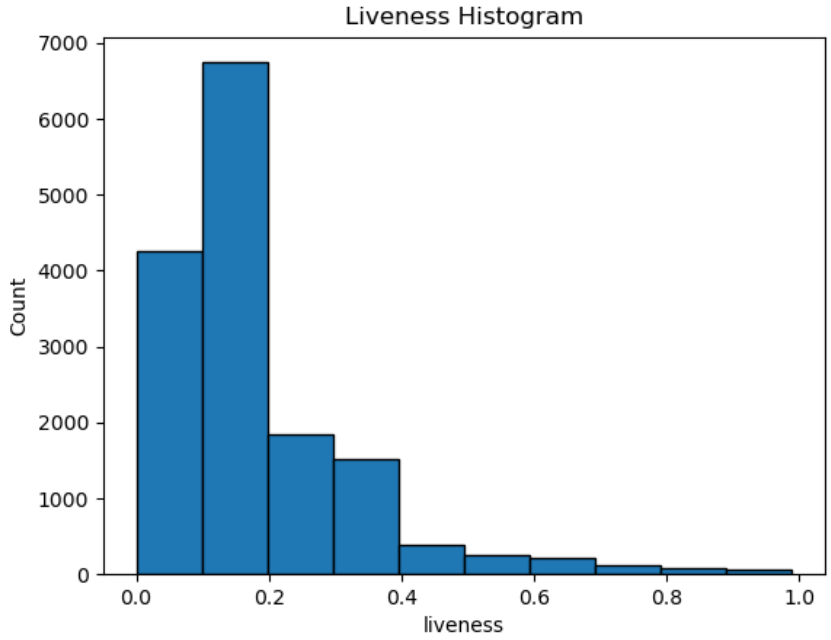
Instrumentalness:

Distribution is right skewed which implies that the majority of songs have less instrumentalness.



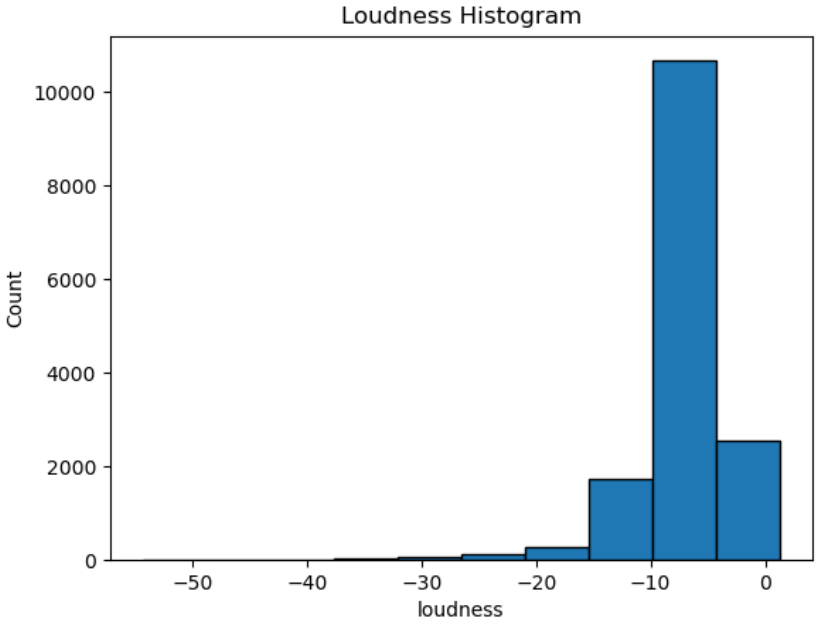
Liveness:

Distribution is right skewed which implies that the majority of songs have less liveness.



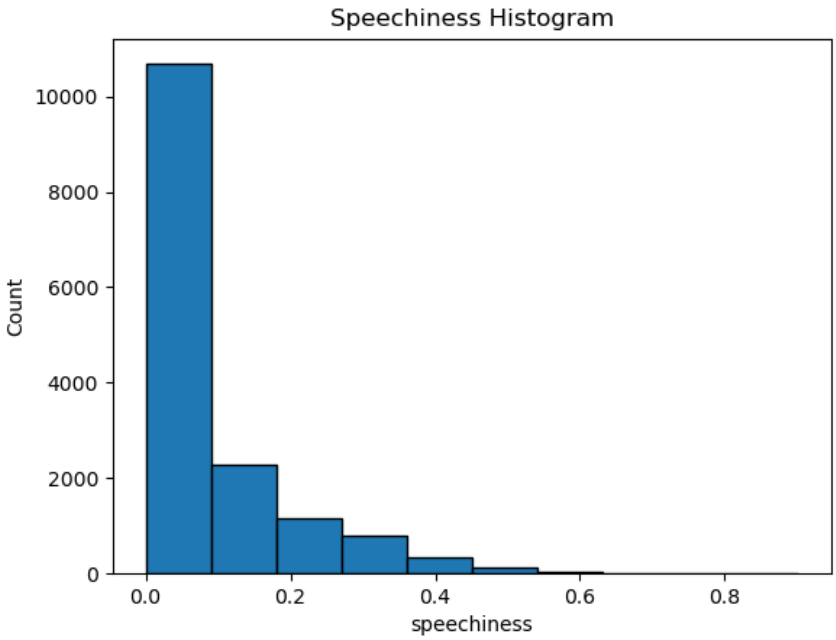
Loudness:

Distribution is left skewed which implies that the majority of songs have a high level of loudness.



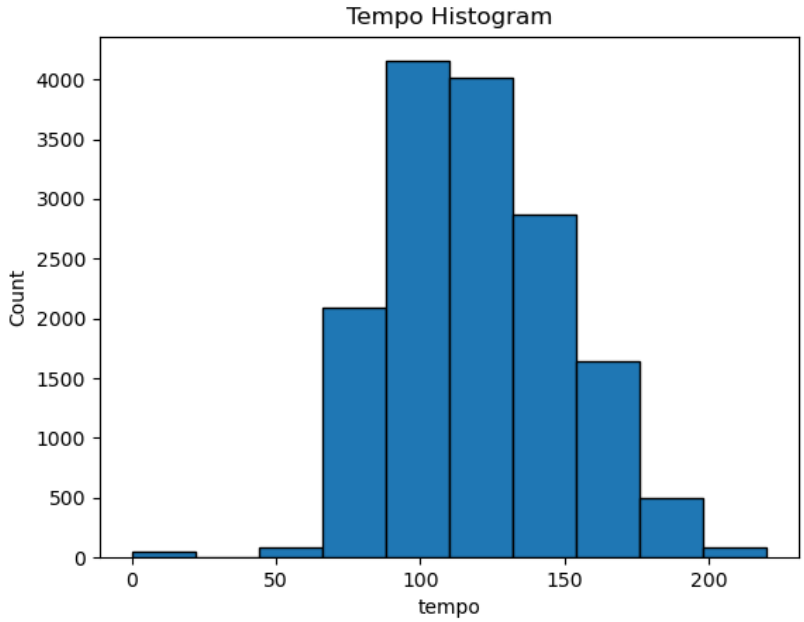
Speechiness:

Distribution is right skewed which implies that the majority of songs have less speechiness.



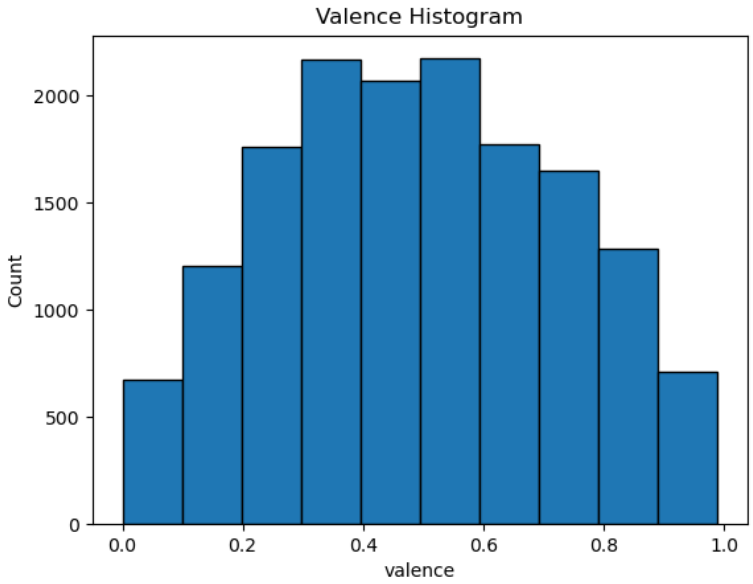
Tempo:

Distribution is normal which implies that the temp values near the mean occur more frequently.



Valence:

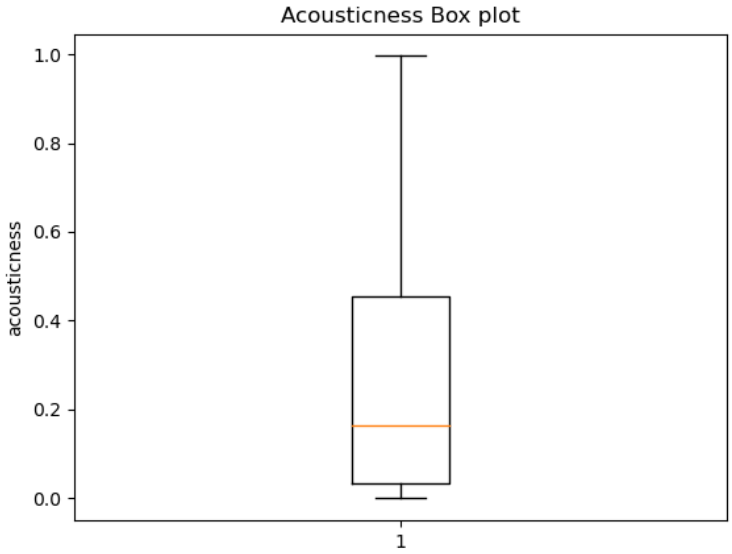
Distribution is normal which implies that the valence values near the mean occur more frequently.



Box plots:

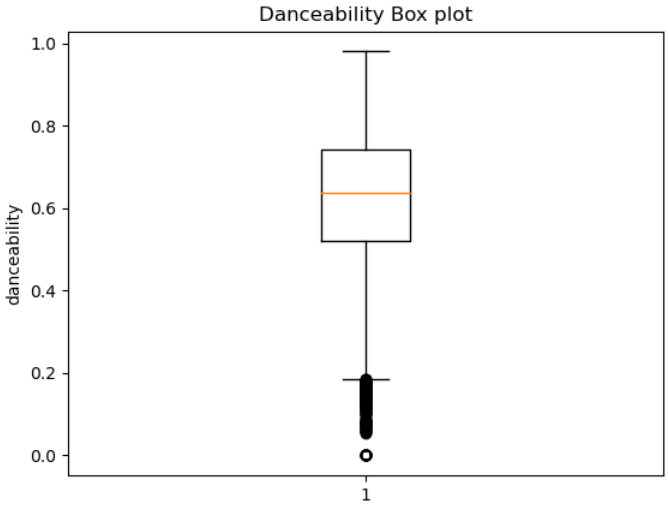
Acousticness:

There are no outliers.



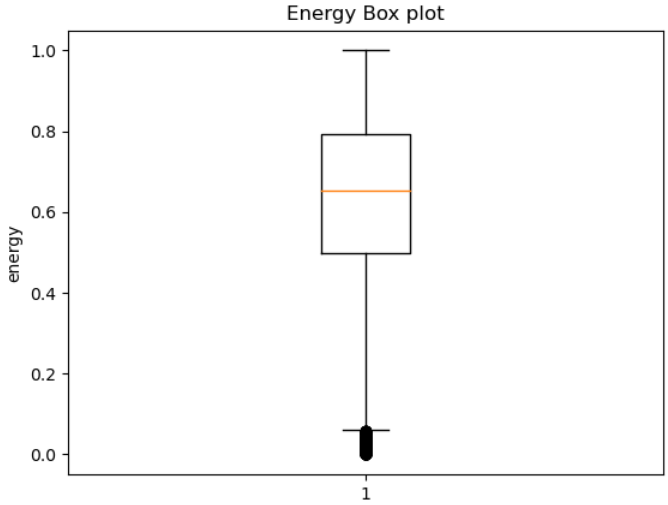
Danceability:

Some outliers can be seen which have an extremely low danceability value.



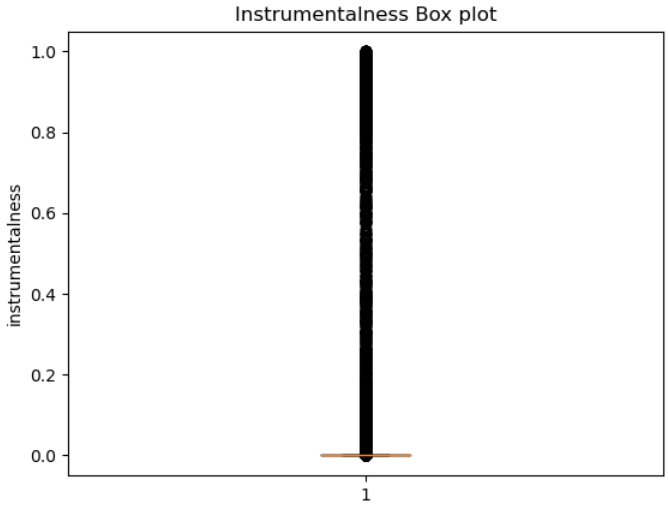
Energy:

Some outliers can be seen which have an extremely low energy value.



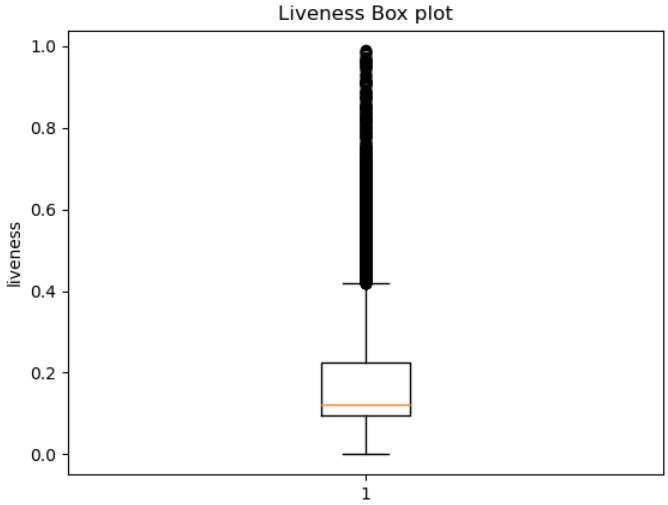
Instrumentalness:

There are many outliers spread over a wide range.



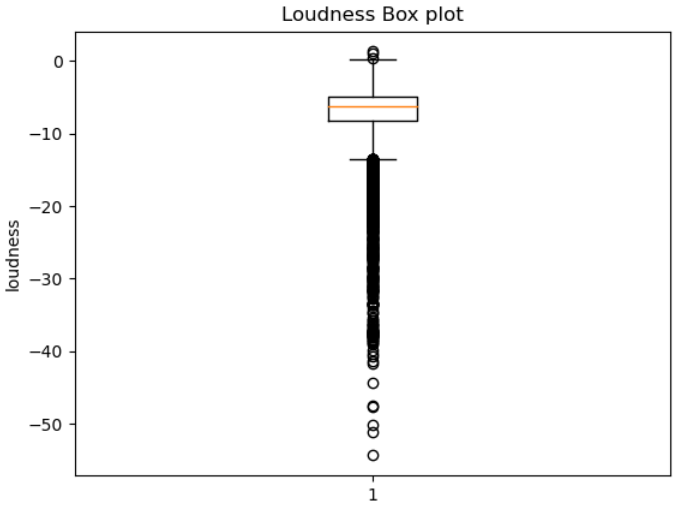
Liveness:

Many outliers can be seen which have moderate to high liveness.



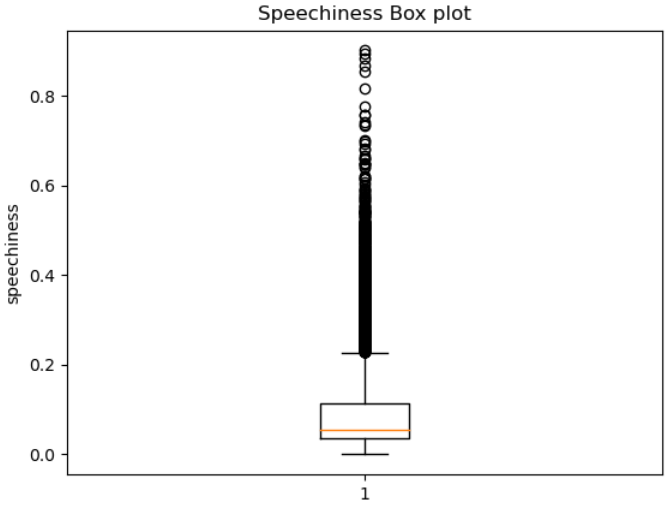
Loudness:

Many outliers can be seen having moderate loudness while some have low loudness.



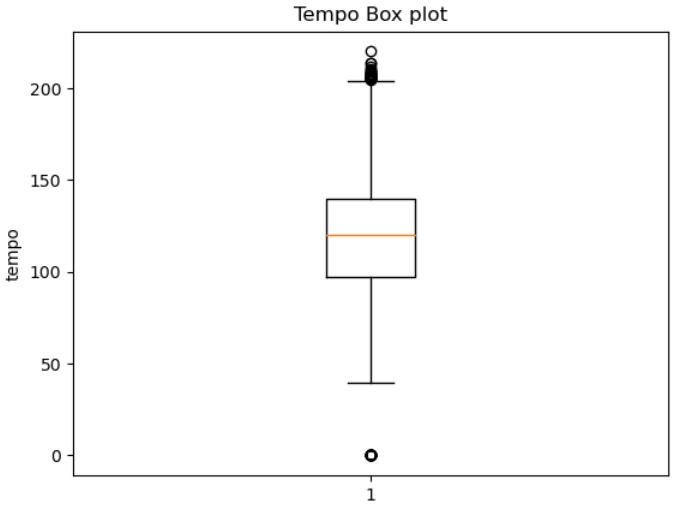
Speechiness:

Many outliers can be seen having moderate loudness while some have high loudness.



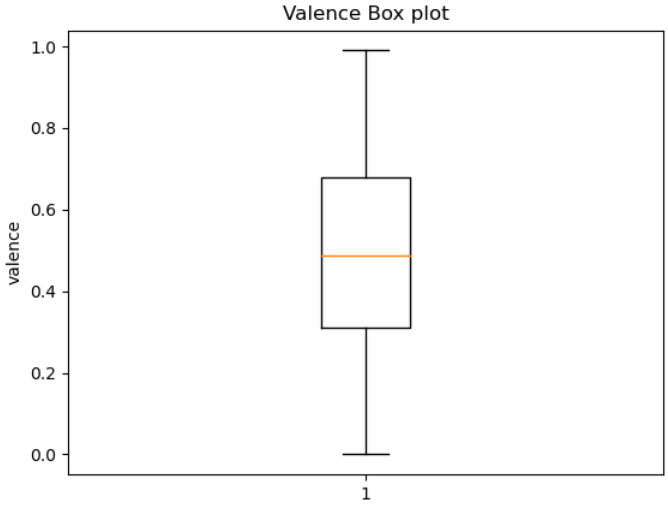
Tempo:

Some outliers have high tempo whereas a few outliers can be seen having low tempo.



Valence:

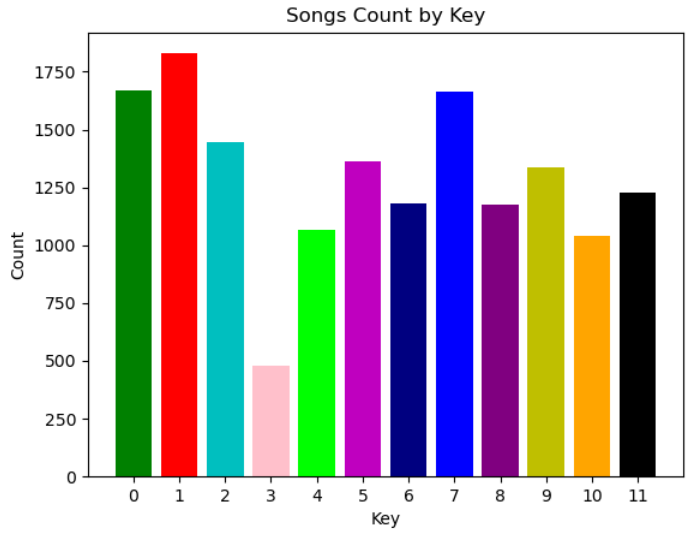
There are no outliers.



Bar chart:

Key:

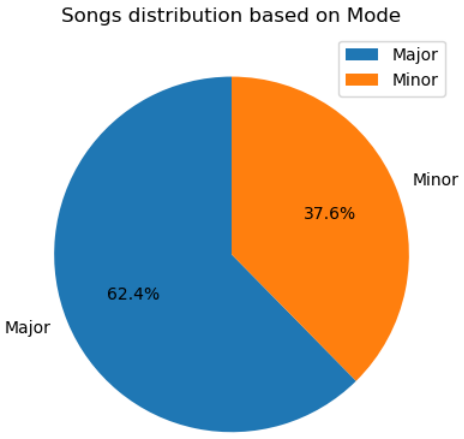
Shows that the majority of songs used key 1 followed by key 0, key 7, key 2, key 5, key 9, key 11, key 6, key 8, key 4, key 10 and finally the least used key 3.



Pie chart:

Mode:

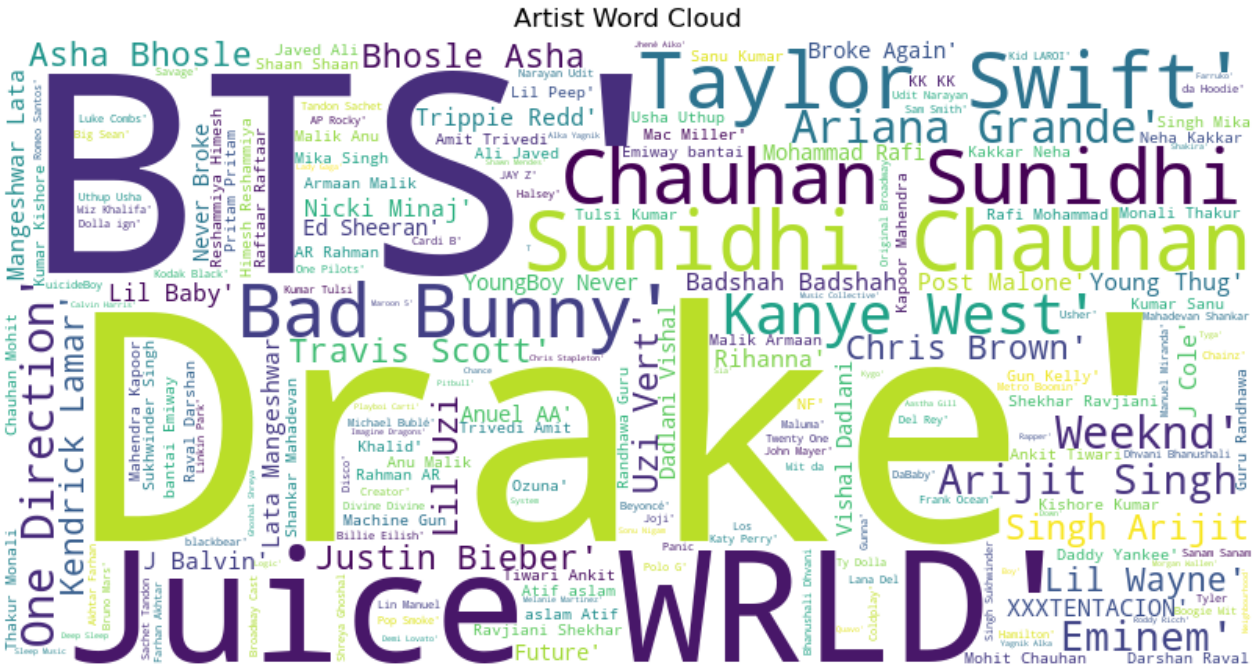
Shows that 63.4% of songs utilize the major mode whereas 37.6% of songs utilize minor mode.



Word cloud:

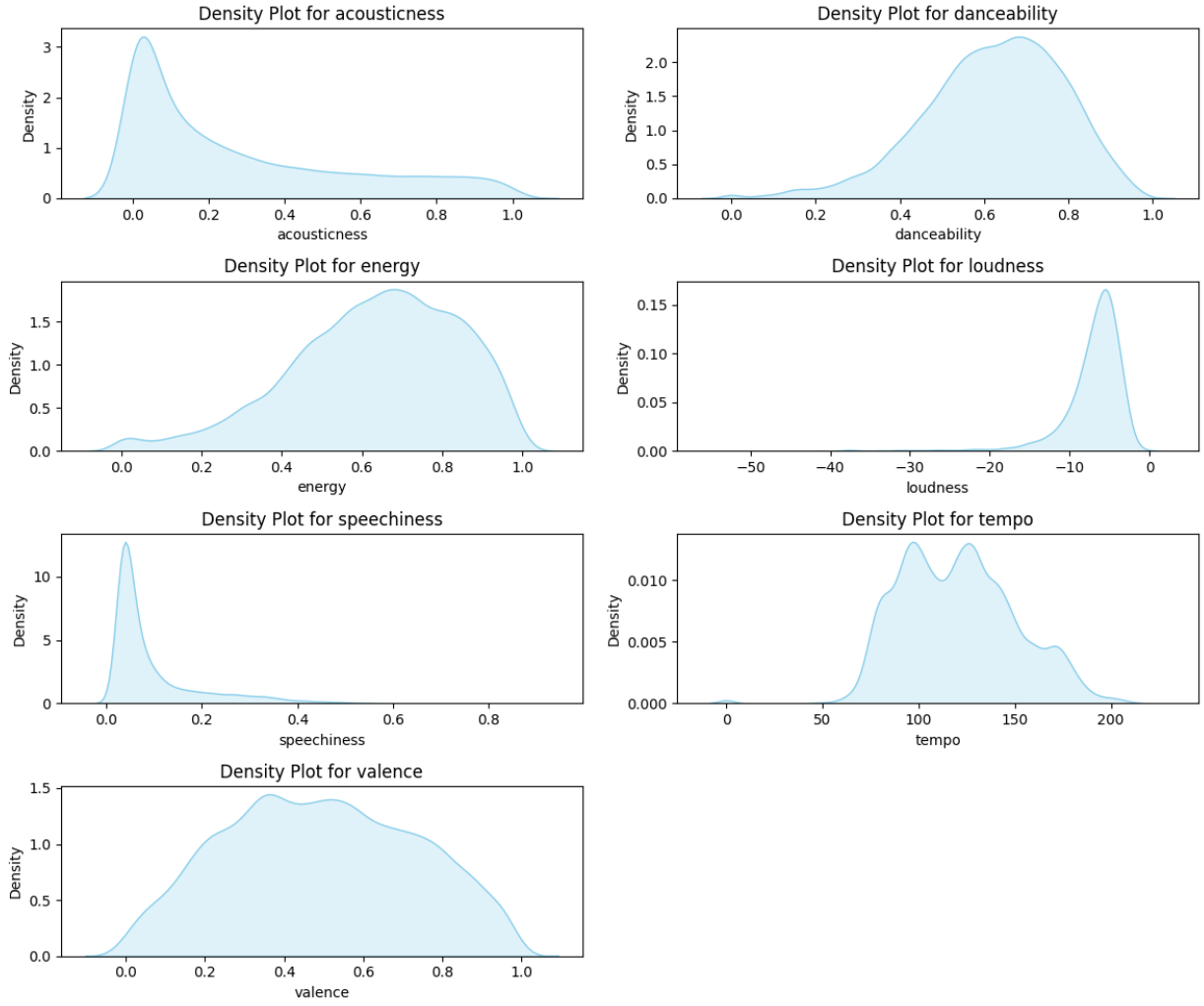
Artists:

Shows the artists that have higher frequency (i.e., have a greater number of songs in the dataset) in larger font.



Density Plots:

Density plots are a valuable tool for understanding the distribution of numeric variables within a dataset. The width of the density plot at different points represents the variability or spread of the data. A wider plot suggests greater variability, while a narrower plot indicates less variability and the highest point on the density plot corresponds to the variable's central tendency.

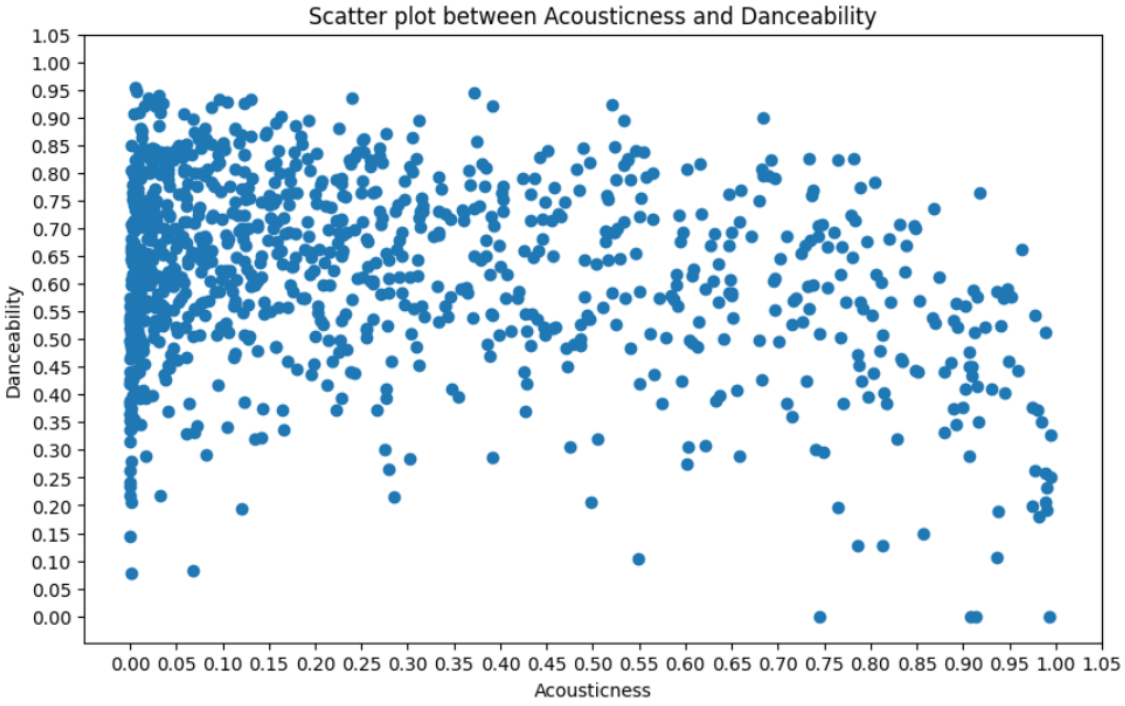


Bivariate Analysis

Scatter Plots:

Accosticness vs Danceability

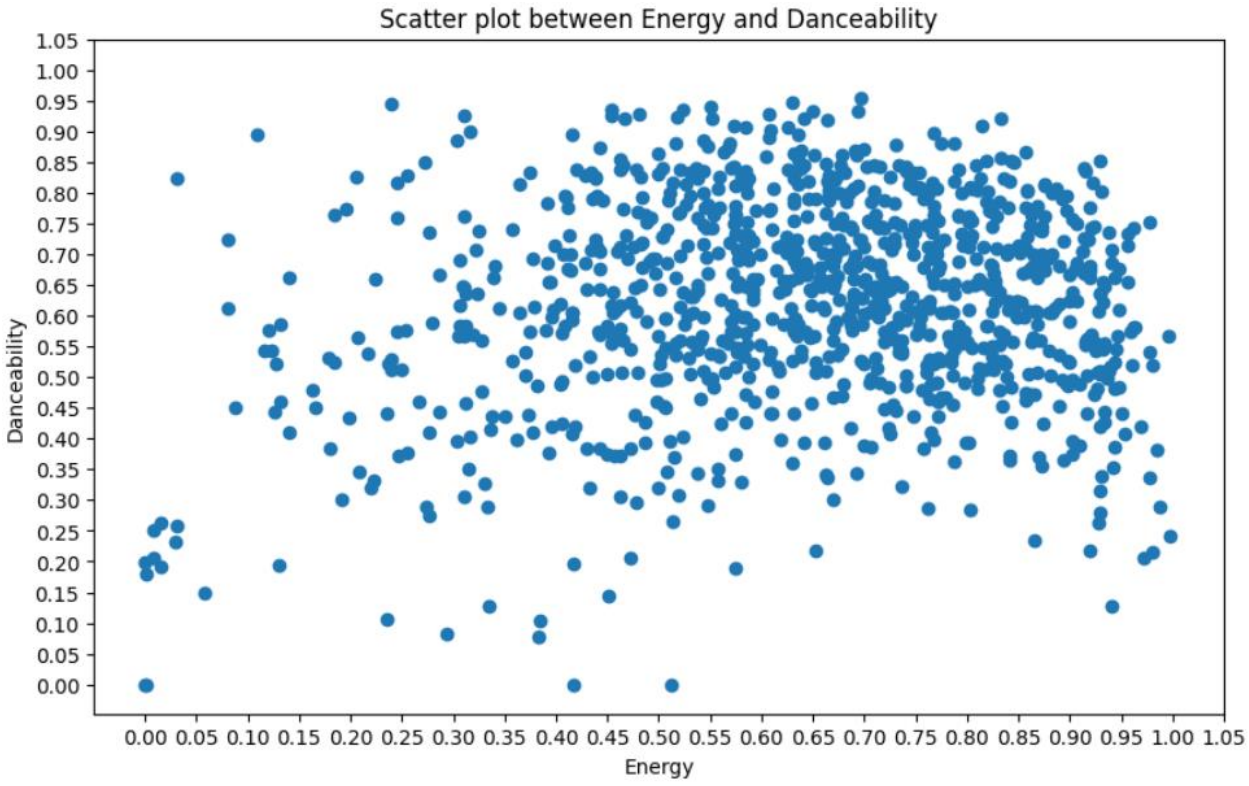
The relationship between acousticness and danceability can be particularly interesting when studying music genres. For example, users who prefer danceable music might not be interested in highly acoustic songs, and vice versa.This graph shows of negative correlation. This means that as one variable increases, the other tends to decrease, and vice versa. In other words, there is an inverse relationship between the two variables and since the gap between the values are more which suggests that the relationship between the two variables is not very strong. The variables do show some tendency to move in opposite directions, but the relationship is not highly pronounced. It also has some outliers and have a higher concentration of data at 0.00.



Energy vs Danceability

The relationship between "energy" and "danceability" can help you identify patterns in your dataset related to different music genres and styles. For example, you might observe that songs in genres like techno or EDM tend to have both high energy and high danceability scores, while ballads or acoustic songs have lower scores in both attributes. The points or graph moving upwards indicates a positive correlation. This means that as one variable increases, the other tends to increase as well, and vice versa. In other words, there is a direct relationship between the two variables.

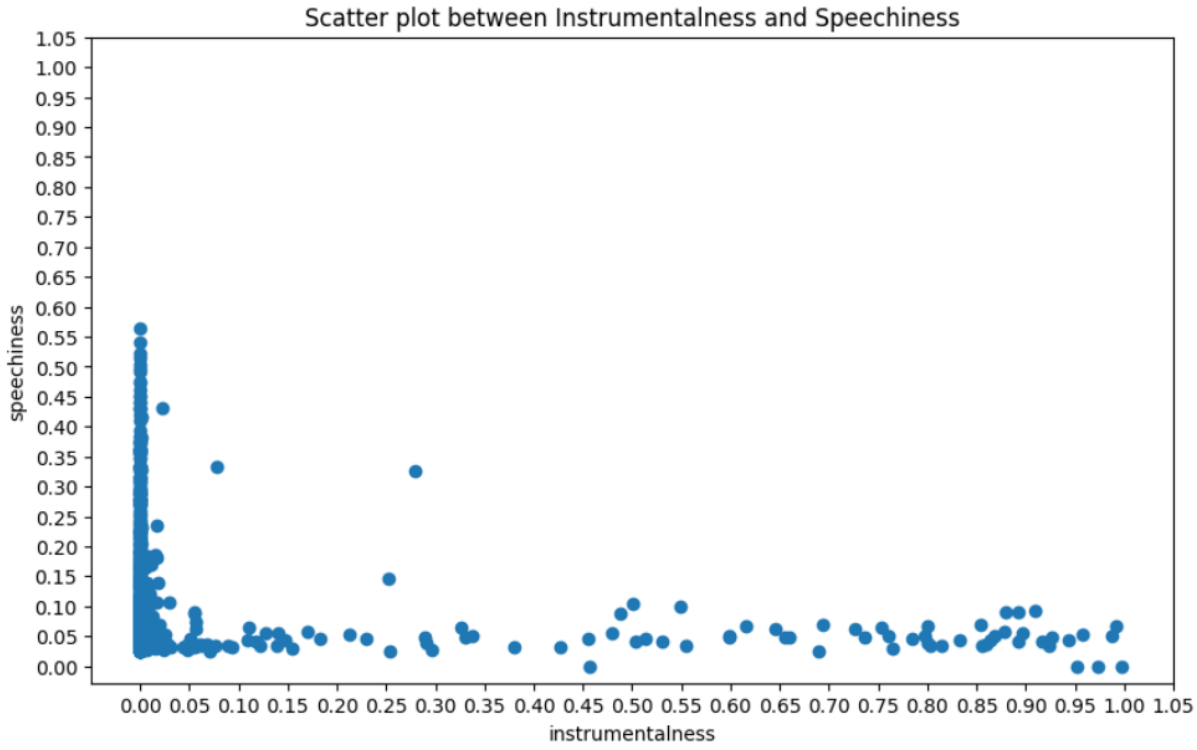
Since the points are somewhat scattered, it suggests that the relationship between the two variables is not very strong. The variables do show some tendency to move in the same direction, but the relationship is not highly pronounced. There are also some outliers, and the data is more concentrated in the range from 0.5 to 0.9.



Speechiness vs Instrumentalness

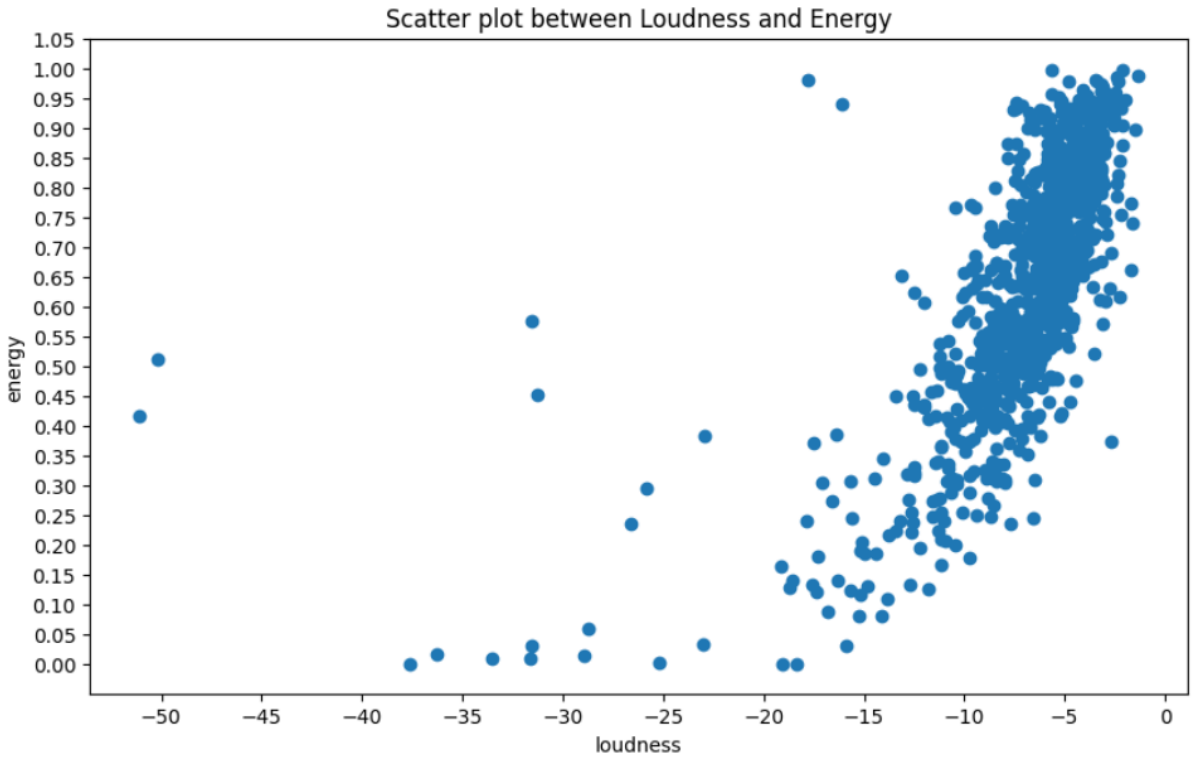
The combination of "speechiness" and "instrumentalness" can provide insights into the lyrical content of songs. For example, in sentiment analysis, one might investigate how the emotional content of lyrics is related to other musical characteristics. The points moving downward from a straight line indicate a negative correlation. This means that as one variable increases, the other tends to decrease slightly, and vice versa. However, the relationship is very weak and not very pronounced, suggesting that there is hardly any discernible relationship between the two variables. The variables do show a slight tendency to move in opposite directions, but the correlation is extremely weak.

There are also some outliers, and the data is more concentrated in the range from 0.0 to 0.025.



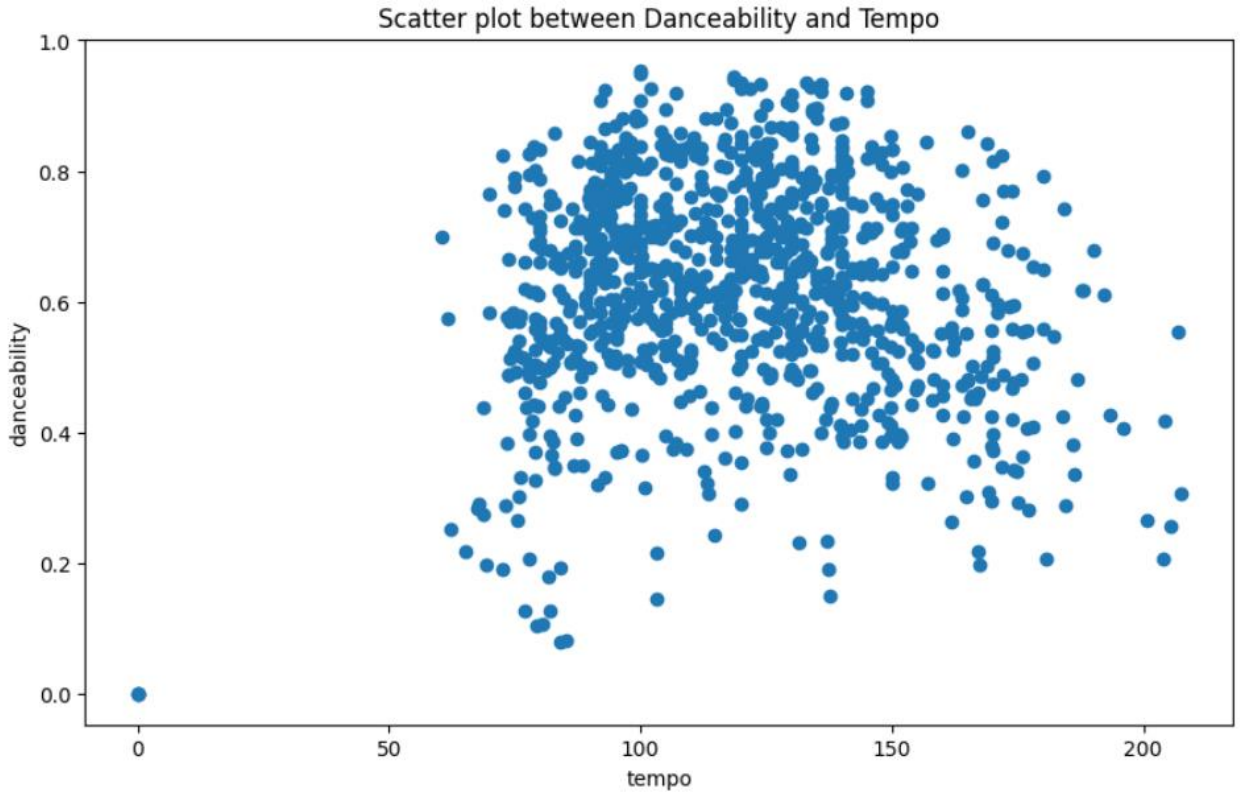
Loudness vs Energy

The relationship between "loudness" and "energy" can help identify patterns related to different music genres. For example, rock and heavy metal genres often have high loudness and high energy, while classical music may have lower loudness and varying energy levels. In our case, a positive loudness value indicates a sound that is louder or more intense, while a negative LU value indicates a softer or quieter sound. The upward-going trend between loudness and energy indicates a positive correlation, signifying that as one variable increases, the other tends to increase as well, and vice versa. In other words, there is a direct and positive relationship between the two variables. The distance between points suggests that the relationship between the two variables is strong and pronounced. The data also includes some outliers and has a higher concentration within the range of -7 to 0.



Danceability vs Tempo

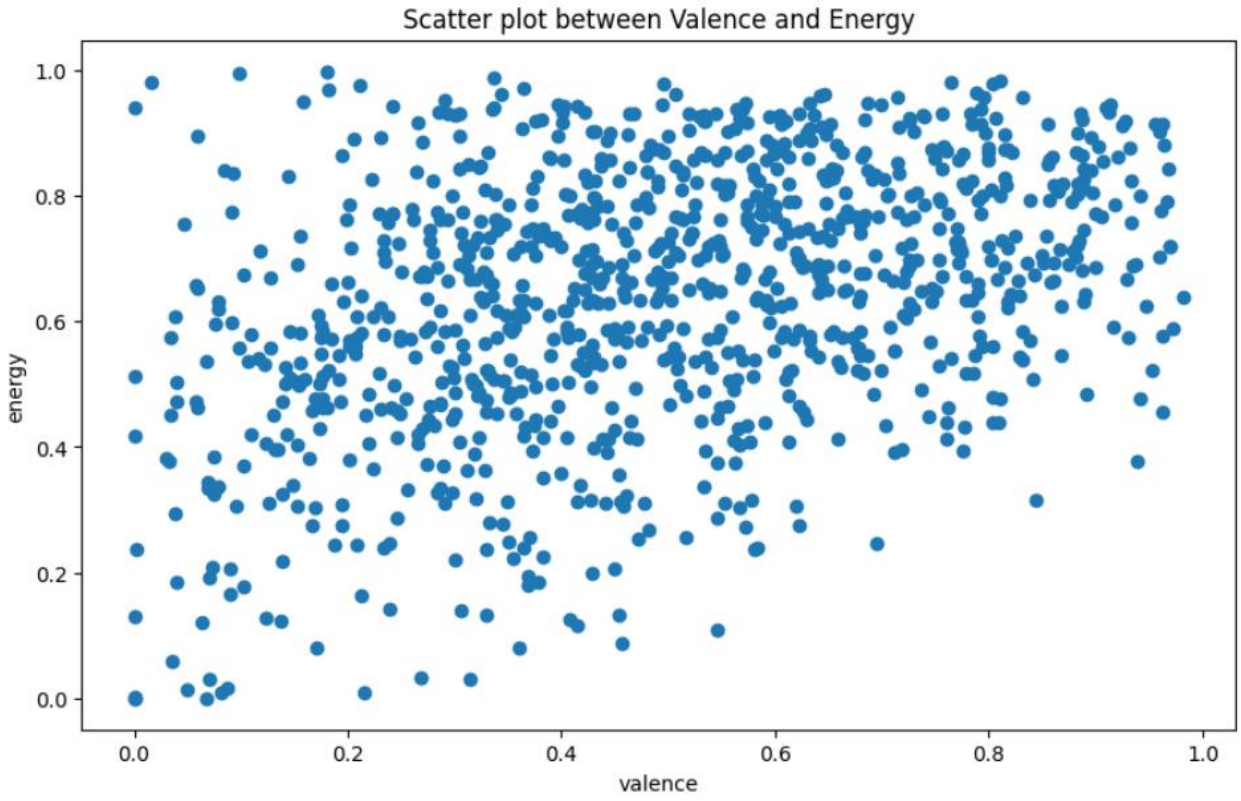
Understanding the relationship between "tempo" and "danceability" can be valuable for music recommendation systems. Users who enjoy dancing to music may appreciate recommendations that consider tempo and danceability. The downward-going points indicate a negative correlation, meaning that as one variable increases, the other tends to decrease very slightly, and vice versa. However, the relationship is almost negligible. Since the points are widely scattered, it suggests that there is virtually no discernible relationship between the two variables. The variables exhibit an almost imperceptible tendency to move in opposite directions. The data also includes some outliers and does not have a higher concentration of data.



Valance Vs Energy

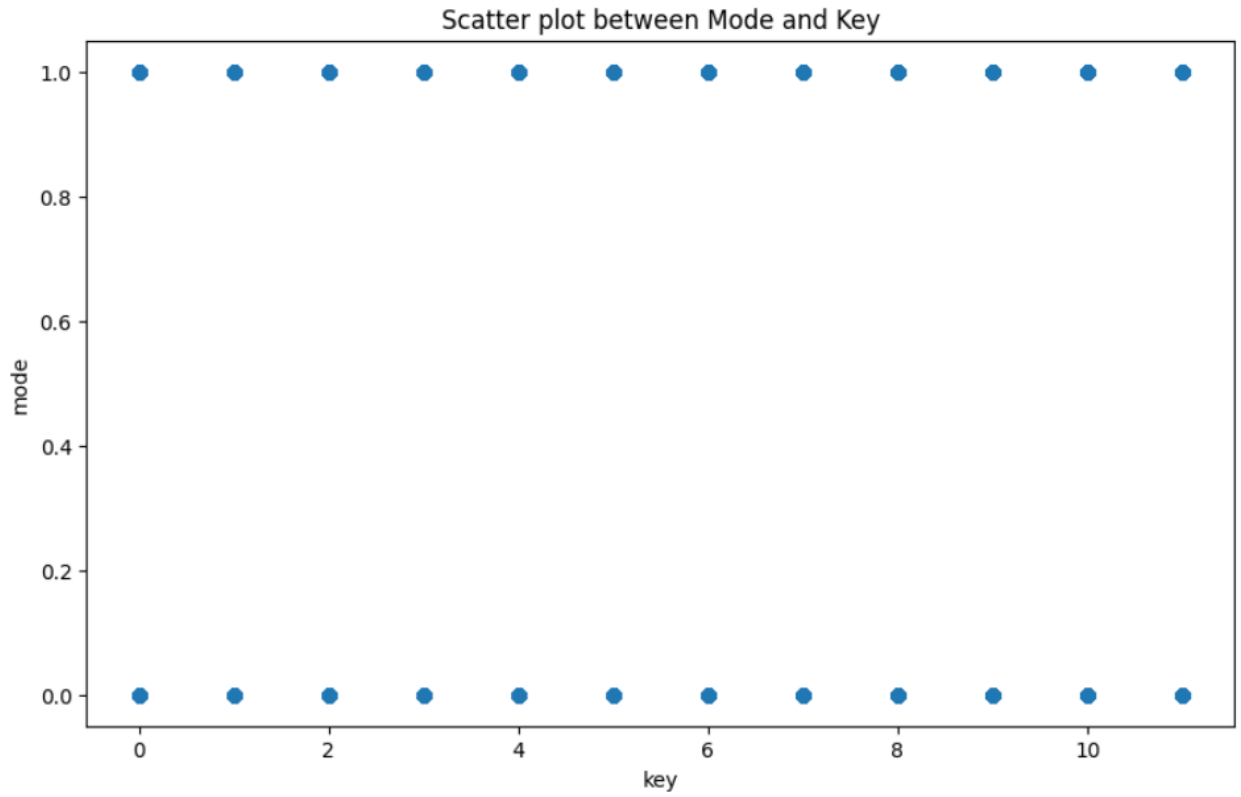
Exploring the connection between "valence" and "energy" proves to be highly valuable for music recommendation systems, particularly in catering to users who favor music with a combination of positive emotional content and high energy. Recommendations that consider the interplay of these attributes can effectively resonate with the preferences of those seeking both positivity and vibrancy in their music choices.

The upward trend of the data points reveals a positive correlation, indicating that as one variable increases, the other tends to follow suit, and as one decreases, the other does as well. This direct and positive relationship between valence and energy is evident, and the spacing between data points suggests a moderate level of strength in this correlation. While it may not be as robust as a perfect correlation of 1, it certainly isn't weak either. The variables exhibit a noticeable propensity to move in the same direction, emphasizing their interconnectedness.



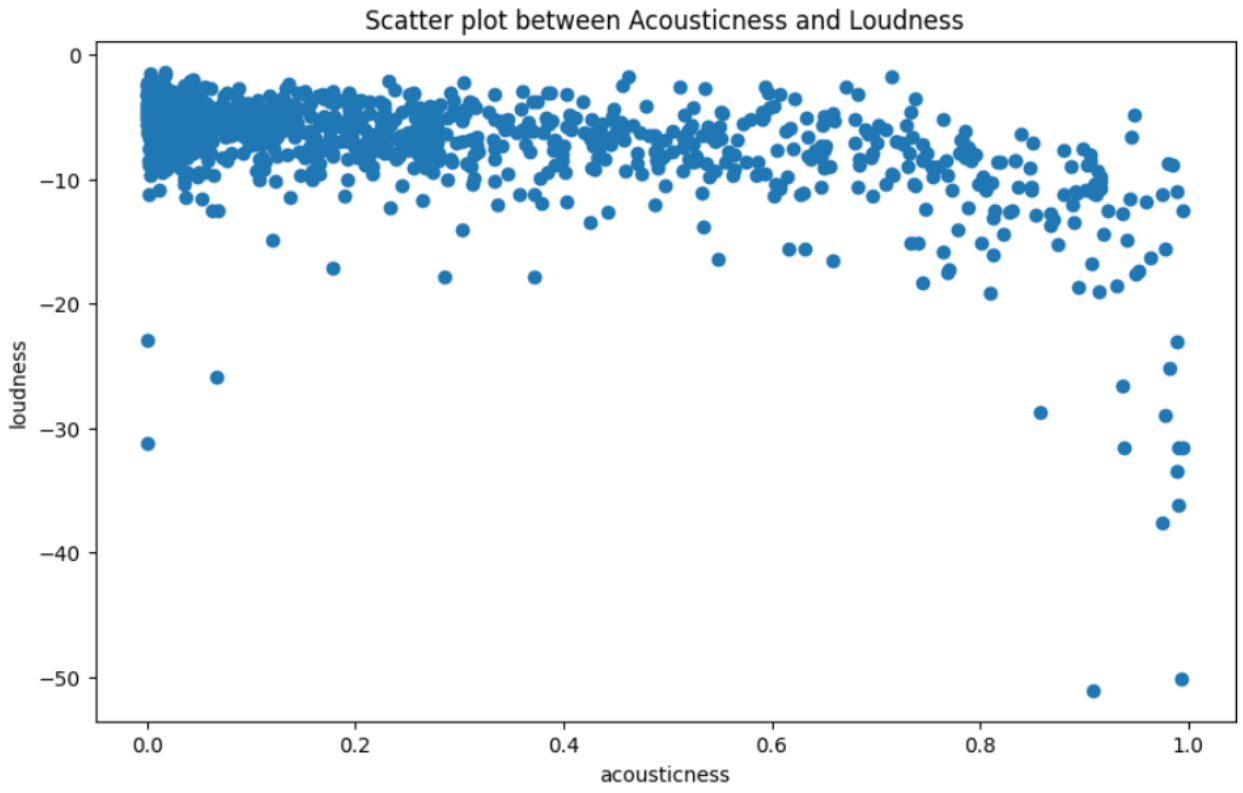
Key Vs Mode

Understanding the relationship between "key" and "mode" can be valuable for analyzing music genres and styles. Certain genres, such as classical and jazz, may explore a wide range of keys and modes, while pop and rock music often adhere to specific tonal centers and modes. The points are slightly moving downwards, indicating a negative correlation, signifying that as one variable increases, the other tends to decrease slightly, and vice versa. However, the relationship is weak and not very pronounced. Since the points are far away, it suggests that the relationship between the two variables is weak and barely noticeable. The variables show a minor tendency to move in opposite directions. There are only two lines along which horizontal points are available, indicating categorical data. It also has some outliers and does not have a higher concentration of data at any point.



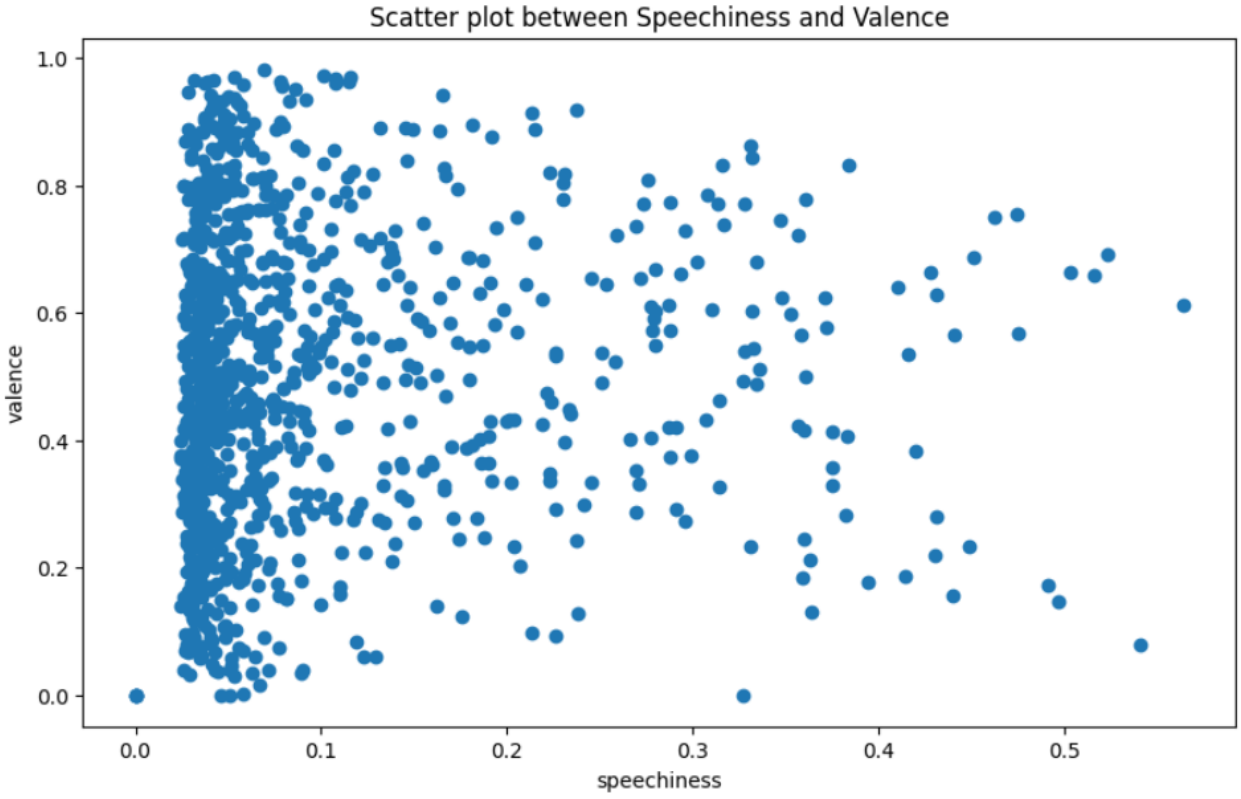
Acousticness vs Loudness

Understanding the relationship between "acousticness" and "loudness" can be valuable for music recommendation systems. Users who prefer acoustic or quieter music may appreciate recommendations that consider these features. The downward movement of points indicates a negative correlation, signifying that as one variable increases, the other tends to decrease, and vice versa. There is a strong and inverse relationship between the two variables. The relatively close proximity of the data points suggests that the relationship is robust and pronounced. When one variable goes up, the other tends to go down with a high degree of certainty. It's also important to note the presence of some outliers within the dataset and a higher concentration of data within the range of 0.0 to 0.2.



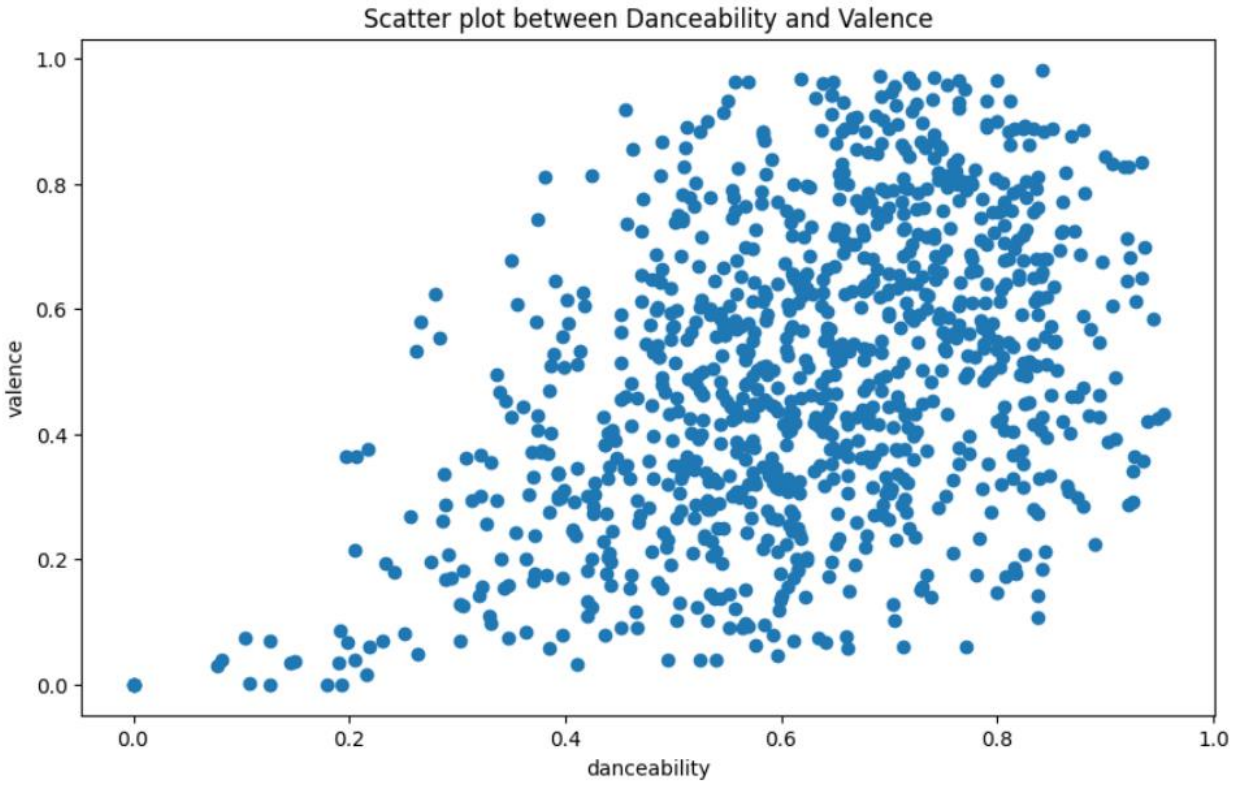
Speechiness vs Valence

Understanding the relationship between "speechiness" and "valence" can be valuable for music recommendation systems, particularly for users who favor songs with lyrics or spoken words and wish to align those choices with specific emotional tones. The predominantly upward trajectory of data points reveals a positive correlation, signifying that as one variable increases, the other tends to increase slightly, and vice versa. However, this correlation is exceedingly subtle, with the points widely dispersed, indicating an almost imperceptible relationship between the two variables, both showing a nearly indiscernible tendency to move in the same direction. Additionally, there are outliers within the dataset, and a notable concentration of data falls within the range of 0.03 to 0.06, offering valuable insights into the interplay of speechiness and valence in music.



Danceability vs Valence

Understanding the relationship between "danceability" and "valence" holds significance for music recommendation systems, particularly for users seeking energetic and positive music to enhance their dancing experience. The consistent upward movement of data points unveils a positive correlation, wherein an increase in one variable corresponds with an increase in the other, and vice versa, indicating a direct and positive connection between these musical attributes. The spacing between the points suggests a moderate strength in this relationship, not as robust as a perfect correlation of 1 but far from being weak. Both variables distinctly exhibit a tendency to move in the same direction, reinforcing their interdependence. It's worth noting the presence of outliers within the dataset, and there isn't a specific concentration of data at any particular point, underlining the versatile nature of the relationship between danceability and valence.



Descriptive Analysis

Correlation Matrix

A correlation matrix is a table that helps us understand how two or more things are related to each other. Let's see the relationship between different features of our dataset by using correlation matrix.

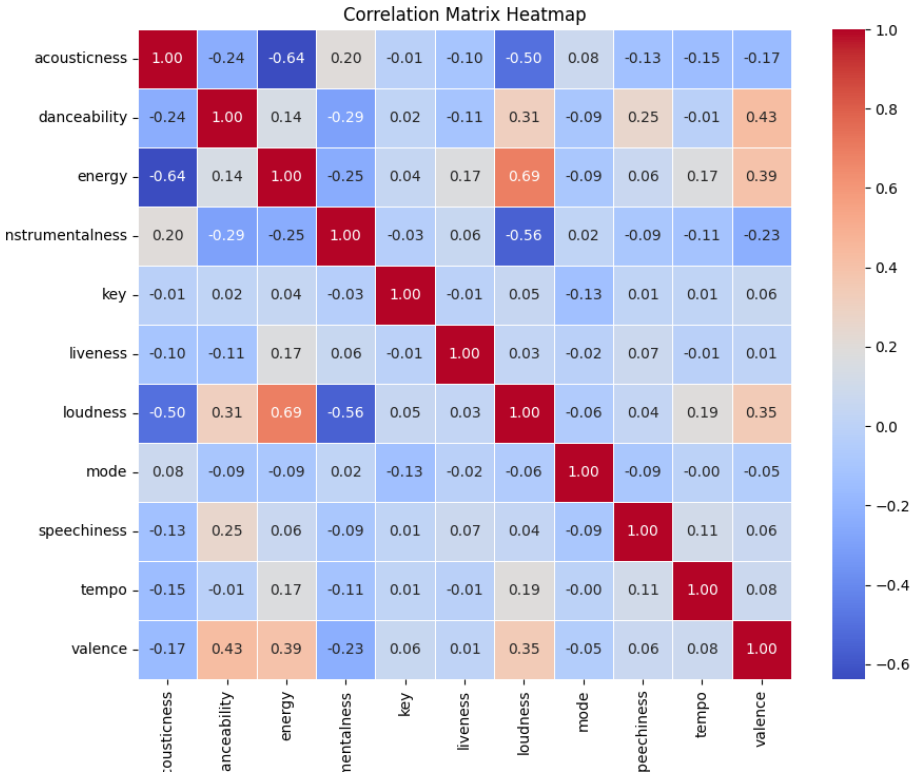


Fig 1. Correlation Matrix

In the generated heatmap, each cell represents the correlation between two variables. The values in each cell indicate the strength and direction of the correlation:

- Positive values (closer to 1) represent a positive correlation
- Negative values (closer to -1) represent a negative
- Values close to 0 indicate a weak or no correlation between variables.

As we can see in correlation matrix, the strong positive correlation is observed between loudness and instrumentalness and strong negative correlation is observed between energy and acoustiness, loudness and instrumentalness.

Statistical Analysis

The provided statistical summary gives us valuable insights into the attributes of a dataset related to music, which includes measures like acousticness, danceability, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, and valence. These insights can help us understand various aspects of the music in the dataset.

Diversity in Acousticness:

The acousticness values vary widely in the dataset, ranging from a minimum of 0 to a maximum of 0.996. This suggests a broad spectrum of music, with some tracks being highly acoustic while others are almost entirely non-acoustic.

Danceability and Energy:

On average, the tracks in the dataset exhibit a moderate level of danceability (mean = 0.622) and energy (mean = 0.632). This indicates that the music is generally suitable for dancing, with a moderate level of energy.

Instrumentalness:

The instrumentalness values have a relatively low mean (0.046) and a wide distribution, suggesting that most tracks contain vocals, but there is some variation, with certain tracks being more instrumental.

Key and Mode:

The key and mode attributes indicate the musical key and mode of the tracks, respectively. The key ranges from 0 to 11, with 0 representing C, and the mode is binary (0 or 1), representing major or minor keys. These values provide information about the musical characteristics of the tracks.

Loudness and Speechiness:

The dataset shows an average loudness of approximately -7.16, which suggests that the music tends to be fairly loud. Speechiness has a mean of 0.623, indicating that a considerable portion of the music contains spoken words or lyrics.

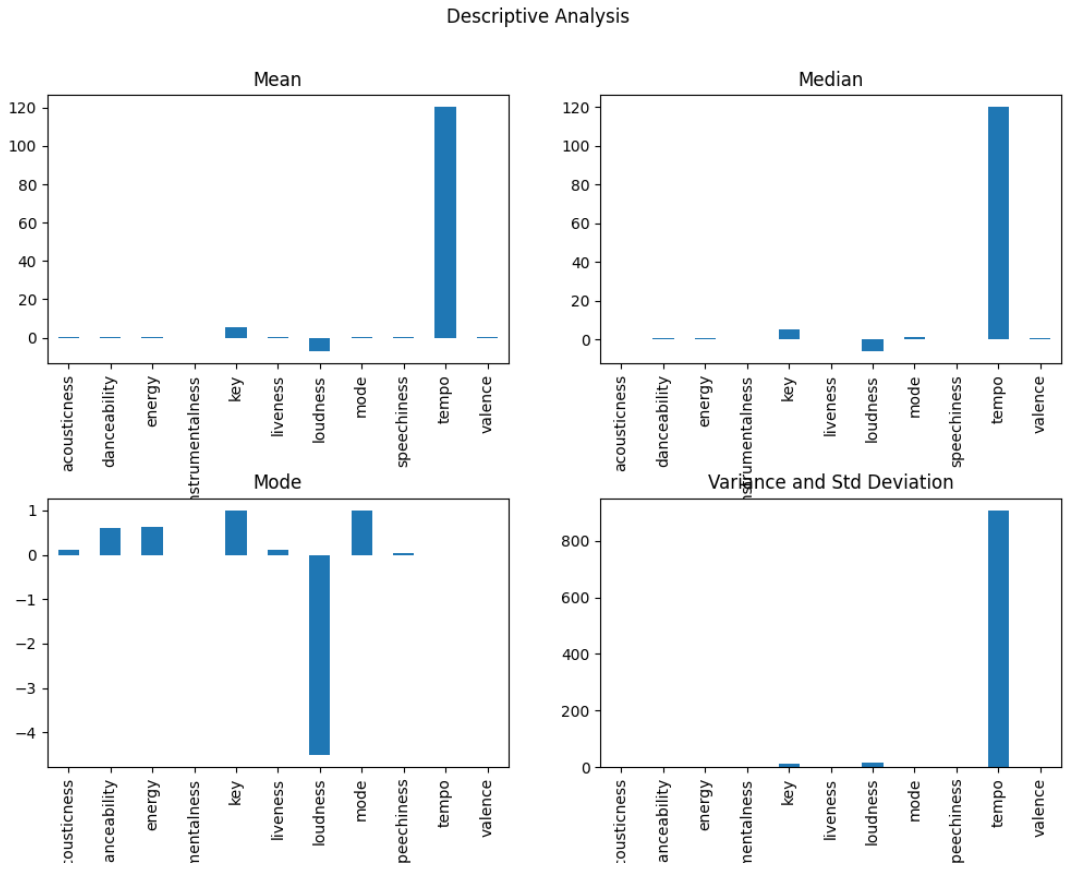
Tempo:

The tempo attribute reveals that the tracks in the dataset have an average tempo of around 120.31 beats per minute (BPM), with a considerable standard deviation. This implies a variety of tempo styles in the music.

Valence:

Valence measures the positivity or happiness of the music. The dataset's valence has a mean of 0.493, indicating that, on average, the music tends to have a moderate positive emotional tone.

In conclusion, this dataset contains a diverse range of music, with variations in attributes such as acousticness, danceability, energy, instrumentalness, key, and tempo. The analysis provides valuable information for music enthusiasts, researchers, and recommendation systems. The data reflects that the music is generally danceable and energetic, with a mix of instrumental and vocal tracks. The variation in musical key and mode, loudness, speechiness, and valence contributes to the richness and diversity of the dataset. These findings can inform music recommendation systems and offer insights into the characteristics of different music genres and styles.



Dataset Description Table

	acousticness	danceability	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	valence
count	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000	15475.000000
mean	0.276240	0.622192	0.632248	0.046076	5.254669	0.182581	-7.155861	0.623845	0.099286	120.306547	0.492853
std	0.287896	0.166186	0.207826	0.178125	3.558730	0.145066	4.152240	0.484435	0.102389	30.102370	0.238314
min	0.000000	0.000000	0.000020	0.000000	0.000000	0.000000	-54.376000	0.000000	0.000000	0.000000	0.000000
25%	0.035250	0.520000	0.499000	0.000000	2.000000	0.096050	-8.289500	0.000000	0.036700	96.915000	0.310000
50%	0.163000	0.637000	0.654000	0.000001	5.000000	0.123000	-6.290000	1.000000	0.054300	119.966000	0.488000
75%	0.456000	0.743000	0.792000	0.000216	8.000000	0.225000	-4.821000	1.000000	0.113000	140.024500	0.680000
max	0.996000	0.980000	1.000000	1.000000	11.000000	0.989000	1.342000	1.000000	0.902000	220.099000	0.990000

K-Means Clustering

The K-Means clustering is applied to song dataset to make clusters of different songs. Each cluster as shown in the below graph contains the songs which are similar to each other. For example, songs with higher energy and dancebility are clustered together. This cluster would "*High Energy Dance Cluster*". Similarly, the songs with low energy, low valance and less danceability are clustered together. These types of songs would be "*Sad Songs*".

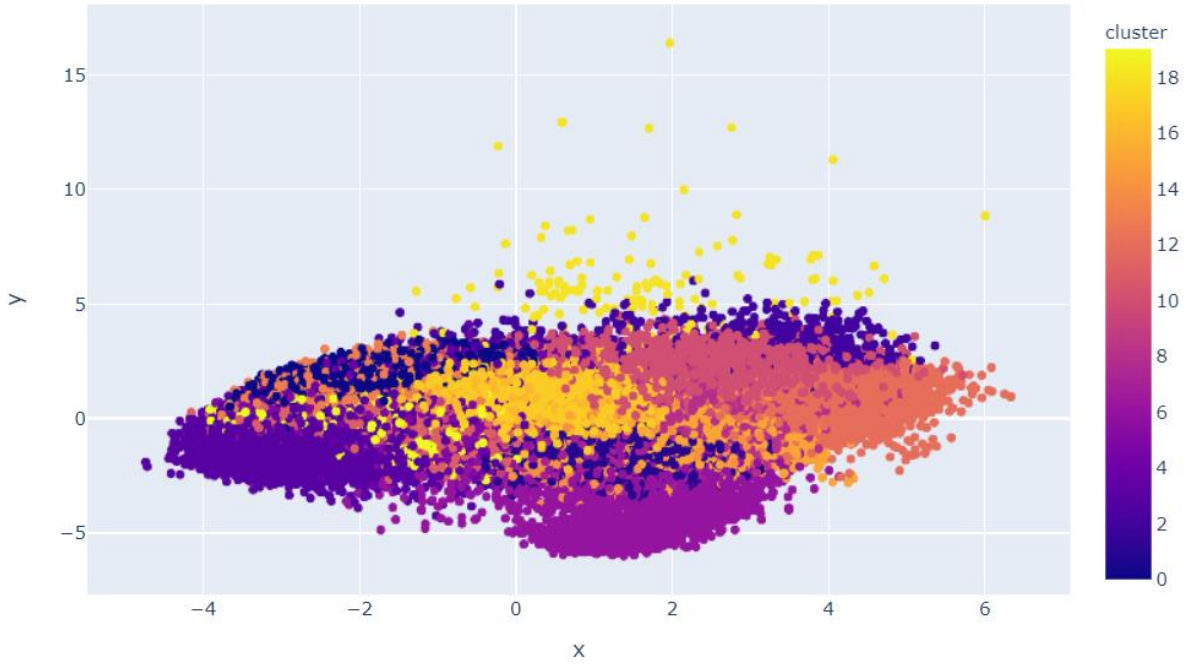


Figure: K-Means Clustering

Conclusion:

In summary, this dataset presents a valuable resource for anyone interested in delving into the world of music analysis. We have done a proper data exploration and visualization techniques that have uncovered hidden patterns and insights within the dataset, contributing to a deeper understanding of the songs, artist and different attribute of audio. We have realized the correlation between loudness and valance, tempo and danceablity and energy and loudness. Additionally, we have seen the word cloud which perfectly portrays the artists having a significant career in music considering the number of songs. In addition to this, it is observed in K-means clustering, how different songs are similar to each other based on different audio feature. Further, the boxplot depicts the outliers and histograms portrays the frequency distribution of the values of different features which will be helpful in training the model for the recommendation engine. Overall, our visualization played an important role in understanding the dataset and finding the patterns among different feature.