Data Analysis on Top songs on Spotify

Reyhaneh Saffar _99222064

1. Data Exploration

This dataset contains information on the top 10,000 songs on Spotify from 1960 to now. It includes details like track name, artist name, album name, release date, popularity, genres, dance ability, energy, and more. This dataset can be a valuable resource for analyzing music trends and characteristics for research and recommendations.

1. **Loading the dataset**: I used pandas to read the CSV file and store it in a data frame. The dataset has 10,000 rows and 35 columns, containing information about the top 10,000 songs on Spotify from 1960 to 2019

   Here is a detailed explanation of each column in the dataset:

   - **Track URI:** A unique identifier for the track on Spotify. (object)
   - **Track Name:** The name of the track. (object)
   - **Artist URI(s):** A unique identifier for the artist(s) on Spotify. (object)
   - **Artist Name(s):** The name(s) of the artist(s). (object)
   - **Album URI:** A unique identifier for the album on Spotify. (object)
   - **Album Name:** The name of the album. (object)
   - **Album Artist URI(s):** A unique identifier for the artist(s) of the album on Spotify. (object)
   - **Album Artist Name(s):** The name(s) of the artist(s) of the album. (object)
   - **Album Release Date:** The date when the album was released. (object)
   - **Album Image URL:** The URL of the album cover image. (object)
   - **Disc Number:** The disc number (usually 1 unless it's a multi-disc album). (int64)
   - **Track Number:** The track number on the album. (int64)
   - **Track Duration (ms):** The duration of the track in milliseconds. (int64)
   - **Track Preview URL:** A URL to a 30-second preview (MP3 format) of the track on Spotify. (object)
   - **Explicit:** Whether or not the track has explicit lyrics (1 = yes, 0 = no). (bool)
   - **Popularity:** A measure of how popular a song is on Spotify (0 to 100). (int64)
   - **ISRC:** International Standard Recording Code, a unique identifier for a specific recording. (object)
   - **Added By:** The Spotify user who added the track to their library or playlist. (object)

- **Added At:** The date and time when the track was added to their library or playlist. (object)
- **Artist Genres:** A list of genres associated with the artist(s). (object)
- **Danceability:** A measure of how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity (0 to 1). (float64)
- **Energy:** A measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy (0 to 1). (float64)
- **Key:** The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. (float64)
- **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 between tracks. (float64)
- **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 represented by 0. (float64)
- **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 . (float64)
- **Acousticness:** A measure of whether or not a track is acoustic. Acoustic tracks are those that primarily feature acoustic instruments as opposed to electronic ones (0 to 1). (float64)
- **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 instrumentalness is to 1.0, the greater likelihood that a track contains no vocal content. (float64)
- **Liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. (float64)
- **Valence:** A measure of 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) (0 to 1). (float64)
- **Tempo:** The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is typically measured in BPM. (float64)
- **Time Signature:** 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. (float64)
- **Album Genres:** A list of genres associated with the album. (float64)

- **Label:** The record label that released this album. (object)
- **Copyrights:** Copyright information for this album. (object)

As it can be suspected based on above attributes, some of the features are useless or illegible, which means they must be discarded during the preprocessing.

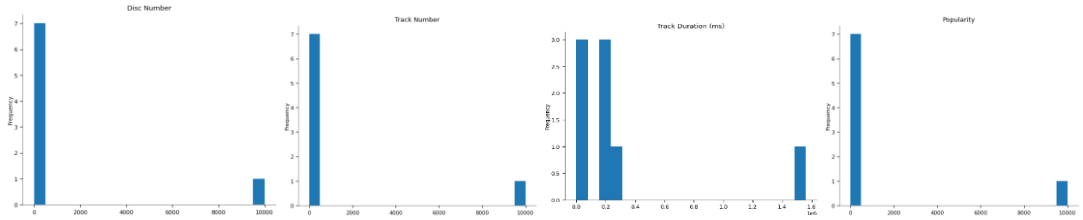
```
df.nunique()
```

```
Track URI               9951
Track Name              8258
Artist URI(s)           4134
Artist Name(s)          4129
Album URI               7462
Album Name              6636
Album Artist URI(s)     3298
Album Artist Name(s)    3294
Album Release Date      3332
Album Image URL         7460
Disc Number               10
Track Number              57
Track Duration (ms)     7320
Track Preview URL       6889
Explicit                   2
Popularity                99
ISRC                    8948
Added By                   1
Added At                 609
Artist Genres           2815
Danceability             779
Energy                   876
Key                       12
Loudness                6329
Mode                       2
Speechiness             1059
Acousticness            2746
Instrumentalness        3028
Liveness                1361
Valence                  994
Tempo                   8621
Time Signature             5
Album Genres               0
Label                   1465
Copyrights              5378
dtype: int64
```

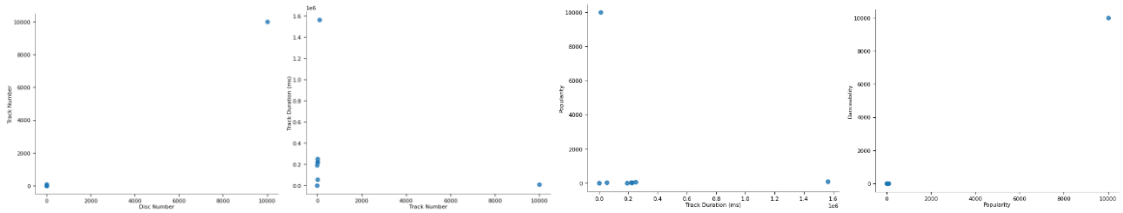
This section shows the number of unique values for each column in the dataset. It can help to identify the diversity and variability of the data, as well as potential issues such as missing values, duplicates, or outliers. For example, the column “Track URI” has 9951 unique values, which means that there are 49 duplicates in the dataset. The column “Album Genres” has zero unique values, which means that it is empty and can be dropped. The column “Explicit” has only two unique values, which means that it is a binary variable indicating whether the track has explicit lyrics or not. The column “Popularity” has 99 unique values, which means that it is a numerical variable ranging from 0 to 100. The column “Artist Genres” has 2815 unique values, which means that it is a list of genres associated with the artist(s).

The df.describe() function is a method in Pandas that generates descriptive statistics for a DataFrame. It provides a summary of the central tendency, dispersion, and shape of the distribution of a dataset, excluding NaN values. The output includes the count, mean, standard deviation, minimum, maximum, and quartile values for each numeric column in the DataFrame.

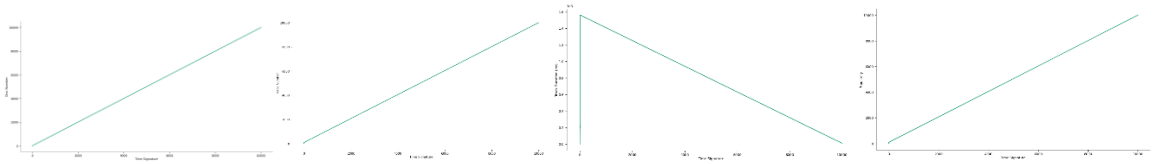
1. **Distributions:**



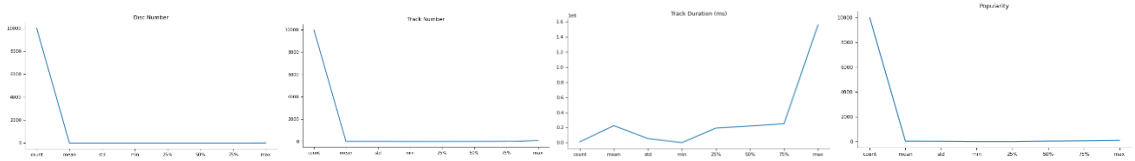
2. **2D- Distributions:**



3. **Time Series:**



4. **Values:**



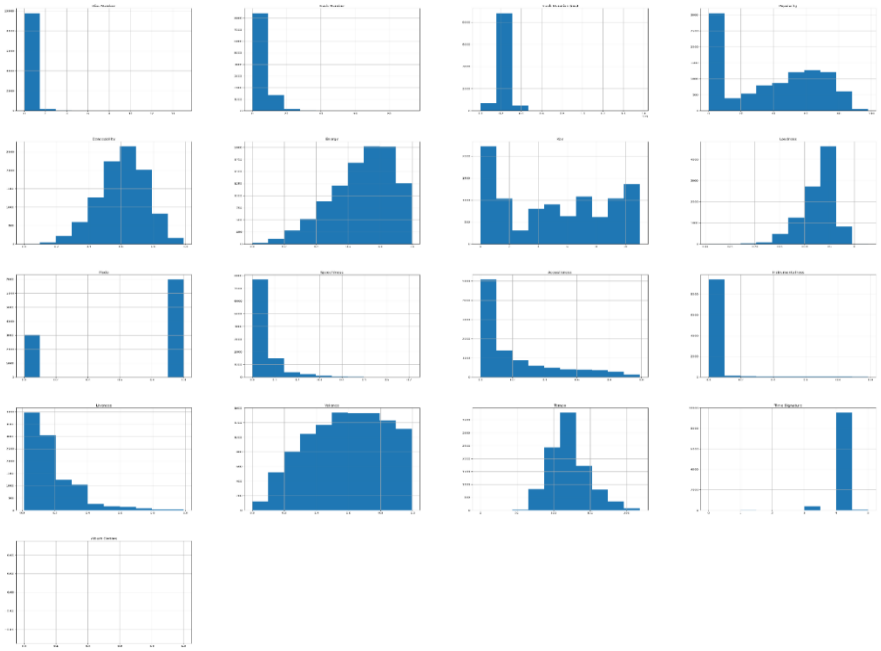
2. Data Processing:

Handling missing values: I checked for any missing values in the dataset using :

```
pd.DataFrame(df[columns_].isna().sum(), columns=['Number of null values'])
```

This table shows the number of null values and their corresponding percentages for each column in the dataset. Null values can indicate missing or incomplete data, which can affect the accuracy and reliability of your analysis. For example, the column “Track Preview URL” has 2897 null values, which means that almost 29% of the data is missing in this column. The column “Album Genres” has 9999 null values, which means that it is empty and can be dropped. The column “Popularity” does not appear in this table, which means that it does not have any null values.

index	Number of null values	Percentage
Track Name	1	0.01
Artist URI(s)	2	0.02
Artist Name(s)	1	0.01
Album URI	2	0.02
Album Name	1	0.01
Album Artist URI(s)	2	0.02
Album Artist Name(s)	2	0.02
Album Release Date	2	0.02
Album Image URL	4	0.04
Track Preview URL	2897	28.97
ISRC	3	0.03
Artist Genres	550	5.5
Danceability	2	0.02
Energy	2	0.02
Key	2	0.02
Loudness	2	0.02
Mode	2	0.02
Speechiness	2	0.02
Acousticness	2	0.02
Instrumentalness	2	0.02
Liveness	2	0.02
Valence	2	0.02
Tempo	2	0.02
Time Signature	2	0.02
Album Genres	9999	100.0
Label	6	0.06
Copyrights	24	0.24



3. **Statistical Tests and Analysis:**

- **The Shapiro-Wilk test:**

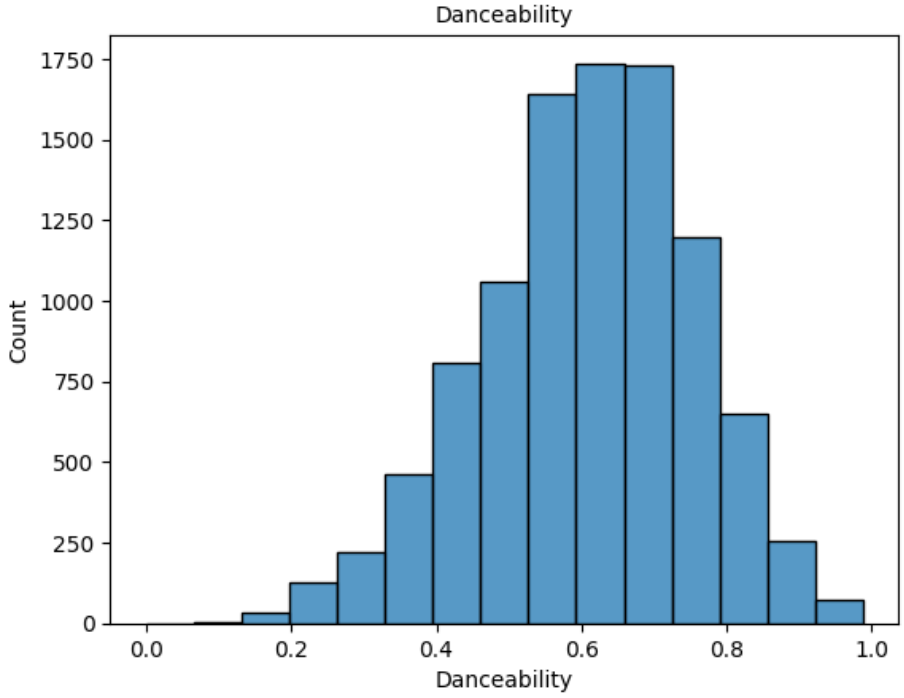
  is a statistical test used to determine whether a given dataset follows a Gaussian (normal) distribution. I have applied the test on different features like to the "Danceability" feature. Here's what the result means:

  Test Statistic (stat): the test statistic (stat) is approximately 0.993.
  P-value (p): The p-value is approximately 0.993.

  Now, let's interpret this result:

  Null Hypothesis (H0): The null hypothesis of the Shapiro-Wilk test is that the data follows a Gaussian (normal) distribution.
  Alternative Hypothesis (H1): The alternative hypothesis is that the data does not follow a Gaussian (normal) distribution.

  P-value (p): The p-value is approximately 0.993. In other words, based on this test, it suggests that the "Danceability" feature in the dataset is not likely normally distributed.



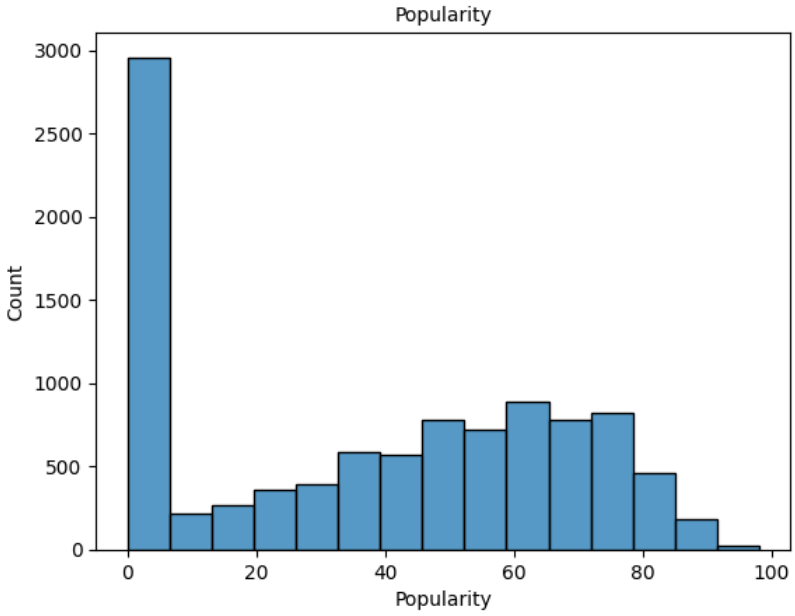
I also have applied the test to the " Popularity" feature. Here's what the result means:

Test Statistic (stat): the test statistic (stat) is approximately 0.888.
P-value (p): The p-value is approximately 0.888.

Now, let's interpret this result:

Null Hypothesis (H0): The null hypothesis of the Shapiro-Wilk test is that the data follows a Gaussian (normal) distribution.
Alternative Hypothesis (H1): The alternative hypothesis is that the data does not follow a Gaussian (normal) distribution.

P-value (p): The p-value is approximately 0.888. In other words, based on this test, it suggests that the " Popularity" feature in the dataset is not likely normally distributed.



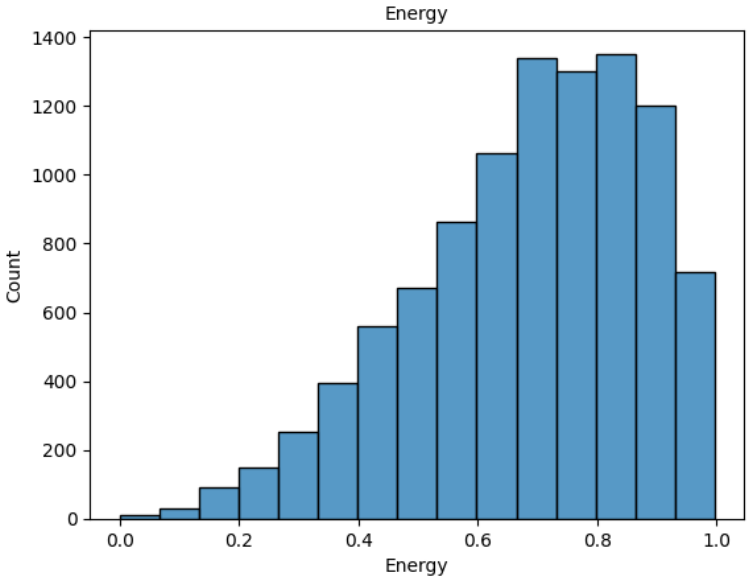
I also have applied the test to the " Energy" feature. Here's what the result means:

Test Statistic (stat): the test statistic (stat) is approximately 0.960.
P-value (p): The p-value is approximately 0.960.

Now, let's interpret this result:

Null Hypothesis (H0): The null hypothesis of the Shapiro-Wilk test is that the data follows a Gaussian (normal) distribution.
Alternative Hypothesis (H1): The alternative hypothesis is that the data does not follow a Gaussian (normal) distribution.

P-value (p): The p-value is approximately 0.960. In other words, based on this test, it suggests that the " Energy" feature in the dataset is not likely normally distributed.



- **Correlation Analysis Report:**

  The objective of this analysis is to determine whether there is a statistically significant correlation between the 'Danceability' and 'Energy' features in the dataset. The null hypothesis (H0) states that the samples are correlated, while the alternative hypothesis (H1) states that the samples do not have any correlation.

  To test the correlation between 'Danceability' and 'Energy', we used Pearson's Correlation Coefficient, which measures the linear relationship between two continuous variables.

  After conducting the correlation analysis, we obtained the following results:

  - Pearson's Correlation Coefficient (stat) = 0.135
  - p-value (p) = 0.135

  1. Correlation Coefficient (stat): The Pearson's Correlation Coefficient (stat) is approximately 0.135. This value represents the strength and direction of the linear relationship between 'Danceability' and 'Energy'. A positive value indicates a positive correlation, while a negative value would indicate a negative correlation. In this case, the positive value suggests a weak positive correlation.
  2. p-value (p): The p-value obtained from the test is also approximately 0.135.

  Based on the results of the correlation analysis, we make the following conclusions:

  The analysis suggests that there is weakley statistically significant correlation between 'Danceability' and 'Energy' based on the given data. Further analysis or a larger sample size may be needed to draw more conclusive results.

- **ANOVA test**:

  The objective of this analysis is to investigate whether there is a statistically significant difference in the popularity of the top 10,000 songs on Spotify between the decades from 1960 to the present. We will use an ANOVA test to examine this.

  We conducted an Analysis of Variance (ANOVA) test to determine whether there is a significant difference in song popularity across different decades. To do this, we divided the data into the following decades: 1960s, 1970s, 1980s, 1990s, 2000s, and 2010s.

  Results:

  The ANOVA test yielded the following results:

  F-value: 36.44

  P-value: 4.24e-37 (very close to zero)

  1. F-value: The F-value measures the ratio of the variance between the group means to the variance within the groups. In this case, the F-value is 36.44, indicating that there is a substantial difference in song popularity between the decades.

  2. P-value: The P-value is extremely close to zero (4.24e-37), which is well below the commonly used significance level of 0.05. This suggests strong evidence to reject the null hypothesis, indicating that there is a statistically significant difference in song popularity between the decades.

  Based on the results of the ANOVA test, we can confidently conclude that there is a statistically significant difference in the popularity of the top 10,000 songs on Spotify between the decades from 1960 to the present. The P-value, being very close to zero, provides strong evidence to support this conclusion.

- **Correlation Analysis between Danceability and Energy:**

  The primary objective of this analysis is to investigate whether there is a significant correlation between two musical attributes: "Danceability" and "Energy."

  Null Hypothesis (H0): There is no significant correlation between danceability and energy in Spotify's top 10,000 songs.

  Alternative Hypothesis (H1): There is a significant correlation between danceability and energy in Spotify's top 10,000 songs.

  The Pearson correlation coefficient for the danceability and energy attributes in the dataset is approximately 0.1346. The p-value associated with this correlation is approximately 1.2943e-41.

  1. Correlation Coefficient: The correlation coefficient (r) of 0.1346 suggests a positive but relatively weak correlation between danceability and energy. This indicates that there is some degree of association between these two attributes, but it is not particularly strong.

  2. P-value: The p-value of 1.2943e-41 is significantly smaller than the conventional significance level (e.g., $\alpha = 0.05$). Therefore, we reject the null hypothesis (H0) in favor of the alternative hypothesis (H1), indicating that there is a statistically significant correlation between danceability and energy in Spotify's top 10,000 songs.

  Based on the results of the Pearson correlation test, we conclude that there is a statistically significant but relatively weak positive correlation between danceability and energy in the top 10,000 songs on Spotify from 1960 to the present. While this relationship exists, it is not strong enough to make precise predictions or draw definitive conclusions about the causal relationship between these attributes.

- T-test:

  we investigate whether there is a statistically significant difference in the popularity of songs across different music genres. The analysis employs a t-test to compare the popularity of songs in five selected genres: pop, rock, hip hop, country, and jazz.

  Hypotheses:
  Null Hypothesis (H0): There is no significant difference in popularity between the genres.
  Alternative Hypothesis (H1): There is a significant difference in popularity between the genres.

  Results:
  The t-value is approximately 3.3691.
  The p-value is approximately 0.0008.

  The t-value is 3.37 and the p-value is 0.00076. Since the p-value is less than 0.05, we can reject the null hypothesis and conclude that there is a statistically significant difference in popularity between the genres. The t-value indicates the difference between the means of the two groups relative to the variance within the groups. A higher t-value indicates a greater difference between the means of the two groups.

- **Mann-Whitney U test**:

  The objective of this analysis is to determine whether there is a significant difference in the popularity of songs with different energy levels. We used a Mann-Whitney U test to compare the means of two independent groups based on energy levels.

  Methodology:

  1. Imported required libraries: Pandas and SciPy's Mann-Whitney U test.
  2. Cleaned the dataset by removing rows with missing data in the "Popularity" and "Energy" columns.
  3. Split the data into two groups:
     - **High Energy:** Songs with an energy level greater than 0.712.
     - **Low Energy:** Songs with an energy level less than or equal to 0.712.
  4. Performed a Mann-Whitney U test to compare the popularity of these two groups.

  The Mann-Whitney U test yielded the following results:

  Statistics: 12,349,741.000

  p-value: 0.318

  With a p-value of 0.318, we fail to reject the null hypothesis at a typical significance level of 0.05. This means that there is no significant difference in the popularity of songs with different energy levels. In other words, we do not have enough evidence to conclude that songs with high energy levels are more or less popular than songs with low energy levels.

- **Kruskal-Wallis test:**

  The objective of this analysis is to determine whether there is a significant difference in the popularity of songs with different acousticness scores. We conducted a Kruskal-Wallis test, a non-parametric test suitable for comparing the medians of more than two independent groups, to investigate this relationship.

  We divided the dataset into two groups:

  High Acousticness Group: Songs with an acousticness score greater than 0.0956.

  Low Acousticness Group: Songs with an acousticness score less than or equal to 0.0956.

  The Kruskal-Wallis test yielded the following results:

  Test Statistic (Statistics): 1.584

  p-value (p): 0.208

  Based on the results of the Kruskal-Wallis test, we conclude that there is no significant difference in popularity between songs with different acousticness scores. In other words, the acousticness score does not appear to have a statistically significant impact on song popularity within the specified groups.