Analysis of Spotify Dataset

Sarah Hosseini Feshtami - 400222026

Shahid Beheshti University

1. Data Exploration

The "Top 10000 Songs on Spotify 1960-Now" dataset sourced from Kaggle provides a comprehensive collection of songs spanning several decades. This dataset comprises a rich compilation of attributes and features associated with each song. The exploration of this dataset offers a valuable opportunity to gain insights into the evolution of popular music over time. The dataset likely contains a wide range of variables, including song title, artist, release year, duration, danceability, energy, popularity, and many more. By examining the distribution and relationships among these variables, we can uncover trends, patterns, and characteristics that define the most successful and influential songs in the Spotify platform. This data exploration aims to shed light on the key attributes that contribute to a song's popularity and to identify any notable shifts or preferences in musical styles and genres throughout the years. The dataset contains the following attributes:

- ‘Track URI’
- 'Track Name'
- 'Artist URI(s)'
- ‘'Artist Name(s)',
- 'Album URI',
- 'Album Name'
- 'Album Artist URI(s)’
- 'Album Artist Name(s)'
- 'Album Release Date'
- ’'Album Image URL'
- 'Disc Number'
- 'Track Number'
- 'Track Duration (ms)'
- ‘'Track Preview URL'
- 'Explicit'
- ‘'Popularity'
- ‘'ISRC'
- ’'Added By’
- 'Added At'
- 'Artist Genres'

- 'Danceability'
- 'Energy'
- 'Key'
- 'Loudness'
- 'Mode'
- ‘'Speechiness'
- 'Acousticness'
- 'Instrumentalness'
- 'Liveness'
- 'Valence'
- 'Tempo'
- 'Time Signature'
- 'Album Genres'
- 'Label'
- 'Copyrights'

2. Exploratory Data Analysis (EDA):

We perform an analysis on the DataFrame to determine the number of unique values present in each column. By examining the unique values, we gain insights into the diversity and variety within the dataset. The code iterates through each column in the DataFrame and calculates the number of unique values using the len(df[c].unique()) expression.

To enhance readability, the code categorizes the columns as either categorical or non-categorical based on the number of unique values. If a column has 20 or fewer unique values, it is considered categorical. Upon identifying a categorical column, the code appends the label "CATEGORICAL" to the output. For each column, the code then displays the column name followed by the corresponding count of unique values.

This analysis provides valuable information about the uniqueness and distribution of values within the dataset. By understanding the number of unique values in each column, we can identify categorical variables that may require specific handling or further analysis.

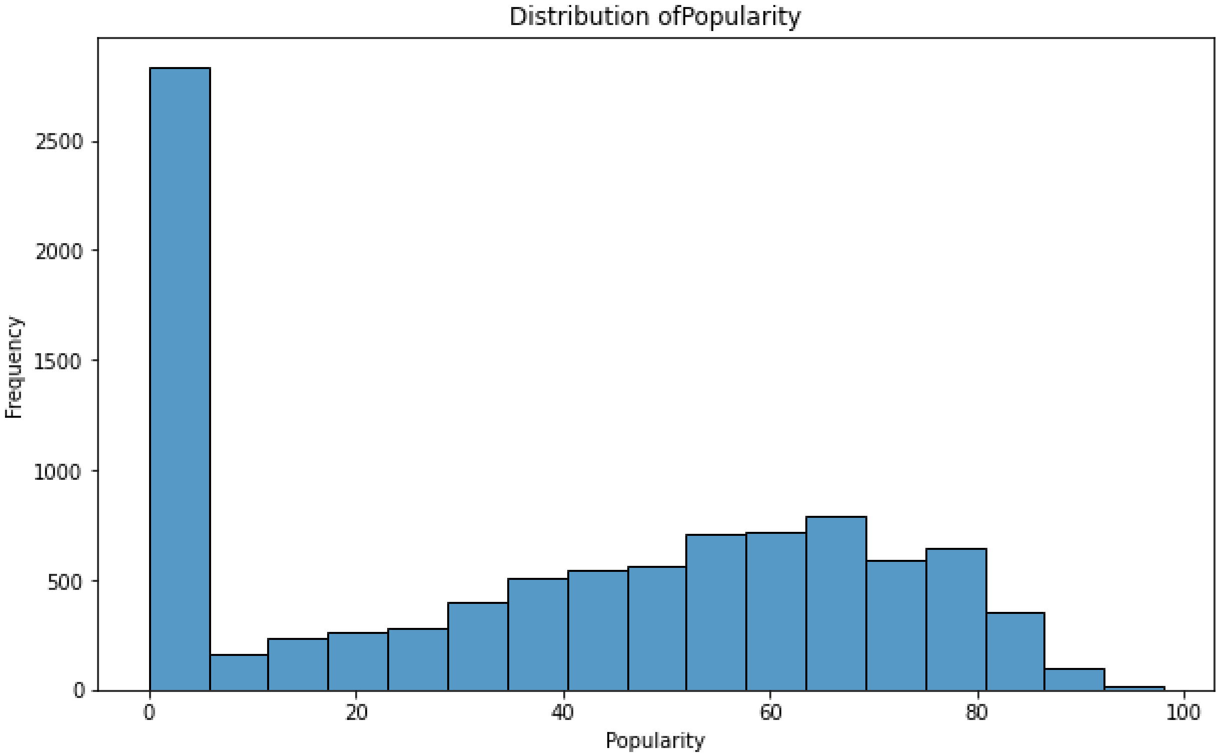
As can be seen in this analysis, the ‘Added by’ and ‘Album Genres’ columns were useless since they only respectively had one and zero unique values. So we drop those two from our dataframe.

Then, we look for missing values. We fill the numerical missing values with the mean of that column and non-numerical ones with the mode.

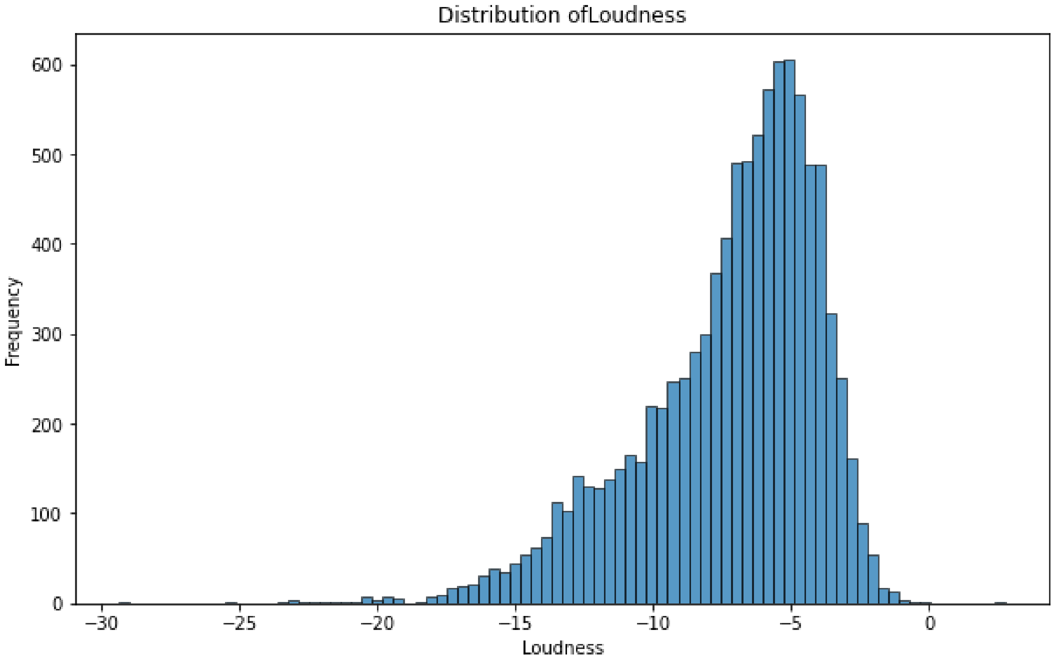
Finally, we remove the outliers which are data that were 3 standard deviations further from the mean.

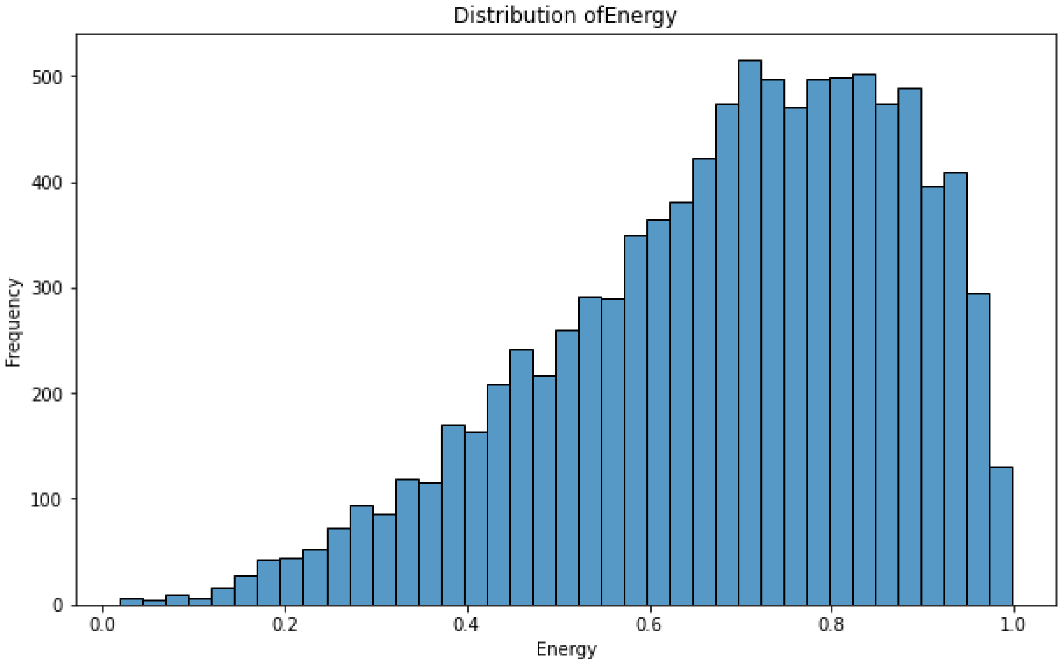
3. Exploratory Visualization

The following plot shows the distribution of popularity in the dataset. As shown below, the majority of data is in the least popular songs.

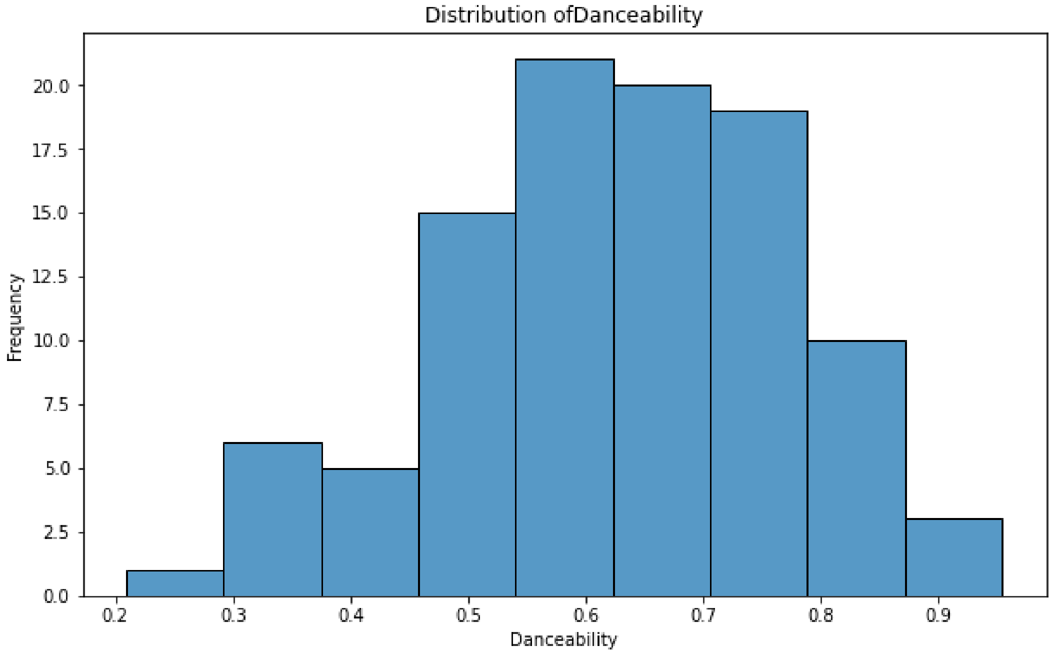


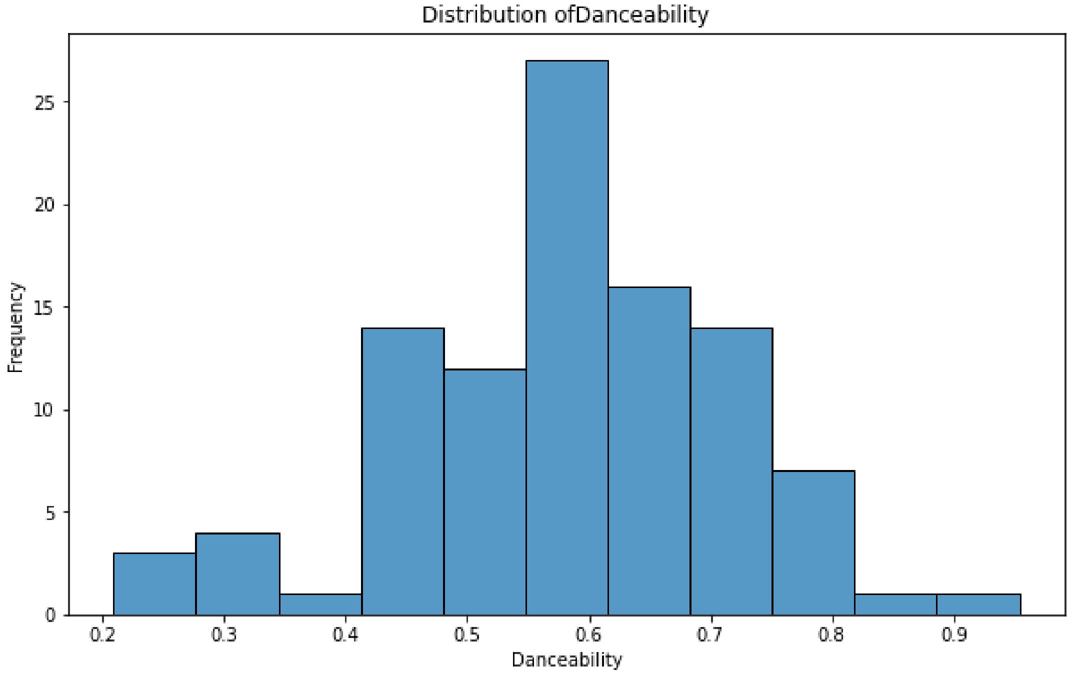
The plots below demonstrate the distribution of loudness and energy. We can infer that most data is energetic and generally loud.





We also grouped the top 100 and low 100 songs and plotted Danceability for each. We can probably say the top songs have a higher danceability.

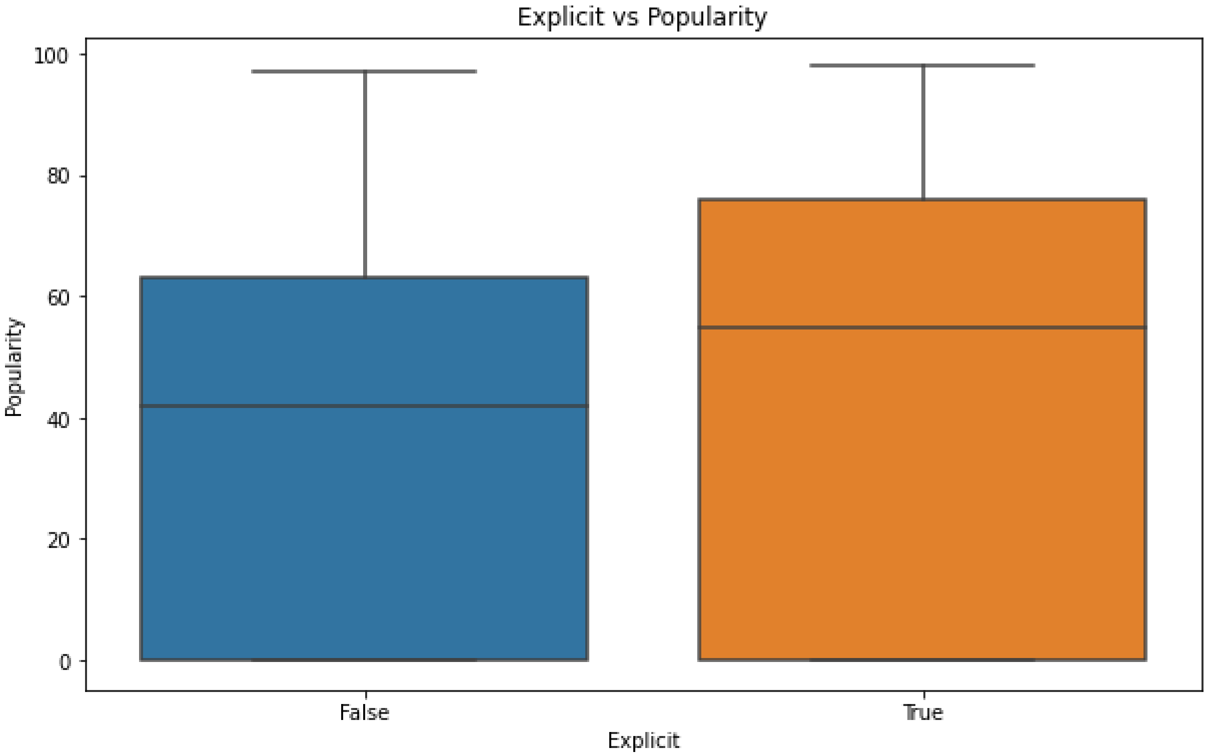




4. Statistical Tests and Analysis

4.1. Testing for the the difference of popularity between explicit and non-explicit songs

To visualize the relationship between explicitness and popularity, we create a box plot using sns.boxplot() from seaborn. This plot shows the median, quartiles, and outliers of the popularity for each state of ‘Explicit’. The x-axis represents the presence (orange) and absence of explicit content (blue), and the y-axis represents the popularity.



We guess that popularity of explicit songs are higher. Now we perform a test to reject or validate this claim. This test focuses on examining the variances in mean values across different explicit states based on popularity. The DataFrame was grouped according to the 'Explicit' column, and the mean popularity values were compared between the groups using an ANOVA (Analysis of Variance) test.

The purpose of the ANOVA test is to determine if there are significant differences in means among the groups. In this analysis, the F-statistic and its corresponding p-value were calculated using the f_oneway() function from the scipy.stats module.

For the provided dataset, the ANOVA test yielded an F-statistic of 29.63 and a p-value of e−08. The F-statistic measures the ratio of between-group variability to within-group variability, while the p-value indicates the statistical significance of the observed mean differences.

Based on the results obtained, when the p-value is lower than the chosen significance level (typically 0.05), it suggests the presence of significant mean differences among the groups. Conversely, if the p-value exceeds the significance level, it indicates insufficient evidence to conclude that the means are different.

In this analysis, the obtained p-value is smaller than the significance level. Consequently, we reject the null hypothesis that the means of the target variable are identical across the groups. This implies that there are likely significant mean differences among the groups.

4.2. Testing for the relationship between popularity and loudness

This is an analysis of the potential relationship between the a song's loudness and its popularity. Based on our analysis, we obtained a correlation coefficient of 0.03 between loudness and popularity. The associated p-value was calculated as 0.001. The p-value is commonly used to determine the statistical significance of a relationship.

Upon conducting a hypothesis test, we found that the p-value less than the significance level of 0.05. Therefore, we reject the null hypothesis that the loudness and popularity are independent. This indicates that there is likely a significant relationship between loudness and popularity.

4.3. Testing for the higher danceability of top 100 songs in comparison to low 100 songs

Here we focus on examining the variances in mean values across different top 100 and lowest 100 songs, based on Danceability. The DataFrame was grouped into more popular and less popular songs, and the mean Danceability values were compared between the groups using an ANOVA (Analysis of Variance) test.

The ANOVA test yielded an F-statistic of 4.5 and a p-value of 0.03. Based on these results, since the obtained p-value is smaller than the significance level, we reject the null hypothesis that the means of the Danceability are identical across the groups. This implies that there are likely significant mean differences among the groups.

4.4. the correlation of Track Duration (ms) and popularity

This is an analysis of the potential relationship between Track Duration (ms) and popularity. Based on our analysis, we obtained a correlation coefficient of 0.02 betweenTrack Duration (ms) and popularity. The associated p-value was calculated as 0.008. The p-value is commonly used to determine the statistical significance of a relationship.

Upon conducting a hypothesis test, we found that the p-value was less than the significance level of 0.05. Therefore, we reject the null hypothesis that Track Duration (ms) and popularity are independent.

4.5. testing whether track duration distribution is Gaussian

We perform an analysis of the normality of a variable using a normality test. The variable under consideration is 'Track Duration (ms)' from the given DataFrame. The normality test used in this analysis is the normaltest from the scipy.stats module.

The normaltest calculates a statistic and p-value to assess whether the variable follows a Gaussian (normal) distribution. In this analysis, the normaltest was conducted on the 'Track Duration (ms)' variable, resulting in a statistic value of 307 and a p-value of e-67..

To interpret the results, the p-value is compared to a significance level (commonly chosen as 0.05). If the p-value is greater than the significance level, it suggests that the variable is likely to follow a Gaussian distribution. On the other hand, if the p-value is less than the significance level, it indicates that the variable is probably not Gaussian.

For the given dataset, the normaltest resulted in a p-value of e-67. Since the p-value is greater than the significance level, we cannot conclude that the Track Duration (ms) variable is normally distributed.