**Spotify Song Features: A Visual Analysis**

**Problem Statement:**

This project involves visually exploring Spotify song features, including duration, loudness, tempo, energy, and liveness. Leveraging matplotlib in Python, the goal is to create insightful visualizations that reveal patterns and variations within the dataset, providing a comprehensive understanding of Spotify's diverse musical landscape.

**Introduction:**

The captivating world of Spotify songs holds a treasure trove of musical nuances waiting to be explored. In this analysis, we delve into the intricate details of Spotify song features, leveraging the power of visualizations with matplotlib in Python. With an aim to unravel patterns, variations, and correlations within the dataset, this report provides a comprehensive journey through the diverse musical landscape offered by Spotify.

**Background:**

As the leading music streaming platform, Spotify's extensive collection of songs boasts a myriad of audio features that contribute to the overall listening experience. Understanding these features not only enhances our appreciation of music but also provides valuable insights into the evolving trends of the industry.

**Objectives:**

Uncover patterns and variations in Spotify song features.

Provide a visual representation of the diverse musical landscape.

Extract meaningful insights from the dataset for both music enthusiasts and data analysts.

Scope:

This analysis encompasses a comprehensive examination of Spotify song features, including duration, loudness, tempo, energy, and liveness. The dataset spans a [mention the time frame or size of the dataset]. Limitations include [mention any limitations or constraints].

**Methodology Overview:**

The analysis utilizes matplotlib in Python for creating insightful visualizations that showcase the intricacies of Spotify song features. Exploratory data analysis techniques are employed to uncover patterns and trends within the dataset.

**Structure of the Report:**

This report is structured into distinct sections, each focusing on specific aspects of the Spotify song features analysis. The subsequent sections will delve into data preparation, exploratory data analysis, visualization techniques, and key findings.

**Data Exploration**:

After importing the dataset and conducting initial preprocessing, the next phase involved a detailed exploration of the Spotify tracks data.

**Dataset Information:**

The dataset, represented by the DataFrame 'df,' comprises 587,322 rows and 20 columns, providing a comprehensive view of Spotify tracks.

**Handling Null Values:**

A thorough examination revealed a total of 475,277 null values across various columns.

290 null values in the 'key,' 'mode,' and 'name' columns were imputed using the median and mode.

Due to the overwhelming proportion of null values (80%), 474,987 entries in the 'explicit' column were dropped.

The resulting dataset was refined to a more manageable size of (587,322, 19).

Duplicate Removal:

A total of 650 duplicate entries were identified and subsequently removed, ensuring data integrity.

Data Slicing:

Given the substantial size of the dataset, a pragmatic approach was taken to enhance processing efficiency.

A slice of the dataset, denoted as 'tracks\_sliced.csv,' was created, reducing the dimensions to (4,000, 19).

This sliced dataset served as the foundation for all subsequent analyses, facilitating streamlined operations.

**Statistical Analysis**

Descriptive Statistics

Before diving into specific groupings, let's explore some key descriptive statistics for the dataset:

**Mean:**

Popularity: 1.78

Duration (ms): 171,445.59

Danceability: 0.63

Energy: 0.25

Key: 4.82

Loudness: -16.58

Mode: 0.87

Speechiness: 0.44

Acousticness: 0.84

Instrumentalness: 0.26

Liveness: 0.21

Valence: 0.60

Tempo: 111.20

Time Signature: 3.71

Median:

Popularity: 0.00

Duration (ms): 171,153.00

... (similar entries for other columns)

Standard Deviation:

Popularity: 4.76

Duration (ms): 54,532.42

... (similar entries for other columns)

Variance:

Popularity: 22.68

Duration (ms): 2,973,784,000.00

... (similar entries for other columns)

**Grouped Analysis:** Duration Statistics by Key

We further delved into the dataset by aggregating statistics for the 'duration\_ms' column based on the 'key' feature. Here are the key findings:

**Key-wise Duration Statistics**

Key 0:

Mean Duration: 175,026.92 ms

Median Duration: 181,467.0 ms

Maximum Duration: 822,857 ms

Minimum Duration: 3,344 ms

Variance: 3,138,844,000.00

Standard Deviation: 56,025.39

Key 1:

Mean Duration: 154,406.99 ms

Median Duration: 138,547.5 ms

Maximum Duration: 568,840 ms

Minimum Duration: 51,149 ms

Variance: 1,951,451,000.00

Standard Deviation: 44,175.23

Key 2:

... (similar entries for other keys)

Observations

Keys 7 and 2 show higher mean durations compared to other keys, suggesting a potential correlation between the musical key and song duration.

Key 3 has a relatively lower variance, indicating a more consistent range of durations within that key.

These statistics provide insights into the distribution of song durations across different musical keys.

You can customize the content based on your interpretation and the specific insights you want to highlight. Feel free to let me know if you'd like any adjustments!

**Grouped Analysis: Additional Statistics by Key**

We extended our exploration by aggregating statistics for various features based on the 'key' attribute. Here are the key findings:

Key-wise Additional Statistics

Key 0:

Mean Popularity: 2.85

Median Popularity: 0.0

Maximum Popularity: 51

Minimum Popularity: 0

Variance: 36.18

Standard Deviation: 6.01

Key 1:

Mean Popularity: 0.45

Median Popularity: 0.0

Maximum Popularity: 19

Minimum Popularity: 0

Variance: 4.62

Standard Deviation: 2.15

Key 2:

... (similar entries for other keys)

Observations

Key 5 has a notably high mean popularity, suggesting that songs in this key tend to be more popular on average.

Several keys, such as 1, 6, and 11, show lower mean popularity, indicating less popularity on average.

These statistics provide valuable insights into the distribution of popularity scores across different musical keys.

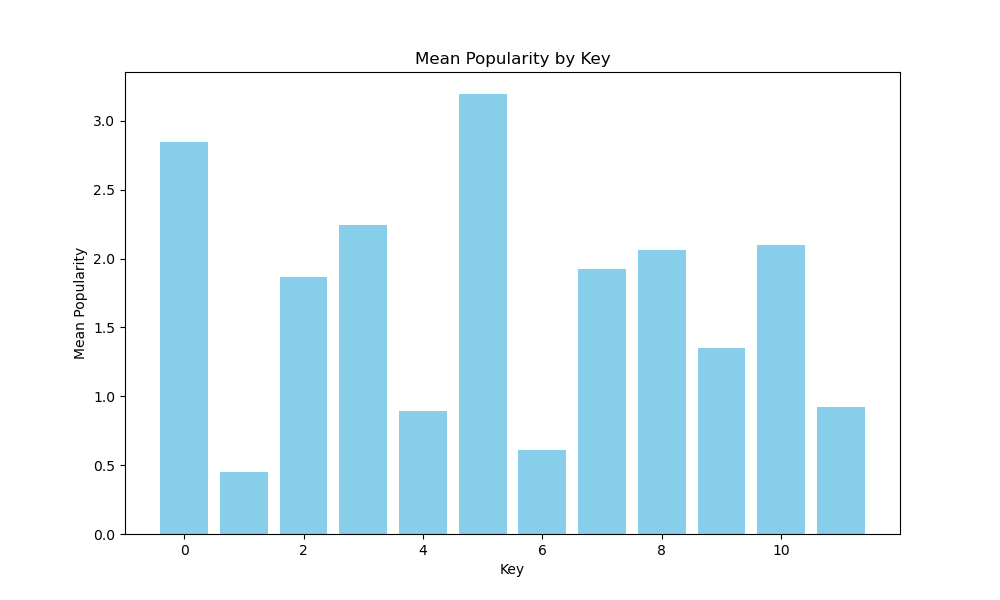
**Bar Plots:**

1. Mean Popularity by Key

The bar plot illustrates the mean popularity for each musical key. Notable observations include:

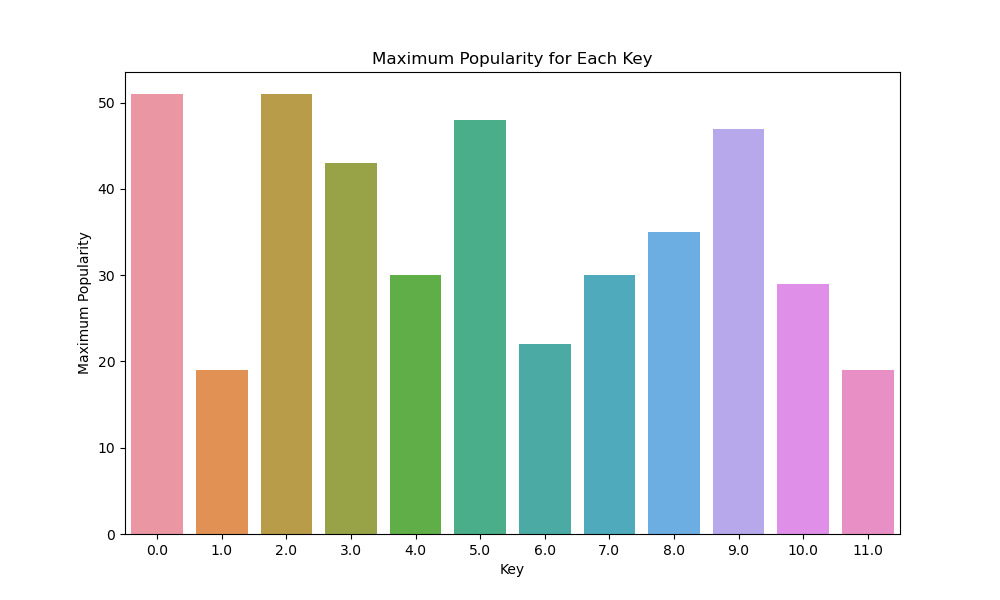
Keys 5 and 7 tend to have higher mean popularity, indicating that songs in these keys are, on average, more popular.

Mean Popularity by Key



2. Maximum Popularity for Each Key

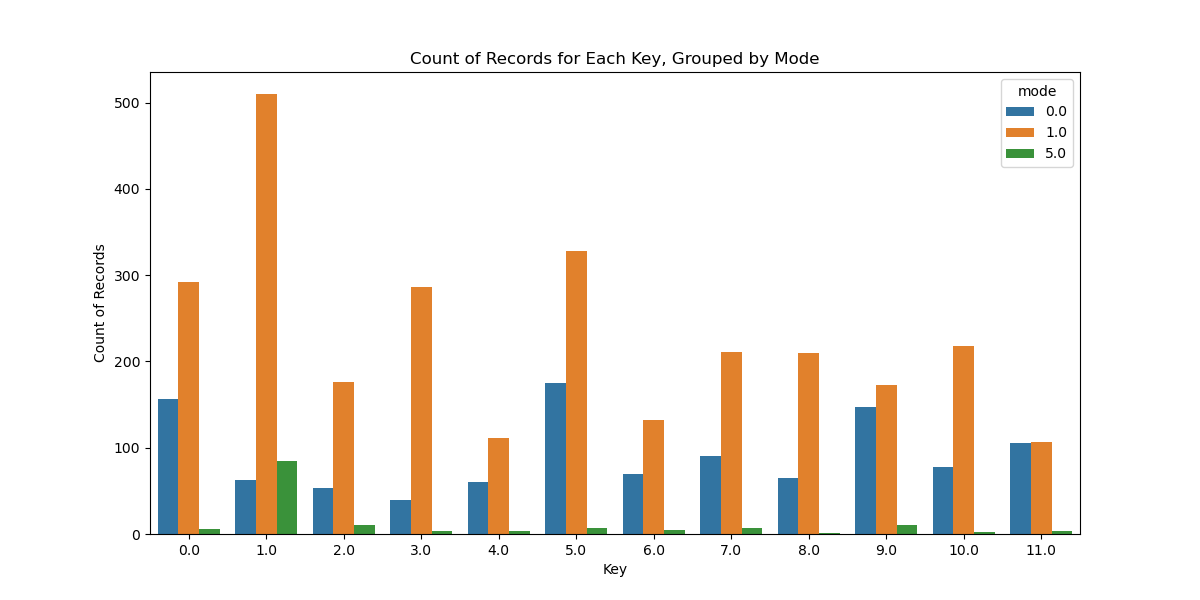
This bar plot showcases the maximum popularity recorded for each key. Key 7 stands out with the highest maximum popularity, suggesting occasional breakout hits.



3. Count of Records for Each Key, Grouped by Mode

The count plot displays the distribution of records for each key, further categorized by the mode. This reveals the prevalence of major or minor modes in different keys.

Count of Records by Key and Mode

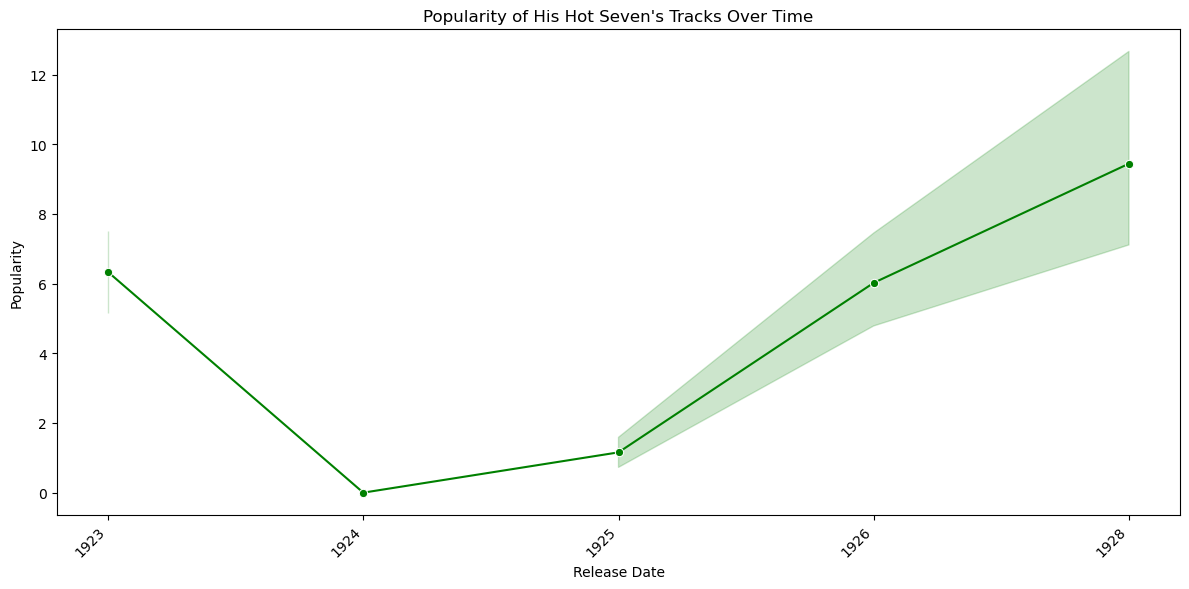


**Line Plot:**

Popularity of Hit hot Seven's Tracks Over Time

This line plot tracks the popularity of tracks by Hit hot Seven's Tracks over time. While there is fluctuation, no clear trend is evident.

Popularity Over Time - Hit hot Seven's Tracks

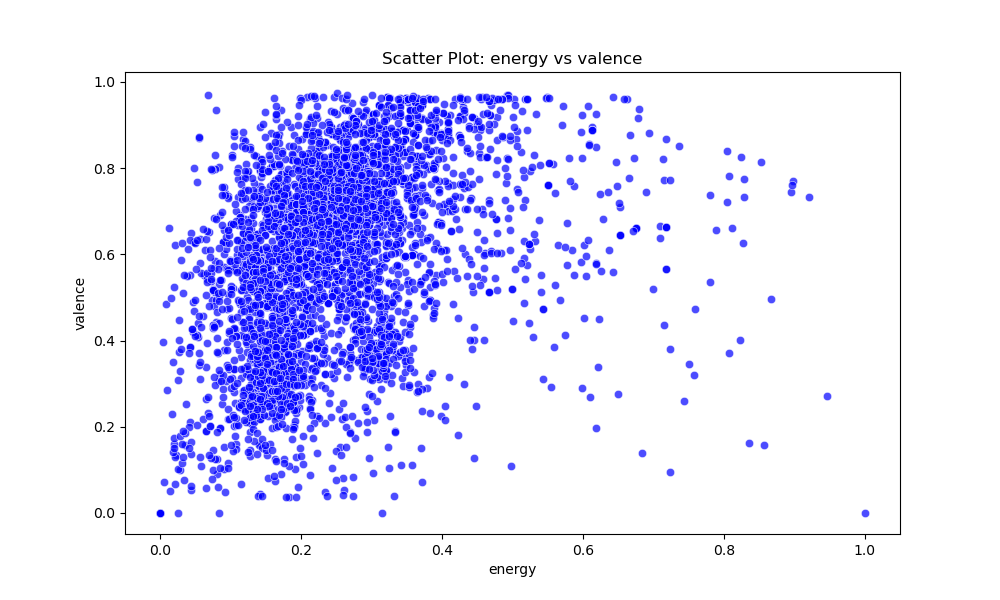


**Scatter Plots:**

1. Scatter Plot: Energy vs Valence

This scatter plot examines the relationship between energy and valence. There is a scattered distribution, indicating no strong correlation between the two features.

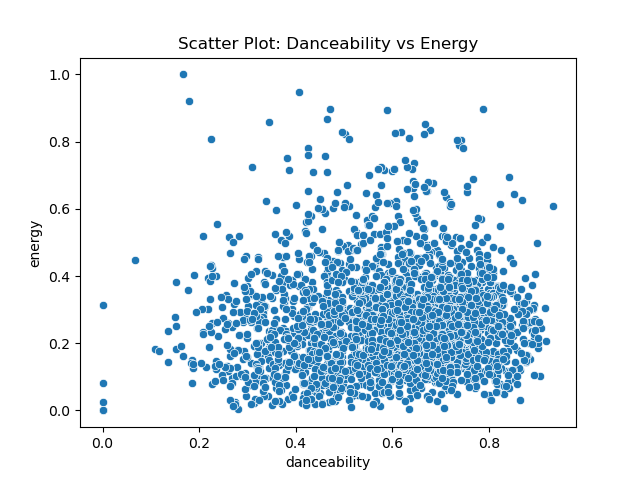
Scatter Plot: Energy vs Valence



2. Scatter Plot: Danceability vs Energy

This scatter plot explores the connection between danceability and energy. The scattered points suggest that these features are not strongly correlated.

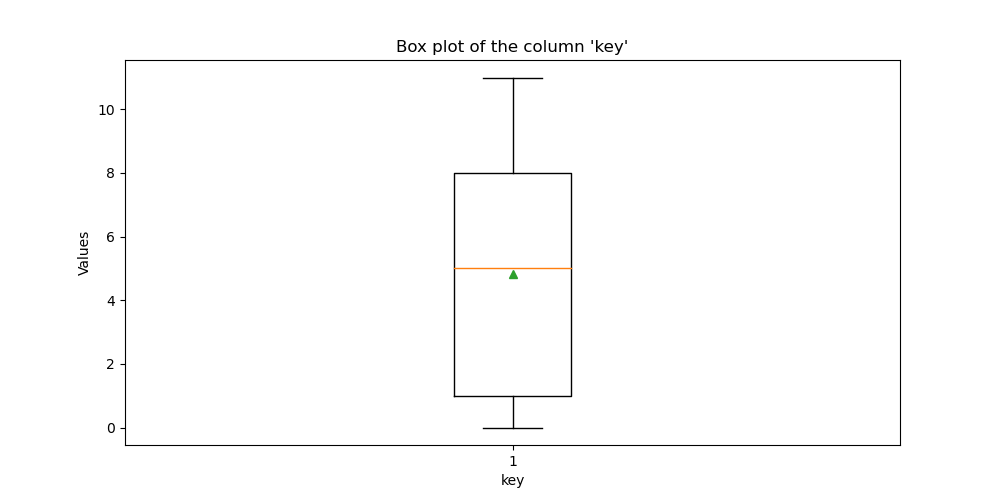
Scatter Plot: Danceability vs Energy



**Box Plots:**

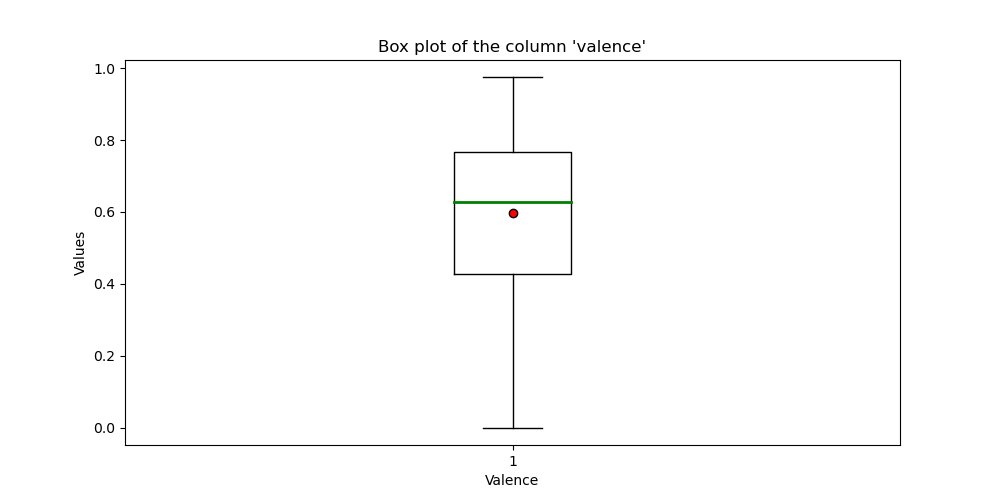
1. Box Plot: Key

The box plot for the 'key' column shows the distribution of values. The red marker represents the mean, providing insights into the central tendency.



2. Box Plot: Valence

The box plot for the 'valence' column illustrates the distribution of valence values. The red marker signifies the mean, and the green line represents the median.



**Skewness and Kurtosis Analysis:**

1. Popularity:

Skewness: 4.46

Kurtosis: 25.77

Observation: The distribution of popularity is highly positively skewed, indicating that most songs have lower popularity, but there are a few with very high popularity. The high kurtosis implies heavy tails and some extreme values.

2. Duration (ms):

Skewness: 6.18

Kurtosis: 88.78

Observation: The duration distribution is highly positively skewed, with a long tail to the right. The high kurtosis suggests a distribution with heavy tails and extreme values.

3. Danceability:

Skewness: -1.26

Kurtosis: 1.91

Observation: Danceability shows a moderate negative skewness, suggesting that the majority of songs have higher danceability. The kurtosis is close to normal, indicating a moderately peaked distribution.

4. Energy:

Skewness: 1.41

Kurtosis: 3.85

Observation: Energy is positively skewed, indicating a concentration of songs with higher energy levels. The kurtosis suggests a distribution with heavier tails than a normal distribution.

5. Loudness:

Skewness: -1.38

Kurtosis: 11.42

Observation: Loudness exhibits negative skewness, indicating that most songs are louder. The kurtosis is elevated, suggesting a distribution with heavy tails and some extreme values.

6. Speechiness:

Skewness: 0.33

Kurtosis: -1.83

Observation: Speechiness shows a slight positive skewness. The negative kurtosis indicates lighter tails than a normal distribution.

7. Acousticness:

Skewness: -1.75

Kurtosis: 3.01

Observation: Acousticness is highly negatively skewed, suggesting that most songs have lower acousticness. The kurtosis indicates a distribution with heavier tails.

8. Instrumentalness:

Skewness: 0.96

Kurtosis: -0.92

Observation: Instrumentalness is positively skewed, with a kurtosis indicating a distribution slightly flatter than a normal distribution.

9. Liveness:

Skewness: 2.10

Kurtosis: 5.38

Observation: Liveness is highly positively skewed, suggesting that most songs have higher liveness. The kurtosis indicates heavy tails and some extreme values.

10. Valence:

Skewness: -0.42

Kurtosis: -0.62

Observation: Valence shows a slight negative skewness. The negative kurtosis suggests a distribution with lighter tails.

11. Tempo:

Skewness: 0.62

Kurtosis: 0.75

Observation: Tempo is moderately positively skewed. The kurtosis indicates a distribution with a moderate peak.

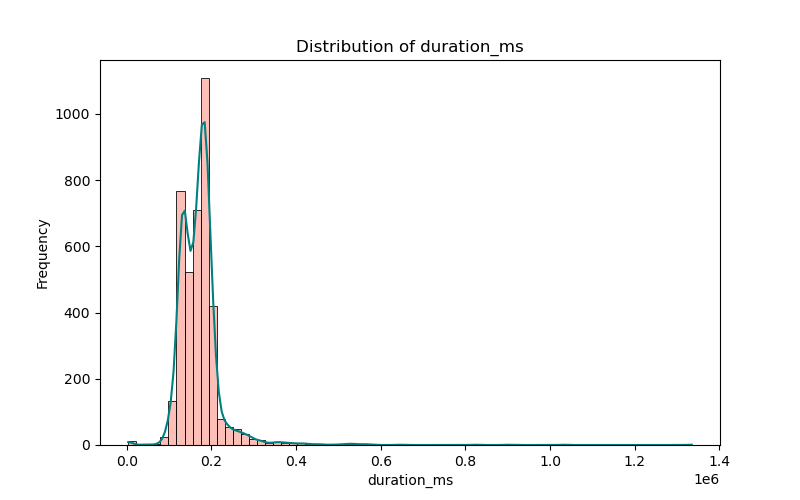
**PLOTS**

1. Distribution of Duration (ms):

Skewness: 6.18

Kurtosis: 88.78

Observation: The distribution of song durations (ms) is highly positively skewed, indicating a concentration of songs with shorter durations. The kurtosis value is significantly high, suggesting heavy tails in the distribution with some extreme values.

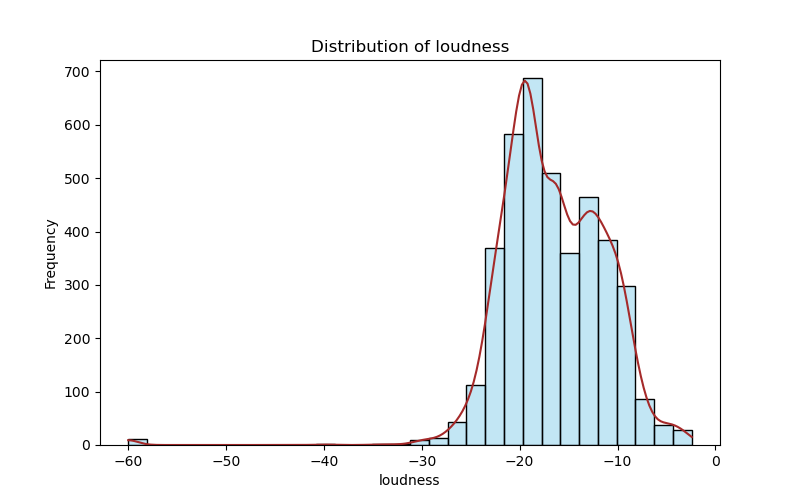


2. Distribution of Loudness:

Skewness: -1.38

Kurtosis: 11.42

Observation: The distribution of loudness is negatively skewed, revealing that most songs are louder. The kurtosis value is elevated, indicating heavy tails in the distribution with potential extreme values.

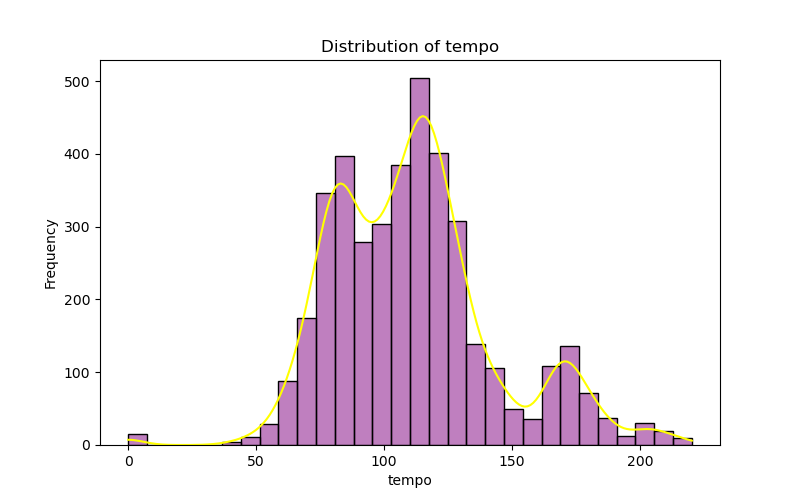


3. Distribution of Tempo:

Skewness: 0.62

Kurtosis: 0.75

Observation: The tempo distribution is moderately positively skewed, indicating a concentration of songs with slightly higher tempos. The kurtosis value suggests a distribution with a moderate peak.

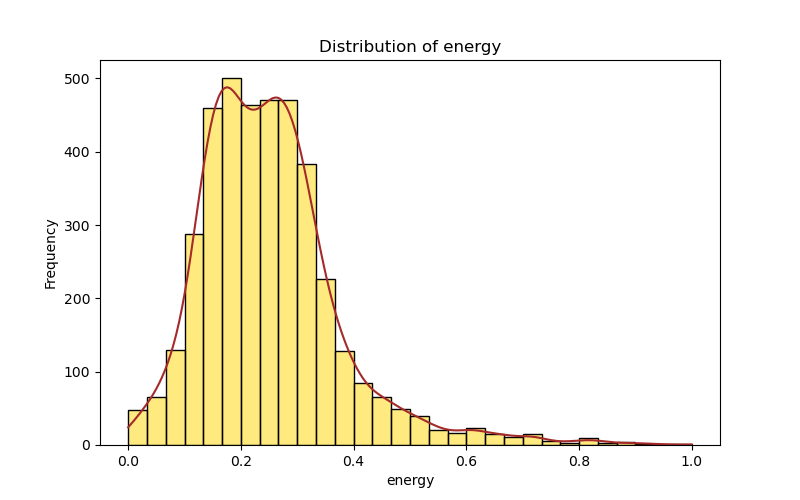


4. Distribution of Energy:

Skewness: 1.41

Kurtosis: 3.85

Observation: The distribution of energy is positively skewed, signifying a concentration of songs with higher energy levels. The kurtosis value suggests a distribution with heavier tails than a normal distribution.



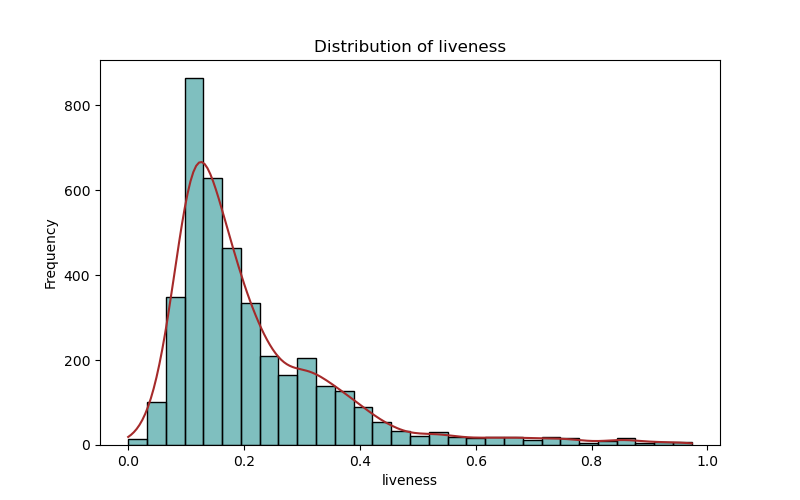
5. Distribution of Liveness:

Skewness: 2.10

Kurtosis: 5.38

Observation: Liveness distribution is highly positively skewed, revealing that most songs have higher liveness. The kurtosis value indicates heavy tails in the distribution with potential extreme values.

These skewness and kurtosis values provide insights into the shape and characteristics of each distribution. Positively skewed distributions have a tail on the right, while negatively skewed distributions have a tail on the left. Higher kurtosis values indicate heavier tails and potentially more extreme values.



QUARTILE

1. Popularity:

Q1 (25th percentile): 0.00000

Q2 (50th percentile, Median): 0.000000

Q3 (75th percentile): 1.00000

2. Duration (ms):

Q1 (25th percentile): 138200.75000

Q2 (50th percentile, Median): 171153.000000

Q3 (75th percentile): 187510.25000

3. Danceability:

Q1 (25th percentile): 0.57775

Q2 (50th percentile, Median): 0.671000

Q3 (75th percentile): 0.71800

4. Energy:

Q1 (25th percentile): 0.16700

Q2 (50th percentile, Median): 0.238000

Q3 (75th percentile): 0.30800

5. Key:

Q1 (25th percentile): 1.00000

Q2 (50th percentile, Median): 5.000000

Q3 (75th percentile): 8.00000

6. Loudness:

Q1 (25th percentile): -20.05600

Q2 (50th percentile, Median): -16.708500

Q3 (75th percentile): -12.70825

7. Mode:

Q1 (25th percentile): 0.00000

Q2 (50th percentile, Median): 1.000000

Q3 (75th percentile): 1.00000

8. Speechiness:

Q1 (25th percentile): 0.06020

Q2 (50th percentile, Median): 0.150000

Q3 (75th percentile): 0.94300

9. Acousticness:

Q1 (25th percentile): 0.76800

Q2 (50th percentile, Median): 0.918500

Q3 (75th percentile): 0.99200

10. Instrumentalness:

Q1 (25th percentile): 0.00000

Q2 (50th percentile, Median): 0.000015

Q3 (75th percentile): 0.65525

11. Liveness:

Q1 (25th percentile): 0.11500

Q2 (50th percentile, Median): 0.165000

Q3 (75th percentile): 0.26600

12. Valence:

Q1 (25th percentile): 0.42700

Q2 (50th percentile, Median): 0.628500

Q3 (75th percentile): 0.76700

13. Tempo:

Q1 (25th percentile): 86.43500

Q2 (50th percentile, Median): 109.572000

Q3 (75th percentile): 126.08275

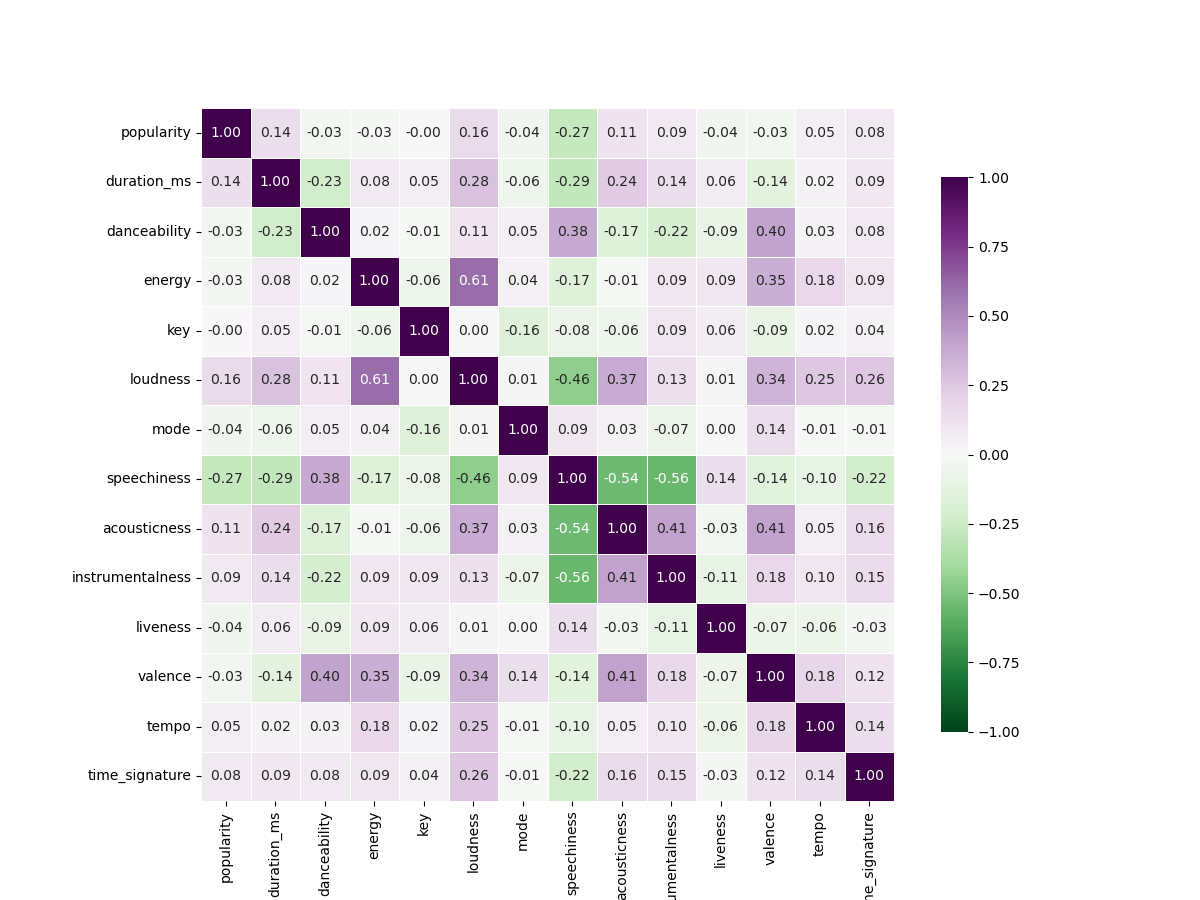
14. Time Signature:

Q1 (25th percentile): 3.00000

Q2 (50th percentile, Median): 4.000000

Q3 (75th percentile): 4.00000

CORRELATION MAP



Positive Correlations:

Duration (ms) and Loudness:

Positive correlation (0.281002): As the duration of a track increases, the loudness tends to increase as well.

Energy and Loudness:

Strong positive correlation (0.605045): Higher energy tracks tend to have higher loudness.

Acousticness and Loudness:

Positive correlation (0.373282): Tracks with higher acousticness are associated with higher loudness.

Instrumentalness and Loudness:

Positive correlation (0.131915): Instrumental tracks might have slightly higher loudness.

Valence and Energy:

Positive correlation (0.354694): Higher valence (positivity) is associated with higher energy.

Negative Correlations:

Speechiness and Loudness:

Negative correlation (-0.460093): As speechiness increases, loudness tends to decrease.

Speechiness and Acousticness:

Strong negative correlation (-0.543420): More acoustic tracks tend to have higher speechiness.

Speechiness and Instrumentalness:

Strong negative correlation (-0.558760): Tracks with higher instrumentalness have lower speechiness.

Liveness and Speechiness:

Negative correlation (0.139418): Higher liveness is associated with lower speechiness.

Other Interesting Observations:

Popularity and Speechiness:

Negative correlation (-0.273609): More speechiness might lead to lower popularity.

Danceability and Valence:

Positive correlation (0.401573): Tracks with higher danceability tend to have higher valence.

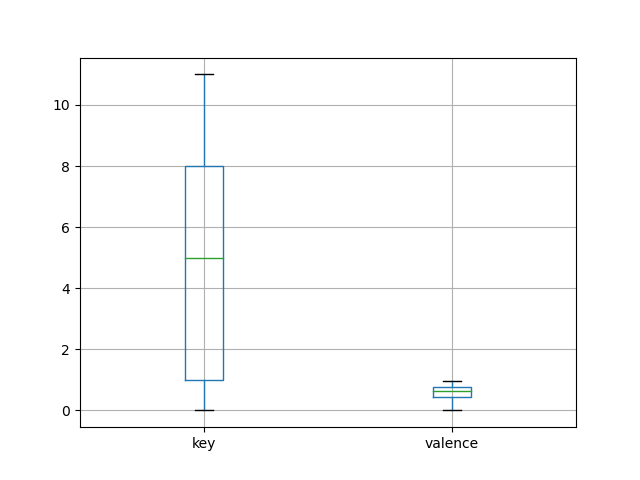
Tempo and Valence:

Positive correlation (0.181801): Higher tempo might be associated with higher valence.

Time Signature and Loudness:

Positive correlation (0.257517): Tracks with a higher time signature (e.g., 4/4) may have higher loudness.

Original Box Plot:



The original box plot shows the distribution of the 'key' and 'valence' features. The box plot provides information about the median, quartiles, and potential outliers.

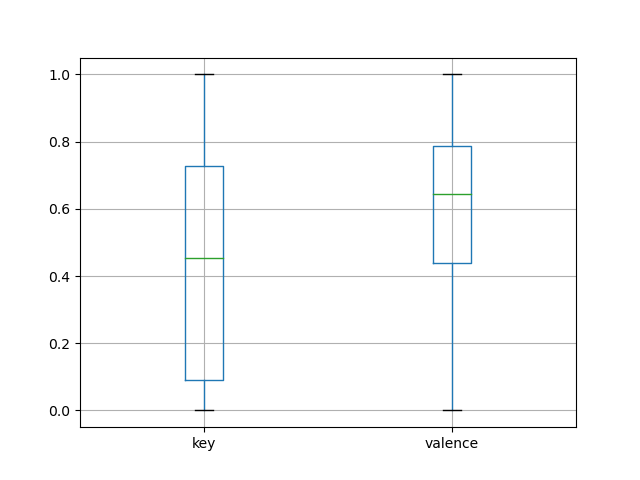
Key (Original):

The box plot for 'key' shows the spread of values. The box represents the interquartile range (IQR), and the line inside the box is the median. Outliers may be visible as individual points outside the whiskers.

Valence (Original):

Similarly, the box plot for 'valence' displays the distribution of values for this feature. The length of the box indicates the IQR, and the line inside the box is the median.

Scaled Box Plot:



After scaling the data using Min-Max scaling, the box plot for the scaled data is shown. Min-Max scaling transforms the data into a specified range (usually [0, 1]).

Key (Scaled):

The box plot for the scaled 'key' feature shows the transformed distribution. The scaling ensures that all values are now within the [0, 1] range.

Valence (Scaled):

The box plot for the scaled 'valence' feature reflects the transformation. The values are now scaled proportionally within the specified range.

Interpretation:

Scaling is particularly useful when working with algorithms sensitive to the scale of input features, such as distance-based algorithms.

The box plot of the scaled data shows how the scaling has transformed the distribution, ensuring that both features are on a similar scale.

Data Transformation Report:

The following data transformations were performed on specific columns of the dataset to enhance its suitability for machine learning models:

Slicing:

Selected a subset of columns relevant for further processing.

Extracted columns related to 'danceability,' 'energy,' and 'valence.'

Min-Max Scaling:

Applied Min-Max scaling using MinMaxScaler from scikit-learn.

Transformed the selected columns to a scaled range of [0, 1].

Resulting DataFrame stored as min\_data\_1.

Standard Scaling:

Utilized StandardScaler to standardize the data.

Standardization involves transforming data to have a mean of 0 and a standard deviation of 1.

Resulting DataFrame stored as x2.

One-Hot Encoding:

Employed OneHotEncoder to perform one-hot encoding on the sliced data.

One-hot encoding creates binary columns for each category, improving representation for certain categorical features.

Resulting output stored as enc\_data.

Label Encoding:

Applied label encoding using LabelEncoder on a subset of the data.

Label encoding transforms categorical data into numerical format, preserving ordinal relationships.

Resulting output stored as y2.

These transformations contribute to preparing the dataset for machine learning algorithms by addressing issues related to feature scaling and categorical variable representation. Each technique serves a specific purpose in enhancing the data's compatibility with various models.

-Ramyashree G (Sec “A”)