**Spotify Exploratory Data Analysis & Songs Recommendation**

A green circle with black lines in it

Description automatically generated

# Exploring and Pre-processing Data

1. **Import Libraries**
   * import warnings
   * import numpy as np
   * import pandas as pd
   * import matplotlib.pyplot as plt
   * import seaborn as sns
   * from sklearn import preprocessing
   * from scipy.spatial import distance
2. **Load dataset**
   * df = pd.read\_csv('dataset.csv',index\_col=0)
   * df
3. **Basic information**
   * How many rows and columns?
     + Dataset has 114000 rows and 20 columns
   * What is the meaning of each row?
     + The Spotify track dataset comprises individual records, each representing a distinct song. The dataset encompasses identifying attributes such as track ID, title, and artist alongside a range of audio features including danceability, energy, loudness, and speechiness.
   * Are there duplicated rows?
     + There are 450 rows that are duplicated so we need to drop those 450 rows
     + After drop duplicated rows, there are 113550 rows left
   * What is the meaning of each column?
     + A screenshot of a computer

       Description automatically generated
   * What is the current data type of each column? Are there any columns having inappropriate data types?
     + A screenshot of a computer

       Description automatically generated
     + A screenshot of a computer code

       Description automatically generated
     + An analysis reveals that three columns – artists, album\_name, and track\_name – exhibit inconsistent data types, containing both string (str) and floating-point (float) values. This discrepancy is likely attributed to the presence of NaN values within these columns, as NaN is a floating-point data type.
   * Missing values in each row
     + 113549 row(s) have 0 missing values
     + 1 row(s) have 3 missing values
     + Total number of rows with missing values: 1

# Data Distribution

1. **Numerical columns**

* We need to extract all numerical columns and store them in the variable numerical\_cols
* (113550, 14)
* There are 14 numerical columns
* What is the percentage of missing values?

A screenshot of a computer code

Description automatically generated

Following this, we computed the minimum and maximum values for each numerical column, accompanied by their respective missing value percentages. Given that the 'key' column employs -1 to indicate undetermined song keys, the count of missing values for this column was derived by tallying the occurrences of -1.

A screenshot of a data

Description automatically generated

The results indicate a complete absence of missing values within all numeric columns. To delve deeper into the dataset, we utilized the describe() function to generate a comprehensive statistical summary of these numerical attributes.

* Distribution of numerical columns

A screenshot of a graph

Description automatically generated

**Popularity:**

* The distribution is right-skewed, with a long tail towards higher popularity values. This indicates that most tracks have lower popularity scores, while a smaller number of tracks achieve very high popularity.

**Duration (ms):**

* The distribution is also right-skewed, with most tracks having durations under 250,000 milliseconds (approximately 4 minutes). There are some outliers with much longer durations.

**Danceability:**

* The distribution is relatively uniform, with a peak around 0.6-0.7. This suggests that a significant portion of tracks have moderate danceability.

**Energy:**

* The distribution is also relatively uniform, with a slight peak around 0.5-0.6. This indicates that a wide range of energy levels are present in the dataset.

**Key:**

* The distribution is relatively uniform across the 12 keys, suggesting no particular preference for any specific key.

**Loudness:**

* The distribution is left-skewed, with most tracks having loudness values between -5 and -10 dB. This suggests that a majority of tracks are moderately loud.

**Mode:**

* The distribution is heavily skewed towards mode 1, indicating that most tracks are in the major key.

**Speechiness:**

* The distribution is right-skewed, with most tracks having low speechiness values. This suggests that the majority of tracks are predominantly musical.

**Acousticness:**

* The distribution is right-skewed, with most tracks having low acousticness values. This indicates that a majority of tracks are not primarily acoustic.

**Instrumentalness:**

* The distribution is heavily right-skewed, with most tracks having very low instrumentalness values. This suggests that the vast majority of tracks contain vocals or other non-instrumental elements.

**Liveness:**

* The distribution is right-skewed, with most tracks having low liveness values. This indicates that most tracks were recorded in a studio environment.

**Valence:**

* The distribution is relatively uniform, with a slight peak around 0.4-0.5. This suggests a wide range of moods represented in the dataset.

**Tempo:**

* The distribution is relatively uniform, with a slight peak around 120 beats per minute (BPM). This indicates a wide range of tempos in the dataset.

**Time Signature:**

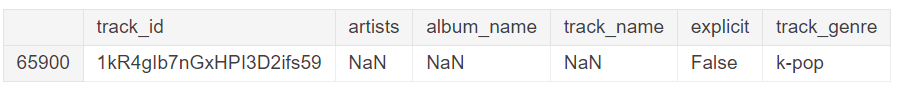
* The distribution is heavily skewed towards 4/4 time signature, indicating that this is the most common time signature used in the tracks.

1. **Categorical columns**
   1. Extract all categorical columns and store them in the variable categorical\_cols
   2. (113550, 6)
      1. There are 6 categorical columns
      2. A screenshot of a music album

         Description automatically generated

* **Track Count:** The dataset comprises a substantial number of tracks, totaling 113,549. This suggests a diverse and extensive collection.
* **Artist Diversity:** The presence of 31,437 unique artists indicates a wide range of musical styles and genres represented in the dataset.
* **Album Variety:** With 46,589 unique albums, the dataset showcases a substantial diversity in album releases.
* **Track Uniqueness:** The high number of unique tracks (89,740) compared to the total count suggests a considerable number of tracks with multiple versions or appearances.
* **Explicit Content:** The binary nature of the "explicit" column with only two values (likely True and False) suggests a clear distinction between tracks with explicit content and those without.
* **Genre Diversity:** The 114 unique genres highlight the broad spectrum of musical styles covered in the dataset.
  1. What is the percentage of missing values?
     1. To start our analysis, we first examined the categorical columns using the info() method.
     2. A screenshot of a computer code

        Description automatically generated
     3. We calculated the number of missing values and percentage of missing values for each categorical column.
     4. A screenshot of a computer

        Description automatically generated
     5. We will examine the rows with missing values to consider how to handle them
     6. 
     7. So, all three columns with missing values are in this particular row. We will proceed to drop that row
     8. Rows with missing values dropped. Updated DataFrame shape: (113549, 20)
     9. A screenshot of a white background

        Description automatically generated
     10. A screenshot of a computer

         Description automatically generated
     11. Some tracks have multiple entries, as the count is higher than the unique

"The Beatles" is the top artist with a relatively high frequency.

The top album is "Alternative Christmas 2022" with a frequency of 195.

"Run Rudolph Run" is the most frequently occurring track name.

The majority of tracks are not explicit (lyrics).

"Acoustic" is the most frequent track genre.

* 1. Visualize unique values
     1. Each song has a unique track\_id, so visualizing it might not provide meaningful insights.

Since explicit has only 2 unique values, we can visually analyze and explore the distribution of these values.

For columns like artists, album\_name, track\_name, and track\_genre with a considerable number of unique values, it's impractical to check them all. Therefore, we will visualize the top 10 most frequently occurring values in each column.

* + 1. **Distribution of Explicit Tracks**

A pie chart with a number of percentages

Description automatically generated

**Observations:**

* **Explicit Content:** The chart clearly shows the distribution of explicit tracks within the dataset.
* **Dominance of Non-Explicit Tracks:** A significant majority of tracks (91.44%) are classified as non-explicit (False).
* **Explicit Tracks Proportion:** A smaller portion (8.56%) of tracks contain explicit content (True).

**Inferences:**

* The dataset predominantly comprises tracks suitable for a general audience without explicit content.
* While explicit content is present, it represents a minority within the overall collection.
* This information can be valuable for users seeking to filter tracks based on their preferences regarding explicit content.
  + 1. **Top of artists, album\_name, track\_name, track\_genre**

A group of blue and green bars

Description automatically generated**Top Artists and Albums**

* **Dominance of English-language artists:** The majority of the top artists are from English-speaking countries (The Beatles, Stevie Wonder, Linkin Park, Ella Fitzgerald, OneRepublic, Chuck Berry).
* **Diverse genres represented:** The top albums showcase a variety of genres, including Christmas music, pop, rock, and Latin music (reggaeton).
* **Potential regional bias:** The presence of artists like Prateek Kuhad and Håkan Hellström might indicate a regional focus or a specific target audience for the data.

**Top Tracks and Genres**

* **Seasonal influence:** The top tracks include several Christmas-themed songs, suggesting the data might be biased towards a specific time period.
* **Genre diversity:** The top genres cover a broad spectrum, including acoustic, emo, rock-n-roll, reggaeton, disco, and more.
* **Language diversity:** The inclusion of tracks in Spanish (CÓMO SE SIENTE -Remix, RUMBATON, X ÚLTIMA VEZ) indicates a diverse listener base.

**Overall Observations**

* The visualization provides a snapshot of popular artists, albums, tracks, and genres on Spotify.
* The data appears to be influenced by seasonal factors, particularly with the predominance of Christmas-related content.
* The inclusion of artists and tracks from different languages suggests a global audience for the platform.

1. **Abnormal values and outliers**
   * 1. We can observe that the distribution as well as the range of values in the numerical columns are **no abnormal values** so we will check outliers for some columns. We will use boxplot to visualize outliers

A chart with lines and numbers

Description automatically generated with medium confidence

**Observations**:

* Analysis of popularity, acousticness, energy, and valence revealed an absence of significant outliers.
* While a single outlier was detected in popularity, its impact was deemed negligible.
* Conversely, the remaining columns exhibited numerous outliers, which is expected given their continuous nature, broad value ranges, and non-normal distributions.
* These outliers are attributable to the diverse characteristics inherent in musical compositions and are thus retained to preserve dataset integrity and complexity.

1. **Correlation Between Variables**

A graph of numbers and a ruler

Description automatically generated

**Strong Positive Correlations:**

* **Loudness and Energy:** Songs with higher energy levels tend to be louder.
* **Danceability and Energy:** Danceable songs are often more energetic.
* **Valence (positivity) and Danceability:** Positive and upbeat songs are often more danceable.
* **Duration and Instrumentalness:** Longer songs tend to have a higher instrumental content.

**Strong Negative Correlations:**

* **Acousticness and Energy:** Acoustic songs tend to have lower energy levels.
* **Acousticness and Loudness:** Acoustic songs are typically less loud.
* **Acousticness and Danceability:** Acoustic songs are generally less danceable.
* **Speechiness and Instrumentalness:** Songs with more speech tend to have less instrumental content.

**Other Notable Correlations:**

* **Popularity and Loudness:** There's a slight positive correlation, suggesting popular songs might be louder.
* **Tempo and Energy:** There's a weak positive correlation, indicating a possible link between faster tempos and higher energy levels.

**General Observations:**

* The correlation matrix reveals interesting relationships between audio features and musical characteristics.
* These insights can be valuable for music recommendation systems, playlist generation, and understanding music preferences.

# Question 1: How do different genre-related characteristics affect the popularity of songs?

1. Understanding audience trends is instrumental in tailoring music content and events to resonate with consumer preferences. Insights into popular genres can inform event planning and programming strategies, optimizing audience engagement. Additionally, this knowledge is crucial for refining advertising and marketing initiatives. For music artists, comprehending genre trends and characteristics enables informed strategic decisions, from genre selection to cultivating a target audience-aligned image.
2. **Analyzing genre-related characteristics:**
   1. Number of songs for each genre.
   2. Average popularity score for each genre.
   3. Explicit ratio.
   4. Genres in the top ~110 most popular songs.
   5. Ratio of live songs for each genre.
   6. Comparing the correlation between loudness and energy in the top 3 most popular genres.
3. **Solution Method**
   1. **Number of genres in the dataset**
      1. 114
   2. **Top 10 genres with the highest number of songs**
      1. **Methodology:**

Calculate the frequency distribution of values within the track\_genre column.

Extract the top 10 genres based on their frequency counts.

Visualize the frequency distribution of these top 10 genres for clear interpretation.

A screenshot of a graph

Description automatically generatedThe analysis reveals a consistent count of 1,000 songs within each of the top 10 genres. This indicates a deliberate dataset construction methodology, where approximately 1,000 songs per genre were included in the collection.

* 1. **Which genre has the highest average popularity score?**
     1. **Methodology:**

Group the dataset by track\_genre to create genre-specific subsets.

Calculate the mean popularity value for each genre group.

Transform the results into a DataFrame format and reset the index for subsequent analysis.

Identify and visualize the top 10 genres with the highest mean popularity scores.

A graph with blue and orange stripes

Description automatically generated with medium confidence**Popularity Distribution:** The average popularity scores among the top 10 genres exhibit relatively small variations, indicating a competitive landscape.

**Genre Diversity:** The presence of diverse genres, including pop-film, K-pop, chill, and sad, suggests a broad appeal to a variety of listener preferences.

**Genre Characteristics:**

* **Pop-film:** Leverages familiarity and emotional connections through its association with popular films.
* **K-pop:** Combines catchy music with visual and image elements to attract a global audience.
* **Chill:** Offers a relaxing and soothing auditory experience, promoting stress reduction.
* **Sad:** Evokes strong emotional responses through its expressive and poignant content.
  1. **Which genre has the highest rate of explicit songs?**
     1. **Methodology:**

Group the dataset by track\_genre to create genre-specific subsets.

Calculate the mean value of the explicit column within each group, representing the proportion of explicit songs per genre.

Convert the results into a DataFrame format and reset the index for subsequent analysis.

Determine and visualize the top 10 genres with the highest mean explicit values.

A graph with blue and orange bars

Description automatically generated**Comedy Dominance:** The comedy genre exhibits a significantly higher average explicit ratio compared to other genres.

**Genre-Specific Language Use:**

* **Comedy:** Leverages explicit language for comedic effect and audience engagement.
* **Emo and Sad:** Employs explicit language to convey intense emotions and authenticity.
* **J-dance:** Potentially utilizes explicit language as a form of expression and rebellion.

**Audience Preferences:** Genres with higher explicit ratios may appeal to listeners seeking music with edgier and more explicit content.

* 1. **Which genre has the highest number of songs in the top 0.001% (~110) most popular songs?**
     1. **Methodology:**

Rank songs by popularity in descending order.

Select the top 0.001% of the ranked songs to form an elite subset.

Group the elite subset by genre and count the occurrences of each genre.

Transform the genre counts into a DataFrame for further analysis.

Rank genres by song count in descending order and visualize the results.

A graph with blue and orange bars

Description automatically generated

**The genre with the most songs in the top 0.001% is pop with 22 songs.**

* The provided chart illustrates a diverse range of genres represented within the top 0.001% of most popular songs. While a variety of genres are present, the **pop** genre significantly outnumbers all others, indicating its dominant position in the realm of popular music.
* Pop music's broad appeal can be attributed to its ability to incorporate elements from diverse genres such as dance, hip-hop, rock, and R&B, creating a versatile and inclusive sound. Coupled with extensive marketing and promotion, pop artists often achieve significant commercial success, solidifying the genre's prominence in popular music charts.
  1. **Which genre has the highest ratio of live songs?**
     1. **Methodology:**

Group the dataset by track\_genre to create genre-specific subsets.

Calculate the mean liveness value for each genre group, representing the average live recording characteristic.

Transform the results into a DataFrame format and reset the index for subsequent analysis.

Identify and visualize the top 10 genres with the highest mean liveness values.

A graph with blue and yellow bars

Description automatically generated with medium confidence

**The genre with the highest liveness ratio is comedy with an average liveness ratio of 66.22%.**

* + - * The chart indicates that the comedy genre exhibits the highest average liveness ratio among the listed genres. This can be attributed to several factors:
        + **Performance-Centric Nature:** Comedy is inherently a live performance art form that thrives on audience interaction and spontaneity.
        + **Audience Engagement:** The dynamic nature of comedy necessitates live performances to fully engage the audience and maximize comedic impact.
        + **Creative Flexibility:** Live performances offer comedians greater freedom to adapt their material based on audience reactions, enhancing the overall comedic experience.
      * These factors collectively contribute to the dominance of the comedy genre in terms of live recordings.
  1. **Regression plot of the correlation of energy vs loudness**

A graph with blue dots

Description automatically generated

**Observations:**

* Positive Correlation: The upward slope of the regression line indicates a positive correlation between energy and loudness. This suggests that as loudness increases, there is a tendency for energy levels to also increase.
* Scatter: The data points exhibit a fair amount of scatter around the regression line, implying that while there is a positive relationship, it is not perfectly linear. This suggests that other factors besides loudness might influence energy levels.
* Outliers: There are some data points that deviate significantly from the general trend, appearing as outliers. These could represent songs with unique characteristics that don't follow the typical pattern.

**Inferences:**

* Songs with higher energy levels tend to be louder, and vice versa.
* Other factors besides loudness play a role in determining a song's energy level.
* The presence of outliers suggests that there might be exceptions to the general trend.
* Objective: Examine the correlation between loudness and energy within the top three most popular genres to identify potential trends.

**Methodology:**

* + Determine the average popularity for each genre.
  + Identify the top three genres based on their average popularity scores.
    - Filter the dataset to include only tracks from the top three genres.
  + Calculate the correlation between loudness and energy for each of the three genres.
  + Visualize the correlation for comparative analysis.

A screen shot of a graph

Description automatically generated

* + - * + The scatter plot comparing loudness and energy across the top three popular genres reveals a strong positive correlation between the two variables. This suggests that songs with higher energy levels tend to be louder, and conversely, louder songs are often perceived as more energetic.
        + The relationship between loudness, often measured as overall sound pressure level, and energy, a subjective measure of perceived intensity and dynamism, is likely due to several factors:

**Perceptual Association:** Human perception tends to associate louder sounds with greater intensity and excitement, leading to a higher perceived energy level.

**Production Techniques:** Music production often employs techniques that enhance loudness and energy simultaneously, such as dynamic range compression and equalization.

**Genre Conventions:** Certain genres, particularly those within the top three (chill, K-pop, and pop-film), may have stylistic conventions that favor a combination of high loudness and energy.

These findings suggest that loudness is a significant factor influencing the perceived energy of a song within these popular genres.

1. **Answer: How do different genre-related characteristics affect the popularity of songs?**
   1. **Key Findings:**
      1. **Genre Popularity:** Pop-film, K-pop, chill, and sad genres emerged as the most popular, with pop significantly dominating the top-tier songs.
      2. **Genre Characteristics:** Comedy exhibited the highest explicit content ratio, while genres like pagode, sertanejo, and samba showcased a preference for live performances.
      3. **Audio Features:** A strong correlation between loudness and energy was observed, suggesting a potential link between these attributes and perceived popularity.
   2. **Implications:**
      1. **Audience Preferences:** Listeners demonstrate a clear inclination towards pop music, particularly within the top-tier songs. Additionally, there's a demand for both explicit content (as seen in comedy) and live performances (as evident in genres like pagode and sertanejo).
      2. **Artist Strategy:** Balancing artistic integrity with commercial appeal is crucial for achieving widespread popularity. Understanding audience preferences for specific genres, audio features, and content types can inform artist development and music creation.
      3. **Industry Trends:** The data suggests a potential trend towards louder, more energetic, and potentially explicit content in popular music. Industry stakeholders, including record labels, radio stations, and streaming platforms, may benefit from analyzing these trends to optimize their strategies.

# Question 2: How should artists choose their music genres?

1. By examining genre preferences and characteristics, this analysis aims to provide artists with insights to inform their genre selection process. This information can assist artists in aligning their musical style with personal expression, capitalizing on strengths, exploring diverse musical territories, or strategically targeting popular trends.
2. To effectively guide artists in genre selection, we will explore two key considerations:
   1. **Artistic Integrity vs. Commercial Viability:** Should artists prioritize personal musical expression or align their sound with popular trends?
   2. **Genre Specialization vs. Diversification:** Is it advantageous for artists to focus on a specific genre or explore a broader musical range?
3. Should artists pick genres they like or go for what's popular?
   1. **Objective:** Analyze the musical preferences of top artists and compare them to overall genre popularity trends.
   2. **Data Preparation:** To achieve this, we will conduct the following preprocessing steps:
      1. **Identify Top Genres:** Determine the most popular genres based on the average popularity of songs within each genre.
      2. **Identify Top Artists:** Determine the most popular artists based on the average popularity of their songs.
      3. **Analyze Top Artist Genres:** Extract the genres associated with the songs of the top-ranked artists.

**Top Genres**

A screenshot of a computer

Description automatically generated

**Top Artists**

A list of music bands

Description automatically generated

**Genres of top artists**

A white background with black text

Description automatically generated

* + 1. **Data Analysis**

To understand the relationship between top genres and the genres favored by top artists, we will employ the following analytical approach:

* + - * **Genre Overlap:** A Venn diagram will be utilized to visually compare the sets of top genres and genres produced by top artists, highlighting shared and unique genres.
      * **Genre Popularity Comparison:** Bar charts will be employed to contrast the average popularity of three genre categories:
        + Top genres not produced by top artists
        + Top genres produced by top artists
        + Genres produced by top artists, regardless of overall popularity
    1. Top genres vs. genres of top artists

A yellow and orange chart with text

Description automatically generated

**Analysis of Genre Overlap**

* + - * The Venn diagram illustrates a significant intersection between the most popular genres and the genres produced by top artists.
      * A total of 35 genres are shared between these two groups, indicating a strong alignment between audience preferences and the creative choices of leading musicians. This overlap suggests that successful artists often tap into prevailing musical trends while also incorporating their unique artistic visions.
      * Conversely, the diagram highlights 22 genres that are widely popular but not extensively explored by top artists. This presents potential opportunities for emerging artists to carve out a niche and differentiate themselves by focusing on these underrepresented genres.
      * Finally, the 17 genres exclusive to top artists demonstrate their role as musical pioneers, venturing into uncharted territories and potentially influencing future trends.
      * Overall, the Venn diagram provides valuable insights into the dynamic relationship between popular music and artistic innovation.
    1. **Popularities of each set of genres**

**Objective: To visualize the average popularity of genres across three categories:**

* + - * Genres exclusively in the top genres list
      * Genres exclusively in the top artists' genre list
      * Genres present in both lists

A graph with blue and yellow text

Description automatically generated with medium confidence

A graph with blue and yellow bars

Description automatically generated with medium confidence

A graph with blue and yellow stripes

Description automatically generated with medium confidence

**Analysis of Genre Popularity**

**Overview of Genre Popularity**

* + - * + The provided bar chart offers a visual representation of average popularity across three genre categories: top genres, genres preferred by top artists, and the intersection of both.

**Key Findings**

* + - * + **High Popularity of Top Genres:** Genres such as pop-film, sad, Indian, and anime demonstrate exceptionally high average popularity scores, exceeding 40. This indicates a strong preference for these genres among the general listening audience.
        + **Lower Popularity of Artist-Preferred Genres:** While still popular, genres predominantly favored by top artists tend to exhibit lower average popularity scores compared to the top genres. This suggests that while these genres resonate with influential artists, they may have a narrower appeal to the broader public.
        + **Intersection of Popularity and Artist Preference:** Genres present in both categories, such as K-pop, chill, and pop, often achieve high popularity levels. This indicates a strong alignment between audience preferences and the creative choices of top artists.

**Implications**

* + - * + These findings suggest that while there is a degree of overlap between popular genres and those favored by artists, there are also distinct differences. To achieve widespread success, artists may need to consider a balance between pursuing personal artistic expression and aligning their music with established popular trends. Additionally, identifying underrepresented yet popular genres could offer opportunities for artists to differentiate themselves and cultivate a unique audience.

1. **Does genre specialization or diversification correlate with artist popularity?**
   1. **Methodology:** To investigate this question, we will analyze the relationship between the number of genres an artist encompasses and their overall popularity.
      1. **Genre Range Determination:** Calculate the number of unique genres associated with each artist.
      2. **Artist Categorization:** Establish a benchmark based on the genre range of top-performing artists to classify artists as either "genre specialists" or "genre diversifiers."
      3. **Genre Range Distribution:** Visualize the distribution of genre counts across all artists and compare it to the distribution among top artists.
      4. By examining these patterns, we aim to identify potential correlations between genre focus and artist success.

A graph with numbers and bars

Description automatically generated

A graph with numbers and a bar

Description automatically generated

**Analysis of Genre Distribution**

**Overview**

* + - * The provided histograms visualize the distribution of the number of genres associated with artists in the overall dataset and among top-performing artists.

**Key Findings**

**Overall Artist Population:**

* + - * Genre Specialization: A significant majority of artists exhibit a specialized approach, focusing on 1 to 3 genres. This indicates a prevalence of artists with a defined musical identity.
      * Genre Diversity: A smaller but notable proportion of artists demonstrate a broader range, covering 4 to 9 genres. This suggests a more eclectic approach to music creation.
      * Genre Experimentation: A limited number of artists explore an extensive range of genres, with a maximum of 19 genres. These artists represent a niche group characterized by experimentation and genre-bending.

**Top Artists:**

* + - * Genre Focus: Similar to the overall population, most top artists concentrate on a specific set of genres, typically 1 to 3. This suggests that genre specialization can contribute to commercial success.
      * Genre Diversification: While present, genre diversity among top artists is less pronounced compared to the overall population. This indicates that while exploring multiple genres can be beneficial, a strong focus on a particular style might be more critical for achieving top-tier status.
      * Genre Experimentation: The number of top artists venturing into 8 or more genres is considerably lower, suggesting a more strategic approach to genre selection among successful musicians.

**Implications**

* + - * These findings highlight the complex relationship between genre focus and artist success. While genre specialization appears to be a common characteristic among top artists, a certain degree of diversity can also contribute to broader appeal. Understanding the trade-offs between these approaches can inform artists' strategic decisions and label strategies for talent development.

1. **Data Preparation**
   1. **Identify Artist Groups:** Create two lists: one containing artists with a narrow genre range (less than the benchmark) and another with artists having a broad genre range (equal to or greater than the benchmark).
   2. **Filter Data:** Use the isin function to filter the original dataset based on these artist lists.
2. **Calculating Average Popularity**
   1. **Group by Artist:** Group the filtered dataframes by the 'artist' column.
   2. **Calculate Mean Popularity:** For each artist group, compute the mean 'popularity' value.

A graph of different sizes and shapes

Description automatically generated with medium confidence**Analysis of Artist Popularity and Genre Diversity**

**Overview of Findings**

* + - The provided visualizations offer insights into the relationship between an artist's genre diversity and their overall popularity.

**Key Observations**

* + - **Similar Popularity Distributions:** The distribution of popularity scores among both diverse and less diverse artists exhibits a striking resemblance. This suggests that the number of genres an artist explores does not significantly impact their overall commercial success.
    - **Overlap in Top Artists:** The presence of renowned artists such as Harry Styles, Olivia Rodrigo, and Eminem in both groups further reinforces the notion that genre diversity is not a sole determinant of popularity.
    - **Diverse Paths to Success:** The data suggests that artists can achieve significant popularity through various strategies. Some may excel by specializing in a particular genre, while others can build a substantial following by exploring multiple musical styles.

**Implications**

* + - These findings challenge the common assumption that artists must diversify their genres to achieve widespread success. The data indicates that both genre specialization and diversification can lead to popularity. Consequently, artists should prioritize creating high-quality music that resonates with their audience, rather than solely focusing on genre breadth.
    - Ultimately, an artist's success is influenced by a combination of factors including musical talent, effective marketing, and alignment with audience preferences. While genre diversity can be a valuable tool for expanding an artist's reach, it is not a prerequisite for achieving popularity.

1. **Question: How should artists choose their music genres?**
   1. **Artist Genre Selection: A Strategic Perspective**
      1. The decision to specialize or diversify within musical genres is a complex one, influenced by both artistic inclination and market dynamics. Our analysis indicates that while genre specialization can lead to significant success, as evidenced by the popularity of certain artists within specific genres, a diversified approach can also yield positive results.
   2. **Key Considerations:**
      1. **Personal Passion:** Artists should prioritize genres that resonate with their personal artistic vision and evoke genuine enthusiasm.
      2. **Market Analysis:** Understanding current trends and audience preferences can inform genre selection without compromising artistic integrity.
      3. **Skill Set:** An artist's strengths and weaknesses should be carefully considered when choosing a genre.
      4. **Audience Development:** Building a dedicated fanbase within a specific genre can provide a solid foundation before exploring other musical territories.
      5. **Evolution and Experimentation:** Artists should remain open to experimentation and genre evolution to sustain long-term career growth.

Ultimately, the most effective approach to genre selection is likely a hybrid strategy that combines elements of specialization and diversification. By carefully considering these factors, artists can increase their chances of achieving both artistic fulfillment and commercial success.

# Question 3: How can we suggest songs based on a user's current listening preferences?

1. **Objective: Song Recommendation System Based on Similarity**
   1. **Goal:** Develop a recommendation system that suggests songs similar to a user's preferred track.
2. **Methodology:**
   1. **Similarity Measurement:** Employ Euclidean distance to quantify the similarity between songs based on relevant features.
   2. **Nearest Neighbor Identification:** Identify the 'K' songs with the smallest Euclidean distance to the target song.
   3. **Recommendation Generation:** Present the identified songs as recommendations to the user.

By calculating the proximity of songs in feature space, this approach aims to provide users with music that aligns with their existing tastes.

A graph of a distribution of a distance

Description automatically generated

**Recommend the top 5 tracks for the user in descending order of similarity, based on the calculated Euclidean distances**.

A screenshot of a computer

Description automatically generated