Music Trend Analysis of Million Songs Dataset

By: Madison L Rundell, Michael A Byrd, Sudha Rani Seeli, Jheel Jagani.

MIS 548 – Final Project Report

Dataset Selection:

Our group selected the Million Song Dataset for Music Trend Analysis due to its extensive coverage of contemporary music tracks, offering a diverse range of genres and artist information. The comprehensive dataset allows us to explore evolving music characteristics, identify emerging patterns, and understand the factors influencing listener preferences over time.

Business Problem:

Our goal with Music Trend Analysis is to uncover insights into the dynamic and evolving landscape of the music industry. This analysis caters to artists, producers, record labels, and streaming platforms. By leveraging this rich dataset, we aim to provide valuable insights that inform strategic decisions within the music industry. The project contributes to a deeper understanding of current trends, guiding marketing efforts and supporting the creation of music that resonates with current and emerging audience preferences.

Dataset Overview:

- The Million Songs Dataset comprises of key data points such as track titles, release dates, artist identifiers, and durations. This dataset provides a comprehensive lens into the dynamic world of music.
- Artist-related metrics, including familiarity and scores, offer nuanced insights into both the overarching recognition of musicians and the specific popularity of their tracks.
- By augmenting the dataset with new columns, such as the duration in minutes and the disparity between artist familiarity and track scores, we aim to refine our analytical approach.

Data Quality Assessment:

Our dataset is extensive, diverse, and well-prepared, laying a solid foundation for robust music trend analysis.

- Correlation Matrix: Strong positive correlation between "artist_familiarity" and "artist_hotttnesss," highlighting the relationship dynamics in the dataset.
- Distinct Counts: Distinct counts provide insights into the uniqueness of values within each column.
- Count: The count section indicates that there are a million entries across all columns, affirming the dataset's completeness and consistency.
- Data Types: Examining data types is crucial for understanding how each attribute is represented. The majority are objects, while numerical features are appropriately represented as float64 or int64.
- Shape: The dataset has 1,000,000 rows and 14 columns.
- Nulls: The absence of null values across all columns suggests a well-prepared dataset with no missing information.

Data Cleaning and Filtering:

Our journey into Music Trend Analysis begins with meticulous data cleaning and filtering, ensuring the integrity and relevance of our insights derived from the Million Songs Dataset.

- **Sorting:** The DataFrame is sorted based on the 'year' column in descending order. This helped in analyzing the most recent data first.
- **Column Selection:** A subset of columns was selected, focusing on essential attributes like song details, artist information, and relevant metrics.
- **Duration Conversion:** The 'duration' column is converted from seconds to minutes, providing a more user-friendly metric for analysis.
- **Column Renaming:** The 'artist_hotttnesss' column is renamed to 'score' for brevity and clarity.
- **New Columns:** Two new columns, 'difference_fam_score', and 'artist_release', are created to facilitate specific analyses or enhance data presentation.
- **difference_fam_score:** Score subtracted from familiarity. This shows how close the popularity of the artist & song are. A positive value in this column indicates that the 'artist_familiarity' is higher than the 'score'. This suggests that the artist is more well-known than the overall score of the song. A negative value indicates that the 'score' is higher than the 'artist_familiarity', suggesting that the song's popularity might exceed the artist's general popularity.
- **artist_release:** This column provides a combined representation of the artist and the release (presumably a song or album). It's useful for creating a concise label or identifier for each record in the DataFrame, making it easier to reference and understand.
- **Data Filtering:** Songs with a year value of 0 were omitted from the dataset. These often represented more niche and peculiar categories, such as movie scores or jingles, which could disrupt visual coherence.
- To enhance visual representation, the dataset was confined to songs released between the years 1922 and 2010. The exclusion of the year 2011, with only one song, helped maintain balance and improve visual aesthetics.
- **Negative Values Removal:** Instances where familiarity or score values were negative, potentially indicative of data errors, were excluded. These outliers were often associated with more peculiar or unconventional songs, and their removal contributed to the overall quality and reliability of the dataset.

After applying these filters, the dataset was refined to approximately 500,000 rows, ensuring a focus on more mainstream and widely recognized artists and songs.

Data Exploration: Various techniques have been employed to unveil meaningful information.

- **Distinct Years:** The distinct_years variable captures all unique values in the 'year' column. The resulting array shows years ranging from 1922 to 2010.
- **Song ID Occurrences:** The count of occurrences for each unique song ID is obtained using song_id_counts. Songs with the same ID but different release details are identified, and those occurring more than once are displayed.
- **Song-Title-Artist Occurrences:** Using song_artist_counts, the occurrences of combinations of 'title' and 'artist_name' are counted. Entries with more than one occurrence are filtered and displayed.
- **Sorting by Duration:** The DataFrame is sorted by 'duration' in descending order, displaying relevant columns such as year, release, title, artist_name, familiarity, score, difference_fam_score, and duration_minutes.
- **Songs per Year:** Using groupby on the 'year' column, the count of songs for each year is obtained and sorted in descending order. The result provides insights into the distribution of top songs over the years.

Insights:

- The number of top songs per year appears to peak around the mid-2000s, after which it gradually declines. This trend may be attributed to Diversification of Music Outlets, Shifting Consumption Patterns, Evolving Measurement of "Top Songs".
- This trend also presents interesting opportunities for targeted music marketing. Knowing that listeners often gravitate towards music from their formative years, targeting high-scoring songs from different eras for specific age groups could be an effective strategy.
- Platforms could curate personalized playlists based on user demographics and musical preferences, further enhancing listener engagement.

Data Visualization: We employed various visualizations to unravel key patterns and insights within the million songs dataset. From identifying prolific artists and popular releases to uncovering trends over the years, these visualizations shed light on the multifaceted dynamics of music popularity, artist familiarity, and the diverse characteristics of top-ranking songs.

- **Top 10 Artists with the Most Songs:** The bar chart displays the top 10 artists with the most songs. **Insights:** Identifies prolific artists based on the number of songs.
- **Top 10 Artists by Cumulative Score:** The bar chart illustrates the top 10 artists based on the cumulative score of their songs. **Insights:** Highlights artists with the highest overall scores.
- **Top 10 Releases by Score:** The bar chart displays the top 10 releases based on cumulative score. **Insights:** Identifies releases (combination of artist and song) with the highest scores.
- **Distribution of Song Durations:** The histogram shows the distribution of song durations in minutes **Insights:** Reveals the common range of song durations.
- **Boxplot of Artist Familiarity:** The boxplot visualizes the spread of artist familiarity values. **Insights:** Provides insights into the variability of artist familiarity in the dataset.
- **Trend of Song Releases Over the Years:** The line plot depicts the trend of song releases over the years. **Insights:** Shows the evolution of song releases, highlighting periods of growth or decline.
- **Scatter Plots:** Several scatter plots explore relationships between variables, including artist familiarity vs. score, artist familiarity vs. song duration, and song duration vs. score. Insights: Visualizes relationships and potential patterns in the data.

Insights:

- **Quantity vs. quality:** high scores favor not just songs, but musical impact.
- **Different strokes for different folks:** artist scores might favor critic favorites, while release scores celebrate personal playlists.
- **Shorter reigns supreme:** Songs under 5 minutes are more prevalent than longer jams.
- **Booming Music:** The number of songs released has skyrocketed over the decades, from fewer than 5,000 in 1920 to nearly 40,000 by 2000.
- **Familiarity Matters:** The more familiar listeners are with an artist, the higher they tend to rate their music.

Recommendation:

- Promote artists with the most top songs and best overall scores.
- These artists are clearly resonating with listeners and achieving high ratings.

- Use "release by score" to identify specific albums with exceptional scores. Highlighting them in playlists, recommendations, and other promotional materials.
- Score remains the primary driver for music recommendations, ensuring quality content takes precedence.
- Introducing familiarity as a secondary filter promotes a diverse music experience while still prioritizing high-quality tracks.
- Execute targeted campaigns to introduce artists with high scores but low familiarity, utilizing curated playlists and promotional initiatives.
- Artists with high familiarity but lower scores may benefit from strategic breaks or shifts in musical direction. Analyze listener feedback for valuable insights into improvement.
- Consider focusing on the "optimal length" range that tends to perform well based on the data.
- Be mindful of listener preferences and avoid overly short or long tracks unless it aligns with the artist's creative vision.

Conclusion:

In our journey through the "Music Trend Analysis of Million Songs Dataset," we've uncovered fascinating patterns in the music world. Exploring artist scores and familiarity, we found the keys to promoting diverse music effectively. Our visuals on top artists, releases, and song length provide a clear tune for strategic decisions. This project isn't just about numbers; it's a guide for navigating the ever-changing landscape of music trends, ensuring that every note resonates with the audience.