Music Factors Analysis

McAelan Remigio, Benjamin De La Torre, Mary-Grace Knepper,
Vince Alihan, Sebastian Carrillo, Shizu Takura

Abstract

This study investigates the factors contributing to a song's success on music charts, such as the Top 50 or Top 100. Leveraging datasets from Spotify and Apple Music, we analyze tempo, danceability, and overall popularity to uncover patterns to distinguish chart-topping hits. Following a structured data science process, we preprocess and clean the data, explore key trends, and utilize SAS for predictive modeling. Through this analysis, we aim to understand the elements that drive musical success, offering meaningful insights into the relationship between song attributes and listener appeal.

Methodology

This research investigates the factors contributing to a song's popularity and its rankings on music charts. Specifically, we aim to analyze musical attributes such as genre, tempo, danceability, energy levels, and artist characteristics to identify patterns determining a song's commercial success. The primary objective is to derive data-driven insights to assist artists, producers, and marketers in their music production strategies by understanding listener preferences

The relevance of this study lies in its potential to provide actionable insights within the competitive music industry, where data analysis can inform marketing and production decisions. This paper is organized as follows: we will begin with a review of existing literature on music popularity determinants, followed by a methodology section detailing our data collection and analysis processes. We will then present our findings, discuss their implications, and conclude with recommendations for future research and practical applications in the music sector.

Importing Two CSV Data Files using SAS Studio

The datasets used in our analysis were sourced from two leading platforms in the music streaming industry. Our primary dataset was from Spotify while our secondary source was Apple Music. These two datasets were selected due to their relevance and scale, making them ideal candidates for exploration. The Apple Music dataset contains 10,000 rows, while the Spotify dataset has over 100,000 rows, providing a rich foundation for analysis. The primary goal of this initial step—importing the CSV files—was to gain an overview of the data's structure and content, which would inform our deeper exploration. To facilitate this, we utilized the PROC CONTENTS procedure in SAS to generate metadata, allowing us to inspect column names, data types, and variable attributes, ensuring that we understood the underlying structure of the datasets before proceeding with further analysis.

Performing Exploratory Data Analysis (EDA) to Understand Data along with any Issues

To begin the first step of our analysis, we were supposed to perform PROC PRINT on the original datasets. However, missing and non-compatible characters caused multiple errors due to not being a part of the ASCII dictionary. This led to an executive group decision to clean both datasets before continuing which will be discussed further on.

Once the datasets were cleaned, to perform exploratory analysis on the Spotify dataset, more rearrangement/cleaning was done to see specific results. For the track_genre column, many rows had two to four genres listed, so we narrowed the number of genres in each row to one. This generalized a lot of songs into main genres to decrease the amount of genres and subgenres that only differed by one or a couple of subgenres. We also added a column called duration_grouped that grouped the duration of songs into three categories. One for Short songs which are 1-3 minutes long, another for Medium which are 3-5 minutes long, and the third, Long, for songs over 5 minutes. This categorization helped look at the data and know roughly how long a song was instead of trying to look at the length in terms of milliseconds. Both of these cleaning steps made PROC FREQ work for certain variables.

To begin performing procedures, we produced one- and two-way frequency distributions on the variables track_genre, explicit, and popularity. This shows the distributions of how many songs are in each genre, how many tracks are explicit or not, and the general popularity of the tracks (more tracks at certain popularity scores). Two-way frequency distributions compared popularity against duration_grouped and track_genre against explicit. This reveals how many songs are in each duration category (short, medium, or long) at each popularity, and how many songs are explicit or not in each genre.

Continuing, descriptive statistics were performed on 11 numerical variables using PROC UNIVARIATE and producing individual and overlapping histograms to visualize the distribution and its statistics. These variables were popularity, duration_ms, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. For popularity and duration_ms, outliers were found by calculating Q1, Q3, IQR, upper, and lower bound, then filtering the data to output rows below or above the outlier cut-off.

PROC MEANS, PROC SORT, and PROC PRINT were performed to create tables to compare popularity scores and track genre against the numerical variables excluding duration_ms. We found the four genres and scores with the highest and lowest averages. Finally, we found the averages of the variables for when the frequency of songs in a category was the highest including the four popularity scores and genres with the highest number of songs.

Performing Data Cleaning and Data Standardization

Cleaning data should occur after initially importing and exploring the dataset, however, this was impossible. Our Spotify dataset contains over 100,000 rows of songs from all over the world. This naturally meant that some songs and their information were written with non-ASCII characters. SAS does not handle non-ASCII characters as many functions do not recognize said characters, thus in the interest of time, any row containing non-readable characters had to be deleted. At this point, cleaning started with the deletion of those rows. Our team then decided to eliminate songs less than 60000 milliseconds and a popularity level of less than 50. This decision was taken in the interest of time and as a way of filtering the less important data. After that, we noticed that the dataset contained duplicates of the exact song with different genres which was known because of the track_Id. To prevent valuable data from being lost we combined the genre of duplicate songs. We rearranged the columns to improve readability and facilitate analysis for our final Spotify data cleanup step.

The Apple dataset did not contain non-ASCII characters or have missing values. However, the dataset filled the character string with “music” for anything that did not contain values for those variables. Naturally, those entries containing said string were deleted. Additionally, this dataset did contain duplicates identified by music URLs which were decidedly deleted. We also decided to change true and false variables into binary form. Finally, for the merge step to work properly, additional standardization was needed to ensure proper behavior when it came to merging. Changes done solely for this purpose included renaming certain variable names to avoid conflict with variable names from the Spotify dataset.

Performing Data Transformation

We focused on merging the datasets from Spotify and Apple Music to create a unified view of the music landscape. After cleaning both datasets to remove duplicates and irrelevant entries, we sorted them by track names to prepare for the merge. The merging process involved using a left join based on the track names, allowing us to align songs that appeared on both platforms. This was crucial because it ensured we could analyze the same tracks across different services.

Before the merge, we utilized PROC CONTENTS to generate metadata for both datasets. This step provided us with valuable information about the structure of the datasets, including variable names, types, and lengths. Understanding the metadata helped us ensure that the columns we intended to merge were compatible and correctly formatted, reducing the risk of errors during the merging process.

To facilitate the merge, we created a frequency table for the track_name column in the Spotify dataset, which helped us identify how many times each track appeared. We then performed a left join using PROC SQL, combining the cleaned Spotify dataset with the frequency table. This step ensured that we retained all relevant information from the Spotify dataset while adding the

corresponding data from Apple Music.

After merging, we filtered the resulting dataset to keep only the tracks with a popularity score of 50 or higher, ensuring that we focused on more successful songs. Additionally, we handled cases where tracks had multiple genres by creating a new variable that combined these genres into a single column by using a RETAIN statement to accumulate genre information for tracks with the same ID. This ensured all relevant genres were captured without losing any details.

Finally, we sorted the merged dataset to ensure all entries were organized and ready for further analysis. This meticulous merging process combined with PROC CONTENTS was essential for creating a comprehensive dataset that accurately represented the music available on both platforms, setting the stage for deeper insights into the combined data.

Producing PDF, HTML, and Excel Results using an Output Delivery System (ODS)

Given that SAS has several ODS destinations we decided to focus our attention on Excel, PDF, and HTML. We developed our code for merging two datasets and conducted a comprehensive analysis of the merged data to uncover key patterns and trends. These pieces of information were subsequently directed to the Output Delivery System (ODS) for formatting and presentation in the desired report format.

Producing Macro Variables and Developing Macro Routines

We also developed macro programs using SAS to analyze a merged dataset of songs from both Spotify and Apple Music, concentrating on various musical factors contributing to a track's success. The process started with importing the merged dataset and creating three essential macro variables: the number of tracks, unique artists, and the average track duration in minutes. We achieved this by using PROC SQL queries that leverage the *count()* and *means()* functions to extract these important statistics from the dataset.

Next, we implemented two macro routines. The first, *genre_report*, generates a detailed report for a specific genre with the PROC MEANS procedure to calculate summary statistics, including the mean, standard deviation, minimum, and maximum values for attributes such as popularity, danceability, and energy. Additionally, it employs PROC PRINT to list the tracks within that genre, providing a clear view of the songs that fit the criteria. The second, *artist_tracks*, lists tracks by a specific artist. We used the *findw()* function to manage cases where delimiters separate artists.

This approach is particularly significant as it streamlines data analysis and reporting, allowing us to uncover valuable insights into key attributes across different genres and artists. By analyzing the average popularity and energy levels of tracks within specific genres, we can identify trends that may indicate what makes certain genres more successful. Ultimately, these macro programs

enhance our ability to make informed decisions in the music industry by revealing patterns and relationships within the data from both Spotify and Apple Music. This information can be crucial for artists, producers, and marketers looking to optimize their strategies and understand audience preferences better.

Producing a Data-driven Program

We then developed a data-driven program using SAS to analyze the merged dataset of songs from Spotify and Apple Music, which allowed us to investigate various musical factors that influence a song's success. The program begins by importing and transforming data to create new variables, such as converting track durations from milliseconds to minutes and flagging explicit content. It also addresses missing values to ensure the accuracy and reliability of our analysis.

Furthermore, the program generates summary statistics in multiple formats, including PDF, HTML, and Excel, which provide insights into key attributes like popularity, danceability, and energy levels across different genres. This multi-format reporting can be a detailed report or a quick summary.

The significance of this program lies in its ability to demonstrate how data analysis can uncover patterns and relationships within music data. This program helped us identify trends indicating what characteristics contribute to a song's popularity and examine average energy levels and danceability of tracks within specific genres. These insights are invaluable for artists, producers, and marketers, as they can inform strategies for song production, marketing campaigns, and playlist placements in the highly competitive music industry. Ultimately, this program enhances our understanding of the factors that drive success in music, enabling stakeholders to make more informed decisions based on data-driven evidence.

Producing Data Visuals and a Dashboard

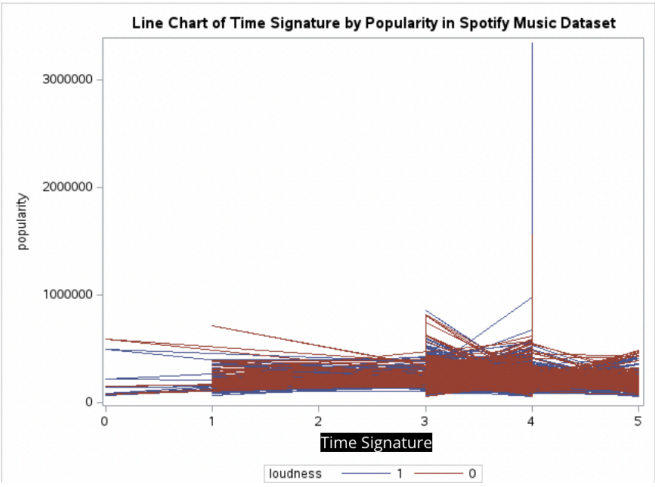

Figure 1: Line Chart illustrates the dataset includes time signatures ranging from 0-5.

There is 1/4, 2/4, 3/4, 4/4, 5/4, and for 0, it is a different time signature. It follows a slightly skewed left distribution which shows that the closer the time signature value approached 4, the higher it got, and after it reached 4, it descended quickly. The correlation is moderate up to time signature 4, as there is a steady upward trend. The sharp spike at 4 and quick descension after suggest that the overall correlation between time signature and popularity is not strictly linear across the entire range. Contributing Factors: Time signature 4 is the most common in mainstream music, which could skew the correlation.

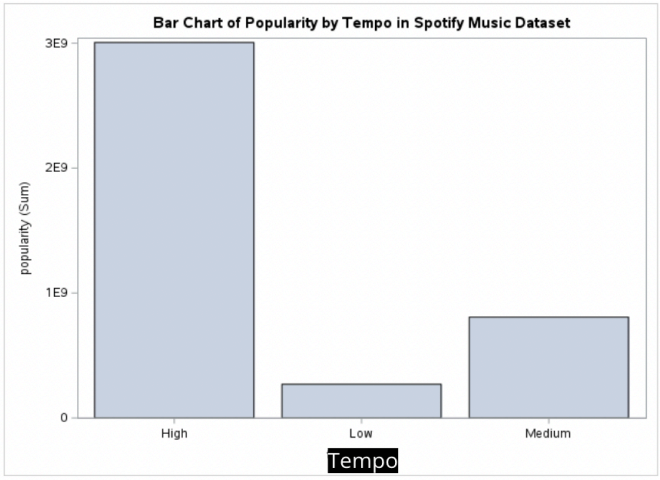

Figure 2: Bar chart illustrates the popularity of three types of tempos: High, Medium, and Low.

The figure shows that songs with a higher tempo are likely to be more appealing to audiences while songs with a medium tempo lead next and songs with a lower tempo fall last to popularity.

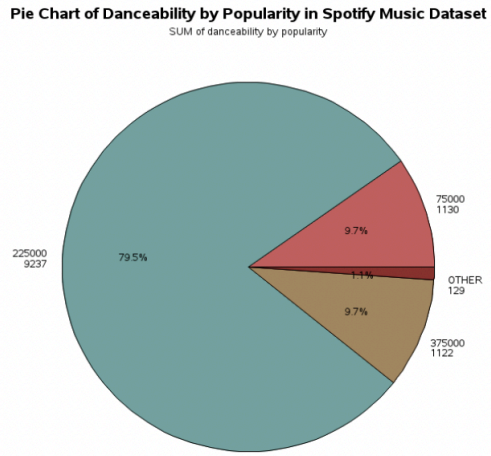

Figure 3: Pie Chart shows the danceability of songs and its correlation to the popularity of the same song.

This data investigated 18,600 observations and the figure shows that the danceability of a song has a strong positive correlation with the popularity of the song. The higher the danceability of a song, the more popular a song is likely to be, and vice versa. The lower the danceability of a

song, the less popular a song is likely to be. Therefore, the Pie Chart shows how the danceability of a song has a strong positive correlation to a song's popularity.

Results

EDA Results

By performing one-way frequency distributions, we see 110 generalized genres with "pop-film" having the most songs: 728. Chill, k-pop, and grunge are the next consecutive genres with 645, 564, and 539 songs respectively. The genre with the fewest number of songs is IDM (intelligent dance music) with 2 songs. Salsa, samba, and Detroit-techno all have the next lowest number with 5. Our popularity scale goes from 50-100 so we are only looking at the upper half of popularity. A popularity score of 56 has the most songs with 1025. 50 and 57 with the next highest with 988 and 984 songs. The popularity scores of 100, 99, 97, and 96 are all tied with 1.

Looking at the two-way frequency distributions reveals that the popularity score of 56 has 545 songs in the 3-5 minute range which is the highest of the three categories. Short (1-3 minutes) had 347 songs while long songs(5+ minutes) had 133 songs. All the songs in the genre of pop-film are not explicit, which is 728 songs while the genre with the most explicit songs is emo with 247 explicit songs. The 20 genres with the highest number of songs in a category (explicit or not) happen to be all false.

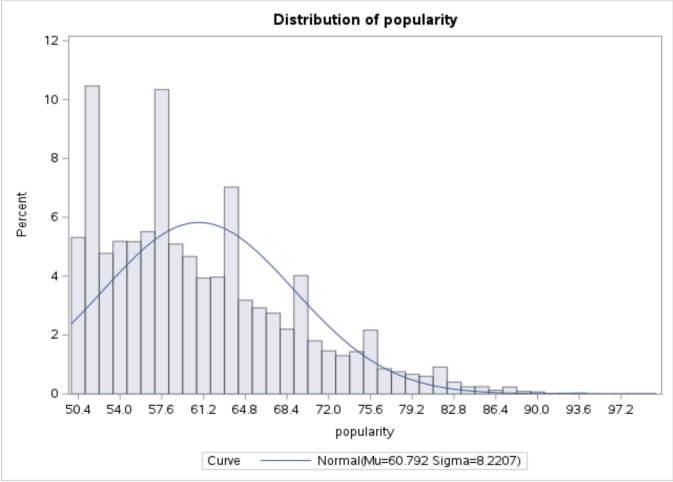

Figure 4: Histogram of Popularity Scores

From our descriptive statistics, the variable popularity has a mean score of 60.79 which makes sense because earlier we saw that the popularity score with the highest number of songs is 56. The histogram displayed shows the distribution as right-skewed compared to the normal distribution. The statistics correlate with this visual as the skewness value is 0.899, the mean and highest number of songs is around the high 50s, a few high popularity scores such as 100, 99, and 98 are present, and the occurrence of 165 outliers.

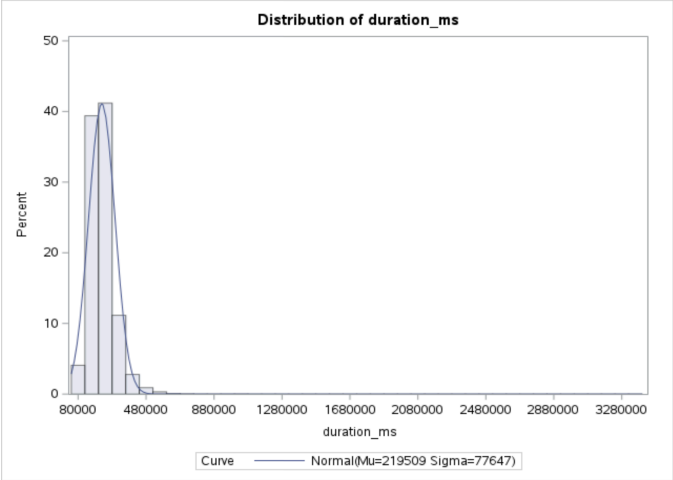

Figure 5: Histogram of Duration (in milliseconds)

For the variable duration_ms, the mean is 219509.18 milliseconds which correlates to a song length of 3.66 minutes. Having the value fall in the song length category of medium (3-5 minutes) aligns with how this category had the most number of songs compared to the two other categories. The maximum is 3340672 milliseconds which is 55.68 minutes. The 698 outliers in this distribution are all longer than 6.12 minutes. The histogram displayed shows the distribution very much right skewed with the majority of the data clustered toward the lower end of the distribution. This right skewness makes sense due to the skewness value of 5.38 and the large number of outliers of very high song lengths.

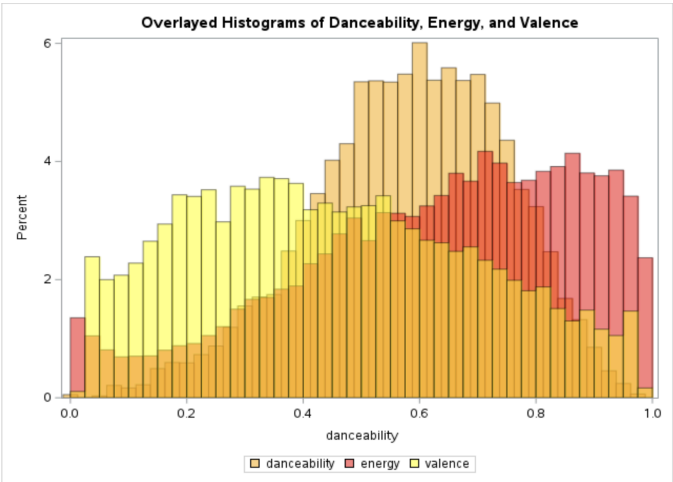

Figure 6: Overlapping Histogram of Danceability, Energy, and Valence

Doing some more comparisons on danceability, energy, and valence resulted in the means being roughly close to each other and a lot of overlap between the three variables. The means for danceability and energy result in a mid-tempo which mirrors what the mean for tempo suggests.

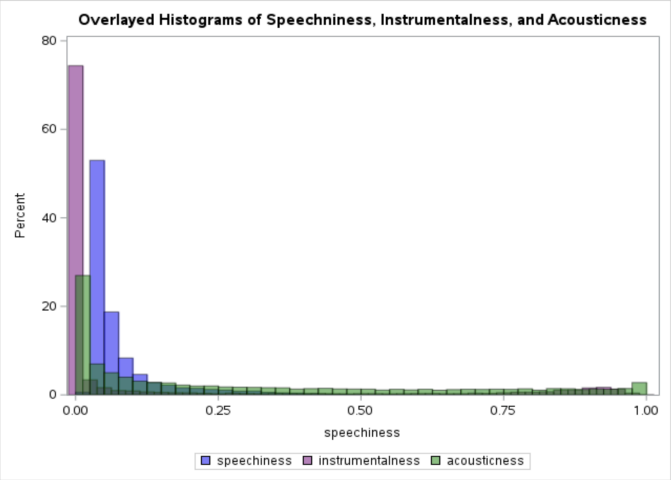

Figure 7: Overlapping Histograms of Speechiness, Acousticness, and Instrumentalness

Comparing speechiness, acousticness, and instrumentalness results in a correlation with each other. The lower speechiness level means songs are mostly singing while instrumentaness indicates most songs are mostly vocals which coincide with each other.

Averages for Popularity Score with Largest Number of Songs

popularity	FREQ	mean_danceability	mean_energy	mean_loudness	mean_speechiness	mean_acousticness	mean_instrumentalness	mean_liveness	mean_valence	mean_tempo
56	1025	0.58180	0.62068	-8.40911	0.073749	0.31577	0.13935	0.18057	0.44117	122.724
50	988	0.56452	0.62026	-8.43593	0.074397	0.31041	0.19211	0.17972	0.38831	123.684
57	984	0.56759	0.60440	-8.71622	0.084016	0.34605	0.14435	0.17685	0.44756	120.640
51	977	0.56088	0.61006	-8.74925	0.072664	0.33015	0.22211	0.18142	0.38059	120.840

Figure 8: Averages for Variables with Top Number of Songs in Popularity Score

A popularity score of 56 has the most songs in a score category whose mean for all the variables roughly follows the means for the dataset as a whole. This indicates that the majority of moderately popular songs are ones with a medium-tempo, more melodic rhythm, whose entirety is singing/vocals, produced in a studio, and have a balanced range of emotions.

Averages for Track Genres with Largest Number of Songs

track_genre	FREQ	mean_danceability	mean_energy	mean_loudness	mean_speechiness	mean_acousticness	mean_instrumentalness	mean_liveness	mean_valence	mean_tempo
pop-film	728	0.59566	0.60161	-7.9844	0.06085	0.43255	0.00929	0.16988	0.52704	117.420
chill	645	0.66602	0.42853	-10.3751	0.10670	0.52815	0.16876	0.16901	0.41322	114.444
k-pop	564	0.64797	0.71193	-5.6296	0.09107	0.23069	0.00832	0.17934	0.56125	119.600
grunge	539	0.45171	0.78623	-5.8851	0.05937	0.05968	0.04493	0.19813	0.39783	129.976

Figure 9: Averages for Variables with Top Number of Songs in Track Genre

The track genre with the most songs is pop-film. The averages deviate from the averages from the overall dataset slightly from popularity scores but it suggests that most songs that have a popularity score of around 56 are in the genre of pop and fall between three to five minutes.

Merged Results

After merging and transforming both datasets, we were able to find the key factors that make a song appealing and popular to audiences. Three factors (tempo, danceability, and time signature) we believed to have an important influence on the popularity of songs, and conducted a data-driven analysis with the merged dataset. By doing so, we found that the correlation of the variables– tempo, danceability, and time signature– to the popularity of songs is a high, positive correlation, and closer to 4/4 signatures, respectively.

Using the merged dataset containing 18,600 observations, we found that songs with higher tempos were more likely to become popular, while songs with lower tempos were less likely to become popular. High-tempo songs were estimated to be ~3 * 10^9, medium-tempo songs were estimated at ~1 * 10^9, and low-tempo songs were estimated to be well below ~1 * 10^9. We found that there is a positive correlation between the tempo of a song and popularity, as higher tempos were associated with more popular songs. High-tempo songs are often correlated with high energy and engagement, therefore making them suitable for environments such as workouts, parties, and social gatherings. The higher the tempo, the more likely a song reaches higher popularity due to its likeness and favorable environmental situations.

We also used the dataset to determine the correlation between danceability and popularity in songs. The analysis of danceability focused on the relationship between tempo and popularity between the 18,600 observations. We found that there was a strong positive correlation between the danceability of songs and their popularity. The biggest section of the analysis sits at 79.5% and its top popularity value is 225,000 while its # of instances is 9,237. The biggest section suggests that most songs fall into a category of high danceability which accounts for a majority of the popular songs. The medium danceability slice is smaller, at 9.7% and its top popularity value is 75,000 while its # of instances is 1,130. The medium-sized slice indicates that songs with a medium danceability range manage a medium amount of popularity but not as much as the high danceability slice. The low danceability slice contains 9.7% of the values, with the top popularity value being 375,000 and the # of instances being 1,122. The lower slice having a top value of 375,000 suggests that the group is still popular despite having a small number of instances. However, this data can be skewed and indicate there is a small amount of popular songs for danceability. At the extremely low danceability slice, at 1.1% the top popularity value is deciphered as “OTHER” and the # of instances of “OTHER” is 129. This value indicates that there is a small number of songs with extremely low danceability.

Lastly, we analyzed the factor of time signature in popularity and found that quiet songs dominate overall popularity, but loud songs achieve a dramatic peak at the widely used time signature 4, highlighting their potential for significant impact in certain rhythmic structures. The analysis considers the differentiation between quiet and loud dynamics. From time signature values $0 \leq x < 4$, both quiet and loud dynamics (red and blue lines) show a steady increase in

popularity with quiet songs dominating the chart visually. At time signature value 4, there is a spike in popularity. Initially, quiet dynamic songs remained prominent, but as the spike grew higher, it quickly turned into loud dynamics which took over and dominated high popularity. After the spike, $4 < x < 5$, the popularity level returns to match those seen at time signatures $3 \leq x < 4$.

Ultimately, we found a strong positive correlation between our three variables (tempo, danceability, and time signature) to popularity.

Conclusion

This study aimed to explore the key factors that make a song appealing by taking datasets from two of the most popular streaming services: Apple Music and Spotify. By utilizing SAS Studio and its tools to compare, analyze, and explore both datasets, we were able to conclude which variables were ideal to make a song appealing to its listeners.

We first cleaned both Apple Music and Spotify datasets by deleting, combining, and simplifying duplicate variables. We found that cleaning both datasets simplified our data to a more workable and understandable model, which stood as the basis for the rest of our data transformations.

With the cleaned dataset, we were able to merge and produce a data-driven program to find the conclusions to our driving question, “What are key features of songs that predict their likelihood of reaching charts like the top 50, top 100, etc?”. The three variables we used to analyze using our merged dataset were tempo, danceability, and time signature. Using SAS tools such as PROC MEANS, PROC IMPORT, and PROC EXPORT, we were able to find the relationship between each of the three variables and song popularity.

To conclude, we found that our three variables – tempo, danceability, and time signature – play a crucial role in shaping the popularity of songs, which offer useful insights for listeners and producers who aim to create music that resonates with its listeners.

References

Kanchana1990. *Apple Music Dataset: 10,000 Tracks Uncovered*. no. 2023. *Kaggle*, https://www.kaggle.com/datasets/kanchana1990/apple-music-dataset-10000-tracks-uncovered.

TheDevastator. *Spotify Tracks Genre Dataset*. no. 2023. *Kaggle*, https://www.kaggle.com/datasets/thedevastator/spotify-tracks-genre-dataset.