Business Description and Data Summary

We plan to help Universal Music Group (UMG) increase their number of streams -- the Spotify data can help identify trends and other useful information. The dataset can help us answer what song characteristics get more streams and what the current landscape of music streaming is; especially with Spotify reporting 286 million active month-to-month users in 2020. We plan to use the data to find correlations among core 'characteristics' that can lead to insight in user preferences. Based on the premise that Spotify's popularity is gauged by the number of streams, we plan to answer questions such as, which genres are the most popular, why is it popular, and what patterns emerge from highly streamed songs.

To get an overall picture of the representation of the data, we first looked into the simplest level of categorization directly given: genre. Based on this division of songs, we found that most genres are evenly represented -- however, Indie, Pop, Rock, Rap, and Children's music are underrepresented, with A Capella being the most underrepresented by a substantial margin (Figure 1). Further delving into genre, popularity analysis via box-plot showed that the most popular genres based on streaming metrics are Dance, Hip-Hop, and Pop (Figure 2). From this, Pop seems to be a potential outlier. Even though it's one of the least represented genres, it has the most streams, implying a greater number of streams per song. Judging by the most popular genres, we speculate that listeners trend towards the younger demographic.

We also noticed that there were also some correlations that exist in the data. From the correlation matrix of streams to other audio characteristics (Figure 3), strongly correlated variables included: instrumentalness and acousticness - negatively correlated, danceability and loudness - positively correlated, and loudness and energy - positively related.

Recommendations

Before discussing our recommendations, we want to preface that there were some difficulties in developing effective models from the data. After finding that many of our regression models had low R^2 values (Figure 4,5), along with the understanding that music popularity is never based on a single characteristic, we determined that there isn't a single rule which contributes to streamability. There's a lack of context with looking solely at data points -- circumstances such as the pre-existing popularity of artists, artists comebacks, and other situations which all have an influence on streams. Residual points of our potential models were also not formless, meaning that there may be better models than linear regression (Figure 6). Furthermore, after building multiple linear models, the QQ plots seemed to violate normality since the points didn't follow the trend of the QQ line. Each time, the QQ line and its points lined up only to a certain point when they suddenly exponentially diverged (Figure 7,8). This suggests skewed data and a different distribution, which is important to keep in mind. All these factors prevented the development of a simple model.

The first trend we noticed is that in the music industry, songs have generally been getting louder over the years (Link 1), so it's an important variable to look at in this dataset. Earlier, we found a positive correlation between loudness and energy, making it worth looking into energy as a potential variable. After plotting energy against streams categorized from high to low, in a histogram, the distribution of high streams became left-skewed compared to danceability, suggesting that energy might be a confounding variable (Figure 9,10). We then created a bar graph which concluded that high energy and loud songs tend to have the highest streams, indicating a positive correlation (Figure 11). Then, from the interaction model between loudness and categorical energy level, we confirmed that for both energy levels, loudness and streams are positively correlated (Figure 5). Even though the R^2 was low: 0.01, all the variables and the interaction term were significant ensuring that the model has some relevance. Listeners are more likely to stream songs if it tends to be louder, regardless of energy. UMG should consider this when compressing songs and aim for a higher decibel level because it seems to increase streams.

After looking into the most popular genres' audio traits, another pattern we noticed was that characteristics which were the most representative of song popularity often related to the genre, especially the characteristics of energy and danceability -- for example, pop songs should have high energy and higher danceability levels to have a higher potential of getting streams (Figure 12). However, this analysis was not statistically significant for every genre, meaning only certain genres, such as pop, dance, and rap, require certain levels of these traits. However, UMG should strive to match energy and danceability to the genre if they plan on promoting a song in a particular genre.

Future Considerations

From the information given, we don't know when this data was collected -- having said data would allow us to answer questions such as the growth of these categories or songs' streams over time, allowing us to find current and upcoming music trends. In the future, we recommend looking into release years since some songs in the dataset came out later on Spotify than others, possibly making the number of streams disproportionate since older songs may have more streams than others. Being able to filter songs on a variable such as "timeless" or "seasonal" would be helpful, since, in comparison to other songs which have a limited lifespan of popularity, these songs may be considered outliers since they are played every year. Furthermore, the trends of TikTok and other social media streaming platforms have an effect on virality on streaming. For instance, songs are becoming popular again because of social media, which makes social media a confounding variable that data can be collected on.