Song Survey

Masters in engineering management - Thayer School of Engineering

Written by
Pranav Dharmadhikari

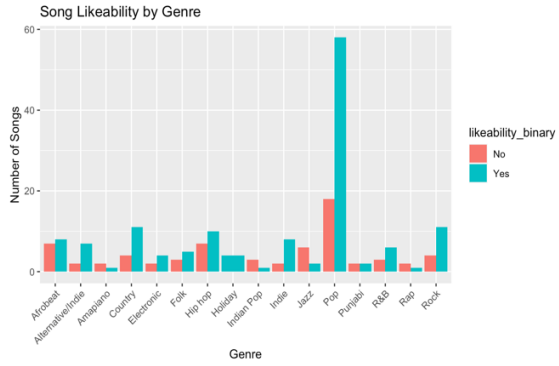
Table of contents

1. Introduction

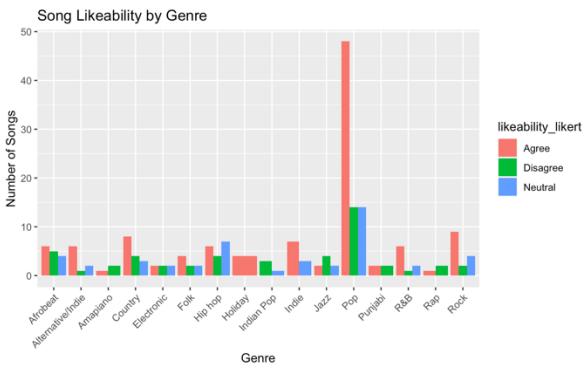
- This report aims to understand the likeability of different types of songs based on myriad factors such as song name, year of release, length etc. Through comprehensive analyses, we visualize patterns of likeability based on genre, country, and language and assess the potential influence of song length and release year, shedding light on temporal and structural trends impacting likeability. Moving beyond descriptive analyses, we develop a predictive model to estimate the probability that a song is likeable, highlighting the challenges associated with capturing the essence of music preferences. Lastly, by suggesting additional data sources and analytical approaches, we aim to make more accurate predictions and visualizations to understand the song likeability based on the factors provided.

2. Likeability based on data points

- **Genre:** In my examination of song likeability, I observed that pop garnered the highest number of responses, emerging as the most liked genre among the surveyed class, based on both the 'likeability_likert' and 'likeability_binary' columns. Despite pop's overall popularity, my analysis accounts for potential variations in sentiment by considering the two distinct columns. Notably, when exploring a niche genre like "electronic," I recognized that visualizations based solely on 'likeability_binary' might not accurately reflect the class's sentiment. This distinction is crucial as it allows us to capture more nuanced responses where a student might endorse a song ('likebility_binary' = "yes") but hold a neutral stance ('likeability_likert' = "neutral"). This approach ensures a more accurate portrayal of the class's preferences, particularly in genres where sentiments may vary.

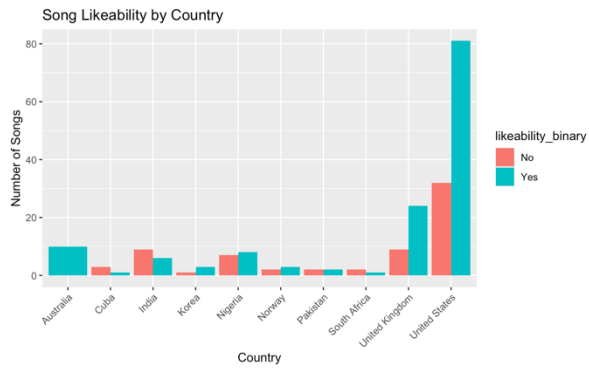


2.1.1Binary Likeability of songs based on genre

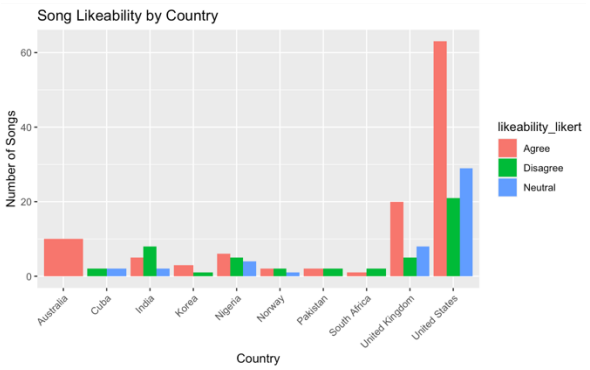


2.1.2 Binary Likeability of songs based on genre

- **Country:** Issue with the data: To address data inconsistencies in country entries, particularly in the case of the United States, variations like "Canada, United States" and "United States " were identified. To ensure accuracy in visual representations, both "Canada, United States" and "United States " were combined under a unified country name, "United States." This consolidation resolves potential inaccuracies and ensures a more precise and cohesive dataset for subsequent analyses. United States-produced songs dominate in likeability based on the data. However, relying solely on binary responses may overlook nuances, especially in countries like Cuba, where a "dislike" response might reflect a neutral sentiment. Utilizing more granular responses, such as the Likert scale, proves crucial in capturing these subtleties for a nuanced understanding of song preferences.

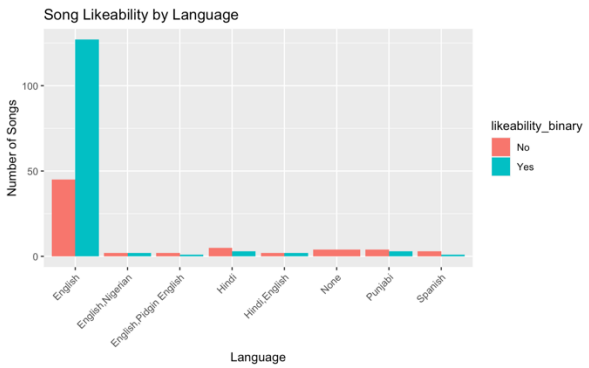


2.2.1Binary Likeability of songs based on country

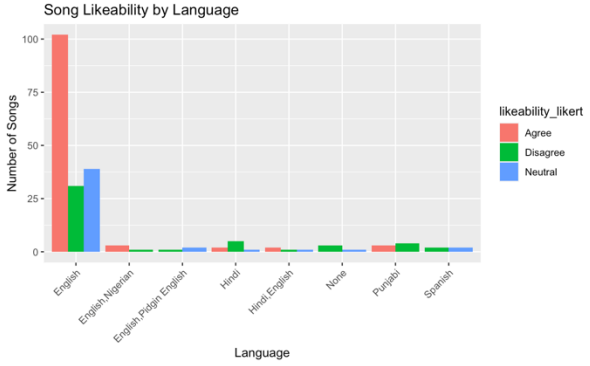


2.2.2 Binary Likeability of songs based on country

- **Language:** English-language songs are the most liked in the dataset. The data hasn't been extensively cleaned for better visualization due to the complexity of songs with multiple languages, like "English, Nigerian" and "English, Pidgin English." While this approach may introduce some skewness, it represents diversity for a more accurate understanding of likeability trends based on language.



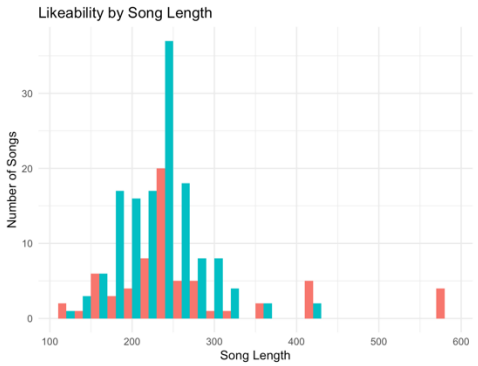
2.3.1Binary Likeability of songs (language)

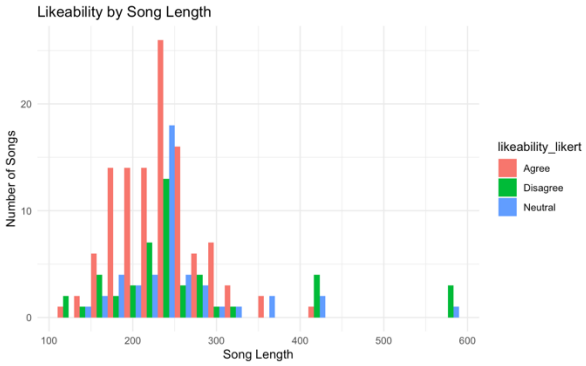


2.3.2 Non- Binary Likeability of songs (language)

3. Likeability based on song length and year of release

- **Length vs binary response for Likeability**: Correlation Value: -0.1941259
- **Interpretation:** There is a weak negative correlation (-0.19) between the length of songs and the "Yes" and "No" response of likability. This suggests that, on average, longer songs tend to have slightly likeability. This correlation is also supported by the data since the data suggests that the songs that are in between 200 to 300 seconds are the most liked.
- **Length vs non-binary response for Likeability:** Correlation Value: 0.1705389
- **Interpretation:** There is a weak positive correlation (0.17) between the length of songs and the "Agree" "Disagree" "Neutral" scale. As the binary response we see a similar spike for the song length across 200 to 300 seconds. Having said that, when it comes to the songs with longer lengths the sentiment is "Neutral" along with "Disagree" rather than a simple "No" in case of the binary responses which justifies the weak positive correlation.



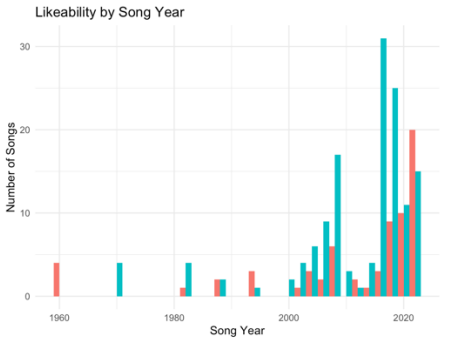


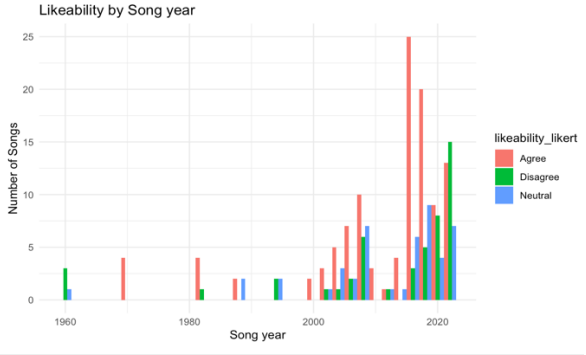
3.1.1Binary Likeability of songs based on length *3.1.2 Non-Binary Likeability of songs based on length*

- **Year vs (binary response for Likeability) & (non-binary response for Likeability):**

Correlation Value for binary: 0.01523782, Correlation Value for non-binary: 0.00894837

Interpretation: There is a very weak positive correlation (0.015) & (0.009) between the release year of songs and the binary and non-binary response of likeability. This suggests that there is almost no discernible relationship between the release year and likability. Even though the visualizations represent that the most liked songs are between the decade 2010 to 2020 there is also a significant level of likeability for the songs that were released between 2000 to 2010 for both binary and non-binary responses which justifies the correlation value and the fact that it has a weak positive correlation.

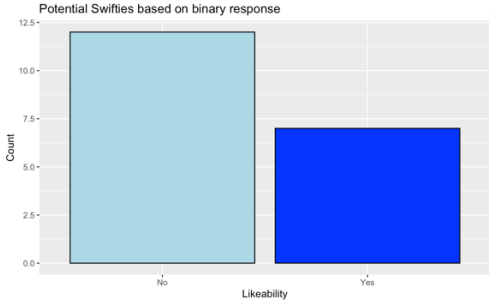


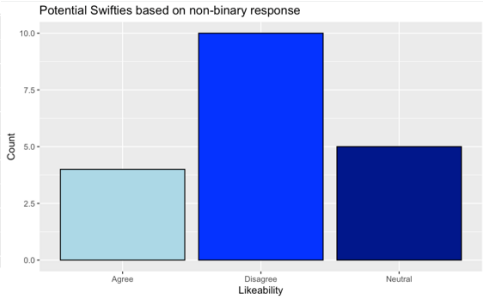


3.2.1Binary Likeability of songs based on Year *3.2.2 Non-Binary Likeability of songs based on Year*

4. Potential Swifties assessment

- Based on the data analysis, it appears that the likeability for Taylor Swift's songs is low among the surveyed class, as indicated by both binary and non-binary responses. Despite the presence of potential Swifties enrolled in the class, the overall conclusion is that the class does not generally favor Taylor Swift's music. This observation suggests that individual preferences within the class may not align with Taylor Swift's musical style or genre, contributing to a lower overall likability for her songs.





4.1.1 Likeability of Swifties in Binary. 4.1.2 Likeability of Swifties in Non-Binary

5. Model Analysis

- The logistic regression model1 was fitted to predict the binary likability of songs based on the variable's 'year' and 'length' and model2 was fitted with 'year' 'length' 'country' and 'genre'.
- The negative coefficient for 'year' (-0.036) suggests that for every one-unit increase in the 'year,' the log-odds of likability decrease by 0.036, and this effect is statistically significant (p-value = 0.048) since effects are said to be statistically significant if the p-value is less than 0.05.
- For a one-unit increase in the 'year' in model2, the log-odds of the response variable decrease by approximately 0.04839 and p-value associated with the 'year' variable is 0.0907, which is greater than the common significance level of 0.05. Therefore, based on a significance level of 0.05, we can conclude that this correlation is not statistically significant.
- For a one-unit increase in the 'length' in model2, the log-odds of the response variable decrease by approximately 0.0107 and p-value associated with the 'year' variable is 0.1230, which is greater than the common significance level of 0.05. Therefore, based on a significance level of 0.05, we can conclude that this correlation is not statistically significant.
- The negative coefficient for 'length' (-0.0096) in model1 implies that for every one-unit increase in 'length,' the log-odds of likability decrease by 0.0096, and this effect is also statistically significant (p-value = 0.0015).
- Overall, this logistic regression model provides insights into the predictors influencing the likability of songs in a binary context, with both 'year' and 'length' demonstrating statistically significant effects.
- The mean predicted probability in this logistic regression model is 0.6747573 for model1, suggesting that, on average, the songs in the dataset have a likelihood of approximately 67.4% of being classified as likable based on year and length.
- The accuracy of model1, as determined by the confusion matrix, is 0.6893. Moreover, a 95% confidence interval for the true accuracy is calculated to be between 0.6213 and 0.7518, providing a range within which we can reasonably expect the model's accuracy to fall.
- The accuracy of model2, as determined by the confusion matrix, is 0.773. Moreover, a 95% confidence interval for the true accuracy is calculated to be between 0.6671 and 0.7921, providing a range within which we can reasonably expect the model's accuracy to fall.

6. Issues with the modelling

- In fitting the logistic regression model using the formula `binaryconvertedtonumeric ~ year + length` in the glm function, it's important to note that the binary converted table was utilized due to limitations inherent in the binary nature of the logistic regression model.
- The glm function assumes a binary response variable with values between 0 and 1, representing the two possible outcomes. Unfortunately, this binary approach doesn't allow for the representation of a "Neutral" category, which may exist in a non-binary likability table.
- In cases where likability is measured on a Likert scale with a "Neutral" option, the binary conversion might not fully capture the nuances of the dataset, potentially resulting in skewed predictions.
- The issue with the model arises when introducing additional constraints in model2. While the confusion matrix indicates higher accuracy compared to model1, a concern emerges as the predicted probabilities in model2 are significantly lower.

- This discrepancy in predicted probabilities suggests instability in the model, highlighting that the addition of more constraints has led to a wider and potentially problematic range of predicted values. Consequently, this inconsistency defines model's instability when subjected to increased constraints.

7. Potential improvements to the survey design

- **Power Analysis:** We would conduct power analysis calculation used to estimate the smallest sample size needed for finding the likeability, given a required significance level, statistical power, and effect size.
- **Sample size adjustment:** Increasing the sample size can enhance the statistical power and reliability of the survey results. This exercise will also help us to calculate the probabilistic predictions for the likeability.
- **Standardized Identifiers:** Clean data necessitates the use of standardized identifiers, such as respondent IDs instead of NetIDs. This ensures uniqueness and clarity in participant identification, mitigating confusion and facilitating accurate data analysis.
- **Consistent Categorical Values:** To decrease ambiguity and provide more definite results, it is crucial to standardize language and country names. Using consistent categorical values, achieved through clear instructions and predefined response options, enhances the reliability and interpretability of survey data.

8. Suggestions for additional data or analyses that may be useful for evaluating the likeability of songs

- **Familiarity:** Familiarity with a song or elements can have direct correlation with likability. People often gravitate towards liking the songs that they are familiar with. So having a binary response to familiarity can have a positive influence in determining the likeability of the song.
- **Personal Relevance:** The extent to which the lyrics or theme of the song resonates with the personal experiences, emotions, or beliefs of the listener can influence likability. Having said that, having a field with non-binary responses would help us determine the likeability of the song.
- **User Demographics:** Insights on the user demographics would help us understand the factors that we can use while computing the model to find the likeability of the model.

9. Conclusion

- In this analysis, I explored song likeability comprehensively, using data visualizations to unveil preferences based on genre, country, and language. Assessments of song length, release year, and Taylor Swift's music affinity were accompanied by visual representations. I tackled the challenge of developing a predictive model, acknowledging associated challenges. A critical reflection on survey design led to comments on enhancements, addressing issues like identifier standardization. Proposed additional data and analyses, including lyrics sentiment and cultural influences, offer insights for future research and methodology improvements. This work contributes to informed discussions and guides future analyses succinctly.