# Spotify Final Report

By: Saiem Ilyas, Onell Yago, Ryan Venerao, Gleb Lavrentyev, Hamid Qureshi

## Non-Technical Executive Summary

The goal our team decided to explore the correlation between track features and its popularity and how artist collaborations affects positive attributes such as danceability, energetic or even acousticness. We predicted that different collaborations could introduce different attributes to the music thus affecting its appeal. By analyzing data from a music dataset that included attributes such as danceability, energy, loudness, and popularity, we sought to answer the following questions:

- **Does collaboration between artists affect the popularity of a song?**
- **How do features like energy, danceability, and acousticness vary between collaborative and solo tracks?**
- **What is the correlation between different musical features within albums, and how does popularity fit into this?**

By following the process of data cleaning, EDA, and statistical modeling, the developments of detailed conclusions were achieved. Our results have shown that collaborations are more likely to have a different acoustic profile than solo tracks, specifically in terms of instrumentalness and acousticness, but the effect on popularity was not very clear. We also determined specific musical features including energy and loudness by which we are also able to predict the level of popularity of that song.

## Technical Exposition

### Wrangling and Cleaning Process

Before applying feature selection, we also preprocessed the dataset with regard to some data quality problems like missing values. In order to remove uncertainties in records which were not complete we decided to delete any row with missing data. We also created new variables to aid in our analysis:

- **Collaboration Indicator:** Based on the above two columns, we created one more column called is_collaboration which informs us whether a track is also a collaboration if there were more than one artist. This was done using the condition that if the number of artists (separated by semicolon) is more than one.
- **Popularity Indicator:** Another new column is_build is_highly_popular which has been created by adding flag to songs whose popularity score is greater than 70.

These transformations made it possible that we had all the features needed to study the effects of collaboration, especially on musical features and their popularity.

### Investigative Depth

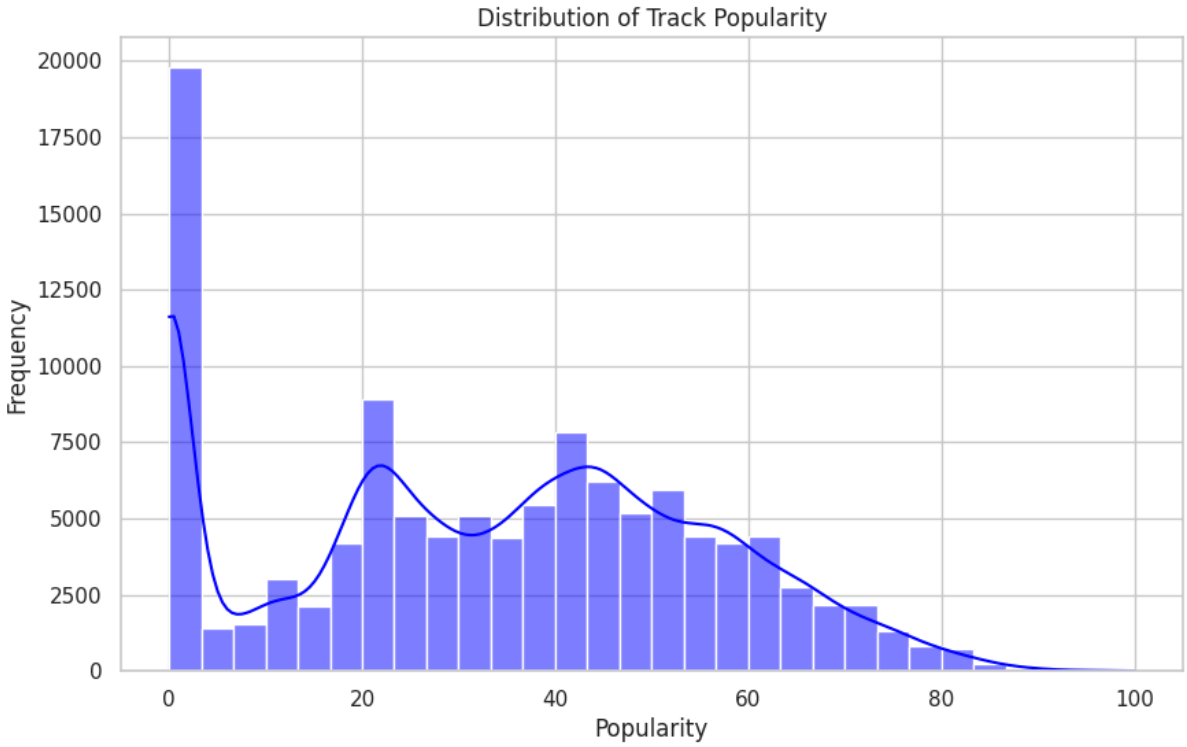
Our exploratory data analysis (EDA) began by examining the distribution of popularity and its relationship with other musical attributes like danceability and energy. Using visualizations like histograms and box plots, we gained insights into the distribution of these features.

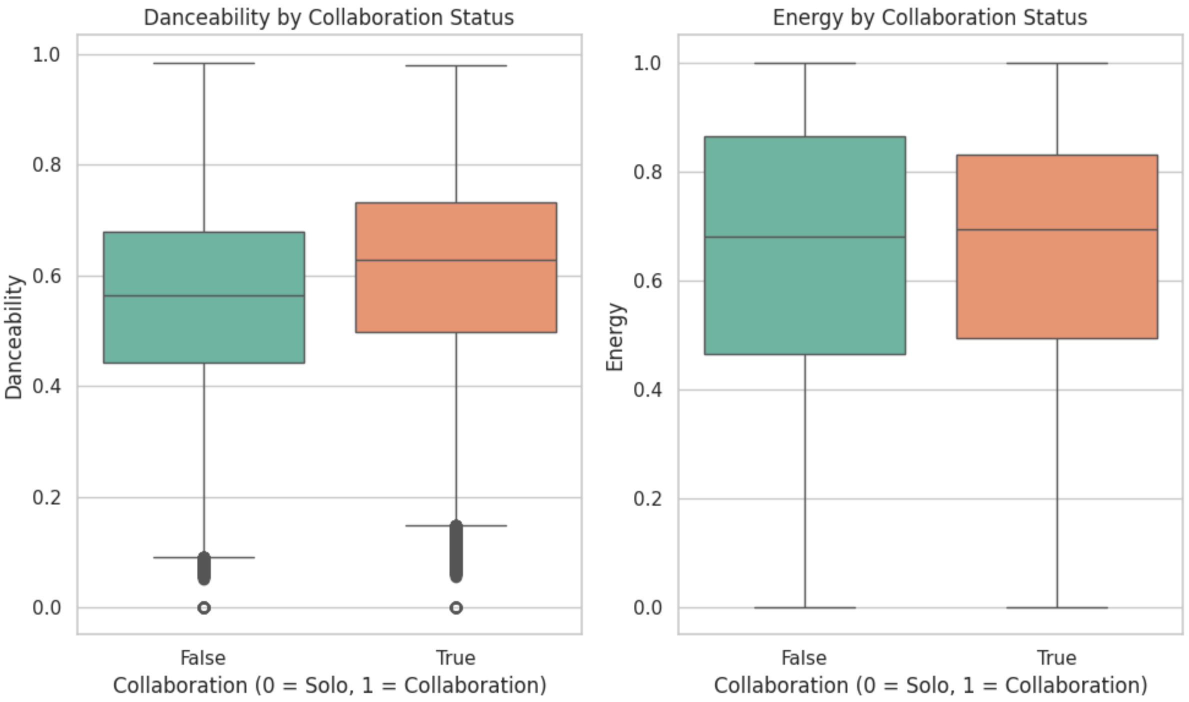
We further examined the relationship between collaboration and musical features. Using box plots, we observed that collaboration seemed to influence acousticness and instrumentalness, with collaborative tracks showing different distributions compared to solo tracks.

Additionally, we performed t-tests to check if there were statistically significant differences between solo and collaborative tracks in terms of musical features such as danceability, energy, and acousticness. The t-tests revealed the following:

- **Danceability**: Highly significant differences (p-value ≈ 0) were found between solo and collaborative tracks, indicating that collaborations tend to have higher danceability.
- **Energy**: No significant difference was found (p-value ≈ 0.1), suggesting that collaboration doesn't significantly impact energy levels.
- **Acousticness and Instrumentalness**: Both features showed highly significant differences (p-value < 0.0001), indicating that collaborative tracks tend to have lower acousticness and instrumentalness compared to solo tracks.

These results provided key insights into the influence of collaboration on various musical features, which later informed our modeling choices.





**Analytical & Modeling Rigor**

We performed feature selection by focusing on features that were most likely to have an impact on a track's popularity, namely danceability, energy, loudness, acousticness, and valence. We then built regression models to predict track popularity based on these features. We used two models for comparison:

1. **Decision Tree Regressor**: A simple but interpretable model to capture non-linear relationships in the data. However, its performance was suboptimal with a **mean squared error** (MSE) of 446.66 and an **R-squared** value of 0.1.
2. **Random Forest Regressor**: An ensemble method that improved predictive performance with an **MSE** of 246.51 and an **R-squared** of 0.5.

The Random Forest model was able to better capture the complexity of the data and provided a more accurate prediction of track popularity.

**Summary of Results**

- **Collaboration**: Collaboration tends to influence certain features of a song (like acousticness and instrumentalness), but its effect on popularity is not clear-cut. While there is a clear pattern in some features, the relationship with popularity is more complex.
- **Popularity and Features**: Our regression models showed that danceability and energy are good indicators of popularity, but the overall accuracy of our models suggests that other factors not captured in our dataset (such as lyrics or marketing strategies) play a significant role in determining a track's popularity.
- **Statistical Tests**: T-tests confirmed that collaboration affects specific musical features, particularly acousticness and instrumentalness, although no significant effect was found for energy.

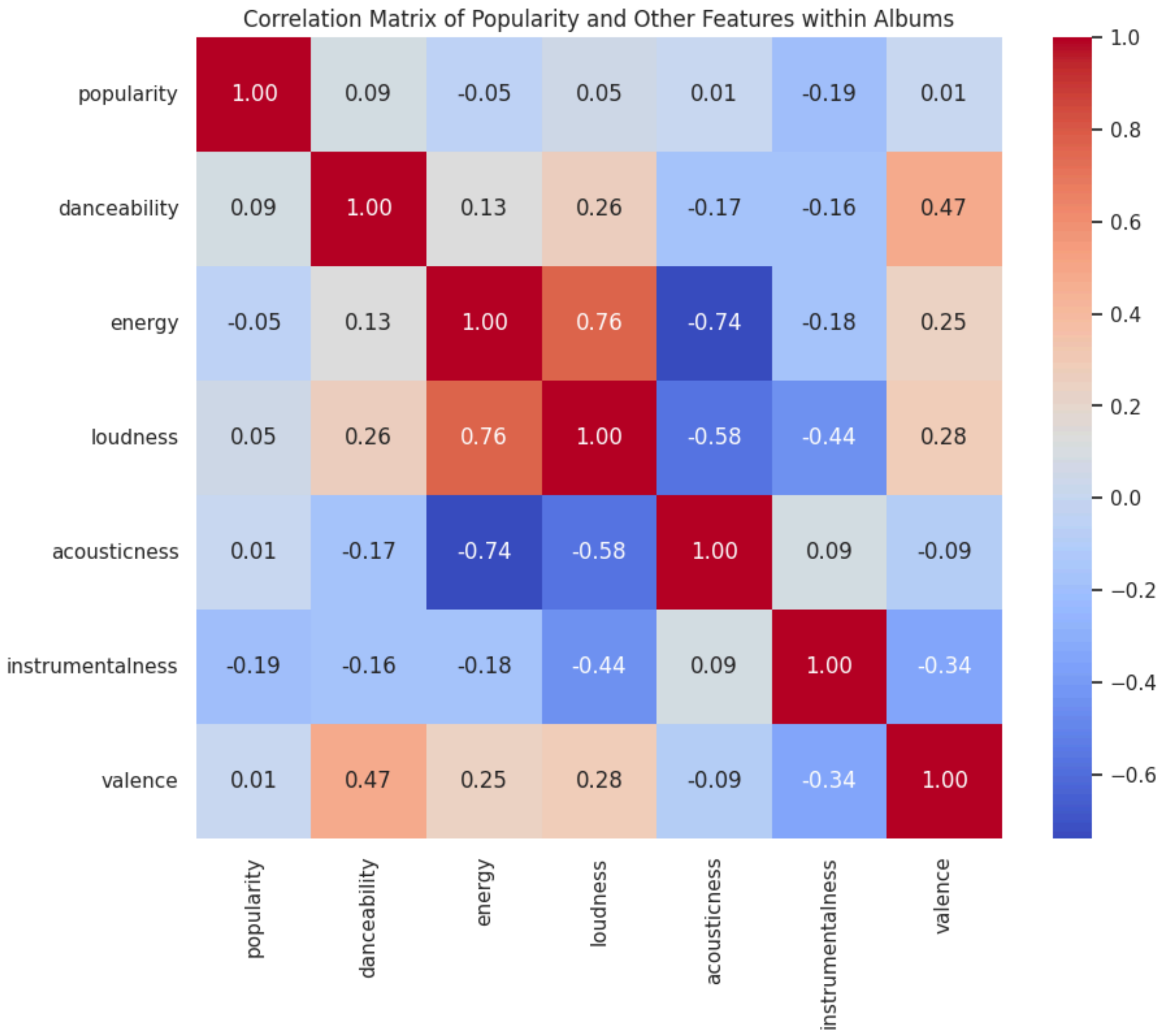


*Figure 1 Correlation Heatmap*

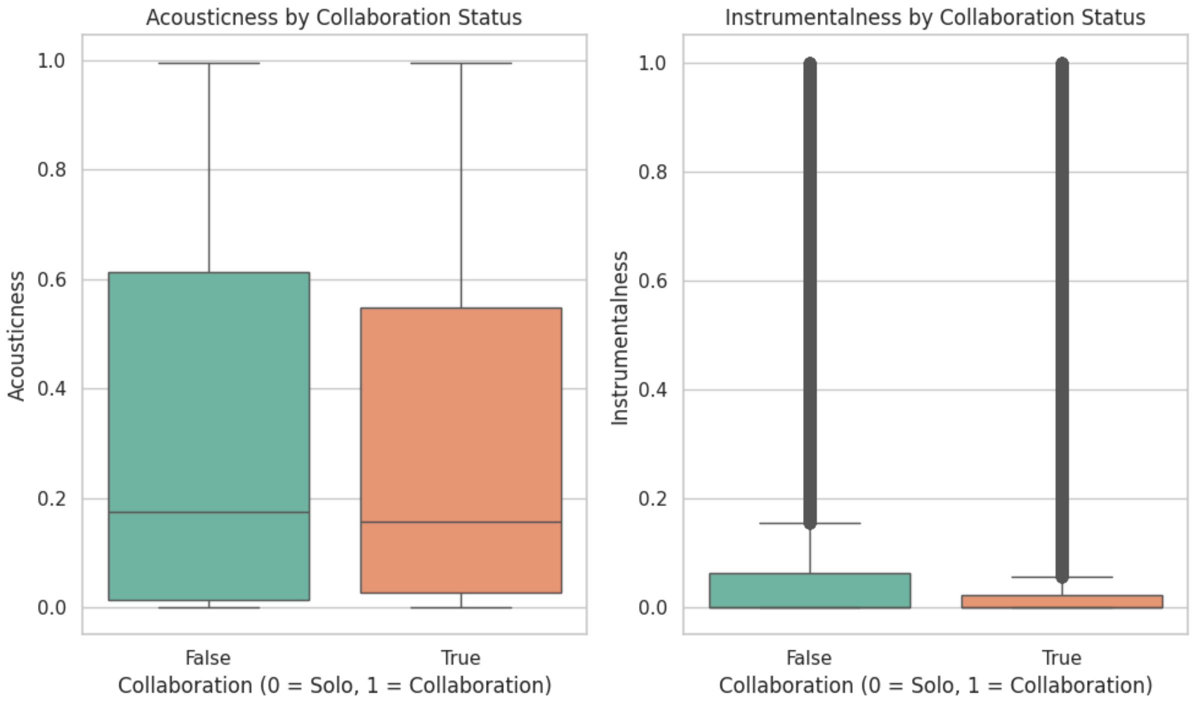


*Figure 2 Impact of Collaboration*

### Recommendations for Further Investigation

- **Include Other Features**: Further analysis could benefit from including additional features, such as song lyrics, genre, or social media engagement metrics, which may help to improve the predictive power of the models.
- **Refine Model Selection**: More advanced machine learning models like Gradient Boosting or Neural Networks could further improve prediction accuracy.
- **Expand Hypothesis Testing**: Additional hypothesis tests, such as analyzing the effect of collaboration across genres, could provide deeper insights into how collaboration influences different types of music.