The statistical question I sought to answer was whether I could predict the popularity of songs based on features like danceability, energy, and genre. I hypothesized that certain musical characteristics would significantly influence a song's popularity. However, as the analysis progressed, it became clear that predicting song popularity is far more complex than initially anticipated due to the nature of the data and the rapidly changing culture surrounding music.

The outcome of my Exploratory Data Analysis (EDA) revealed several critical insights that shaped my understanding of the problem and the necessary steps for refining the model. One of the most important findings was that the dataset lacked sufficient features to accurately predict song popularity. While some of the features I used, such as danceability, energy, and genre, were statistically significant, they did not fully account for the variance in song popularity. This gap highlighted the need for additional variables that could capture more nuances, such as song release date, artist popularity, and marketing factors, which likely influence a track's success.

During the analysis, I observed that the R-squared values for my models were quite low, suggesting that the model was unable to explain a significant portion of the variation in song popularity. The adjusted R-squared values were particularly disappointing, signaling that the current features were not sufficient to provide a good fit for the model. Although the features themselves were statistically significant, the low R-squared values pointed to the need for more complex models and additional data.

One of the major challenges during the analysis was dealing with the distribution of the dependent variable—song popularity. The variable I wanted to predict was not normally distributed, which posed a challenge for using linear regression models. The skewed distribution suggested that I needed to apply a transformation to the data in order to normalize it. By doing so, I would have been able to improve the model’s accuracy and ensure better predictions.

In addition to this, I realized that the relationship between song features and popularity might be non-linear. Songs are influenced by cultural trends that shift rapidly, making them subject to abrupt changes in popularity. Given this, I hypothesize that using a non-linear model (such as K-nearest neighbors (KNN)) might produce better results, as it can handle more complex, non-linear relationships in the data.

The Cumulative Distribution Function (CDF) for song popularity revealed that most songs are not popular. The CDF showed that only a small percentage of songs reach high popularity scores, which aligns with the real-world observation that few songs become hits. This is an important insight because it informs the model’s approach—predicting popularity is inherently difficult due to the skewed nature of the data.

During the analysis, I took steps to reduce multicollinearity, which can negatively impact the quality of a regression model. By addressing this issue, I believe the model is relatively non-biased, although I must acknowledge the potential for omitted variable bias. Without accounting for external factors like cultural context or promotion strategies, the model may still overlook crucial influences on song popularity.

The analysis highlighted the challenges of predicting a subjective and complex outcome like song popularity. While some features were statistically significant, the model’s low R-squared values suggest that additional features and potentially more sophisticated modeling techniques are necessary. I am now considering exploring ensemble models or non-linear algorithms to better capture the complexities of the data. Future work would involve incorporating more features related to culture, marketing, and other external factors to improve the accuracy and robustness of the model.