678 Midterm Project Proposal

Huaijin Xin

**Dataset:**

I get the dataset from Kaggle.com. The data consist Almost 30000 Spotify songs from the Spotify API. Here is the URL for Datacard:

<https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs/data>

**Question**:

The question I aim to investigate is: "What features of a song predict its popularity on Spotify, and how do these features vary across different genres and subgenres?"

**Literature review:**

Matt Devor conducted a capstone project to understand what makes a song popular on Spotify. Using over 116,000 tracks' data, he explored 17 attributes against popularity scores and found the distribution heavily skewed towards lower popularity. His initial linear regression model had a low predictive power (R^2 of 0.08). He then applied undersampling to address the imbalance, which improved model performance, but reduced the significance of the independent variables. Shifting to logistic regression with a popularity cutoff yielded a more accurate model, highlighting danceability as a key predictor of a song's popularity, alongside energy, instrumentalness, and loudness. His findings suggest that while certain song characteristics can predict popularity, there are unquantified factors like artist recognition that also play a crucial role.

Understanding the attributes that contribute to a song's popularity can be incredibly valuable for artists, producers, and record labels. It helps in making informed decisions about which songs to promote and invest in, potentially leading to higher success rates in the charts. Moreover, for streaming services like Spotify, improving the understanding of song popularity can refine their recommendation algorithms, leading to better user experiences by suggesting tracks that listeners are more likely to enjoy.

Matt Devor’s research gives me inspiration of how I could deliver my investigation. I am going to do the EDA first. The data for my research seems good, so I just need to do a few works on data processing. Then I will do the Modeling and Validation which could hopefully help me get the answer to my question. Lastly, I would write up a final report.  
  
**Reference:**  
Matt Devor, *Predicting Spotify Song Popularity*

<https://github.com/MattD82/Predicting-Spotify-Song-Popularity/blob/master/README.md>