**Group 4 - Project Proposal**

Trieu Vo

Abid Khan

Ashutosh Kumar Singh

College of Professional Studies, Northeastern University

ALY 6040: Data Mining

Prof. Joseph M. Reilly

**Introduction**

We decided to use the Spotify dataset. This dataset provides audio information for almost 600,000 Spotify tracks. The data is divided into around 20 columns, each of which describes the track and its characteristics. Some noteworthy and popular musicians' songs from 1922 to 2021 are featured.

Data source: <https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets>

This data covers acoustic characteristics such as Danceability, Energy, Acousticness, Valence (Positivity), and many more. Each feature has been described in full below.

* Index: ID
* Name:
* Title: Name of the Track
* Artist: Name of the Artist
* Year: Release Year of the track
* Energy: The energy of a song - the higher the value, the more energtic. song
* Danceability: The higher the value, the easier it is to dance to this song.
* Loudness: The higher the value, the louder the song.
* Valence: The higher the value, the more positive mood for the song.
* Length: The duration of the song.
* Acoustic: The higher the value the more acoustic the song is.
* Speechiness: The higher the value the more spoken words the song contains
* Popularity: The higher the value the more popular the song is.

**Questions to Investigate**

We intended to be able to answer several questions using this dataset, such as:

* Which genres were more popular from the 1950s through the 2020s?
* Which genre's songs were most likely to be in the Top 2020s?
* Which musicians had a higher chance of having a hit song?
* Which terms are most commonly used in songs?
* What has been the average pace of songs over the years?
* Is there a tendency in which acoustic songs were more popular in the 1960s than they are now?
* Is there a difference in genre preferences between then and now?
* How to predict the popularity of a song based on its statistics information?
* ...

**Machine Learning Models**

The following models will be used to examine the data:

* Linear Regression
* Logistic Regression
* Decision Tree
* Random Forest
* XGBoost
* Neural Networks