GitHub Repository: https://github.com/USC-DSCI-510/final-project-sydneyhu1016

Analyzing the determining factors of indie songs' popularity
by Sydney Hu and Stephanie Shaw

The project objective is to solve the **problem** that many underexposed indie artists, including Sydney, face while releasing their songs to the Spotify platform: **excluding external promotion, what are the in-platform determinants of a song's streaming count?** To answer the question, we plan on comparing two groups of artists under the "Indie genre" – popular and underexposed, determined by the artist's popularity score on Spotify. Typically, an indie artist is defined as someone who writes and produces their own music without a major label. For the purpose of this project and to ensure the reliability of our data, we will define artists using Spotify's "indie" genre label.

To **collect** our data, we plan to utilize Spotify Web API (https://developer.spotify.com/documentation/web-API) and the library Spotipy (https://spotipy.readthedocs.io/en/2.22.1/). After a little more exploratory data stage, we will determine two appropriate ranges of popularity scores. We will sample the top 5 songs from each of the 50 indie artists within these two ranges. In other words, the final dataset will contain data about 500 songs from 100 unique artists.

As part of our **analysis**, we will be conducting the following steps:

1. comparing the characteristics of the two groups of artists (popular and underexposed)
    a. features include the distribution of their popularity scores
2. comparing the audio features of songs from these two groups
    a. audio features include tempo, danceability, energy, duration, acousticness, instrumentalness, etc.
3. comparing other in-platform factors
    a. the completeness of their metadata (images), etc.

For **visualization**, we will create histograms for the popularity scores of artists to provide context for the two sample groups. In addition, we will provide a side-by-side comparison of the distributions for each audio feature's metrics. We will then conduct z-tests to test the hypothesis that the two distributions have similar audio features.