**Intro:** This project represents my final analysis for my social statistics course, where I explored patterns in some of the most successful songs from 1990–2021. The goal was to uncover the shared characteristics of hit songs and investigate whether a formula for success exists.

#### **Data:** The dataset consisted of the top 10 Billboard Hot 100 songs from each year (ranked by weeks at #1). If a year had fewer than 10 #1 hits, we included only the available songs. This resulted in a population of 300 songs. Using R, a random sample of 120 songs was selected for analysis. The study examined factors such as genre, popularity, production, critical reception, and narrative tone.

#### **Instructions:** Access the **"SCRIPT.R"** file for the code used in the analysis.. Open the **"RAWDATA - sheet1.csv"** file for the dataset.

#### **Operationalization of Variables**

#### **Popularity:** Popularity was defined as a song's widespread appeal and operationalized using several measurable factors:

#### Streams: Total Spotify streams, obtained directly from the platform.

#### Album Sales: Certified digital sales from the RIAA. Standard physical sales data were used for 1990s songs without digital metrics.

#### Gender: A categorical variable with values from 1–9, classifying artist(s) by gender identity. Examples include 1 = women, 2 = men, 3 = nonbinary, etc.

#### Fanbase Size: The average number of followers on Spotify, YouTube, and Pandora, calculated by summing followers across platforms and dividing by three.

#### Media Use: Total instances of the song appearing in movies and TV shows, sourced from WhatSong.com. Multiple appearances in a single series counted as one.

#### Genre: Each song was categorized into one or more genres using binary variables (1 = yes, 0 = no) for Pop, Rock, Rap, Country, Hip-Hop, Alternative, Reggae, and R&B.

#### **Critical Acclaim:** This concept was operationalized by tracking a song's recognition in five major music award shows (Grammys, VMAs, MTV Music Awards, Billboard Music Awards, and AMAs). Each show was represented as a categorical variable with three values:

* 0 = Not Nominated
* 1 = Nominated (but didn’t win)
* 2 = Won

**Narrative Tone:** This variable represents the emotional vibe conveyed by a song through its lyrics, melody, and tempo. binary variables were created, each corresponding to a specific tone: sad, happy, fearful, angry, hopeful, romantic, shameful, and desperate (1 = yes, 0 = no).

**Song Production:** Song Production refers to the technical aspects of a song's creation process, like its composition, arrangement, and recording, with features like BPM, time signature, length, key signature, release year, and instrument count being created during the song's production phase. The technical aspects of creating a song were operationalized through the following variables:

* BPM: Beats per minute, measured as a discrete quantitative variable using songbpm.com.
* Time Signature: A categorical variable indicating rhythm, with each time signature assigned a unique number (e.g., 1 = 2/4).
* Length: Total song duration, measured in seconds. Data from Spotify (in minutes) was converted into seconds for consistency.
* Instrument Count: The total number of instruments used, determined by careful listening to each track.
* Key Signature: The starting key of the song, categorized into one of 21 options, we did not account for a minor or major distinction (e.g., F♯, G, A). Changes in key during the song were not considered.
* Release Year: The year the song was released, treated as a discrete quantitative variable ranging from 1990–2021.