# Stat 333 - Final Project Report

Kyan Cox, Kexin Wen, Shivani Potnuru

2025-05-07

## Major vs. Indie: Analyzing Label Impact on Billboard Hot 100 Reappearances

### Introduction

The Billboard Hot 100 is one of the most influential music charts in the U.S., representing the pinnacle of commercial success and cultural relevance in popular music. However, making an initial appearance on the chart is only the beginning; the true challenge lies in whether an artist can return to the chart, demonstrating sustained popularity. This not only reflects the appeal of the artist's work but also highlights the underlying impact of resource allocation and promotional strategies.

Within this context, a compelling question arises: Are artists signed to major record labels more likely to reappear on the Billboard Hot 100 than their independent counterparts? Major labels such as Universal, Sony, and Warner typically possess greater financial resources and promotional infrastructure, which in theory could help sustain an artists visibility. However, with the rise of digital platforms and social media, an increasing number of independent artists have managed to gain widespread recognition through online channels alone.

Using historical data from the Billboard Hot 100, this study investigates that artists signed to major labels do not have a statistically significant chance in reappearing on the chart after their initial breakout—even after controlling for variables such as initial chart position, musical genre, and release timing. In contrast, factors like the debut songs performance and genre play a more decisive role in shaping an artist's continued visibility. These findings can be said to challenge conventional assumptions about the dominance of major-label backing and offer important implications for marketing allocation in the American music industry.

### Methods

To determine whether affiliation with a major record label influences an artist's likelyhood of repeatedly appearing on the Billboard hot 100, we integrated multiple publicly avaliable data sources. Our primary dataset consists of the Billboard Year-End Hot 100 charts spanning 1970-2024, gathered through web scraping Wikipedia (e.g., Billboard Year-End Hot 100 Singles of 1995). Intially, this yielded approximately 5,400 song entries containing rank ("No."), song title, artists, and year.

Given our focus on sustained popularity after initial breakthrough, we condensed the raw file to a single record per artist. The `Artist(s)` field was tokenized on common delimiters (commas, "feat.", "&", "x", etc.), with the first token being treated as the `primary artist`. If multiple songs from a primary artist debuted simultaneously, we selected the song with the highest initial rank, treating subsequent songs as reappearances. After these refinements, our dataset comprised 2,043 unique artist–song combinations. Derived indicators included `is_first_appearance` (indicator) and `reappearance_count` = `total_appearances` - 1.

To accurately categorize artists by label status, we queried each debute track via the Spotify Web API (Spotify Web API) using the spotifyr R package. From Spotify, we collected: (1) the official label name(s), (2) track duration (`duration_sec`), and (3) release date, used only for validation purposes (e.g., identifying reissues/remastered songs). Label strings were standardized through trimming, punctuation removal, and

lower-casing, then cross-referenced against decade-specific major label lists gathered from Wikipedia (e.g., 1950s-1970s: RCA Records, Columbia Records; 1980s-1990s: MCA Records, PolyGram; 2000s: Universal, Sony, Warner) compiled from industry histories, while also considering for sub-labels (e.g., Rhino Atlantic under Warner in the 1980s).This created our binary variable, is_major_label, classifying 475 debute artists as "Major" and 962 as "Indie" DOUBLE CHECK THESE NUMBERS

Recognizing the critical role genre plays in sustained musical popularity, we also decided to annotate each track's genre. We initially matched each song to the Million Song Database's (track_metadata.db) titles and artists using exact and fuzzy matching techniques, retrieving track ID's subsequently linked to one of 15 high-level genres provided in the Tagtraum annotations (msd_tagtraum_cds.cls). Because the Million Song Dataset primary covers songs up to 2010, recent entries frequently lacked genre annotations. To address these gaps, we used the Last.fm API to fetch user generated top-five tags, accepting the first match from the same predefined genre categories that Tagtraum used. After deduplication (preferring Tagtraum over Last.fm when both were available) CHECK THIS only approximately 8% of songs remained "Unknown". The final categorical variable of genre therefore blends MSD labels with curated user tags while preserving a single genre per song.

Our outcome variable, `reapperance_count`, showed considerable right skewness with a median of 0, and max of 15. To address this, we decided to use log_reappearance = log(reappearance_count + 1) as the dependent variable in all models; the +1 helps keep the 0-reappearance "one-hit-wonders" on the scale.

Our primary analysis applied an ordinary least squares regression:

```
## log_reappearance ~ is_major_label + No. + decade + duration_sec +
##     genre
```

Here, `No.` represents initial chart position (1 being the highest), chosen to control for initial commercial success; `genre` adjusts for lister-base loyalty; `duration_sec` indirectly captures streaming and radio-friendliness; and decade serves as a cateogrical fixed effect accountinf for systemic industry changes across time. Additionally, when running linear regression, categorical variables required explicit baseline categories. For the decade variable, we chose the 1970s as the baseline, allowing us to interpret the regression coefficients of subsequent decades relative to this earlier period. Similarly, for genre, we selected 'Unknown' as the reference category, to ensure we were comparing all known genres consistently against an undefined baseline.

All data preparation steps—including scraping, fuzzy matching, label verification, standardization, and genre annotations—were implemented through replicable R scripts detailed in the accompanying .Rmd document.

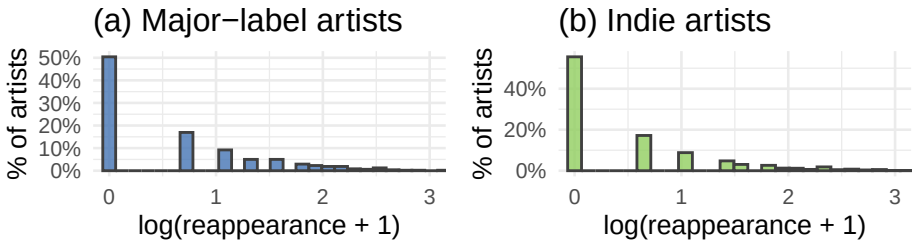


Figure 1: Distribution of log-transformed re-appearance counts for first-time Billboard Hot 100 artists by label type (n = 475 major-label, n = 962 indie). Bars show the percentage of artists in each group; the spike at 0 reflects 'one-hit-wonders'. The two distributions are nearly identical, reinforcing that label affiliation does not predict sustained chart success.