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 whether artists signed to major labels have a statistically significant advantage in reappearing on the chart after their initial breakout—even after controlling for variables such as initial chart position, musical genre, and release timing. Our analysis indicates that affiliation with a major label is associated with only a slight increase in the likelihood of reappearance. In contrast, factors such as the debut song's initial performance and genre exert a more substantial influence on an artist's sustained visibility. These findings challenge conventional assumptions regarding the predominant role of major-label backing and offer important implications for resource allocation and promotional strategies within the American music industry.

Methods

To determine whether artists affiliated with major record labels have a higher likelihood of repeatedly appearing on the Billboard Hot 100, we integrated multiple publicly accessible datasets. Our primary data source consisted of Billboard Year-End Hot 100 charts from 1970 to 2024, gathered through web scraping Wikipedia pages (e.g., Billboard Year-End Hot 100 Singles of 1995). This initial dataset contained roughly 5,400 entries, each including the rank ("No."), song title, artist(s), and year.

Since our research focuses on sustained popularity after an initial breakthrough, we condensed the data to a single record per artist. We tokenized the `Artist(s)` field by common delimiters (e.g., commas, "feat.", "&", "x"), designating the first token as the `primary_artist`. If an artist had multiple debut songs in the same year, we retained only the highest-ranking song as their initial appearance, with subsequent songs classified as reappearances. After these adjustments, our final dataset comprised 2,042 unique artist-song pairs. We derived two key indicators: `is_first_appearance` (a binary indicator marking initial appearances) and `reappearance_count` (defined as the total appearances minus one).

To accurately categorize artists by their record label type ("Major" or "Indie"), we queried each debut track via the Spotify Web API (Spotify Web API) using the spotifyr R package. For each track, Spotify provided: (1) the official label name(s), (2) track duration (`duration_sec`), and (3) release date, which was used for validation (e.g., identifying reissues or remastered tracks). Label names were standardized through lowercasing, punctuation removal, and trimming whitespace. Standardized labels were then cross-referenced with major-label lists compiled from Wikipedia and industry histories for each decade, also considering sub-label affiliations (e.g., Rhino Atlantic under Warner). This process yielded a binary variable, `is_major_label`, classifying 478 debut artists as "Major" and 1,564 as "Indie."

Recognizing the importance of genre in predicting sustained popularity, we annotated each track with a genre label by integrating metadata from the Million Song Dataset (MSD) and the Tagtraum genre annotations. Specifically, we leveraged the msd_tagtraum_cd2.cls file, which maps over 900,000 unique `track_ids` from the MSD to one of 16 genre categories. To match our Billboard dataset to these genre annotations, we first cleaned and normalized both song titles and artist names, and then attempted to join them to the MSD's track metadata database, which contains track_id, title, and artist_name for all tracks in the collection. Matching was conducted in two stages. We first performed exact string joins between our cleaned Billboard entries and MSD metadata. For any records that remained unmatched, we used fuzzy string matching (with a maximum Levenshtein distance of 2) to link the most similar MSD entries based on title and artist name. Once a match was established, we used the track_id to merge in the Tagtraum genre label. Because the MSD largely covers music released before 2010, many recent Billboard entries were not represented in the dataset. To fill these genre gaps, we queried the Last.fm API for each unmatched track, retrieving the top five user-generated tags. We then selected the highest-ranking tag that matched one of Tagtraum's predefined genres. After deduplication, giving priority to MSD labels when available, approximately 8% of debut tracks remained labeled as "Unknown", meaning no reliable match could be found in either dataset.

Our primary outcome variable, `reappearance_count`, exhibited substantial right-skewness, with many artists never reappearing after their initial chart debut. To manage this skewness and effectively model the data, we applied a logarithmic transformation: log_reappearance = log(reappearance_count + 1). This transformation maintained the interpretability of zero reappearances (one-hit wonders).

We employed ordinary least squares (OLS) regression for our analysis, modeling log-transformed reappearances as follows:

```
## 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 listener-base loyalty; duration_sec indirectly captures streaming and radio-friendliness; and decade serves as a categorical fixed effect accounting for systemic industry changes across time. When running linear regression, categorical variables required explicit baseline categories. For the decade variable, we chose the 1970s as the baseline, allowing interpretation of regression coefficients for subsequent decades relative to this earlier period. Similarly, for genre, we selected 'Unknown' as the reference category, ensuring all known genres were consistently compared against those not categorized.

Figure 1 visually represents the distribution of log-transformed reappearance counts, split by label type. We opted for a percentage to account of the large variation in counts between both categories. Both distributions display prominent peaks at zero, indicating many one-hit wonders, and overall similarity between major-label and indie artists. These visual patterns align with our hypothesis that major-label affiliation alone does not significantly predict sustained Billboard Hot 100 success.

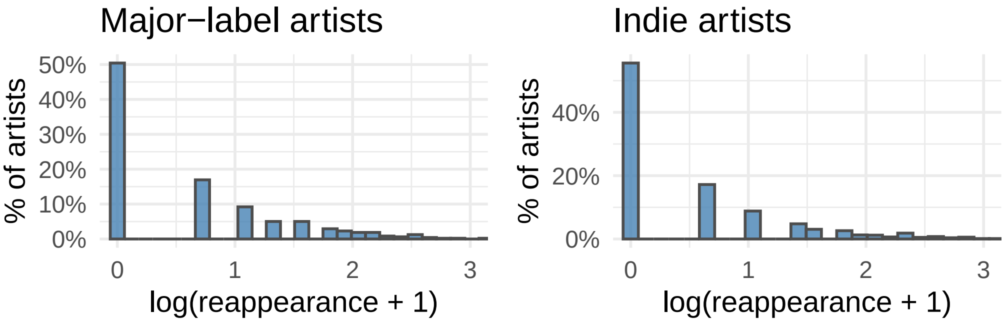

Figure 1: Distribution of log-transformed re-appearance counts for first-time Billboard Hot 100 artists by label type (n = 478 major-label, n = 1,564 indie). The spike at zero shows all the one-hit wonders.

Detailed procedures for data scraping, matching, label classification, and genre annotation are available and replicable via R scripts provided in the accompanying .Rmd file.

Results

We fit a linear model predicting log_reappearance using label affiliation, chart position, decade, genre, and track duration. The overall model was statistically significant (F(22, 2019) = 6.21, $p < 0.001$) but explained only a small portion of the variation in reappearance counts (Adjusted $R^2 \approx 5.3\%$), consistent with the view that many other artist-level or industry-level factors influence sustained popularity beyond those included in this model.

Table 1: Table 1: Coefficients from OLS model predicting log(reappearance + 1).

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.502	0.084	5.953	0.000	0.337	0.668
is_major_labelMajor	0.089	0.041	2.192	0.029	0.009	0.169
No.	-0.004	0.001	-7.427	0.000	-0.006	-0.003
decade1980	0.046	0.053	0.870	0.385	-0.058	0.151
decade1990	0.051	0.053	0.954	0.340	-0.054	0.155
decade2000	0.164	0.056	2.941	0.003	0.055	0.274
decade2010	0.218	0.060	3.637	0.000	0.100	0.335
decade2020	-0.085	0.077	-1.103	0.270	-0.235	0.066
duration_sec	0.000	0.000	1.358	0.174	0.000	0.001
genreBlues	-0.227	0.187	-1.210	0.226	-0.594	0.141
genreCountry	0.309	0.084	3.676	0.000	0.144	0.474
genreElectronic	-0.202	0.115	-1.752	0.080	-0.427	0.024
genreFolk	0.037	0.147	0.251	0.802	-0.251	0.325
genreJazz	-0.011	0.170	-0.066	0.947	-0.345	0.323
genreLatin	0.202	0.245	0.825	0.410	-0.278	0.682
genreMetal	-0.222	0.761	-0.292	0.770	-1.714	1.270
genreNew Age	-0.310	0.347	-0.892	0.372	-0.990	0.371
genrePop	0.153	0.057	2.711	0.007	0.042	0.264
genrePunk	-0.503	0.539	-0.934	0.351	-1.561	0.554
genreRap	0.175	0.077	2.289	0.022	0.025	0.325
genreReggae	0.069	0.191	0.359	0.720	-0.306	0.443
genreRnb	0.266	0.070	3.789	0.000	0.128	0.403

term	estimate	std.error	statistic	p.value	conf.low	conf.high
genreRock	0.239	0.064	3.700	0.000	0.112	0.365

Major-label affiliation had a small but statistically significant effect: artists signed to major labels were predicted to have about 9% higher reappearance counts than indie artists, controlling for all other variables ($\beta \approx 0.089$, $p = 0.029$). However, the strongest predictor in the model was initial chart position. Artists who debuted higher on the chart were significantly more likely to return to it, even after adjusting for genre and decade. Some genre effects were also notable—RnB and Pop showed elevated return rates—but genre was generally a noisier variable due to inconsistent classification. Songs from the 2010s slightly outperformed the 1970s baseline in terms of reappearance, whereas those from the 2020s underperformed, possibly due to recency bias (i.e., not enough time has passed to observe reappearances).

To evaluate model assumptions, we examined four standard diagnostic plots. The Residuals vs. Fitted plot revealed mild heteroscedasticity—residual spread increased at higher fitted values—suggesting non-constant variance, a violation of OLS assumptions. The Normal Q-Q plot showed modest right-tail deviation, indicating that extreme values were slightly heavier-tailed than assumed under normality. The Scale-Location plot echoed the heteroscedasticity issue, with variance spreading more as fitted values increased. Lastly, the Residuals vs. Leverage plot identified a few moderately influential points but no clear outliers with extreme leverage. Together, these diagnostics suggest that while the model provides useful estimates and interpretable coefficients, it may benefit from a transformation of variables or the use of a more flexible modeling approach.

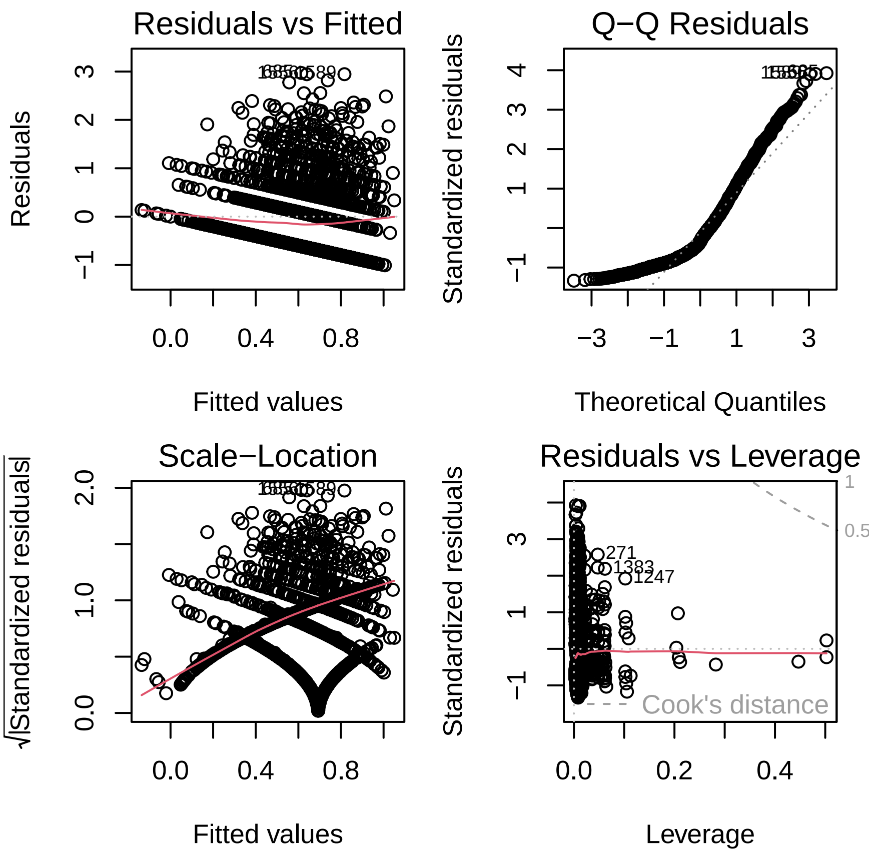

Figure 2: Figure 2: OLS regression diagnostic plots: (1) Residuals vs Fitted, (2) Normal Q-Q, (3) Scale-Location, (4) Residuals vs Leverage.

Conclusion

This study examined a common assumption: that artists signed to major record labels are more likely to reappear on the Billboard Hot 100 following their initial chart entry. Regression analysis reveals that after controlling for initial chart position, release decade, song duration, and genre, affiliation with a major label does exhibit a statistically significant but relatively modest positive association with repeated appearances ($p = 0.0285$). Although statistically significant, the small effect size (approximately 0.09 on a log scale) suggests that while major labels have an advantage, it is not overwhelmingly decisive in predicting long-term chart success. In contrast, other variables play a more pivotal role. Initial chart position emerges as the strongest predictor, showing a significant positive association with reappearance likelihood, suggesting that the initial impact of a song is critical to an artists continued visibility. Additionally, both genre and release decade exhibit meaningful trends: for instance, RnB tracks and songs released in the 2010s are more likely to re-enter the chart, whereas those from the 2020s are less likely to do so.

However, this study also presented several limitations that need to be considered. The linear regression model demonstrated a relatively low degree of explanatory power (adjusted $R^2 \sim 5.3\%$), indicating that important influencing factors were likely omitted. In terms of data quality, genre classification based on Last.fm tags may have introduced considerable noise, especially with the increasing prevalence of “Unknown” tags in recent years, which diminish the reliability of genre as a variable. Meanwhile, the classification of labels as “major” or “independent” is complicated by decades of industry consolidation and ownership changes, making categorization across time difficult. The analysis also focused solely on primary artists, potentially underestimating the role of featured collaborators in contributing to a songs success. Additionally, more recent Billboard entries may not have had sufficient time to exhibit reappearance behavior, which could bias the results in favor of earlier releases. Diagnostic tests further suggested violations of key linear model assumptions, such as non-linearity and heteroscedasticity, indicating that alternative approaches like non-linear modeling or survival analysis may better capture reappearance dynamics. These limitations provide valuable insights for future research, highlighting the need to improve research methods, incorporate more variables, and explore more appropriate modeling approaches

This study carries practical implications for stakeholders in the music industry. Relying solely on label affiliation as a guarantee of sustained success might be misguided. Instead, a songs intrinsic appeal, particularly its initial impact and its compatibility with contemporary cultural and musical trends, plays a more decisive role in an artist’s long-term visibility on the Billboard charts. Future research could refine the classification of record labels, incorporate multidimensional variables such as marketing expenditures, social media engagement, and streaming metrics, and adopt alternative modeling techniques better suited to the complexities of chart performance. Enhancing the precision of genre categorization may also contribute to more robust analytical outcomes. In conclusion, this study suggests that while major record labels hold distinct advantages, continued chart success is a multifaceted phenomenon, where song’s initial impact and its resonance with audience preferences are more statistically significant than label type.

Works Cited

- Billboard Year-End Hot 100 Singles pages. Wikipedia. https://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles

- Spotify Web API. Spotify Developers. https://developer.spotify.com/documentation/web-api/

- Last.fm API. Last.fm. https://www.last.fm/api

- Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011). http://millionsongdataset.com

- Schreiber, H., & Müller, M. (2011). Tagtraum genre annotations. Tagtraum Industries. https://www.tagtraum.com/msd_genre_datasets.html