Taylor Swift Case Study Report

Alex Katopodis, Emi Pollard, Lucas Bulczynski, & Colin Song

April 12, 2024

1 Introduction

The name Taylor Swift is not new, "but this year, something shifted. To discuss her movements felt like discussing politics or the weather – a language spoken so widely it needed no context." The "Taylor effect" phenomenon is observed in many ways, from at least ten college classes devoted to Swift, including one at Harvard, to the Federal Reserve comments on the economic impact of The Eras Tour[5]. Swift is primarily known for her work as a singer-songwriter and one of the most influential artists in contemporary music[4]. Swift has garnered numerous accolades, including fourteen Grammy Awards, among others. Her music covers a variety of genres and has continuously topped the charts. Examining Taylor Swift's discography through a data-driven lens may uncover patterns that contribute to her immense success and reveal insights into the broader music and academic industries.

This report analyzes data from three sources. The `taylor` package available on CRAN provides us with access to data from Taylor Swift's songs, including lyrics and audio characteristics. The data comes from Genius and Spotify API. There are four main datasets in the package, which include songs from Swift's official studio albums, Swift's entire discography, Swift's albums, and The Eras Tour surprise songs[7]. One of the external data sources our group decided to use is from the Billboard Hot 100 chart history for Taylor Swift. This data provides us with information such as song title, peak chart position, and number of weeks the song was on the chart[1]. The other source was YouTube, where we scraped mp3 files for every available song. The mp3 files enable us to perform analysis on the raw audio beyond the features given by the `taylor` package.

There is a variety of information online about Taylor Swift's "most popular" songs ranked by streaming numbers and subjective opinions. One study enlisted quantitative analysis by applying mathematical models to analyze the lyrical diversity in Swift's music over time[3]. Another source enlists machine learning techniques that analyze thematic and linguistic elements in her lyrics[2]. We found little previous research that defines or measures the success of Swift's songs. Our analysis strives to enlist statistical methods to predict the success of Taylor Swift's discography and build a high-quality visualization.

The analysis of this report may provide insights into the musical industry, benefit academic research, add to cultural analysis, and enhance fan engagement. Our work targets the music industry with empirical data on factors that contribute to an artist's success, which may help to inform marketing strategies, talent scouting, and production decisions. This report may also contribute to academic fields such as musicology and data science by modeling success in a creative industry. Another possible aim of this report is to assist cultural analysis and further understanding of the impact of societal trends on music consumption and the evolution of popular music over time. Finally, our analysis may offer fans a deeper understanding of Swift's music and enhance their engagement with her work.

This report has two key objectives: creating a high-quality visualization of Swift's discography and building a model to predict the "success" of her songs. The intention of this analysis is to assess and visualize the success of Taylor Swift's discography. For the purposes of this report, we have defined success as the number of weeks a song has spent on the Billboard Hot 100 chart. We plan to analyze success with a Poisson regression model to understand the significance and effect size of different factors. We will additionally use deep learning to attempt to uncover patterns in the lyric and audio data of her songs and incorporate them into our regression model.

2 Data

In the data-cleaning phase of our report, we prepared the data sourced from the `taylor` package on CRAN, Billboard Hot 100 charts, and associated audio files. This process included preparing the lyrics for text analysis by transforming them into a standardized format and consolidating them into a single string per track. We created three new variables: one corresponding to the album category, delineating the phase of Swift's career during which an album was released; a binary variable indicating whether a track is a Taylor's Version re-recording; and another binary variable signifying whether a track is a vault song, which denotes newly released tracks that accompany Taylor's Version albums. The Billboard Hot 100 chart data (as of 3.27.2024) was cleaned manually and recorded in a CSV file using identical track names to those in the existing dataset. This curated chart was merged with our primary dataset to include information regarding our success metric. We also collected audio data by scraping the mp3 files of all songs in the dataset. We used YouTube's API to retrieve video URLs, and a YouTube to mp3 python library to extract the audio data. The mp3 files were not preprocessed after being downloaded. These modifications equipped us with an informative dataset tailored for a nuanced exploration of Taylor Swift's discography.

We have chosen the number of weeks a song remains on the Billboard Hot 100 chart as our metric for gauging success. Recognized as an industry standard, the Billboard Hot 100 offers a comprehensive view of a song's popularity in the United States by aggregating data from sales, radio play, and online streaming. The duration a song spends on this chart directly reflects consumer listening habits and purchasing behavior, providing significant insights into audience engagement. While the Billboard Hot 100 is a quantifiable indicator of a song's appeal, it is critical to note the significant changes that have been made to chart methodology in recent years. In the mid-2010s, the charts were updated to integrate digital track sales and on-demand audio streams, a change that reflects the shift toward digital music consumption. From 2018 onward, Billboard introduced multi-level streaming tiers, assigning more weight to streams from paid subscription services like Apple Music and Amazon Music compared to those from ad-supported services and free tiers.[6]

These two significant changes span Taylor Swift's entire career and have implications for our selected metric. The first change, integrating digital sales into streaming chart calculations, played to Swift's strengths, as she had strong sales in singles and albums and targeted a younger audience. The more recent change also aligns well with Swift's release strategies. Swift's exclusive releases and re-recording of old albums tap into streaming metrics that favor paid subscriptions. These methodological changes likely enhance the chart performance of artists like Taylor Swift, who have strong fan bases who are willing to pay for their music. Although Swift's career trajectory has aligned well with the changes in Billboard metrics, it is important to recognize that these changes have occurred during her career, and the metric of success used to compare her songs has changed. The music industry is constantly evolving, and for this reason, a metric of success that captures changes is critical.

It is important to note the limitations of our selected metric. A song's billboard success is a reflection of commercial performance and may not capture the cultural and ethical dimensions of consumer choices. Streaming trends may better represent listening behavior as they offer a more immediate and granular view of how consumers interact with music daily. Swift's re-recordings challenge traditional industry models and are an important characteristic of her artistry. Streaming trends may better reflect the popularity of success among fans of these releases. As we were concerned with commercial success as well as song popularity, we settled on Billboard success as it is quantifiable, accessible, and adaptive to industry changes. However, utilizing a different metric of success or combining metrics like streaming trends, sentiment analysis, and Billboard success could be an area for future work.

We list all independent variables of interest and rationale for inclusion below:

- Musical attributes: danceability, energy, speechiness, acousticness, liveness, valence, tempo
  - These audio features are of interest because they may capture nuanced musical attributes associated with successful songs. We are interested in exploring the relationship between audio features and song success.
- Audio Clusters
  - We clustered the songs based on a latent representation of their audio. We include this to see if there are any notable clusters of songs that are more successful than others based on

abstract musical attributes.

- Lyric Clusters
  - In the same manner as the audio, we clustered songs based on a latent representation of the lyrics. We did so to discern if any lyrical clusters were notably more successful than others.
- Album Group
  - The dataset contains 18 possible album designations that songs may belong to. Instead of comparing all 18 albums, we are more interested in what stage of Swift's career a song was written during, as song popularity varies within albums. Albums were grouped into 5 categories: early career, early-middle of her career, later-middle of her career, later career, and miscellaneous. This predictor provided insight into how Swift's evolution during her career has influenced the success of her songs over time. It is important to note that Taylor's Version albums were grouped with the original version of the album, not the release date because the songs were written at that stage during her career. We wanted album grouping to reflect the creative period of the songs rather than their re-release context.
- Is a song part of Taylor's Version album?
  - We created a binary variable indicating whether a track is a Taylor's Version re-recording. This variable was created to account for the re-release context of these albums since the album group variable buckets are based on the inception period. We believe that the release date is still an important factor and should be accounted for.
- Is a song a vault song?
  - We created a binary variable signifying whether a track is a vault song, which denotes newly released tracks that accompany Taylor's Version albums. We wanted to account for vault songs because newly released tracks on Taylor's Version albums may influence song success and popularity in a different way.

2.1 Exploratory Data Analysis

For initial EDA, we plotted variables of interest, such as danceability and energy against our metric of success, weeks on chart, to identify any potential correlations. From this analysis, we only identified weak negative correlations among the columns acousticness and speechiness (see figure 1 and 2). This would suggest that catchier and less acoustic songs do better in terms of weeks on chart than wordier, more acoustic songs. We also looked at the popularity of each of her albums in Figure 3, which we will reference in our discussion as we analyze the popularity of albums across cluster assignment.

In our analysis, we examined the distribution of various predictors across the 15 clusters to identify potential patterns. We plotted the distribution of each predictor within every cluster in order to ascertain whether any of the clusters exhibited discernible patterns or trends that could be indicative of underlying correlations with the predictors.

After examining each plot side by side, we concluded that the audio and lyric clusters formed independently of these variables. This independence implies that the clusters might contain information not captured by the other features, highlighting their potential utility as an external source of data.

3 Methodology

Our first step was processing the lyric and audio data to extract information. We used HuggingFace transformers to encode the unstructured data. We used Facebook's wav2vec2 model for audio data and a version of Open AI's GPT2 fine-tuned on lyric data (including Taylor Swift's) for language data. Both types of data were encoded into 784 dimensional vectors, followed by a UMAP transformation

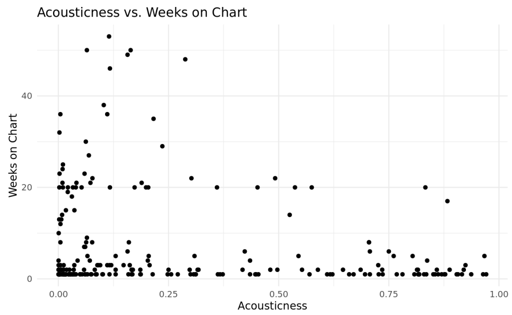

Figure 1: Scatter plot of a song's acousticness versus its weeks on chart.

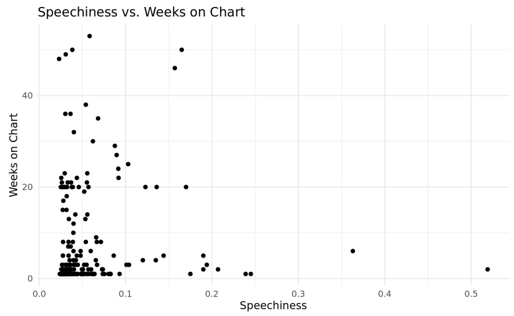

Figure 2: Scatter plot of a song's speechiness versus its weeks on chart.

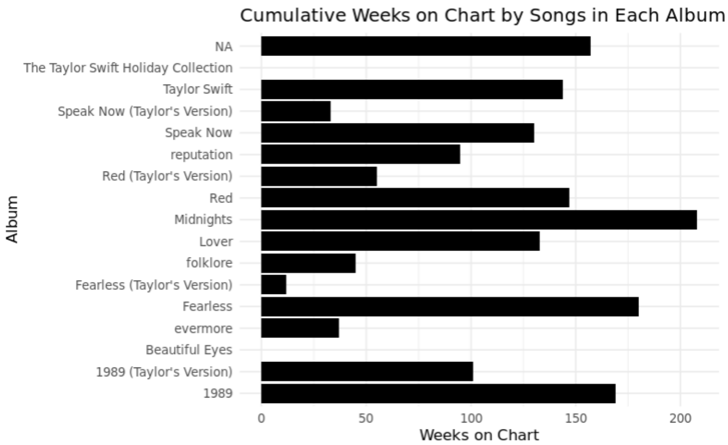

Figure 3: Cumulative number of weeks on chart spent by songs in each respective album. Note, NA represents singles.

to reduce the dimension to 40. We chose 40 for the final dimension size since it was the largest dimension we found to produce roughly even-sized and reasonable clusters. We used sci-kit learn's KMeans class to cluster the UMAP reduced data and chose the value K based on the elbow method with within cluster sum of squares (WCSS) as the metric. We found 15 clusters to be the optimal number in both cases.

Following the clustering, we opted to use a Poisson regression to model the success of Swift's songs. The number of weeks a song remains on the charts is count data, which naturally fits the Poisson distribution's scope. Poisson regression allows us to easily adjust for confounding variables and provides straightforward, interpretable results, which was a fundamental goal of our analysis. The aim of this approach is to directly model the relationship between the number of weeks a song remains on the Billboard Hot 100 chart and a variety of song features to offer a framework for understanding what factors are associated with a song's success. We plotted the deviance residuals to assess model fit. The plot can be seen in the appendix 9.

We predetermined the covariates we wished to implement before modeling the data to eliminate bias due to common variable selection methods. After cleaning our data and consolidating our covariates into a single data frame, we fit the Poisson regression in R using the `glm` function from the `stats` package. The selected model is shown below:

$$
\begin{aligned}
\log(E[\text{weeks on chart}_i]) = \beta_0 &+ \beta_1 \cdot \text{danceability}_i + \beta_2 \cdot \text{energy}_i + \beta_3 \cdot \text{speechiness}_i \\
&+ \beta_4 \cdot \text{acousticness}_i + \beta_5 \cdot \text{liveness}_i + \beta_6 \cdot \text{valence}_i \\
&+ \beta_7 \cdot \text{tempo}_i + \beta_{\text{TV}} \cdot \text{TV}_i + \beta_{\text{vault}} \cdot \text{vault}_i \\
&+ \beta_{\text{album group,early}} \cdot \text{album group}_{\text{early},i} \\
&+ \beta_{\text{album group,middle1}} \cdot \text{album group}_{\text{middle1},i} \\
&+ \beta_{\text{album group,middle2}} \cdot \text{album group}_{\text{middle2},i} \\
&+ \beta_{\text{album group,late}} \cdot \text{album group}_{\text{late},i} \\
&+ \sum_{j=1}^{14} \beta_{\text{audio},j} \cdot \text{audio cluster}_{i,j} \\
&+ \sum_{k=0,k\neq 3}^{14} \beta_{\text{lyric},k} \cdot \text{lyric cluster}_{i,k}
\end{aligned}
$$

Because the audio and lyric clusters were categorical, we had to decide on a baseline cluster. Our objective was to make the baseline clusters the ones whose distributions of weeks on chart most accurately reflected the overall distribution of weeks on chart. We computed the KL divergence of each cluster's distribution of weeks on chart compared to the overall, and selected the clusters with the lowest KL divergences to be the baselines. The baseline audio cluster was cluster 0 and the baseline lyrics cluster was cluster 3.

4 Results

We performed encoded, dimension reduced, and visualized the lyric and audio data.

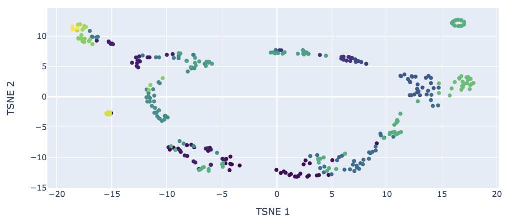

Figure 4: Audio clustering visualized via tSNE. Different colors are different audio embedding clusters.

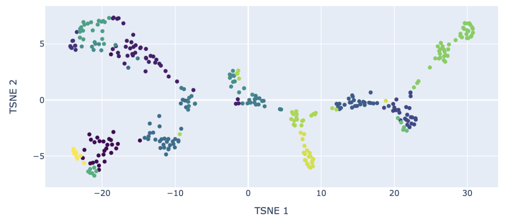

Figure 5: Lyric clustering visualized via tSNE. Different colors are different lyric embedding clusters.

From our Poisson regression model, we saw clear associations between the predictor variables and our chosen success metric, weeks on chart. The results can be found in table 1.

Audio features, including energy, acousticness, liveness, valence, and tempo, were statistically significant at the $\alpha = 0.05$ level and had exponentiated coefficients that were less than one. Every one-unit increase in acoustics is associated with an 83.5% decreased expected number of weeks that the songs stay on the chart, holding all else constant. Speechiness was the only audio feature that was statistically significant with an exponentiated coefficient greater than one, meaning that a one-unit increase in speechiness is associated with a 2658% increase in the expected number of weeks a song stays on the chart by holding all else constant.

All audio clusters were statistically significant at the $\alpha = 0.05$ level except clusters 9, 10, and 11. Most had exponentiated coefficients that were greater than one. The interpretation of cluster 3, for example, is that holding all else constant, cluster 3 is associated with an expected number of weeks on chart 3.883 times or 388.3% higher than the expected number of weeks for songs in cluster 0. Audio clusters 4, 5, 6, and 13 were statistically significant and had exponentiated coefficients less than one. The interpretation of cluster 13, for example, is that holding all else constant, cluster 13 is associated

Table 1: Poisson Regression Model Results

Musical Features			Audio Clusters			Lyric Clusters		
Predictor	Exp(Coef)	P Value	Predictor	Exp(Coef)	P Value	Predictor	Exp(Coef)	P Value
Danceability	0.671	0.219	Audio 1	1.734	<0.001	Lyric 0	2.561	<0.001
Energy	0.501	0.015	Audio 2	1.860	0.002	Lyric 1	2.211	<0.001
Speechiness	26.575	<0.001	Audio 3	3.883	<0.001	Lyric 2	4.270	<0.001
Acousticness	0.165	<0.001	Audio 4	0.555	<0.001	Lyric 4	6.251	<0.001
Liveness	0.472	0.018	Audio 5	0.601	0.015	Lyric 5	4.114	<0.001
Valence	0.715	0.046	Audio 6	0.597	0.011	Lyric 6	1.862	<0.001
Tempo	0.988	<0.001	Audio 7	1.575	<0.001	Lyric 7	1.507	0.033
Later Era	0.982	0.849	Audio 8	1.524	<0.001	Lyric 8	1.838	<0.001
Early-Middle Era	0.825	0.019	Audio 9	0.814	0.161	Lyric 9	3.347	<0.001
Late-Middle Era	0.228	<0.001	Audio 10	0.805	0.260	Lyric 10	0.387	<0.001
Misc. Album Era	0.641	<0.001	Audio 11	1.123	0.677	Lyric 11	19.585	<0.001
TV: True	0.164	<0.001	Audio 12	1.732	<0.001	Lyric 12	3.110	<0.001
Vault: True	2.020	<0.001	Audio 13	0.180	0.018	Lyric 13	3.225	<0.001
			Audio 14	1.835	<0.001	Lyric 14	0.420	<0.001

with an expected number of weeks on chart 82% lower than the expected number of weeks on the chart for songs in cluster 0.

All lyric clusters were statistically significant, and nearly all had exponentiated coefficients greater than one. Lyric clusters 10 and 14 had exponentiated coefficients less than one. The interpretation of the lyric cluster exponentiated coefficients is similar to the audio clusters, but the reference category is cluster 3.

All of the album group categories had exponentiated coefficients of less than one, and all but the group that categorizes albums created later in Swift's career were statistically significant. The reference category was Swift's earliest album group. The interpretation of the album groups is similar to that of the audio and lyric cluster interpretation.

Finally, the expected number of weeks on the chart for songs that were released as a Taylor's Version re-recording is associated with an expected number of weeks on the chart that 84% lower than songs that were not released as a Taylor's Version re-recording. The expected number of weeks on the chart for songs that were vault songs is associated with an expected number of weeks on chart that is approximately 2 times larger than the expected number of weeks on the chart for songs that were not. Both of these indicator variables were found to be statistically significant in the model.

In the context of our Poisson regression, the statistical significance of our predictors means that we have enough evidence to reject the null hypothesis. For the quantitative variables, this means that we reject the null hypothesis that there is no association between the predictor variable and the response variable - the expected number of weeks on the chart - holding all other variables constant. For the categorical variables, this means we reject the null hypothesis that there is no statistically significant difference in the expected weeks on the chart associated with the different levels of the predictor variable, holding all other variables constant.
We used our domain expertise to analyze clusters for patterns and try to uncover potential trends between the unstructured data and song success. We pick the 4 most interesting audio and lyric clusters to further explore and preface this by acknowledging that we do not know why particular clusters were formed. Our patterns certainly could be purely coincidental and have no relationship to clustering.

4.1 Audio Clusters

Audio clusters had interesting patterns and were easier to analyze since we could consider the sound of the song in general.

Audio Cluster 2: These songs all were in the key of G, C, and D major. Interestingly, these keys fall next to each other on the circle of fifths, meaning the keys have at least 10 notes in common and can be tough to discern for the average ear.

Audio Cluster 4: The overwhelming majority of this cluster were ”unpopular” (not on the chart) songs from Fearless, Speak Now, and Red. These are her early albums and suggest that her style has changed over time.

Audio Cluster 5: The songs here all come from her three most recent albums and are also unpopular. In conjuction with cluster 4, there is definitely a suggestion that her new songs sound different from her old songs.

Audio Cluster 6: These songs all have high energy. They're her more ”pop” songs and have a generally positive feel.

4.2 Lyric Clusters

Lyric clusters were far more difficult to analyze. Removing bias of the sound of the song and only focusing on lyrics was challenging. Nonetheless, a few exhibited some more high level lyrical patterns.

Lyric Cluster 4: This cluster consists of entirely Taylor's one off songs that aren't apart of an album. There are some Christmas songs, which would make sense given they have specific lyrical themes. But, there are still non-Christmas songs, and it's suggestive that her albums might all have lyrical themes that her one offs don't match.

Lyric Cluster 10: The songs here were all of her saddest songs off of her three most recent albums, indicating the cluster might have picked up on depressing themes in the lyrics.

Lyric Cluster 11: These songs were all popular, and all had an interesting theme of being particularly upset with a man, and it's possible the cluster picked up on the themes of anger.

Lyric Cluster 13: These songs all came from the album Fearless. Not only this, but they were also all unpopular. This is suggestive of a lyrical pattern in her unpopular songs on the album.

5 Discussion

Our Poisson regression model leads to several interesting findings. Firstly, there seems to be a strong connection between lyrics and song success. Every lyric cluster was statistically significant at the $\alpha = 0.05$ level, and some with drastic coefficients. Take, for example, lyric clusters 2 and 11, with huge exponentiated coefficients of 4.27 and 19.59, respectively. These clusters contain her mega-hits (14 songs altogether, or 4% of her discography). We found with our domain knowledge that they were mostly from the pop genre of her discography. On the other hand, the comparatively unpopular cluster 10 was large and packed with slow ballads. While we cannot tell for certain which lyrical aspects led to the formation of these clusters, it does suggest that lyrical patterns matter to fans.

Audio clusters did not turn out to be nearly as interesting in the context of determining weeks on chart. Some were not statistically significant, and most had exponentiated coefficients close to 1, meaning they did not do a good job of distinguishing weeks on the chart. A few clusters did have larger effect sizes and a somewhat discernible pattern, like 3 and 13. 13 was exclusively Fearless songs that did not make the charts, and 3 had lots of successful Red and eponymous songs. Most of the other clusters were uninteresting. We hypothesize that this could be due to our choice of encoder. We used Facebook's audio model wav2vec2, which is trained largely on speech data. Musically trained encoders are nearly impossible to come by due to copyright laws, but access to one would have likely made for more interesting clusters and could be addressed in future work.

Other, more concrete audio features offered a bit more value. Her songs, on average, became much more popular as they got more speechy and less popular as they became more acoustic. These are more challenging to directly interpret since they are bounded between 0 and 1, but the general direction offers insight into what audio features might be associated with popularity in Swift's songs.

Perhaps the most surprising results are the exponentiated coefficients for the Taylor's Version Song and Vault Song indicator variables. On average, holding all else constant, vault songs do about twice as well as non-vault. Taylor's Versions, on the other hand, do substantially *worse*, with Taylor's Versions only being approximately 16% as successful as non-Taylor's Version songs. We did not expect this to be the case given the hype surrounding her re-recordings. We explored this further by taking a more in-depth look at Red (Taylor's Version).

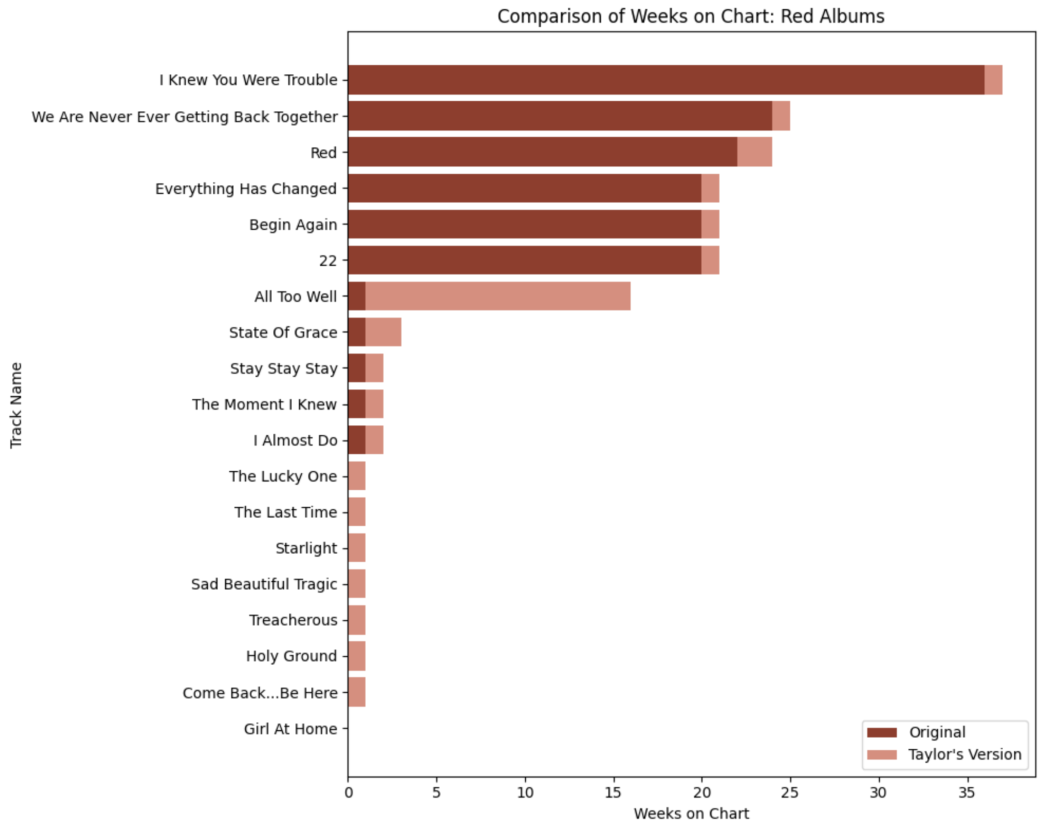

Figure 6: Shown is the success of all songs featured on Red and Red (Taylor's Version). All Too Well is a Taylor's Version enigma, whose re-recording did substantially better than the initial release.

From figure 6, we see that almost every re-recording makes the billboard hot 100 the first week, but only rarely do they stay longer than this. On the other hand, the original versions and vault songs sometimes stay on the charts for much longer. We think this is indicative of Taylor's fans being supportive. They binge her re-recordings on release, but they ultimately want to hear new music. This trend persists for all other Taylor's Version albums.

Our analysis provided insight into the associations between the success, defined by weeks on the Billboard Hot 100 chart, and a variety of song features including audio and lyric clusters. The model revealed the association between song success and the selected features. The results of this analysis provide valuable insight into Swift's discography that have the potential to be applied to the broader music industry for other artist's discography or academic fields interested in exploring the music industry. Additionally, these results may be used to create interesting materials to enhance fan engagement, as is done with the visualization submission of our report.

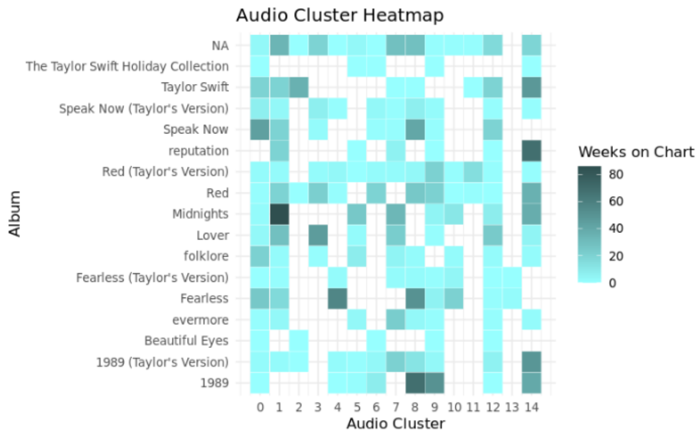

Figure 7: The distribution of albums across audio cluster assignment appears slightly more spread out and has fewer buckets with significant cumulative weeks on chart. Note, NA represents singles.

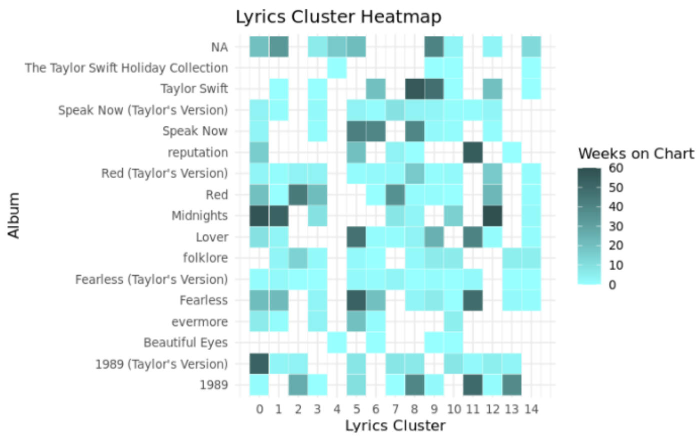

Figure 8: The distribution of albums across lyrics cluster assignment appears slightly less spread out and has more buckets with significant cumulative weeks on chart. Note, NA represents singles.

References

[1] Billboard. Taylor swift chart history. https://www.billboard.com/artist/taylor-swift/chart-history/, 2024. Accessed on 2024-03-24.

[2] Ian Freed. Taylor swift lyrics and machine learning. https://www.codecademy.com/resources/blog/taylor-swift-lyrics-machine-learning/, 2018. Accessed on 2024-03-24.

[3] Nathan Garland. A mathematical exploration of taylor swift's music. https://enlighten.griffith.edu.au/a-mathematical-exploration-of-taylor-swifts-music/, 2024. Accessed on 2024-03-24.

[4] Information Architects of Encyclopaedia Britannica. Taylor swift. Encyclopedia Britannica, 2024. Retrieved from https://www.britannica.com/facts/Taylor-Swift.

[5] S. Lansky. Person of the year 2023: Taylor swift. *Time*, 2023. https://time.com/6342806/person-of-the-year-2023-taylor-swift/.

[6] The Daily Rind. Billboard streaming charts methodology, 2024.

[7] W. J. Thompson. Lyrics and song data for taylor swift's discography, 2023. [R package taylor version 3.0.0].

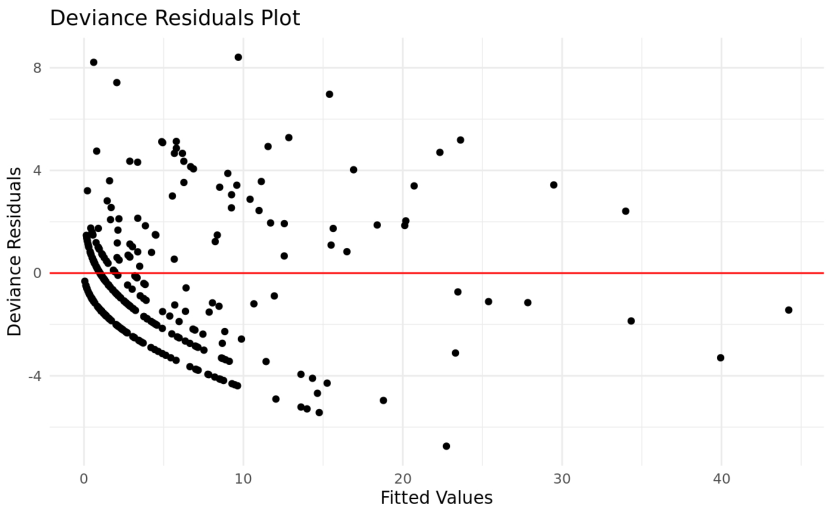

Figure 9: Deviance Residuals Plot

6 Appendix

6.1 C: GitHub Repository

Please follow the link to see our group's GitHub repository (https://github.com/alexkato29/sta440-case2).