Project Report

How to Win a Grammy

Team members :

AKBI Hiba
Dwi Prima Handayani Putri

May 15th, 2025

A report submitted as partial fulfillment of the "DALAS" course

Abstract

Grammy Awards are one of the most prestigious music awards, where singers, songwriters, and producers compete for the best title and recognition of the year. Beyond its significance musically, the Grammys also garners huge mediatic attention. In light of the importance of such an event, as well as the subjectivity of the voting process, utilizing data-driven techniques can help uncover patterns, possible biases, and trends. In this work, we explore the question "How to Win A Grammy" with three objectives: First, to understand the features that differentiate the status of a song. Second, uncover the Grammy trends and how they compare to Billboard. Third, develop a model that is able to accurately and reliably predict winners from nominees. Some of our results include that the longevity of a song on the charts and its debut rank are important, but not strict requirements. Winning rap songs tend to diverge in their topics from the trends of average rap songs. The US dominates the Grammys as well as the charts; English songs also dominated both the Grammys and the charts. And a strong musical infrastructure for US states is strongly correlated with producing more winners and nominees. Our trend analysis shows that the music industry is getting more vulgar over time, and more negative songs are being favored. Also, it seems more females are getting nominated with time. Finally, our model achieved excellent performance and was accurately and reliably able to distinguish between classes.

Contents

1 Introduction

1.1 Background

The 'Grammy Awards' stands for artists as the peak of achievement in the music industry, a coveted recognition that distinguishes itself even among other prominent awards like the MVAs and the American Music Awards. Undoubtedly, the Grammy Awards are synonymous of prestige to both nominees and winners. Since its inception in 1959, **the Grammys have served as a platform for singers, songwriters, producers, and sound engineers, to showcase their musical genius and mastery through competing against their peers for the most esteemed titles**. Over the years, the Grammy Awards have expanded significantly, recognizing excellence in more than 90 diverse categories that represent multiple aspects of the music industry, musical genres, and music careers.

Beyond its significance and recognition for artists, the Grammy Awards garner substantial mediatic attention. Each year, the ceremony records millions of viewers and millions of dollars of profit. For example, the 65th Annual Grammy Awards in 2023 drew an audience of 12.4 million viewers in the United States alone. **The prestige of the Grammy Awards draws fans and industry experts alike to eagerly anticipate the ceremony and its outcomes, often sparking debate and polarized opinions over the outcome of the ceremony.** In 2025, the 67th Grammy ceremony recorded an unprecedented social media engagement, with over 102 million interactions on social media platforms, including tweets, reposts, and hashtags. This undoubtedly solidifies the Grammys Awards as one of the most significant musical events of the year, both for its musical significance and its media impact.

1.2 Problematique

In the cermenony's process, the Recording Academy is responsible for selecting nominees from all eligible songs recorded that year. The same committee of selection then elects a winner in each category. **And while songs must adhere to several eligibility criterias**, such as the recording period, specific formats, and category-specific requirements, **the ultimate decision on nominations and wins remains at the discretion of the voters.**

Understanding which features of a song or an artist make it to the Grammy nominations and wins is significant because it gives artists insights into what makes their songs competitive. For instance, analyzing the tempo, the language, or the vocabulary of the lyrics of nominated and winning songs can reveal which patterns capture the attention of voters.

Additionally, understanding how Grammy selections align with or diverge from broader music industry trends provides interesting insights into how such awards distinguish a certain level of artistic excellence from popular mainstream songs.

Finally, and while we do not doubt the fairness and ethics of voting members, it is worth noting that subjective processes are subject to biases. Thus, a data-oriented analysis can help uncover possible biases and inconsistencies in various aspects, whether demographic, gender, or other.

Thus, the inherent subjectivity of the Grammy voting process is the driving factor to employ systemic and statistical methods to first, gain a better understanding of the factors that influence the outcome, and second, to uncover potential biases and trends within the music industry.

1.3 Our Work

In this work, we aim to answer the question "How to Win a Grammy?". Our purpose is as follows:

- First, identify the features and characteristics of a song and artist that can lead to its nomination or win.
- Second, gain a better understanding of the Grammy Awards trends and analyze how they compare to trends in the music industry.
- Third, develop a model capable of accurately predicting Grammy nominees and identifying winners among the nominees.

Our work follows the pipeline as shown in Figure 1 below. Specifically:

1. **Data Acquisition:** First, we scraped data from multiple sources and used APIs to acquire remaining features. We discussed in Section 2 the different challenges encountered in data scrapping, as well as the reasoning behind the choice of our sources.

2. **Data Preprocessing:** Here, we cleaned our data from missing values and inconsistent entries. We also identified primary and foreign keys in our tables to merge data from different sources. Finally, we create additional features that will be relevant in our analysis, specifically for lyrics. We discuss this in Section 3.

3. **Exploratory Data Analysis (EDA):** We studied the relationships between variables through graphical analysis and statistical tests. Discussed in Section 4.

4. **Prediction Modeling:** Here, we aim to predict winners from a pool of nominees by applying and comparing different models. We use models adapted to our classification task, such as Linear SVMs, Random Forest, and XGBoost, and applying various methods for mitigating class imbalance. We discuss our models and their performances in Section 5.

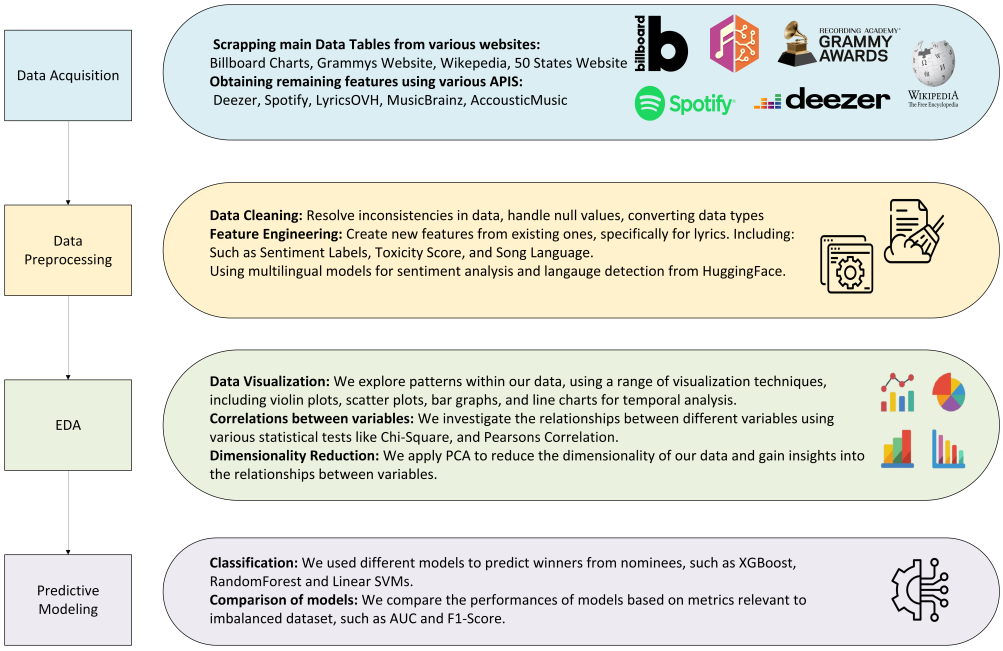


Figure 1: Pipeline of our Data Science Process

2 Data Acquisition

2.1 Data Scraping and APIs

In this section, we enumerate the different websites and API used to acquire our data. For this we have used 4 websites to scrape data tables, and 6 APIs to scrap additional features. **We obtained all our datasets by either scraping or using APIs, rather than relying on readily available CSV files from the internet.** The table below summarizes the different sources and API used to acquire our data, as well as their URLs.

Then, we explain the reasoning behind our choices for choosing each platform to acquire specific features, and the different challenges imposed, leading to these choices.

Source	Description	URL
Billboard	Official Billboard website. Provides charting data of songs, including weekly ranks.	Billboard Charts
Grammy Awards	Official Grammy Awards Website. Provides data on nominees and winners for every category.	Grammy Website
Wikipedia	Provides data related to artists, such as nationality and gender.	Wikipedia website
50 States of Music	Provides data related to the music industry per US State.	50 States Website

Table 1: Websites for Data Scrapping

Source	Description	URL
Spotify	Provides some song features like song duration.	Spotify for Developpers
Genius	Provides some song features like song release date.	Genius API
Music Brainz	Provides MBDI (identifiers) for songs, used for MusicBrainz.	Music Brainz
Accoustic Brainz	Provide song features like tempo and loudness.	Accoustic Brainz
Deezer	Provides some song features like song genre.	Deezer for Developpers
Lyrics OVH	Provides lyrics for songs.	Lyrics OVH Website

Table 2: API for Other Features Extraction

2.2 Pipeline for Data Acquisition

In this section, we explain the steps undertaken to acquire our data, and issues faced, and our solution to handle them.

2.2.1 MusicBrainz

We needed song features from the AcousticBrainz API, which requires a MusicBrainz ID (MBID) for each song, obtainable from MusicBrainz. However, getting the MBID for each song presented several problems. We often failed to match artist and title due to MusicBrainz search limits. Also, **many MBIDs lacked song features**, only offering alternative versions. We also faced issues with special characters in titles and slow retrieval due to many API requests.

To fix these issues, we used a better search approach with standardized artist and title names, and **only accepted MBIDs that had song features (prioritizing them). If an exact match failed, we would accept alternative versions or any recording with valid features**. This helped us get MBIDs more accurately and quickly, and made sure all matched songs had real sound data.

2.2.2 Deezer

Unfortunatelly, low-level of Acoustic Brainz song features does not provide the genre of the song. So we need to retrieve it from another website, and for that, we are using Deezer to complete it. In general, we implemented a pipeline that queries the Deezer public API based on each song's title and main artist. Deezer typically assigns one or more genre tags at the album level rather than per individual track.From the album metadata, the first genre name is selected and assigned to the corresponding song in the dataset. If no genre is available, the value is 'None'.

3 Data Preprocessing

3.1 Data Cleaning

Since our data was scraped, it required significant cleaning on multiple levels. In this section, we detail the techniques used for data cleaning as well as the different challenges faced when handling the data.

3.1.1 Common Data Cleaning Techniques

The following steps were included:

- **Cleaning Data Inconsistencies:** Specifically, inconsistencies were in the Countries, where many variations of the same name exist (e.g., "USA", "US", "US.", and "United States").
- **Handling Null Values: Given the limited data and multiple features, rows with single null values were retained to preserve information. For the exploratory data analysis of specific features, rows were discarded as needed.** For predictive modeling, null values are handled accordingly (see Section 5).
- **Fixing Errors:** Specifically, errors in Birthdates were corrected.
- **Lyrics Preprocessing:** For lyrics, we paid attention to using appropriate stop words. We employed a multilingual stop word list and included domain-specific stop words such as singing noises(words like "mmm", "yeah", and "aaah"). Finally, we applied lemmatization.

3.1.2 Censored Words

A challenge we faced when attempting to find songs via APIs for feature extraction was **the presence of censored words in their titles.** Because of the asterisks in censored words, APIs often failed to fetch the corresponding songs, as they often required near-exact term matching.

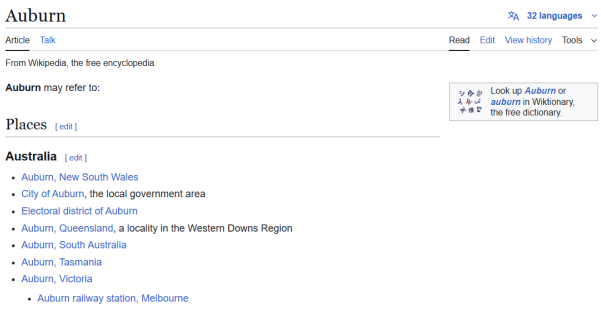
The challenge of this task stemmed from two problems:
First, we could not simply discard the censoring with a simple "delete all special characters" preprocessing, as deleting it would result in an incorrect word, and thus, the song would still not be found.
Second, there exist many variations of the censoring for the same word. For example ("f*ck", "f***", "f**k", "f**kin", "f****g", "fuck**g" ...etc).

To resolve this issue, we created a **dictionary that maps censored words to their uncensored version, then we applied regex to handle differences in variants** like plural and singular forms.

This preprocessing was also reused for lyrics preprocessing, in order to accurately calculate the vulgarity ratio of songs.

3.1.3 Ambiguous Artist Names

To scrape artists' information like gender, country of origin, and birthdate, we initially searched for them by name on Wikipedia. However, name ambiguities frequently prevented Wikipedia from returning the correct URL, as illustrated in Figure 2a. To address this, we iteratively searched for variations of the name. For artist X, this involved looking up terms such as 'X (singer)', 'X (band)', and 'X (musician)', which proved effective in resolving the issue.



(a) Example of Ambiguity for Singer "Auburn"

```
1277 Auburn
try for Auburn
https://en.wikipedia.org/wiki/Auburn
try for Auburn (singer)
https://en.wikipedia.org/wiki/Auburn_(singer)
**********
```

(b) Our code iterating to resolve ambiguity

3.2 Feature Engineering

For our analysis, we required certain features that could not be found readily available in the websites and API. In this section, we explain these features and the methods employed to calculate these new features.

- **Lyrics Language:** To determine the language of a song, we utilized a language detection library from the `langdetect` package.
- **Lyrics Sentiment:** To determine the sentiment of a song, we employed the multilingual sentiment classification model 'tabularisai/multilingual-sentiment-analysis' from Hugging Face. This model provided five labels: [Very Positive, Positive, Neutral, Negative, Very Negative], along with confidence scores for each prediction. For our analysis, we first grouped these categories into three main ones: [Positive, Neutral, Negative], and then applied the confidence scores as weights to the data points.
- **Lyrics Vulgarity:** To measure the vulgarity of a song's lyrics, we counted the number of vulgar words it contained and divided this by the total length of the lyrics. Prior to this, we preprocessed the text to uncensor vulgar words. Our profanity list was extracted from the Bad Words Dataset from Carnegie Mellon University.

Additionally, we faced the challenge that **artists' gender information was not provided as a standalone field on the Wikipedia Infoboxes**. To handle this, we extracted gender from the "External Links InfoBox" section of Wikipedia pages by identifying terms such as "woman", "female", "male", and "man" within the links.

Figure 3: Infox on Wikepedia does not contain Gender

Figure 4: External Links include keywords that hint to gender

3.3 Summary Table of Datasets and Features

In this section, we describe a **summary of our data tables**, their most important features, alongside their types, their description, their sources, and their values.

Table/Feature	Description	Type	Source	Nb Unique Values	Missing Values	% Missing Values
Artist Data: 2305 rows x 5 columns						
Artist ID	Unique identifier of an artist	Categorical	Created	2305	0	0
Country	Country of origin of the artist	Categorical	Scrapped	56	242	10.50%
Gender	Gender of the artist	Categorical	Scrapped	2	873	37.87%
Birthdate	Birthdate of the artist in yy-mm-dd format	Categorical	Scrapped	1443	873	37.87%
State ID	Unique ID for the U.S state	Categorical	Scrapped	2305	764	6.65%

Table 3: Artist Data Table

Table/Feature	Description	Type	Source	Nb Unique Values	Missing Values	% Missing Values
Grammys Data: 413 rows × 6 columns						
Grammy	ith Grammy Ceremony	Numerical	Scrapped	24	0	0
Year	Year of the grammy ceremony	Numerical	Scrapped	25	0	0
Artist ID	Unique identifier of an artist	Categorical	Created	2305	0	0
Song ID	Unique identifier of the song	Categorical	Scrapped	56	242	10.50%
Category	Category of the award	Categorical	Scrapped	4	0	0%
Win	Status of the song (1:winner, 0: nominated)	Categorical	Scrapped	2	0	0%

Table 4: Grammys Data Table

Table/Feature	Description	Type	Source	Nb Unique Values	Missing Values	% Missing Values
Hot100 Data: 135800 rows × 6 columns						
Artist ID	Unique identifier of the artist	Categorical	Created	2305	0	0
Song ID	Unique identifier of the song	Categorical	Scrapped	56	242	10.50%
Charting Week	The specific charting week for the song rankings	Categorical	Scrapped	2	873	37.87%
Last Week	The song's position on the chart during the previous week	Numerical	Scrapped	1443	0	0%
Peek Pos	The highest position this song has reached on the chart to date	Numerical	Scrapped	100	0	0%
Rank	Rank of the song in the current week	Numerical	Scrapped	100	0	0%

Table 5: Billboard Hot100 Data Table

Table/Feature	Description	Type	Source	Nb Unique Values	Missing Values	% Missing Values
US States Data: 51 rows x 7 columns						
State ID	Unique identifier to represent a U.S. state	Categorical	Created	51	0	0
State	U.S. state name	Categorical	Created	51	0	0
Budget	The Music Industry's Contribution to America's GDP	Numerical	Scraped	44	0	0
Jobs	Number of jobs supported by the music industry in the state	Numerical	Scraped	51	0	0
Music Establishments	Number of music-related businesses or organizations (e.g., studios, labels)	Numerical	Scraped	51	0	0
Events and Festivals	Number of music events and festivals held annually in the state	Numerical	Scraped	41	0	0
Music Instituitions	Number of educational or cultural institutions focused on music	Numerical	Scraped	17	0	0

Table 6: US States Table

Table/Feature	Description	Type	Source	Nb Unique Values	Missing Values	% Missing Values
Songs Data: 11.489 rows x 15 columns						
Song ID	Unique identifier of a song	Categorical	Created	11489	0	0
ArtistID	Unique identifier of the singer	Categorical	Created	2267	0	0
Title	Title of the song	Categorical	Scrapped	11489	0	0
Artist	The main singer of the song and the featured artist(s)	Categorical	Scrapped	5125	0	0
Main Artist	The main singer of the song	Categorical	Scrapped	2313	0	0
Feat	The featured artist(s)	Categorical	Scrapped	1900	7857	68.39%
Song Writers/Credits	People involved in the process of writing and producing songs	Categorical	Scrapped	9675	40	0.35%
Producer	Producer of the song	Categorical	Scrapped	5904	40	0.35%
Label	Song recording company	Categorical	Scrapped	2182	50	0.44%
Album	Album of the song	Categorical	Scrapped	5664	167	1.45%
Duration	Duration of the song in seconds	Numerical	Scrapped	8784	167	1.45%
Release Date	Release date of the song	Categorical	Scrapped	3916	627	5.46%
Genre	Genre of the song	Categorical	Scrapped	26	103	15.43%
Tempo	Tempo of the song	Numerical	Scrapped	10115	1114	9.70%
Lyrics	Song's lyrics	Categorical	Scrapped	9655	1773	15.43%
Song Language	Language of the song	Categorical	Calculated	100	0	0%
Sentiment Label	Sentiment that dominates the song lyrics	Categorical	Calculated	100	0	0%
Vulgarity Score	Ratio of vulgar words in the song lyrics	Numerical	Calculated	100	0	0%

Table 7: Songs Data Table

4 Exploratory Data Analysis

4.1 Song Features

First, we explored song-related features to identify any significant differences between Grammy songs and non-Grammy songs, as well as between nominees and winners.

4.1.1 Tempo

Tempo is the speed or pace of a song. One of the units of tempo is BPM. Here are several analyses according to the song's tempo.

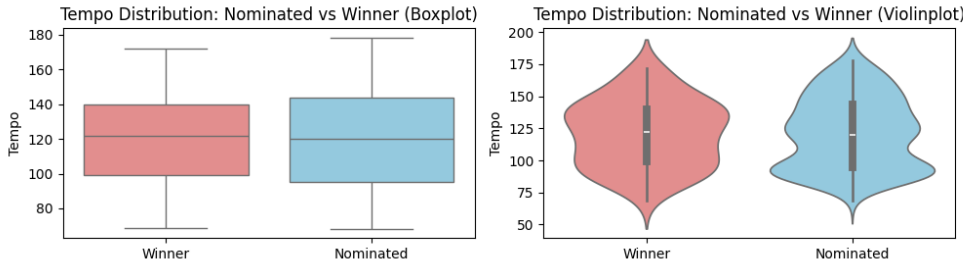


Figure 5: Distribution of Tempo in Grammy Songs (2000–2024)

In Figure 5, we can see that the winners tend to have a more homogeneous tempo (130-140 BPM). While nominated songs seem to be slightly more spread out. Now we are interested to looking at Tempo Distribution in Grammy vs Non-Grammy Songs.

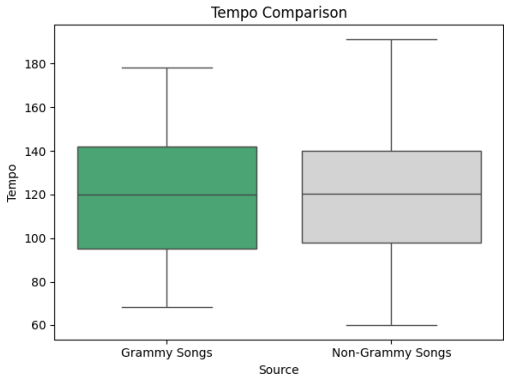


Figure 6: Distribution of Tempo in Grammy Songs and Non-Grammys Song(2000–2024)

In Figure 6, we can see that the tempo ranges on Grammy and Non-Grammy songs are very similar, around 60–190 BPM. Over the years, compared to non-Grammy songs, Grammy songs show much greater year-to-year variation in average tempo. While non-Grammy songs maintain a relatively steady tempo around 117–124 BPM.

4.1.2 Loudness

Loudness is one of the song attributes that determines the intensity of auditory sensation produced. Here are several analyses of Grammys loudness.

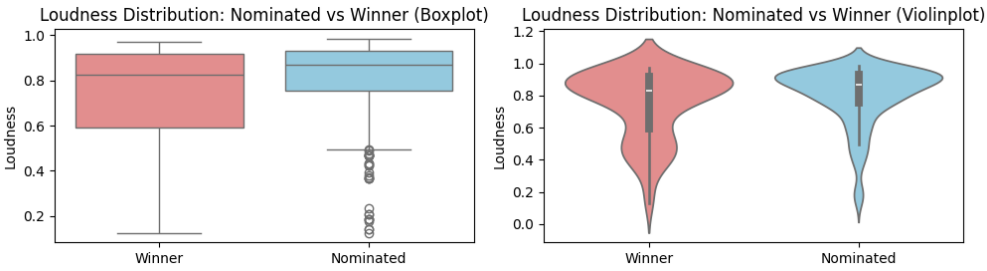


Figure 7: Distribution Loudness in Grammy Songs per Year (2000–2024)

In Figure 7, we can see that nominated songs have many outliers (with very low loudness), which means

softer-sounding songs tend not to win. After this, we will see you loudness separation between Grammys vs Non-Grammys songs.

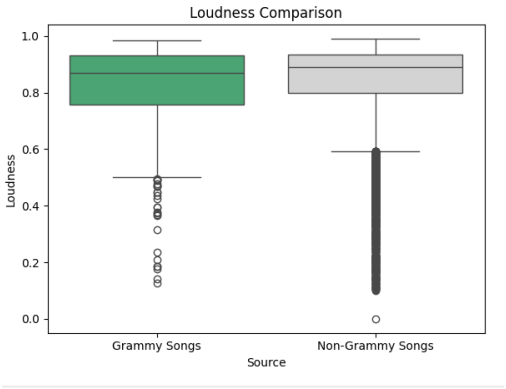


Figure 8: Distribution Loudness of Grammys vs Non-Grammys Songs

Figure 8 shows us that Grammy songs has more consistent loudness level than Non-Grammys songs. This means that Grammys' songs are polished to certain loudness variations.

4.1.3 Genre

Every song belongs to a genre. The Grammys feature numerous genre-specific categories, such as Best Pop Performance/Song, Best Rap Song, and Best Country Album. Our analysis focuses on three song categories [Best Rap Song, Song of the Year, Record of the Year].

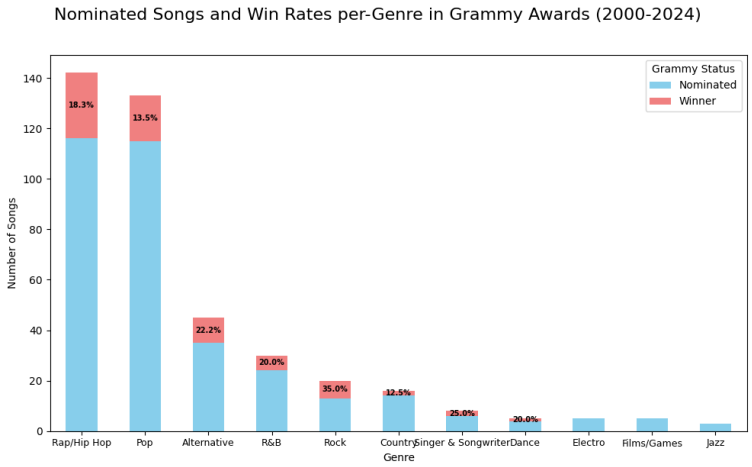


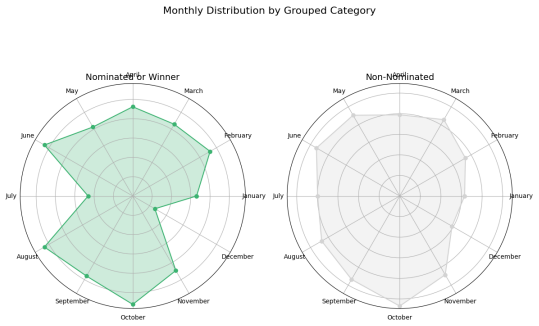
Figure 9: Nominated Songs and Win Rates per-Genre in Grammy Awards (2000-2024)

From a total of 26 genre types, the genres nominated from year to year (2000-2024) were only 11 but only eight genres have won Grammys at least once. In Figure 9, we can see that the most popular genres are Rap/HipHop and Pop. Interestingly, the "Singer & Song Writer" and "Dance" genres have 25% and 20% win rates, even though they rank only seventh and eighth in terms of nomination count. This suggests that some less frequently nominated genres may still have a strong chance of winning when they do appear. We want to see how genre evolve through each Grammys Category, except "Best Rap Song".

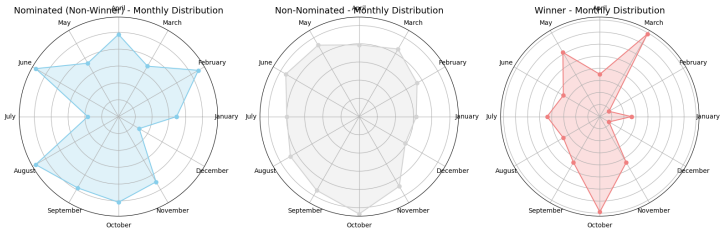
However, what stands out is the performance of the Rock. It has the highest win rate in the Record of The Year category. It also ranks second in win rate for the Song of The Year category. Despite receiving fewer nominations, Rock songs have a higher chance of winning once nominated.

4.1.4 Time Release

Does the time of release give a competitive advantage? We **aim to explore whether there is any pattern in the months and seasons winnings songs get released** compared to their nominated counterparts.

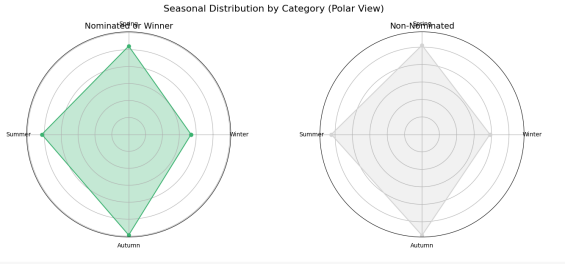


(a) Monthly Distribution by Category (Billboard)

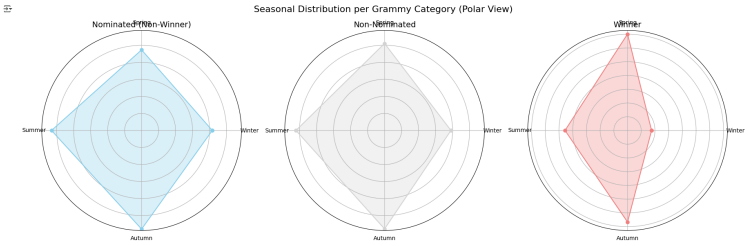


(b) Monthly Distribution by Category (Grammys)

Our data shows that nominated songs tend to avoid December and July releases. While winning songs peak in both October and March. **The October peak is explained by the Grammy submission deadline**, which is also around October or September. This mean that songs released around this time benefit of full years to make it to the chart and increase their sales. We find it harder to explain the March Peak.



(a) Seasonal Distribution by Category (Billboard)



(b) Seasonal Distribution by Category (Grammys)

We attempt a larger level of granularity, by grouping months by season. When it comes to seasons, it seems that winner are less likely to release in winter and summer. **Probably because winter is clogged with Christmas related songs, and summer is too close to the Grammy deadline to make an impact.**

4.1.5 Chart Performances and Debut Rank

An important question for us is whether strong Billboard performance is a requirement for a song to get nominated and to win. **Is charting longetivity is a requirement for getting nominated for a Grammy? If yes, what is the minimum required?**

We start by calculating the number of weeks each song remained on the chart, then plot the distribution for both nominees and winners.

We also looked at the top 20 longest-charting songs on Billboard. Surprisingly, we found that only a fourth of them had been nominated for a Grammy, and only one had won.

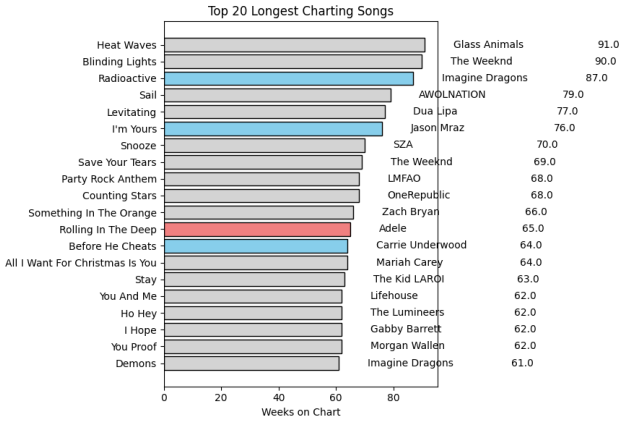
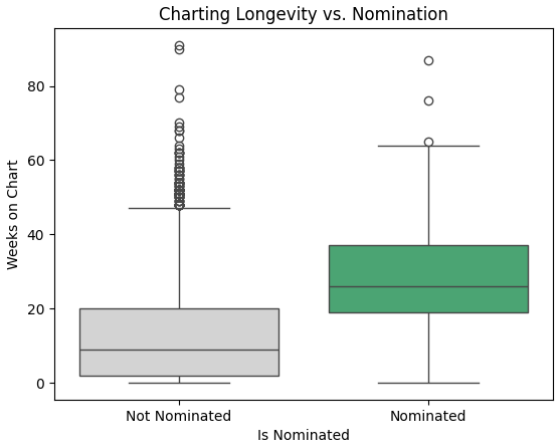
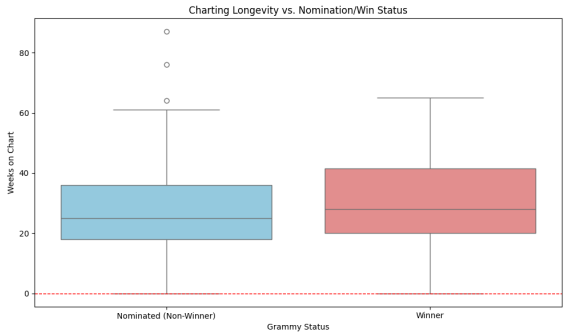


Figure 12: Top 20 Longest Charting Songs on Billboard



(a) Longevity: Nominated vs. Non-Nominated Artists



(b) Longevity: Nominated vs. Winning Artists

The data shows that while there is a big variance in chart longevity between nominated and non-nominated songs, it seems pretty homogeneous between winners and nominees. **However, a detail catches our attention: the bottom whisker of the box plots for winners and nominees being at 0**. We investigated this further, **as we found that a number of Grammy-nominated (10.1%) and even Grammy-winning songs (8.5%) have never charted on Billboard**.

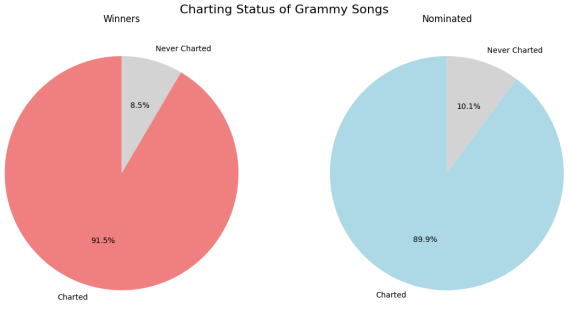


Figure 14: Grammy Songs That Never Charted on Billboard HOT100

How could this be the case? We investigated this further by retrieving the artists behind these win-

s/nominations, and we extracted the artists' discographies. We noticed that indeed, f**or a majority of these artists, they had an extensive discography with highly performing songs, some even with wins and nominations**. This shows that these artists were quite popular and well-known in the music industry, and their win for a non-charting song was attributed to that, but still makes them legitimate.

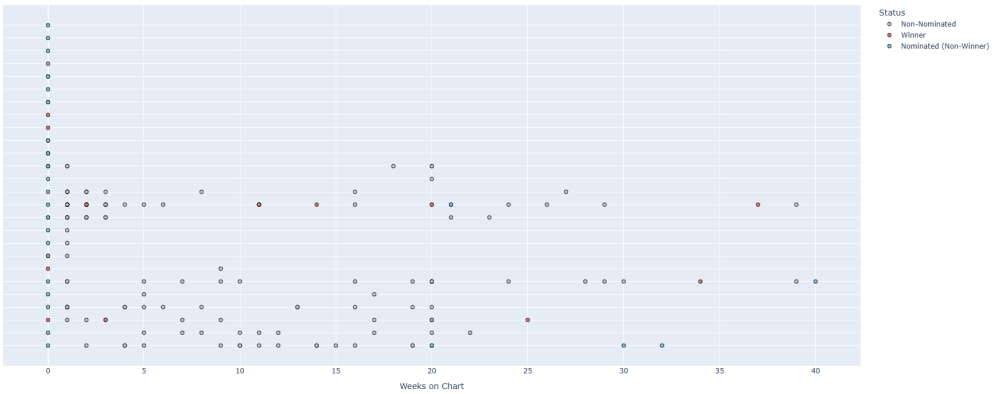


Figure 15: Discography of Artists with Uncharted Music Recognition

Debut Rank Finally, we also examined the debut rank and concluded that winning songs tend to debut higher on the charts (75th percentile around the top 20) compared to nominees, who in turn debut higher than non-nominated songs (around debut rank of top 60).

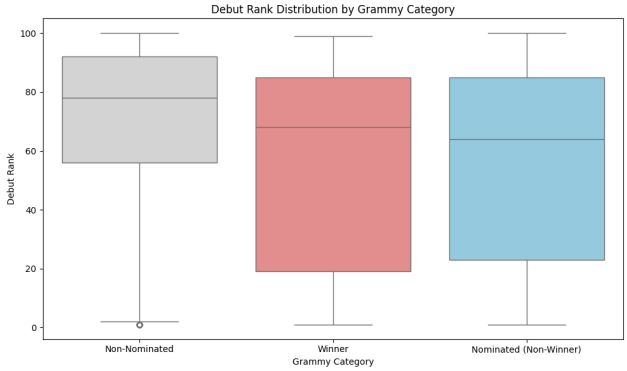
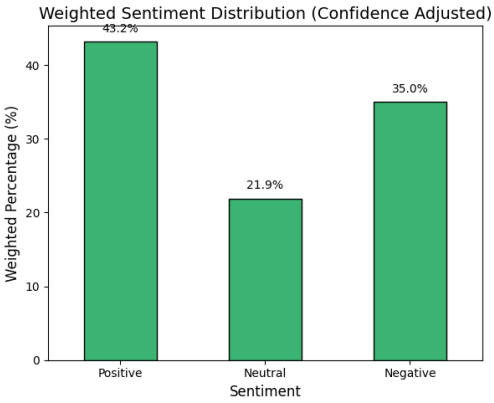


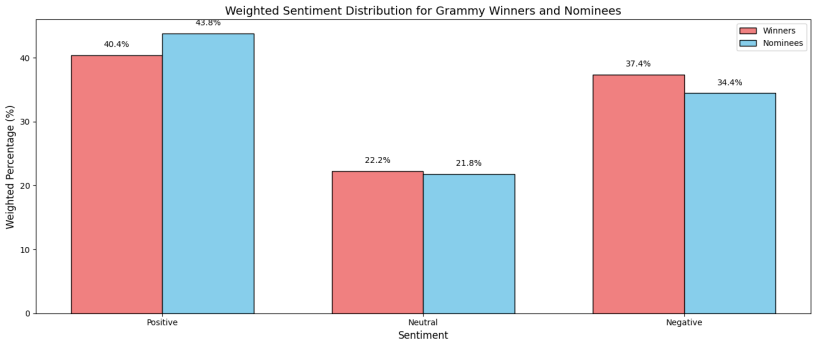
Figure 16: Debut Rank by Status

4.1.6 Lyrics

A key component of our analysis includes examining the content of song lyrics. We are interested in understanding **the themes and sentiments expressed in winning songs** and whether their distinction from nominees and Billboard songs lies in **writing quality or solely in sonic features**.

Next, we examine song lyrics in order to see if the content and themes in songs influence their status. First, looking at sentiment we notice that **the dominate sentiment in Grammy is positive** , with this trend also into breaking down data into winners and nominees, although with the lower variance.





(a) Weighted Sentiment Distribution on Lyrics Globally) (b) Weighted Sentiment Distribution on Grammy's Lyrics

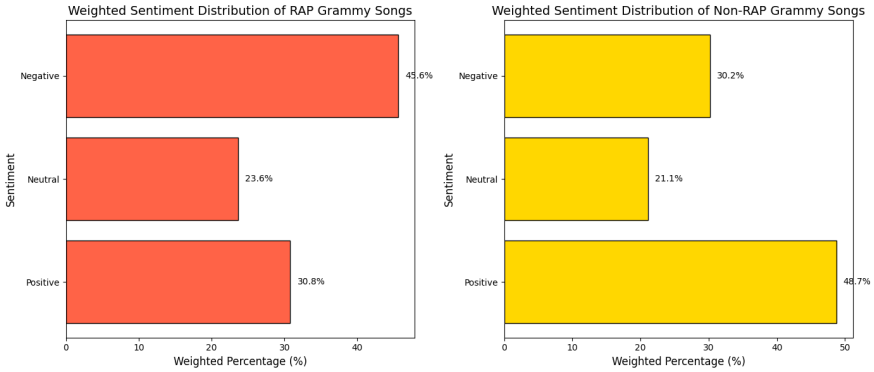
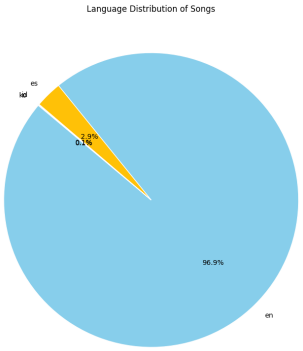
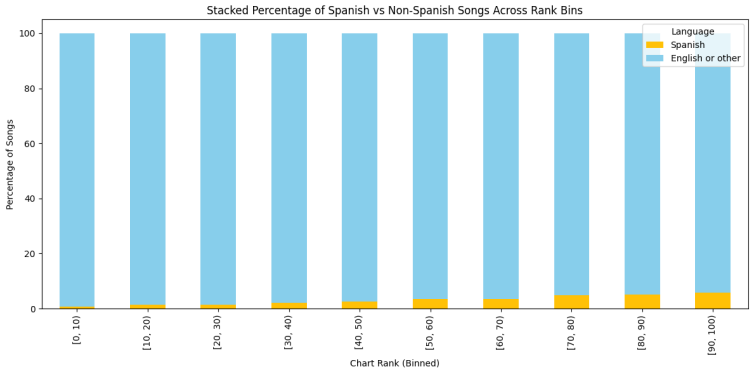


Figure 18: Lyrics Seniment of Rap Song's Category

However, when breaking down data into categories, we notice a different pattern with **Rap songs exhibiting a notably more negative sentiment**, compared to other Genres.

Language-wise, Grammy songs that are exclusively in English, unlike Billboard hits, which contain a small percentage of Spanish tracks. However, this does not indicate a language bias. Because when we see the specificity of non-English songs, we see that **they tend to rank much lower on a chart, compared to their English counterparts**, **making them less competitive for nominations.**



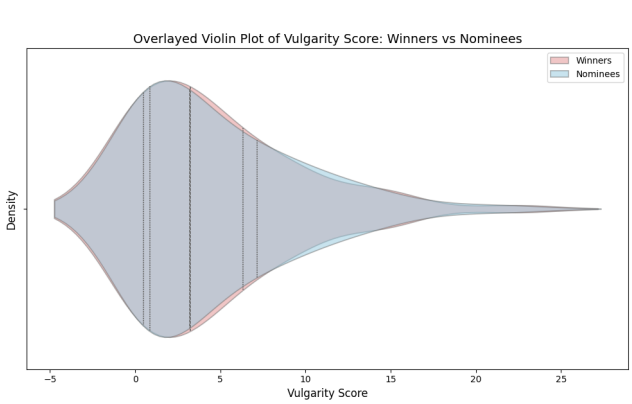


(a) Language Distribution of Songs in Billbaord Data

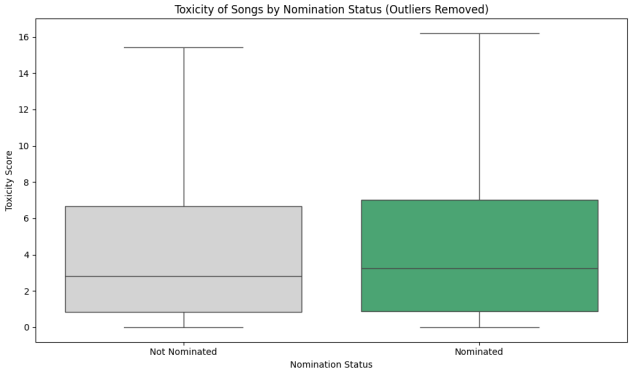
(b) Stacked Percentage of Spanish vs Non-Spanish Songs Across Rank Bins

Figure 19: Language Use in Songs and Their Popularity Distribution

We also aimed to explore vulgarity and its potential effect on wins. **We aim to understand whether voters tend to favor clean lyrics in winning songs or if vulgarity is considered an artistic choice with no impact on the outcome.**



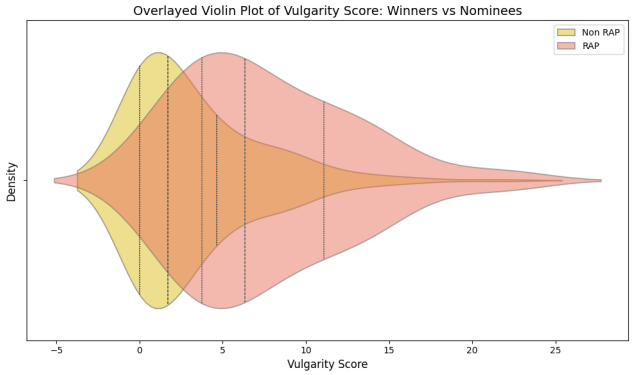
(a) Vulgarity Score: Winners vs Nominees



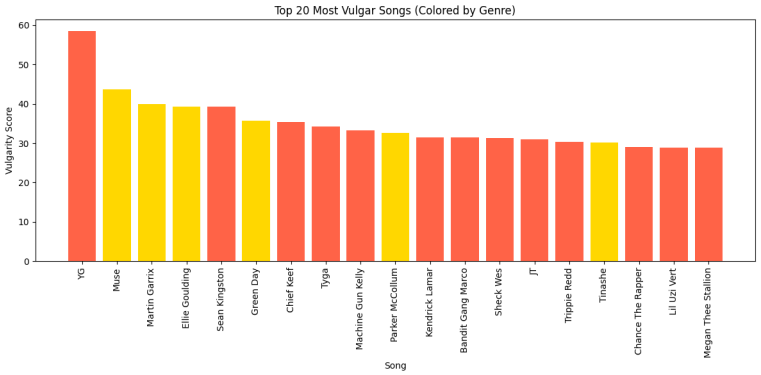
(b) Toxicity of Songs by Nomination Status (Outlier Removed)

Figure 20: Comparison of Vulgarity and Toxicity by Grammy Nomination Outcomes

Our data shows that vulgarity was not a majority characteristic of Grammy songs overall, as it has a similar level between winners and the nominees, as well as Grammy and non-Grammy songs. However in here we could see again that RAP Song has significantly more vulgar than other Genres.



(a) Vulgarity Scores in Rap vs Non-Rap Songs



(b) Top 20 Most Vulgar Songs by Genre

Figure 21: Analysis of Vulgarity in Songs Based on Genre and Ranking

Finally, our last component of our lyrical analysis of topics. Our aim is to understand if there are any recurring topics that distinguish winners from nominees. What do Grammys songs tend to talk about?

For this, **we use an LDA with 3 topics, and get the 10 top words of each topic.** We settle on 3 topics as we have tried a few experiments, and 3 topics provided with the neatest semantics.

Then, we use the distribution of topics per song to assign each song to the topic it most belongs to. This way, we can map each song to either topic 1, 2 or 3. Our three topics are as follows:

- **Topic 1: Romance, Love, Feelings, Longing**
  - love, way, right, see, time, well, need, think, god, feel
- **Topic 2: Romance, Love, Fun, Party, Energy**

– baby, tell, love, take, ever, time, never, name, work, dance

- **Topic 3: Slurs, Harsh Language, Aggressive**
  – bitch, fuck, nigga, shit, put, back, see, niggas, look, take

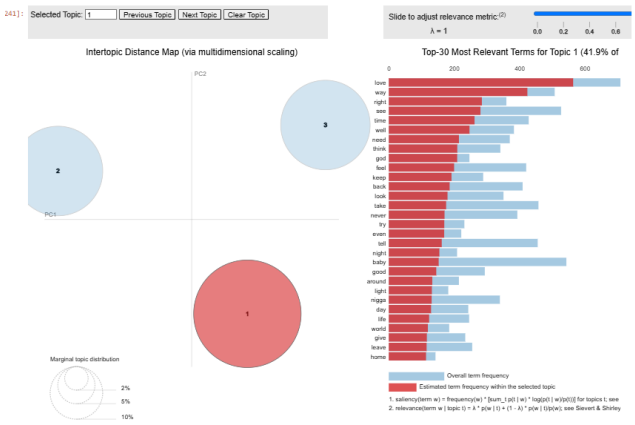


Figure 22: Topic Distribution Visualization with PyLDAvis

Topic 1 is about profound emotion, such as romance, Topic 2 centers around love and fun, and Topic 3 mostly made of harsh language and toxic language. We start with studying the distribution of topics across documents, in terms of winners vs nominees.

We see that in terms of whole distributions, higher win rates comes from a songs of the first topic.

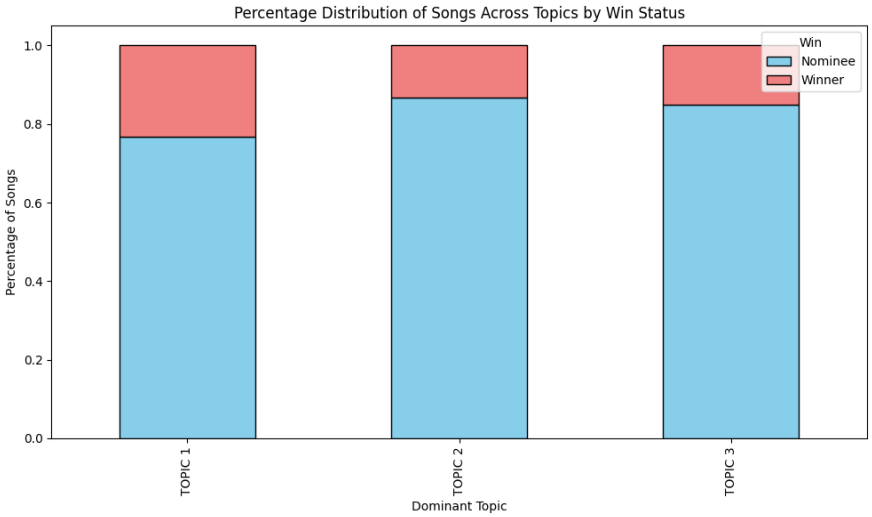
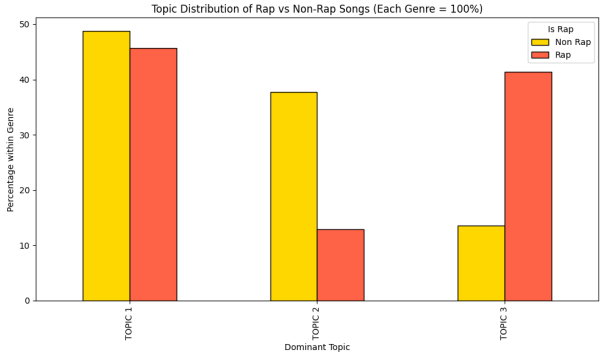


Figure 23: Percentage Distribution Songs Across Topic by Win Status

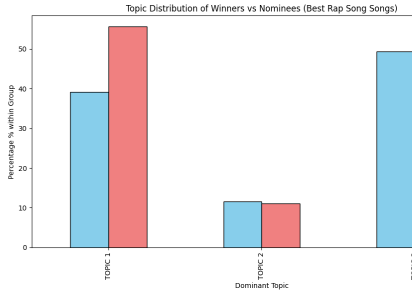
(a) World Cloud of RAP

(b) World Cloud of POP

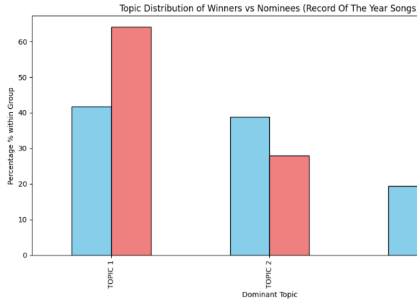


(c) Topic Distribution of Rap vs Non-Rap Song

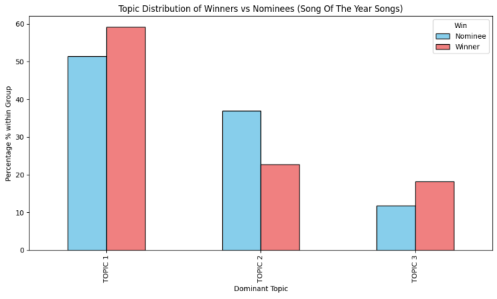
In terms of Rap songs, the prominent features is both for the topic of love and romance as well as round more explicit Rap themes. With world cloud also showing more slurs compare for example to Pop.



(a) Best Rap Song



(b) Record of the Year



(c) Song of the Year

Figure 25: Topic Distribution of Grammy Winners vs Nominees Across Categories

Intrestingly, when aligning the distribution of songs across Grammy Categories and win status, we observe that **winner Rap Songs surprisingly lean toward love and romance** teams, while nominated Rap Songs tend to stick to the traditional Rap topics. Non Rap winners on the other hand follow the more expected trend of love and romance. **This indicates that RAP winners distinguish themselves from other nominees by having more original topics, more Pop-like topics.**

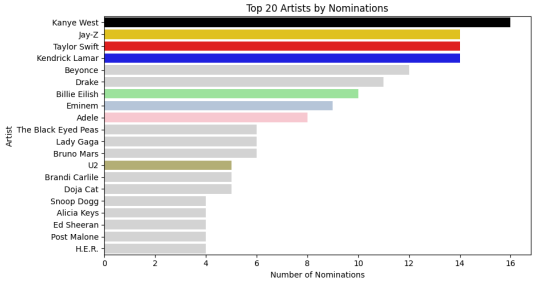
4.2 Artists Features

In this section, we aim to determine **what about artists give them a competitive advantage in Grammy nominations and wins ?**. We explore whether an artist's age or popularity plays a role, and if any biases exist regarding gender or origin.

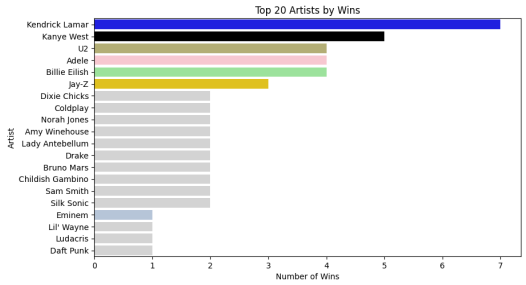
4.2.1 Artist Nominations

First, we are interested in uncovering **who is winning Grammys**? Do we have a plethora of artists who win, or is it the same well-known icons of the music industry that steal all the wins?

Our data highlights the prestige of this award with only 1.7% of best selling artists securing a win. First, we do a general analysis where we plot artists by their number of wins and number of nominations as shown in figure 26a. We see some well-knwon names, including Jay-Z, Kanye West, Kendrick Lamaar, and U2.

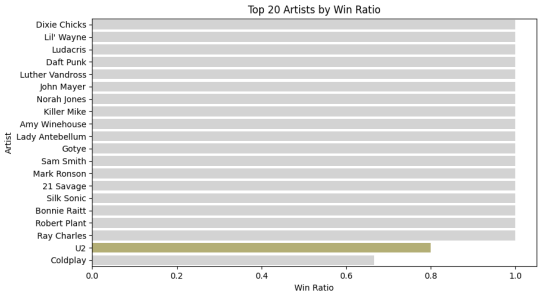


(a) Top 20 Artists by Nominations

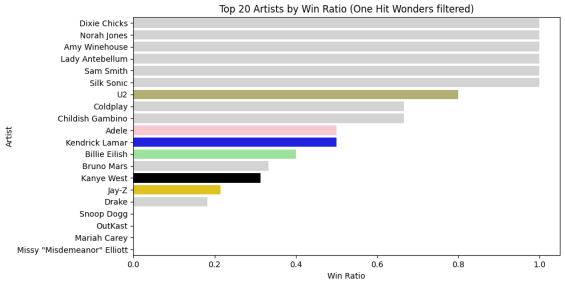


(b) Top 20 Artists by Wins

Notably, a comparison between the total number of wins and the win rate indicates that **nearly one-third of our winners are 'One-Hit Wonders': artists who received a single nomination and one win, then never made it to Grammys again**.Such names include Gotye with "Somebody That I Used to Know", among other artists.



(a) Top 20 Artists by Win Rate Ratio



(b) Top 20 Artists by Win Rate Ratio (One Hit Wonders Filtered)

We also explored whether an artist's discography matters. Do Grammy-recognized artists have outstanding chart performances, and do they dominate the awards? We calculated the total number of an artist's songs that hit the Billboard charts, as well as their cumulative weeks on the chart. We found that Grammy recognized artists show **significantly stronger Billboard performances with more charting songs, and overall longer cumulative weeks on the chart.** This result was statistically significant.

We observed a similar pattern when comparing winning artists to just nominated artists, but the effect was weaker and not statistically significant. **This highlights that having a strong discography influences an artist's nomination, but is not a decisive factor for securing a win.**

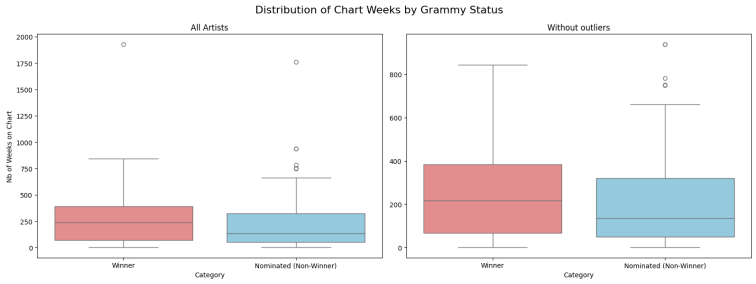


Figure 28: Distribution of Chart Weeks by Grammy Status Nominated vs Win

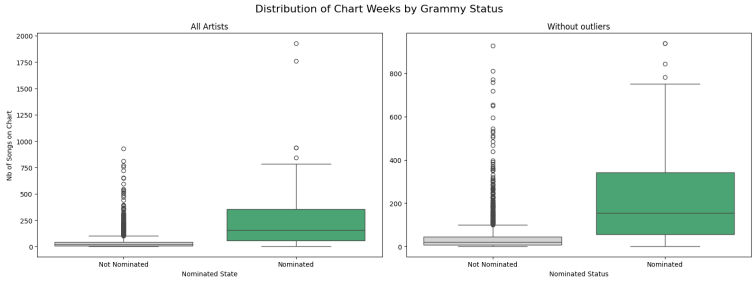


Figure 29: Distribution of Chart Weeks by Grammy Status Win vs Win

4.2.2 Age

Age is often an overlooked factor that determines the winner or nominee artist in the Grammy Awards. Perhaps experienced people tend to gain more attention from the jury.

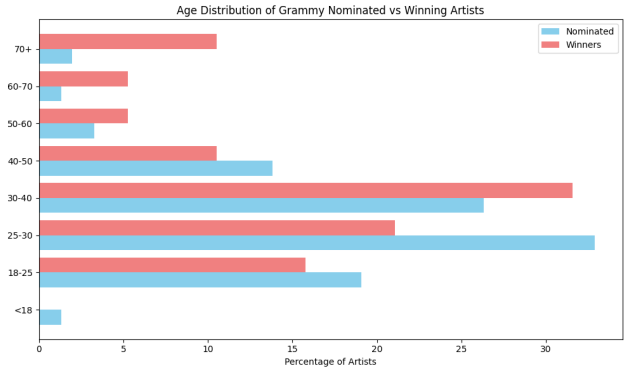


Figure 30: Age Distribution of Grammy Nominated vs Winning Artists)

In Figure 30, we see that both nominees and winners come mainly from the same age range, with the highest is in the 25-30 and 30-40 age groups. These two age group has the highest number of nominations and also the most wins, showing that artists in their 30s tend to be the most successful at the Grammys.

We also explore the data by breaking it down into Grammy category, specifically, we look at the Best New Artist category separately.

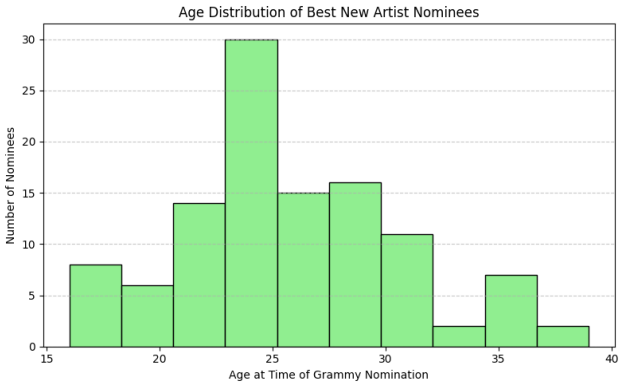


Figure 31: Age Distribution of Best New Artist Nominees

Figure 31 tells that most Best New Artist nominees are in their early to mid-20s. There is a clear concentration around 23 to 25 years old, meaning new artists getting Grammy attention tend to be quite young. It is very rare to see someone above 35 in this category.

4.2.3 Gender

Gender representation has been a long-standing topic of discussion, particularly at the Grammy Awards. **Examining gender distribution among nominees and winners can provide valuable insight into the evolution of inclusivity in music recognition over time.**

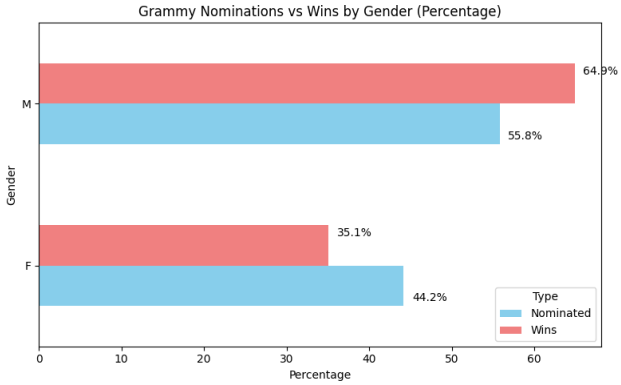


Figure 32: Grammy Nominations vs Wins by Gender (Percentage))

Figure 32 tells us that male are nominated and win more often, their win rate is also higher than women. We also display the gender of the best new artist winners.

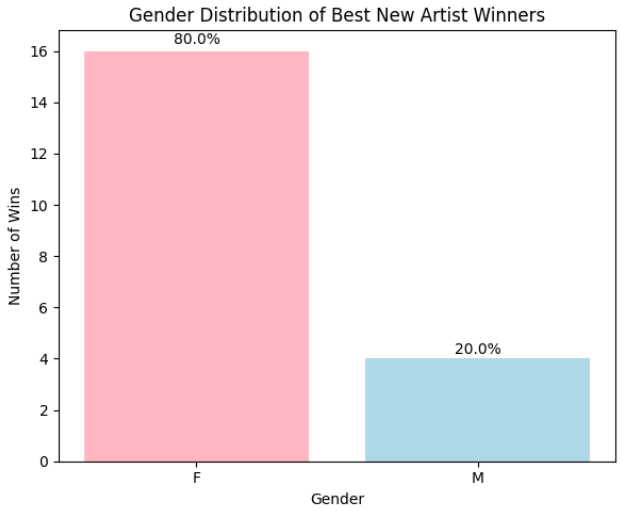


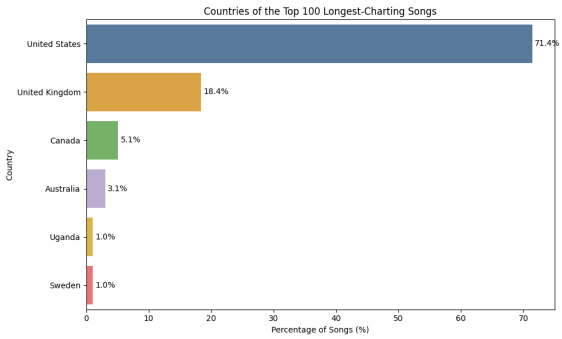
Figure 33: Gender Distribution of Best New Artist Winners

Surprisingly, women leads the victory in this category. This is a contrast from other Grammy categories. **Best New Artist category appears to provide greater recognition for emerging female talents,** possibly reflecting shifting industry dynamics and increasing visibility for women in music.

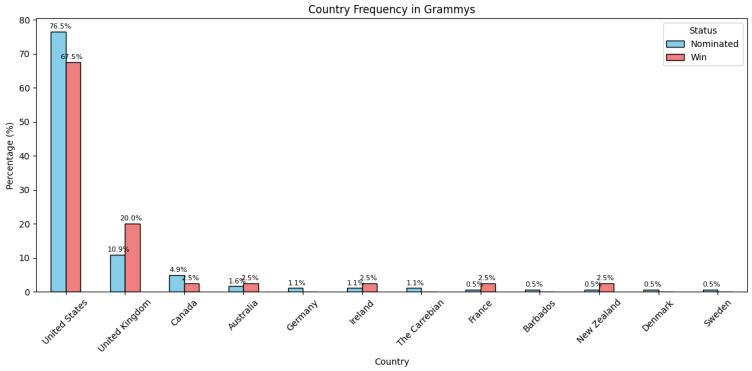
4.2.4 Country of Origin

By analyzing the country of origin, we aimed to uncover potential biases in country representation. Given that the Grammys are a US-based competition, we wanted to see the extent of international artist participation.

Unsurprisingly, we found that **US artists dominate Grammy nominations with over 75%**. Interestingly, the UK follows, securing 20% of the wins despite only receiving 10% of the nominations. This trend is also evident in Billboard data and persists when analyzing Grammy winners and nominees separately, **solidifying the US as the most represented country**.



(a) Bilbaord Performance by Country



(b) Grammy Wins and Nominations by Country

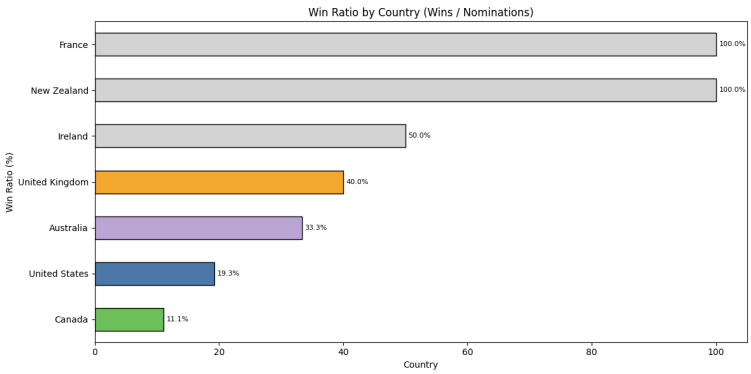
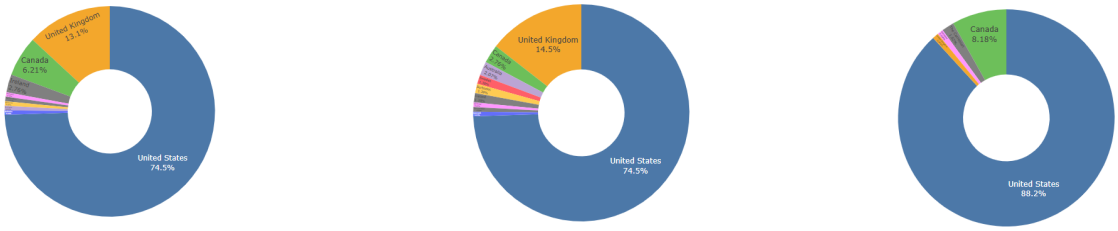


Figure 35: Win Ratio by Country (Wins/Nominations)

Interestingly, when plotting the win rate, we noticed many other country have 100% of rate, such as France and New Zealand. **However, these are mostly due to 'One Hit Wonder' (Daft Punk) or just a smaller number of nominations.**

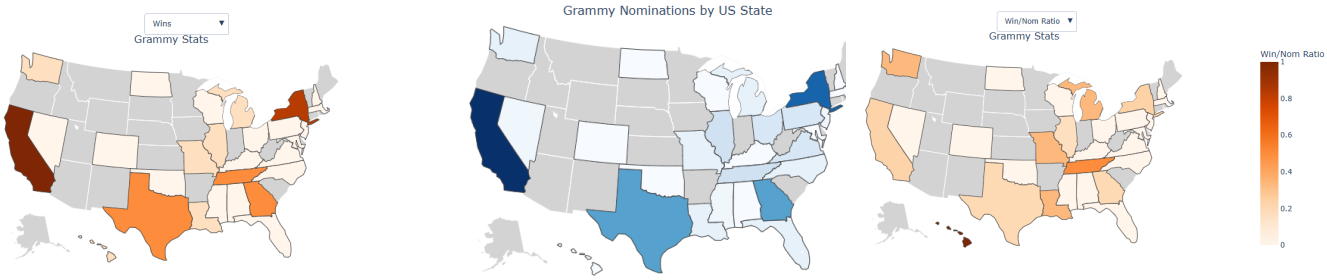


(a) Song of the Year (b) Record of the Year (c) Best Rap Song

Figure 36: Country Distribution of Grammy Winners Across Categories (2000–2024)

Splitting data by category allows us to **pinpoint the origin of our diversity, which mostly comes from Song Of The Year and Record Of The Year,** but not so much from 'Best Rap Song', which in the overwhelming majority includes only the US and Canada.

Given the predominance of US artists, we conduct a second analysis, excluding international artists and focusing on a finer level of granularity, by US states.



(a) Wins by State (b) Nominations by State (c) WinRates by State

Figure 37: Country Distribution of Grammy Winners Across Categories (2000–2024)

Focusing solely on the US, **only 52% of states are represented**, with **California, New York, Georgia, and Texas** leading in both wins and nominations. Hawaii and **Tennessee have the highest**

win rates; however, Hawaii's is likely due to its low number of nominations, in contrast to Tennessee's.

We then explored the relationship between US states' Grammy performance and their music infrastructure. Specifically, we examined five factors: the number of festivals hosted, the budget allocated to music, the number of music establishments (like recording studios), the number of music-related jobs, and the number of music institutions (like universities). To analyze the relationships between our music infrastructure metrics and Grammy performance, we generated a correlation matrix, a table showing the pairwise linear relationships between these variables, to identify any associations.

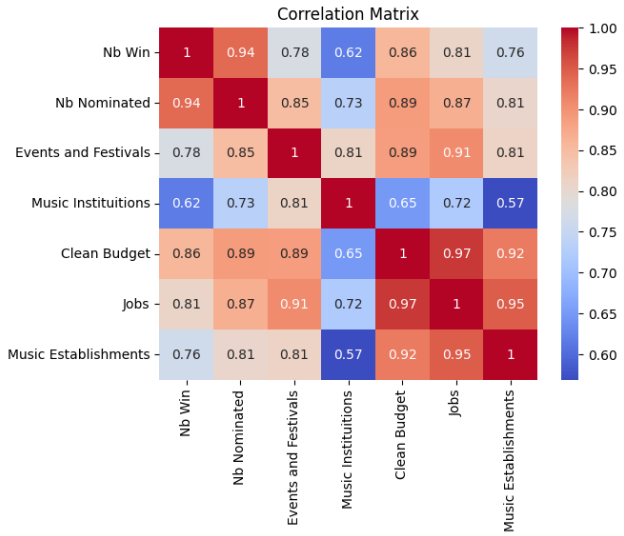


Figure 38: Correlation Matrix Between Grammy Success in US State and Robust Music Infrastuctures

Table 8: Pearson Correlation Coefficients with Number of Wins and Nominations

Variable Correlated With	**Pearson r**	**p-value**
Nb Win and correlation to other variables		
(Clean Budget)	0.86	< 0.05*
(Events and Festivals)	0.78	< 0.001*
(Jobs)	0.81	< 0.001*
(Music Establishments)	0.76	< 0.001*
(Music Instituitions)	0.62	< 0.001*
Nb Nominations and correlation to other variables		
(Clean Budget)	0.89	< 0.001*
(Events and Festivals)	0.85	< 0.001*
(Jobs)	0.87	< 0.001*
(Music Establishments)	0.81	< 0.001*
(Music Instituitions)	0.73	< 0.001*

We found a **strong and statistically significant link between Grammy success in US states and robust music infrastructure**. States with more music-related jobs, higher budgets, more music establishments, and more festivals tend to produce more nominees and winners. Interestingly, music institutions such as universities showed the weakest linear correlation, which aligns with the understanding that most artists are not college-educated. **These findings suggest that aspiring artists may benefit from relocating to states with strong music infrastructure to up their careers.**

4.3 Time Analysis

For our project's third objective, **we analyze our features with a focus on their evolution over time.** This approach aimed to **reveal potential trends** that become apparent through the years.

4.3.1 Tempo

Here we can see whether tempo varies over the years for winner or nominated songs in the Grammys.

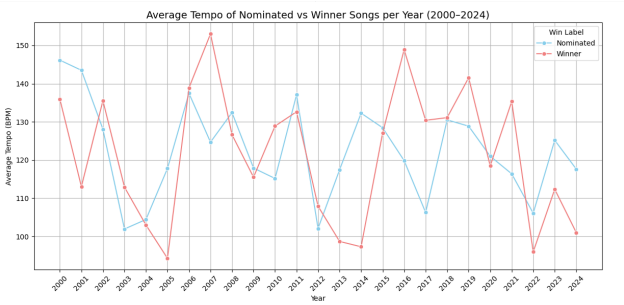


Figure 39: Average Tempo of Nominated vs Winner Songs per Year (2000–2024)

In Figure 39 we see that yearly trends reveal that there are several years where the average tempo of winning songs is either significantly higher or lower. This also applies similarly in every Grammy categories.

4.3.2 Loudness

Loudness also play crucial role in modern music production. Seeing the trend will be beneficial for-preparing future songs that singer want to produce.

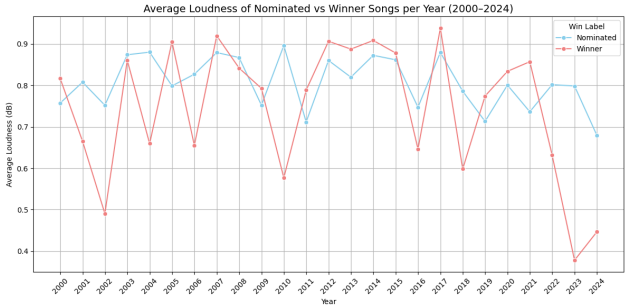


Figure 40: Average Loudness of Nominated vs Winner Songs per Year (2000–2024)

In Figure 40 we see that the average loudness for nominated songs fluctuates from year to year, but compared to winner songs, nominated songs have a bit more consistent average, ranging from 0.7–0.9 dB range.

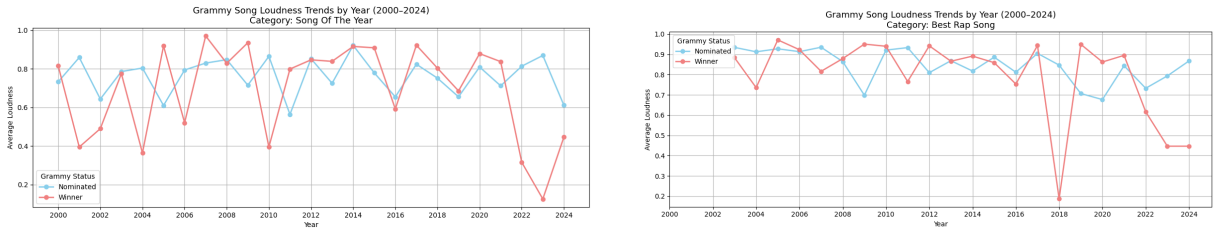


Figure 41: Gender Distribution of Grammy Winners Across Categories (2000–2024)

In Figure 41 For the Grammy categories, we have an interesting analysis part on just in two categories. In Song of the Year, louder songs tend to win more frequently. In Best Rap Song, loudness is almost a baseline requirement.

4.3.3 Genre

Genre plays an important role in the music industry. Analysing the evolution could also have a big impact on the industry and the music industry.

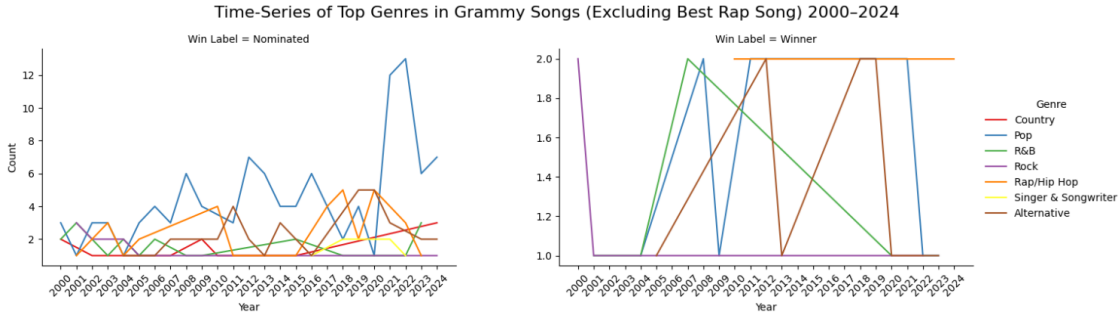


Figure 42: Time-Series of Top Genres in Grammy Songs (Excluding Best Rap Song) 2000–2024

We can see from Figure 42 that Pop always dominates Grammy nominees. However, the genre for the winner seems to fluctuate. This also the case of the two categories (Song Of The Year and Record of The Year).

4.3.4 Age

Age evolution also needs to be taken into account because from previously, Grammys tend to give the throne to adult.

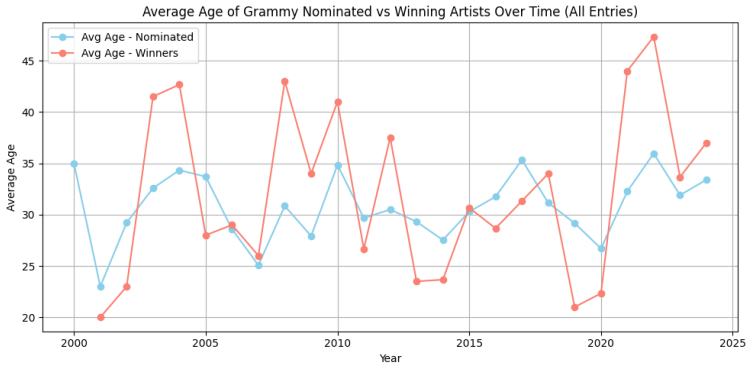


Figure 43: Average Age of Grammy Nominated vs Winning Artists Over Time (All Entries)

From Figure 43, we observe that nominees tend to be centered around the 25–30 age range, whereas winners show a more fluctuating distribution. However, both nominees and winners are mostly above the age of 25.

4.3.5 Gender

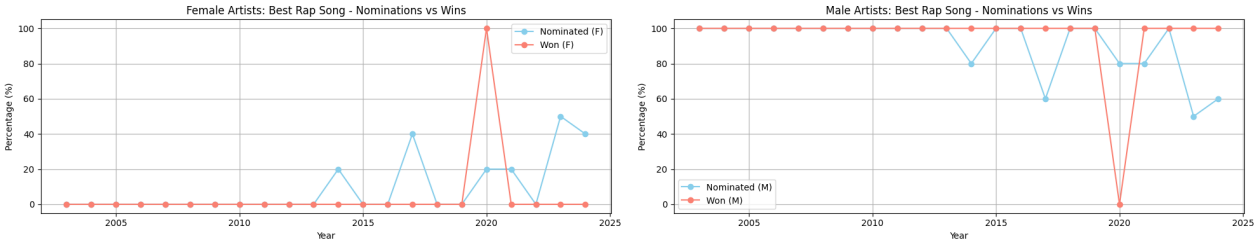
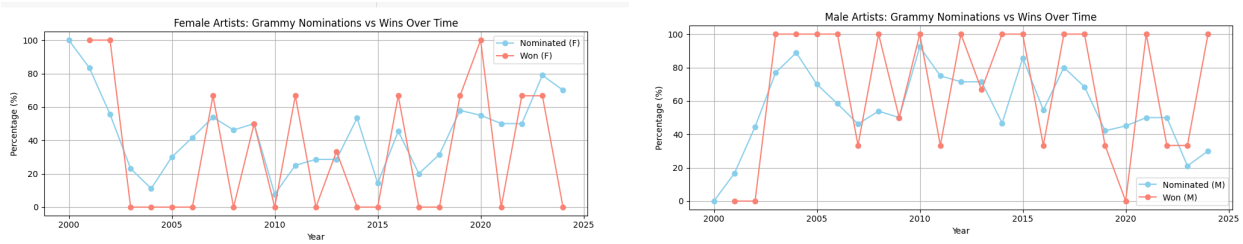


Figure 44: Gender Distribution of Grammy Winners Across Categories (2000–2024)

From Figure 46, from previous analysis, we know that male artists consistently dominate in terms of wins. This also applies to Record Of The Year, and Song Of The Year categories.

However, an interesting part comes to the Best Rap Song category. In Best Rap Song, the disparity is most pronounced: female artists are rarely nominated, and almost never win, with a single win around 2020 standing out as an anomaly.

These trends suggest that while nomination practices have become more inclusive, actual award recognition still tends to favor male artists, particularly in traditionally male-dominated categories like rap.



(a) Female Artists: Grammy Nominations vs Wins Over Time

(b) Male Artists: Grammy Nominations vs Wins Over Time

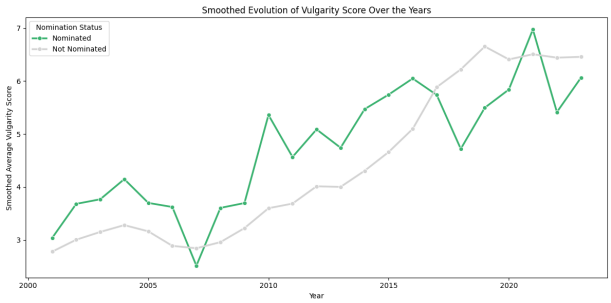
Figure 45: Gender Distribution of Grammy Winners Across Categories (2000–2024)

Here, we also seeing an increasing of female Grammy nomination, especially starting in 2015. However, it seems male still dominate the industry.

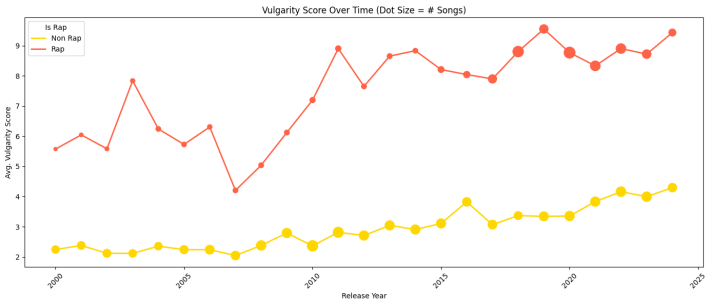
4.3.6 Lyrics

In terms of trends, we have 3 main findings when it comes to lyrics time evolution:

- **The music industry is getting more tolerant of vulgarity,** with both Billboard songs and Grammy songs showing an increase in vulgarity score over time, specifically starting from the year 2007.
- The music industry also seems to be **leaning towards more negative songs.** Although this did not show in Grammy data, Billboard charts show that starting from the year 2015, there have been more negative songs than positive songs charting on Billboard.
- In terms of language, non-English songs, specifically Spanish songs, seem to be regaining popularity, especially during the years 2021 to 2024, with multiple Spanish hit songs like "TQG" by Karol G and Shakira (2023) and Latino artists like Bad Bunny and Rosalía rising in popularity. This, however, did not reflect in the Grammy data, which remained 100% English songs.

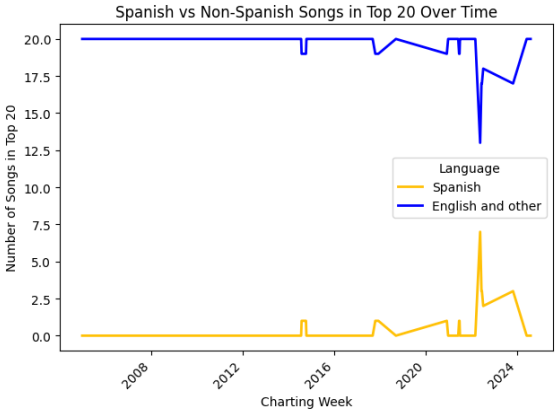


(a) Smoothed Evolution of Vulgarity Scores by Nomination Status

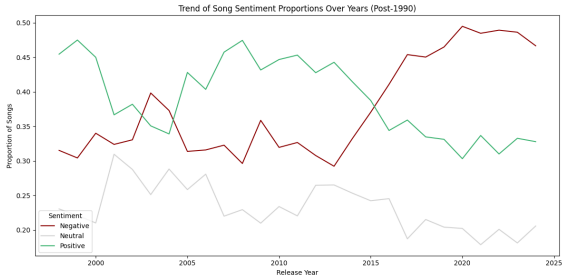


(b) Vulgarity Scores Over Time (Dot Size = Number of Songs)

Figure 46: Temporal Trends in Vulgarity by Nomination Status and Genre



(a) Spanish vs Non Spanish Songs in Top 20 Over Time



(b) Evolution of Emotions in Songs Over Time

5 Predictive Modeling

Our third goal is to develop a model capable of accurately and reliably differentiating between Grammy winners and nominees. To achieve this, we will use the Grammys Dataset as our training set. This section will first provide a brief overview of our data preprocessing steps before detailing the modeling process. Subsequently, we will expand on the various experiments conducted and compare their results.

We note the following key details regarding our experimentation and the rationale behind our modeling choices:

- **Task: Binary Classification** - Our objective is to classify entries as either winners (class 1) or nominees (class 0).
- **Cross Validation** - Given the limited size of our dataset, we will employ 5-fold cross-validation and average the results to ensure a more robust interpretation.
- **Oversampling** - The dataset exhibits a class imbalance (82.57% wins and 17.43% nominations). To mitigate the potential for the model to be biased towards the majority class, we will evaluate two techniques: utilizing class weight parameters and applying SMOTE Oversampling.
- **Feature Selection** - We will compare the performance of our models using two approaches: incorporating all available features and employing Principal Component Analysis (PCA) for our numerical features.

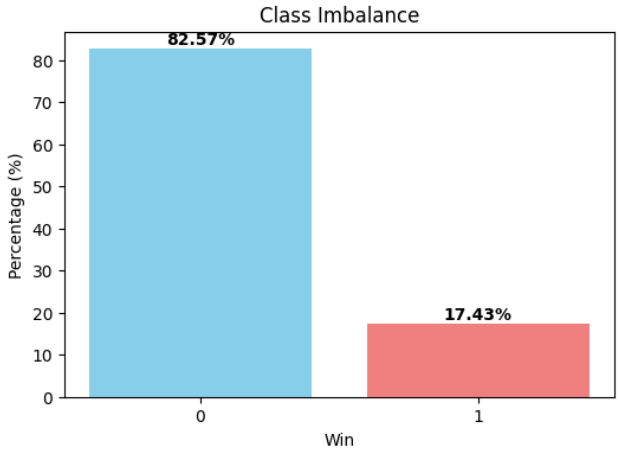


Figure 48: Target Class Imbalance

5.1 Preprocessing and Data Preparation

The dataset contains a mix of categorical and numerical features. First, based on the insights from our EDA, we perform a preliminary selection of features to include only those showing a variance between the winner and nominee classes.

In total, we have 16 features:

- **Numerical Features:** Inlcuding 'Vulgarity Score', 'Duration', 'Tempo', 'Loudness', 'Nb of Weeks on Chart', 'Debut Rank', 'age', and 'Release Month'.
- **Categorical Features:** Including 'ArtistID', 'Song Language', 'Genre', 'Combined Sentiment', 'Dominant Topic', 'Country', 'Gender', and 'Lyrics'.
- **Target Feature:** This is 'Win'.

Then, we applied preprocessing steps that are relevant to the type of each feature.

- **Numerical Features:** We normalized and standardized using StandardScaler. Missing values were handled by replacing them with the mean.
- **Categorical Features:** We applied One-Hot encoding. Given the small size of our dataset, removing rows with missing values would have resulted in a significant loss of information. Missing values were handled by assigning them a distinct "missing" category.
- **Lyrics Data:** For the lyrics, we use the same preprocessing detailed in our previous lyrics analysis. Mainly, we apply a lemmatization and multilingual stop words removal.

5.2 Principal Component Analysis

Due to our large number of features, we want to improve the performance of our model to avoid falling into the curse of dimensionality. For this, we go for dimensionality reduction using principal component analysis for our numerical features.

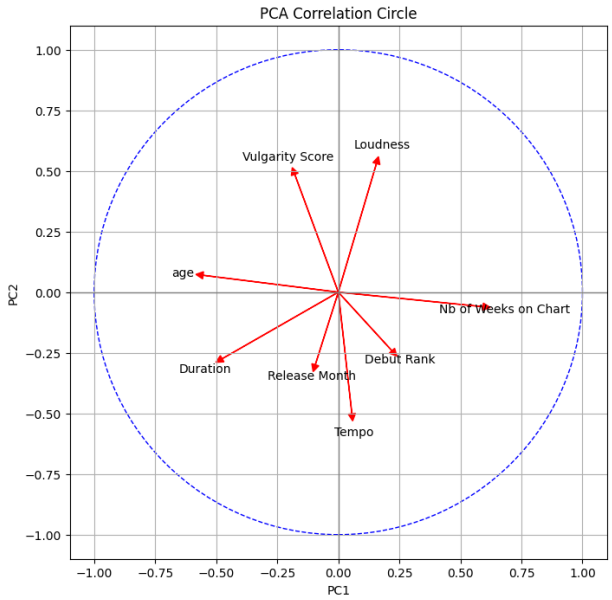


Figure 49: Projection of Features onto PCA Components

Our PCA results reveal the following:

- **Component 1:** Mainly represented by chart performance (positively) and age and duration (negatively). 'Debut Rank' also contributes, but its shorter vector suggests it is less strongly represented by PCA1.
- **Component 2:** Mainly represented by 'Vulgarity Score' and 'Loudness' (positively) and 'Tempo' (negatively). 'Release Month' also has a positive projection but is less well represented.
- 'Loudness' and 'Vulgarity Score' are positively and strongly correlated, likely due to songs in genres like hip-hop and rock.
- 'Number of weeks on chart' and 'Debut Rank' are positively correlated. This suggests that songs with longer charting periods tend to have a higher debut rank, which might seem counterintuitive.
- The age of an artist and the duration of the song appear to be correlated, though we don't have an intuitive explanation for this relationship.
- 'Number of weeks on chart' and 'age' are negatively correlated, which aligns with our previous finding that popular artists (with popular and long charting songs) tend to be younger.
- 'Loudness' and 'Tempo' are negatively correlated, which is an interesting observation.

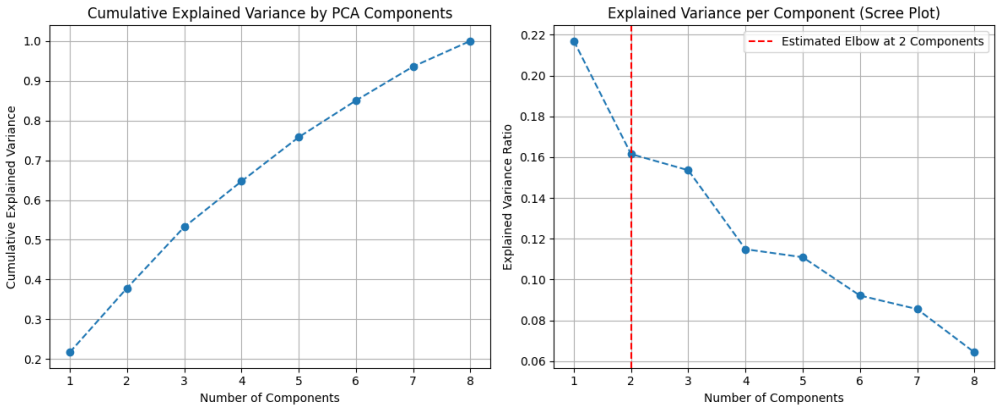


Figure 50: Cumulative Variance and Explained Variance

To determine the optimal number of principal components, we examine the cumulative and explained variance as shown in Figure 24 50. Although the elbow method suggests 2 components, we notice it captures less than 40% of the variance. Thus, we will settle for using 6 components (+80% explained variance) and 7 components (+90% explained variance) to ensure sufficient information retention.

5.3 Models Performances

We experimented with several algorithms, including HistGradientBoosting, Random Forest, and Linear SVM. For each training scenario, we consistently applied 5 fold cross-validation and averaged the results.

We showcase the differences in our features and models in the tables and graphs below.

5.3.1 Experiments on mitigating class imbalance

We attempt 3 methods for mitigating class imbalance: SMOTE Oversampling, and parameterized class weight balancing. We compare them against the baseline of imbalanced classes. The results below were run on a HistGradientBoostingClassifier model.

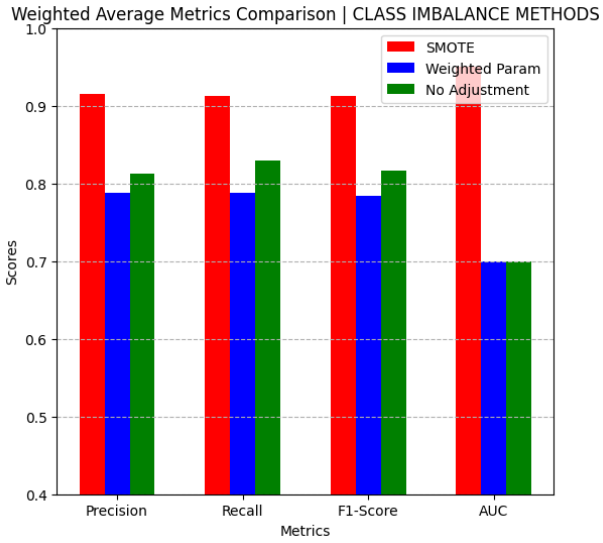


Figure 51: Comparison of class imbalance techniques

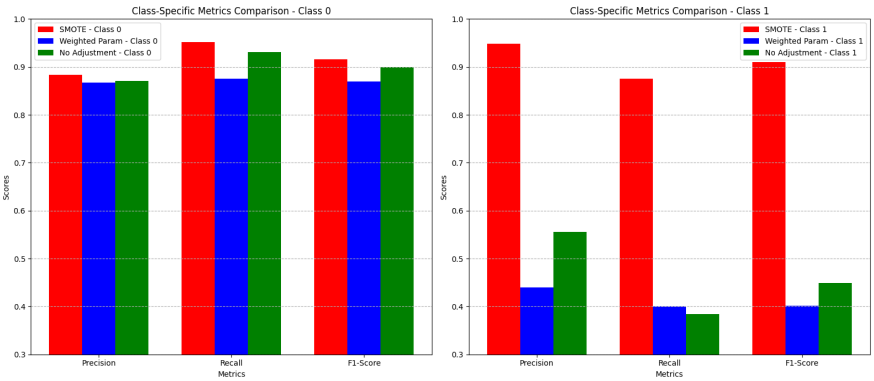


Figure 52: Comparison of class imbalance techniques per class

The results indicate that **SMOTE oversampling significantly enhances performance**. This is apparent in Graph 2 of Figure 51, where SMOTE is the only model demonstrating **balanced performance across both class 0 and class 1**. It's important to note that averaging results can be misleading, as highlighted by a comparison of both charts.

5.3.2 Experiments with models

We compare 3 models: Linear SVMs, HistGradientBoostingClassifier, and Random Forest. **Our results show that Linear SVMS give us the best results,** with the highest balance between classes and overall average metrics. The three models have roughly similar performances.

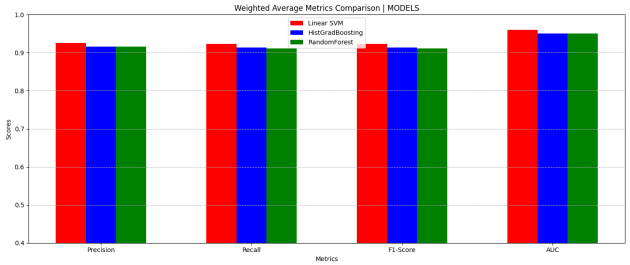


Figure 53: Comparison of Models

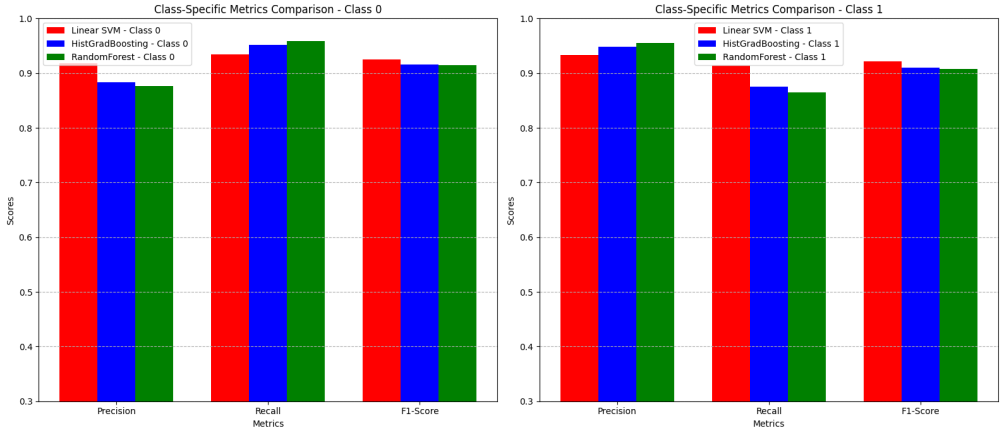


Figure 54: Comparison of Models Performances by class

5.3.3 Experiments with features

Here, we will compare raw features to PCA components. We will compare the PCA with 6 and 7 components, and compare them to the baseline of raw features.

Results show PCA components do not bring any improvements.

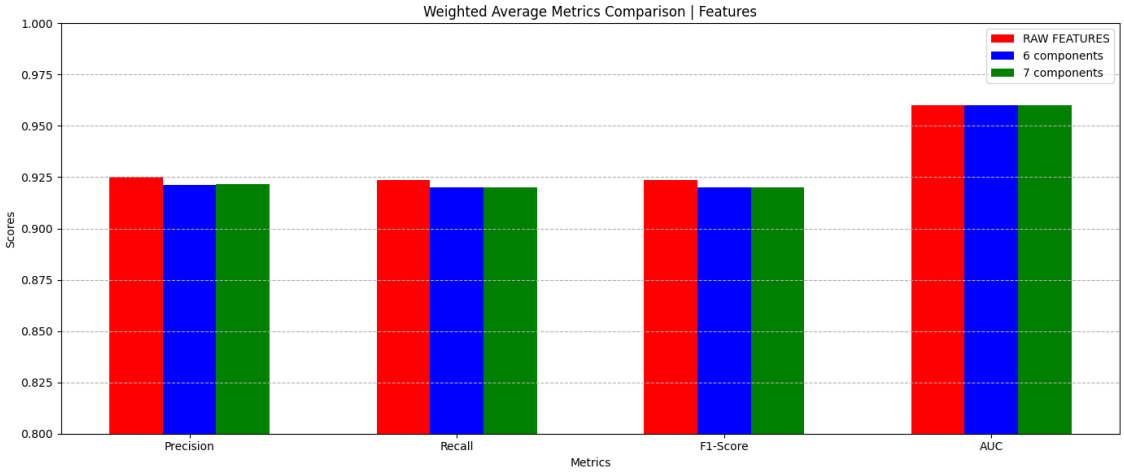


Figure 55: Comparison of input Features (nb of PCA components vs RAW)

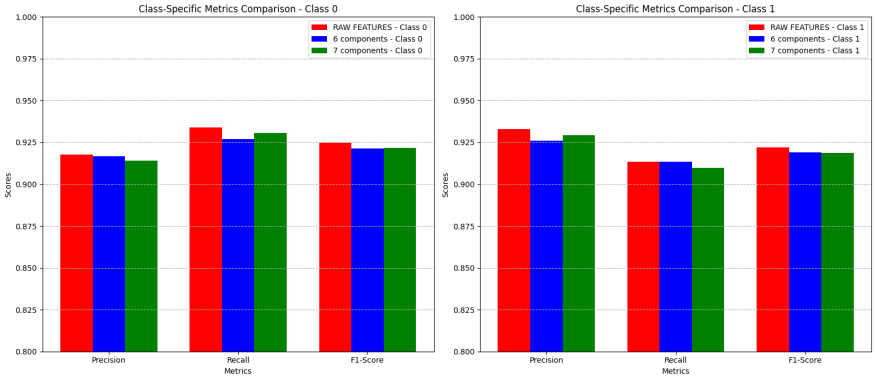


Figure 56: Comparison of input Features per class

5.3.4 Discussion

Our best-performing model is a Linear SVM, utilizing raw features with SMOTE oversampling. The results, depicted in Figure 57 and Table 9, demonstrate the model's **high accuracy in distinguishing between classes**, achieving an AUC of 96%. Notably, the model exhibits balanced performance, maintaining scores of approximately 92% for F1-score, precision, and recall, indicating **no bias towards the majority class**.

Table 9: Classification Report

	Precision	Recall	F1-Score	Support
0	0.9177	0.9340	0.9249	57.6
1	0.9330	0.9132	0.9222	57.6
macro avg	0.9254	0.9236	0.9235	115.2
weighted avg	0.9253	0.9236	0.9236	115.2
Average accuracy over folds: 0.9236				

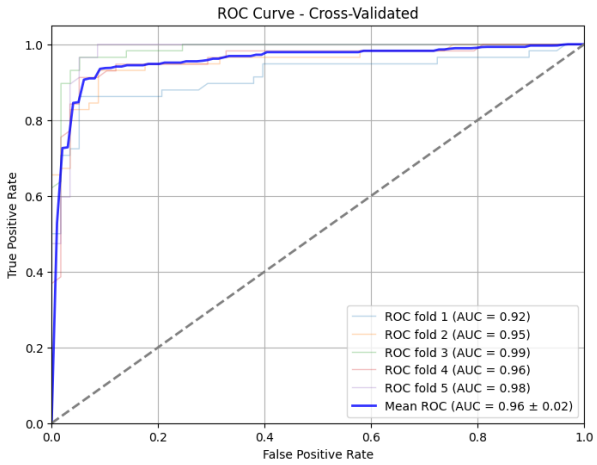


Figure 57: Best Model AUC Curve

6 Limitations

Our analysis is pretty solid with many significant result to determine whether someone could be a Grammy winner or not. However, when it comes to differentiating winners from non-winners, the findings are less conclusive. Several limitations contributed to this.

- **Limited Category Scope** Our project focused on four Grammy categories, which potentially leads to misleading results. For example, Taylor Swift has never a a Grammy according to our data, however, the reality is that she has won multiple times in Album of the Year (a category that was not included in this work). Similarly, when discussing songs nominated/winning Grammys that never charted on Billboard, this doesn't imply they didn't chart on iTunes or Hot 200.
- **Language Constraint in Lyrics Analysis** In this project, the analysis of the lyrics is only for English-language songs, excluding potentially rich insights from non-English tracks.
- **Missing data and Bias** Several entries had missing values in key features such as tempo, lyrics, loudness, and country, which may have introduced bias. Also, the limitation of just four categories analysis in Grammy leading us to deal with another selection bias.
- **Country Attribution Issues** To assign nationality, we relied on the artist's places of birth. This does not fully capture the industry realities. For example, Canadian-born artists like Drake are labeled as Canadian, despite having built his career in the U.S, and being affiliated with a US labels.
- **Sentiment Analysis Limitations** The sentiment analysis part, we possibly missed nuances such as irony, metaphores, or deeper feelings, making the analysis approximative. This is due to our sentiment model not being trained specifically on song lyrics.

7 Conclusion

In this work, we aimed to answer the question "How to Win a Grammy Award" with three core objectives: (1) identify key features differentiating successful songs, (2) analyze Grammy trends against Billboard data, and (3) develop a robust predictive model for winners.

Overall, our analysis provided relevant insights and reliable predictions. We first show that while song longevity on charts and high debut ranks are beneficial, they aren't strict requirements for a Grammy win. Then, our results revealed that winning rap songs often diverge thematically from general rap trends. We also notice the dominance of the U.S and English-language songs in both the Grammys and charts, with a strong correlation between robust U.S state musical infrastructure and increased nominations and wins.

Furthermore, our trend analysis indicated a shift towards more vulgar and negative lyrical content being favored over time, alongside an encouraging increase in female artist nominations. Finally, our predictive model achieved excellent performance, accurately and reliably distinguishing between classes.

However, limitations remain due to inherent biases, such as selection bias and missing data bias, which could impact the generalizability of our findings. Other improvements, like incorporating a wider range of data sources could be used to tackle these challenges.