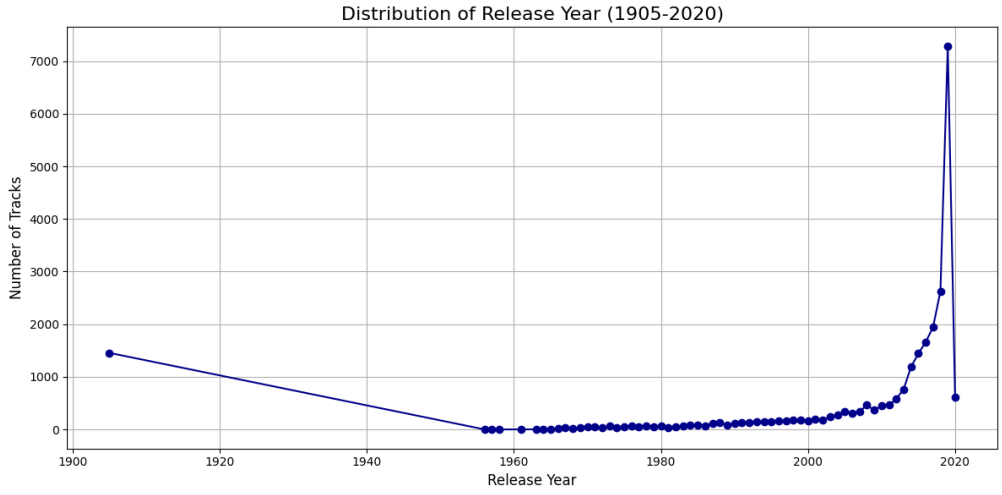
[Spotify Streaming Analysis and Marketing Investment Strategy]

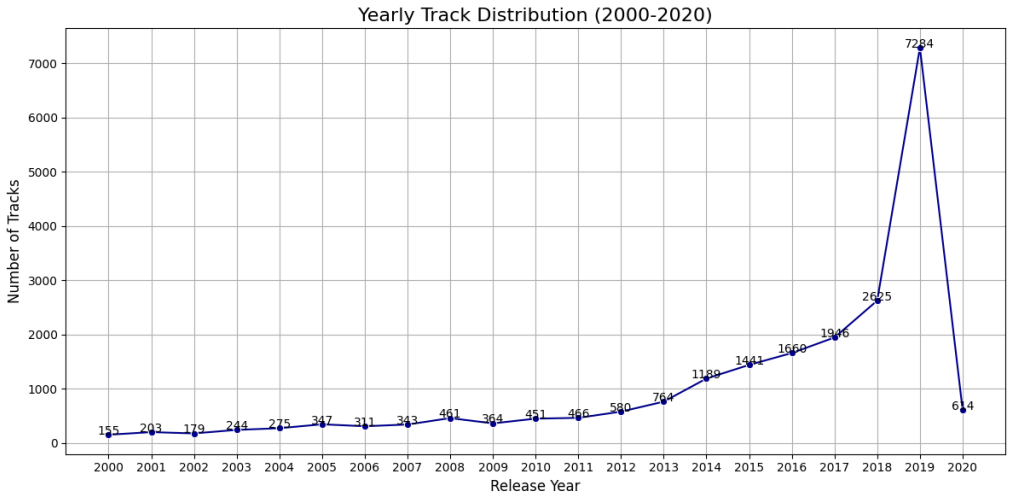
This project analyzes the key characteristics of successful songs to help Universal Music enhance its promotion strategies using Spotify data. The dataset, Spotify 1, contains 26,266 entries of historical songs released before March 2020, capturing long-term trends. This analysis provides insights into how music trends have evolved over time and identifies the factors that contribute to a song's success in the current market.

The exploratory data analysis identified significant trends in the music landscape. Genres such as pop, Latin, and R&B dominate, with Latin music experiencing remarkable growth in recent years, even surpassing rap and pop in some instances. Popular songs share distinctive characteristics, including high danceability, energy, and frequent use of major keys like Db/C#, C, and G, which evoke positive and uplifting emotions. Most successful tracks are studio-produced, feature prominent vocals, and align with modern listening habits, with an average duration of approximately 3:20 minutes. Key release months—January, June, and November—were identified as critical periods for maximizing exposure, coinciding with promotional cycles and seasonal engagement. External factors, such as artist fame, viral trends on platforms like TikTok and Instagram, and collaborations with well-known artists, also play a significant role in boosting song visibility and success. Below are the visualizations for each variable.

[Release Year]
From 1900 to 2000, the number of music releases remained steady. However, focusing on the period between 2000 and 2020, there is a noticeable **upward trend in releases starting in 2014**, culminating in a **dramatic peak in 2019**. Several factors contribute to this phenomenon.

- **Growth of Streaming Platforms**: The expansion of streaming platforms played a pivotal role. According to MIDiA Research, 2019 was marked by growth and consolidation in the streaming industry, with Spotify retaining a significant market share despite competitors like Apple, Amazon, Tencent, and Google gaining ground. Counterpoint Research further highlights that global online music streaming subscriptions increased by 32% in 2019, reaching 358 million subscribers. This rise in streaming incentivized artists and labels to release more music to meet growing consumer demand.
- **Growth of User-Generated Content Platforms**: The surge in popularity of platforms like TikTok and YouTube significantly impacted music releases. Artists increasingly tailored tracks for viral challenges and short-form content, resulting in a higher volume of releases to cater to these trends.



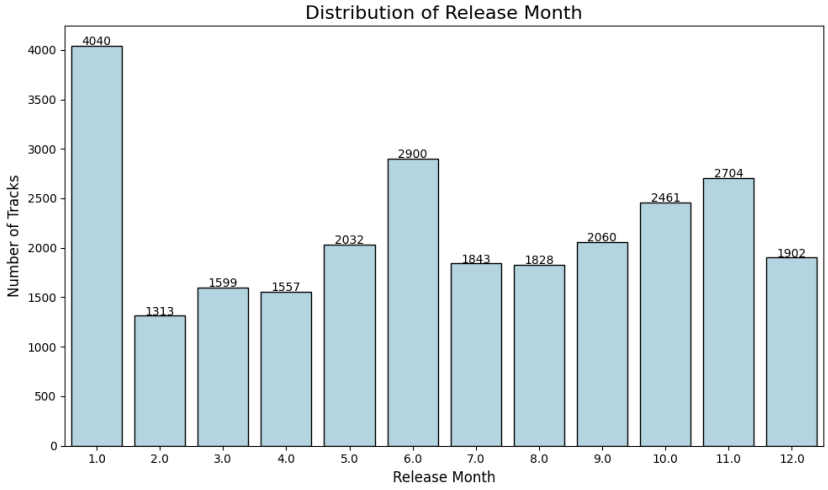


[Release Month]
Top 3 Months for releases are **January, June, and November**.

1. January
   - **Industry Characteristics**: Labels often launch promotional campaigns for singles or albums with long-term potential. Early releases can also qualify for year-long award eligibility, ensuring ample time to build momentum.
   - **Budget Allocations**: New fiscal year budgets encourage labels to fund new projects and marketing initiatives.
   - **Strategic Marketing Purpose**: December is crowded with holiday content, so January offers a less competitive window for visibility.

2. June
   - **Seasonal Appeal**: Summer months see increased activity in outdoor events, festivals, and vacations, creating a high demand for new music that fits the season's upbeat mood.
   - **Engagement Opportunities**: Audiences are more active during this period, making it an ideal time to maximize reach and engagement.

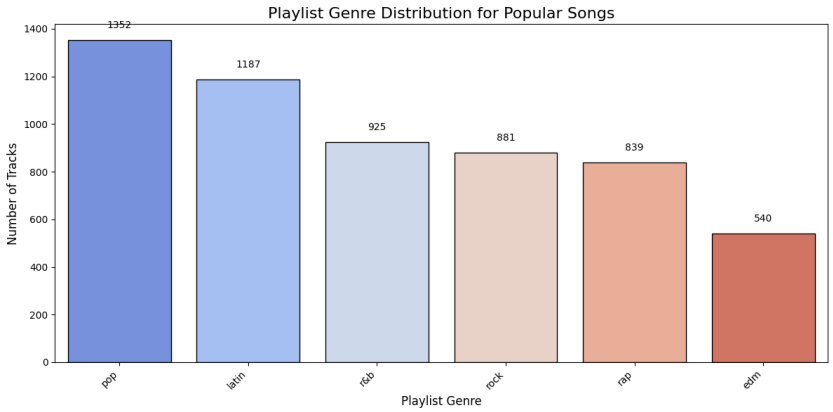
3. November
   - **Holiday Influence**: November sets the stage for holiday-themed releases, playlists, and campaigns, making it a prime time for artists and labels to capitalize on seasonal demand.
   - **Gift-Giving Season**: Physical and digital music purchases often spike as people shop for gifts.

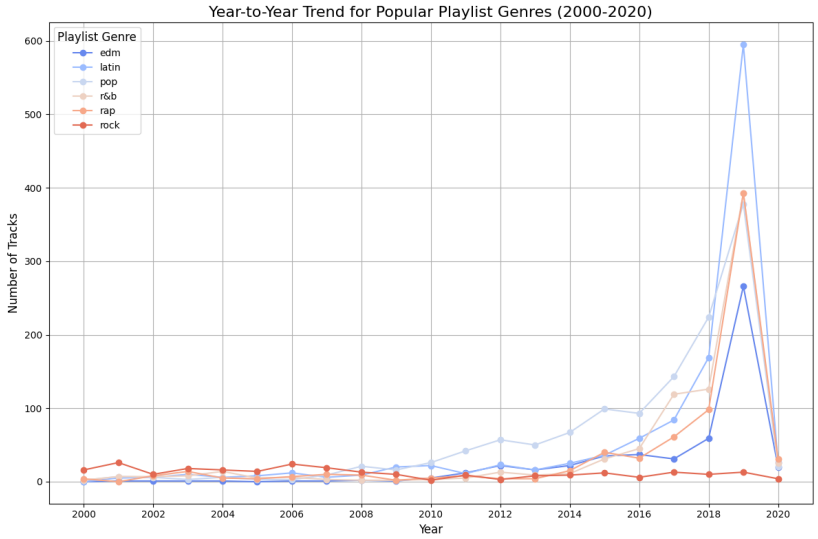


[Playlist Genres]

The genre distribution of popular songs is dominated by **pop, Latin, and R&B**, contrasting with the broader trend of EDM, rap, and pop. This reflects **the universal appeal of pop and R&B to a wide audience**, while the **rise of Latin music** highlights its growing influence and diverse fanbase, contributing to higher streaming numbers in recent years.

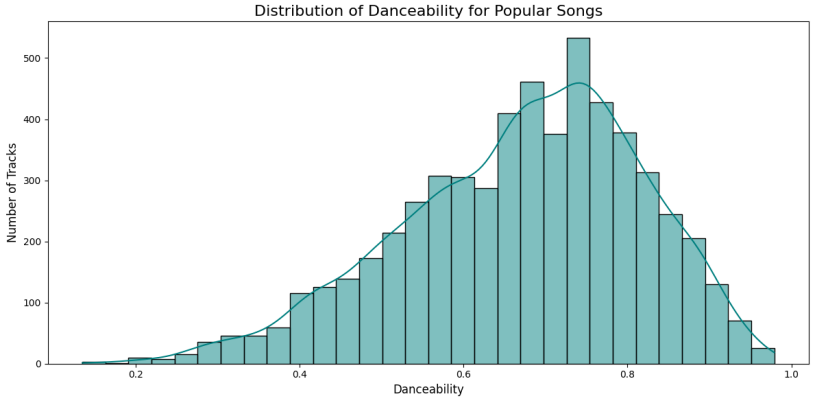
The more recent trend from 2018 to 2020 shows a **significant surge in the popularity of Latin music**, making the largest difference compared to other genres. This is followed by rap, pop, and R&B, which all experienced similar growth.

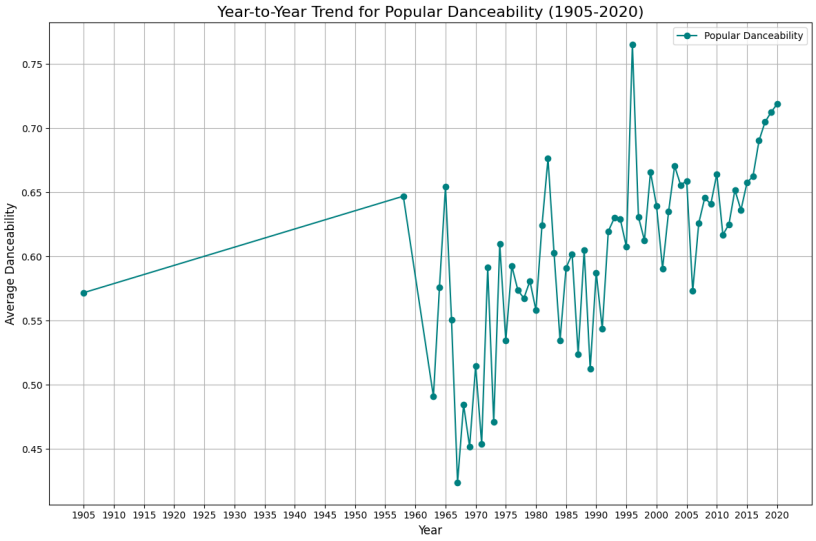




[Danceability]

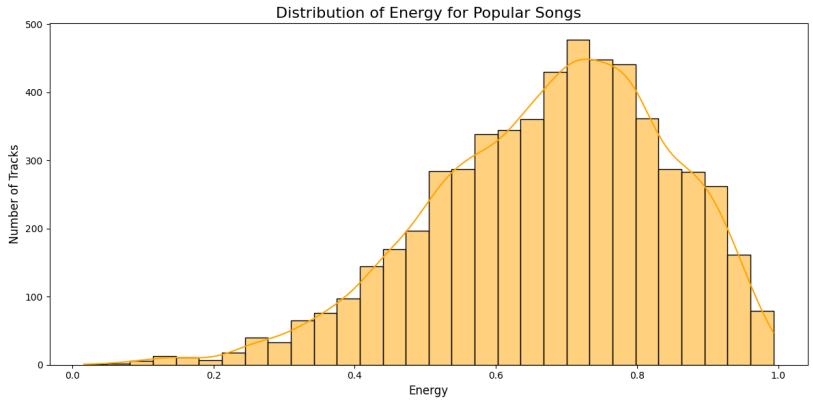
The danceability of popular songs aligns with the general music trend, mostly around 0.7, indicating a **rhythmically engaging beat.** The trend shows an increase in danceability over time, especially in the modern music landscape, highlighting a growing focus on creating danceable music.

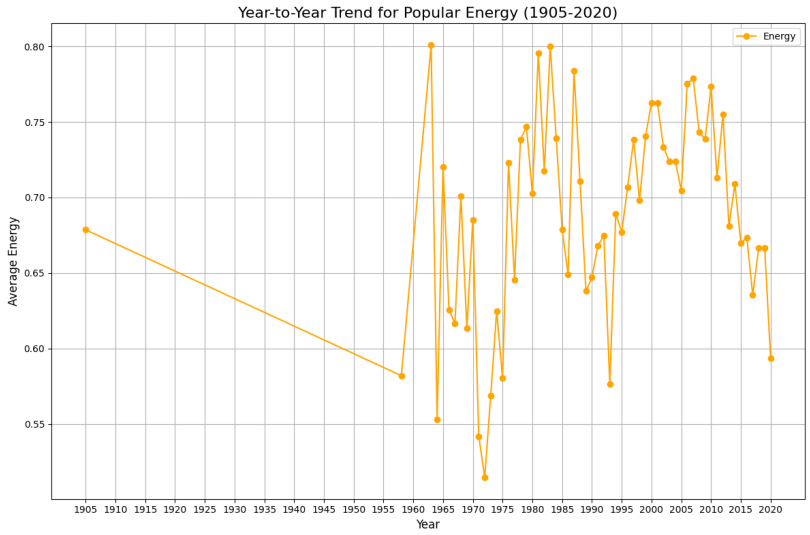




[Energy]

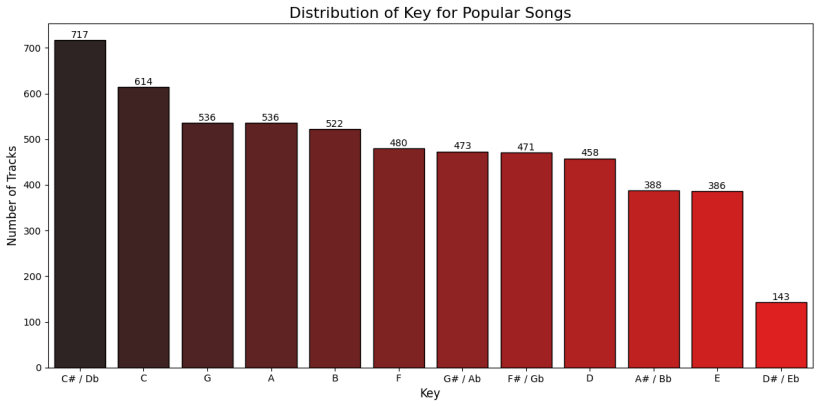
The graph shows that popular songs are more concentrated around 0.7 while the overarching tendency was widely spreaded out across largely 0.6 to 0.9. Throughout the years in Spotify history, the popualr songs have maintained the enrgy level of around 0.7.





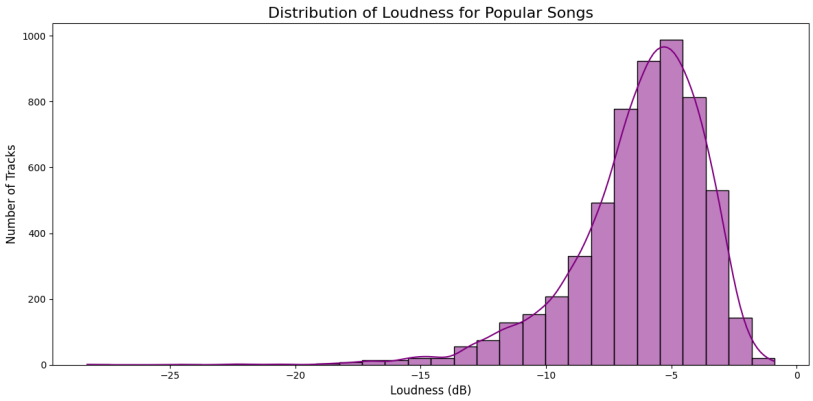
[Key]

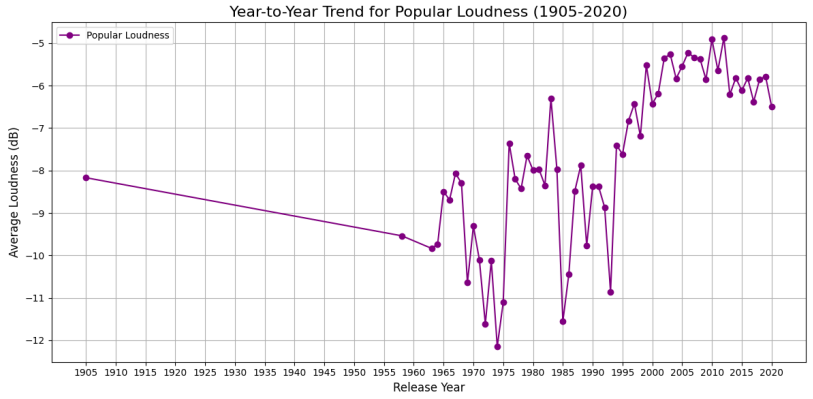
The key distribution for popular songs shows that the top three keys are **Db/C#, C, and G**. This **aligns with the observed trends in playlist genres, danceability, and energy**. Db Major, known for its depth and sophistication, complements expressive and introspective tracks. C Major, with its clarity and versatility, is well-suited to mainstream genres like pop and EDM. G Major, with its uplifting and resonant qualities, corresponds to the high energy (0.6–0.8) and danceability (0.6–0.8) trends, supporting the vibrant nature of genres like EDM, Latin, and rap.



[Loudness]

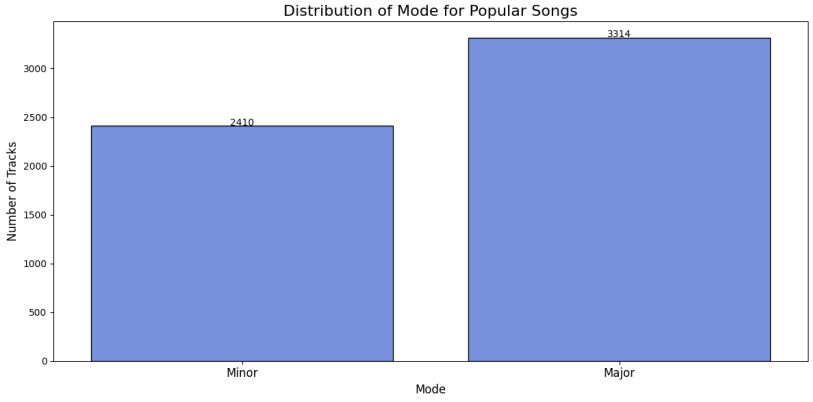
The **loudness of popular songs closely aligns with the overall trend on Spotify**. However, popular songs tend to exhibit a stronger concentration around -5 dB, reflecting the dominance of **high-energy, danceable genres like EDM, pop, and rap**. This consistency in volume levels may also be attributed to industry standards and the mastering process, where tracks are optimized to be loud enough without distortion or clipping. As a result, most tracks fall within this range to ensure consistent sound quality across various listening environments, whether on streaming platforms, radios, or headphones.





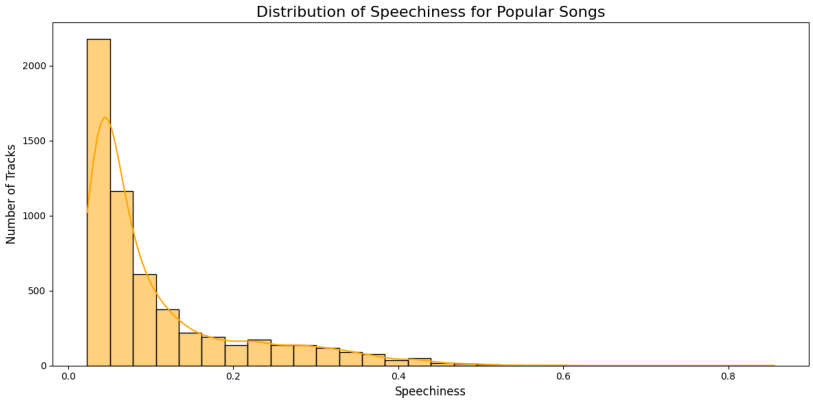
[Mode]

Most popular songs are in the major mode (58%), likely due to the dominance of **energetic and uplifting genres like EDM and pop** in the dataset. These genres, known for their **positive and high-energy sound**, often align with major mode characteristics. While danceability, energy, and loudness contribute to the overall feel of a track, they have less direct influence on the choice of mode.



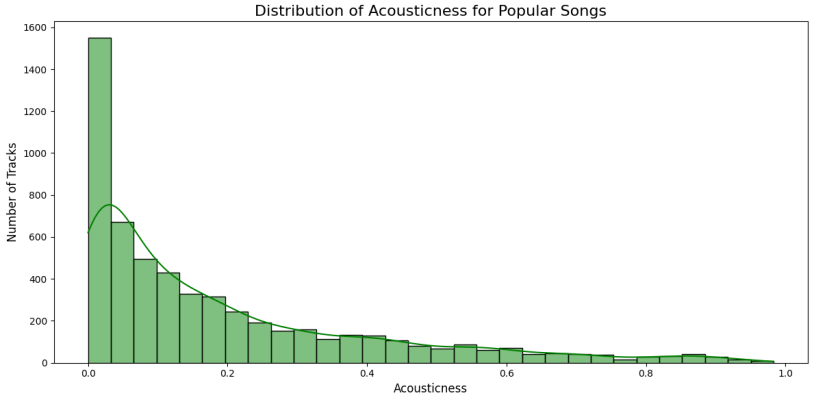
[Speechiness]

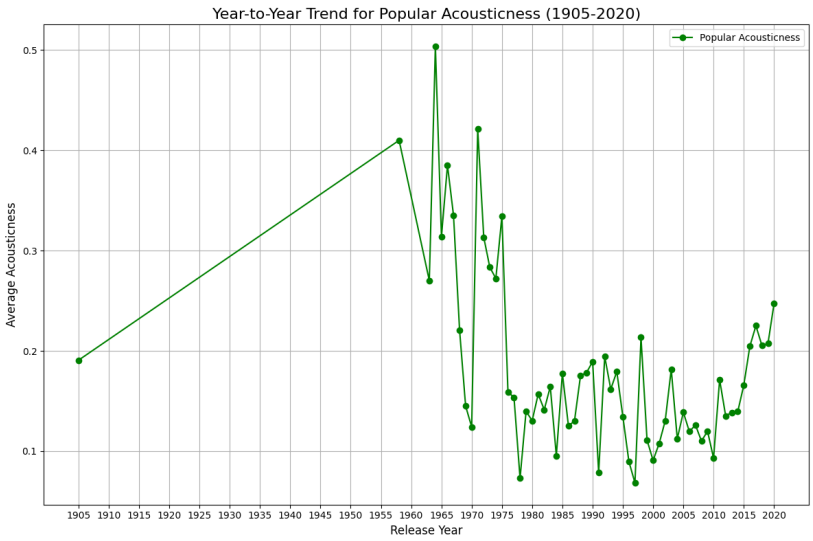
The distribution is heavily skewed toward low speechiness values, reflecting the dominance of traditional music tracks in Spotify's catalog. Most tracks have speechiness values below 0.33, indicating they **primarily consist of music with minimal speech-like elements**. Very few tracks exceed 0.66, representing purely spoken-word recordings such as audiobooks, podcasts, or poetry readings.



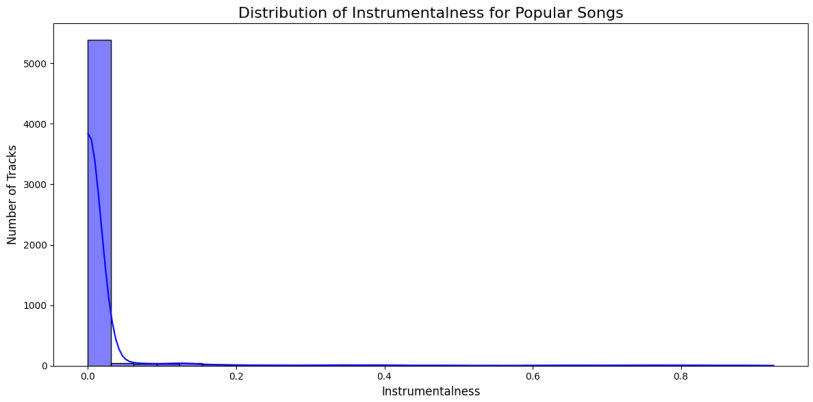
[Acoustiness]

The acousticness of popular songs follows a similar trend to the overall dataset. The majority of tracks have acousticness values close to 0.0, indicating that **most popular songs are primarily non-acoustic**. From 2015 to 2020, there is a slight increase in acousticness, peaking at around 0.25, though the change remains modest. **This trend is consistent with other factors such as playlist genres, danceability, and energy**, which tend to prioritize high-energy, rhythmic, and electronically produced tracks over acoustic elements.

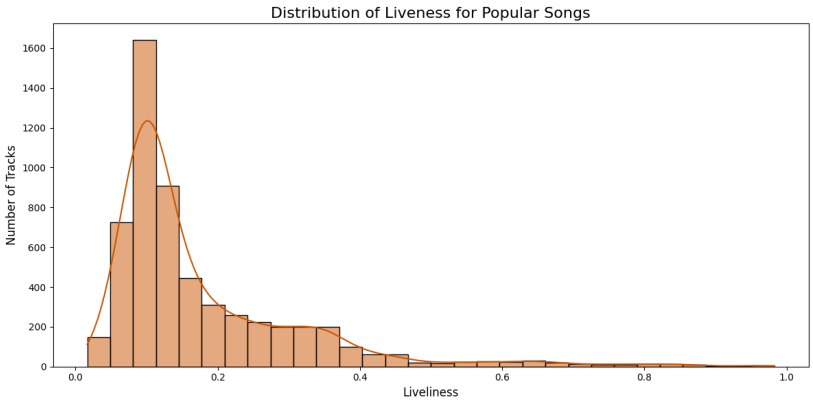


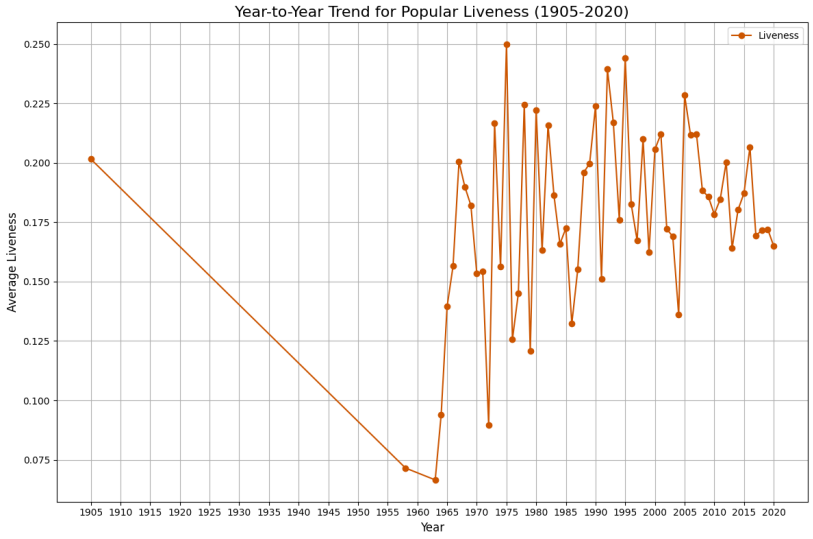


[Instrumentalness]
The high concentration near 0.0 suggests that most tracks in the dataset include **prominent vocal content**, such as singing, rap, or spoken word.



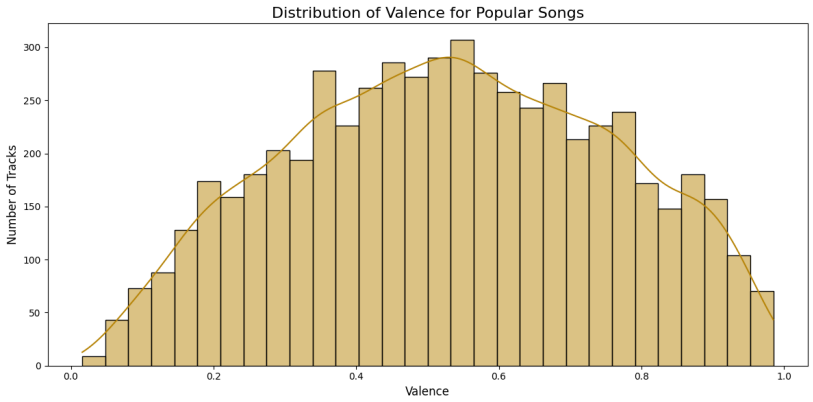
[Liveness]
The distribution of liveness for popular songs is heavily concentrated at **lower values**, especially around 0.1, indicating that most tracks are **studio recordings with little to no audience or live elements**. In recent years, however, the trend has shown increased fluctuation in liveness. Over the last five years, the liveness of popular songs has consistently decreased, reinforcing the dominance of studio-produced tracks.

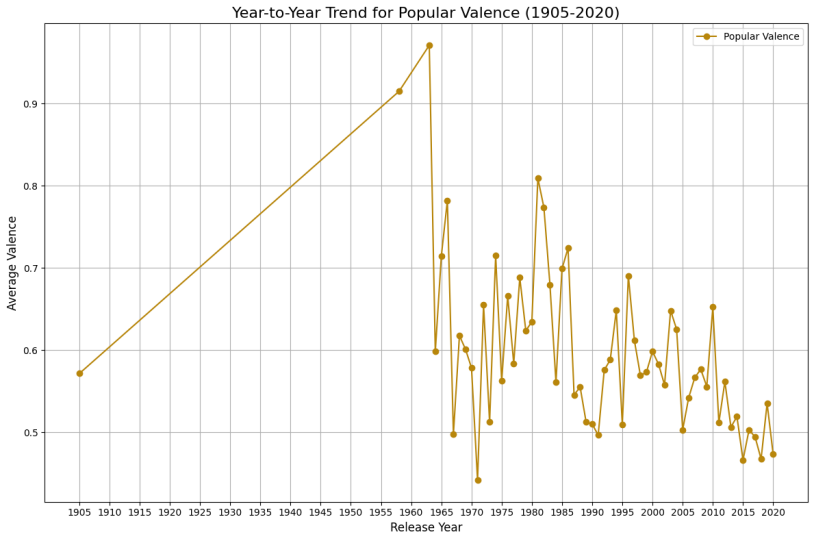




[Valence]

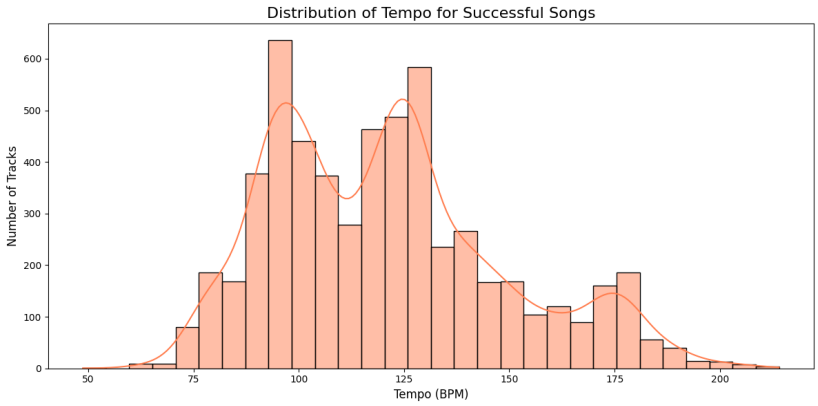
Unlike the overall trend, popular songs tend to be slightly more **concentrated in the range of neutral to more positive songs**. Similarly, the year-to-year trend for the last five years has remained relatively stable with a generally neutral or slightly positive emotional tone and valence values consistently range between 0.45 and 0.53.

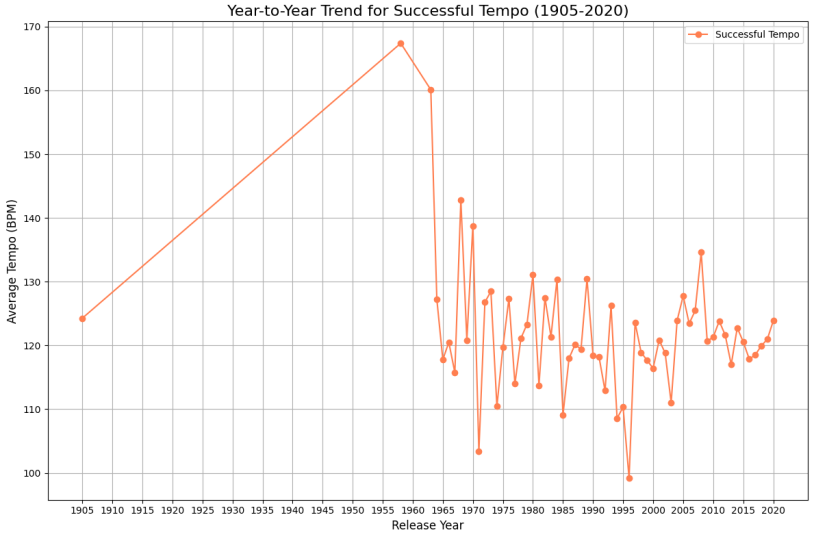




[Tempo]

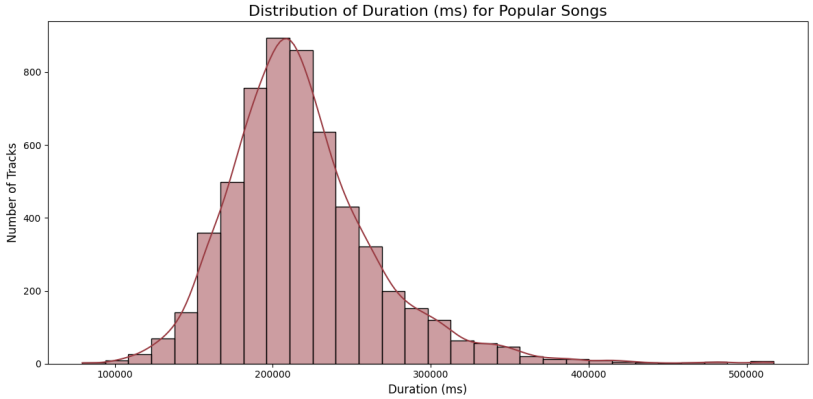
The tempo of popular songs is mostly concentrated around 95 to 130 BPM, which represents the most common tempo for successful tracks. In the past five years, there has been a **slight shift towards a higher tempo**, with tracks aiming for a range closer to 120 BPM. This trend reflects the increasing emphasis on **more upbeat and energetic rhythms in modern popular music**.

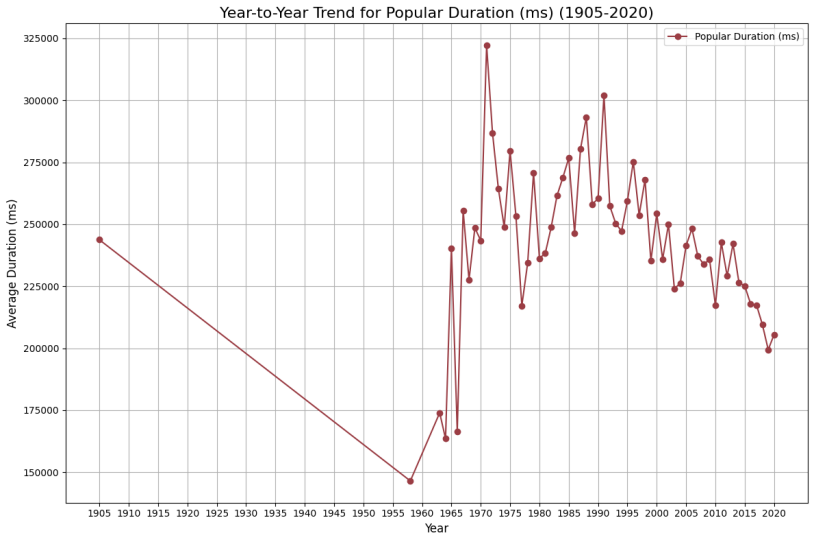




[Duration Time]

The **trend for popular songs' duration is almost identical to the overall music trend**, with most tracks revolving around 200,000 milliseconds (**3 minutes and 20 seconds**). The trend throughout musical history shows a steady decline in track durations since 2016, reflecting a broader shift in modern media consumption, where **shorter content is prioritized to capture attention quickly** in a highly competitive digital environment, contrasting with traditional radio and streaming norms.



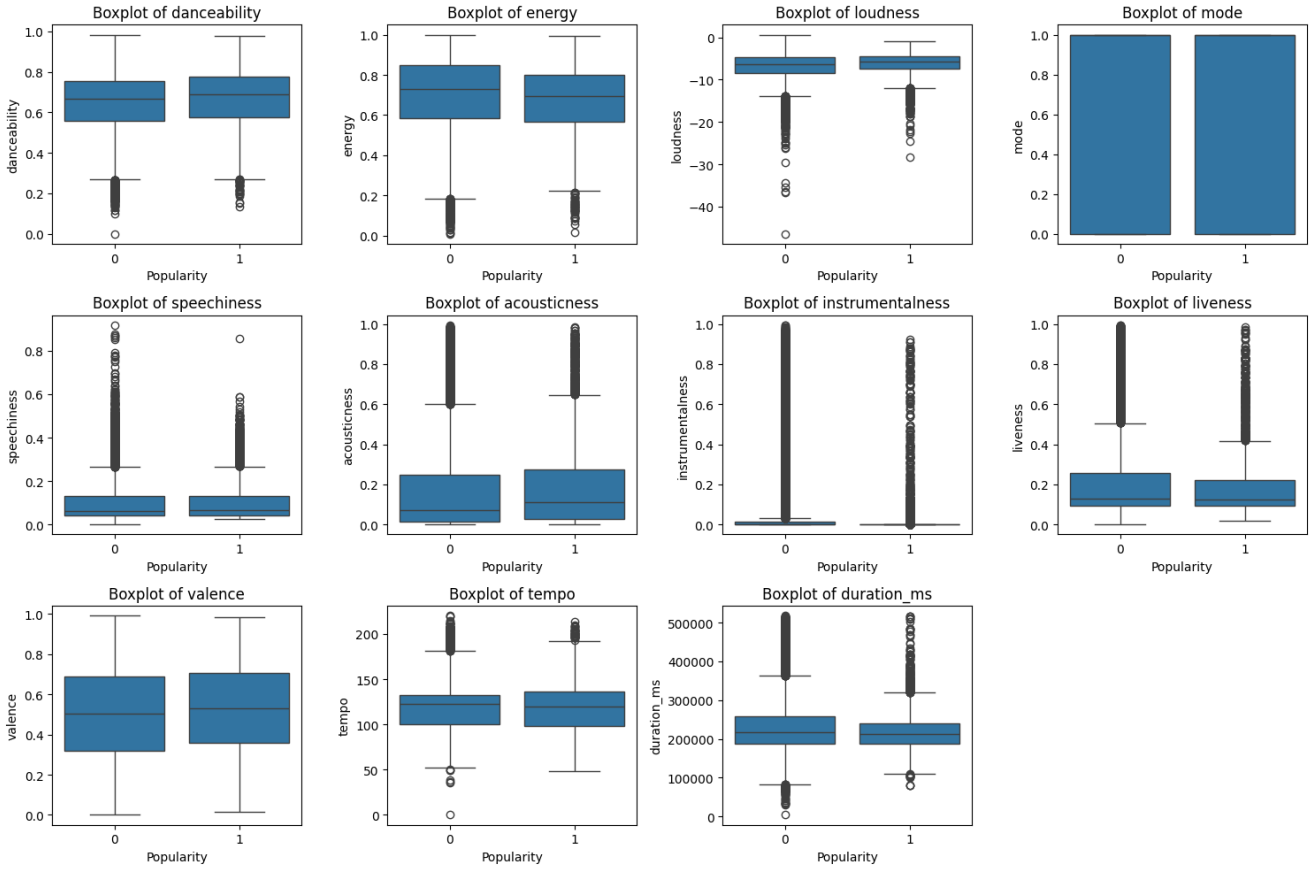


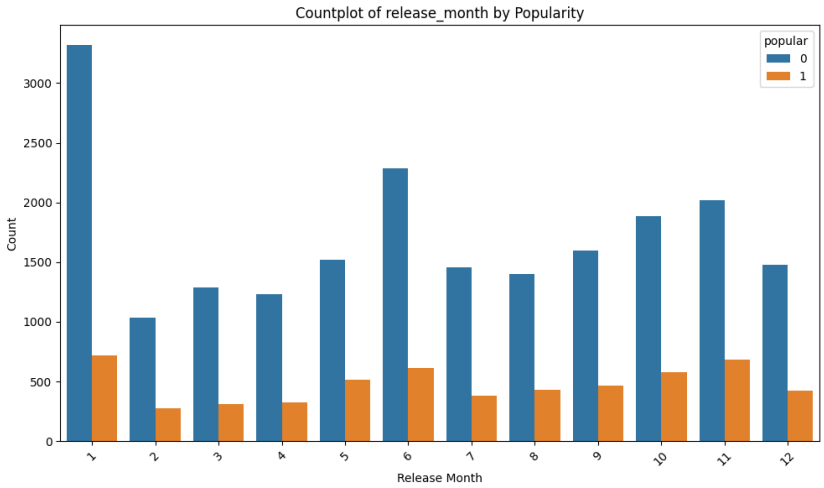
[Popular vs. Non-popular songs]

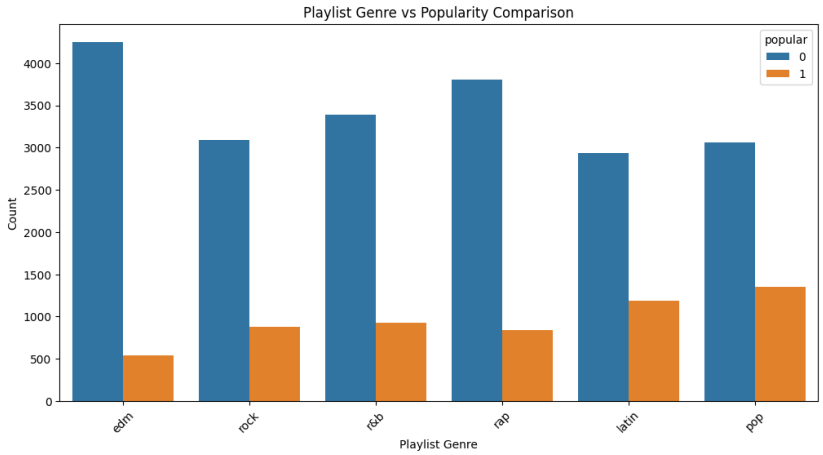
For numerical variables, these boxplot comparisons show that while there are some general trends, there are also a lot of overlapping features. The key takeaway is that popular songs tend to have higher energy, danceability, and a focus on vocals, while non-popular songs are more likely to be instrumental or acoustic, with a greater diversity in terms of tempo and loudness.

Regarding categorical variables, the release month shows that both popular and non-popular songs share similar release trends, particularly in January, June, and November, which aligns with the overall distribution. This suggests that release month may not be a critical factor in determining whether a song will become popular.

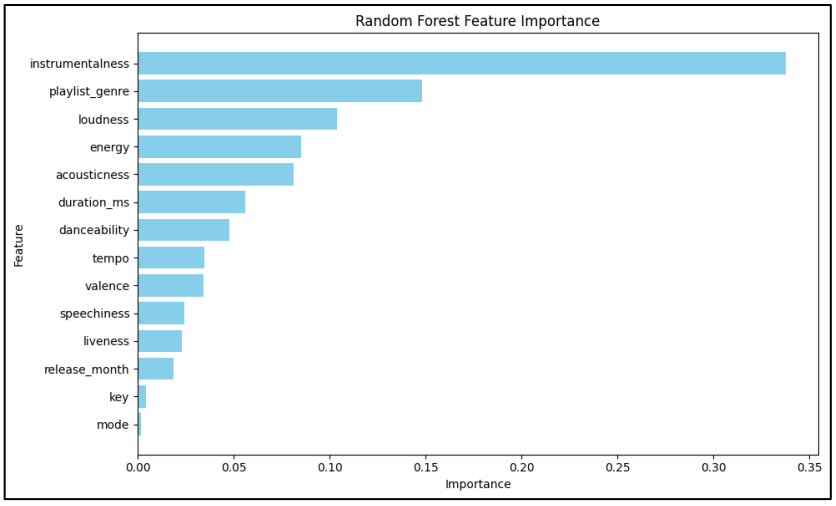
The playlist genre, however, shows a clear distinction between popular and non-popular songs. Non-popular songs are predominantly from genres like EDM, Rap, and R&B, whereas popular songs are more likely to be from genres such as Pop, Latin, and R&B. This indicates that playlist genre may be a more relevant factor in distinguishing songs with the potential for higher popularity.







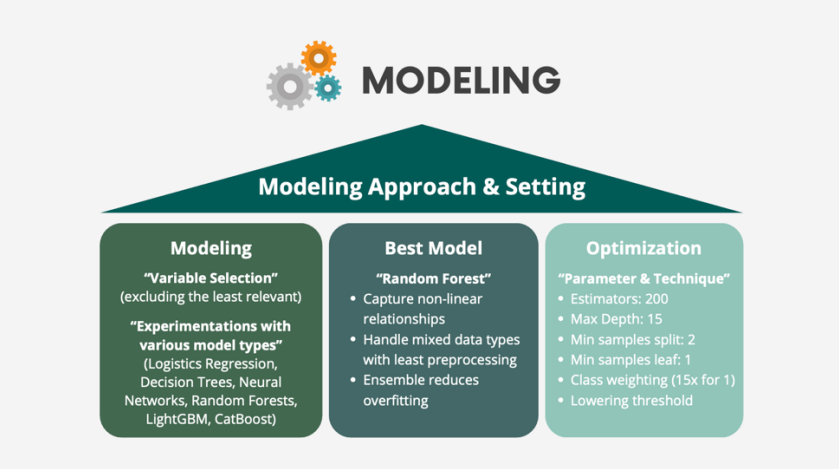
Based on the feature importance analysis, key variables contributing to song success include instrumentalness (34%), playlist genre (15%), loudness (10%), and energy (9%), highlighting the significant role of musical intensity and genre. Other features such as acousticness (8%), duration minute (6%), and danceability (5%) also play a moderate role. In contrast, variables like tempo (3%), speechiness (2%), liveness (2%), release month (2%), key (0%), and mode (0%) showed minimal or no impact and were excluded from the modeling process. This analysis shows that the musical characteristics (instrumentalness, loudness, energy, etc.) and genre are the key drivers of popularity, while more technical features like tempo, speechiness, liveness, key, and mode have less influence.



Feature Importance:

	Feature	Importance
9	instrumentalness	0.34
1	playlist_genre	0.15
5	loudness	0.10
3	energy	0.09
8	acousticness	0.08
13	duration_ms	0.06
2	danceability	0.05
12	tempo	0.03
11	valence	0.03
7	speechiness	0.02
10	liveness	0.02
0	release_month	0.02
4	key	0.00
6	mode	0.00

The Random Forest model was selected for predictive modeling. This tree-based ensemble method is well-suited for capturing complex, non-linear relationships and handling mixed data types without extensive preprocessing. Songs' popularity can depend on a combination of factors (e.g., energy and loudness interacting with danceability), which Random Forest handles well through its decision tree-based approach. Its ability to aggregate predictions across multiple decision trees reduces overfitting, ensuring robust generalization to unseen data.

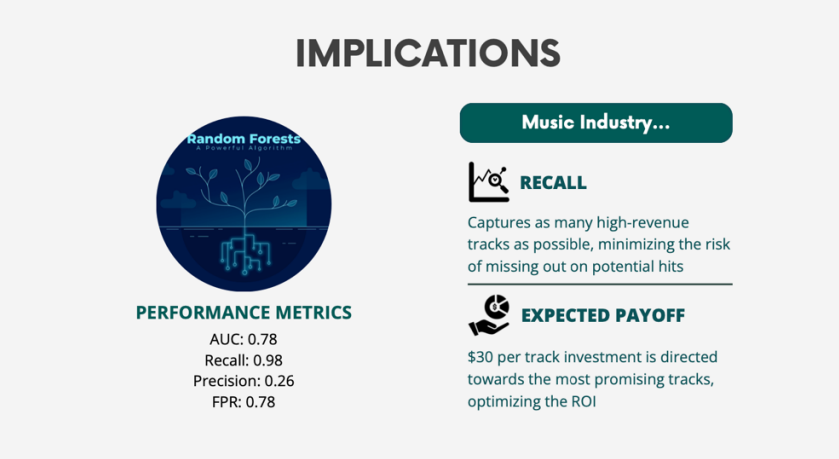


To maximize the model's performance, three approaches were employed. After testing various parameters, it was found that simpler parameter settings led to higher metrics and expected payoff. As a result, in this model, the number of estimators was set to 200 and the max depth to 15, with default values used for the minimum samples split and minimum samples leaf. Class weighting was also applied to address the class imbalance, assigning a higher weight of 15 to popular songs to improve recall. Then the decision threshold was lowered to 0.2 to prioritize recall, ensuring that 98% of actual popular songs were correctly identified, though this resulted in a higher False Positive Rate (FPR) of 78%. The model achieved an AUC of 0.78, indicating reasonable performance in distinguishing between popular and non-popular songs.

```
Random Forest AUC: 0.7772117731963107
Random Forest Recall (with Class Weighting and Adjusted Threshold): 0.9785879629629629
Random Forest Precision (with Class Weighting and Adjusted Threshold): 0.2595548733691481
Random Forest False Positive Rate (FPR) (with Class Weighting and Adjusted Threshold): 0.7841352405721717

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.22      0.35      6152
           1       0.26      0.98      0.41      1728

    accuracy                           0.38      7880
   macro avg       0.62      0.60      0.38      7880
weighted avg       0.82      0.38      0.37      7880


Confusion Matrix:
[[1328 4824]
 [  37 1691]]
```

After obtaining the probability output from the model, it was time to evaluate Universal Music's investment strategy, as described in the Case Description in the Appendix. With a cutoff of 0.22, the model generated an annual expected payoff of $155,480,000 on testing data. An adjustment was made to account for a 20% probability that initially unpopular songs could later gain popularity. This adjustment increased the expected payoff to $290,304,000 at a cutoff of 0.10, representing an improvement of $135,250,000. This refinement reflects a more optimistic forecast by acknowledging the potential for delayed success and reducing the risk of misclassifying songs as unpopular.

In the real business world, particularly in the music industry, investment in promotions is a significant financial commitment. For a company like Universal Music, promoting a track involves not only marketing efforts but also contracting the song, which requires substantial investment. This is where the metrics of recall and expected payoff become critically important. Maximizing recall ensures that the company captures as many high-revenue tracks as possible, minimizing the risk of missing out on potential hits. By also focusing on expected payoff, the company ensures that its $30K per track investment is directed towards the most promising tracks, optimizing the return on investment. Therefore, balancing recall to identify potential hits and using expected payoff to allocate the promotion budget effectively is critical for making profitable decisions and driving revenue in the competitive music industry.



Overall, this project highlights the importance of aligning data-driven strategies with evolving trends in the music industry. In such a fast-paced industry, where consumer tastes and trends change rapidly, understanding what is recent or even anticipating future trends is crucial. This is evident in the feature importance analysis, which shows that while most variables contribute to the story, their significance alone may not be strong enough to drive critical financial investment decisions for large corporations like Universal Music. Therefore, in industries like entertainment, it is essential to complement quantitative analysis with strong domain knowledge. By integrating deep industry insights with data-driven strategies, companies can make more informed decisions that align with both market trends and consumer preferences.

Appendix

1. Data Description

• **Genre**: The genre of the song (rap, rock, country, etc.).
• **Artist_name:** The name of the artist of the song.
• **Track_name**: The name of the song.
• **Track_id**: A unique id for the song.
• **Release_year**: Year released.
• **Release_month**: month of the year released.
• **Popular**: 1 if the song is popular and 0 if not. Popular is evaluated based on various factors, including Spotify popularity score. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past.
• **Acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
• **Danceability**: How suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
• **Duration_ms**: The duration of the track, in milliseconds.
• **Energy**: A measure from 0.0 to 1.0 that represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
• **Instrumentalness**: Detects whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
• **Key**: The key of the song.
• **Liveliness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
• **Loudness**: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entiretrack and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
• **Mode**: Whether the song is in a major or minor key.
• **Speechiness**: Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
• **Tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
• **Valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

2. Case Description

Investing in music promotion is essential for artists and labels looking to reach wider audiences, stand out in a competitive market, and build lasting fan loyalty. In today's streaming-driven music industry, effective promotion helps artists secure placements on popular playlists, gain traction on social media, and attract the attention of fans and industry influencers. Targeted marketing and digital tools can significantly enhance an artist's visibility, deepen their connection with listeners, and ultimately drive streams, ticket sales, and revenue. As the music landscape continues to evolve, strategic promotion is critical for maintaining growth and relevance.

Equally important, however, is identifying music with strong potential for success. Investing in tracks or artists with high potential increases the likelihood of a successful promotional campaign, leading to better returns on investment. Beyond traditional promotional strategies, companies like Universal Music recognize the growing importance of data in shaping music promotion.

While data now plays a pivotal role in the industry, many organizations are still lagging in effectively using it. To address this gap, Universal Music has brought you on board as a data analyst to help them leverage Spotify's extensive data to their advantage. By analyzing this data, the company aims to identify the characteristics of successful songs, optimize promotion strategies, and better support their artists.

Based on the recent popular songs data and historical data, the company estimated that if the track can become popular within two months of release, the track can generate on average 150K annual revenue from Spotify streaming, and an unpopular song generates on average 20K annual revenue. (For simplicity, you may assume all songs labeled popular in the data sets became popular within two months of release.) The annual investment budget is 30K per piece of music. The ideal situation is that the company can perfectly identify and contract music that will be popular within two months of release. However, the company understand that this is difficult to reach perfection given music, as art, often has unexpected outcome. The company would like to see if there is any statistical learning model that can help improve the prediction.
2a. Given this information above, what model would you recommend and how would you apply the model to help select music? Please include quantitative outcome such as annual expected payoff on testing data.

2b. With further consideration and experience, the company would like to add an adjustment to the model proposed in part 2a. Specifically, even if newly promoted music does not become popular immediately, there is a 20% chance that, due to promotion, the song will gain popularity a few months later. Please suggest how to revise the model to incorporate this condition and calculate the expected payoff.

Unfortunately, there is no reliable estimate for how long it may take for these songs to become popular. For simplicity, the company prefers to assume no additional investment if a song has not become popular within one year of release. Additionally, for simplicity, assume that once a song is labeled as popular, it remains so for the duration of this study.