MUSIC POPULARITY ANALYSIS

DSO 528 FINAL TEAM PROJECT

Blended Data Business Analytics for Efficient Decisions

PRESENTED BY TEAM 3

Yeunbin Cho, Sherleen Lee, Yangruiqi Li,
Jesslyn Noorjono, and Stanley Toh

TABLE OF CONTENTS

PROJECT OVERVIEW

01

Objective

Identify the key characteristics of successful songs to help Universal music enhance its promotion strategies

Approach

- Understanding the music industry landscape
- Determines key features driving past success
- Develops models to predict song popularity
- Evaluating investment returns

DATA DESCRIPTION

02

DESCRIPTION OF THE DATA

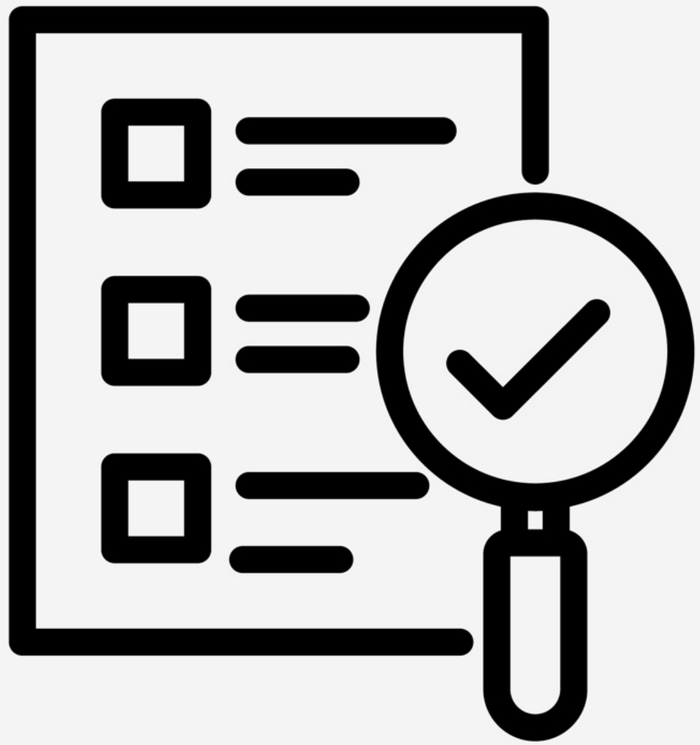
Spotify 1: Long-term trends with 26,266 entries and 19 columns, featuring historical songs released at least two months before March 2020

Spotify 2: recent popular songs, with 952 entries and 17 columns, providing data updated through 2023 to assess the relevance of historical trends in the current market.

Categorical Variables: Genre, Artist_name, Track_name, Track_id, Release_year, Release_month, Key, Mode

Numerical Variables: Acousticness, Danceability, Duration_ms, Energy, Instrumentalness, Liveliness, Loudness, Speechiness, Tempo, Valence, Stream

Target Variables: Popular (y)



DESCRIPTIVE ANALYSIS

Spotify 1: Numerical variables

	Field Name	Field Type	# Records Have Values	\
danceability	danceability	Numeric	26266	
energy	energy	Numeric	26266	
loudness	loudness	Numeric	26266	
mode	mode	Numeric	26266	
speechiness	speechiness	Numeric	26266	
acousticness	acousticness	Numeric	26266	
instrumentalness	instrumentalness	Numeric	26266	
liveness	liveness	Numeric	26266	
valence	valence	Numeric	26266	
tempo	tempo	Numeric	26266	
duration_ms	duration_ms	Numeric	26266	

	% Populated	# Zeros	Min	Max	Mean
danceability	100.0%	1	0.00000	0.983	0.654880
energy	100.0%	0	0.00814	1.000	0.698126
loudness	100.0%	0	-46.44800	0.642	-6.733465
mode	100.0%	11439	0.00000	1.000	0.564494
speechiness	100.0%	1	0.00000	0.918	0.106995
acousticness	100.0%	1	0.00000	0.994	0.175092
instrumentalness	100.0%	9666	0.00000	0.994	0.084425
liveness	100.0%	1	0.00000	0.994	0.190961
valence	100.0%	1	0.00000	0.991	0.509772
tempo	100.0%	1	0.00000	220.252	120.929737
duration_ms	100.0%	0	4000.00000	517810.000	225878.447232

	Standard Deviation	Most Common
danceability	0.145427	0.733
energy	0.180970	0.828
loudness	2.994345	-5.608
mode	0.495833	1.000
speechiness	0.100875	0.102
acousticness	0.219371	0.128
instrumentalness	0.224070	0.000
liveness	0.154650	0.111
valence	0.232955	0.516
tempo	26.962447	127.992
duration_ms	59631.215366	240000.000

Spotify 1: Categorical variables

	Field Name	Field Type	# Records Have Values	\
track_id	track_id	Categorical	26266	
track_name	track_name	Categorical	26263	
track_artist	track_artist	Categorical	26263	
release_year	release_year	Categorical	26266	
release_month	release_month	Categorical	26266	
playlist_genre	playlist_genre	Categorical	26266	
popular	popular	Categorical	26266	

	% Populated	# Zeros	# Unique Values	Most Common
track_id	100.0%	0	23184	7BKLCZ1jbUBVqRi2FVlTVw
track_name	100.0%	3	19547	Poison
track_artist	100.0%	3	9405	Martin Garrix
release_year	100.0%	0	63	2019
release_month	100.0%	0	13	1
playlist_genre	100.0%	0	6	edm
popular	100.0%	20542	2	0

DATA PREPARATION

Spotify 1

```
--- DF1 (Historical Data): Null Values and '#VALUE!' Entries ---
Null Values:
track_id            0
track_name          3
track_artist        3
popular             0
release_year        0
release_month       0
playlist_genre      0
danceability        0
energy              0
key                 0
loudness            0
mode                0
speechiness         0
acousticness        0
instrumentalness    0
liveness            0
valence             0
tempo               0
duration_ms         0
dtype: int64


Rows with '#VALUE!' entries:
track_id             0
track_name           0
track_artist         0
popular              0
release_year        24
release_month       24
playlist_genre       0
danceability         0
energy               0
key                  0
loudness             0
mode                 0
speechiness          0
acousticness         0
instrumentalness     0
liveness             0
valence              0
tempo                0
duration_ms          0
dtype: int64
```

Spotify 2

```
--- DF2 (Historical Data): Null Values and '#VALUE!' Entries ---
Null Values:
track_name           0
artist(s)_name       0
artist_count         0
released_year        0
released_month       0
released_day         0
streams              0
tempo                0
key                 95
mode                 0
danceability         0
valence              0
energy               0
acousticness         0
instrumentalness     0
liveness             0
speechiness          0
dtype: int64


Rows with '#VALUE!' entries:
track_name          0
artist(s)_name      0
artist_count        0
released_year       0
released_month      0
released_day        0
streams             0
tempo               0
key                 0
mode                0
danceability        0
valence             0
energy              0
acousticness        0
instrumentalness    0
liveness            0
speechiness         0
dtype: int64
```

- **Dropping Rows**
  - Not enough information
- **Missing/Incorrect Values: Manual filling**
  - release year and month (qualitative research)
  - key (Spotify API, songbpm.com website)
- **Conversion: Categorical -> Numerical**
  - Key mapping (industry standard)
  - Mode mapping (number matching)
- **Outliers**
  - Include them to capture unique trends and standout tracks that break the norm

DATA EXPLORATION

03

DATA SELECTION

For both EDA and modeling, **Spotify 1** is the primary dataset. Its key metric, popular, reflects **both the total number of plays and how recent those plays are**, offering a dynamic measure of a song's current success. In contrast, **Spotify 2**'s metric, streams, represents **cumulative success** up to the data collection point **without accounting for whether a song is currently trending.**

Merging is avoided to prevent inconsistencies due to differing focuses: recent plays (Spotify 1) vs. cumulative success (Spotify 2).

POPULAR RELEASE TIME

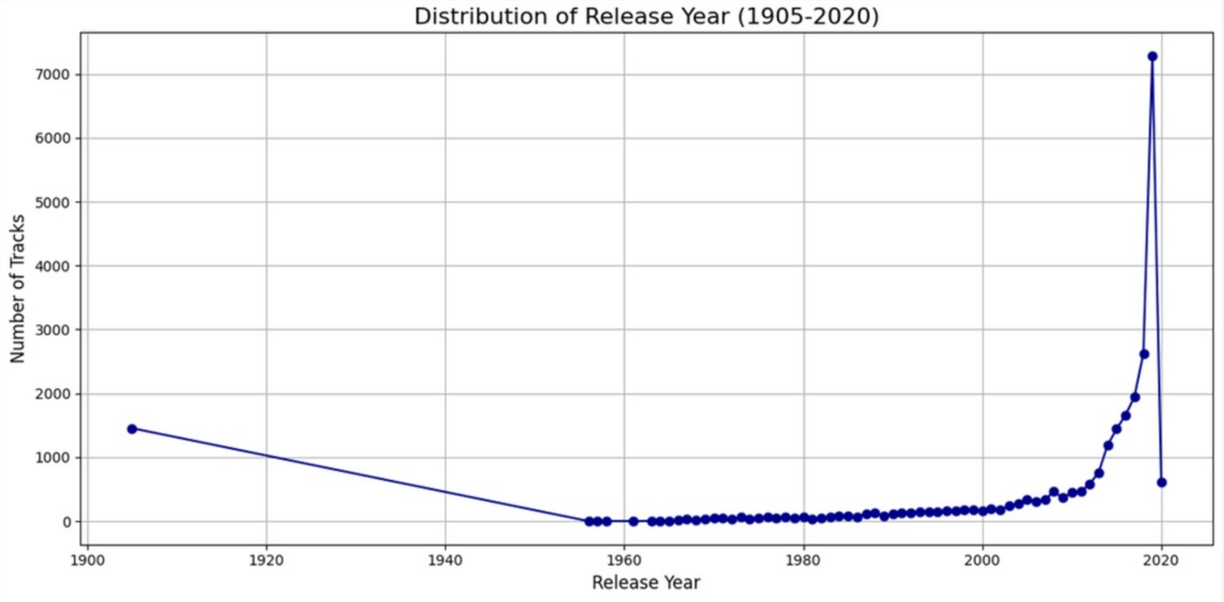
Release Year:

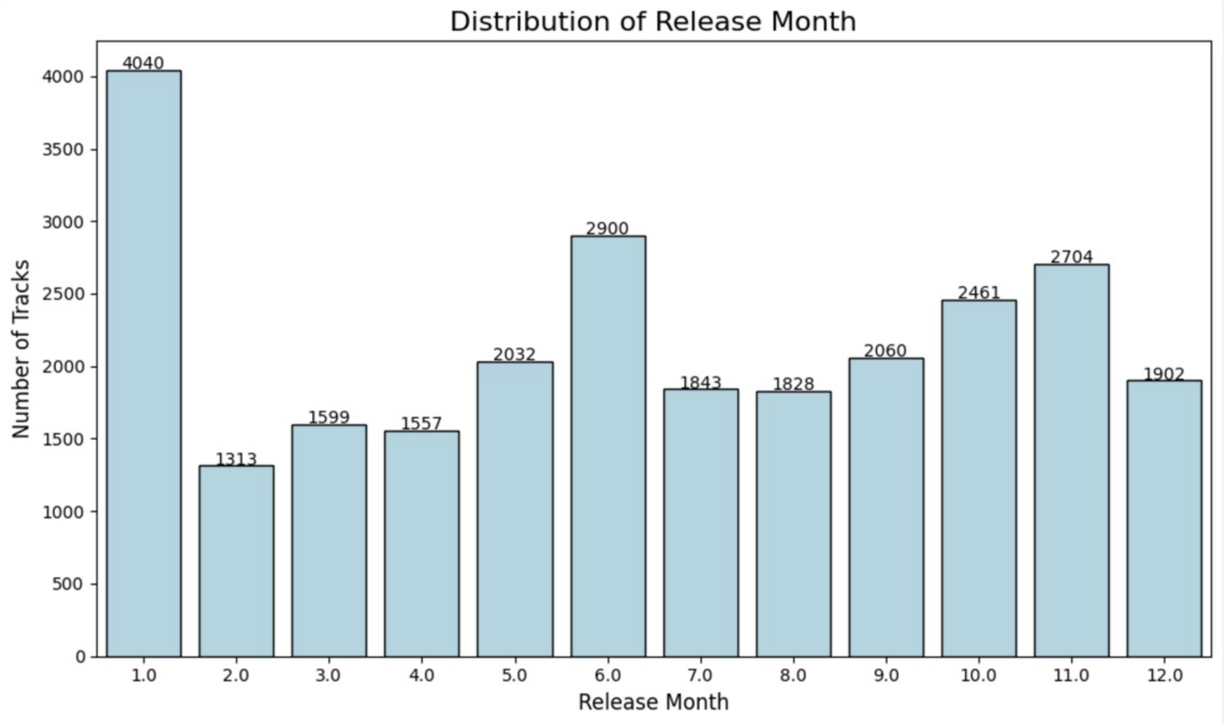
- From 1900 to 2000, the trend remained steady
- Starting in 2014, there was a noticeable upward trend, culminating in a dramatic peak in 2019 (due to growth of streaming and user-generated content platforms)

Release Month:

Top 3 Months are

- **January:** Industry Characteristics, Budget Allocations, Strategic Marketing Purpose
- **June:** Seasonal Appeal, Engagement Opportunities
- **November:** Holiday Influence, Gift Season





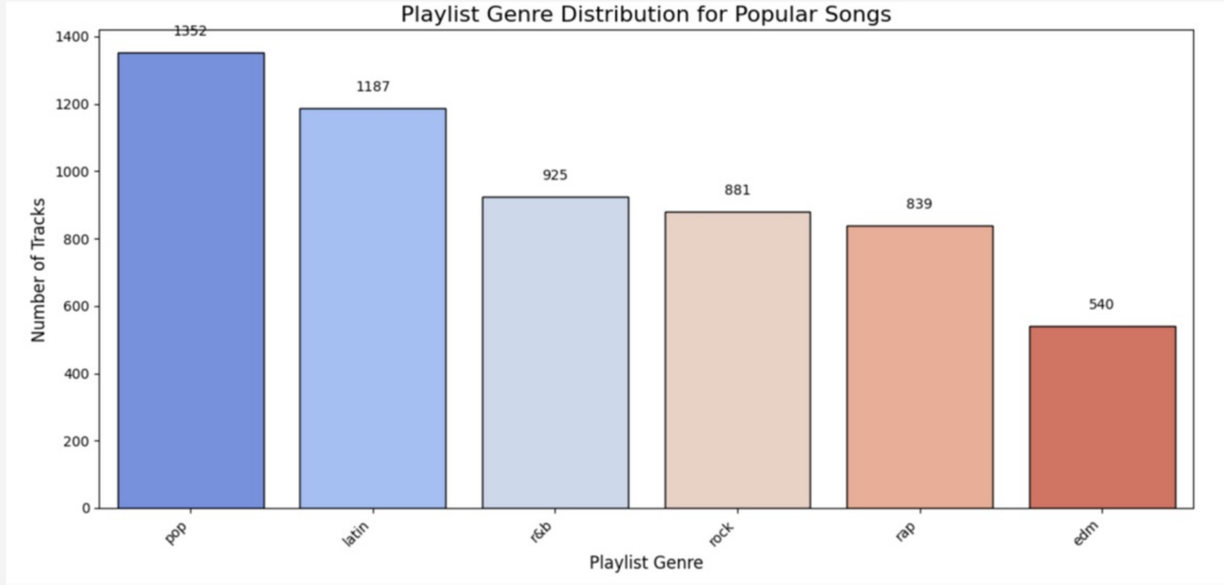
POPULAR PLAYLIST GENRE

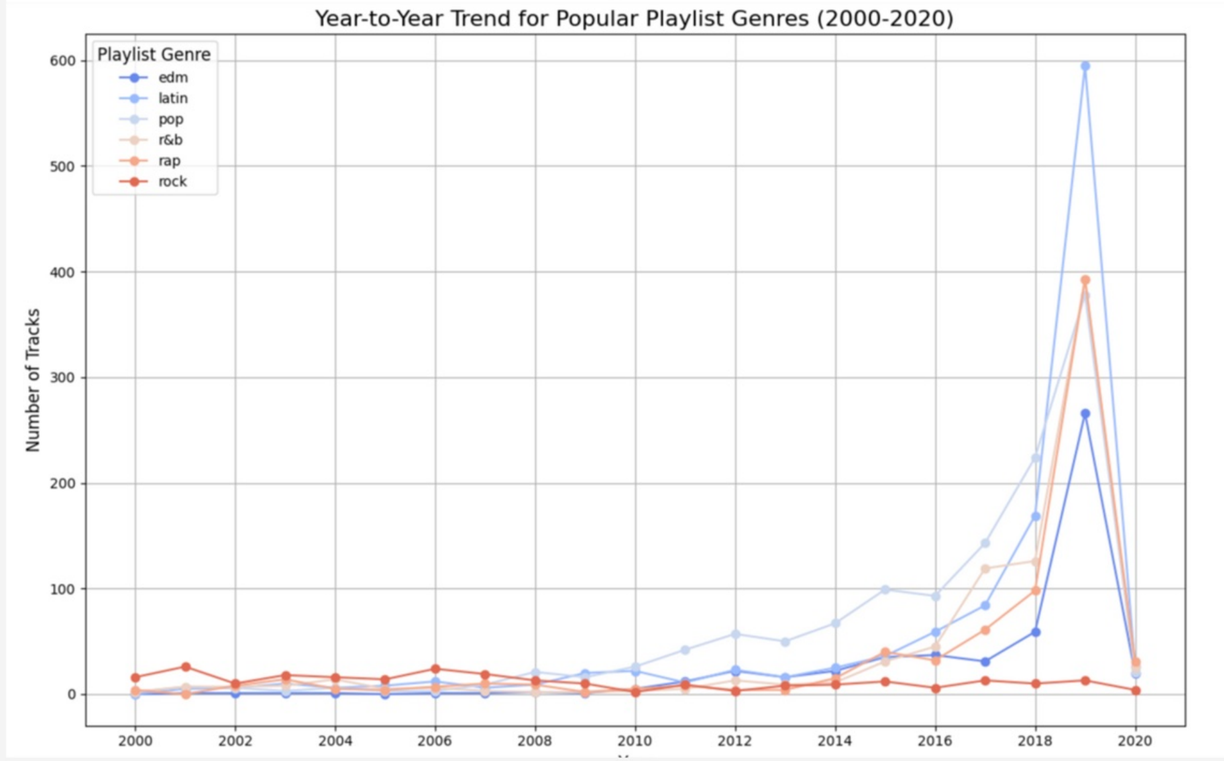
Key Points:

- **Dominant Genres:**
  - Popular songs are led by Pop, Latin, and R&B genres
  - Contrasts with broader trends of EDM, Rap, and Pop

- **Insights:**
  - Universal appeal of Pop and R&B to a wide audience
  - The rise of Latin music due to its growing influence and diverse fanbase

- **Recent Trends (2018–2020):**
  - General trend of noticeable rise in EDM, followed by Latin
  - Popular songs trend of significant surge in Latin music, followed by Rap and Pop





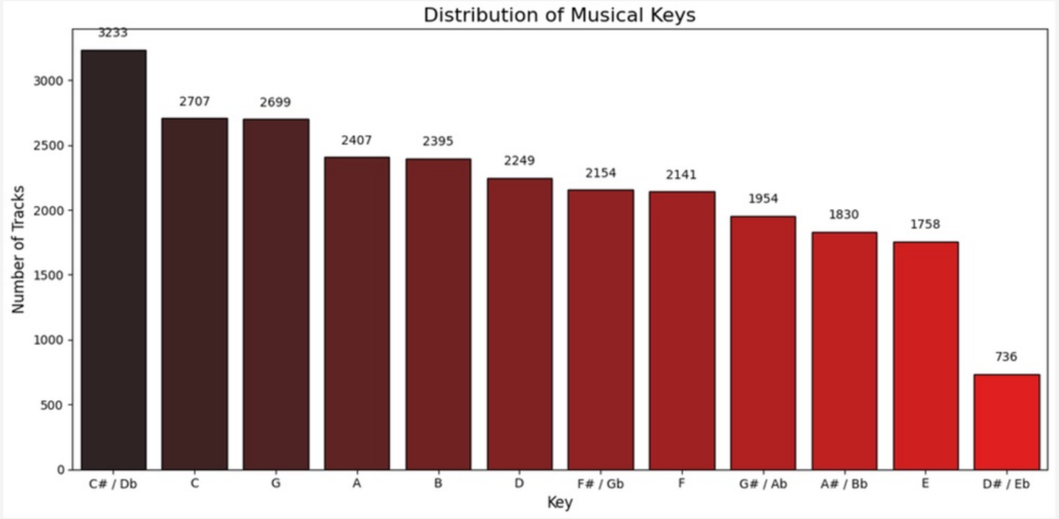
POPULAR KEY

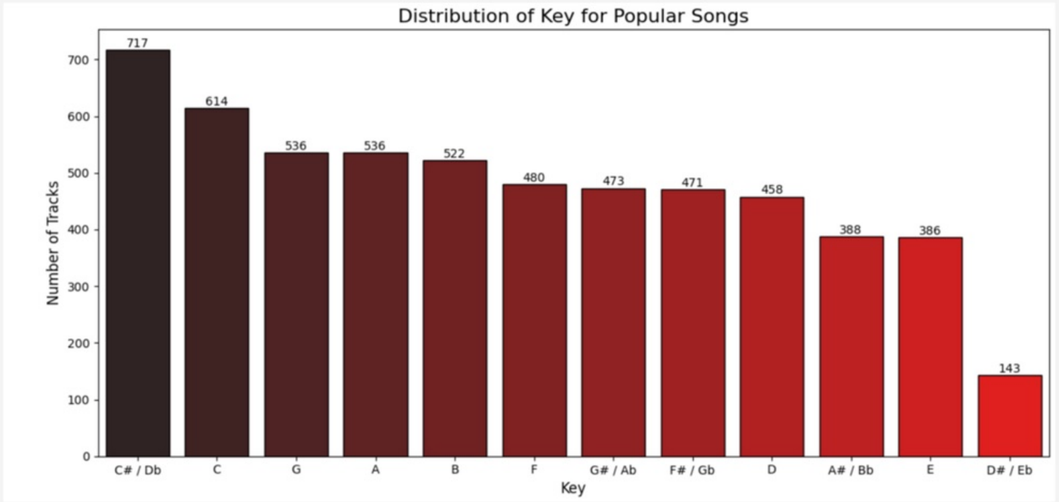
Key Points:

- **Top Keys:**
  - The most common keys for both general trend and popular songs are Db/C#, C, and G

- **Insights:**
  - Db Major: Known for depth and sophistication, complements introspective and expressive tracks
  - C Major: Offers clarity and versatility, ideal for mainstream genres like pop and EDM
  - G Major: Uplifting and resonant, aligns with high-energy and danceable trends

- **Genre Alignment:**
  - The preferred keys support vibrant and dynamic genres such as EDM, Latin, and rap, matching trends in energy (0.6–0.8) and danceability (0.6–0.8).





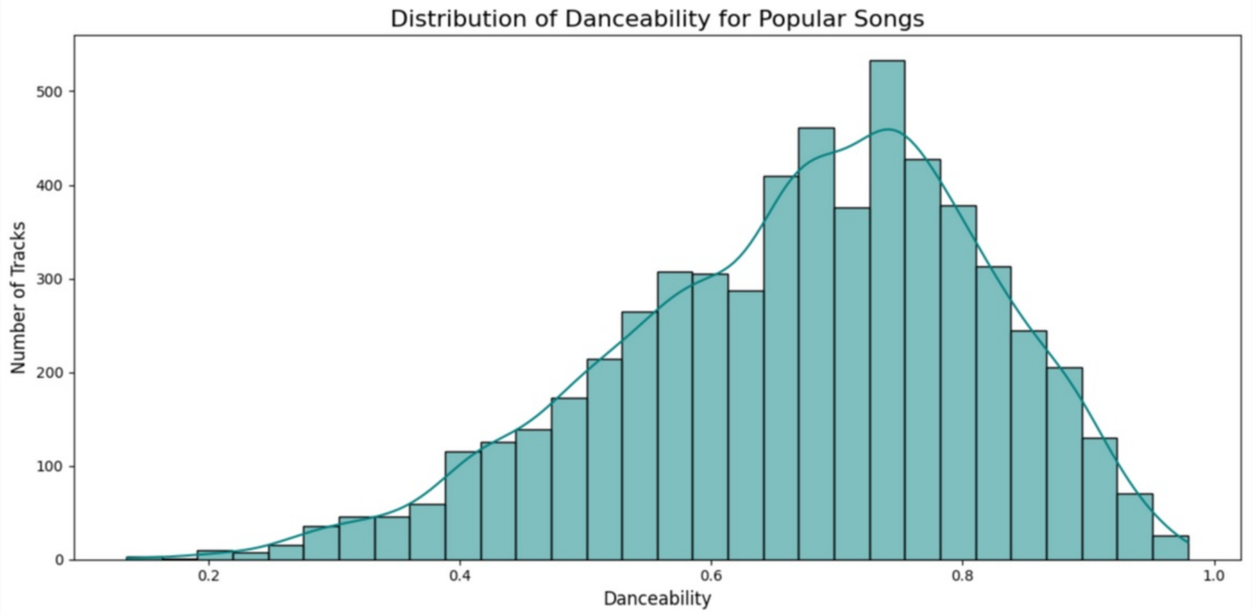
POPULAR DANCEABLITIY

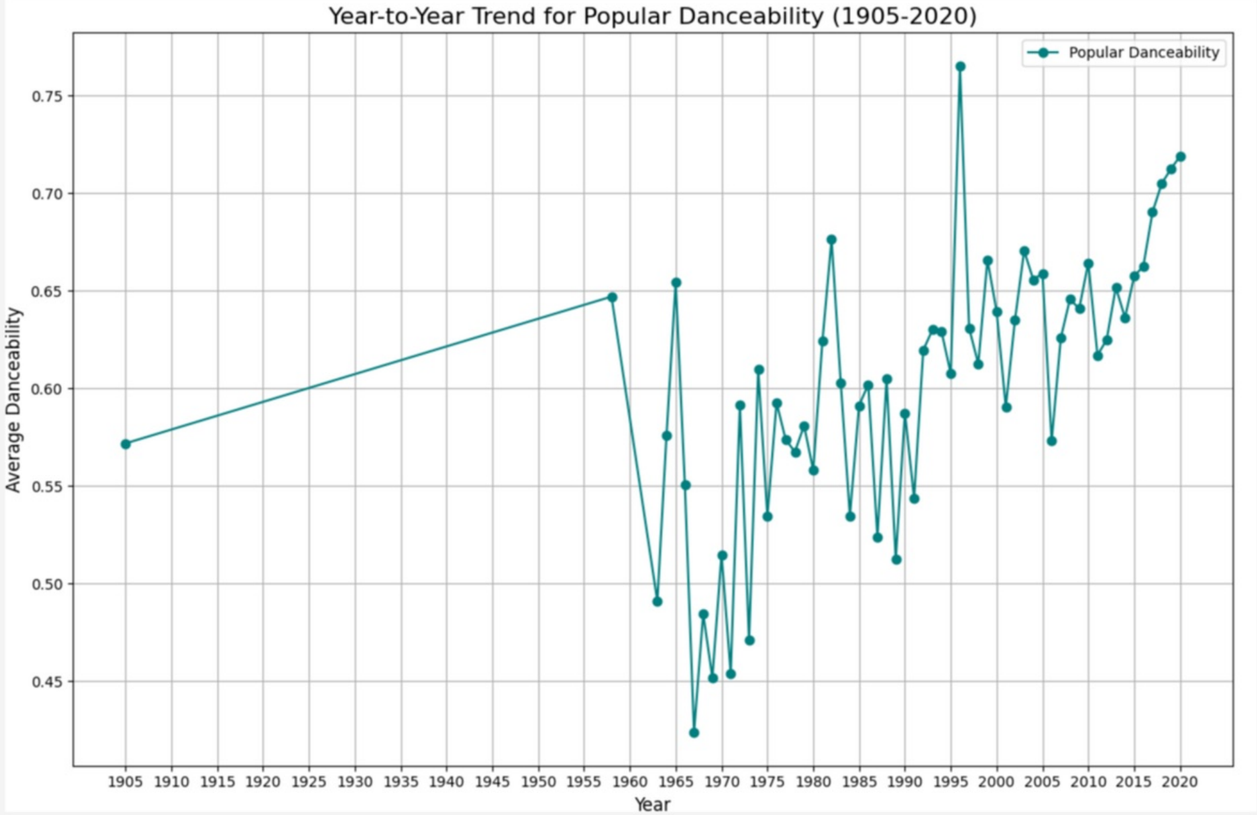
Key Points:

- **Typical Danceability:**
  - Most popular songs have a danceability score around 0.7, indicating a rhythmically engaging beat

- **Trends Over Time:**
  - Danceability has increased over the years, especially in the modern music era

- **Focus on Danceable Music:**
  - Highlights a growing emphasis on creating tracks suited for rhythmic and danceable experiences





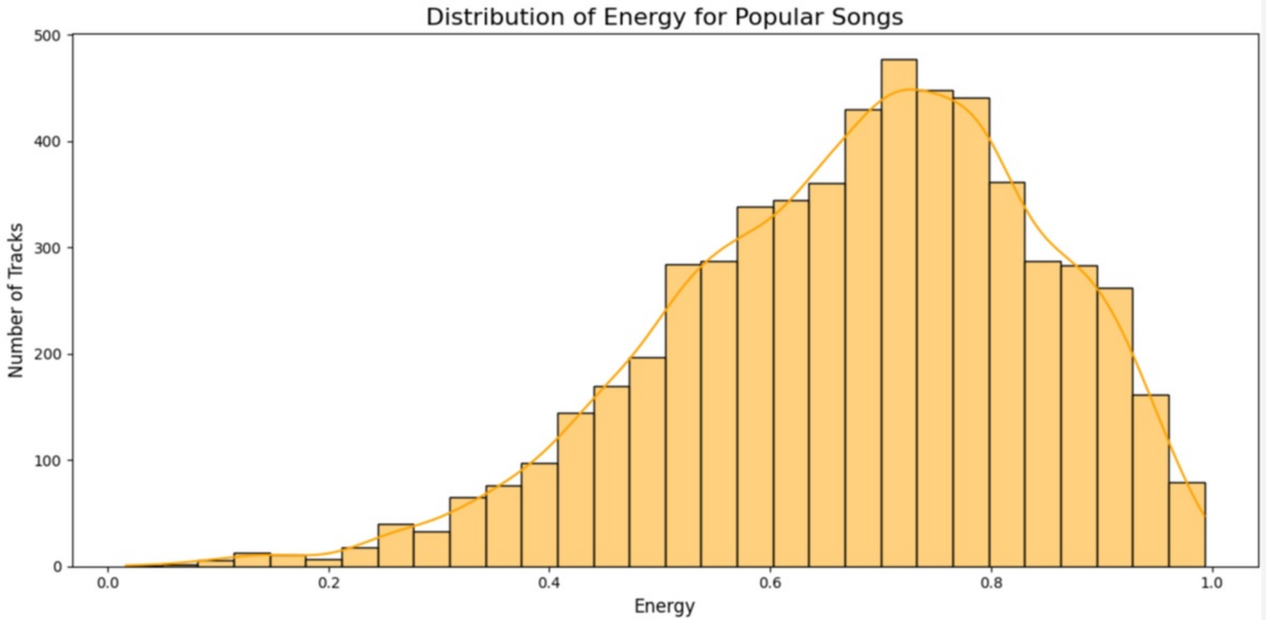
POPULAR ENERGY

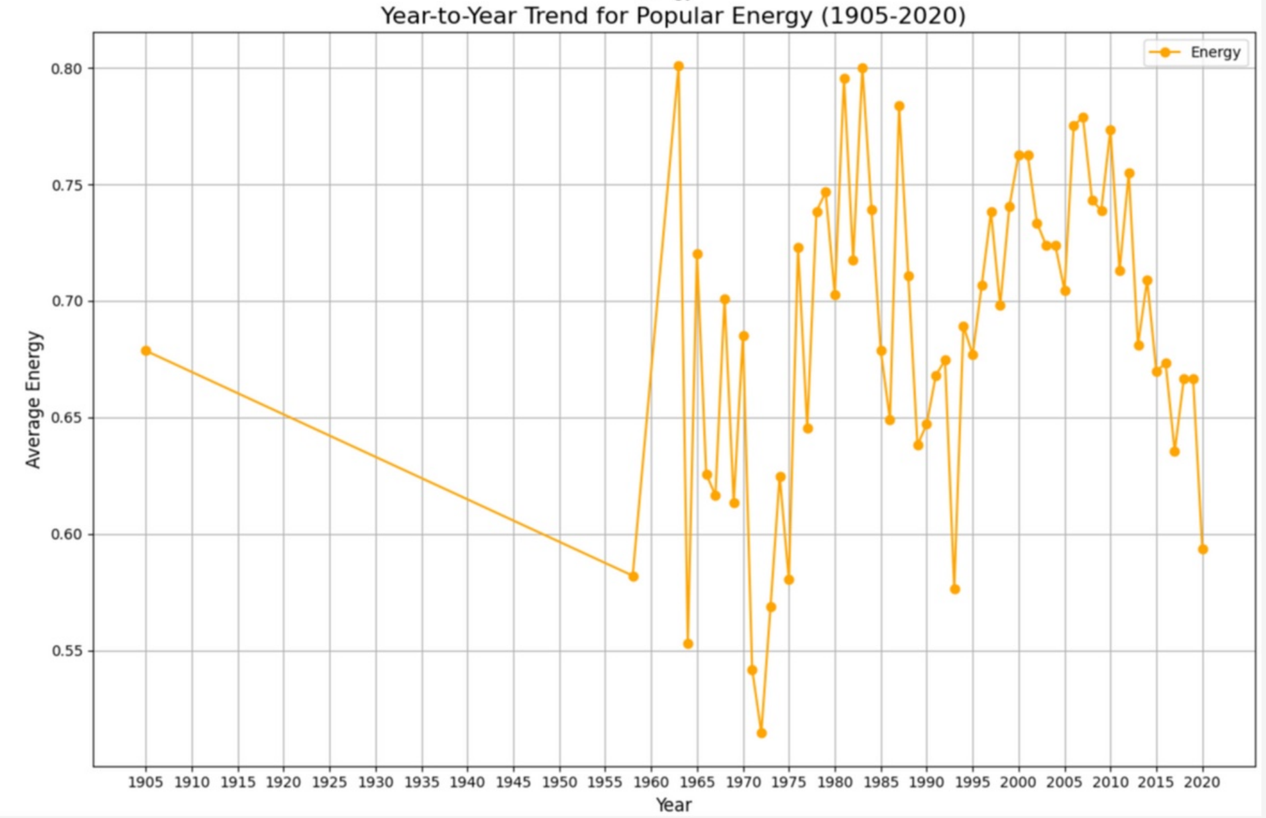
Key Points:

- **Energy Concentration:**
  - Popular songs are mostly concentrated around an energy level of 0.7.
  - The overall range of energy values spans from 0.6 to 0.9.

- **Consistency Over Time:**
  - Throughout Spotify's history, popular songs have maintained a relatively stable average energy level near 0.7.

- **Tendency:**
  - Energy levels in popular music reflect a balance of upbeat and dynamic elements.

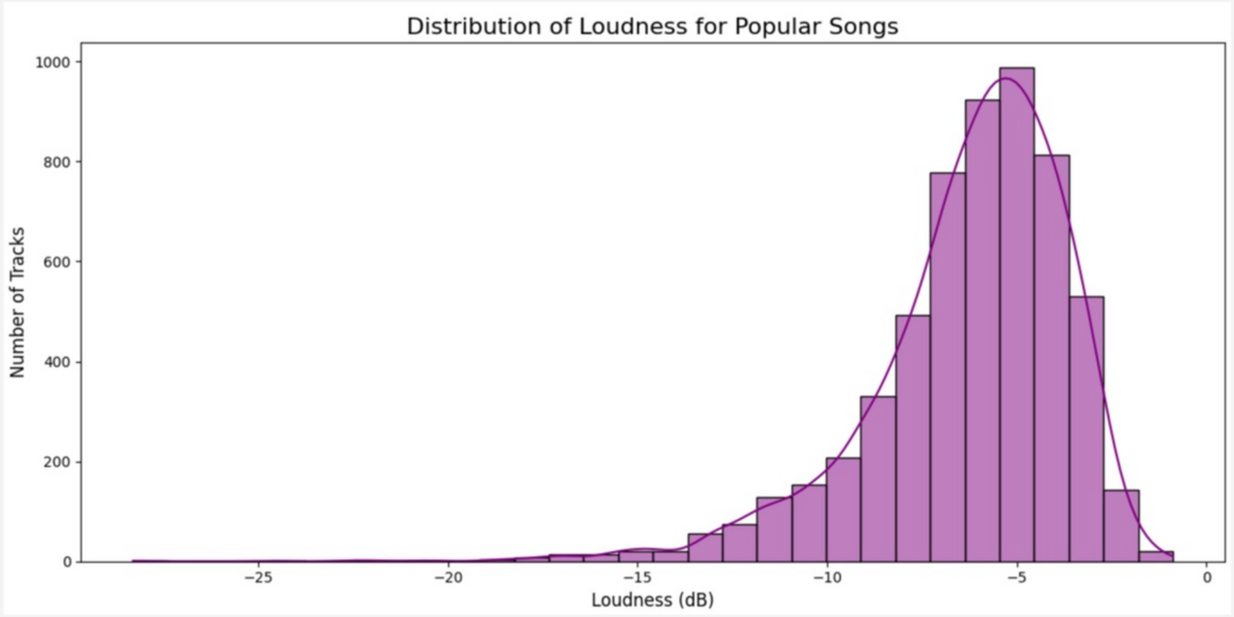


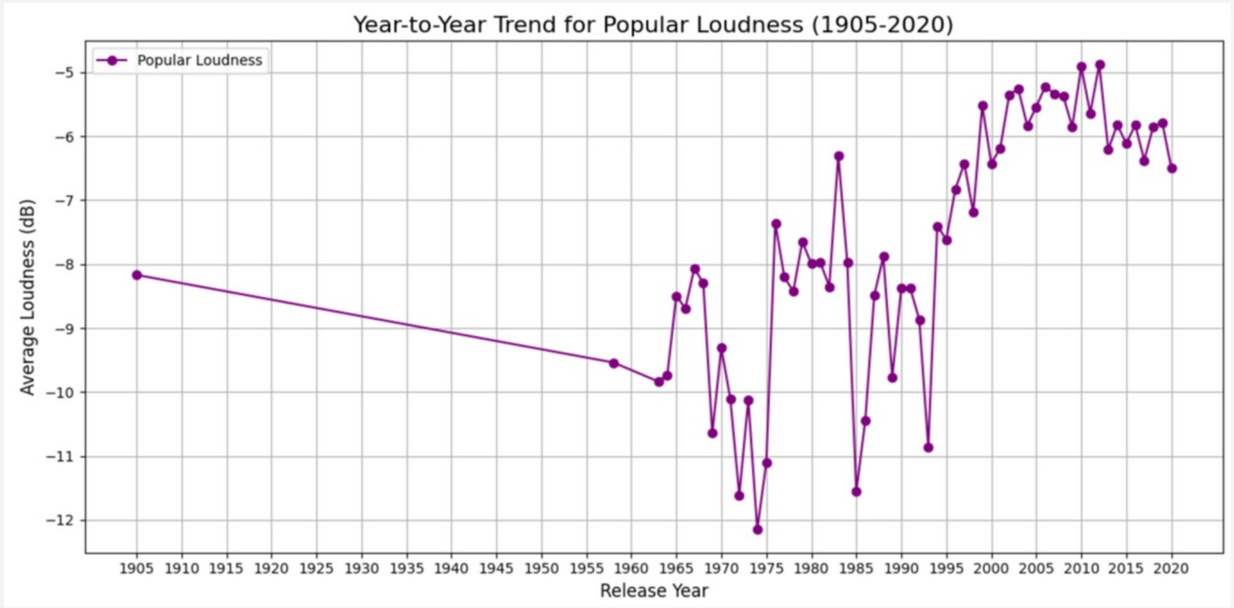


POPULAR LOUDNESS

Key Points:

- Popular songs concentrate around -5 dB in loudness, reflecting industry standards.
- High-energy genres like EDM, pop, and rap dominate this trend.
- Mastering ensures tracks are loud enough without distortion, maintaining consistent sound quality across platforms.
- Loudness increased significantly after the 1980s, peaking in the 2000s.

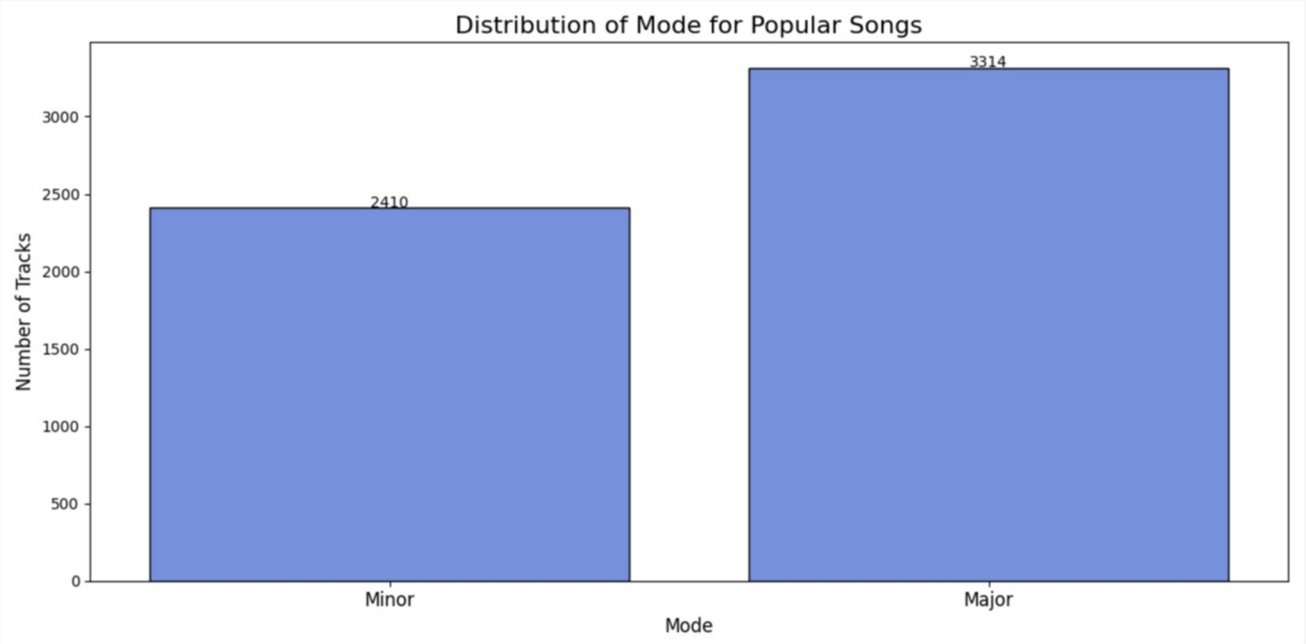




POPULAR MODE

Key Points:

- **Major Mode Dominance:**
  - 58% of popular songs are in the major mode.
  - Associated with energetic, uplifting genres like EDM and pop.
- **Genre Characteristics:**
  - Positive and high-energy sounds align with the major mode.
- **Influence on Mode:**
  - Danceability, energy, and loudness affect a track's feel but have less impact on mode choice.



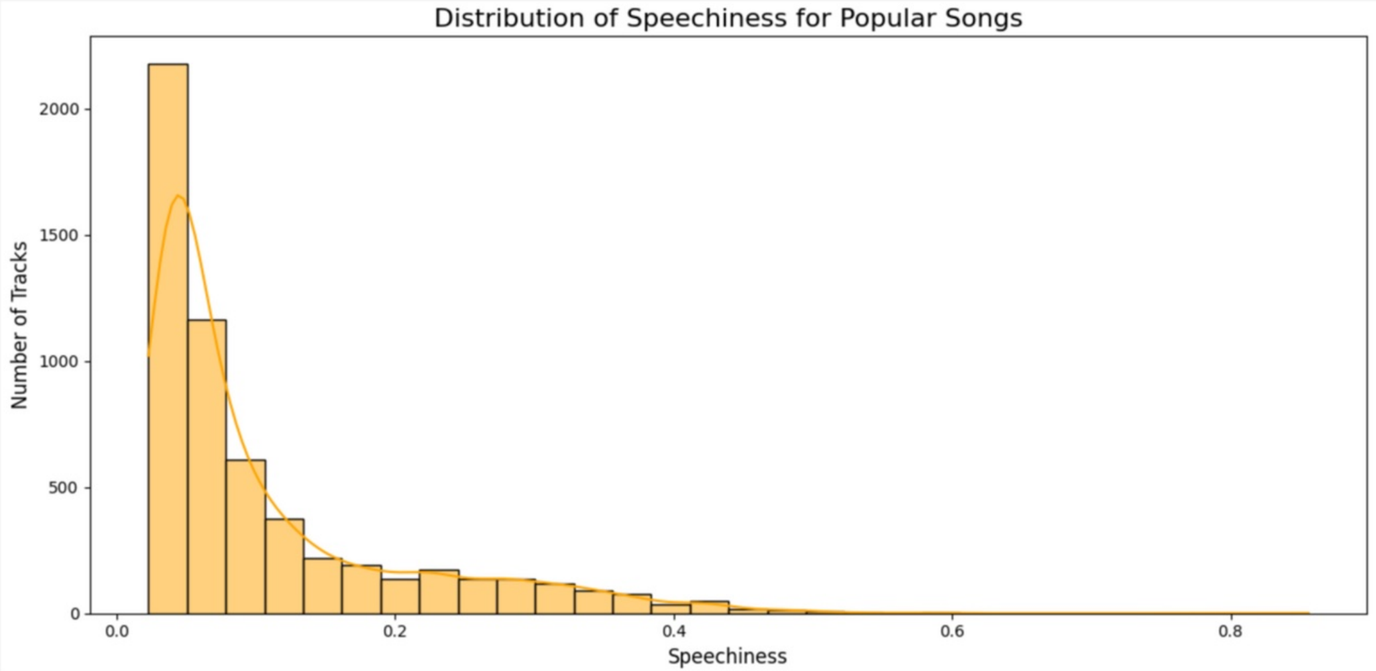
POPULAR SPEECHINESS

Key Points:

- **Skewed Distribution:**
  - Speechiness values are heavily skewed toward low levels, indicating dominance of traditional music tracks.

- **Typical Range:**
  - Most tracks have speechiness values below 0.33, representing minimal speech-like elements.

- **High Speechiness Tracks:**
  - Very few tracks exceed 0.66, which include spoken-word recordings like:
    - Audiobooks
    - Podcasts
    - Poetry readings.



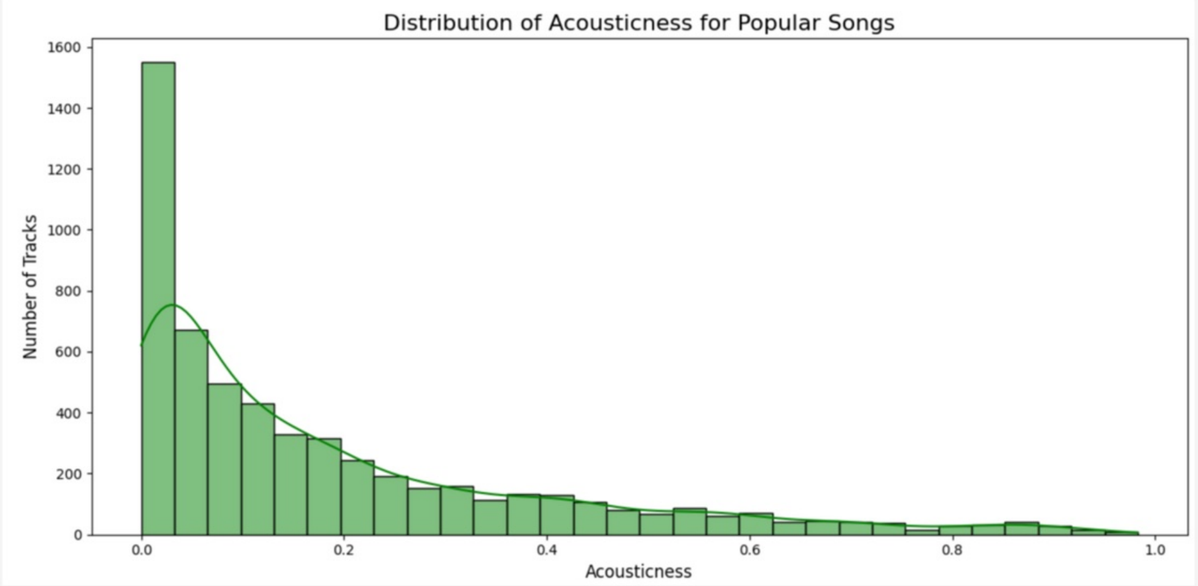
POPULAR ACOUSTICNESS

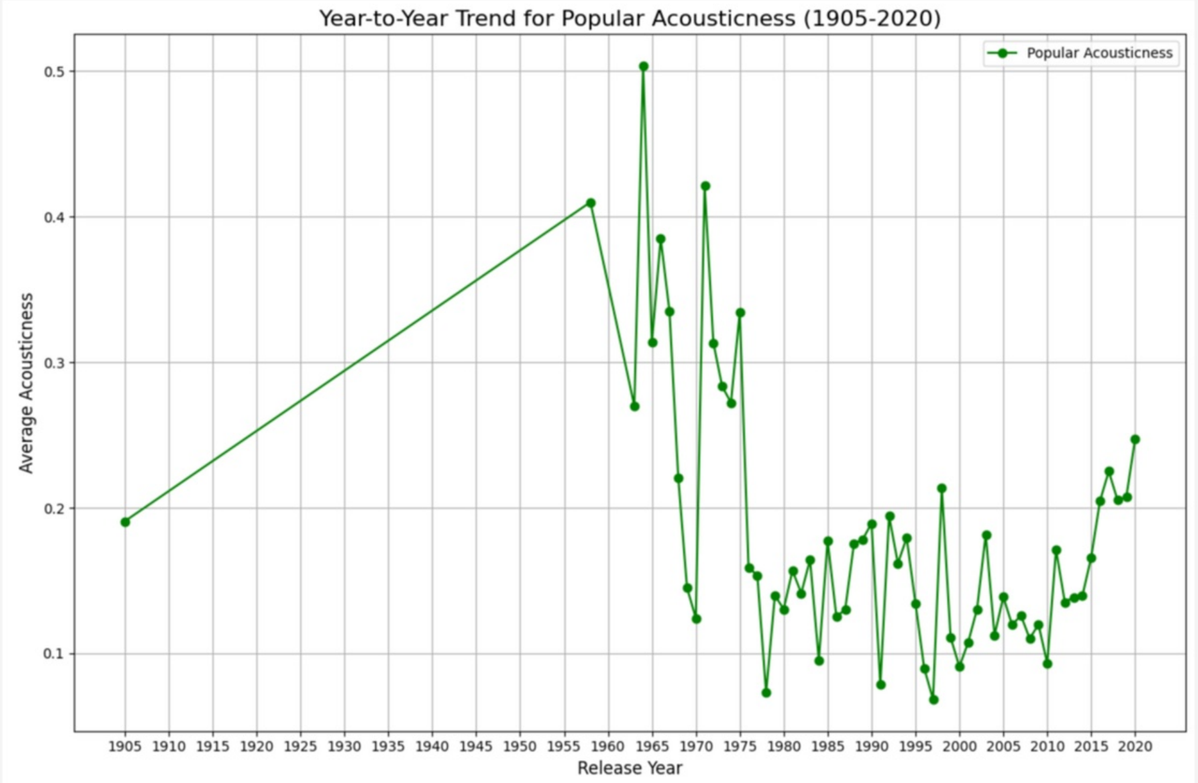
Key Points:

- **Low Acousticness:**
  - Most popular songs have acousticness values close to 0.0, indicating they are primarily non-acoustic.

- **Recent Trends:**
  - A slight increase in acousticness is observed from 2015 to 2020, peaking at around 0.25.

- **Consistent with Popular Genres:**
  - High-energy, rhythmic, and electronically produced tracks dominate over acoustic elements.
  - Trends align with genres prioritizing danceability and energy.





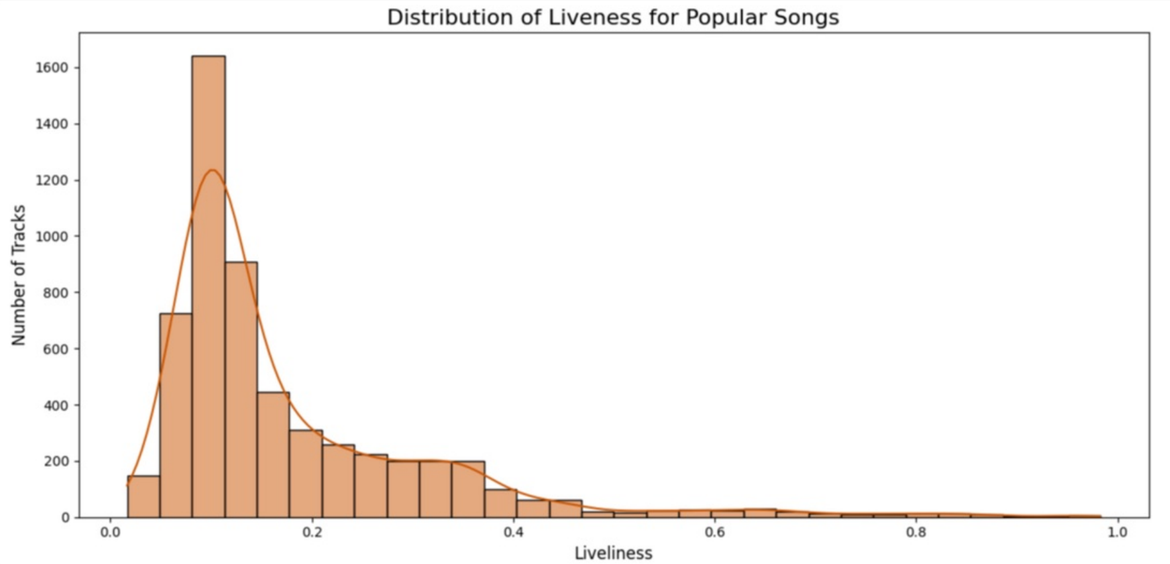
POPULAR LIVENESS

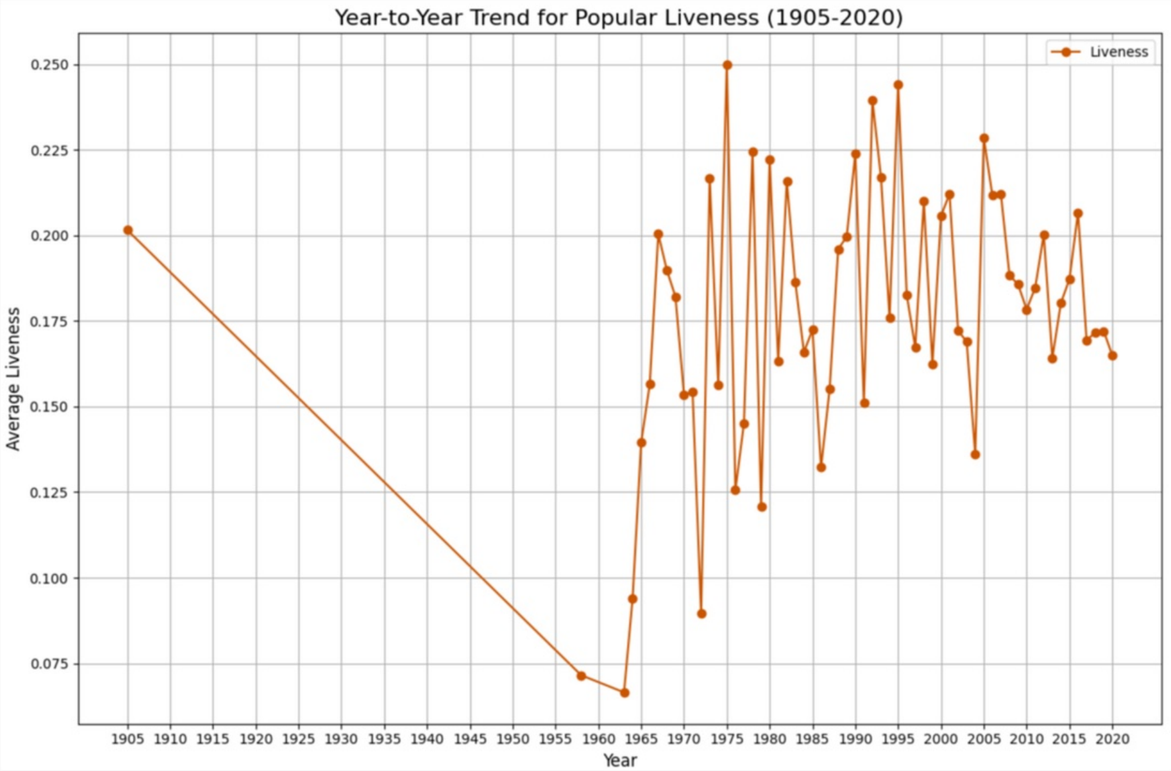
Key Points:

- **Low Liveness Dominance:**
  - Liveness values are concentrated around 0.1, indicating most tracks are studio recordings with minimal live elements.

- **Recent Trends:**
  - Increased fluctuations in liveness observed in recent years.
  - Over the last five years, a consistent decrease in liveness highlights the preference for studio-produced tracks.

- **Dominance of Studio Tracks:**
  - Tracks with little or no audience presence dominate popular music trends.





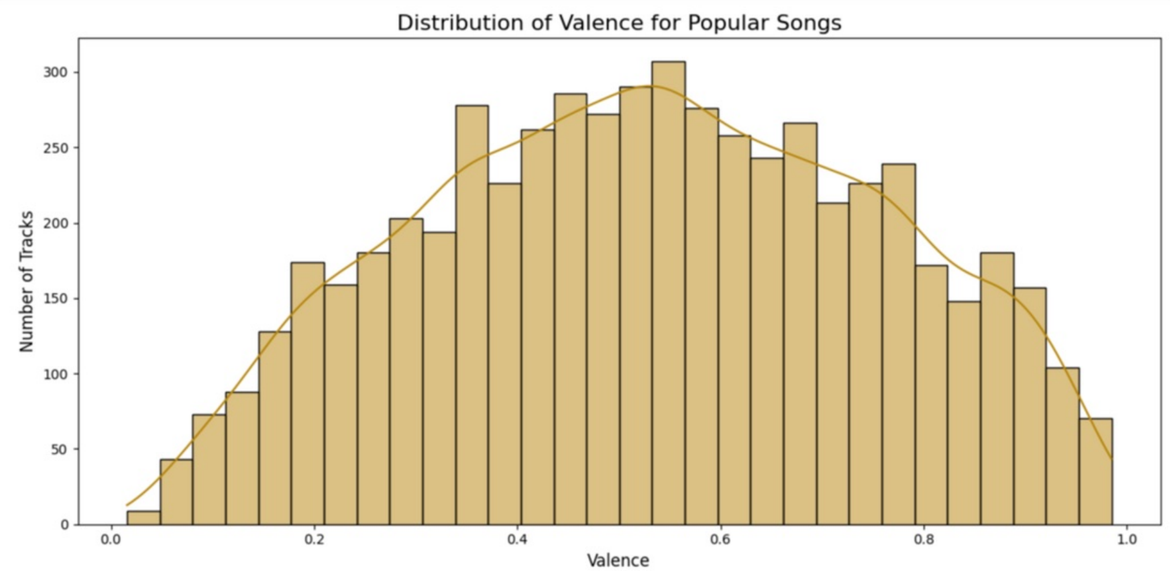
POPULAR VALENCE

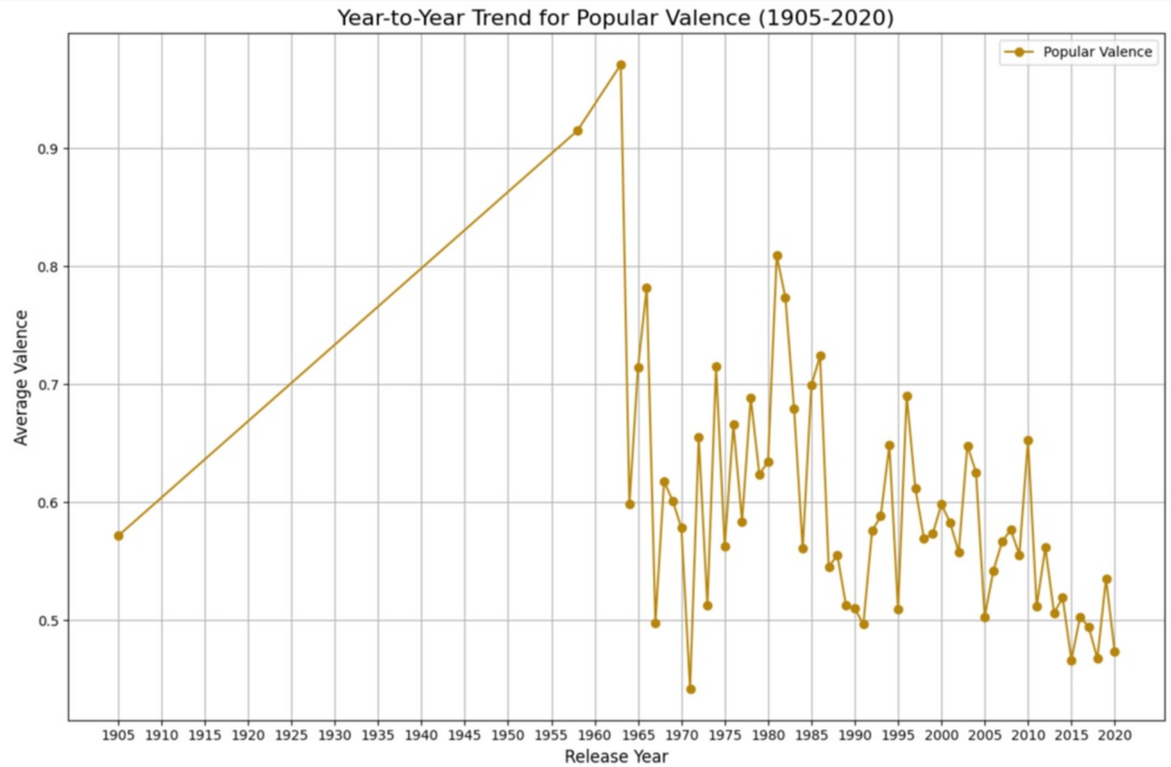
Key Points:

- **Neutral to Positive Valence:**
  - Popular songs are concentrated in the neutral to slightly positive range of emotional tone.

- **Valence Stability:**
  - Over the last five years, valence values have been stable, ranging between 0.45 and 0.53.

- **Emotional Tone:**
  - Reflects a generally neutral or slightly positive emotional tone in popular music trends.





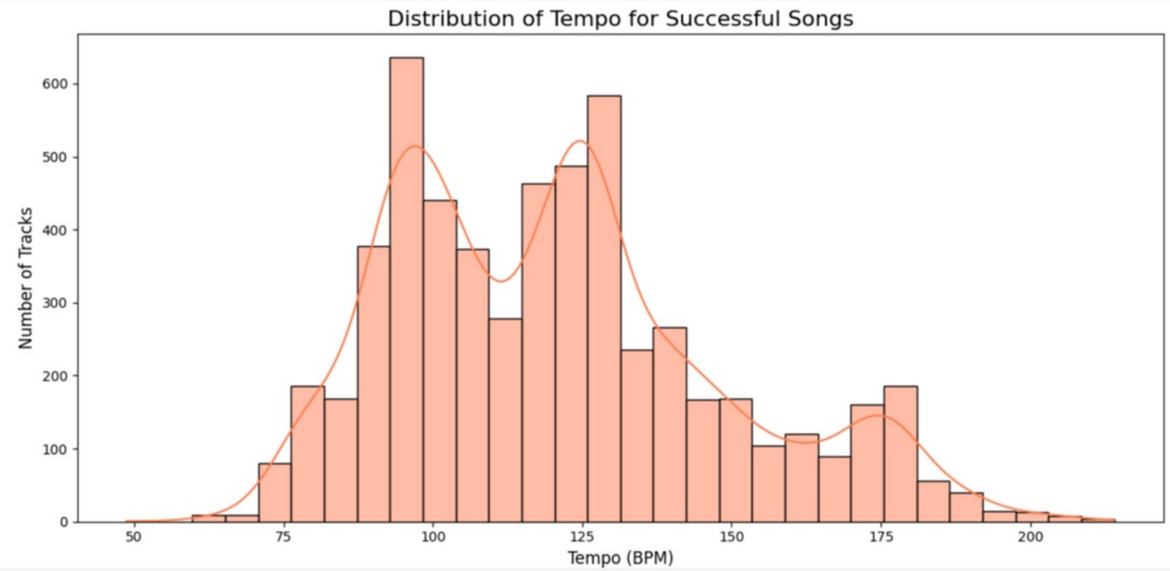
POPULAR TEMPO

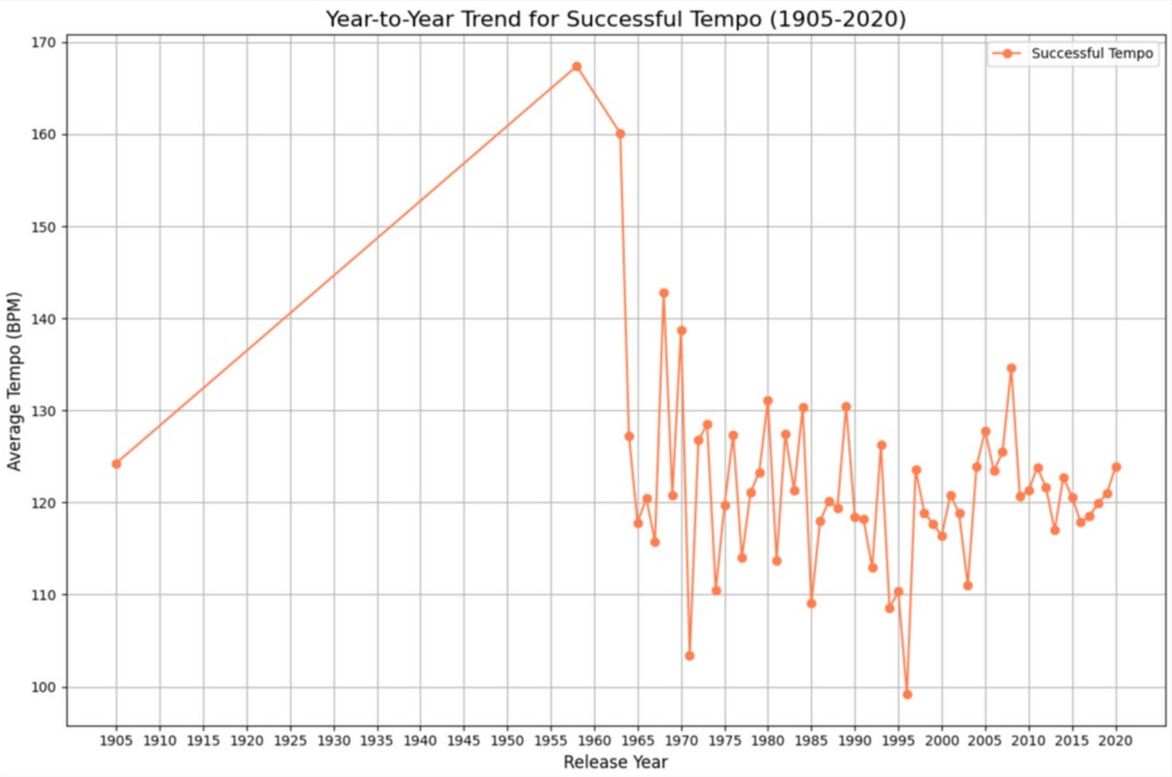
Key Points:

- **Tempo Range:**
  - Popular songs are concentrated between 95 and 130 BPM, the most common tempo for successful tracks.

- **Recent Shift:**
  - In the past five years, there has been a slight trend toward higher tempos, clustering around 120 BPM.

- **Rhythmic Trends:**
  - Reflects an increasing focus on upbeat and energetic rhythms in modern popular music.





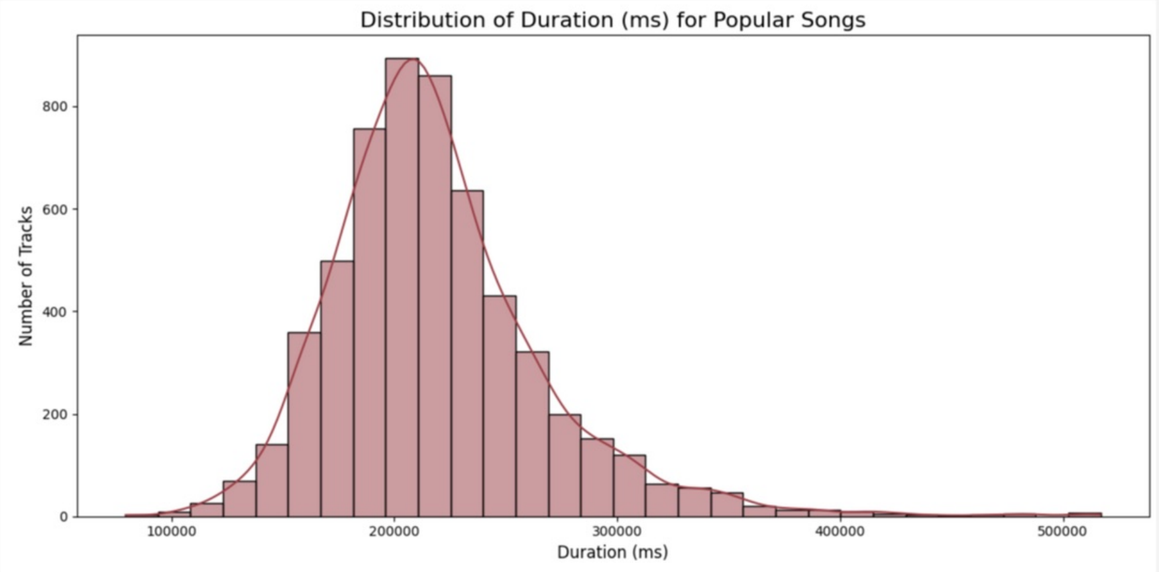
POPULAR DURATION TIME

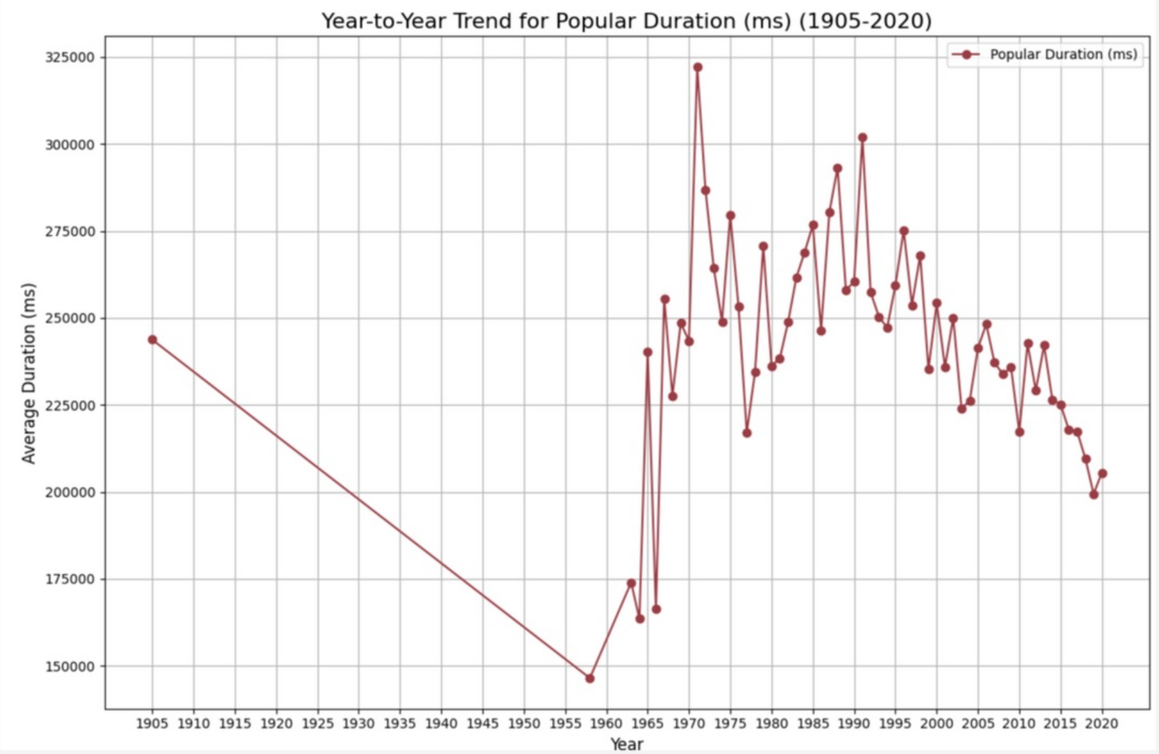
Key Points:

- **Typical Duration:**
  - Most popular songs have a duration around 200,000 milliseconds (3 minutes and 20 seconds).

- **Recent Decline:**
  - Track durations have steadily decreased since 2016, reflecting modern media trends.

- **Media Consumption Shift:**
  - Shorter tracks are prioritized to capture attention quickly in a competitive digital landscape, diverging from traditional radio and streaming norms.





FEATURE SIGNIFICANCE

04

FROM RANDOM FOREST: FEATURE IMPORTANCE

Feature Importance:

	columns_to_include	Importance
9	instrumentalness	0.301001
5	loudness	0.139467
1	playlist_genre	0.129933
3	energy	0.098396
2	danceability	0.065320
8	acousticness	0.059825
13	duration_ms	0.054055
12	tempo	0.038570
11	valence	0.035699
7	speechiness	0.027741
10	liveness	0.023532
0	release_month	0.019671
4	key	0.005495
6	mode	0.001294

- **Key Drivers:** Instrumentalness, loudness, playlist genre, and energy.
- **Moderate Impact:** Acousticness, danceability, and duration.
- **Low Impact:** Release month, key, and mode.

"**Conclusion:** Musical characteristics (energy, danceability, loudness) and genre are the main drivers of popularity while more technical features like key and mode have less influence"

MODELING

05

MODELING

Modeling Approach & Setting

Modeling

“Variable Selection”
(excluding the least relevant)

“Experimentations with various model types”
(Logistics Regression, Decision Trees, Neural Networks, Random Forests, LightGBM, CatBoost)

Best Model

“Random Forest”

- Capture non-linear relationships
- Handle mixed data types with least preprocessing
- Ensemble reduces overfitting

Optimization

“Parameter & Technique”

- Estimators: 200
- Max Depth: 15
- Min samples split: 2
- Min samples leaf: 1
- Class weighting (15x for 1)
- Lowering threshold

IMPLICATIONS

PERFORMANCE METRICS

AUC: 0.78
Recall: 0.98
Precision: 0.26
FPR: 0.78

Music Industry...

RECALL

Captures as many high-revenue tracks as possible, minimizing the risk of missing out on potential hits

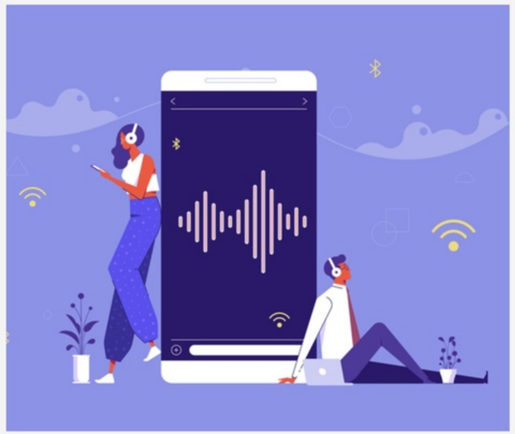
EXPECTED PAYOFF

$30 per track investment is directed towards the most promising tracks, optimizing the ROI

EXPECTED PAYOFF

2a.
Cutoff: 0.22
Expected payoff: S155,480,000.

Actual \ Predicted	0	1
0	1654	74
1	4154	1998



20% ADJUSTMENT APPLIED

Cutoff: 0.10
Expected payoff: S290,304,000

Actual \ Predicted	0	1
0	1700	28
1	4929	1223

CONCLUSION

06

FACTORS FOR SUCCESS

- **Genre Trends:** Latin music has become a leading genre, surpassing pop, R&B, and rap, reflecting its increasing global influence and diverse fanbase.
- **Musical Features:** Popular songs are characterized by being danceable, energetic, and uplifting. They often feature major keys (e.g., Db/C#, C, G), upbeat tempos (95-130 BPM), and a typical length of around 3:20 minutes, catering to modern consumption habits.
- **Production Style:** Most successful tracks are studio-produced, focusing on prominent vocals, positive vibes, and emotionally uplifting elements that resonate with broad audiences.
- **Release Timing:** January, June, and November are strategic months for music releases, aligned with promotional cycles, seasonal demand, and holiday gift-giving trends.
- **External Factors:** Artist fame, viral social media trends (e.g., TikTok, Instagram), and collaborations with influential artists significantly impact a song's success and virality.

CONCLUSION

From EDA and running random forest on the dataset, we conclude that:

- **Market Trends:** Popular songs are shaped by global shifts, with genres like Latin, Pop, and R&B driving listener engagement and streaming growth.
- **Popular Songs**: High energy, danceability, prominent vocals, and upbeat tempos, often falling within mainstream playlist genres.
- **Popular vs Unpopular:** Popular songs are more likely to feature positive vibes, major keys, and rhythmic elements, whereas unpopular songs lean towards acoustic or instrumental styles.
- **Statistical models** or data analyses can help explain the key factors driving a song's success

THANKS!