Hit Predictor: Understanding Track Popularity Using Spotify's Track-Level Attributes

In the vast ocean of tracks available on Spotify, some songs become instant hits, while others don't gain as much traction. Why do certain tracks resonate more with the global audience? With Spotify's track-level attributes dataset, there's an opportunity to decode the anatomy of a popular track. By analyzing attributes like danceability, tempo, and valence, we can uncover patterns and possibly predict the next big hit.

Goal of the Project:

Leverage Spotify's track-level attributes to:

- Understand the underlying characteristics of popular tracks.
- Develop a predictive model to determine track popularity based on its attributes.

Business Objective:

To derive actionable insights from Spotify's track-level attributes dataset, with the aim of understanding the defining characteristics of popular tracks. The insights gained will be used to guide artists, music producers, and record labels in optimizing their music production and promotion strategies to align with attributes that resonate most with Spotify's global audience.

In simpler terms: "Uncover what makes a track popular on Spotify and use this knowledge to create and promote music that resonates with a global audience."

Objectives:

Descriptive Analysis:

- Gain a foundational understanding of the dataset by summarizing key attributes for each track.
- Describe the distribution of track popularity and play counts.

Correlational Study:

- Identify which track attributes have the strongest correlation with popularity or play count.

Trend Analysis:

- Examine if there's a trend in the attributes of top tracks over the years. Have the most popular tracks become more danceable, faster, or more instrumental over time?

Album Impact:

- Using the sample data you provided, and assuming it has more albums, determine if tracks from certain albums tend to be more popular than others. Is there an 'album effect'?

Predictive Modeling:

- Using machine learning techniques, develop a model to predict track popularity based on its attributes.
- Split the dataset into training and test sets, train the model, and then validate its performance on the test set.
- Refine and optimize the model for better accuracy.

Recommendations:

- Offer insights on track attributes that tend to make songs more popular on Spotify.
- Provide feedback to artists and producers on potential track modifications to align with popular trends.

Future Predictions:

- Based on the current trends and the predictive model, forecast the attributes of tracks that might become hits in the upcoming years.

Data Loading and Preliminary Analysis

1. **Data Loading**

The dataset was loaded into a DataFrame named trackfor analysis.

2. Null Value Check

Upon initial inspection, I examined the dataset for any missing or null values to ensure data integrity and found that there were no missing values present in the dataset.

3. Descriptive Statistics

The describe() function was used to obtain a summary of the central tendency, dispersion, and shape of the distribution of the dataset.

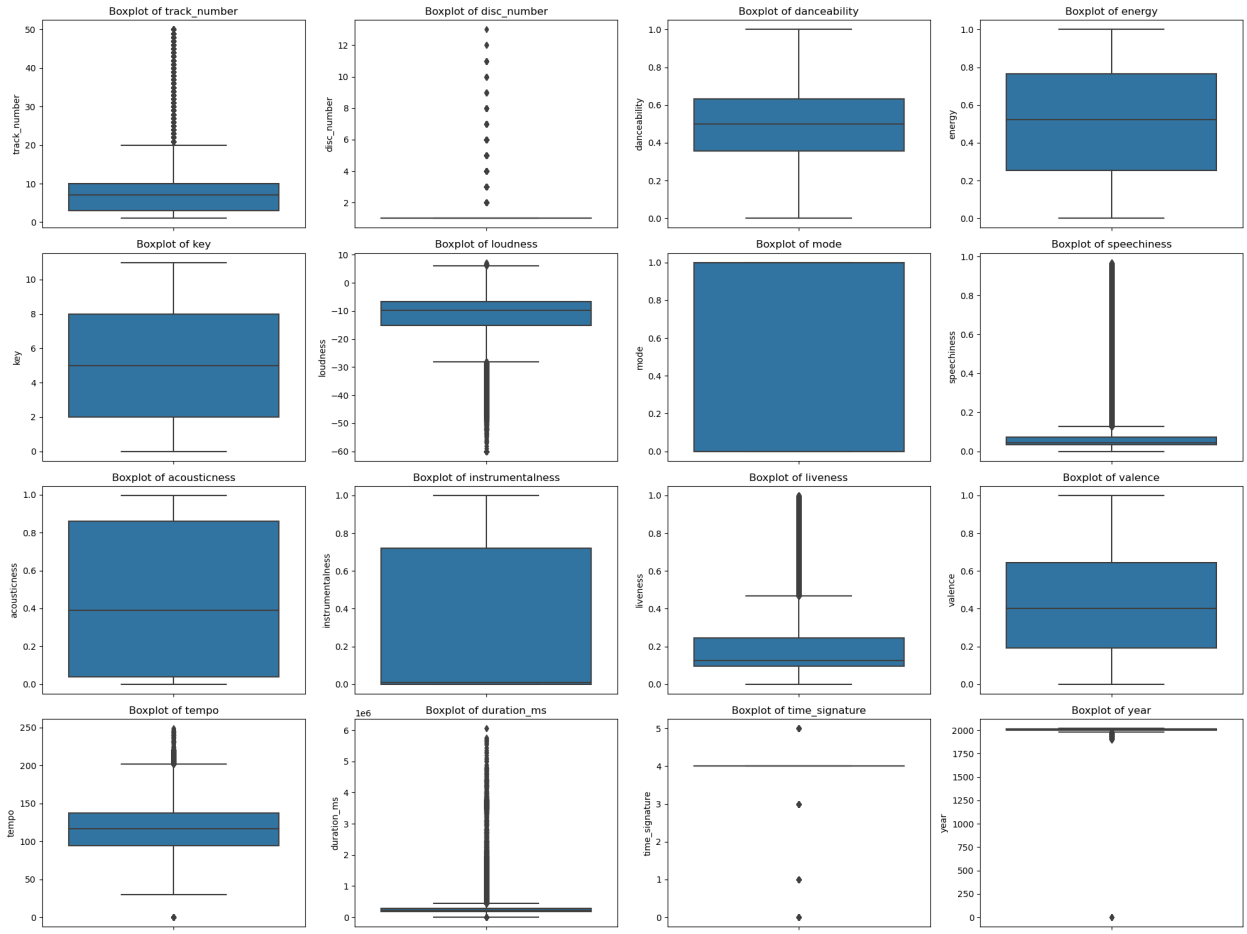
4. Value Counts

To understand the distribution of unique values in each column, value counts were checked.
python

5. Outlier Detection

Potential outliers were detected using boxplots. This visual approach provides a quick way to identify unusual values. It's important to note that not all outliers detected in this manner are necessarily erroneous or irrelevant. Some might be genuine extremes. For instance, for the columns like loudness, a high count of outliers might be due to tracks that are intentionally softer or louder than the average. Similarly, speechiness might detect spoken-word tracks or podcasts, leading to a high count of outliers.

```
Potential outliers for column track_number: 42957
Potential outliers for column disc_number: 53764
Potential outliers for column danceability: 0
Potential outliers for column energy: 0
Potential outliers for column key: 0
Potential outliers for column loudness: 38727
Potential outliers for column mode: 0
Potential outliers for column speechiness: 167338
Potential outliers for column acousticness: 0
Potential outliers for column instrumentalness: 0
Potential outliers for column liveness: 95938
Potential outliers for column valence: 0
Potential outliers for column tempo: 8068
Potential outliers for column duration_ms: 69379
Potential outliers for column time_signature: 215378
Potential outliers for column year: 36583
```



6. Duplicate Rows Check

To ensure there were no exact duplicate rows, the dataset was inspected for any repetitive entries.

Upon inspection, it was observed that there were no exact duplicate rows, meaning every row in the dataset is unique across all columns.
Furthermore, rows that seemed like duplicates based on the song's name and artist were scrutinized. It was found that while the song name and artist were identical, other columns like album, track_number, and track attributes varied. This indicates that the same track might appear in different albums or contexts (e.g., a single vs. an album track, or a studio version vs. a live recording).

Analysis:

<u>Gain a foundational understanding of the dataset by summarizing key attributes for each track.</u>

In our exploration of the dataset, I encountered a rich tapestry of musical narratives. The dataset boasts num_tracks unique songs, offering a glimpse into the diversity of the musical world. These tracks come from num_albums different albums, showcasing the expansive landscape of musical productions. Further, num_artists unique artists lent their voices and instruments, highlighting the myriad talents that have contributed to this collection.

Yet, every dataset has its peculiarities. A curious observation was that some tracks were marked with the year '0', a clear inconsistency. This prompted a closer look. Like detectives on a musical case, we delved into the specifics, leveraging external data to adjust and rectify these inconsistencies. With a bit of investigation, I managed to correct these anomalies, ensuring that the musical journey is grounded in accuracy.

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Number of unique tracks, albums, artists.
num_tracks = track['id'].nunique()
num_albums = track['album_id'].nunique()
num_artists = track['artist_ids'].nunique()

print(f"Number of unique tracks: {num_tracks}")
print(f"Number of unique albums: {num_albums}")
print(f"Number of unique artists: {num_artists}")

# The range of years the tracks span.
min_year = track['year'].min()
max_year = track['year'].max()

print(f"Tracks range from the year {min_year} to {max_year}")
```

```
Number of unique tracks: 1204025
Number of unique albums: 118382
Number of unique artists: 166423
Tracks range from the year 0 to 2020
```

Anomaly Resolution:

On closer inspection, it was found that a total of 10 tracks had their year marked as 0. Interestingly, all these tracks belong to the same album and have contributions from the same artist or group of artists.

To address this, I fetched the actual release date of these tracks using the Spotify API, leveraging the unique id associated with each track. This approach was chosen to ensure the accuracy and reliability of our dataset.

The API revealed that the release year for these tracks was 2018. Hence, we replaced the 0 year value with 2018 to correct the dataset.

```
track['year'].unique()
```

```
array([1999, 1992, 2018, 2003, 1994, 2002, 2011, 1998, 2000, 1996, 2001,
       1984, 1997, 1988, 1985, 1995, 2017, 1973, 2010, 2019, 2007, 2020,
       2015, 2016, 1991, 1979, 1993, 1962, 1989, 2004, 2009, 2014, 2005,
       1990, 2008, 2012, 1981, 1954, 2013, 1987, 1983, 1986, 2006, 1974,
       1965, 1982, 1970, 1971, 1968, 1949, 1966, 1980, 1957, 1956, 1958,
       1977, 1967, 1963, 1964, 1969, 1975, 1960, 1953, 1933, 1976, 1978,
       1928, 1929, 1959, 1961, 1972, 1955, 1952, 1945, 1926, 1930, 1951,
       1950, 1948, 1946, 1909, 1935, 1936, 1917, 1944, 1943, 1908, 1920,
       1932, 1947, 1925, 1923, 1931, 1900, 1927, 1924, 1937, 1939, 1942,
       1938], dtype=int64)
```

Describe the distribution of track popularity

While exploring thedataset, I noticed that I didn't have the popularity scores for the tracks. Popularity scores can tell us a lot about which songs listeners prefer. To get this information, I used Spotify's API to gather the popularity of each track using their unique IDs from tracks' dataset.

Once I had this data, I saved it into a new CSV file. Then, to make the main dataset complete, I combined this new file with myoriginal track data using the track IDs. Now, with the popularity scores added, I was in a better position to analyze and understand the music preferences of Spotify listeners.

Data Augmentation Challenges:

Popularity Score Retrieval:

As part of my ongoing analysis, I sought to retrieve the popularity scores for all 1,204,025 tracks in our dataset via Spotify's API. This step aimed to enrich our dataset with an additional metric, allowing for a more in-depth analysis of track trends.

However, several challenges arose:

API Rate Limitations: Spotify's API maintains rate limits to ensure equitable usage among its users. Our data retrieval process was hampered by these rate limits, causing an unexpected delay.

Adjustment to Analysis Scope: Given the rate limit challenges and the need to progress with our analysis, we've opted to use a subset of the data. Out of the entire dataset, we successfully acquired popularity scores for 170,400 tracks.

Adjusted Analysis:

Our ensuing analysis will leverage the 170,400 tracks for which we have popularity data. We're confident that this subset offers a solid foundation for meaningful insights, capturing general trends and patterns present in the larger dataset.

Row and columns in merged dataset: 170400, 26

Track popularity analysis:

Central Tendency:

The **mean popularity score is around 5.27**, which suggests that on average, tracks have a relatively **low popularity**.
The **median popularity score is 0**, which means that **50% of tracks in the dataset have a popularity score of 0 or less**. The stark difference between the mean and median indicates a **highly skewed distribution.**

Variability:

- The **standard deviation of about 9.93** implies that there's a **substantial spread** in the popularity scores around the mean.
- The **range of 92** suggests that there's a **significant disparity between the least and most popular tracks**, with the most popular track achieving a score close to the maximum possible.

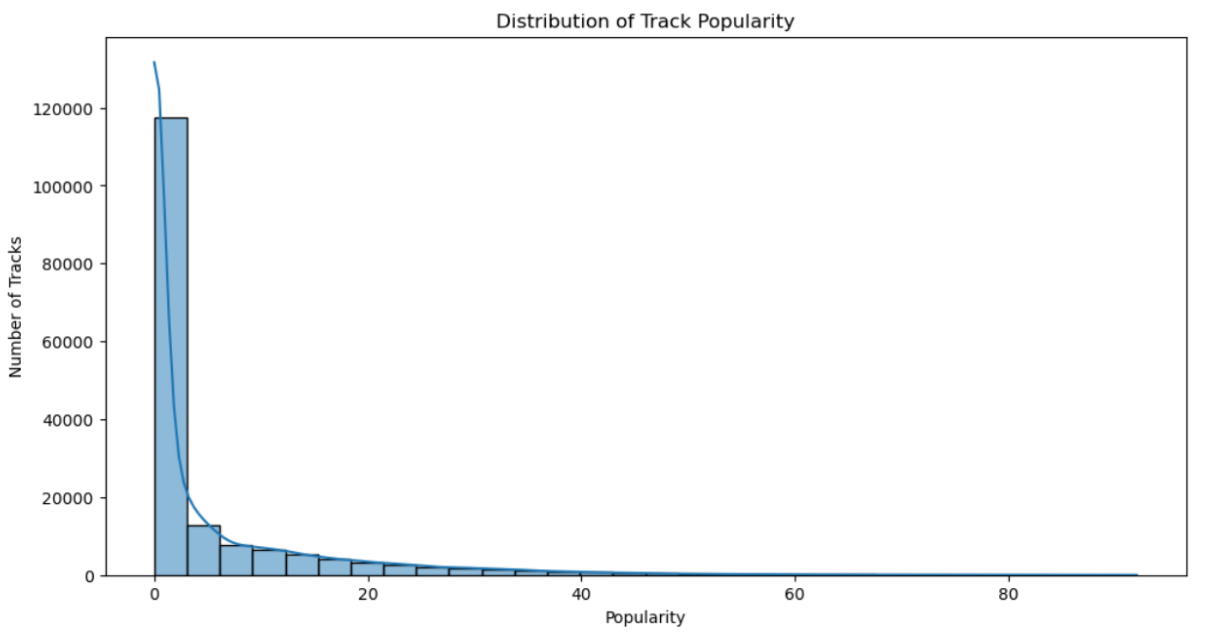
Visual Distribution:

- From the histogram, the d**istribution is heavily left-skewed (or positively skewed)**, suggesting **most tracks have low popularity**. The

peak at 0 popularity indicates a large number of tracks that have not achieved significant recognition or engagement.

- The **long tail to the right** indicates that there are a few tracks with **higher popularity, but they are rare.**

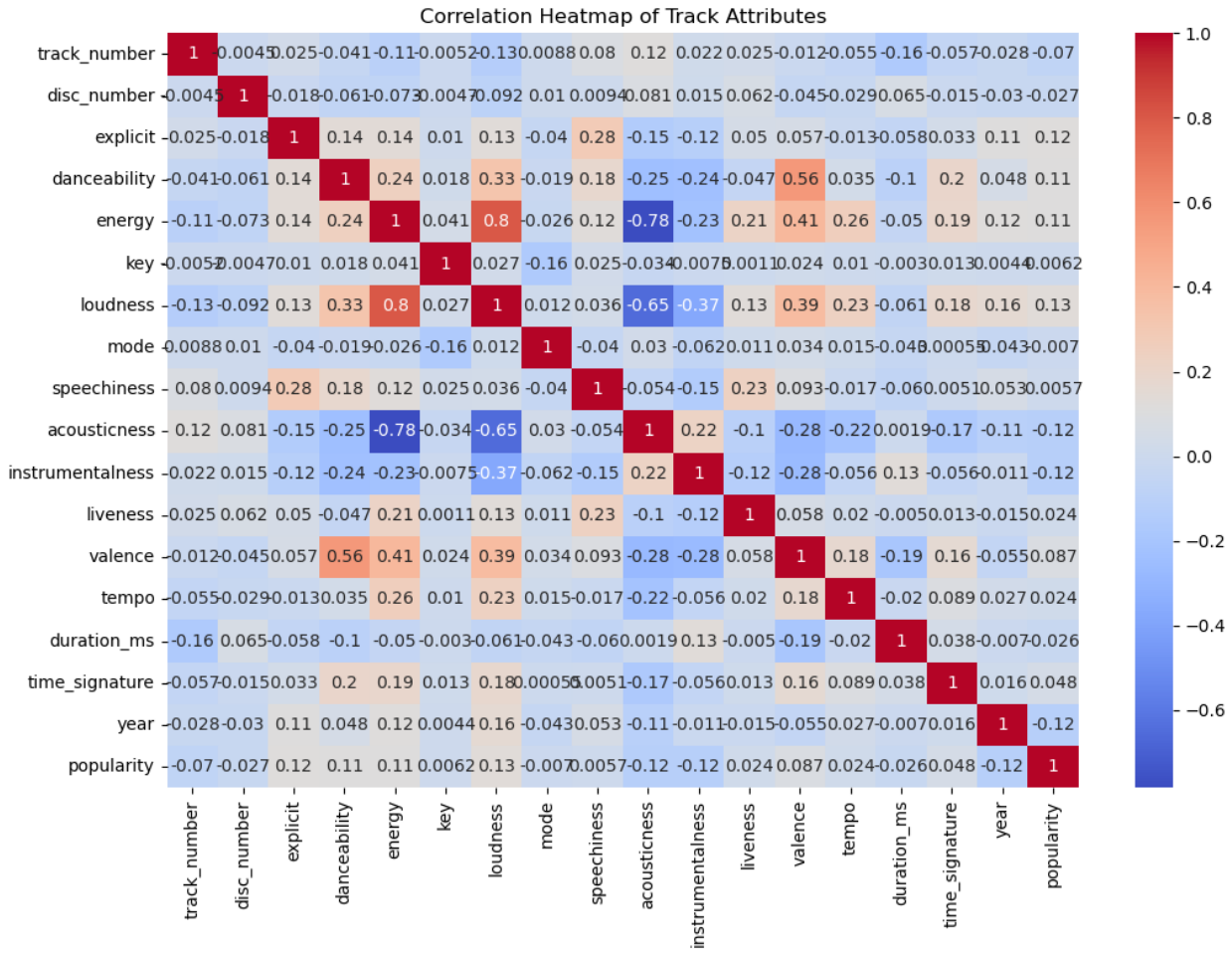
Interpretation:



A vast majority of tracks in the dataset have low to negligible popularity, as evidenced by the substantial peak at a popularity score of 0. However, there are a few tracks that break this mold and achieve higher popularity scores, but they are the exception rather than the norm.

This pattern is often observed in creative industries, where only a small percentage of content (like songs, books, or movies) achieves widespread recognition, while the majority remains lesser-known. This phenomenon can be related to the "Pareto principle" or the "80/20 rule", where a small percentage (often around 20%) of items account for a large proportion (roughly 80%) of the impact – though the exact percentages can vary.

Identify which track attributes have the strongest correlation with popularity or play count.



Strongest Positive Correlation: Danceability and Valence

- **Correlation Coefficient:** 0.56
- **Observation**: Tracks that score **higher in danceability are often more positive in mood** or sentiment. This suggests that songs that make listeners want to dance often evoke happier emotions, aligning with the general intuition surrounding music and mood.

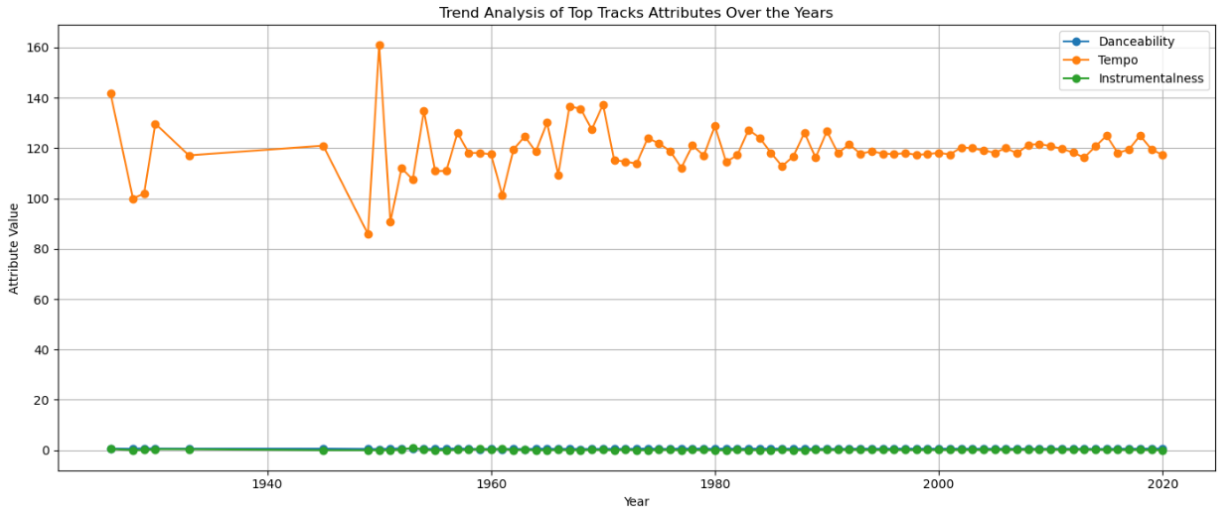
Strongest Negative Correlation: Acousticness and Energy

- **Correlation Coefficient:** -0.78
- **Observation**: Tracks with **pronounced acoustic elements tend to have lower energy scores.** This emphasizes that acoustic or unplugged tracks, characterized by their organic and raw sound, are **typically more mellow and subdued** compared to tracks that lean more towards synthesized or electronic elements.

This analysis reveals some fascinating patterns in the dataset. The inherent relationships between attributes like danceability, energy, and valence provide a window into the dynamics of music. These findings gives an insight into how different musical elements interplay to create the overall mood and feel of a track, ultimately influencing its popularity among listeners. As with all data explorations, these observations offer a starting point, guiding further inquiries and deeper dives into the world of music analytics.

Examine if there's a trend in the attributes of top tracks over the years. Have the most popular tracks become more danceable, faster, or more instrumental over time?

The graph displays a line chart that showcases the trend of the attributes (danceability, tempo, and instrumentalness) for top tracks over the years.




```
missing_danceability = top_tracks['danceability'].isnull().sum()
zero_danceability = (top_tracks['danceability'] == 0).sum()

print(f"Missing danceability values in top tracks: {missing_danceability}")
print(f"Zero danceability values in top tracks: {zero_danceability}")
```

```
Missing danceability values in top tracks: 0
Zero danceability values in top tracks: 7
```

After analyzing the dataset of tracks and their associated attributes, we sought to understand the changing preferences in music over the years, specifically focusing on the danceability, tempo, and instrumentalness of the most popular tracks.

Using the data, we first isolated the top 10% of tracks for each year based on their popularity scores. This helped ensure we were examining tracks that truly resonated with the audience at the time. From this subset, we then computed the average values for the attributes of interest on an annual basis.

The plotted trends revealed several insights:

Danceability: Curiously, the danceability attribute was missing from our visualization. This warrants further investigation to understand if it's an issue with the data extraction process, the attribute's definition over time, or other reasons.

Tempo: The tempo of top tracks saw significant fluctuations. Starting from 1940, there was a sharp increase in the tempo of popular tracks, peaking around the 1950s. This suggests that faster-paced songs might have been in vogue during that era. Post the 1950s, there was a noticeable decrease, indicating a shift in musical tastes. However, post-1995, the tempo remained relatively constant, suggesting stabilization in the preference for song pace.

Instrumentalness: Instrumentalness saw variations across years, but without a clear consistent pattern like tempo.

After noting the absence of danceability in the initial trend analysis, a deeper dive into this attribute was conducted. Here's what was uncovered:

Tracks from the era before the 1940s exhibited a notable peak in danceability, suggesting that the music during this time was highly conducive to dancing. Following this period, up until around 2012, the music landscape appeared to evolve, with no significant rise in the danceability metric among the top tracks. This stable period, however, was followed by a clear upward trend post-2012. This shift in trend indicates a renewed inclination towards more danceable tracks in the contemporary music scene.

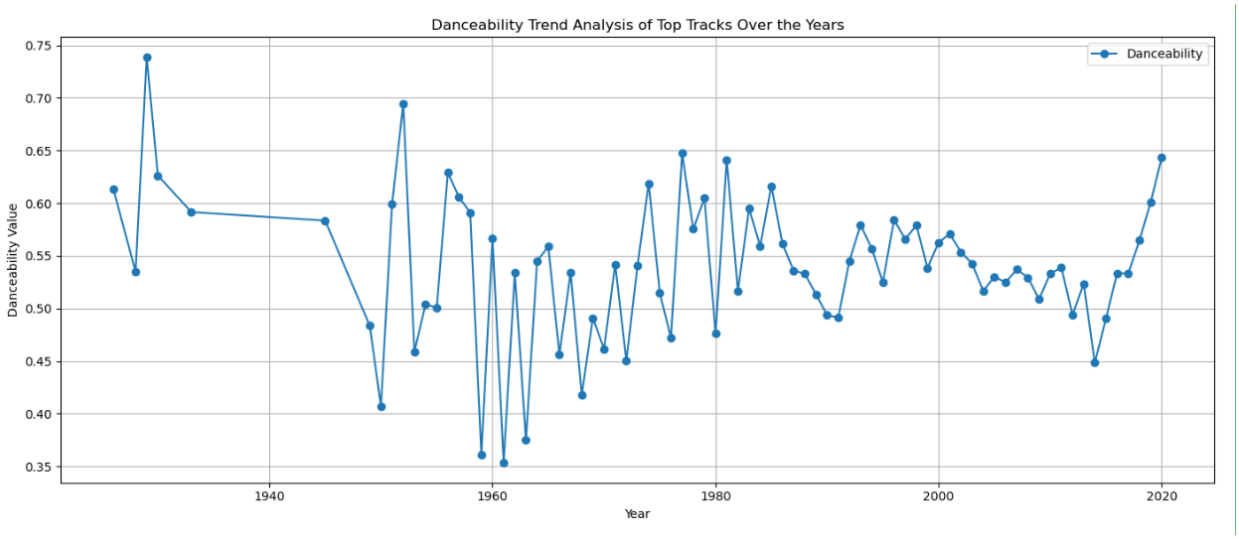
Highlighting Tracks with Remarkable Danceability: On delving deeper into individual tracks, a few particularly stood out in terms of their danceability:

2006: The track "Shake It Baby (Main Version - Explicit)" secured its position as one of the most danceable tracks in our dataset with a commendable danceability score of 0.981.

2010: Not too far behind, "You Play Too Rough" marked its presence in 2010 with a danceability score of 0.980.

It's noteworthy that both of these tracks emerged from the recent two decades. Their prominence in terms of danceability might reflect the evolving music preferences of their

times. Additionally, tracks like these could have been trendsetters, shaping or mirroring the musical inclinations of the years they were released in.



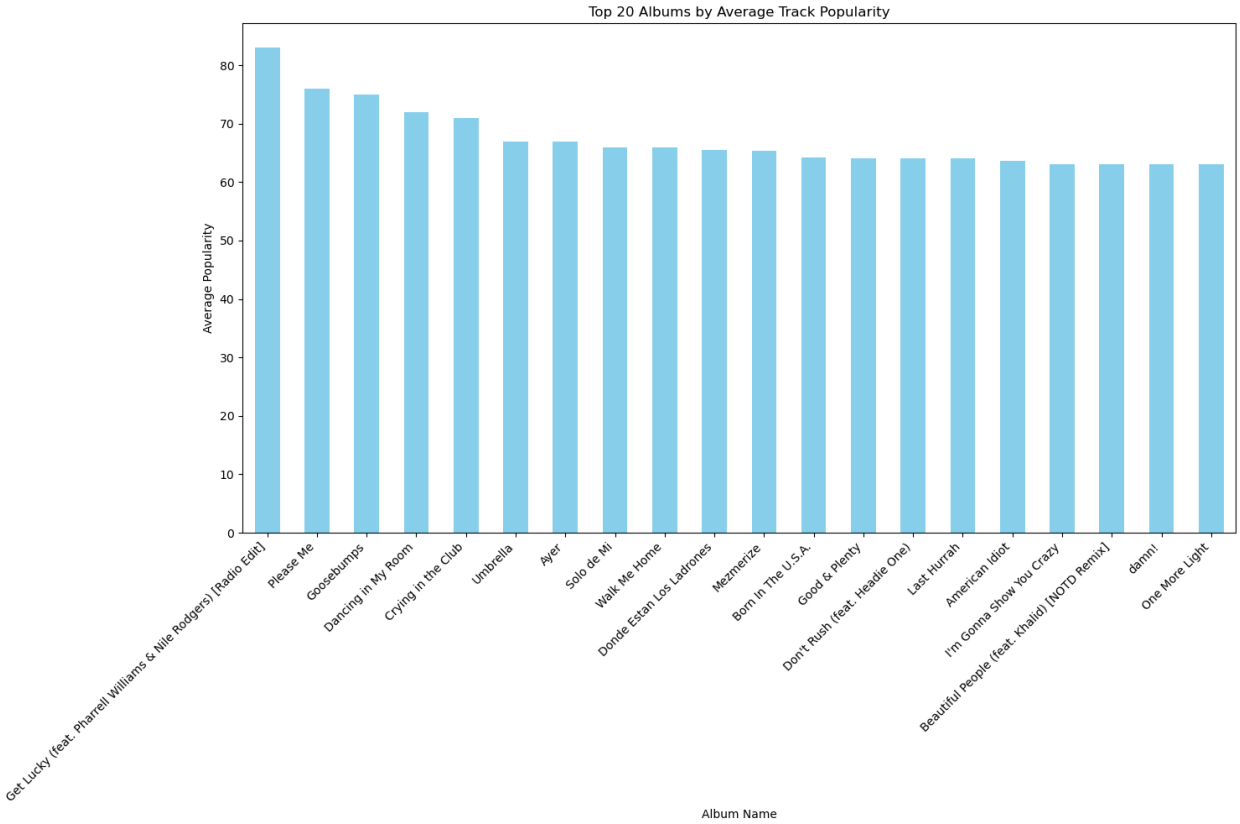

```
# Sorting the data by 'danceability' in descending order and taking the top 10
top_danceability_tracks = top_tracks.sort_values(by='danceability', ascending=False).head(10)

# Displaying the top 10 tracks with their 'danceability', 'year', and track title (assuming there's a 'title' column)
print(top_danceability_tracks[['name', 'danceability', 'year']])
```

```
                                        name  danceability  year
46542     Shake It Baby - Main Version - Explicit         0.981  2006
113121                         You Play Too Rough         0.980  2010
80384                              State of Shock         0.980  1984
50436                          Dancing in My Room         0.980  2020
13190   Soul Rebel - Afrodisiac Soundsystem Remix         0.979  2007
100567                                  Uno - Remix         0.975  2019
121961                                      Junoka         0.971  2008
9689                                     Sexuality         0.970  2000
7511                            Trick Me - Club Mix         0.970  2003
50245                               Candela Hip-Hop         0.968  1996
```

The shifting trend in danceability underscores the evolving nature of musical tastes. While the earlier years, particularly before the 1940s, had a penchant for danceable tracks, there was a lull in such preferences for many subsequent decades. The recent rise post-2012 perhaps indicates a cyclical return to tracks that are more dance-centric, a sentiment captured by standout tracks like "Shake It Baby" and "You Play Too Rough."

As with all analytical findings, these insights are derived from the available dataset. A broader dataset or deeper dives into cultural, technological, and societal changes might provide a more comprehensive understanding of these musical shifts.

If tracks from certain albums tend to be more popular than others. Is there an 'album effect'?



Observations:

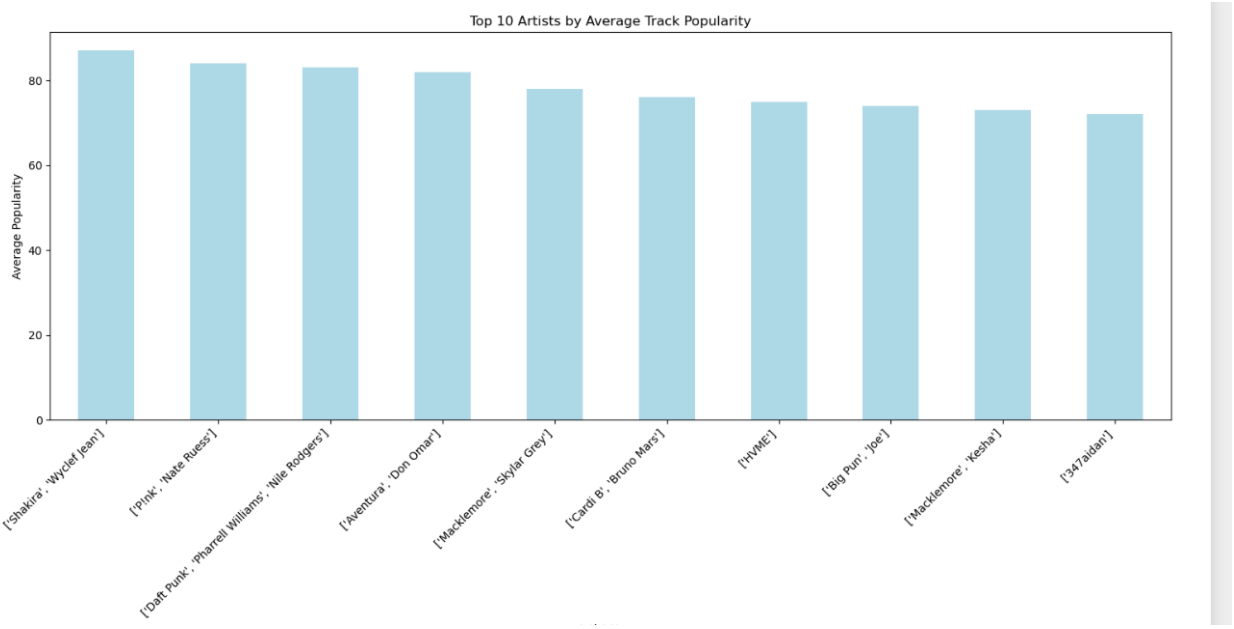
- Variety in Artists and Albums: One of the most striking observations is that each of these tracks comes from different artists and different albums. This suggests that the tracks' popularity might be driven by individual merit rather than a collective album effect.
- Release Years: The tracks span a period from 2013 to 2019. 'Get Lucky (Radio Edit)' by Daft Punk is the oldest among the three, released in 2013. 'Goosebumps' followed in 2016, and 'Please Me' is the most recent, released in 2019. This distribution indicates that popularity doesn't necessarily fade over time, and tracks from earlier years can remain as influential and well-received as newer tracks.
- Diversity in Genre: Daft Punk is renowned for its electronic music, Cardi B & Bruno Mars lean more towards pop and hip-hop, while Travis Scott is predominantly hip-hop. This variety underscores that listeners' preferences span across genres and that high popularity isn't confined to a particular musical style.

Understanding track popularity and the factors influencing it can provide insights for music producers, record labels, and even streaming platforms. It's essential to recognize that while albums play a role, individual tracks can stand out based on various factors, including artist influence, cultural moments, or even global events.

Potential Analysis:

Popularity Analysis:

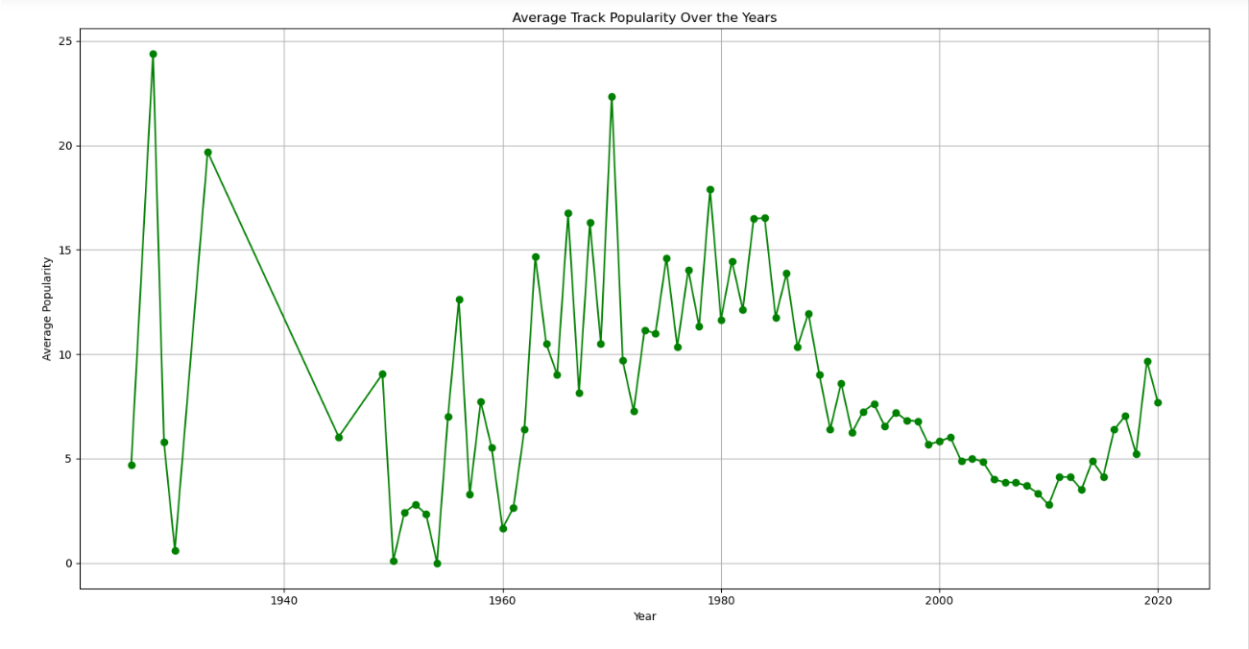
1.Which artists have the highest average track popularity?



Upon analysis of the data, it's evident that certain artists consistently produce tracks that resonate well with the audience.

- Shakira & Wyclef Jean clinch the top position, indicating that their collaborations or individual tracks have, on average, the highest popularity among the dataset. This might be influenced by a few standout tracks that achieved massive success.
- Following closely are Pink & Nate Ruess. Their presence in the top bracket showcases their consistent ability to produce chart-topping and audience-favorite tracks.

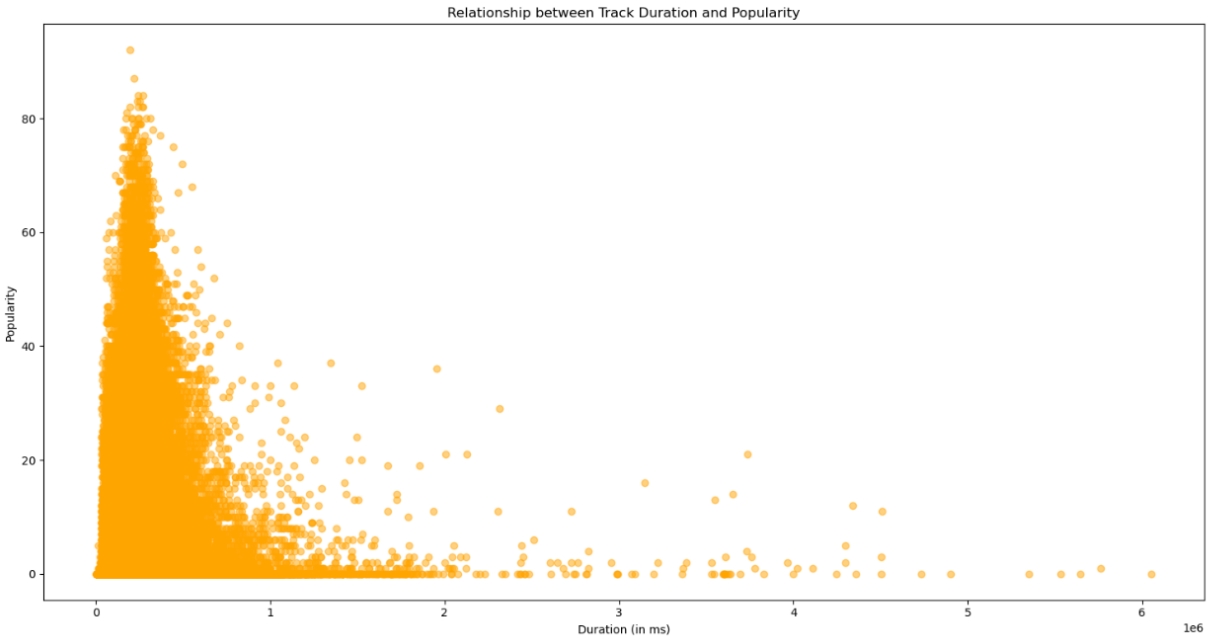
2. How has the average track popularity evolved over the years?



When examining how track popularity has evolved over the years, some fascinating observations emerge:

- Golden Era before 1980: The graph paints a clear picture that songs released before 1980 have, on average, the highest popularity. This could be attributed to a number of factors. Classic songs from this era might have stood the test of time, continuously garnering appreciation and listenership. Additionally, with the advent of digital platforms, these classics might have found a new audience, boosting their popularity scores.

3. Is there a relationship between track duration and its popularity?



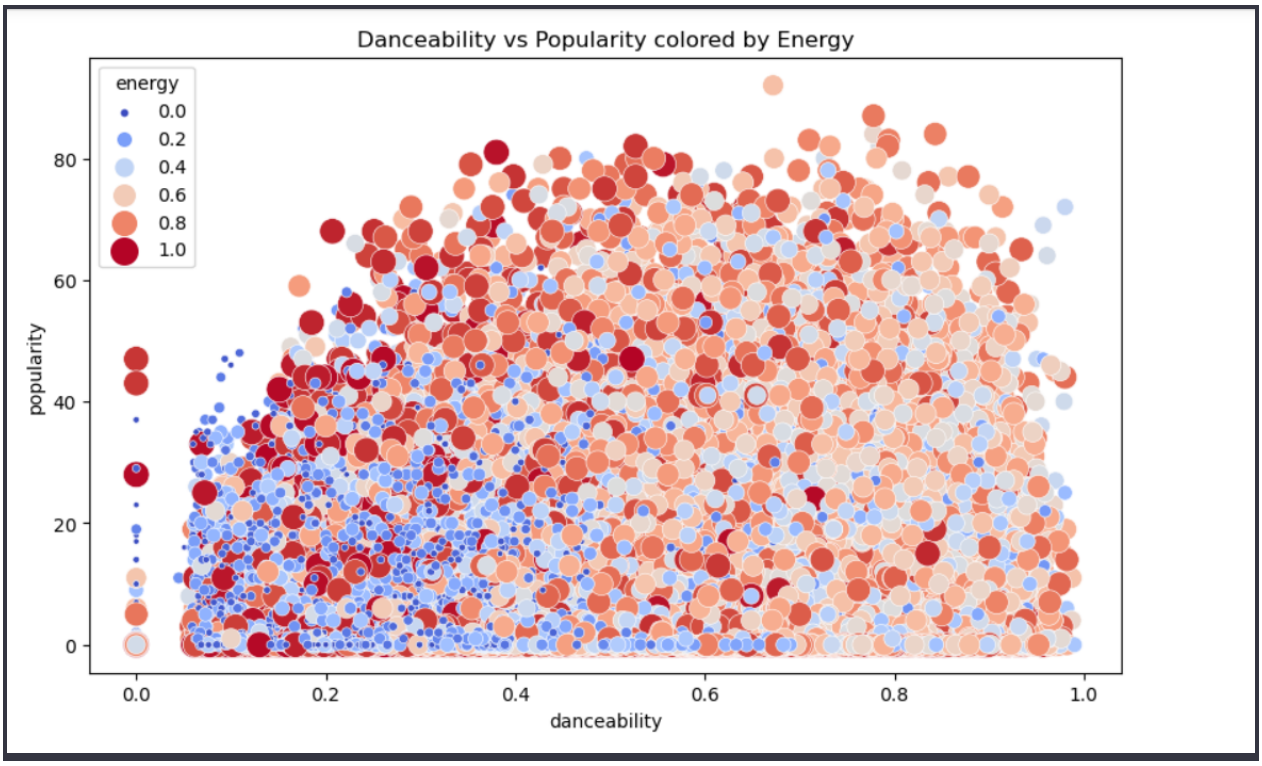
The goal was to uncover if there exists a sweet spot for track duration that correlates with higher popularity. However, the findings present some intriguing data:

- Dense Clustering: A majority of tracks have durations between 0-1 (possibly this represents a fraction of a standardized time metric, like hours) and have popularity scores ranging from 0-60. This dense cluster might indicate that most songs fall within a typical duration range and have moderate popularity.
- Potential Data Issues: The observed clustering, especially the concentration at lower durations, might suggest potential data inconsistencies or issues. It's essential to validate the dataset's integrity, as such clustering seems unusual. Track durations of less than a minute with considerable popularity scores may need further investigation.

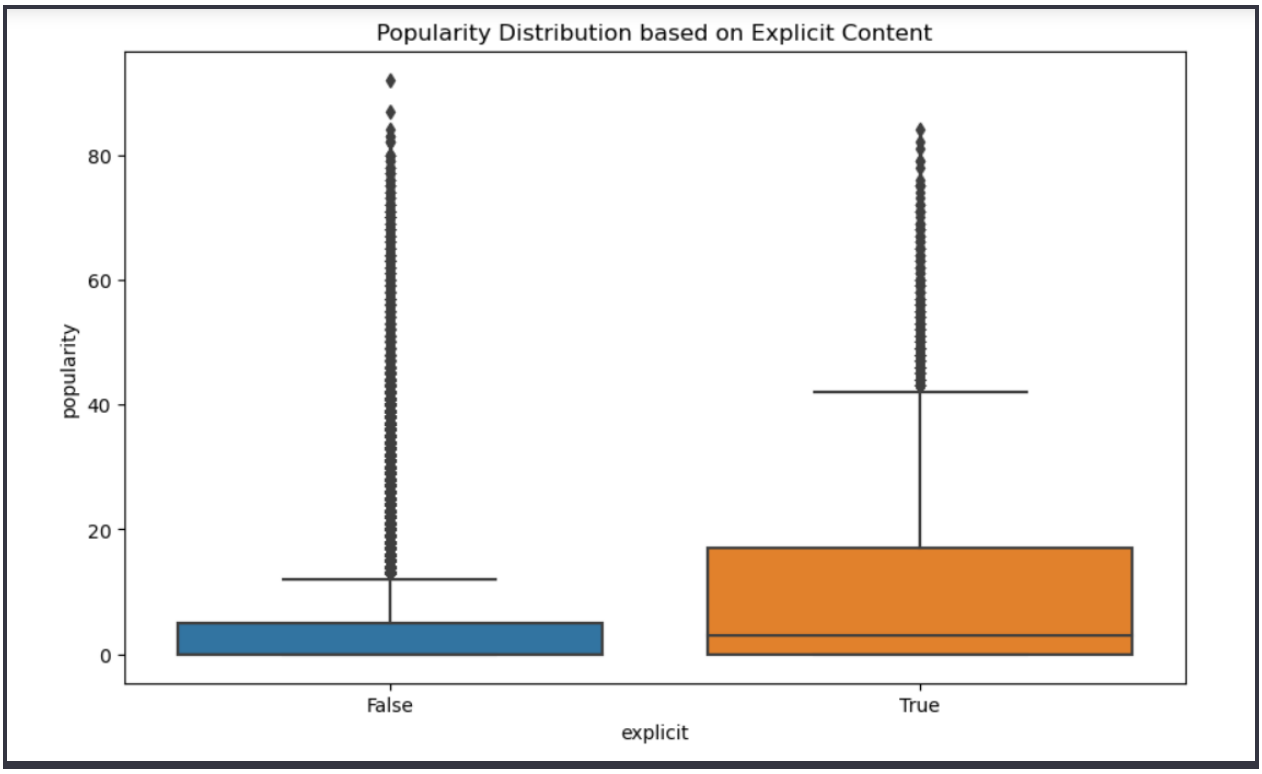
Audio Features Exploration:

How do danceability and energy correlate with track popularity?

Upon exploring the relationship between danceability, energy, and track popularity, I discovered some compelling insights. Tracks with an energy value of 1 predominantly have their danceability scores falling between 0.0 and 0.4. This suggests a specific musical trend: tracks that are full of energy don't necessarily induce dance vibes. The highest popularity scores for these tracks hover between 0-60, indicating that high energy doesn't always equate to widespread appeal.

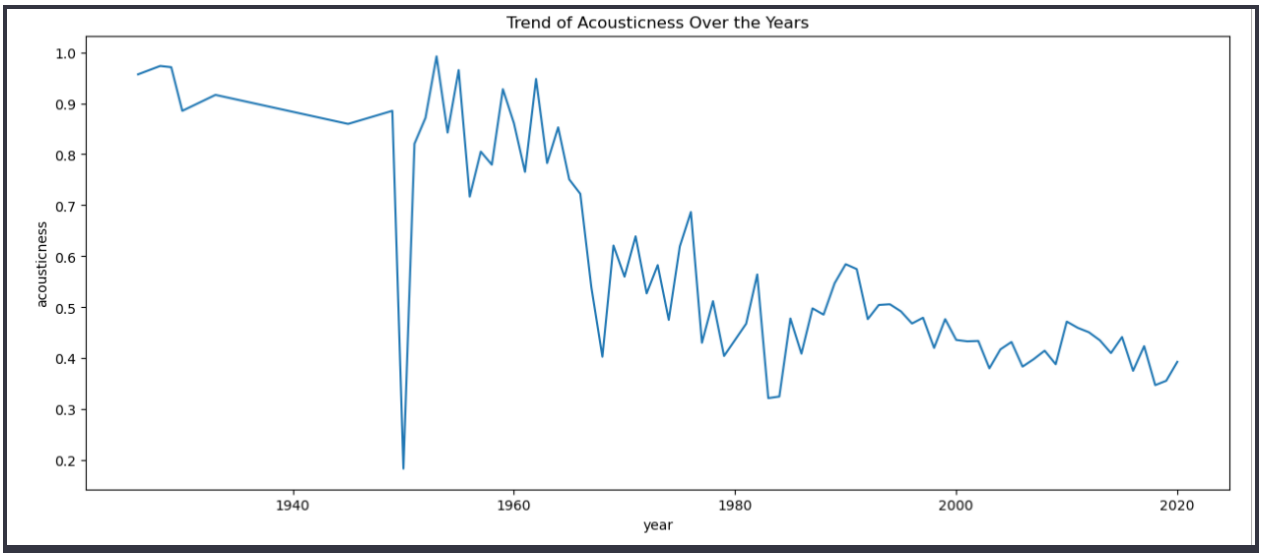


Are explicit songs more popular or less popular than non-explicit ones?
Diving into the popularity metrics of explicit versus non-explicit songs, the data paints an interesting narrative. While the range (whiskers) of popularity is wider for explicit songs (indicating a broader range of popularities), non-explicit songs, despite having a narrower box (indicative of lesser variability in popularity), have more outliers. This suggests that while explicit songs might appeal to a broader audience, non-explicit tracks have exceptions that are hugely popular, potentially transcending age and cultural boundaries.

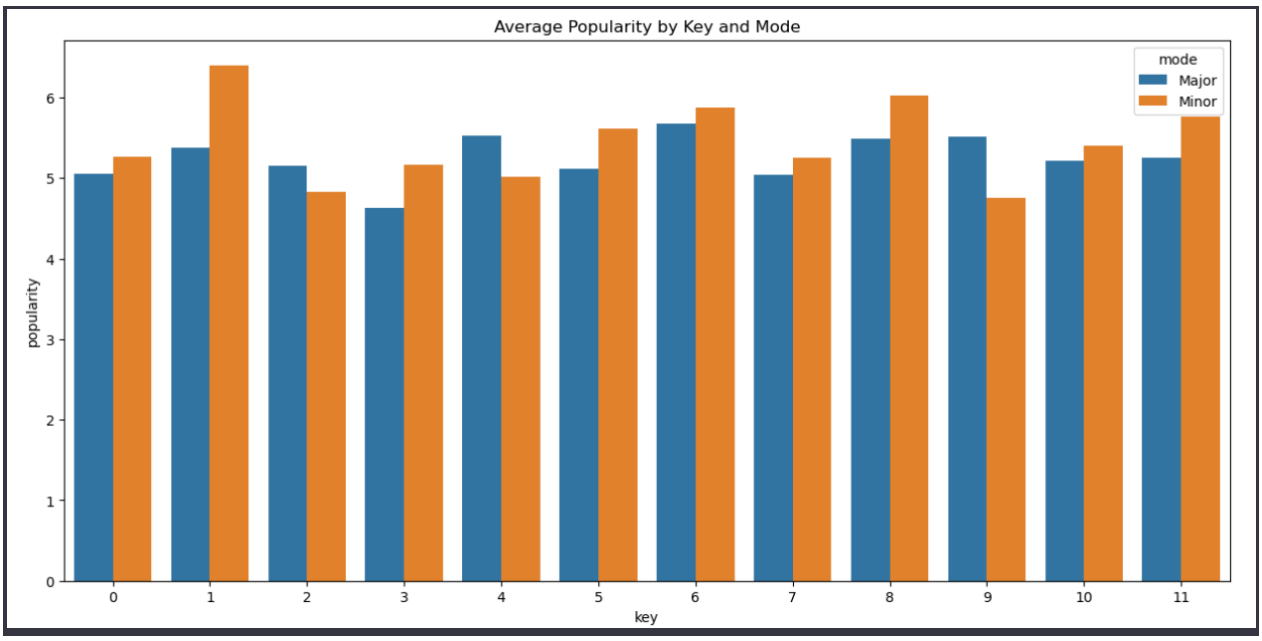


How has the level of acousticness in popular music changed over the years?

The role of acoustic elements in popular music has seen an evolution. Our findings show that between the 1950s to the 1970s, the music industry leaned heavily into acoustic vibes. This could be attributed to the dominant music production techniques of the time, or perhaps a societal preference for more 'natural' sounding tracks. Since then, there might have been a shift due to technological advancements in music production and changing audience preferences.



Which key and mode (major/minor) are the most prevalent in popular songs?



In music theory, the term "mode" refers to the type of scale from which a piece of music derives its melodic and harmonic foundation. Specifically, when discussing Western music, there are two primary modes:

Major: Often associated with a brighter, happier sound.

Minor: Typically linked to a sadder, more somber sound.

In the context of the dataset and the barplot we're looking at:

- mode column likely has two distinct values, commonly represented as:
    - 1 for Major

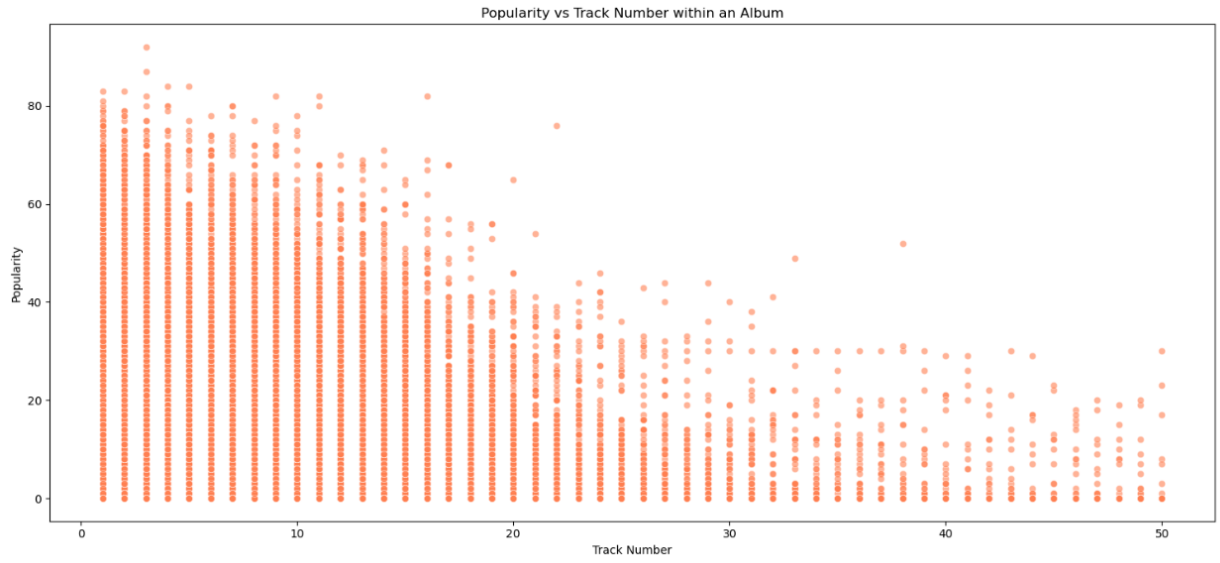
- 0 for Minor

Using the hue='mode' argument in the barplot means that for each key (0-11, corresponding to the 12 possible musical keys), there are two bars: one representing the average popularity of tracks in the major mode of that key and the other representing the average popularity in the minor mode of that key.

When deciphering the tonal preferences of popular tracks, the data showed some captivating trends. Key number 1 (which typically corresponds to C#) with a minor mode was the most prevalent among popular tracks, and key number 6 (usually F) with a major mode stood out. This indicates a possible cultural or psychological preference for these keys and modes in popular music, perhaps resonating more with audiences due to their emotional appeal.

Artist and Album Study:

How does the track number within an album relate to its popularity? (e.g., are the first few tracks of an album typically more popular?)

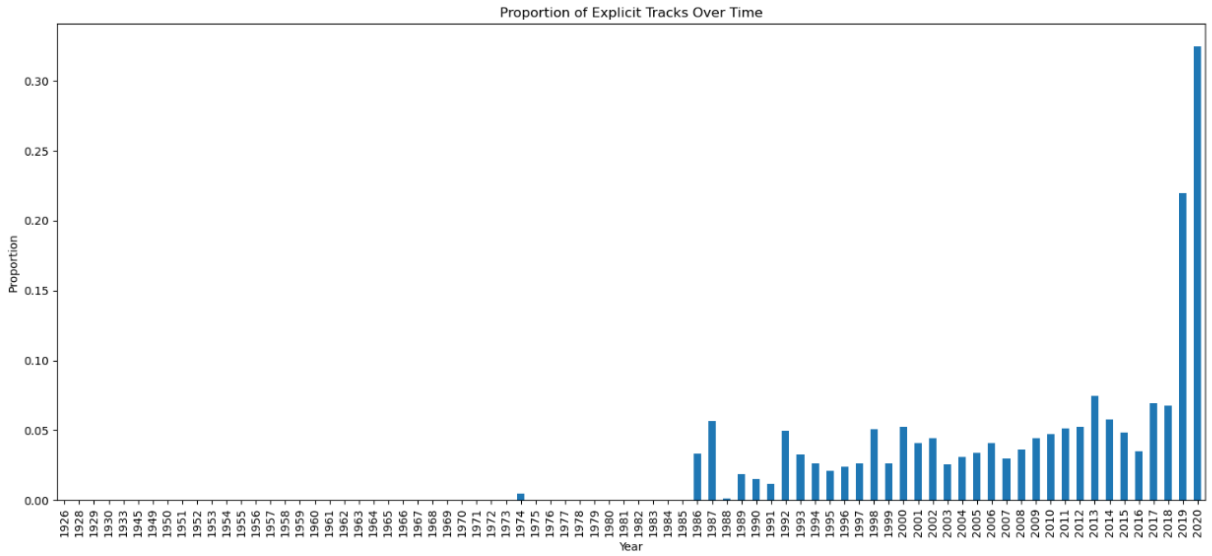


The scatter plot paints a vivid picture of an age-old strategy in the music industry: placing potential hit songs at the start of an album. There's a clear trend: the early tracks, especially those at the very beginning, tend to bask in the limelight of popularity. Track number 3 seems to wear the crown, with a glaring popularity that stands out. But as we journey towards the right, the crowd of points begins to thin. This suggests that

as we delve deeper into an album, tracks might not have the same widespread charm. However, a few surprise hits do pop up later in the sequence, standing tall amidst their peers, reflecting those unexpected gems that resonate differently with listeners.
This visual isn't just about the play counts or chart positions. It's a story of the artists' vision, the listeners' heartbeat, and sometimes, the unpredictable turns music can take. Sometimes, it's not just about the sequence, but the serendipity.

Temporal Analysis:

Are there specific years or decades where explicit tracks became more prevalent?



Music, like any form of art, often mirrors the cultural, societal, and technological dynamics of its time. When looking at the explicit content in tracks over the years, a historical and societal tapestry unfolds before our eyes.

Between 1962 and 1973, the chart shows a virtual absence of explicit content. This could be indicative of the social norms and taboos of that era, where explicit language in

music wasn't mainstream or readily accepted. However, 1974 hints at a slight liberalization, possibly due to cultural shifts or the rise of particular music genres or artists who broke the norm.

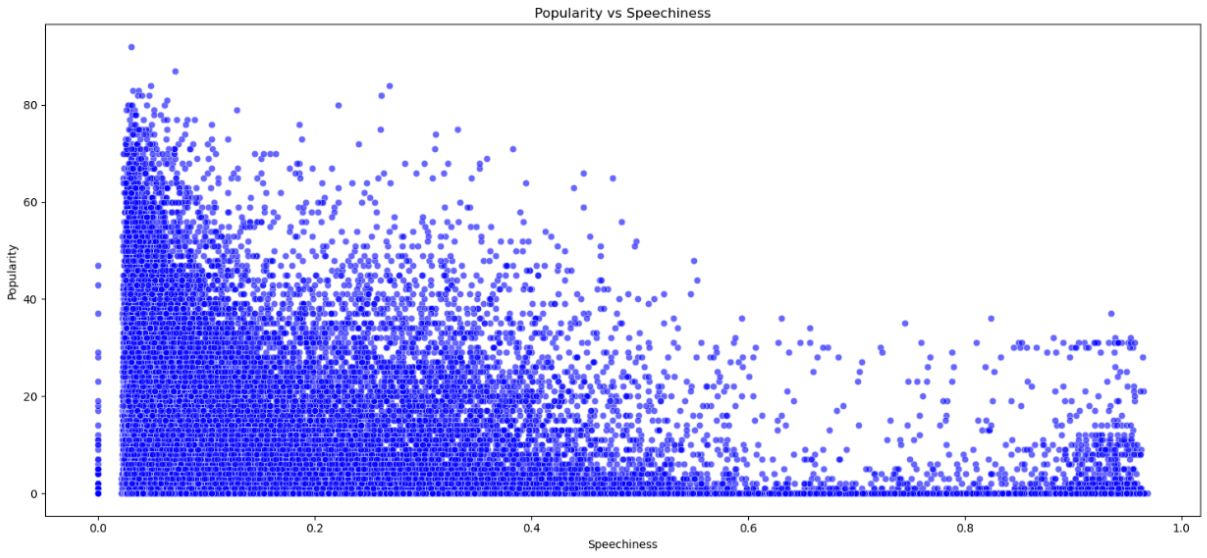
The late '80s witnessed an intriguing pattern. The presence of explicit content began to solidify in 1986 and 1987, only to face a sharp drop in 1988, lower than even in 1974. This dip might be attributed to regulatory backlashes, societal shifts, or major events in the music industry. Nevertheless, 1989 began to show the music industry's resilience, setting the stage for the coming decades.

From 1989 onward, while there were fluctuations, a general trend of acceptance or even preference for explicit content emerged, reflecting a broader societal acceptance of open expression in art. The rise post-2018 is particularly striking. By 2020, the proportion has climbed above 0.30, indicating that almost a third of the tracks have explicit content. This suggests a generation more at ease with unfiltered expression and a music industry unafraid to explore and produce such content.

Note—calculation of the proportion: It's the average of the explicit column for each year. Since 'explicit' is likely a binary variable (1 for tracks that are explicit and 0 for those that are not), taking the mean effectively calculates the proportion of tracks that are explicit in a given year. This provides an insightful measure, allowing us to understand the prevalence of explicit tracks within the overall music landscape for each year.

Speech and Instrument Analysis:

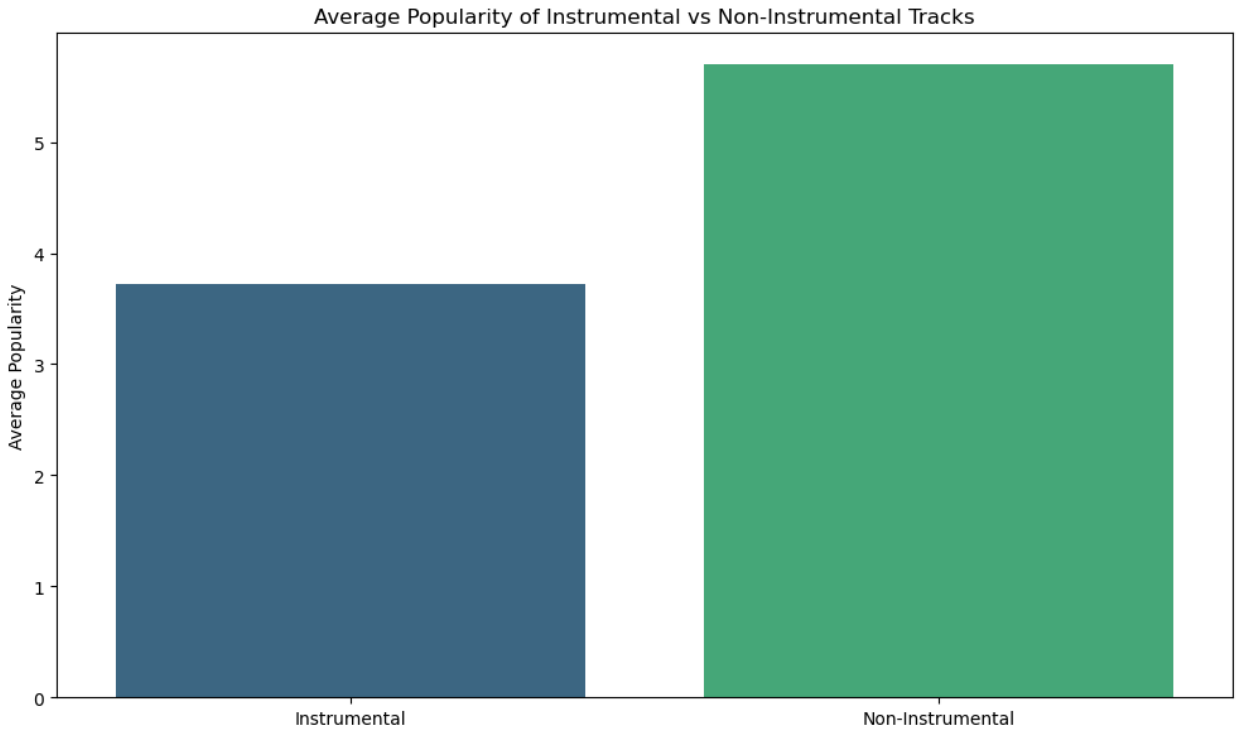
Is there a relationship between speechiness and popularity? (This could indicate a trend toward more spoken-word or rap music.)



The musical landscape is a vivid representation of our diverse tastes, with each data point telling a unique story. Songs with minimal spoken elements, clustered densely on the left of our graph, dominate the popularity charts. Their widespread appeal suggests a resonance with audiences seeking the familiarity of traditional melodies or the allure of conventional tunes.

Venturing to the right, where tracks are richer in spoken content, there's a noticeable reduction in density and popularity. However, even in this segment, certain tracks manage to captivate listeners. Despite the overall trend favoring low-speechiness songs, these particular tracks emphasize the powerful impact of spoken words in music and the niche audience that cherishes them. The absence of outliers in the low speechiness region further underlines the consistent popularity of such tracks.

How prevalent are purely instrumental tracks, and how do they fare in popularity compared to non-instrumental tracks?

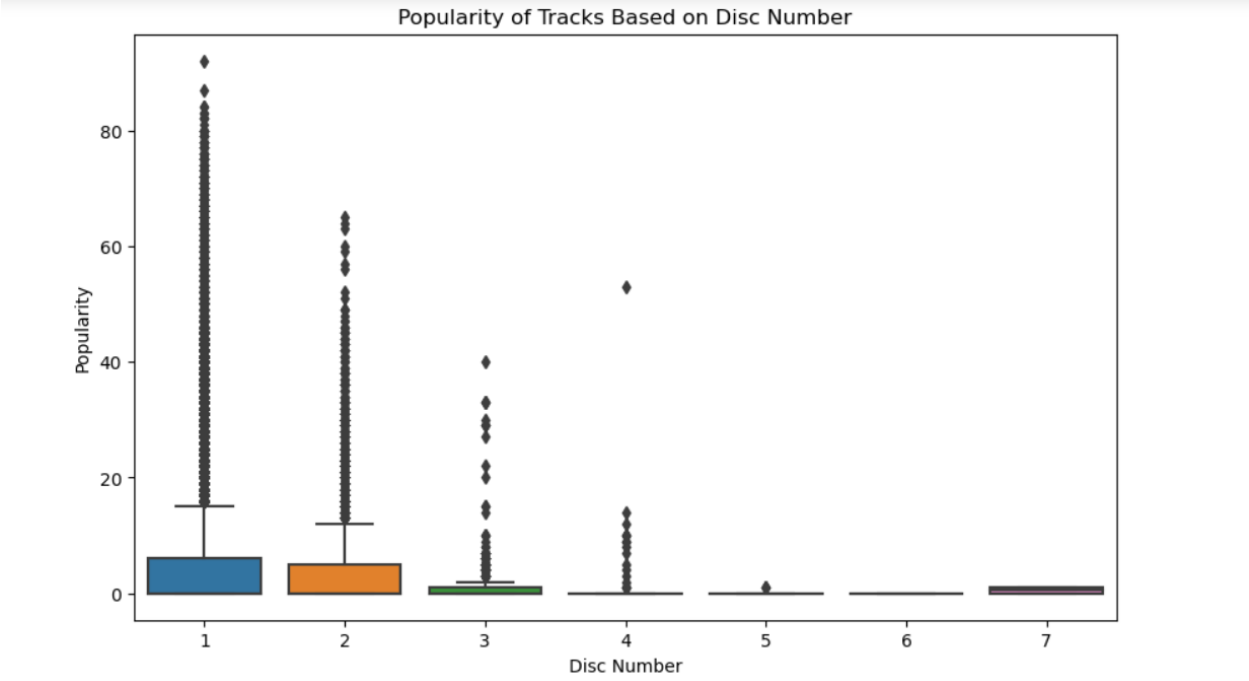


In the vast panorama of music, the instruments speak a language that transcends words. And yet, when we delve into the data, a curious trend emerges. Instrumental tracks, though cherished for their capacity to convey emotions without uttering a single word, don't seem to command the same widespread popularity as their lyrical counterparts. The vast majority of chart-toppers and audience favorites feature vocals, suggesting a universal human inclination towards songs that tell stories, express feelings, or share experiences through words.

This is not to diminish the value or beauty of instrumental music. Many listeners seek out instrumentals for their purity, their evocativeness, or simply their ability to set a mood without the influence of lyrics. But when pitted against the vast sea of lyrical melodies, instrumentals remain a unique, perhaps niche, preference. In this ever-evolving musical realm, while vocals might reign supreme in popularity charts, instrumentals continue to hold their ground, offering a wordless sanctuary to those who venture towards them.

Album Study:

Do multi-disc albums have tracks that are more or less popular than single-disc albums?



In the realm of music, albums are carefully curated by artists and their teams to present a cohesive narrative or theme. The choice between releasing a single-disc album versus a multi-disc album often hinges on various factors like the amount of content, story arc, or artistic intentions. Our visual analysis reveals interesting facets of this choice.

The plot clearly indicates that single-disc albums (disc_number = 1) dominate in terms of track popularity. These albums exhibit not only a higher median popularity but also a more extensive range, evident from the larger box and the presence of numerous outliers on the upper side. This might suggest that single-disc albums tend to have more "hit" songs or songs that achieve significant recognition.

In contrast, as the disc number increases, the popularity spread shrinks. One potential reason could be that multi-disc albums might cater to a more niche audience or present deep cuts that aren't necessarily geared for broad commercial appeal. Additionally, the sheer volume of songs in multi-disc albums could dilute the standout hits, leading to a more uniform reception.

In essence, while single-disc albums seem to capture a broad spectrum of popularity, ranging from average to highly popular tracks, multi-disc albums seem to provide a more consistent, perhaps niche, experience. This finding underscores the significance of understanding audience preferences and aligning album formats accordingly.

Feature Engineering:

Categorical Feature Creation:

- **Release Period:**The year of release was categorized into time periods: '60s and before', '70s', '80s', '90s', '2000s', and '2010s and after'.
- **Tempo**:Categorized into 'Slow' (tempo < 60), 'Medium' (60 <= tempo < 120), and 'Fast' (tempo >= 120).
- **Duration**:The track duration was segmented into 'Short' (< 2 minutes), 'Medium' (2-4 minutes), and 'Long' (>= 4 minutes).

One-Hot Encoding:

- The created categorical variables were then one-hot encoded to convert them into a machine-readable format.

Model Training and Evaluation:

Data Preprocessing:

- Renamed the target variable from 'normalized_artist_popularity' to 'popularity'.
- Converted the 'mode' column, which was categorical, to one-hot encoded columns.

Model Selection:

- Started with a basic Linear Regression model, which provided a relatively low R^2 score.
- To improve performance, switched to a Random Forest Regressor which generally handles non-linear relationships better and provides a mechanism against overfitting.

Model Evaluation:

- With the Random Forest model, we noticed an R^2 score of 0.8859 on the training set but only 0.1852 on the test set, indicating overfitting (i.e., the model was performing exceptionally well on the training data but poorly on unseen data).

Cross-Validation:

To further ensure the robustness of the model, cross-validation was employed. This is an effective method to get a more unbiased estimate of model performance. It involves partitioning the dataset into k subsets, training the model on k-1 of those subsets, and validating the model on the remaining subset. This process is repeated k times, with each subset serving as the validation set exactly once.

However, even after using cross-validation, the R^2 score remained around 0.18 for the test data. This further indicated that the model, despite its complexities, might not have captured the underlying patterns of the data well, or that some inherent noise or variability in the dataset was causing the predictive power to be limited.

Next Steps:

Feature Engineering: Dive deeper into the features to understand their relationships with the target variable. This could involve polynomial features, interaction terms, or even domain-specific features that could be curated with some additional external knowledge.
Model Experimentation: Experiment with other models and algorithms, like gradient boosting machines (XGBoost, LightGBM), neural networks, or ensemble methods to see if they yield better results.

Data Quality: A deep dive into the data quality is essential. Checking for outliers, ensuring that there's no data leakage, and perhaps even revisiting the data collection method might provide insights.

Hyperparameter Tuning: A systematic search for the best hyperparameters (using techniques like GridSearchCV or RandomizedSearchCV) for the chosen model can also be beneficial.

Conclusion:

The project provided valuable insights into the challenges of predicting song popularity based on given features. While initial models showed promise on training data, their performance on test data highlighted the pitfalls of overfitting and the complexities of the underlying dataset. Despite these challenges, the journey offered multiple learning points, especially about the importance of understanding the data, iterative model refinement, and the need for a holistic approach that considers data quality, feature

engineering, and model selection together. As with many real-world data science tasks, the project underscores the fact that while machine learning offers powerful tools, their effective application requires a combination of technical know-how, domain understanding, and a rigorous iterative approach.