Exploring Music Trend:

Comprehensive Analysis of Popularity Factors and Genre Features in Songs for Spotify

NYU MSDS Program: DS-GA 1001 Capstone Project

Group Members Runhan Chen(rc4062@nyu.edu), Jinrui Fang (jf4959@nyu.edu)

Group Name Dafinity (Group 52)

Data Preprocessing Please refer to *Content 1.2*

The Seed N-number 19202232 (Jinrui Fang)

Date 19 Dec 2023

Author Contribution

Our project is the result of a collaborative effort where both of us, Runhan Chen and Jinrui Fang, equally contributed to all phases of the project, with continuous input and shared responsibilities. Throughout the project, ChatGPT provided support by debugging code, enhancing report format, and ensuring grammatical accuracy.

NYU

Contents

NYU

1. Introduction

In the diverse landscape of the music industry, data science is emerging alongside musicians and producers. As data scientists, we are uniquely positioned to bring analytical and data-centric approaches into an arena traditionally guided by artistic intuition. This project, leveraging Spotify's rich database, aims to dissect and understand the factors driving song popularity and the distinctive features of various musical genres. By employing data science techniques, we seek to augment the traditional narrative of music analysis with quantifiable insights, offering a fresh perspective on the evolving patterns and trends that define the music world.

1.1 The Datasets

The primary dataset of our analysis is a comprehensive dataset, *spotify52Data.csv*, provided by Spotify, encompassing a diverse collection of 52,000 songs. This dataset represents a wide range of genres from acoustic to hiphop. Each song in the dataset is described through a series of attributes that capture its unique characteristics, including basic information such as the track ID, artist, album name, and track title. Additionally, for our data-driven analysis, the dataset offers detailed audio features for each song, including duration, key, mode, tempo, and various measures that reflect the song's acoustic properties – such as danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, and valence. More importantly, each track is scored with a 'popularity' metric, an integer value ranging from 0 to 100, provided by Spotify, reflecting the song's reception and frequency of plays. This dataset forms the foundation for a deep dive into the analytics of music popularity and genre classification.

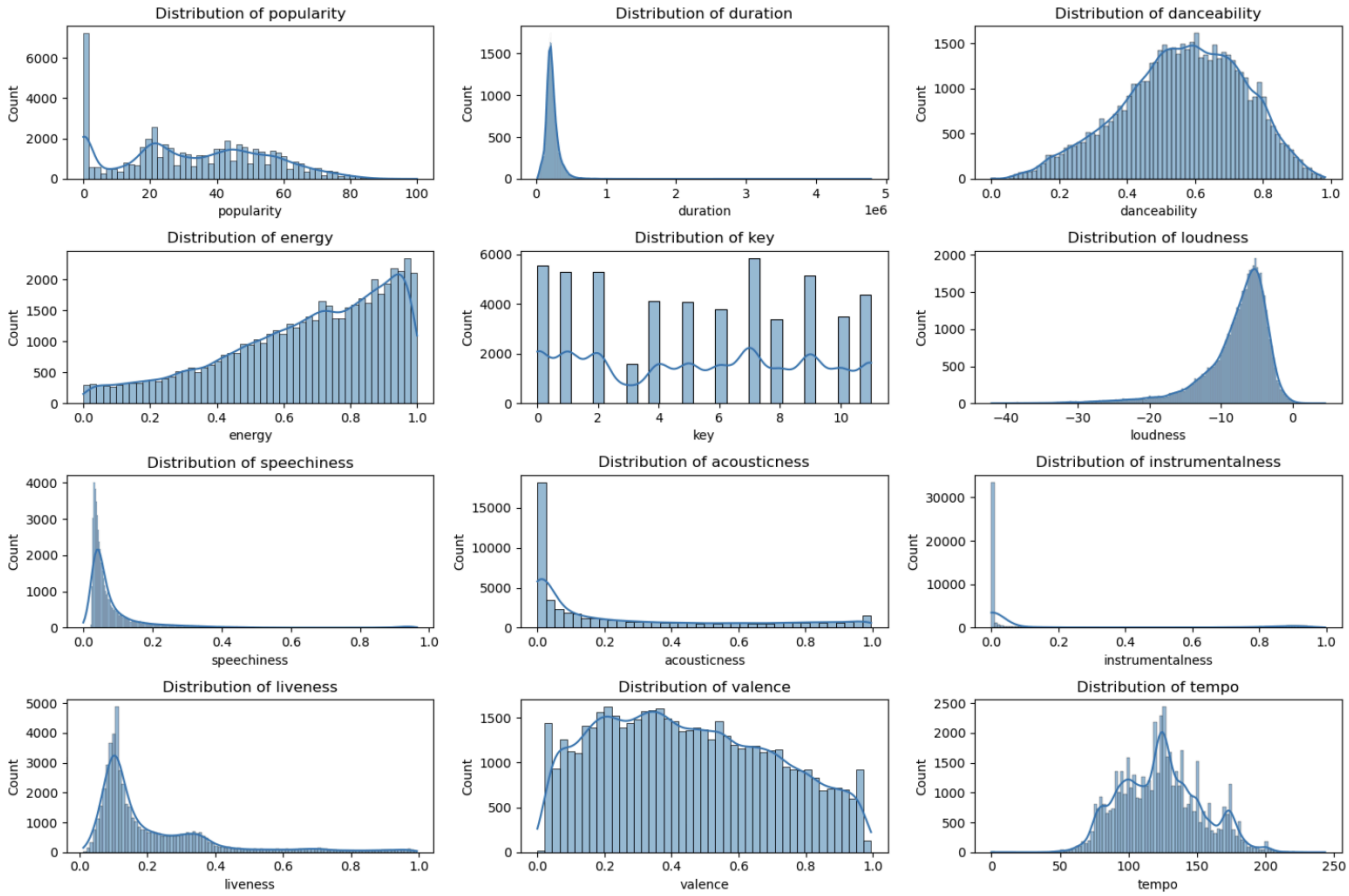
The secondary dataset, *starRatings.csv*, provides user feedback in the form of star ratings for the first 5,000 songs from the Spotify dataset. Ratings range from 0 to 4, given by 10,000 users, offering a direct measure of listener preference. This dataset, which includes some missing values, offers valuable insights into how users perceive and engage with these songs.

1.2 Data Preprocessing

- ***Data Cleaning.*** The initial step in our preprocessing involved a thorough check for missing values across all variables in two datasets. While no missing values were found in *spotify52Data.csv* dataset, we found a significant number of missing values in *starRatings.csv* dataset. To address this, we imputed missing ratings with the average rating of each song across all users. This imputation strategy preserves the overall rating trends and song popularity measures. By imputing the average, we strike a balance between data integrity and practicality, allowing for a comprehensive analysis that includes all available user feedback.
- ***Duplicate Removal.*** Considering the possibility of the same song appearing in multiple genres or albums, we removed duplicates from the dataset based on artist, album, and track names, creating a unique song dataset. This ensures that our analysis is not biased by multiple instances of the same song.

- ***Assessing Distribution.*** We visualized the distribution of various numerical features to identify skewness, as illustrated in Figure 1.2.1. We observed that while 'popularity' exhibited minor skewness, 'duration' was highly positively skewed, indicating a greater number of shorter songs. Additionally, 'loudness' and 'speechiness' showed noticeable deviations from a normal distribution. These findings were substantiated by skewness statistics detailed in Table 1.2.2, prompting considerations for data transformations to normalize these features.

Figure 1.2.1. Distribution Plots of Song Features Related to Popularity



- ***Conclusion.*** Through meticulous data cleaning, duplicate removal, and distribution assessment, we have prepared a robust dataset for the subsequent stages of our analysis. These preprocessing steps are instrumental in ensuring that the data is accurately represented and that any analytical models or insights derived from this data are based on a solid foundation of quality and reliability.

Table 1.2.2. Skewness Table of Song Features

popularity	0.079294
duration	11.565373
danceability	−0.299694
energy	−0.730231
key	−0.012035
loudness	−2.019584
speechiness	4.318930
acousticness	0.966942
instrumentalness	1.477542
liveness	2.074428
valence	0.230075
tempo	0.240884

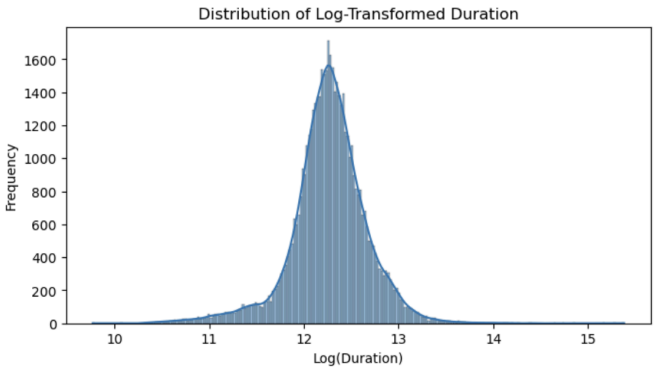
* *Note:* The left column displays Feature names, and right column specifies skewness.

2. Hypothesis Testing

2.1 Song Length and Popularity

1) Is there a relationship between song length and popularity of a song? If so, is it positive or negative?

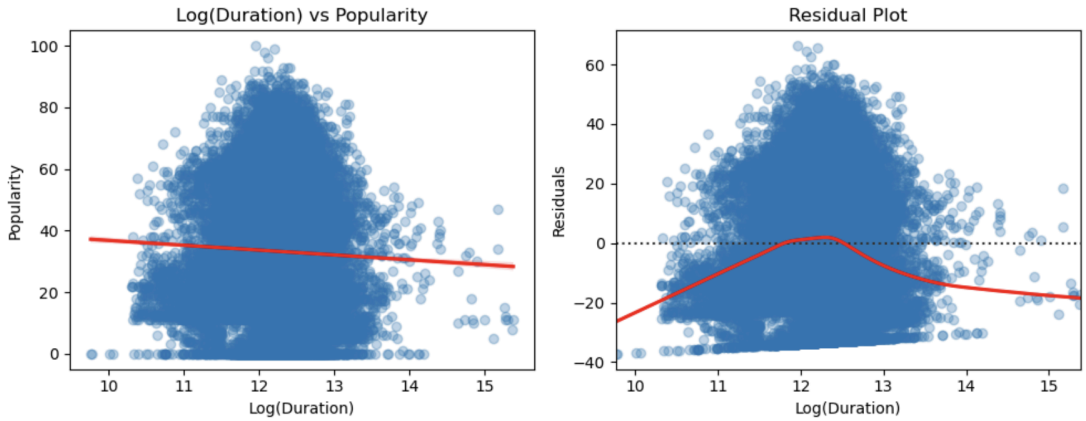
Figure 2.1.1. Distribution of log-Transformed Duration



- ***Methodology.*** In assessing the relationship between song length and popularity, our data exhibited a high positive skew in song duration. A logarithmic transformation normalized the distribution effectively, as illustrated by the histogram of log-transformed durations (See *Figure 2.1.1*). Pearson correlation and simple linear regression were applied to evaluate the relationship between song length and popularity.

- ***Findings.*** The Pearson correlation coefficient was -0.019, indicating a statistically significant, but weak, negative correlation with a p-value of less than 0.001. The linear regression analysis echoed these results; the coefficient for log-transformed duration was -1.0315, significant at $p < 0.001$. However, the R-squared value was approximately 0.000, showing that the duration explains virtually none of the variability in popularity (see *Figure 2.1.2*).

Figure 2.1.2. Relationship Between Log Duration and Song Popularity



* *Note:* The left plot demonstrates the weak correlation between log-transformed song duration and popularity. The right plot illustrates the variance of the residuals and shows the minimal influence of duration on song popularity.

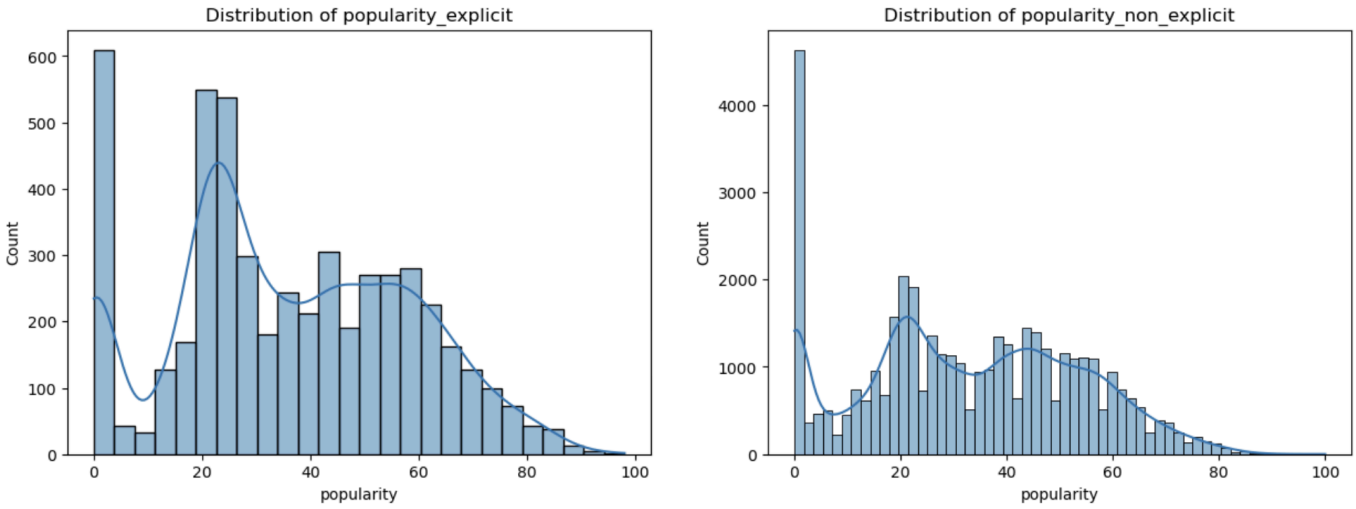
- ***Conclusion.*** While there is a statistically significant negative relationship between song length and popularity, the practical impact is negligible. This suggests that duration is not a strong predictor of a song's popularity on Spotify, and other factors likely play a more substantial role.

2.2 Explicitly Rated Songs and Popularity

2) Are explicitly rated songs more popular than songs that are not explicit?

- ***Methodology.*** To compare the popularity of explicit and non-explicit songs, we segmented the dataset into two groups accordingly. Then we assessed the variance in popularity scores for both groups and visualized the distributions to understand their characteristics (see *Figure 2.2.1*). Given that the data is not normally distributed and not categorical, we decide to compare the medians of the two groups by applying the Mann-Whitney U test, with the assumption of independent song data.

Figure 2.2.1. Distribution of Song Popularity for Explicit and Non-explicit Rated Songs



* *Note:* The left histogram illustrates the popularity distribution of explicit-rated songs, and the right histogram shows that of non-explicit-rated songs, providing a visual comparison of their popularity metrics.

- ***Findings***. The histograms for explicit and non-explicit songs show distinct popularity distributions. Explicit rated songs tend to have higher popularity scores, with the variance difference in popularity being approximately 55. The median in popularity scores for explicit rated songs(median = 34) was higher than for non-explicit rated songs(median = 32). The Mann-Whitney U test yielded a u-statistic of 111578985.5 and a highly significant p-value (1.45×10^{-17}), indicating that the observed differences in popularity are statistically significant.

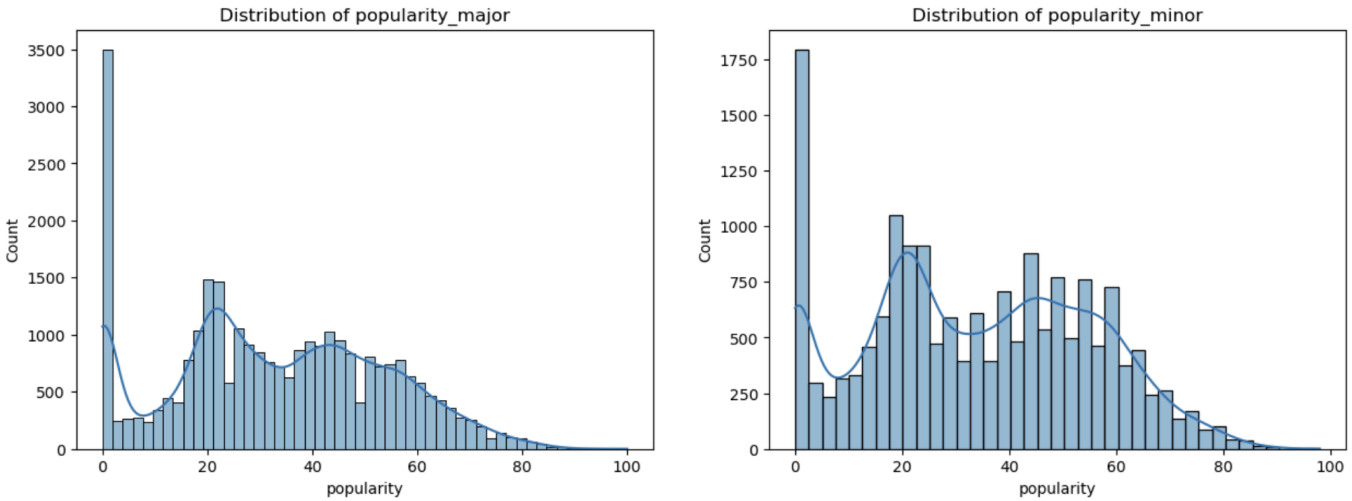
- ***Conclusion***. Our analysis indicates that the explicit rated songs are more popular than non-explicit rated songs, with a statistically significant difference in median popularity.

2.3 Songs in Major Key and Popularity

3) Are songs in the major key more popular than songs in the minor key?

- ***Methodology.*** To compare the popularity of songs in major key and songs in minor key, we segmented the dataset into two groups accordingly. Then we assessed the variance in popularity scores for both groups and visualized the distributions to understand their characteristics (see Figure 2.3.1). Given that the data is not normally distributed and not categorical, we decide to compare the medians of the two groups by applying the Mann-Whitney U test, with the assumption of independent song data.

Figure 2.3.1. Distribution of Song Popularity for Songs in Major and Minor Key



* *Note:* The left histogram illustrates the popularity distribution of songs in major key, and the right histogram shows that of songs in minor key, providing a visual comparison of their popularity metrics.

- ***Findings***. Visual inspection of the histograms indicates that the distribution of popularity for major and minor key songs differs slightly. The variances for major (variance = 426.95) and minor (variance = 444.70) key songs were comparable. The Mann-Whitney U test resulted in a u-statistic of 238963582.5 with a highly significant p-value (9.82×10^{-7}), suggesting a statistically significant difference in popularity. However, the median popularity scores for major and minor key songs were 32 and 33, respectively, with minor key songs being slightly more popular by a mean difference of -1.

- ***Conclusion***. We found that songs in a minor key are marginally more popular than those in a major key, with a small but statistically significant difference in mean popularity scores.

NYU

3. Regression Analysis

3.1 Predictive Factors for Song Popularity

4) Which of the following 10 song features: duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo predicts popularity best? How good is this model?

- ***Methodology***. To determine which song features best predict popularity, we conducted 10 simple linear regressions, one for each of the following features: duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. We split the dataset into training and test sets, ensuring reproducibility with a fixed random state. Each feature was individually regressed against the popularity score to compute the coefficient of determination, R-squared, which measures the proportion of variance in the popularity explained by the feature.

- ***Findings***. The analysis produced R-squared values for each feature, indicating the strength of their linear relationship with song popularity. The results, as shown in the *Figure 3.1*.1, suggest that instrumentalness has the highest R-squared value (0.025973), indicating it is the best predictor among the features tested, although it still explains a relatively small proportion of the variance in popularity. Other features such as danceability and loudness also show a relationship with popularity but to a lesser extent. The Tempo shows the lowest R-squared value(0.000088).

	Features	R_squared
0	duration	0.003769
1	danceability	0.001541
2	energy	0.003898
3	loudness	0.002816
4	speechiness	0.003400
5	acousticness	0.000920
6	instrumentalness	0.025973
7	liveness	0.002437
8	valence	0.000756
9	tempo	0.000088

Figure 3.1.1. R-squared Values for Individual Song Features Predicting Popularity

- ***Conclusion***. Instrumentalness emerged as the strongest predictor of song popularity among the features analyzed, followed by loudness and danceability. However, all features show a low predictive power individually, as reflected by their R-squared values. This suggests that while these features have some influence, they do not strongly predict song popularity on their own. A model incorporating multiple features might better capture the complexity of factors that contribute to a song's popularity.

NYU

3.2 Comprehensive Model Performance

5) Building a model that uses *all* of the song features mentioned in question (4), how well can you predict popularity? How much (if at all) is this model improved compared to the model in question (4). How do you account for this? What happens if you regularize your model?

- ***Methodology***. We constructed a multivariate linear regression model incorporating all song features identified in the previous question: duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. The dataset was split into training and testing sets to validate the model. Then, we employed Ridge and Lasso regression to assess the impact of regularization on model performance, and to see if they help reduce overfitting by penalizing the size of the coefficients.

- ***Findings***. Our complete model resulted in an R-squared(Coefficient of Determination) of about 0.055, indicating a small improvement in explaining the popularity variance compared to the single-feature models from the previous question. While we are trying Ridge and Lasso regressions to refine the model, the R-squared did not improve significantly, suggesting that the additional complexity did not capture more variability in popularity. See the detailed results in Table 3.2.1 below.

Table 3.2.1. Comparison of Regression Models for Predicting Song Popularity

Model Type	R-squared	RMSE
Multivariate Model	0.05465613	20.260946
Ridge Regression	0.05465619	20.260945
Lasso Regression	0.05020713	20.308566

* *Note:* R-squared values indicate the proportion of variance in song popularity explained by the model. Higher values suggest a better fit to the data.
** *Note:* RMSE (Root Mean Squared Error) values indicate the average distance between the predicted popularity scores and the actual scores. Lower values suggest a more accurate model.

- ***Conclusion***. The analysis suggests that while the selected features do provide some insight into what makes a song popular, they only explain a small part of the overall picture. The limited change in performance with regularization techniques like Ridge and Lasso indicates that there may be other factors influencing popularity that are not included in the model. To improve our predictions, we should consider exploring more detailed data or different modeling techniques that can capture the more subtle aspects of what makes a song popular on platforms like Spotify.

4. Dimensionality Reduction and Clustering

4.1 Dimensionality Reduction and Genre Clustering Analysis

6) When considering the 10 song features in the previous question, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for? Using these principal components, how many clusters can you identify? Do these clusters reasonably correspond to the genre labels in column 20 of the data?

- ***Methodology***. We applied Principle Components Analysis(PCA). It reduces feature dimensionality and helps us capture the most variance with fewer components. We also performed KMeans clustering on the PCA-transformed data to find natural groupings of songs. Then we applied silhouette scores to find the optimal number of clusters.

- ***Findings***. PCA revealed that the first two principal components account for approximately 44% of the variance within our song features dataset. The scree plot (See *Figure 4.1.1*) displays a clear elbow after the second component, showing diminishing returns on explained variance with additional components. The silhouette scores (See *Figure 4.1.2*) peaked at two clusters, indicating that two distinct groupings exist within the PCA-transformed feature space.

Figure 4.1.1: PCA Scree Plot

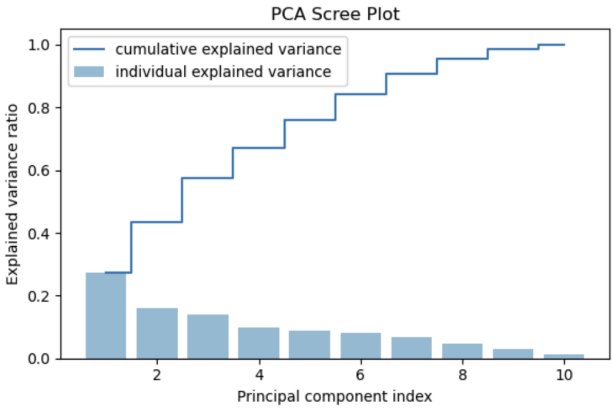
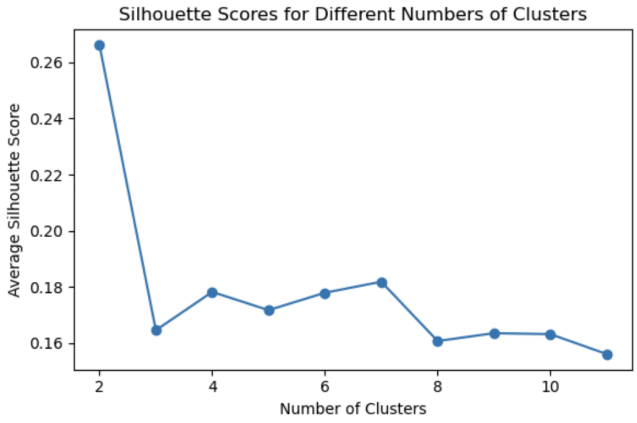


Figure 4.1.2. Silhouette Scores for Optimal Cluster Count



*Note: The scree plot displays the variance each principal component contributes, guiding the selection of components for dimensionality reduction.
**Note: Silhouette scores assess the fit of data points within a cluster compared to other clusters, with the peak indicating the optimal cluster count.

- ***Conclusion***. Our analysis found two principal components to extractThey account for roughly 44% of the total variance. Using clustering, they reveal two distinct groups. But, these clusters do not align with the 52 unique genre labels provided, suggesting that genre classification may involve patterns that are not captured by the features considered in this study. This insight points to the complex nature of musical genres and the potential need for a broader set of data to achieve a more accurate clustering that reflects genre distinctions.

5. Classification Models

5.1 Key Prediction from Valence

7) Can you predict whether a song is in major or minor key from valence using logistic regression or a support vector machine? If so, how good is this prediction? If not, is there a better one?

- ***Methodology***. Our approach began with addressing the imbalance in the dataset through Synthetic Minority Over-sampling Technique (SMOTE) and standardizing the 'valence' feature to accommodate the sensitivity of Support Vector Machines (SVMs) to data scaling. We conducted a comparative analysis using logistic regression, SVM, Random Forest, and Decision Trees to predict the modality (major or minor) of songs based on their valence. To enhance our models, we integrated additional song features identified in Question 4. The performance of each model was evaluated based on the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC).

- ***Findings***. The analysis demonstrated that prior to including additional features, the Random Forest model achieved the highest AUC of 0.532 shown in *Figure 5.1.1*, indicating a marginal ability to distinguish between major and minor keys based on valence alone. Upon incorporating a broader set of song features, the predictive power of the Random Forest model improved significantly, with the AUC increasing to 0.706 shown in *Figure 5.1.2*. This suggests a more robust model that can better capture the complexity of the key modality as related to the combined features of the songs.

Figure 5.1.1. Performance Metrics for Random Forest

The AUC: 0.5325676675538418
Report for Random Forest

	precision	recall	f1-score	support
0	0.53	0.55	0.54	5757
1	0.53	0.52	0.53	5732
accuracy			0.53	11489
macro avg	0.53	0.53	0.53	11489
weighted avg	0.53	0.53	0.53	11489

Figure 5.1.2. Performance Metrics for Random Forest with Multiple Features

The AUC: 0.7065998327713183
Report for Random Forest including mulitple features

	precision	recall	f1-score	support
0	0.71	0.70	0.71	5757
1	0.70	0.71	0.71	5732
accuracy			0.71	11489
macro avg	0.71	0.71	0.71	11489
weighted avg	0.71	0.71	0.71	11489

- ***Conclusion***. The exploration of classification models for key prediction from valence revealed that while individual features offer some predictive capability, a multi-feature model notably enhances performance. The Random Forest model emerged as the most effective, with an improved AUC score reflecting its increased accuracy in predicting a song's key modality. Future work could explore further feature engineering and alternative machine learning algorithms to optimize predictive performance.

5.2 Genre Prediction with Neural Networks

8) Can you predict genre by using the 10 song features from question 4 directly or the principal components you extracted in question 6 with a neural network? How well does this work?

- ***Methodology***. We built two distinct models: one using raw song features and another incorporating principal components as inputs to a neural network. We began with a RandomForestClassifier to establish a benchmark using the raw features. Afterward, we used a Sequential neural network model, structured with dense layers, to predict across the 52 genres. The neural network was compiled to optimize for accuracy. To balance the dataset for the neural network model, we utilized the principal components derived from PCA in question 6. We evaluated the models' performance using the Area Under the Curve (AUC) metric for each class.

- ***Findings***. The RandomForestClassifier model, utilizing the direct song features, achieved an AUC of 0.579, suggesting a modest capability to classify genres correctly. In contrast, the neural network model, which used PCA-reduced features, demonstrated a significant increase in discriminative power with an AUC of 0.889. This indicates a substantial enhancement in the model's ability to differentiate between genres when applying dimensionality reduction techniques before classification.

- ***Conclusion.*** The significant improvement in AUC scores when using PCA suggests that the reduction of feature dimensionality allows the neural network to capture the essence of the data more effectively, leading to more accurate genre classification.

6. Recommender Systems and Popularity Analysis

6.1 Star Ratings and “Greatest Hits”

9) In recommender systems, the popularity based model is an important baseline. We have a two part question in this regard:

a) Is there a relationship between popularity and average star rating for the 5k songs we have explicit feedback for? b) Which 10 songs are in the “greatest hits” (out of the 5k songs), on the basis of the popularity based model?

- ***Methodology***. To analyze the relationship between popularity and average star ratings for the 5,000 songs with explicit feedback, we first computed the average star rating for each song. We then assessed the correlation between these average ratings and the songs' popularity scores using Pearson's correlation coefficient. To visually interpret the relationship, a linear regression model was fitted to the data(see *Figure 6.1.1*). Additionally, to establish a baseline for a popularity-based model, we identified the top 10 songs by sorting them according to their average ratings.

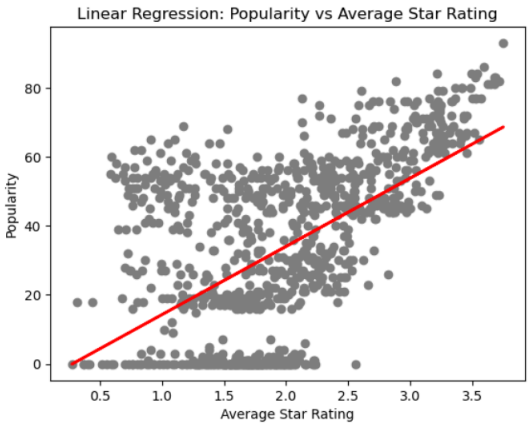


Figure 6.1.1. Linear Relationship between Popularity and Average Star Rating

- ***Findings***. Our analysis yielded a Pearson correlation coefficient of 0.5694, indicating a moderate positive correlation between song popularity and average star ratings. The linear regression's R-squared value was 0.30, suggesting that average star ratings account for approximately 30% of the variation in the songs' popularity. While this demonstrates a notable correlation, it also implies that other factors contribute significantly to a song's popularity. Regarding the "greatest hits," shown in *Figure 6.1.2*, the top 10 songs were determined based on the highest average ratings, reflecting their popularity in the dataset.

	artists	album_name
2260	Red Hot Chili Peppers	By the Way (Deluxe Edition)
2562	The Offspring	Rise And Fall, Rage And Grace
2105	Red Hot Chili Peppers	Californication (Deluxe Edition)
2003	The Neighbourhood	I Love You.
2011	WALK THE MOON	TALKING IS HARD
3253	Gorillaz;Tame Impala;Bootie Brown	New Gold (feat. Tame Impala and Bootie Brown)
3201	Evanescence	Fallen
3007	Linkin Park	Meteora
2009	Nirvana	Nevermind (Remastered)
2770	The Offspring	Americana

Figure 6.1.2. Top 10 Songs in the “Greatest Hits”

- ***Findings***. There is a moderate positive relationship between the popularity and average star ratings. The 10 songs are in the “greatest hits”, on the basis of the popularity based model are Can't Stop ’, ‘You're Gonna Go Far, Kid’, ‘Californication’, ‘Sweater Weather’, ‘Shut Up and Dance’, ‘New Gold (feat. Tame Impala and Bootie Brown)’, ‘Bring Me To Life’, ‘Numb’, ‘Smells Like Teen Spirit’, and ‘The Kids Aren't Alright’.

6.2 Personal Mixtape Recommendations

10) Create a “personal mixtape” for all 10k users we have explicit feedback for. This mixtape contains individualized recommendations as to which 10 songs (out of the 5k) a given user will enjoy most.
a) How do these recommendations compare to the “greatest hits” from the previous question
b) How good is your recommender system in making recommendations?

- ***Methodology***. To craft personalized mixtapes for 10,000 users, we employed collaborative filtering using cosine similarity. Handling missing values by setting them to zero, we constructed a predictive function to estimate a user's rating for songs they hadn't rated, drawing on similar users' preferences. Additionally, we devised a function to extract a user's top 10 song recommendations. Our assessment metric was the hit rate—the proportion of recommendations that overlapped with the top 10 songs from the popularity-based model.

- ***Findings***. An exemplar case using the third user in our dataset demonstrated the effectiveness of the recommendation model. After predicting their ratings for songs they hadn't rated, we identified their top 10 songs. Comparing these to the "greatest hits" from the popularity-based model revealed a 60% hit rate, indicating a substantial overlap and suggesting that our recommender system is capable of reflecting widespread preferences while also providing personalized suggestions.

- ***Conclusion***. The personalized mixtapes show a strong alignment with popular songs, reflecting a balance between individual tastes and overall popularity. The hit rate of 60% confirms the recommender system's proficiency in capturing user preferences, paralleling general trends observed in the broader dataset. This indicates that the system can be a valuable tool for music discovery and personalized user experiences.

7. Extra Credit

7.1 Innovative Analysis Beyond Given Questions

Extra Question: Does the key in which a song is composed affect its popularity?

- ***Methodology***. We classified songs by their musical key and calculated the average popularity within each group. To assess whether the popularity differences across keys were statistically meaningful, we conducted a Kruskal-Wallis test, a non-parametric method suitable for non-normally distributed data or unequal sample sizes. Additionally, we generated visual representations to display popularity distributions across keys, shown *Figure 7.1.1*, aiding in the detection of any notable patterns or exceptions.

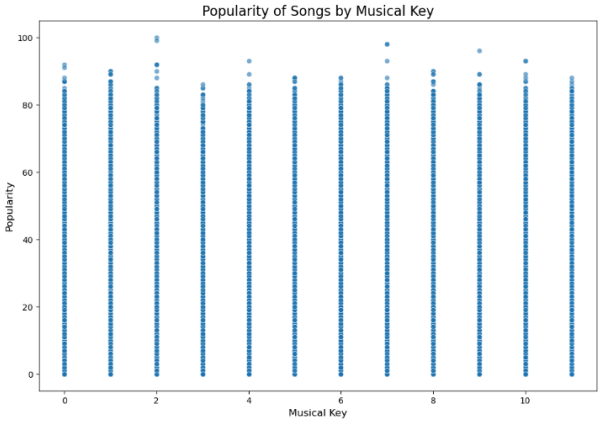


Figure 7.1.1. Popularity Distributions across Keys

- ***Findings***. Shown *Figure 7.1.2*, We can observe that songs composed in Key 4 tend to have the highest average popularity, while those in Key 10 have the lowest. The very low p-value (far below the standard significance level of 0.05) indicates that there are statistically significant differences in the popularity of songs across different keys. This result suggests that the key in which a song is composed might have an effect on its popularity. However, it has only a very minor effect on its popularity. This implies that while the key might contribute to differences in popularity, it is not a major factor, and other elements like genre, artist popularity, and production are likely more influential in determining a song's success.

key	
0	33.195848
1	32.451003
2	33.610869
3	33.312303
4	34.706340
5	33.442082
6	33.644678
7	32.243649
8	33.505003
9	32.447419
10	31.880763
11	33.535870

Figure 7.1.2. Average Popularity Scores by Musical Key

- ***Conclusion.*** The key in which a song is composed does have an influence on its popularity, even though a slight one. The findings imply that other aspects of a song carry more weight in its potential success. While the key might add a nuanced layer to a song's appeal, artists and producers may prioritize other elements in the creative process to maximize popularity.