Spotify Analysis

Student number **S2706063** **S2750388** **S2510135** **S2694903**

Abstract

In this study, we leverage a dataset from Spotify's "Top 200" global playlists, which includes over half a million songs from 2017 to 2023, enhanced with track features and rankings, to predict music trends for 2024. We apply k-means, principal component analysis (PCA) for feature extraction and Gaussian distribution adjustments on the data, then utilize four predictive models: Linear Regression, Random Forest, SARIMA, and LSTM. Our analysis evaluates each model's performance using Mean Squared Error (MSE), Mean Squared Aberration (MSA), and R² to determine their effectiveness in forecasting artist and genre popularity, aiming to guide Spotify's strategic insights into evolving musical preferences.

1 Instructions

In this paper, we utilize four machine learning methods to analyze Spotify's music database from 2017 to 2023, aiming to predict which music genres will dominate in 2024. Predictive models, as widely applied in streaming platforms, have been demonstrated to enhance business strategies and user engagement[1]. As part of an internal team at Spotify, this analysis helps the company gain deeper insights into artists, albums, and user preferences. Initially, we assessed the feasibility of several questions individually, ultimately selecting the prediction of 2024 music trends as our focus.

This endeavor holds substantial importance for the Spotify music streaming platform. By forecasting upcoming music trends, Spotify can strategically enhance its content library by securing collaborations with artists likely to rise in popularity or acquiring copyrights to songs and albums poised for success. Such proactive strategies not only enrich the user experience but also solidify Spotify's competitive edge in the streaming industry. Furthermore, this predictive capability allows for more personalized user experiences and targeted marketing campaigns. Overall, accurate prediction of music trends serves not just to elevate user engagement but also as a crucial tool for data-driven decision-making and strategic market planning at Spotify.

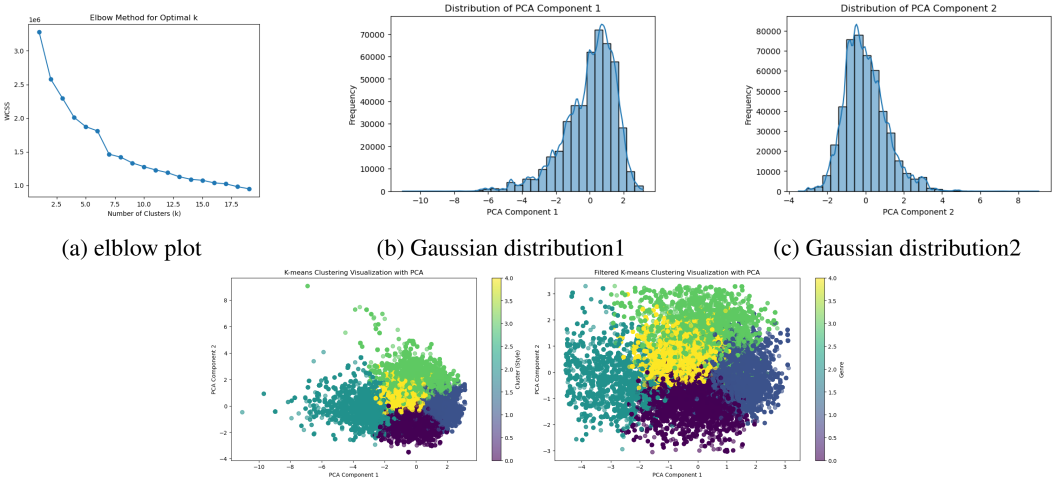
2 Data preparation

The original data set did not show basic problems such as null values after basic test, but there were many features that could not be used in this experiment. Moreover, the music genre of the experiment was not clearly defined in the dataset, and it could not be determined whether there were outliers in the data. We will do data preprocessing in preparation part and prepare for the formal experiment later.

Feature engineering: Our main topic is figuring out what will be the most popular music genre in the future, so we only need to consider the features that can help define music ID information, the features related to music genre, the features that can define the popularity of music, and the time features. So the team dropped the columns *Song URL*, *# of Nationality*, *Nationality* and *Continent* in the original dataset, and we only considered the total point of the music. In order to prevent data bias, we performed a drop_duplicates operation on the dataset. Each music product of specific title, date and id is allowed to appear in the dataset only once. In order to split the training set and the validation set, we use the time information to arrange the data set in the order from early to last.

Clustering: To get the genre of each music, we used k-means to assign similar data points to the nearest cluster[2], it clustered based on 7 characteristics related to music which are *Danceability*, *Energy*, *Loudness*, *Speechiness*, *Acousticness*, *Instrumentalness* and *Valence*. We added a new label *Genre* for each piece of music. First we use the elbow plot(Figure 1a) and Silhouette Coefficient to get the best k value of 5. Based on this we added a new label *Genre* for each piece of music and synced it to our dataset.

Data dimension reduction: In order to understand the distribution of data, we first standardized the selected features, converted the features into the same scale (mean 0, standard deviation 1), and ensured that the weight of each feature's contribution to the principal component was only related to its variance. We then used Principal Component Analysis (PCA) to reduce the dimensions of the data and visualize the distribution of the data(Figure 1d)[3].



(a) elblow plot

(b) Gaussian distribution1

(c) Gaussian distribution2

(d) Original PCA

(e) New PCA

Figure 1: Data preparation

Handling of outliers: From the obtained PCA images, it can be found that there are many outliers in the data, and their Euclidean distance from the center of each cluster are far away, forcing them to be classified into this cluster will adversely affect the effect of models. We check that the distribution of the two principal components is consistent with the characteristics of Gaussian distribution(Figure 1b & 1c), so we use Gaussian distribution to remove data points outside the range out of 3 standard deviations(Figure 1e), and the change is synchronized to the original high-dimensional data set.

Now the team has obtained the new genres and calculated the average of the 7 music-related features corresponding to each genre. By combining the visualization of these values in Table 2 and the music genre records from official website of Spotify [4], we can give specific information to each genre generated. The 5 genres present in the dataset can be summarized as: Rock or Alternative, Dance or Electronic, Jazz or Classical, Hip-pop or Rap, Pop or R&B.

3 Exploratory data analysis

The initial dataset contained 651,936 rows and 20 attributes. After we used Gaussian distribution to filter outliers and delete duplicates caused by multiple creators in data preprocessing, we ended up with a new dataset of 459,485 rows, adding an extra attribute called Genre to classify each song. We compared the average features of five music genres (Dance or Electronic, Hip-hop or Rap, Jazz or Classical, Pop or R&B, Rock or Alternative) using radar charts. These charts show clear differences in features like loudness, energy, and danceability among the genres. For instance, Electronic or Dance music tends towards high energy and loudness, while Jazz or Classical music focus more on instrumental performance and sound quality. These comparisons highlight the core features of each genre and their distinct emphases, also shows that our genre classification has obvious differentiation, Figure 2.

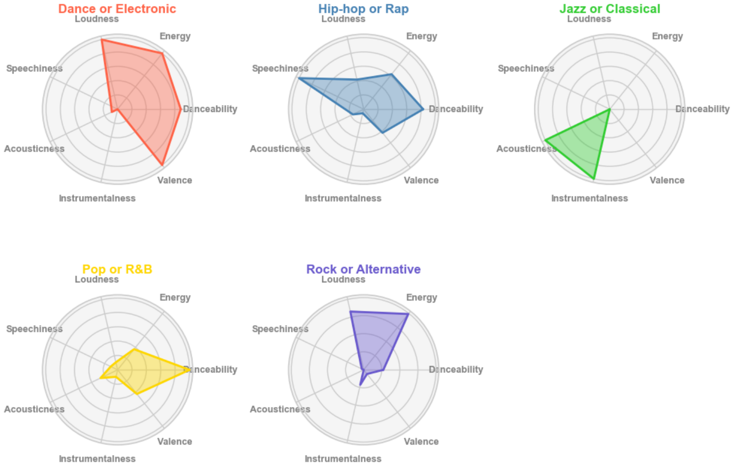


Figure 2: Music Genre Features Comparison

To examine the popularity of different genres from 2017 to May 2023, we calculated the total yearly scores for each genre and divided by the total scores of all songs each year to get annual ratios. During this period, the popularity ratios of the five genres changed somewhat. Generally, Dance or Electronic and Rock or Alternative held larger shares of popularity. Pop or R&B showed an increasing then decreasing trend. Hip-hop or Rap and Jazz or Classical were slightly less popular, with small changes and a stable trend overall, as shown in Table 1. For further experiments, we used the first 80% of data from 2017 to May 2023 as the training set and the last 20% as the validation set.

Table 1: Popularity Proportions of Music Genres from 2017 to 2023

Year	Dance or Electronic	Hip-hop or Rap	Jazz or Classical	Pop or R&B	Rock or Alternative
2017	0.371	0.130	0.050	0.200	0.249
2018	0.318	0.151	0.071	0.209	0.250
2019	0.328	0.175	0.090	0.236	0.171
2020	0.336	0.169	0.111	0.217	0.167
2021	0.367	0.104	0.139	0.207	0.183
2022	0.341	0.113	0.119	0.159	0.268
2023	0.320	0.102	0.134	0.156	0.288

4 Learning methods

To effectively utilize the processed data for predicting the artists or genres that will be popular in 2024, this study will employ four models — Linear Regression, Random Forest, SARIMA, and LSTM. Each model is selected for its ability to address specific facets of the prediction task, contributing to a comprehensive and robust analysis. By focusing on interpretability and accuracy, these models will provide diverse perspectives to enhance the reliability of the forecasts.

4.1 Seasonal Autoregressive Integrated Moving Average

Observing historical data, we noted periodic changes in the popularity scores across genres, so we decided to predict future trends using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model [5]. SARIMA, which adds four seasonal parameters to the ARIMA model (three hyperparameters P, D, Q, and a seasonal period parameter s), can handle time series with seasonal characteristics[6]. We targeted the "Jazz or Classical" genre's popularity trend as our time series, grouped data by "year-month," and calculated the genre's total monthly popularity scores. We set the maximum lag order and seasonal period to 12 to match the 12-month seasonal patterns in the data. We also employed a rolling forecast method to assess SARIMA's performance, where the model predicts one time point in the validation set at a time. After each prediction, that value is added to the training set history, and the model parameters are recalibrated for the next prediction, as shown in Figure3. This method simulates real-world scenarios of updating predictions with new data. To clearly show the results, we created comparison charts of training data, validation data, and rolling forecast outcomes.

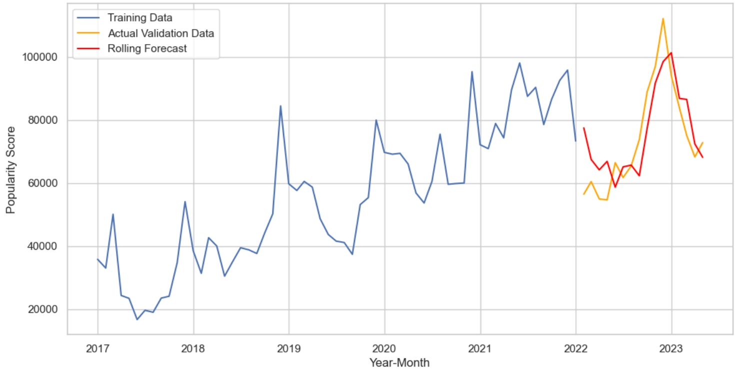


Figure 3: Rolling Forecast vs Actual using SARIMA for Jazz or Classical Music

4.2 Linear Regression

Linear regression models the linear relationship between a dependent variable and one or more independent variables by minimizing the loss function, common techniques discussed in related works [7], which emphasizes the application of regression analysis in modeling relationships. In our study, after the preprocessing phase, we utilize the genre classification of each track as the independent variable, while the associated scores serve as the dependent variables for constructing a linear regression model.

To establish our training and test datasets, we allocate 80% of the chronologically sorted data for training purposes, reserving the remaining 20% for testing, consistent with standard practices in time-series analysis [8], which emphasizes preserving temporal order in datasets. This model aims to predict the trends for the year 2024 based on data from the initial seven-year span. Given the time-series nature of our data, we maintain the sorted sequence of the dataset to preserve temporal relationships. Additionally, we employ the Mean Squared Error (MSE) as our loss function to evaluate the accuracy of the model, ensuring it reflects the deviation between the predicted values and actual data effectively.

4.3 Random Forest

The core idea of a Random Forest is to build multiple decision trees and then get the final predicted value by integrating the predictions of all trees. The reason why Random Forest is chosen here is that it is suitable for high-dimensional data, supports processing of a large number of features, and can evaluate the importance of features. There are 7 different features to describe the genre of music in this experiment, which is very suitable occasion for using Random Forest model. At the same time, the model can reduce the risk of overfitting comparing to train data of a single tree by integrating multiple decision trees. Besides, the operation of adjusting the generalization ability of model is also intuitive, which can be done by adjusting hyper-parameters such as max_depth.

The input of this model are the features related to the music genre, the genre of each song and the time information, the target variable is the score of each musical product, and the resulting output is the prediction result of the most popular music genre in the future. We divided the data set into a training set (80%) and a validation set (20%) in chronological order to evaluate the model performance, and then used the overall data to make final predictions. The loss functions used to judge the performance of the model are mean Squared error (MSE), mean absolute Error (MAE) and r-squared (R^2). We hope to measure the sensitivity of the model to high deviations, understand the average deviation from the predictions and measure the model's ability to explain changes in genre popularity. At the same time, in order to verify the stability and fitting degree of the model, we applied 5-Fold cross-validation.

4.4 LSTM

LSTM (Long Short-Term Memory) networks are a specialized type of recurrent neural network (RNN) designed to handle long-term dependencies in sequential data. In this project, the music data exhibits clear seasonality and trends, where historical patterns significantly influence future predictions. LSTM's gated mechanisms (input gate, forget gate, and output gate) allow it to retain important historical information while filtering out irrelevant noise, making it an ideal choice for modeling complex time series data like music popularity trends.

5 Results

In the result part, we will first evaluate the performance of the four models on training set and validation set by using mean Squared error (MSE), mean absolute Error (MAE) and r-squared (R²), then give the final prediction of each model about which music genre will become the most popular genre in 2024. Besides, considering the fact that SARIMA actually use a totally different logic for prediction comparing to Linear Regression, Random Forest and LSTM, we will demonstrate our result in two parts.

5.1 Seasonal Autoregressive Integrated Moving Average

SARIMA's predictions for the future popularity of the five music genres showed that the mean squared error (MSE) and mean absolute error (MAE) were slightly higher on the validation set than on the training set, indicating the model learned more from the training data but could improve in generalizing to new data. The training set's R² was 0.611, showing good explanatory power, while the validation set's R² was 0.084, indicating limited explanatory power for the validation data, likely because SARIMA is better at predicting short-term changes than long-term changes, as shown in Figure 4.

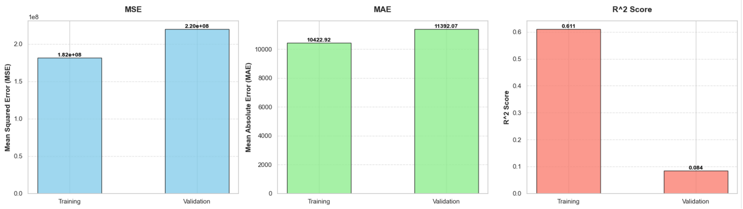


Figure 4: MSE, MAE and R² of SARIMA

Using data from January 2017 to May 2023, we predicted the popularity of different genres after June 2023. Overall, Dance or Electronic and Rock or Alternative were the most popular, with Dance or Electronic peaking in 2019 and then gradually declining, while Rock or Alternative showed an upward trend during the forecast period. Hip-hop or Rap and Pop or R&B showed slight declines, while Jazz or Classical maintained a low level of popularity but showed a rising trend during the forecast period, as shown in Figure 5

5.2 Linear Regression, Random Forest and LSTM

To enhance the assessment of our model's effectiveness, we will conduct a comparative analysis of linear regression, random forest, and LSTM methodologies.
MSE value is shown in Figure 6a, the lowest training MSE appears in Random Forest model(2842.8), which is slightly lower than the value of LSTM(3150.7), and Linear Regression gets the highest training MSE(3331.4). For validation set, Random Forest still gets the lowest MSE with 3276.4, followed by Linear Regression(3325.7), and LSTM has the highest MSE(3604.4).
MAE value is shown in Figure 6b, the highest training MAE appears in Linear Regression (50.0), while the lowest appears in Random Forest (45.6). The best result of validation value appears in

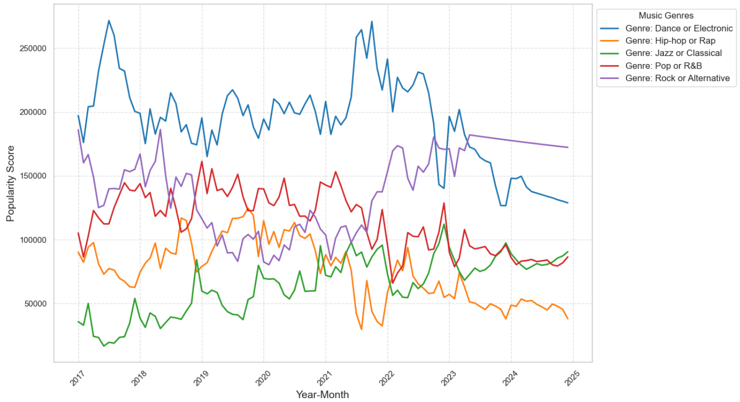
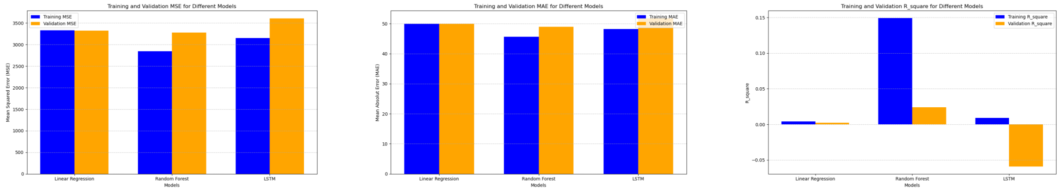


Figure 5: Popularity Trend for Each Genre with Forecast from June 2023 to December 2024

LSTM (51.7), while the value of Random Forest (49) are the lowest.
R2 value is shown in Figure 6c, Random Forest has both the highest training value(.0041) and the validation value(.0022). The lowest training value appears in Linear Regression, lower than that of LSTM(.0092). It is notable that the validation value of LSTM is a negtive number(.059), which is the lowest.

It can be found that Random Forest performs the best, and both Random Forest and LSTM have poor generalization ability, and all those three models have weak explanatory power, the performance of LSTM is even worse than that of direct mean prediction. As for the final prediction, the results of Linear Regression, Random Forest and LSTM are *Dance or Electronic*, *Dance or Electronic* and *Hip-pop or Rap* respectively.



(a) Mean squared error

(b) Mean absolute error

(c) R²

Figure 6: Result of Linear Regression, Random Forest And LSTM

6 Conclusions

In our exploration, the application of four diverse machine learning techniques provided a nuanced understanding of potential shifts in music genre popularity for the year 2024, as deduced from Spotify's extensive dataset covering the period from 2017 to 2023. Our methodology includes fundamental pre-processing techniques, such as the removal of irrelevant data, PCA and K-means for feature extraction, and Gaussian denoising, to enhance the quality and applicability of the data used in the modeling process. The predictive outcomes revealed varied trends across genres, with Linear Regression and Random Forest pointing to a rising trend in Dance or Electronic genres. In contrast, SARIMA predicted a sustained interest in Rock or Alternative music, and LSTM primarily identified Hip-hop or Rap as predominant.

These insights are invaluable for Spotify's strategy in content curation and market positioning. By aligning their portfolio with these predictive insights, Spotify can optimize its engagement strategies and remain competitive in the ever-evolving music industry. The study's outcomes serve as a testament to the utility of machine learning in forecasting market trends, thereby empowering Spotify to tailor its offerings to meet future consumer demands effectively. As the digital music landscape continues to transform, such predictive capabilities will be crucial for navigating and capitalizing on the changing tastes of music listeners worldwide.

References

[1] Rutger Ruizendaal. The predictive power of social media: Using twitter to predict spotify streams for newly released music albums. Master's thesis, University of Twente, 2016.

[2] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[3] Takio Kurita. Principal component analysis (pca). *Computer vision: a reference guide*, pages 1–4, 2019.

[4] Spotify. Getting started with web api, 2024. Accessed: 2024-11-20.

[5] A. C. Harvey. *ARIMA Models*, pages 22–24. Palgrave Macmillan UK, London, 1990.

[6] S.L. Ho and M. Xie. The use of arima models for reliability forecasting and analysis. *Computers Industrial Engineering*, 35(1):213–216, 1998.

[7] David A Freedman. Statistical models: Theory and practice. *Cambridge University Press*, 2009.

[8] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 2015.

Contribution

S2706063(25%): Linear Regression, report writing, K-means.

S2750388(25%): Dealing with data standardization, PCA and implement Random Forest model.

S2510135(25%): Mainly responsible for completing Gaussian distribution screening of abnormal data, Exploratory data analysis and SARIMA model training and prediction.

S2694903(25%): In charge of LSTM model, Spotify API.

Appendix A: Characteristics of clustered music genres

Table 2: Genre Characteristics Table

Genre	Danceability	Energy	Loudness	Speechiness	Acousticness	Instrumentalness	Valence
Rock or Alternative	0.582097	0.684415	-4991.073024	0.063593	0.117468	0.007194	0.343734
Dance or Electronic	0.747473	0.754278	-3688.942739	0.082366	0.165495	0.001889	0.704452
Jazz or Classical	0.528063	0.419775	-7546.066938	0.057832	0.646948	0.033013	0.321288
Hip-pop or Rap	0.734111	0.628750	-5910.092886	0.305196	0.206158	0.003719	0.482767
Pop or R&B	0.776477	0.542914	-7129.651286	0.084010	0.257264	0.005006	0.486125

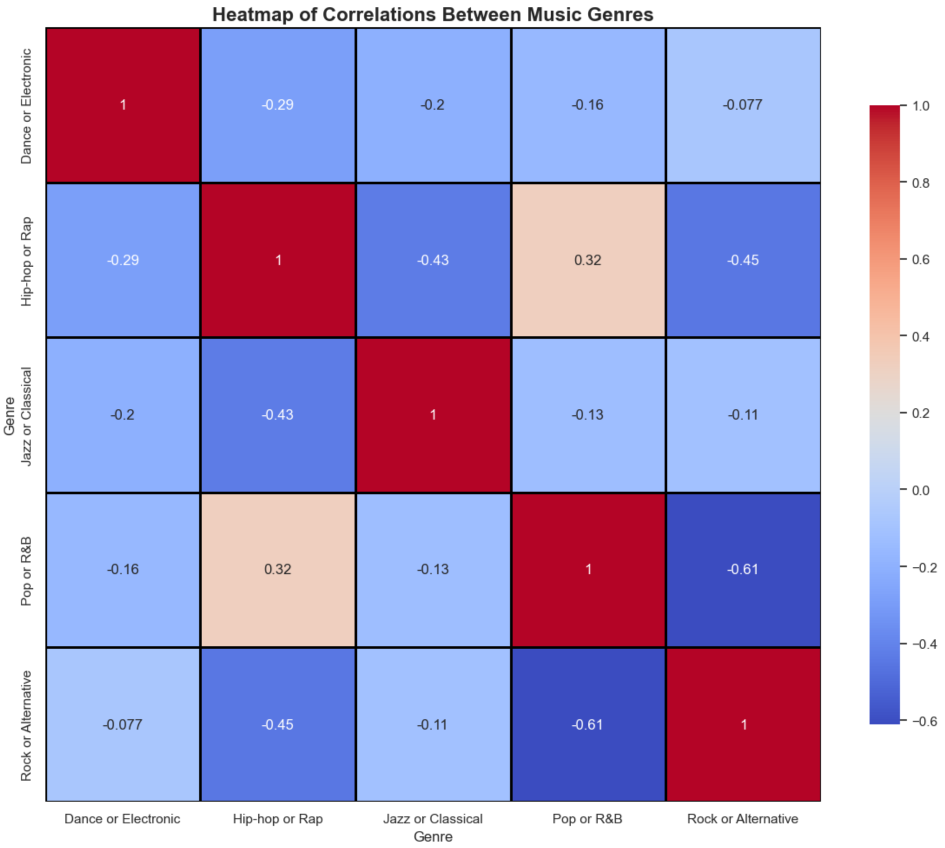


Figure 7: Correlations Between Music Genres

Appendix B: Music Genre Popularity Over the Years

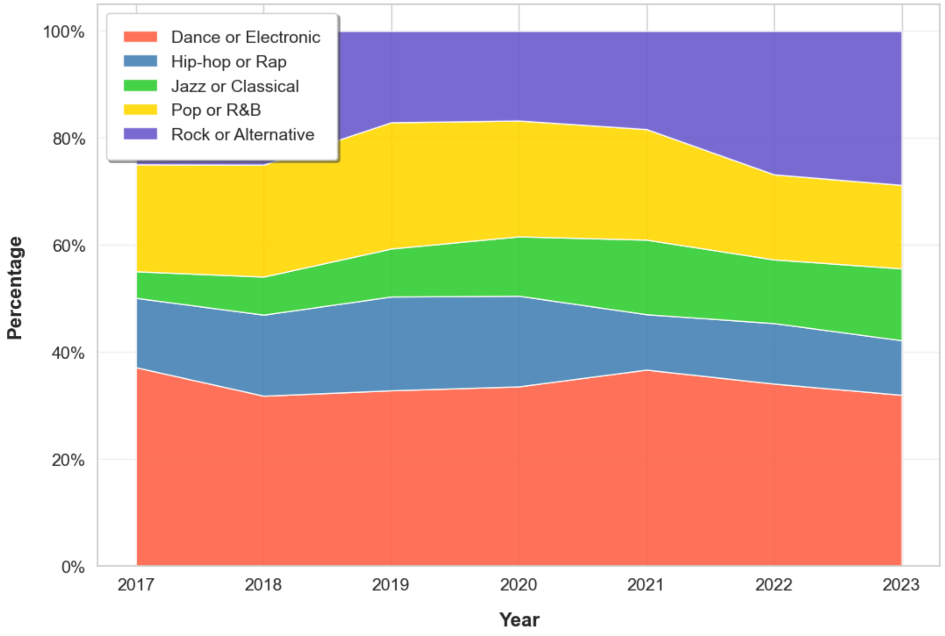


Figure 8: Music Genre Popularity Over the Years

Appendix C: Training lines related to SARIMA

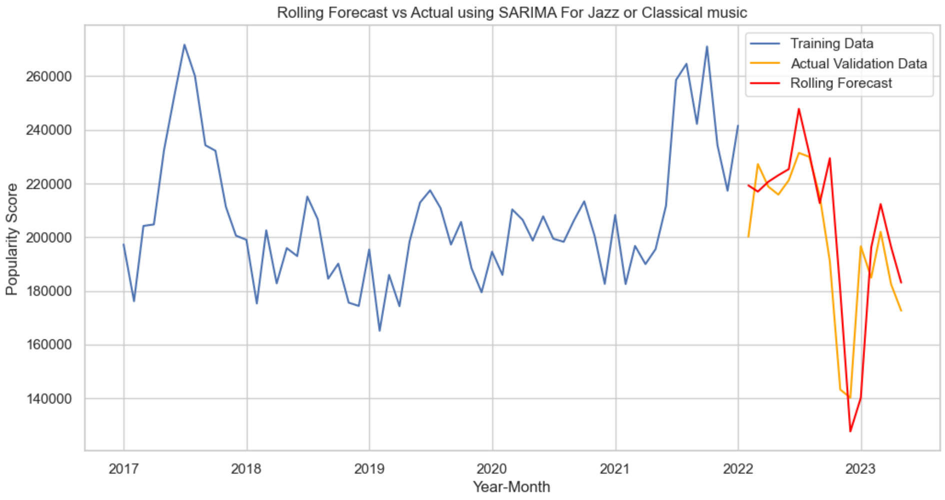


Figure 9: Rolling Forecast vs Actual using SARIMA for Dance or Electronic Music

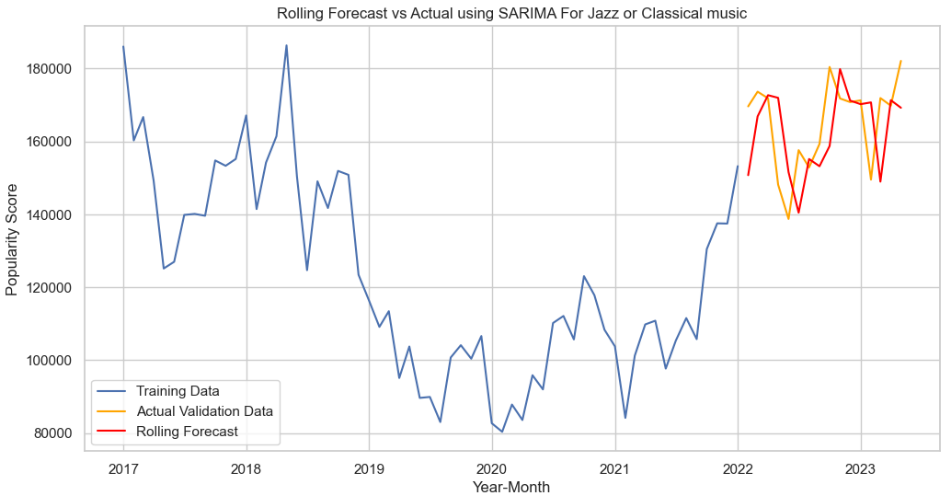


Figure 10: Rolling Forecast vs Actual using SARIMA for Rock or Alternative Music

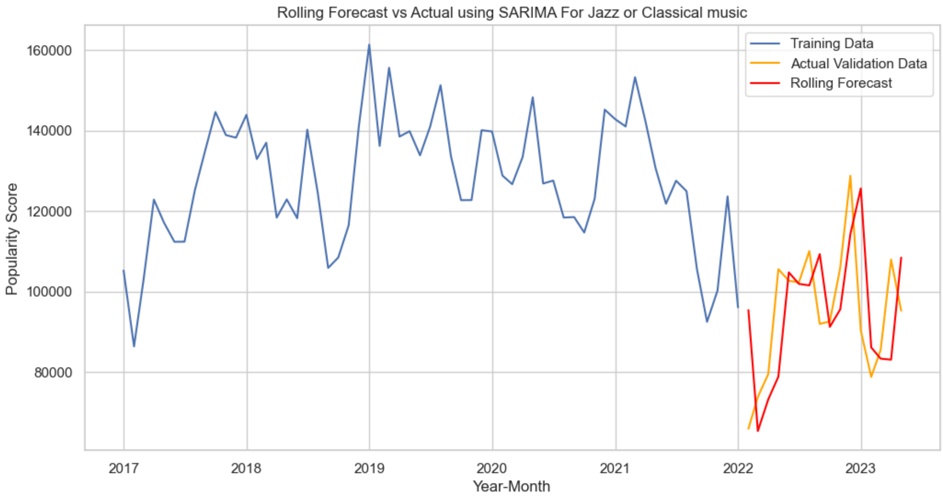


Figure 11: Rolling Forecast vs Actual using SARIMA for Pop or RB Music

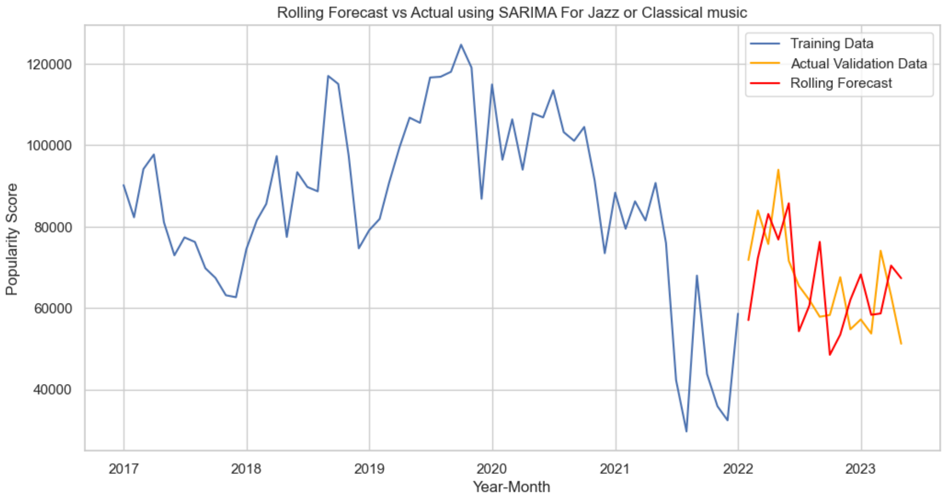


Figure 12: Rolling Forecast vs Actual using SARIMA for Hip-hop or Rap Music