Title Using audio-based features and features derived from cover images to predict Spotify hits

Bruno Nadalic Sotic – 11353473

Big Data & Automated Analysis

Final Assignment

1. Introduction

The digital music landscape has evolved significantly over the past decade, with streaming platforms like Spotify becoming the primary medium for music consumption (Bello, 2021). In this highly competitive market, understanding and predicting song popularity has become a vital aspect for musicians, record labels, and streaming platforms alike. Previous research in musicology has explored various factors that contribute to a song's success, such as genre (Fraser et al., 2021), melodic qualities (Swarbrick et al., 2019), and audio features (Dawson et al., 2021). Additionally, studies have shown that visual elements, such as album cover images, can also impact audience engagement (Libeks & Turnbull, 2011). This research aims to develop a comprehensive understanding of the factors that predict song popularity on Spotify, focusing on both audio and visual features.

The study is divided into two main parts. The first part investigates the question*: To what extent can audio features, as defined by Spotify's API, be used for predicting song popularity? What is the relative importance of each audio feature in predicting song popularity, and how do they compare to each other?* This is primarily a replication of the research conducted by Dawson et al. (2021), who demonstrated that a model built using latent audio features could predict popularity with 81% accuracy. The second part of the study explores the potential of using features derived from cover images to classify songs based on their popularity, drawing from previous research in media audience engagement (Lau & Ajudha, 2021). Hence the question: *To what extent can features derived from a songs cover image be used to classify songs into popularity groups on Spotify?*

The integration of both audio and visual features in this study provides a comprehensive understanding of the factors that contribute to song popularity on Spotify. This research is methodologically relevant but also has practical implications for musicians, record labels, and streaming platforms, as it can inform marketing strategies and content creation processes.

By using causal machine learning techniques to model the relationship between audio features, visual features, and song popularity, this study aims to provide a rigorous and data-driven approach to understanding the factors that contribute to a song's success.

2. Theoretical Framework

2.1. Audio Features and Popularity

The concept of genre has long been a cornerstone of musicological research, with certain musical characteristics believed to be inherently appealing to audiences. For example, rock music is typically characterized by its energy and tempo, while disco music is associated with valence and danceability. These genre-specific audio features can be quantified and analyzed to better understand the factors that contribute to a song's success (Interiano et al., 2018).

Spotify has made these audio features available through their API, which provides metrics such as valence, tempo, and speechiness, acousticness, melody and many more. Dawson et al. (2021) used these metrics to build an xGBoost model capable of accurately predicting song popularity. By replicating this research, this study seeks to validate the findings and explore the impact of audio features on song popularity.

However, instead of treating popularity prediction as a binary classification task, as done by Dawson et al. (2021), this study will aim to predict the actual popularity scores provided by Spotify's API. This approach allows for greater precision in understanding the relationship between audio features and popularity (Reddy et al., 2021).

2.2. Visual Features and Popularity

In addition to audio features, visual elements have been shown to play a significant role in audience engagement across various forms of media, including news articles (Tang et al., 2011), YouTube videos (Koller & Grabner, 2022), and music albums (Venkatesan et al., 2023). This study seeks to build upon this research by investigating the influence of visual features derived from album cover images on song popularity.

By extracting and analyzing features from album cover images, such as color, composition, and subject matter, this study aims to determine whether these visual elements can provide additional explanatory power in predicting song popularity on Spotify.

2.4. Variables & Features

Popularity on Spotify and audio features are defined with regard to Dawson et als. (2021) work as well as the Spotify API documentation. All variables are directly retrievable via the API.

- **Popularity Score**: Popularity score for each track ranging from 0 to 100. Indicates the relative popularity of a track among Spotify users and is determined by factors such as play count, playlist additions, and user engagement.
- **Danceability**: Represents the suitability of a track for dancing. Range: 0 to 1, where 0 is low danceability and 1 is high danceability.
- **Energy**: Measures the intensity and activity level of a track. Range: 0 to 1, where 0 is low energy and 1 is high energy.
- **Loudness**: Measures the overall volume of a track in decibels (dB). Range: -60 dB to 0 dB.
- **Speechiness**: Detects the presence of spoken words in a track. Range: 0 to 1, where 0 is non-speech-like and 1 is primarily spoken words.
- **Acousticness**: Measures the likelihood of a track being acoustic. Range: 0 to 1, where 0 is likely electronic/synthesized and 1 is likely acoustic.
- **Instrumentalness**: Predicts the likelihood of a track being instrumental. Range: 0 to 1, where 0 is likely to contain vocals and 1 is likely instrumental without vocals.
- **Liveness**: Measures the presence of a live audience in a track. Range: 0 to 1, where 0 is likely a studio recording and 1 is likely a live performance.
- **Valence**: Represents the musical positiveness conveyed by a track. Range: 0 to 1, where 0 is negative or sad mood and 1 is positive or happy mood.
- **Tempo**: Measures the beats per minute (BPM) of a track. Range: 0 to infinity, typical values range from 50 to 250 BPM.
- **Duration (ms)**: Measures the length of a track in milliseconds. Range: 0 to infinity, depending on the length of the track.

Popularity score is the main dependant variable, whereas the audio features are independent predictors. One downside to this is that the variables are defined and computed by Spotify, and the exact mechanisms of how that was done is unclear.

3. Part 1: Using Audio Features to Predict Popularity

3.1. Analytic Strategy

Part 1 of this study aims to determine which audio features can be used to predict the popularity of a song. To achieve this, the task is treated as a regression problem. This approach is justified by the fact that popularity (main dependant variable), as well as the audio features, are measured on a numeric scale, allowing for more precise predictions. Regression models, which are

suitable for predicting continuous numerical values, will be employed to estimate the popularity score based on the song's audio features.

In accordance with the recommendations of Reddy et al. (2021), popularity will not be transformed into categorical variables, nor will any other variables undergo transformation. By maintaining the variables in their original numeric scale, the study aims to produce more accurate and precise findings regarding the phenomenon.

The selection of appropriate models for this task and variables will be informed by the outcomes of an exploratory data analysis (EDA). At present, the relationships and directionality between the dependent variable (popularity) and the independent variables (audio features) are unclear. Therefore, the distribution and skewness of the variables will be examined initially. Outliers and skewed data will not be removed, as they may represent genuine extreme values that are representative of the phenomenon rather than measurement errors.

Moreover, to address potential issues of multicollinearity, correlations and variance inflation factor (VIF) scores will be calculated. The modelling section provides a more detailed explanation of these procedures.

3.2. Data Collection

To have a representative sample, this study retrieves data on popular and unpopular music, along with their corresponding audio features, from the Spotify API. The Spotify API was chosen due to its rich resources and comprehensive music collection. The API provides access to a diverse range of information and features related to songs and their popularity.

Predicting music popularity based on audio features requires a substantial corpus and anti-corpus, consisting of popular and unpopular music sets respectively. Spotify conveniently assigns a popularity index to all the songs it hosts, which serves as a measure of their relative popularity.

The data collection process is twofold. Firstly, an API call was made to retrieve the currently most popular playlists, as they contain the currently most popular songs. Subsequently, another API call was made to obtain playlists that contain unpopular songs. These steps yielded a total of 10 playlist IDs, with 5 IDs representing popular playlists and 5 IDs representing unpopular playlists.

A custom scraper was developed to iterate over each song within the playlists using these playlist IDs. This allowed for the retrieval of song data and their corresponding audio features. Consequently, a final dataset comprising 1956 observations, including songs and their associated data and audio features, was obtained.

It is important to acknowledge the limitations of this dataset, as its findings may not be entirely generalizable. However, these limitations are inherent to the restrictions imposed by Spotify's

API, which allows for the gathering of only 100 observations per playlist. Furthermore, given the need for images in the subsequent part of the research, a larger dataset would exceed the computational resources available.

Subsequently, the dataset was carefully examined to identify and address missing values and duplicates. Any instances of missing values and duplicate entries were removed, resulting in a refined dataset consisting of a total of 1492 observations.

It is essential to reflect on potential biases and privacy concerns introduced through this data collection method. One important consideration is the reliance on the recommendation algorithm employed by Spotify's API, which could influence the songs retrieved during the data collection process.

3.3. Modelling

3.3.1. EDA

The dataset was examined through an exploratory analysis to understand its characteristics and the relationships between variables.

The popularity scores are almost evenly evenly split (with a mean = 54 out of 100), indicating a balanced distribution between popular and unpopular songs. On the other hand, duration, loudness, speechiness, acousticness, instrumentalness, and liveness exhibit potential skewness in their distributions, whereas the rest of the dataset is normally distributed.

The correlations between audio features and popularity are generally weak to none. Some variables show very weak positive or negative correlations, while others have moderate negative associations. Overall, the relationships are not strongly linear, meaning that the relationship between popularity and audio features might be non-linear or more complex.

VIF scores were calculated to detect multicollinearity. All variables have VIF scores below 5, indicating no significant multicollinearity issues. Each variable provides unique information in predicting popularity. The full details and visualizations of the EDA can be found in the Appendix section.

3.3.2. Model Selection & Procedure

The EDA helps in understanding the relationships between the independent variables and the dependent variable (popularity). Based on this understanding, models that are appropriate for capturing these relationships can be selected.

A basic linear regression model is chosen as a baseline model for comparison with other models. Linear regression assumes a linear relationship between the independent variables and the

dependent variable. It is a straightforward and interpretable model, providing insights into the individual coefficients and their impact on the outcome.

Random Forest and XGBoost are selected as more advanced ensemble methods. These models are capable of capturing non-linear relationships and interactions between variables. They are robust to outliers and can handle skewed distributions. Random Forest creates an ensemble of decision trees, while XGBoost is an optimized gradient boosting algorithm. These models have shown good performance in various regression tasks and are suitable for handling complex relationships in the data (Jonas et al., 2017).

The models are trained on 80% of the data and tested on the remaining 20%. This split allows for evaluating the performance of the models on unseen data. The goal is to train the models on a representative portion of the data and assess their performance on data thet was not encountered during training.

Grid Search Cross Validation was used to optimize the hyperparameters of the XGBoost and Random Forest models. Hyperparameters are the settings that control the behavior of the models, and finding the best combination of hyperparameters can improve their performance (Jonas et al., 2017). Grid Search Cross Validation systematically tests different combinations of hyperparameters and selects the one that yields the best performance.

The performance of the models is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared values. MAE and RMSE measure the average prediction errors, while R-squared indicates the proportion of variance in the dependent variable explained by the model.

After training the models, a feature importance analysis is conducted to understand which features have the most impact on predicting song popularity. This analysis helps identify the key audio features that strongly influence the popularity of a song.

4. Results

The outputs of the three models are seen in table 1.

	Baseline OLS	Random Forrest Regressor	xGBoost
RMSE	19.87	19.23	19.70
MAE	15.59	15.18	15.39
R2	0.09	0.15	0.11

Table 1: Outcomes for predicting Song Popularity based on Audio Features

The baseline linear regression model has an RMSE of 19.88 and an MAE of 15.59. The R-squared value, which indicates the proportion of variance explained by the model, is 0.10. These results suggest that the baseline model has limited predictive power for song popularity based on the given audio features.

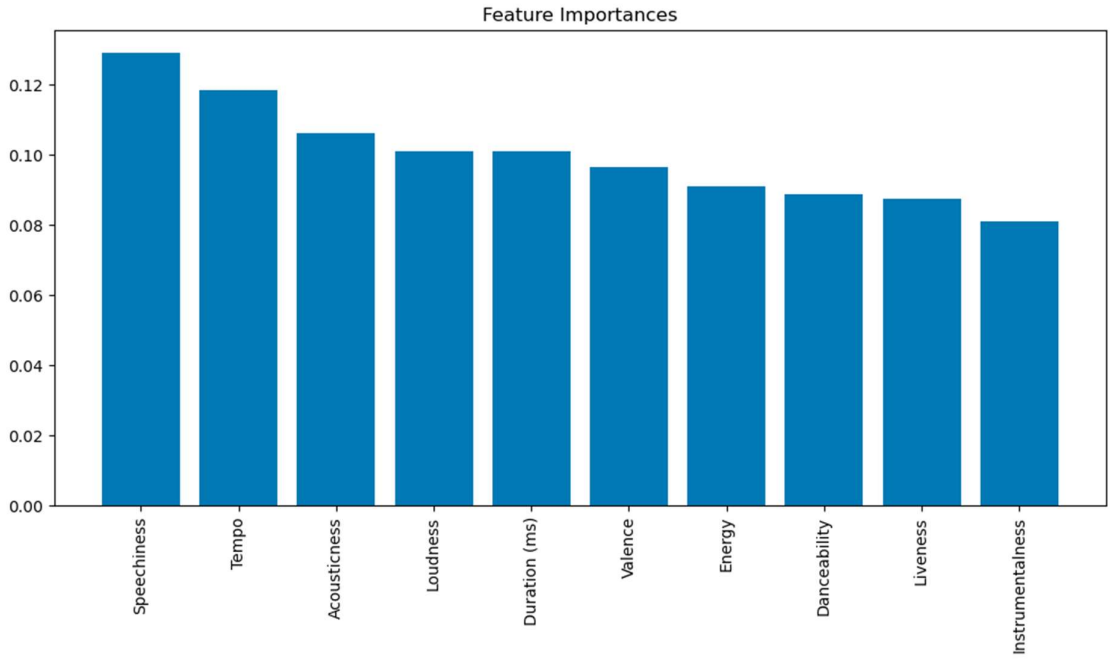
The optimized Random Forest model (*best Parameters: 'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 200*) has a lower RMSE of 19.24 and a lower MAE of 15.19 compared to the baseline model. The R-squared value improves to 0.16, indicating that the Random Forest model explains a slightly larger proportion of the variance in song popularity. This suggests that the Random Forest model captures more complex relationships and interactions among the audio features, leading to better predictive performance.

The optimized XGBoost (*best Parameters: 'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 300*) model has an RMSE of 19.70 and an MAE of 15.40. The R-squared value is 0.12, which is slightly higher than the baseline model but lower than the Random Forest model. The XGBoost model performs better than the baseline model but slightly worse than the Random Forest model in terms of predictive accuracy.

Both the Random Forest and XGBoost models outperform the baseline linear regression model in predicting song popularity. The Random Forest model shows the best performance among the three models, with lower RMSE and MAE values and a higher R-squared value. This indicates that the Random Forest model captures more of the patterns and relationships between audio features and song popularity. However, none of the models has reached a desirable or high accuracy level in predicting popularity scores. This could be due to several factors and is reflected on in section 6. Limitations and Future Research.

4.1. Feature Importance

‘Feature_importances’ form the sklearn.ensemble python library have been used to understand which variables are the biggest contributors for the results for the Random Forrest Model. The bar graph represents the features on the x-axis and their corresponding feature importance scores on the y-axis. The feature importance scores indicate the relative importance of each feature in predicting the popularity of a song.



According to the graph, speechiness (the level of vocal content), tempo (the speed or rhythm), and acousticness (the degree of acoustic elements) are the most influential features in determining popularity, as they have the highest importance scores, with values above 0.10.

Loudness and duration have similar importance scores, followed by valence (the musical positiveness), danceability, and energy, which show a slight decline in importance.

The features that contribute the least to predicting popularity are liveness and instrumentalness. However, it is worth noting that even these features have importance scores greater than 0.08. Overall, while some features have a higher importance than others, the differences between the importance scores are not large.

5. Part 2: Using Features from Cover Images for Popularity Classification

5.1. Approach

The researchs second part explores the potential of features extracted from Spotify cover images in predicting and classifying song popularity. The aim is to examine the visual characteristics of cover images and their relationship with song popularity on the Spotify platform. By analysing said features, this study aims to understand the discriminative power of these features in distinguishing between popular and unpopular songs. The methodology includes scraping and pre-processing the cover images; extracting features using a VGG16 CNN mode; visualizing the feature space; and performing binary classification.

Since this is a classification task, the popularity variable is operationalized into two distinct groups: popular (1) and unpopular (0). The division between these two groups is determined based on the mean value of the popularity variable, which ranges from 0 to 100 as provided by Spotify's API.

5.2. Data Collection

An additional scraper was developed to enhance the dataset. This scraper operates by iterating through the song IDs from the previous dataset to gather the cover images of each song from the API. These images are then saved in a local folder and appropriately labelled as popular or unpopular based on the corresponding popularity score.

5.3. Image Preprocessing & Feature Extraction

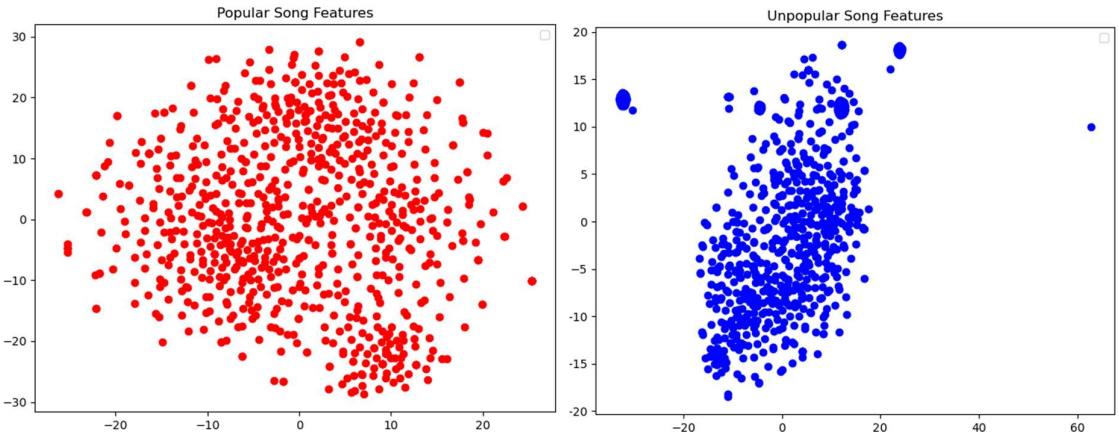
The VGG16 model - a deep convolutional neural network pretrained on the ImageNet dataset - is used to extract features from the cover images. The VGG16 model has shown reliable performance in image classification tasks, and was validated in countless studies (Chen et al., 2015).

To ensure compatibility between the data and the model, cover images were pre-processed by resizing them to 224:224 pixels, as recommended by the VGG16 documentation (Tenserflow.org, n.d.). Additionally, images were converted into NumPy arrays and normalized using the pre-processing function provided by the VGG16 library. Finally, the global average pooling layer was used to aggregate the spatial information of the extracted features into a fixed-length vector representation (Agato et al., 2017).

To gain insights into the distribution and separability of the extracted features, t-SNE (t-Distributed Stochastic Neighbor Embedding) dimensionality reduction was applied. This technique projects high-dimensional feature vectors into a two-dimensional space, allowing intuitive visualization (Agato et al., 2017). The resulting feature space was visualized using scatter plots, with popular and unpopular songs represented by different colors.

5.4. Feature Visualization and Analysis

The scatter plots displays the features of popular and unpopular songs in a two-dimensional space. The x-axis and y-axis represent the two dimensions obtained through the t-SNE algorithm. Each point in the plot represents a song, and its position in the plot corresponds to its feature representation.



The scatter plots reveal that the features of popular songs are scattered throughout the plot, occupying almost every angle. This suggests that popular songs image features exhibit a wide range of values and do not cluster tightly together. On the other hand, the features of unpopular songs are more concentrated and tend to cluster on the left-mid side of the plot. The presence of bigger clusters indicates that there are distinct subsets that share similar features and patterns. It shows that popular songs have a broader range of visual features, while unpopular songs tend to have more consistent patterns in their visual features.

5.5. Classification Model

To assess the predictive power of cover image features, a binary classification task was conducted. The feature vectors obtained from the cover images were divided into two categories based on their popularity labels: popular and unpopular. The classification task's objective was to train a model that could effectively differentiate between these two classes. For this purpose, a Support Vector Machine (SVM) model was used.

SVM is a machine learning algorithm known for its ability to handle high dimensional vectors and its effectiveness in binary classifications. It aims to find an optimal hyperplane that maximally separates data points of different classes in the feature space (Atteveldt et al., 2022).

Before training the SVM model, the features extracted from the images were normalized to a range of 0 to 1, as SVM requires input features to be within this range. The model was then trained on 80% of the data and evaluated on the remaining 20% using an accuracy score and a classification report.

The accuracy score provides an overall measure of the model's performance in correctly classifying popular and unpopular songs. The classification report offers more detailed insights, including precision, recall, and F1-score, for both the popular and unpopular classes.

5.6. Results

The SVM model for predicting song popularity based on image features achieved an accuracy of 0.5886, indicating that the model correctly classified about 58.86% of the songs in the dataset. The full report is seen in table 2.

Accuracy	0.58

	Precision	Recall	F1-score	support
0	0.61	0.42	0.50	145
1	0.58	0.75	0.65	154
Accuracy			0.59	299
Macro avg	0.59	0.58	0.57	299
Weighted avg	0.59	0.59	0.58	299

Table 2: SVM Results

Precision (the ratio of true positive predictions to the total number of positive predictions made by the model), for the unpopular class (0), was 0.61. This means that 61% of the songs predicted as unpopular by the model were actually unpopular. For the popular class (1), the precision is 0.58, indicating that 58% of the songs predicted as popular belonged to this category.

Recall (the sensitivity or true positive rate) is the ratio of true positive predictions to the total number of actual positive instances in the dataset. The recall for the unpopular class(0) is 0.42, indicating that the model identified 42% of the actual unpopular songs. The recall for the popular class (1) is 0.75, meaning that the model captured 75% of the actual popular songs.

The F1-score (the mean of precision and recall provides a balanced measure of the model's performance) for the unpopular class (0) is 0.50, while for the popular class (1), the F1-score is 0.65. These scores consider both precision and recall, providing a summary of the model's overall effectiveness for each class.

Support represents the number of samples in each class. In this case, there were 145 songs in the unpopular class (0) and 154 songs in the popular class (1). Considering the macro average, which calculates the average of precision, recall, and F1-score across both classes, the model achieved an average precision, recall, and F1-score of approximately 0.59. The weighted average takes into account the support of each class and provides a weighted average of precision, recall, and F1-score, yielding similar results of approximately 0.59.

6. Conclusion

The objective of this study was to investigate what audio and visual features (and to what extent) influence music popularity on Spotify. The research was conducted in two parts.

In Part One, the importance of audio features in predicting song popularity was explored, including danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration. By scraping data from the Spotify API, a diverse dataset comprising both popular and unpopular music was collected, aiming to replicate previous research in this area.

The regression task was approached by training three distinct models: Linear Regression as a baseline, xGBoost, and random forest regressor. The results revealed that the random forest model outperformed the other models, followed by xGBoost and OLS. However, it is important to note that none of the models achieved a high level of performance. The random forest model yielded metrics with RMSE: 19.23, MAE: 15.18, and R-squared: 0.15. Despite optimizing the models using grid search cross-validation, the performance remained suboptimal. A feature importance analysis revealed that speechiness was identified as the most important feature, while the remaining audio features exhibited relatively similar levels of importance. The difference in importance between the most significant feature and the least significant feature was approximately 0.2.

As a result, this study was unable to replicate the findings of previous research in this field. Several factors may have contributed to this outcome, including changes in the music landscape and trends, as well as potential issues related to model specifications. These factors will be further discussed in the subsequent section.

In Part Two, this study extended this research area by investigating whether popularity could be predicted or classified based on features derived from the cover images of songs on Spotify. By scraping the images from the API and pre-processing them, a VGG16 model was employed for feature extraction. To visualize the differences between popular and unpopular songs, a t-SNE algorithm was applied for dimensionality reduction, resulting in a 2D scatter plot. The plot revealed that image features from popular songs spanned a wide range, while features from unpopular songs tended to cluster together, forming five distinct clusters.

To classify songs into popularity groups based on image features, a support vector machine (SVM) was trained. The accuracy achieved was approximately 0.59. The interpretation of this accuracy score and its ability to effectively distinguish between popularity groups remains a pending task by comparing it to a baseline accuracy metric.

This study sheds light on the limited success in replicating previous findings regarding the predictive power of audio features for music popularity on Spotify. Furthermore, it introduces a novel exploration of image features in predicting popularity, albeit with moderate success. The report will conclude by discussing alternative approaches, potential areas for future research, and the limitations encountered throughout this study.

7. Limitations and Future Research

The present study has several limitations that should be considered when interpreting the findings. One of the major limitations is the small size of the dataset. The study was conducted using data on approximately 1500 songs and cover images from Spotify, which may not be representative of the entire population. Future research could aim to replicate this study on a larger and more diverse dataset. This could be achieved by implementing techniques such as time delay and pagination in the data scraping process to retrieve a more comprehensive sample.

Another limitation pertains to the regression task and the handling of continuous variables. While models that are not sensitive to different data ranges were employed, future research could explore the performance of models specifically designed for certain types of data or investigate appropriate feature engineering techniques. Additionally, the study attempted to incorporate causal machine learning concepts; however, it is important to acknowledge that external factors may influence the results. For instance, the study did not consider genre as a categorical variable, which could potentially be a significant predictor. Exploring the relationship between genre and popularity in future research would require addressing the challenge of merging the extensive number of Spotify genres into more manageable categories.

Furthermore, although efforts were made to optimize the models and assess multicollinearity, additional diagnostic tests to validate model assumptions could have been employed. It is possible that the observed results may be attributed to overfitting or underfitting of the data. Future research or replications could investigate this aspect in greater depth.

Moreover, it is important to recognize that Spotify represents only one platform for accessing music data, and its user base may not be fully representative of the entire music audience and overall popularity trends. The scraping process itself may introduce biases due to the reliance on Spotify's recommendation algorithm.

Regarding the extraction of features from images, the current study utilized the VGG16 model, but it would be beneficial to compare the performance of feature extraction using alternative models. Furthermore, conducting dimensionality reduction techniques such as principal component analysis (PCA) could provide valuable insights and improve the understanding of the extracted image features.

Additionally, while the features were visualized using a scatter plot, the interpretation of the clusters remains unclear. Future research could aim to investigate the underlying characteristics and meaning of these clusters.

Finally, the performance of the SVM model could be further enhanced by implementing more advanced specifications and comparing it to a baseline model. Additionally, exploring the scope and limitations of the results through the inclusion of robust comparative analyses would contribute to a more comprehensive understanding of the predictive capabilities of the SVM model.

References

Amato, G., Bolettieri, P., Monteiro de Lira, V., Muntean, C. I., Perego, R., & Renso, C. (2017). Social Media Image Recognition for Food Trend Analysis. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1333–1336. https://doi.org/10.1145/3077136.3084142

Arcila, W. van A., Damian Trilling &. Carlos. (2022, March 11). *Computational Analysis of Communication*. https://cssbook.net/

Bello, P., & Garcia, D. (2021). Cultural Divergence in popular music: The increasing diversity of music consumption on Spotify across countries. *Humanities and Social Sciences Communications*, *8*(1), Article 1. https://doi.org/10.1057/s41599-021-00855-1

Chen, J., Chang, M.-C., Tian, T.-P., Yu, T., & Tu, P. (2015). Bridging computer vision and social science: A multi-camera vision system for social interaction training analysis. *2015 IEEE International Conference on Image Processing (ICIP)*, 823–826. https://doi.org/10.1109/ICIP.2015.7350914

Fraser, T., Crooke, A. H. D., & Davidson, J. W. (2021). "Music Has No Borders": An Exploratory Study of Audience Engagement With YouTube Music Broadcasts During COVID-19 Lockdown, 2020. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.643893

Gulmatico, J. S., Susa, J. A. B., Malbog, M. A. F., Acoba, A., Nipas, M. D., & Mindoro, J. N. (2022). SpotiPred: A Machine Learning Approach Prediction of Spotify Music Popularity by Audio Features. *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, 1–5. https://doi.org/10.1109/ICPC2T53885.2022.9776765

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*, *5*(5), 171274. https://doi.org/10.1098/rsos.171274

Koller, T., & Grabner, H. (2022). Who wants to be a Click-Millionaire? On the Influence of Thumbnails and Captions. *2022 26th International Conference on Pattern Recognition (ICPR)*, 629–635. https://doi.org/10.1109/ICPR56361.2022.9956202

Lau, D. S., & Ajoodha, R. (2022). Music Genre Classification: A Comparative Study Between Deep Learning and Traditional Machine Learning Approaches. In X.-S. Yang, S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of Sixth International Congress on Information and Communication Technology* (pp. 239–247). Springer. https://doi.org/10.1007/978-981-16-2102-4_22

Libeks, J., & Turnbull, D. (2011). You Can Judge an Artist by an Album Cover: Using Images for Music Annotation. *IEEE MultiMedia*, *18*(4), 30–37. https://doi.org/10.1109/MMUL.2011.1

Modeling Language Usage and Listener Engagement in Podcasts. (n.d.). Spotify Research. Retrieved May 22, 2023, from https://research.atspotify.com/publications/modeling-language-usage-and-listener-engagement-in-podcasts/

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press. https://library.oapen.org/handle/20.500.12657/26040

Swarbrick, D., Bosnyak, D., Livingstone, S. R., Bansal, J., Marsh-Rollo, S., Woolhouse, M. H., & Trainor, L. J. (2019). How Live Music Moves Us: Head Movement Differences in Audiences to Live Versus Recorded Music. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02682

Tang, Y. (Elina), Sridhar, S. (Hari), Thorson, E., & Mantrala, M. K. (2011). The Bricks That Build the Clicks: Newsroom Investments and Newspaper Online Performance. *International Journal on Media Management*, *13*(2), 107–128. https://doi.org/10.1080/14241277.2011.568420

Tf.keras.applications.vgg16.preprocess_input | TensorFlow v2.12.0. (n.d.). TensorFlow. Retrieved May 24, 2023, from https://www.tensorflow.org/api_docs/python/tf/keras/applications/vgg16/preprocess_input

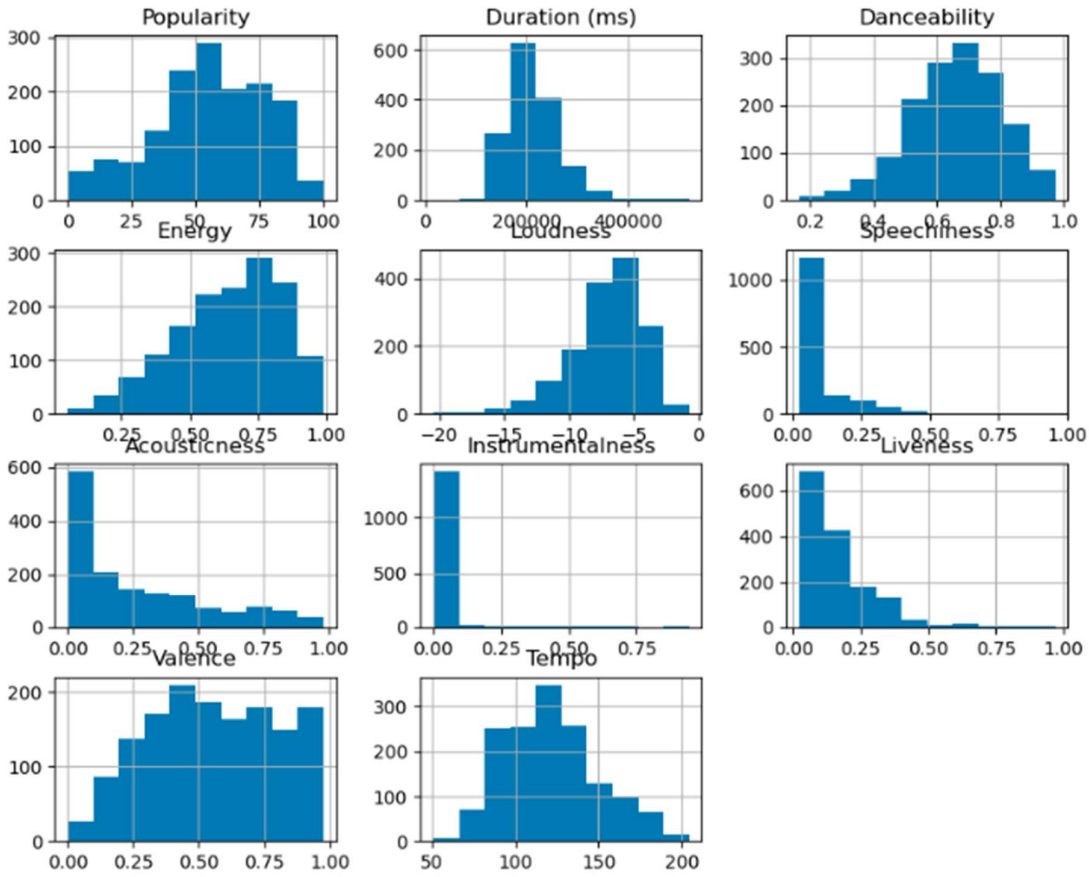
Venkatesan, T., Wang, Q. J., & Spence, C. (2022). Does the typeface on album cover influence expectations and perception of music? *Psychology of Aesthetics, Creativity, and the Arts*, *16*, 487–503. https://doi.org/10.1037/aca0000330

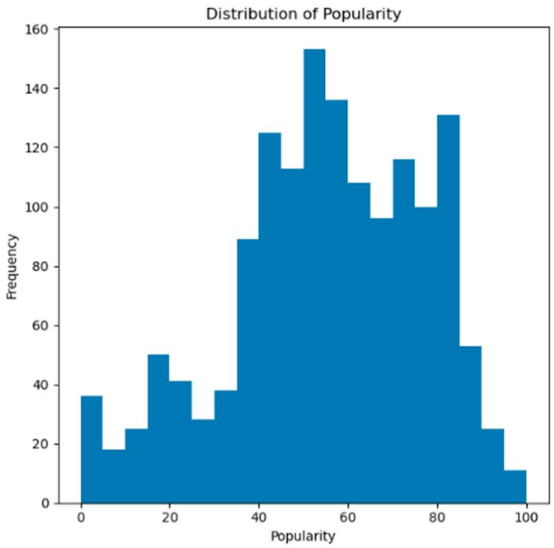
Appendix

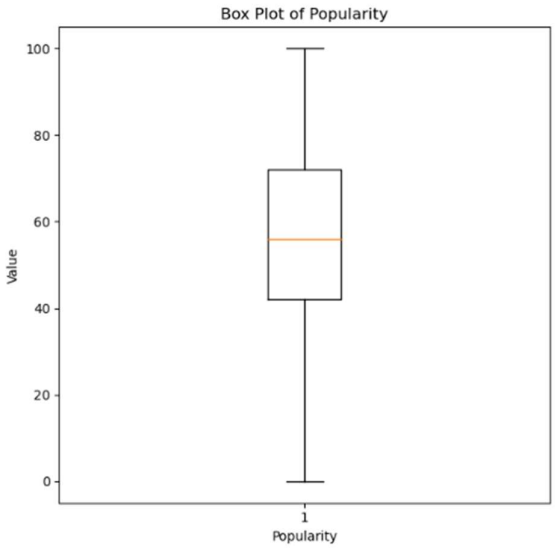
All data and code for replication available at:

https://drive.google.com/drive/folders/1u31x3CUVHfY25WVV6h4_fKINIUJADsmF?usp=drive_link

Distributions

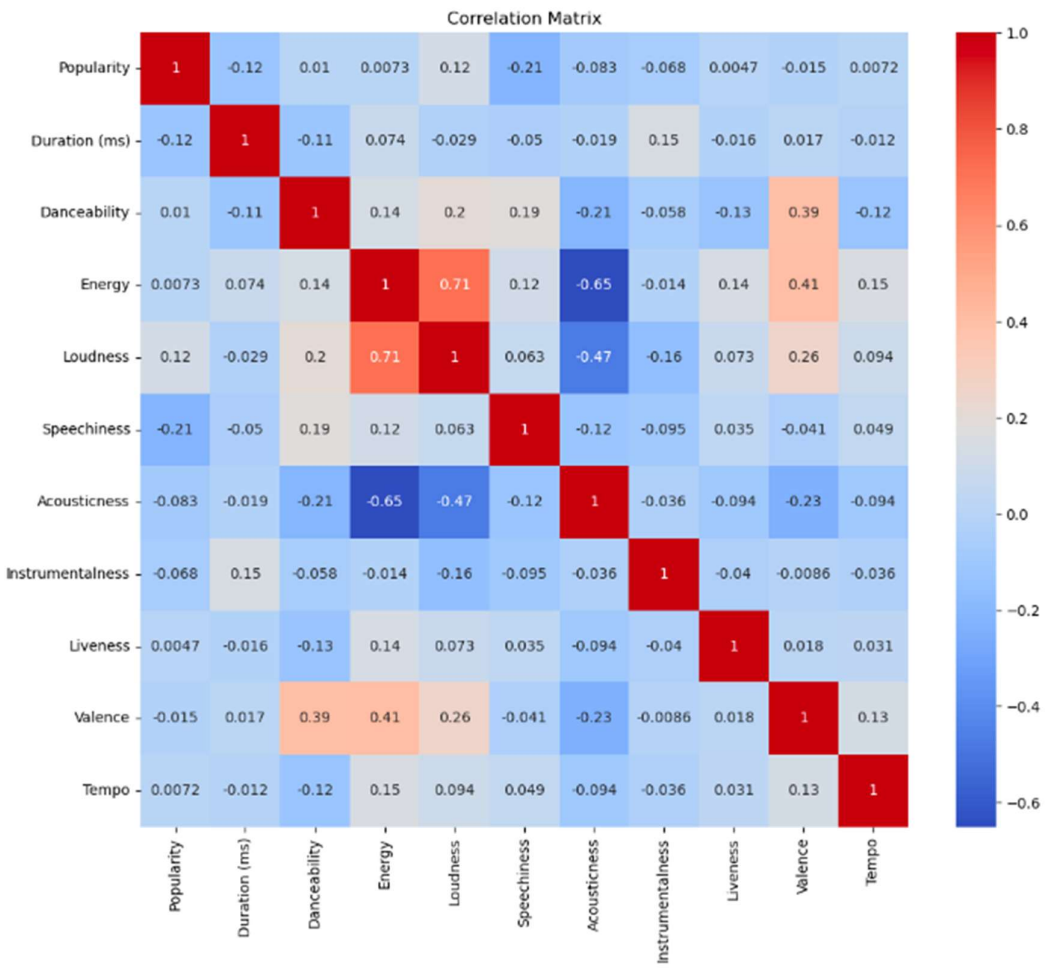






	Popularity	Duration (ms)	Danceability	Energy	Loudness	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo
count	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000	1492.000000
mean	54.707775	212067.668231	0.661139	0.644786	-7.087983	0.093845	0.267860	0.023640	0.171589	0.553716	121.561242
std	21.895499	52402.201980	0.141502	0.193802	2.805988	0.097514	0.266983	0.105750	0.125376	0.243714	27.665917
min	0.000000	16400.000000	0.167000	0.051800	-20.461000	0.024400	0.000025	0.000000	0.022200	0.000000	50.658000
25%	42.000000	175422.000000	0.568750	0.511000	-8.491000	0.037075	0.039200	0.000000	0.092700	0.365000	99.320250
50%	56.000000	205492.500000	0.668000	0.674000	-6.645000	0.051500	0.176000	0.000005	0.122000	0.545000	121.937000
75%	72.000000	238969.000000	0.762000	0.798000	-5.116500	0.102250	0.430000	0.000323	0.217000	0.754250	137.924750
max	100.000000	519665.000000	0.972000	0.991000	-0.722000	0.957000	0.979000	0.945000	0.970000	0.976000	204.396000

Correlations



VIF

	Variable	VIF
0	Popularity	0.102446
1	Duration (ms)	0.062114
2	Danceability	0.308620
3	Energy	0.704541
4	Key	0.006601
5	Loudness	0.562043
6	Speechiness	0.146312
7	Acousticness	0.458059
8	Instrumentalness	0.080757
9	Liveness	0.050563
10	Valence	0.341928
11	Tempo	0.072616