Predicting Album Artwork Colors

The Data Cavaliers
Adina Kugler, John Hope and Adam Kippenhan (Group Leader)
DS 4002
November 9, 2022

Restate Hypothesis

The album art color for song albums can accurately predict the genre of a song at a rate greater than 50%.

Executive Summary

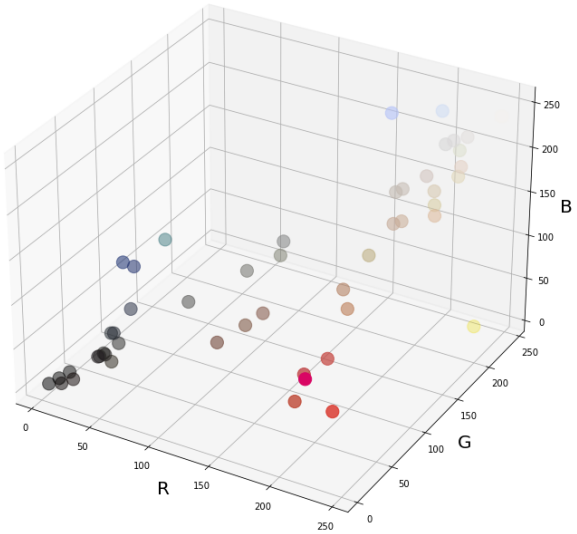
This document contains our data summary and plan for how we will analyze the data. We discuss the finding and establishment of our data with the data dictionary. This plan will dictate how we will move forward in the manipulation and analysis of our data and the basis on which we will build our conclusions and create additional visualizations.

Data Discovery Findings

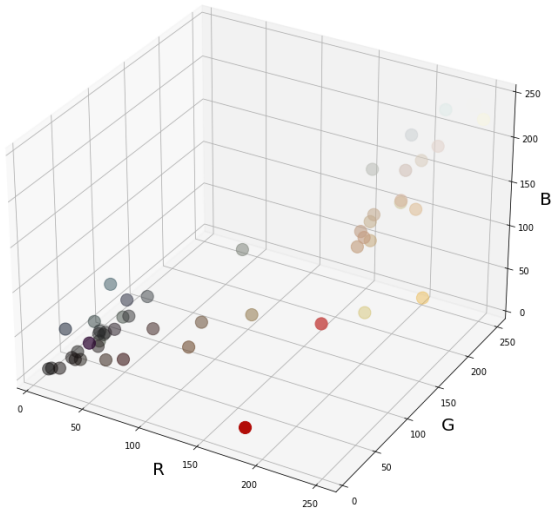
The data for this project was collected using the Spotify API. Data for 6 different genres was collected using Python. For each genre, 20 queries were made with each containing 50 records to reach Spotify's maximum limit of 1000 records. The results from the API requests are in JSON format. There are 3184 total album art images that were downloaded from these request responses. The number of images varies for each genre as many of the image URLS provided in the request responses were duplicates. The data and accompanying data dictionary can be found at the following GitHub repository: https://github.com/akippenhan749/AlbumArtCoverColors.

With the dataset established, we can begin the exploratory analysis phase of the project. Our exploratory data analysis consisted of generating different visuals to get an initial understanding of the overall trends of page visits across different sites. The overarching question for this phase is: Can we see differences in the predominant colors of album covers across different genres? In order to get an initial understanding of this, we will visualize the predominant colors of samples of album covers across the genres. The predominant color of the album covers was found using the Color Thief package in Python, which takes an image, and returns an RGB tuple of the predominant color [1]. The following 3-d scatter plots plot 50 albums' predominant colors, with RGB values on the axes, and each point is colored by the predominant color of the corresponding album. This is done for each genre, and can be seen below:

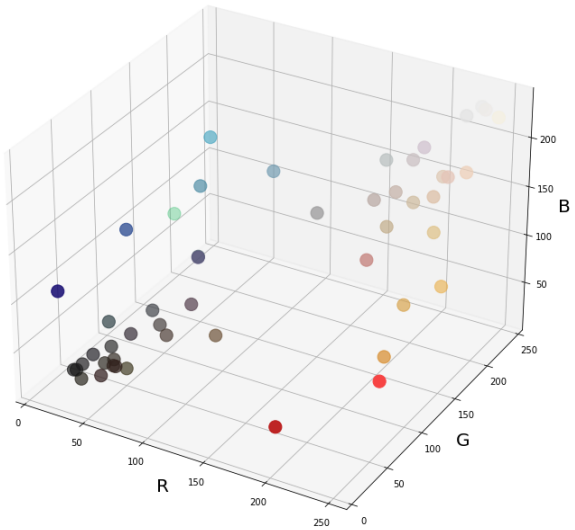
Alternative



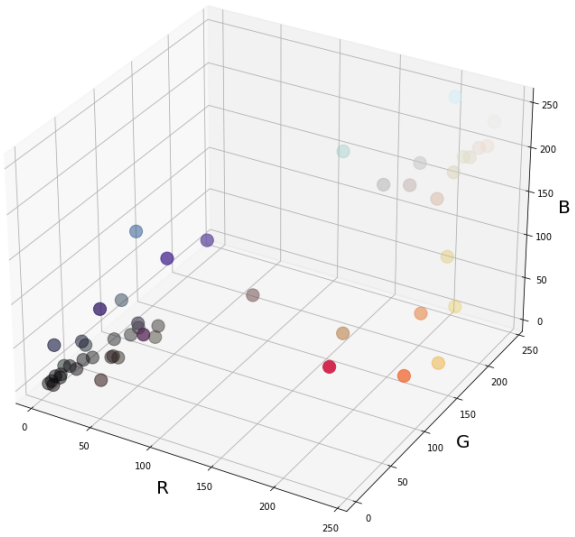
Country



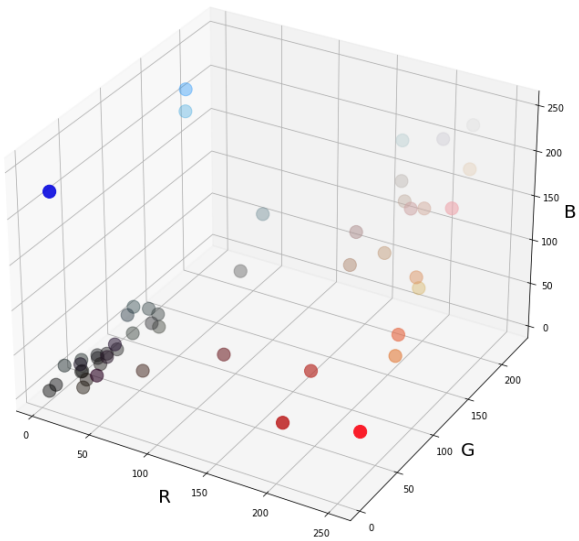
Jazz



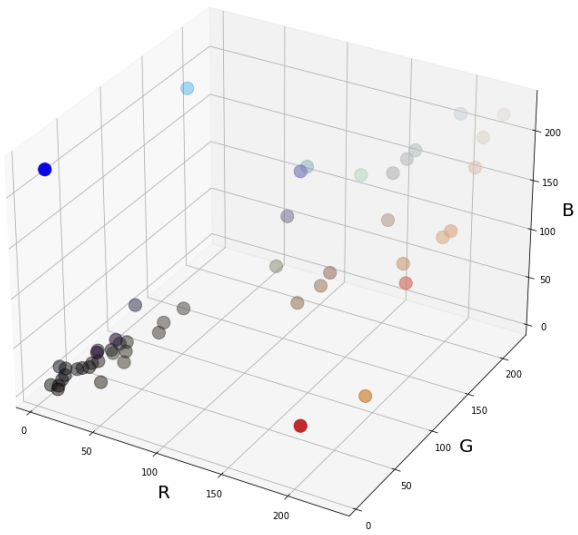
Hip-Hop



Pop



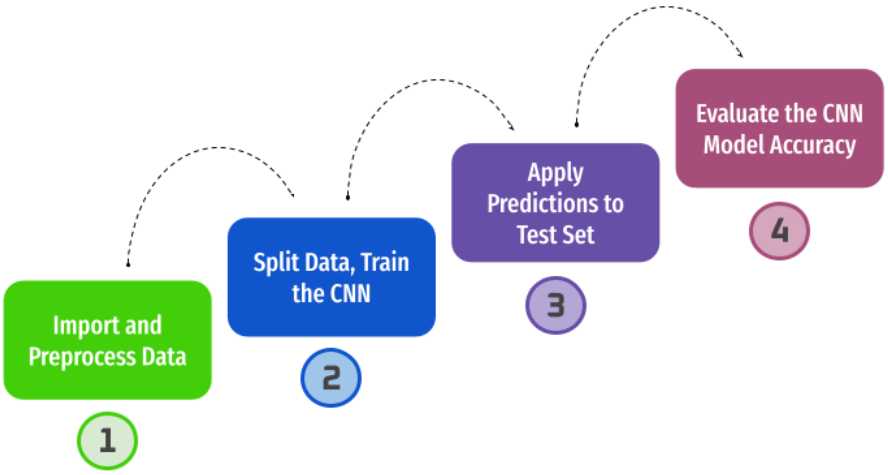
Rock



Although we are working with limited data from each genre, we can see some similarities and differences among the plots. From these plots, one of the biggest takeaways is that there appears to be a cluster of points in the bottom left portions of each graph. What this tells us is that across genres, the most popular predominant color of album covers is black/dark gray. Although this appears to be the most popular color, we see some slight deviations in color palettes across the genres. For example, in the country genre plot, we see that almost all of the predominant colors are more neutral colors (black/gray, tans, browns), whereas in the rest of the genres, there is a slightly larger presence of a wider variety of colors, like more vibrant reds, blues, and greens. Alternative, as we can see, has some more reds compared to the rest of the genres, whereas jazz has more blue/greens present. With this being said, we are working with limited data, so all of the patterns could potentially be different if we visualized all of the album covers on hand. From all of this information, we can take away that there are slight differences in the usage of color in album covers across genres, but there is an overall commonality of the abundance of predominantly black/gray album covers.

With these gained insights in mind, we can move onto our analysis plan, which will outline our process of testing our hypothesis.

Analysis Plan



Our main goal in our analysis will be to compare our findings to our hypothesis to see if cover art is related to genre and if cover art can be used to classify as a specific genre with some level of accuracy.

Import and Preprocess Data

Our first step to begin the analysis will be to import our data into our analysis tool of choice: the data science-oriented programming language R. We chose to use R for our analysis because of our familiarity with the language as well as its ability to easily create visualizations that will effectively convey our findings. Our data, which is currently in folders of JPG files and JSON format for each genre, will need to be compiled appropriately in order to train our convolutional neural network (CNN). Ideally, the data is to be converted to a CSV file format with a data frame. Then we can combine the results from each genre into a single dataframe. After this, we can remove unnecessary columns, missing values, duplicates, etc., and create a new column for the genre.

Split Data, Train CNN

Once our data has been visualized and summarized, our next step will be to split it up to make our future analysis and machine learning model creation and training easier. R will make it easy to accomplish this and put the data into separate data frames. We will split the data with a 75% split as train and 25% split as test. The training data will then be added to our CNN model to train it and training error rates will be recorded.

Apply Predictions to Test Set

The model created from the trained data will be tested against the remainder of the data to assess its accuracy and further improve the model. This process consists of making predictions, in the form of assigning a photo a genre and assessing if our prediction falls within a certain range of error from the observed result. This will allow for testing of the hypothesis.

Evaluate the CNN Model Accuracy

The accuracy will be the primary metric of evaluating that the model predicts the test data at 50% accuracy. Further visualization will be performed at this stage to evaluate the strength of the predictions for each genre. Looking at which genre is predicted at the highest accuracy will allow for analysis of ethical questions surrounding this type of predictive modeling, namely if the models are predicting based on race or gender. Additionally, accuracy is a metric worth using to understand if there are similarities in genres or if cover images do not relate to each other.

References

[1] L. Dhakar, “Color Thief,” *Color thief*. [Online]. Available: https://lokeshdhakar.com/projects/color-thief/. [Accessed: Nov. 9, 2022].