Colin Fitzgibbons

Dr. Chou

CSC 485-001: Data Science Capstone

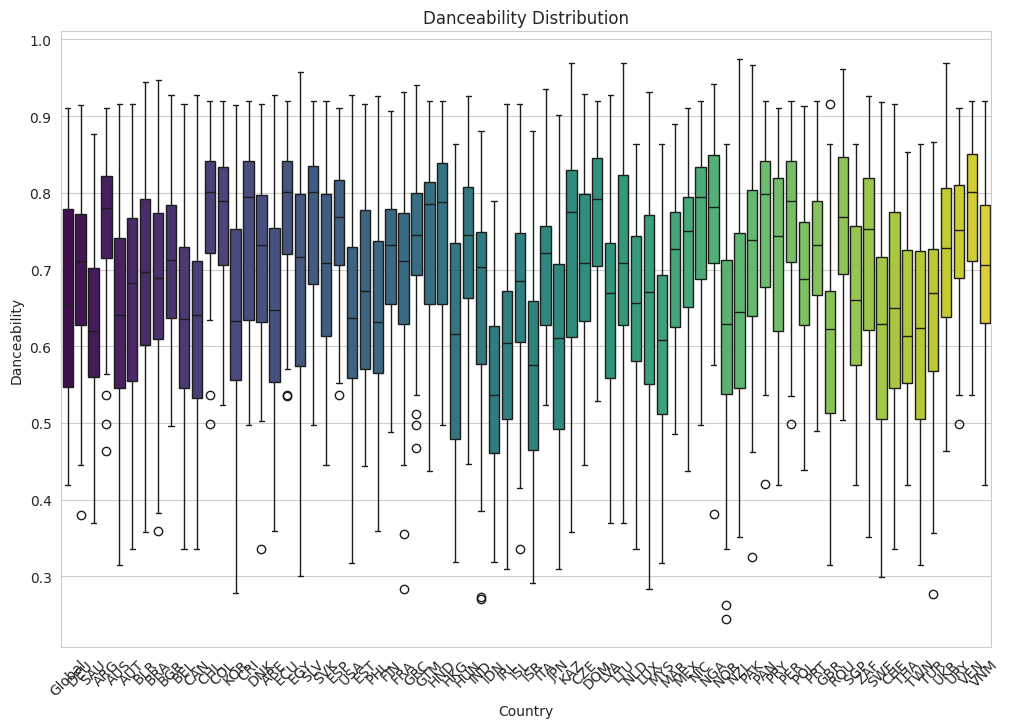
04 March 2024

Predicting Top Songs Within Spotify’s Categories

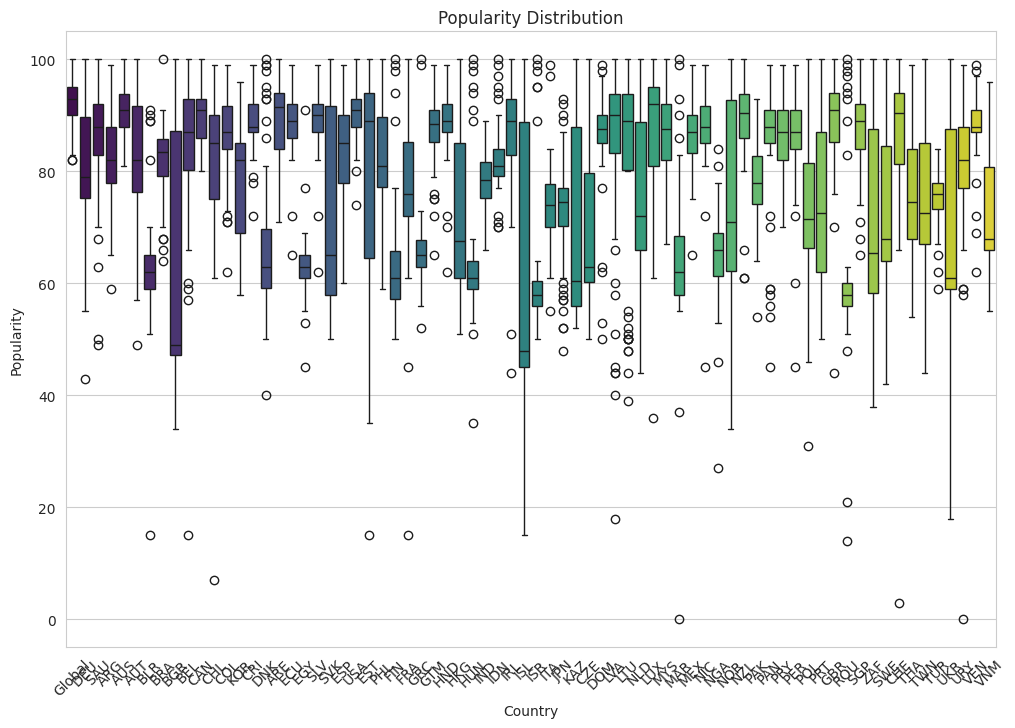
Every day, the streaming giant Spotify hosts millions of users to not only upload and publish content, but also to consume content and boost individual popularity of a piece of media. Spotify is a Swedish media and audio streaming service, known for music, podcasts, and videos that was founded in 2006, with approximately 602 million users worldwide combined from free and premium subscriptions. In the media world, especially with regards to music, Spotify is known as a powerful platform for artists, as the streams generated for a piece of media on Spotify contribute to revenue and virality for artists and their management teams, as well as for the art itself. Most notably, Spotify generates new ways to support artists as well as subscribers, all with different features to attract new markets and further generate virality and revenue for artists and creators. In the modern era, many artists, as well as the industry, gauge their success off of streams and downloads of released songs on streaming platforms such as Spotify, rather than any other form of music media consumption.

Being a daily user of Spotify myself, I was able to find three datasets on Kaggle that contained data generated and harvested by Spotify. All three of the datasets I utilized all contained the categories that Spotify created and designed for grouping and classifying songs. Some of these classifications defined by Spotify include danceability, acousticness, speechiness, energy, loudness, positiveness, instrumentalness, and (most importantly) popularity. All of these variables, apart from popularity, are scored on a scale of 0.0 to 1.0, with 0 being the lowest and 1 being the highest. Popularity on the other hand, throughout all three of the other datasets, was not a score, but a ranking ranging from 0-100 based on a song’s virality. Using these three datasets, I was certain that I had enough data to work with in order to assemble a prediction model to rank songs based on popularity, finding common rankings within the categories they were classified with.

The first dataset I used was the original dataset I was able to use with my first project, the dataset from kaggle titled *World's Spotify TOP-50 playlist musicality data.* This dataset, captured from the 45th week of November 2023 (November 6th-12th), held the top 50 playlist data for 73 countries, as well as the global top 50 song playlist data. The dataset contained a total of 3,588 entries, with thirteen columns of the song information (title, artist, country,...) as well as the scores within the categories defined by Spotify mentioned above. Upon taking the basic data preprocessing and cleaning steps with this data by removing NaN values and removing duplicates, I created two box plots to showcase important factors within the data. The first box plot I created was a distribution of danceability scores ranked by country.

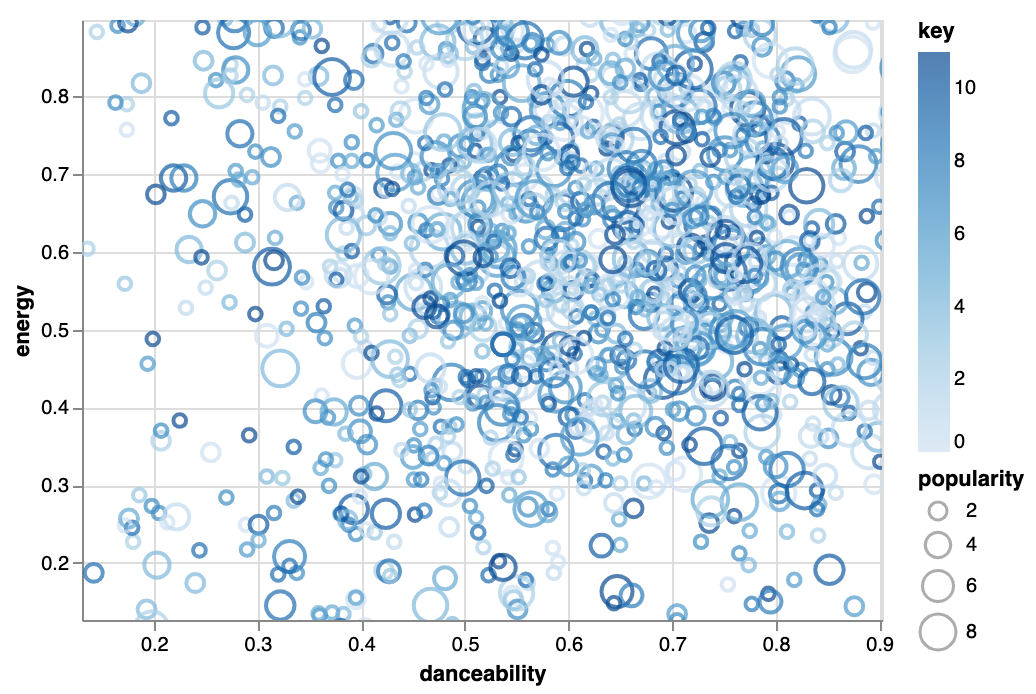


From this chart, I was able to gather information that one of the songs within the dataset with one of the lowest danceability scores came from Norway (outlier), whereas one of the highest danceability scores came from New Zeland, not far below a perfect 1.0. Following this chart, I also created another distribution box plot using the popularity scores for each country.



From this distribution chart, we can see that some of the lowest overall popularity rankings come from Malaysia and Uruguay, whereas many countries hold popular ranking tracks. With these two plots, we can see that there is some correlation with popularity and danceability overall, even though a high danceability score overall does not make a song popular.

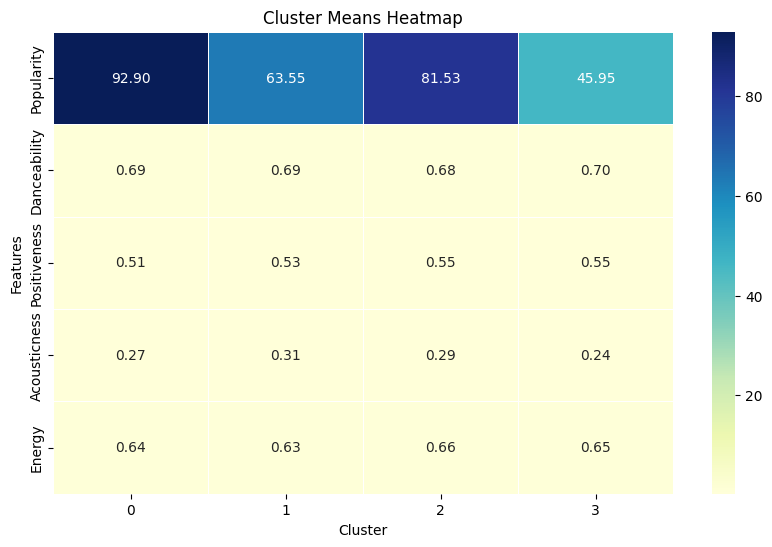
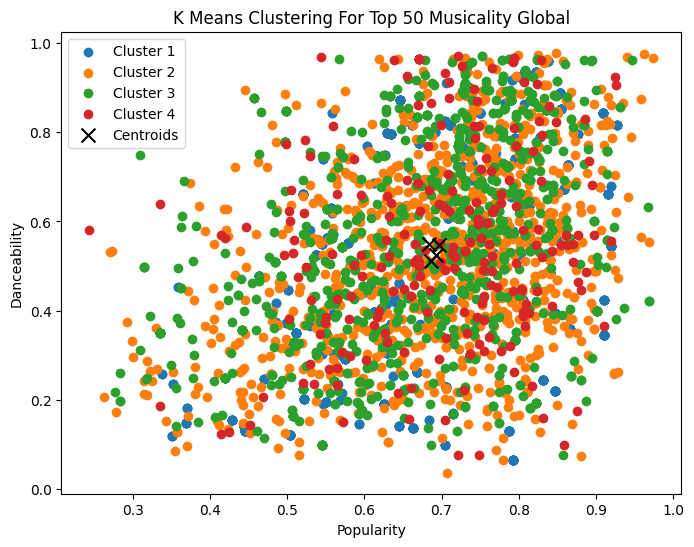
The next dataset I utilized within my project was another kaggle dataset titled *Spotify unpopular songs.* This dataset, unlike the last one, provides scores and rankings for some of the most unpopular songs on the Spotify platform, regardless of country of origin. The popularity within this dataset is described as being based on payback rate overall, as well as total and recent amount of listeners. Despite the dataset not being updated since 8/31/2022, this dataset still provides an important insight into song data of songs that were classified within spotify, despite having a lack of popularity and virality. After performing basic cleaning on this dataset, one plot that I found to be informative, suggested by Google Colab was the interactive scatter plot chart comparing the energy vs danceability of songs mixed with the popularity of the song overall.



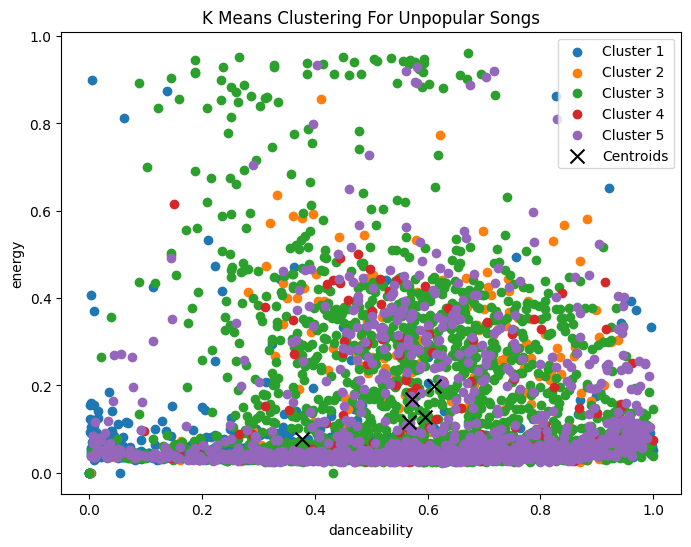
Within this chart, the bigger size of a circle representing a song showcases a higher score of popularity compared to those with a smaller circle, where the place of the circle represents where the song ranks within Spotify’s energy and danceability characteristics. From the chart, I was able to gather that higher popularity rankings were awarded to songs with higher energy and danceability, as well as songs with either just high energy, or just high danceability. While high energy and high danceability alone or together do not automatically mean high popularity, this chart, despite being from unpopular songs on Spotify, shows an insight that there is pattern with a song earning a higher popularity ranking while holding higher ranking energy and danceability scores.

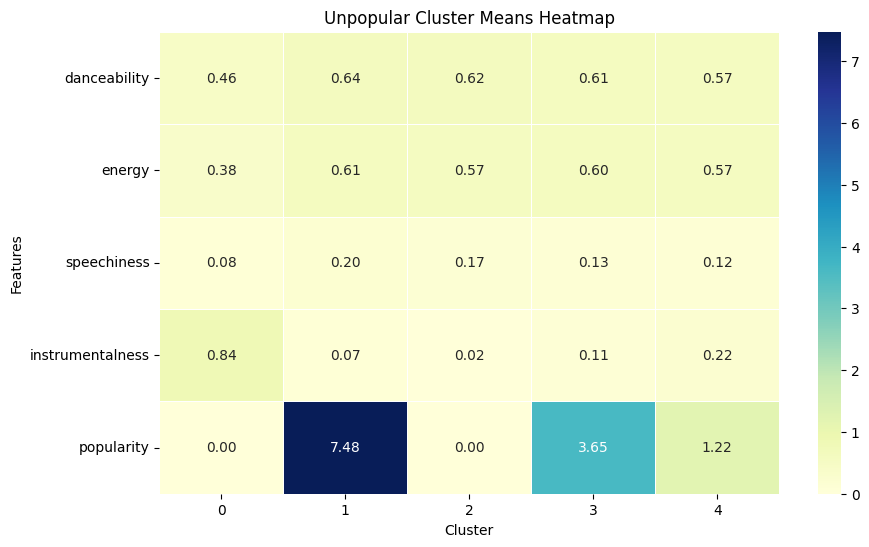
The last dataset I utilized for the project was another dataset from Kaggle titled *Spotify Top 100 Songs of 2010-2019.* This dataset contained the top 100 songs for the nine year period of 2010-2019, and contained al the same categories and rankings, as well as two more columns containing the year the song released as well as the top year of the song’s popularity, where the song gained the most virality. However, this dataset contained abbreviations for the columns of all the defined Spotify categories, for example where instead of “energy” the dataset would have the label “nrgy” for the energy score of all the songs. Overall, this dataset, despite the abbreviations, proved to run consistent in terms of popularity and scoring with songs along with the other two datasets.

With the first two datasets, I also performed a clustering analysis using k-means clustering in order to build two scatter plots with centroids and two heatmaps. For the first dataset of the Top 50 songs globally and for countries, I implemented 4 clusters, and built a scatter plot based on the popularity and danceability of songs, as well as building a heatmap with the feature categories of energy, acousticness, danceability, positiveness, and overall popularity.



For the second dataset of unpopular Spotify songs, I created a scatter plot using 5 clusters, and clustering by energy level and danceability score, similar to the chart above. The heatmap for the unpopular song data also contained the features of instrumentalness, speechiness, energy, and danceability along with the overall popularity score.





From these clustering models, the heatmaps proved to be more helpful and detailed, as the scores that associated with the most popular songs on both reflected higher scores amoung energy and danceability, along with the other features, further supporting that these two categories showed some significance.

For building my predictive model, I strived for an accurate model that would be able to pull from all three datasets and accurately predict popularity based on the scores of a song within the relevant categories awarded by Spotify, as well as gain a greater understanding of songs and performance within the app. However, there were some challenges that came with assembling a model. First, the datasets were unable to be combined properly, as per the differences with structure and modeling within each one, as well as case sensitivity within row and column names, particularly seen within the Top 100 songs from 2010-2019 dataset, as all the column names for the Spotify categories were abbreviated or shortened. Additionally, working with an extensively larger dataset required more computing power, one aspect where Google Colab can fall short for non-premium users. This challenge led me to build three separate prediction models, one for each dataset. After building and testing the accuracy of multiple prediction models for the datasets, such as random forest regression, gradient boosting, K-nearest neighbors regression, and a neural network model, I landed on using a linear regression prediction model for all three of the datasets. All three of the linear regression models were trained and fitted with a size of 0.2, and a random state of 42. The calculated mean squared error, or MSE, for the linear regression models proved to be the lowest and most accurate for the models, leading it to be the best performing for this project, despite having the MSE be somewhat undesirable for the top 50 and top 100 datasets. However, for the unpopular song dataset, the MSE of the model came out to 3.42, showing that the unpopular song dataset to be one of the most helpful datasets with predicting a song’s virality, or lack thereof.

In conclusion, while the predictive modeling with linear regression provided difficulties among the three datasets, the dataset of unpopular songs on Spotify proved to be the most valuable, and the most helpful in terms of predicting based off the given Spotify defined categories. This model used the categories of danceability, acousticness, liveness, and valence crossed with popularity to predict. Overall, while the categories of danceability and postitiveness showed consistency from my first project with Spotify data, the categories based off the unpopular model of acousticness, liveness, valence, and energy can also contribute to a song’s popularity score, or lack thereof.