# Determining What Elements of a Song Make it Popular Based On Trend Studies and K-Means Clustering

Logan Wrinkle
University of Tennessee, Knoxville
lwrinkle@vols.utk.edu

Mahim Mathur
University of Tennessee, Knoxville
mmathur@vols.utk.edu

***Abstract*—For every song on Spotify, details are stored about elements of the song such as genre, dancibility, loudness, instrumentalness, energy, and much more. By relying on a dataset that contains this data for the most popular songs of this century, we can study trends in popularity and the elements that make a song popular. In our study, we visualize the popularity of each genre and how the popularity of each genre has changed over time, we measure and visualize the correlations between song elements by using a heatmap of correlation coefficients, and we use the K-Means clustering method to determine what elements can be put together in a song to make it more popular. Our K-Means study was unsuccessful due to the large number of track elements and the lack of measurement units provided by the dataset.**

***Index Terms*—Spotify, Music Trends, Popular Genres, K-Means Clustering**

## I. Introduction and Motivation

For this project we choose to use a publicly available dataset which can be found here [1]. Music is an almost universal passion which has major differences across cultures and time periods. Our goal is to try and explore some of that history using data. Our dataset contains 2000 of the most popular songs on Spotify over 20 years (100 for each year), which tracks not only genre but elements such as energy, danceability, upbeatness and acousticness for each song.

Our group wants to explore two questions:

- What trends can we spot over the 20 years that the dataset covers?
- What elements, or combination of elements, are associated with popular songs?

Using Python libraries and data mining tools, we can visualize and analyze Spotify data to understand what listeners have liked the most this century.

## II. Methodology

The dataset chosen is a collection of the top Spotify hits from 2000-2019. The dataset includes the track's name, artists, release year, explicitness, track duration, popularity and lots of data on the sound of the track. It includes measures of genre, dancibility, loudness, instrumentalness, energy, speechiness, tempo, and more. This data is collected from the Spotify API, which gives the values for tempo and loudness, etc... The dataset has the top 2000 tracks from the 20 years spanning 2000-2019, and it has 18 columns for characteristics of the tracks. Most of the columns are integers, but strings and boolean values are present as well. An important observation is that a few songs are present from 1998, 1999, and 2020, which is likely because a song can be popular outside the year it was released, and Spotify likely gathered the information soon after the year ended, potentially allowing the 2020 song.

First, we worked to understand the dataset and determine how many songs there are per year and what track elements are included. Once we explored the dataset, we visualized the data with the Pandas and Matplotlib libraries in Python. Specifically, we graphed the average popularity scores over time, the number of songs in each genre per year, and the percent-contribution of each genre per year. Then, we analyzed which track elements work best together using a heatmap of correlation coefficients.

Finally, we applied K-Means clustering by utilizing the Sklearn library to identify any interesting relationships between track elements and popularity. For the K-Means clustering method, we used the elbow-method to find the optimal number of clusters. We visualized the clusters and distinguished them by setting each cluster to a different color, and we highlighted the centroid of each cluster.

## III. Results

### *Trend Analysis*

When graphing the average popularity score for each year, we find that 2018 had the highest popularity score meaning that the most popular songs from this century come from 2018. Looking at genre popularity, Rock, Metal, and R&B fell into a state of decline over the 20 years of the data. Pop had a relatively steady percent of songs over time, around 40%. Hip-Hop and dance/electronic music were on the rise and overall, along with Pop, were very consistently the top genres each year. Additionally, by calculating the correlation coefficients between every song element, we discovered how loosely-related the elements are among each other, which was not expected. By visualizing the correlation coefficient matrix in a heatmap, it is evident that energy and loudness are the most correlated with each other, as shown in Figure 1.

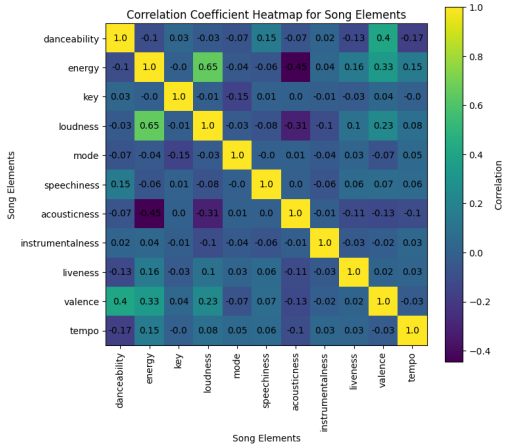


Fig. 1. Correlation Coefficient Heatmap for Song Elements

### *K-Means*

The elbow-method suggested 4 clusters. Clustering with K-Means failed to get any meaningful or distinct clusters, as evidenced by the clusters being on top of each other in Figure 2. The proximity of the clusters together indicates that the K-Means algorithm is not successful in finding clear separations or distinctions in the data. This poor performance was constant throughout all tests run, including reducing the dimensionality of the dataset, changing the number of clusters, and altering the visualization method by using a 3-D representation of the clusters. This would suggest that the dataset is poorly fitted for the K-Means algorithm. The hypothesis is that K-Means is too simple of a clustering algorithm and that music taste is too complex to obtain proper results. Specifically, there are too many variables involved in determining the popularity of a song, and the K-Means algorithm cannot find a clear way to use these track elements to calculate the best track elements for popularity. Another huge problem lies within the dataset: the track element measurements are not labeled. The dataset only contains numeric values but no metric of measurement. Our K-Means algorithm uses scaled values of the data, but without units for the data, we are left with an incomplete understanding of the data. For data analysis, this lack of unit makes it challenging to provide our code with the most accurate information so that we can get valuable information from the data.

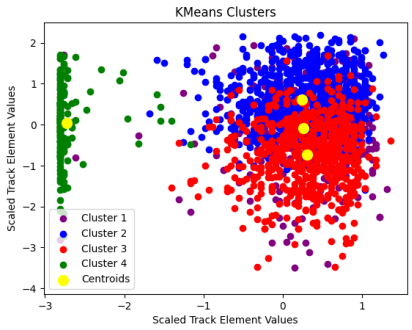


Fig. 2. K-Means Clustering results

As shown in Figure 2, three of the clusters are on top of each other and are in close proximity. The green cluster on the left only appears after introducing popularity scores into the data that K-Means performs on. If popularity is excluded, all four clusters are all on the right and almost indistinguishable. The centroid of each cluster is highlighted in yellow.

## IV. Conclusion, and Future Work

### *Conclusion*

From our results, we identified clear trends with genres over time, including the popularity of each genre every year this century. Although we expected certain track elements to be correlated with each other, we found few correlations among track elements. Importantly, we demonstrated that K-Means was an insufficient method to cluster our music dataset. This study also demonstrates that having datasets with labeled units and measurements is vital for thorough data analysis and for applying clustering methods such as K-Means. Furthermore, this study shows the complexity of music and possibly the weakness of trying to measure it. There are many other factors involved in the popularity of a song such as promotion, touring, and virality. Our code for genre trend analysis can be used to track the popularity of each genre over time. We generated graphs highlighting the average popularity of songs in each genre for every year included in this dataset. Our code also is used to visualize how many popular songs come from each genre per year.

### *Future Work*

In the future, finding a larger and more holistic dataset will be vital to understanding popularity trends. This dataset only contained the top 100 songs on Spotify for 20 years, which is relatively small compared to the thousands of songs released every year. Based on our foundational study of genre trends, more research can be performed to understand why the trends we see are occurring. For example, why is the metal genre not as popular as it was in the past? What causes genre preferences to shift over time? Importantly, more research can be done to find an alternative clustering method to separate the data and determine the elements that make a song more popular. More research can determine if clustering methods are valuable tools in answering these questions. What other data science tools can be used to measure the elements that make a song popular? By continuing this research, we could possibly predict whether a song will be a hit based on the elements of the song. There is one paper that has done some limited research on predicting popularity, but there is only one track element included in the study [2].

## References

[1] M. Koverha, “Top Hits Spotify From 2000-2019,” Kaggle, https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019.

[2] V. Ochi, R. Estrada, T. Gaji, W. Gadea, and E. Duong, ‘Spotify Dance-ability and Popularity Analysis using SAP’, CoRR, vol. abs/2108.02370, 2021.