Analyzing Features of Popular Spotify Songs

David Kozhevnikov **netid: dak255** May 9, 2025

Github: https://github.com/David-Kozh/CS439_Project

Introduction

This unsupervised learning project aims to uncover more information on current music trends by investigating how certain audio characteristics of Spotify songs potentially affect their popularity. The project attempts to answer this question through the usage of dimensionality reduction concepts to reduce the high dimensional space into essential components, clustering methods to uncover trends that repeat throughout the dataset represented by clusters of songs, and data visualization techniques that we have covered.

Motivation

My project is important because music trends are commonly difficult to identify, understand, and keep up with, due to their constantly evolving nature. This analysis can better equip music professionals to make more informed decisions in creating and promoting their music. I'm personally excited about this project as a music enthusiast/producer myself.

Method

Datasets

The primary dataset used for this analysis is obtained from Kaggle, and contains two csv files with data on the audio characteristics of Spotify songs. The data is split into a high_popularity_spotify_data table and a low_popularity_spotify_data, based on the song's popularity score being greater than or less than 70. Therefore the 'low popularity' scores are not necessarily bad or unpopular songs, just not the most popular at the time of the collection of the data from Spotify.

The secondary dataset 'Most Streamed Spotify Songs 2024' was also obtained from Kaggle, and contains performance data on songs across other platforms. Those columns of the dataset were not used for this analysis, but would be a good next step in evaluating performance of songs more holistically. Rather, this dataset is used to cross check the results of the analysis and evaluate them more deeply. This was done by inner-joining this table with the combined high and low popularity tables, to obtain a generally popular subset of the original dataset. This subset was then used to help evaluate the methodology of the analysis.

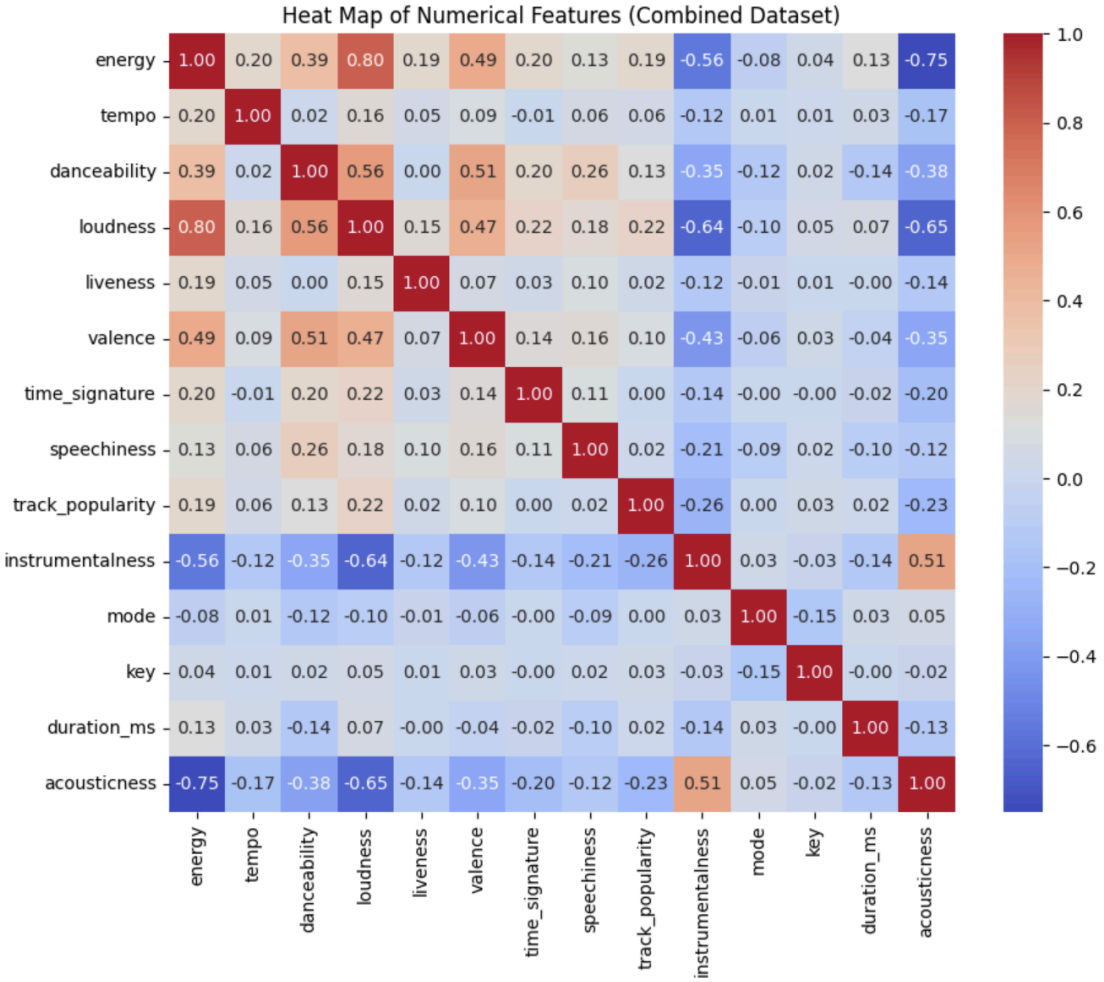
Data Format

The high and low popularity tables are formatted with twelve audio features (energy, tempo, danceability, loudness, liveness, valence, speechiness, instrumentalness, mode, key, duration, and acousticness) and eleven descriptive

features (track name, track artist, track album name, track album release date, track id, track album id, playlist name, playlist genre, playlist subgenre, playlist id, and track popularity). Most of the significant audio features are decimal values from zero to one (energy, danceability, liveness, valence, speechiness, instrumentalness, acousticness). Tempo is measured in beats per minute, and loudness is measured in db. Mode is an integer corresponding to the main musical modes. Key is “an integer from zero to eleven, mapping to standard Pitch class notation”, which represents the twelve major keys and their relative minor keys. The descriptive features are generally represented as Strings, except ‘Track Popularity’ which is an integer from 0 to 100 representing the song’s popularity score given by Spotify at the time the data was collected.

Analysis

After cleaning the dataset of the only 2 rows with missing values, I began by investigating the distribution of popularity scores in the combined dataset (high_popularity and low_popularity). This validated the realisticness of the dataset in that fewer songs had extremely high popularity scores in the range of 90-100. Then I began deeper analysis by creating a heatmap of the correlation matrix of audio features. This provides an overview of how the various audio features correlate with each other, especially with ‘track_popularity’.



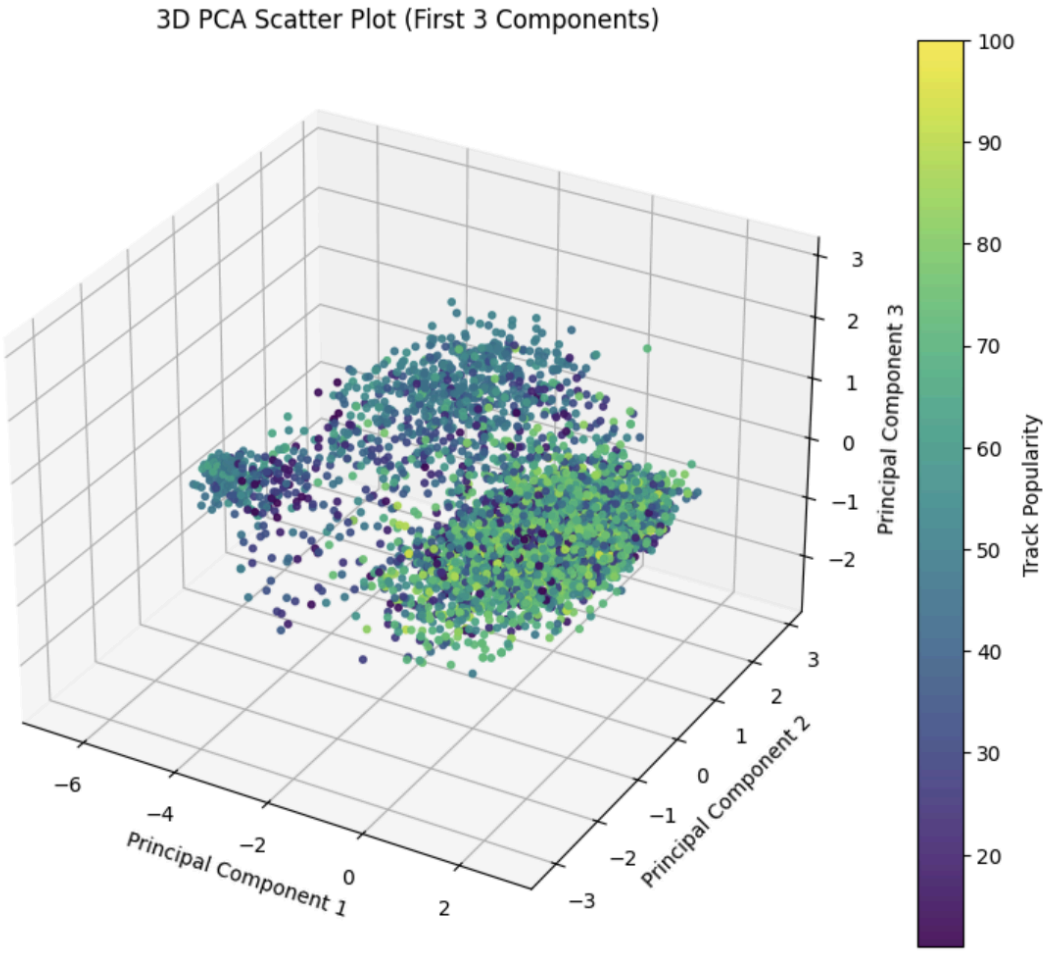
I then conducted Principal Component Analysis (PCA) on the dataset, to reduce the dimensionality, in an effort to ensure clustering was maximally effective. Building off this, I did KMeans Clustering on the space formed by the principal components. Recognizing that the dataset is not well separated and likely contains noise, I proceeded with DBSCAN Clustering. To test and measure success, the separability of the resulting clusters was checked by measuring their silhouette scores, and by testing the differences in the distributions of 'track_popularity' among the clusters.

To gain further insight into a notably large, but more popular, cluster formed by DBSCAN, I did Hierarchical clustering to obtain more precise subclusters. Then, to analyze these subclusters further, the median values for each of the numerical columns (audio features and popularity score), was compared to the median values for both the entire dataset and the high popularity songs, using robust z-scores to measure the differences in Median Absolute Deviation (MAD). The median values and MAD were used rather than the means and Standard Deviations, so that the impact of outliers would be reduced. Success of this methodology was tested by comparing the differences in the distributions of popularity scores among the subclusters, the 'high popularity' subset, and the general dataset. Further validation of results was done by inner-joining the 'Most Streamed Spotify Songs 2024' to the clustered data and checking which clusters are most represented in this 'Most Streamed' dataset.

Results

Principal Component Analysis (PCA)

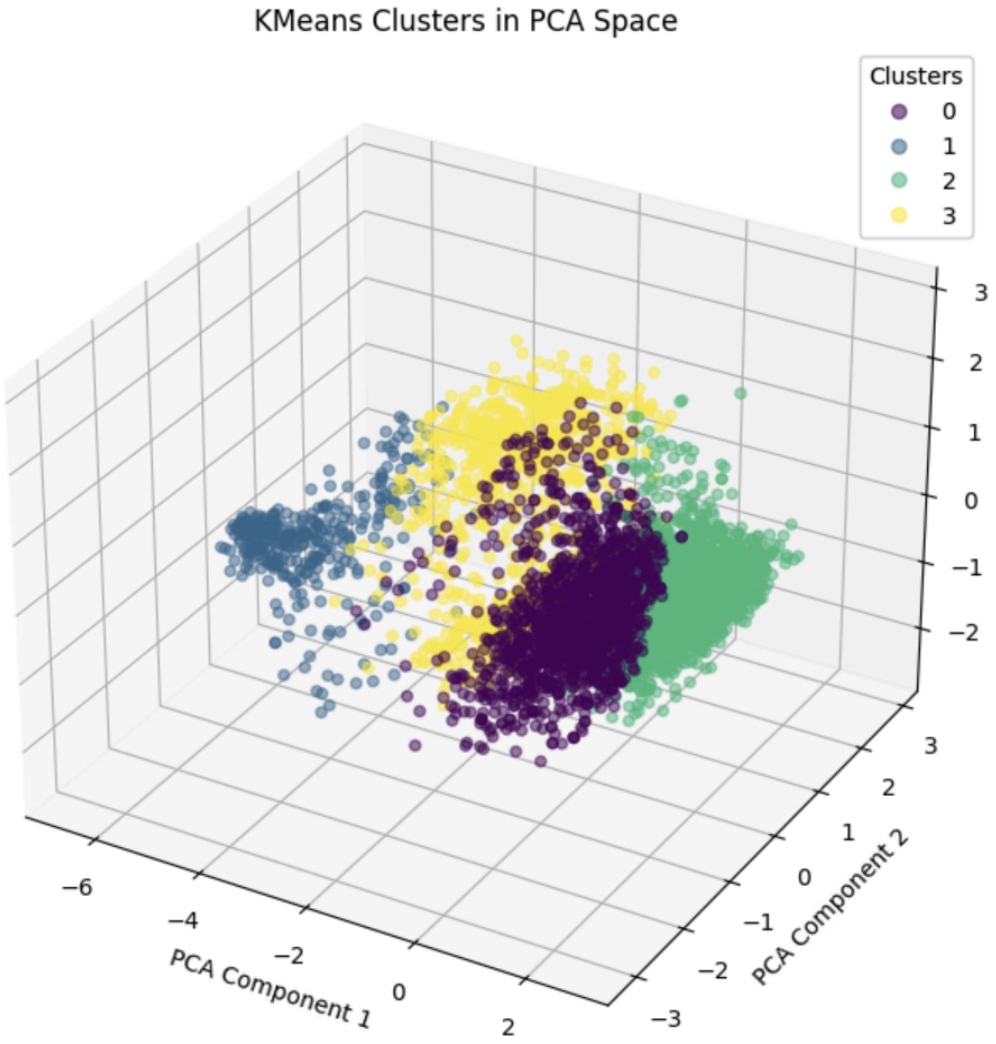
Initially PCA was done on all 12 audio features resulting in 8 components needed to reach near 90% cumulative explained variance. That PCA result was not as conducive to clustering and resulted in clusters with less differences in popularity. Hypothesizing that PCA was explaining variance in audio features that do not affect popularity, six of the audio features with very weak correlations to the 'track_popularity' descriptive feature were removed (tempo, key, mode, duration, liveness, speechiness). Through trial and error, conducting PCA on the remaining audio features (energy, danceability, valence, loudness, instrumentalness, acousticness) led to a better focus on the relationship driving popularity, better separated clusters, and clusters with more distinct differences in popularity. Reasonably, it was not productive for the PCA components to explain the variance in 'key' for example, since it is well known that all twelve pitch classes are equal (songs can be mapped to a different pitch class and still sound fundamentally the same). Based on the cumulative explained variance of the PCA Components exceeding 90%, the number of components was limited to four when conducting PCA on energy, danceability, valence, loudness, instrumentalness, acousticness.

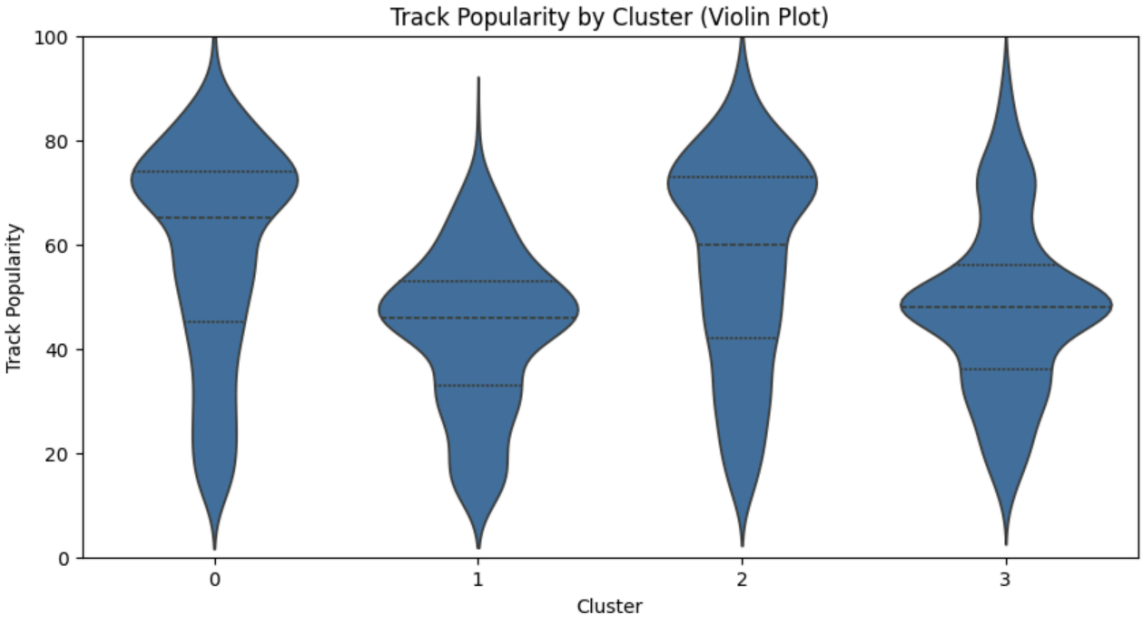


The high popularity songs seemed to occupy a common place in the PCA space, visible in both the two and three dimensional graphs above. Some of the songs with more average popularity scores formed two groups, one smaller and more dense than the other, while the least popular songs in the dataset seemed dispersed throughout the space. This suggests that PCA was effective in reducing the space into more understandable principal components, while preserving the variance in the data. It also suggests that clustering will be an effective next step, however low popularity score songs present throughout the space may present noise as an obstacle.

KMeans Clustering

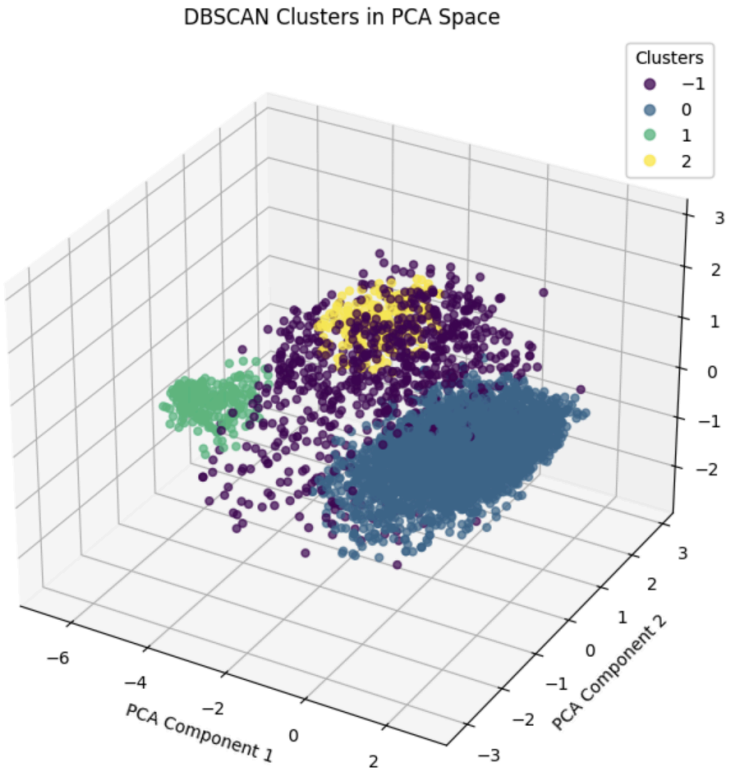
After PCA was done, investigation into the optimal number of clusters was done, using the 'Elbow Method'. Based on the resultant graph, 4 clusters were selected as the optimal option. This method of analysis provided the clusters which are displayed below in a three dimensional representation of the PCA space, with a silhouette score of 0.29 suggesting decent separation.


KMeans Clusters in PCA Space
Clusters
0
1
2
3
PCA Component 1
PCA Component 2


Track Popularity by Cluster (Violin Plot)
Track Popularity
Cluster

DBSCAN Clustering

This portion of the analysis was done to see if a different clustering method would better handle noise present in the dataset that could be watering down the presence of trends. This noise is due to lesser popular songs that occupy a similar place in the PCA space to higher popularity songs, because of their feature similarity, but lack popularity for some reason (possibly not due to audio features).

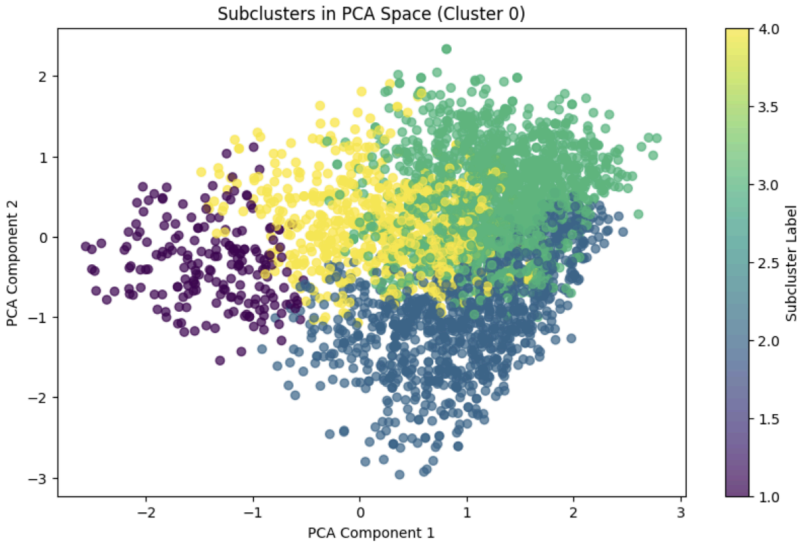


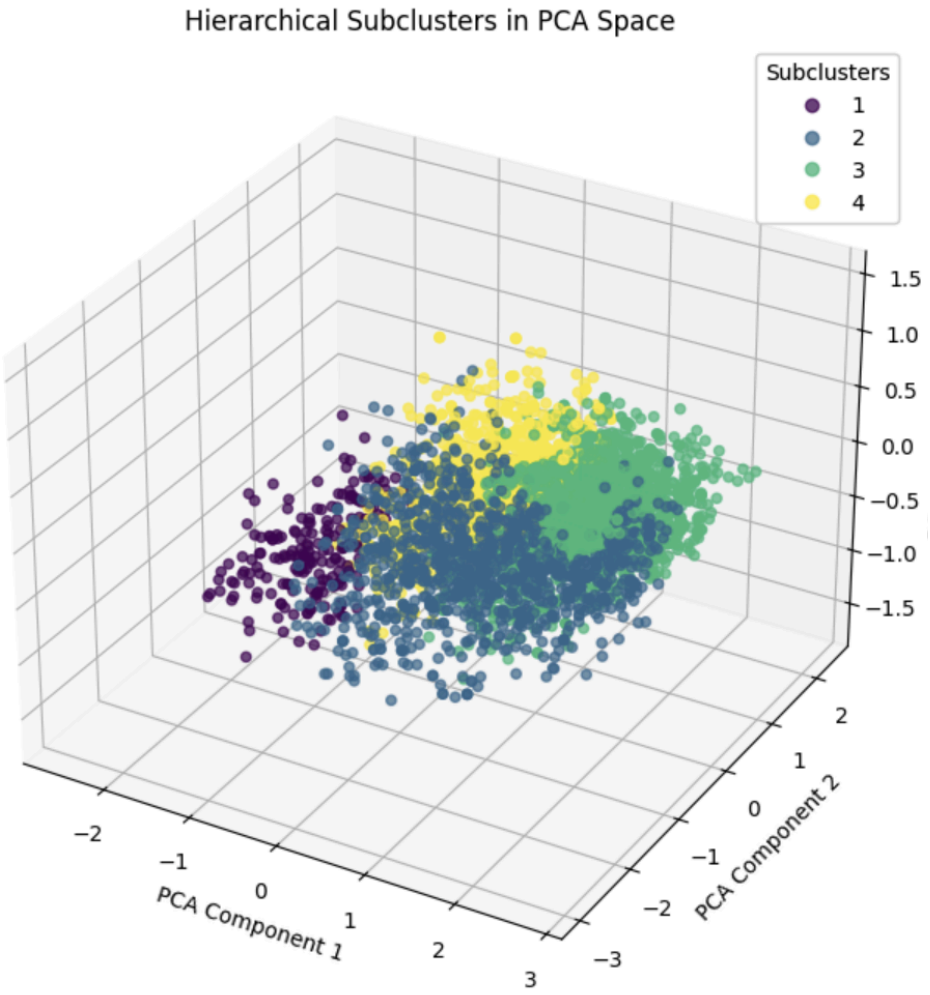
This clustering method netted 3 clusters, with a silhouette score of 0.50, suggesting this method was effective. This is supported by the fact that Cluster 0 exceeds the other two clusters and the noise, by 15 and 16, when looking at their median popularity scores. This means that this method was effective in filtering out many songs that are similar to each other, which are also excessively dissimilar to the average song and which likely suffer in popularity as a result of this. As for features, the most notable difference between Cluster 0 and the other data points, is that Cluster 0 is significantly louder. Its median loudness is -6 db, while next loudest is -11db, and because the decibel scale is logarithmic, this represents a significant difference in loudness. However, Cluster 0 still contains a majority of the data points in the overall cleaned dataset (73% specifically), suggesting that a deeper analysis on Cluster 0 could be insightful.

Hierarchical Clustering

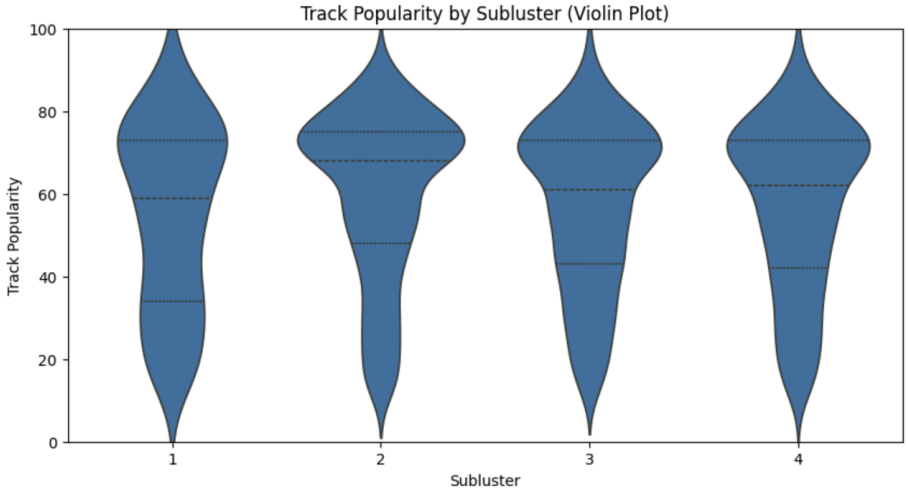
Hierarchical clustering was done on the large, but more popular than average, Cluster 0 from DBSCAN, in an effort to better narrow down this Cluster and the trends

within it. The result was 4 subclusters, visible in the PCA space in the 2D and 3D graphs below, with the 2D graph being a top-down view of the 3D graph.

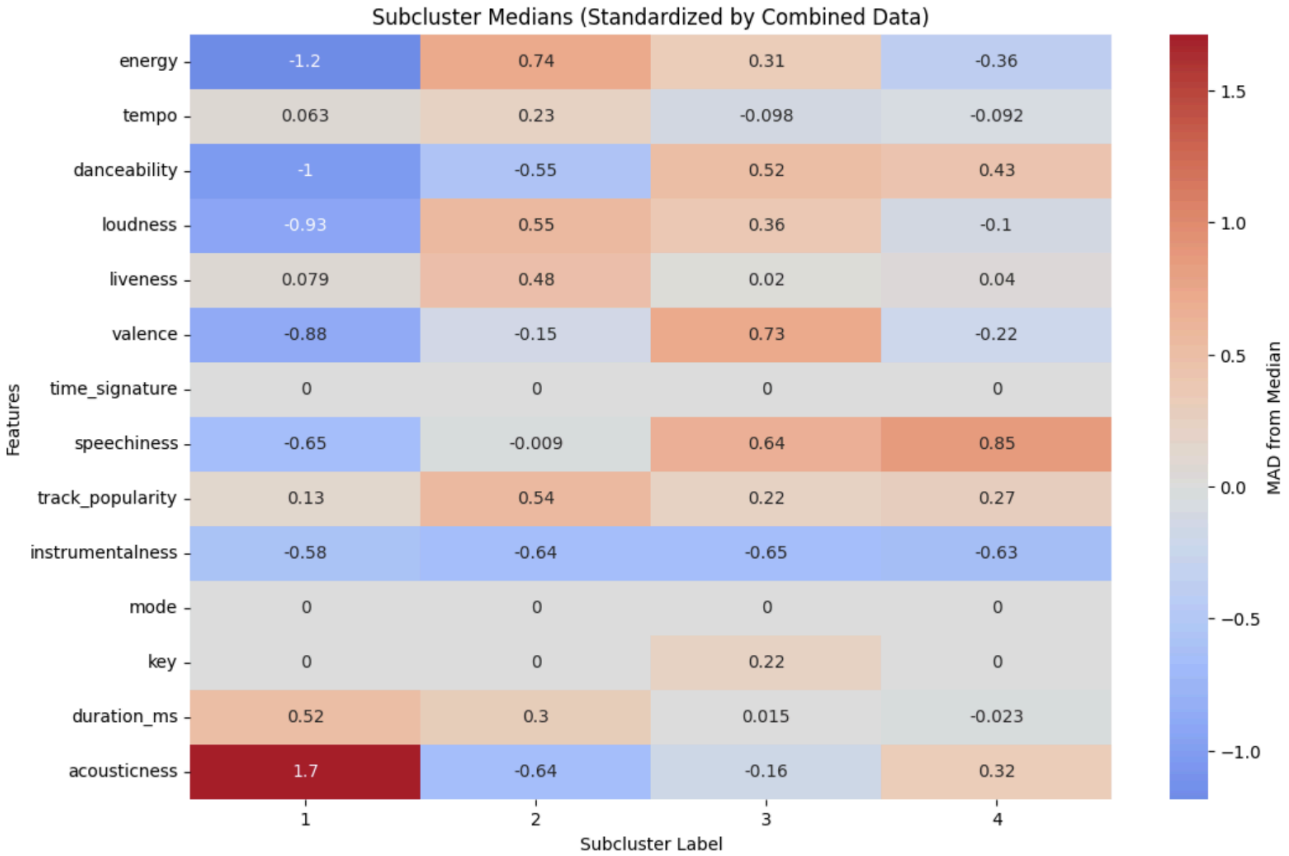




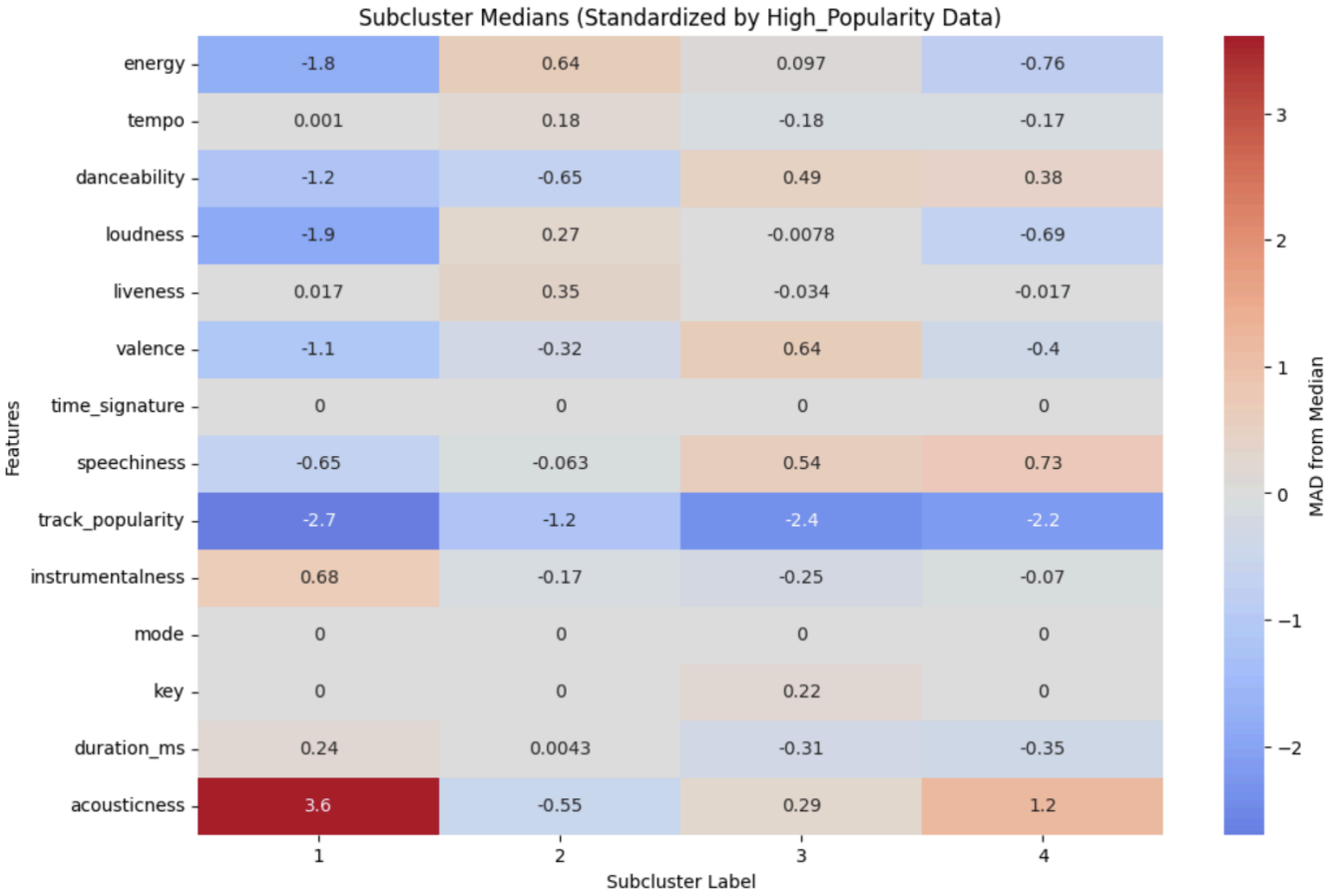
Investigating these subclusters further, I plotted their distributions of popularity scores in another violin plot for comparison.



Subcluster 2 has the distribution that is most focused in the upper range of popularity, with the most slim tail, suggesting that this subcluster represents a group of similar songs with features most conducive to a high track popularity. This is supported by the median track popularity of Subcluster 2 exceeding the other subclusters by 6, 7, and 9 (out of 100). This subcluster's track_popularity also exceeds the cluster it's in (Cluster 0) by 5, which itself already exceeded the other 2 clusters by 15. For a better understanding of how the features of these subclusters differ from the entire dataset, and from the 'high_popularity' subset, I compared the median values of each using Median Absolute Deviation (MAD). I then displayed the results of this comparison in two heatmaps to best visualize the intensity of these differences.



The first heatmap, comparing the subclusters to the overall dataset, shows that Subcluster 1 is defined most by its increased acousticness, and its lack in energy, danceability, and loudness. This is informative considering Subcluster 1 was the one with the least significant presence in the upper range of popularity, and had the most significant tail in the lower range of popularity. Based on this, acoustic music with less energy and danceability than the average song is not conducive to popularity, representing a niche that may be more difficult to grow in.

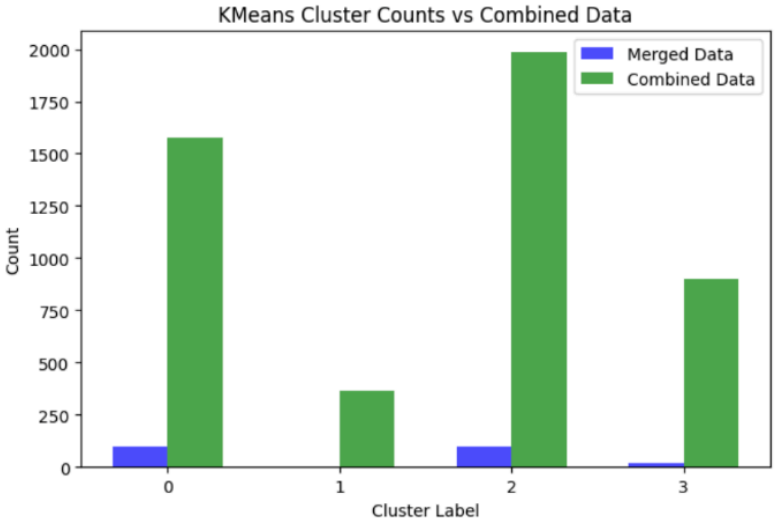


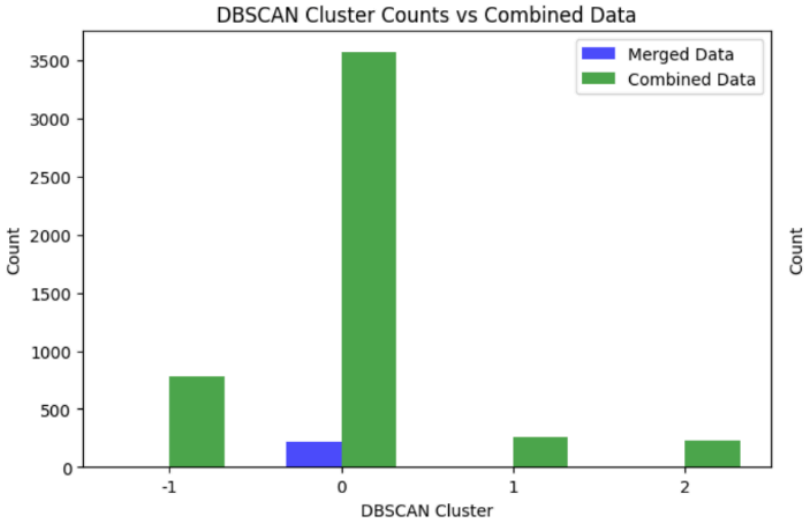
Based on the second heatmap, it is clear that Subcluster 2 is the most similar to the high popularity dataset, and also the closest subcluster in terms of popularity score specifically. This suggests that Subcluster 2 being louder, more energetic, and not very acoustic or instrumental, enables it to keep its popularity higher than the other subclusters. The features being more similar to the high_popularity dataset than the combined dataset leads me to believe that

Comparing the two heatmaps together, Subcluster 2 has features that are more similar to the high_popularity subset than the general dataset. This leads me to believe that my dimensionality reduction and clustering techniques were effective in identifying some of the effects that features can have in combination on a song's popularity score.

Further Validation of Results

Further validation of results was done by inner-joining the secondary dataset, 'Most Streamed Spotify Songs 2024' with the primary dataset and displaying the count of these songs in each cluster for each clustering technique. This essentially provides a random subset of popular songs that can be used to show how present each cluster is in the secondary dataset as a percentage (number of songs from 'Most Streamed Spotify Songs 2024' present in the cluster, over the total number of songs in the cluster). This can be seen in the double bar charts below, with the percentage presence of the clusters below each of them. Based on these evaluations, the KMeans and DBSCAN clustering techniques effectively separated the songs into less and more popular clusters, as KMeans clusters 0 and 2, and DBSCAN Cluster 0 are drastically more represented in the 'Most Streamed Spotify Songs 2024' dataset. The comparison (not shown) with the Hierarchical subclusters of DBSCAN Cluster 0 showed a much more even spread of the merged data, which shows that they likely represent genres, techniques, or feature combinations that are possible to find in popular songs, but difficult on average for songs to replicate at scale while maintaining popularity.





Discussion

I expected clusters to form around songs that particularly focused on one feature. For example, a cluster for very loud songs, a cluster for very danceable songs, a cluster for very acoustic songs, etc. This expected result is somewhat validated by the highly acoustic KMeans_Cluster 1, as well as the loud DBSCAN_Cluster 0 and its acoustic Subcluster 1. However ,many of the clusters represented songs with feature values that are generally average. Interestingly, danceability was not as important as I had expected. 'Energy' and 'liveness' being some of the defining features of the less

danceable Subcluster 2, shows that they may be able to make up for the lack of danceability. As I expected, and noted in the results, high loudness and low ‘acousticness’ and ‘instrumentalness’ were consistently shown to be important across the clusters for someone wanting to obtain higher popularity scores. The strength of this trend, however, may be explained by some less popular genres being mainly acoustic and instrumental, rather than a general statement on the ability of an acoustic or instrumental song becoming popular. It is important to consider the existence of external factors (e.g., marketing, artist popularity, playlist placements) not captured in the dataset, which play a significant role in determining song popularity. Due to this, the presence of more popular and less popular clusters in the PCA space of the audio features, suggests value in these findings even when the difference in the median audio feature values would otherwise be insignificant. Taking this into consideration, I have identified several trends present in the dataset that an artist should consider when evaluating the performance of their own songs.