Cluster Analysis of Spotify Top Hits of 2018

STAT 437 - Project 2

Becky Catlett, Harley Clifton, Natasha Gesker, & Eliot Liucci

2023-12-01

Introduction

The Dataset

The dataset being explored in this project is sourced from Kaggle.com and is called "Top Spotify Tracks of 2018." This dataset includes numerous quantitative variables as well as some qualitative variables, such as the song title, artist, and the Spotify url ID. The quantitative variables are slightly less intuitive, such as danceability, energy, key, loudness (in dB), mode, speechiness, and acousticness. Most of these quantitative variables range from 0 to 1, where 1 is the highest measure and 0 indicates a low/null value. Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. Speechiness is how often words are present in the song, and mode indicates the modality, which is whether the song is written in major (1) or minor (0). Acousticness is how acoustic the song is, versus a non-acoustic song which would be composed of lots of electronics. Energy is essentially describing the tempo of the song, whether is song is slower and sleepier or more upbeat.

Methodology and Statistical Approach

We will be using various methods of clustering analysis to find an model structure of the data that is more easily interpreted. The goal of this is to gain an understanding of music popularity in 2018 using Spotify data. We will be using PCA to reduce the number of dimensions purely for visualization purposes, since there are 16 different variables and only 100 observations. From there, we will explore k-means, hierarchical, and model-based clustering to analyze the data and see what relationships occur between variables that may not be obvious without the application of clustering methods.

Below is code for three different clustering techniques, from which we will chose one to use for our analysis of the Spotify data. Including all three methods ensures that the method we choose for our analysis is one that best maximizes the accuracy of the analysis. Although all statistical results have uncertainty, the goal is to use the best statistical method to get rid of as much uncertainty as possible in the result. Clustering/classification results are useful for artists, songwriters, producers, and many more individuals that are looking for qualities of a hit song.

This data set contained no missing data. It did, however, have a lot of variables, meaning when trying to visualize the data using PCA made the most sense. This plot will be used when exploring cluster methods.

Data Cleaning & Exploratory Data Analysis

This analysis will focus on only 6 variables from the dataset, specifically the quantitative variables. Danceability, energy, key, loudness, mode, speechiness, acousticness will be used in each of the cluster methods, however PCA will be used to visualize the data. While it is possible to visualize the data without PCA, it would result in an excessive number of plots. As seen in Figure 1, PCA allows one to see a version of the numeric data in one plot. Cluster methods can be easily assessed and compared to each other.

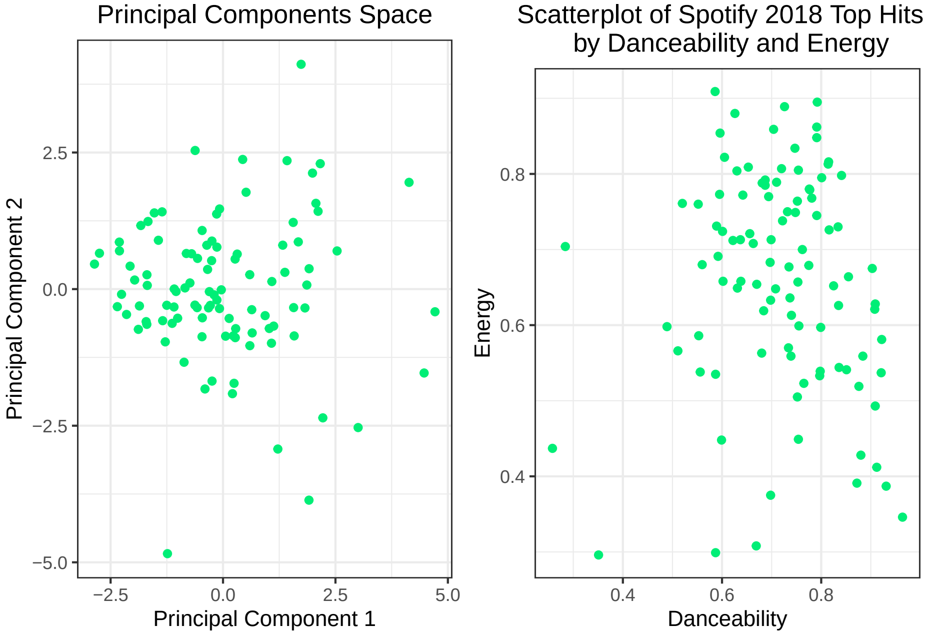

Figure 1: Left: Scatterplot of the Spotify Data (before any clustering) in the Principal Component Space. Right: Spotify 2018 Data plotted by Danceability and Energy.

In Figure 2, created using the `psych` package in `R`, we can see variable pairs that are notably correlated with each other. Energy and Loudness have a moderately strong, positive correlation (r = 0.73). All other pairs of numeric variables in this dataset have weak or non-existent correlations.

Cluster Analysis

To analyze this data, three clustering methods will be explored: Agglomerative Hierarchical, K-Means, and Model-Based.

Agglomerative Heirarchical Clustering

For the hierarchical clustering method, clusters are based on a distance matrix. The *single* method connects the points based on the minimum distance of the clusters. The strange shape of the dendrogram pictured below is due to the Single Linkage which results in "chaining". Chaining occurs because the clusters just grow from a single cluster, instead of new clusters being formed. This can be problematic because we only need a single pair of points to be relatively close for clusters to merge. Therefore, clusters can be too spread out and not compact enough. The *average* method, which connects the points based on average distances of the clusters, has the same phenomenon occurring in this case.

The *complete* method, which connects the points based on the maximum distance of the clusters, does not have any "chaining" issues. When determining what height to cut off the groups, a height 7 was chosen since that is the first time we start to see a lot of tiny groups begin forming into larger groups.

When visualizing these clusters in the principal component space, it seems that cluster 3 starts to take over most of the points. When comparing it to cluster 1, it is not entirely clear why a point should be in one group compared to another. There are also a lot of little groups that are in the outskirts of the plot. While

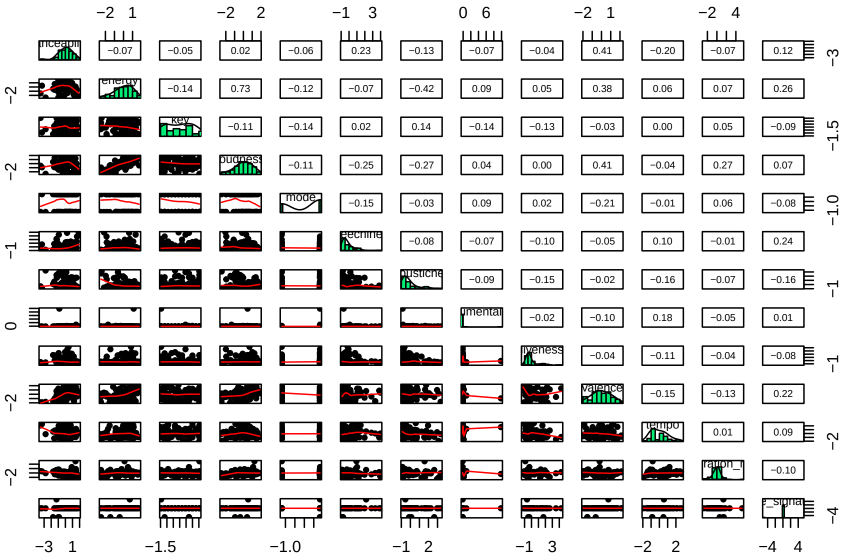

Figure 2: Scatterplot Matix of the Quantitative Variables in the Spotify 2018 Dataset.

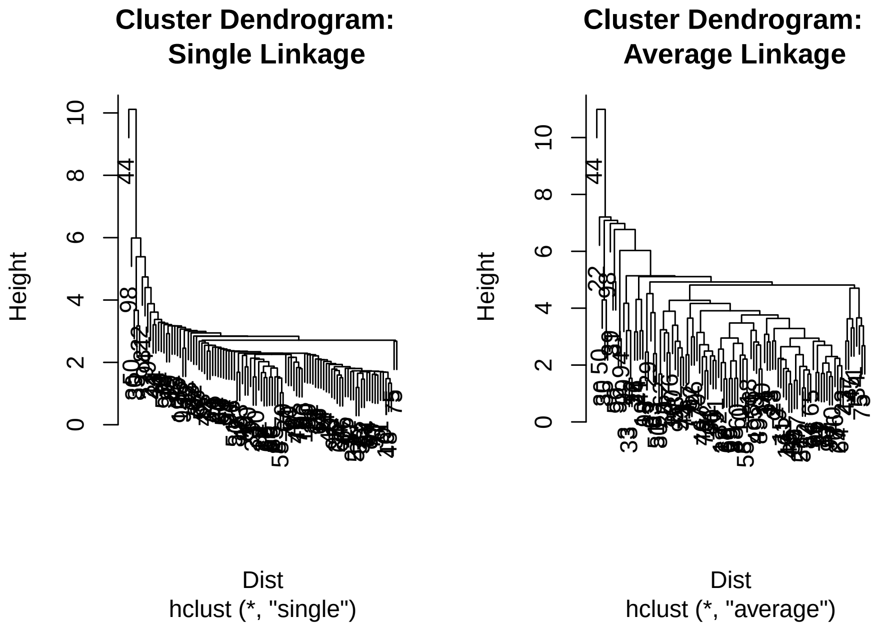

Figure 3: Dendrograms for the Single and Average Linking Methods for Hierarchical Clustering.

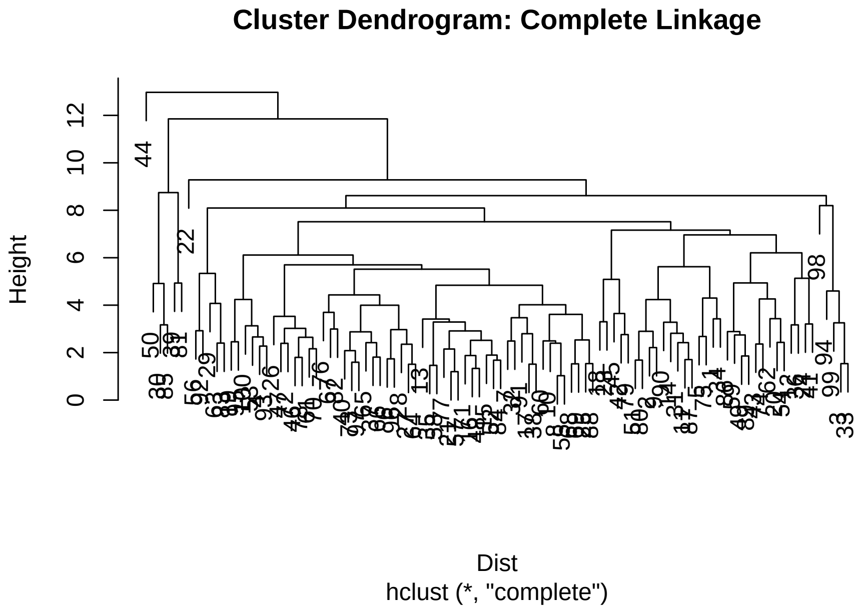

Figure 4: Dendrogram for the Complete Linking Method for Hierarchical Clustering.

you could argue that making less clusters by increasing the height, if the clustering logic already appears to be faulty at this point it may be detrimental to the process by making too few clusters.

K-Means Clustering

K-means clustering starts with choosing a number of clusters before performing the clustering analysis. To decide how many clusters would be ideal, a Within-Group Sum of Squares (WGSS) plot was created using the original data. Based on this plot, one could contend for many appropriate numbers of clusters due to varying perspectives of the cutoff. For the purposes of this analysis, 15 clusters will be created. This number was chosen because having too many groups gets overly confusing, especially considering there are only 100 data points. While 15 is still quite a large number of clusters, it yields a WGSS of under 500, which compared to 1 is much better.

The ‘kmeans’ function was used to assign each original data point to one of the 15 groups. When visualizing these clusters in the PC Plot, one can see from a single glance that these groups do not make a lot of sense for this data set. While groups tend to have clear centers and overall consistent variability within groups, this method is not ideal for interpreting what these results mean for this data set.

For K-Means clustering, there is the ability to create some additional visualizations to assess the chosen number of clusters.

In the additional visualization, the line thickness is representative of the strength of the relationship between clusters. If the line is thin, it means it is far to the next cluster; if it is thick, it is close to the next cluster.

The stripe plot is used to visualize shadow variables by themselves. In the plot above, the stripes represent just the second closest cluster. Since they are pretty far apart within each cluster, we can be confident that they should be separate clusters and do not need to be further merged.

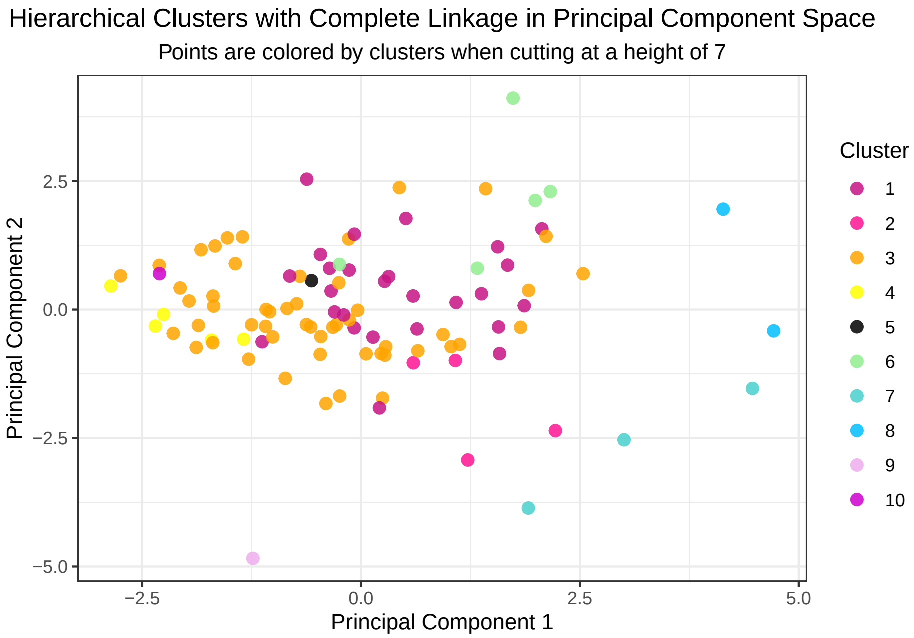

Figure 5: Scatterplot of Clusters from Agglomerative Heirarchical Clustering with Complete Linkage in the Principal Component Space.

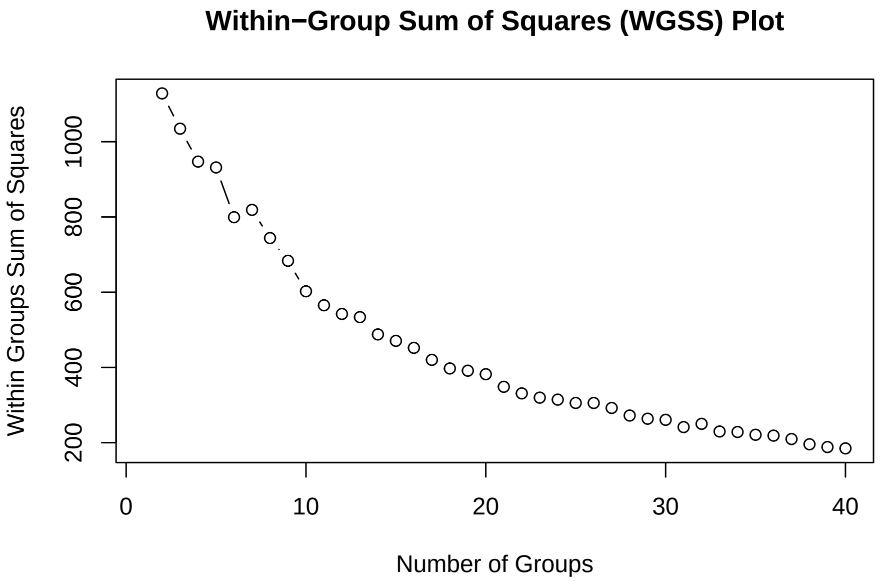

Figure 6: Within-Group Sums of Square vs. Number of Groups Plot. Used to determine the optimal number of clusters with K-Means Clustering.

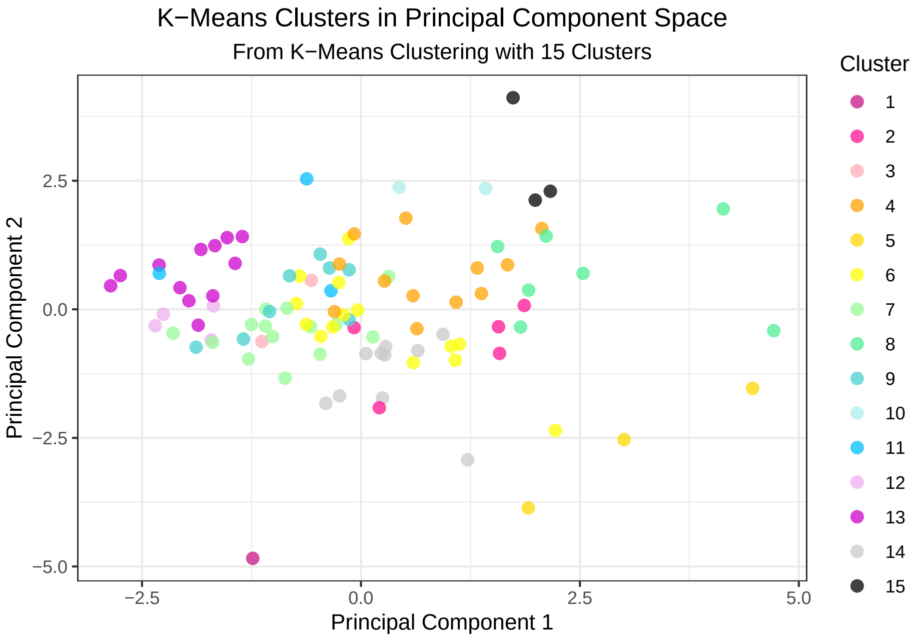

Figure 7: Scatterplot of Clusters from K-Means Clustering with Complete Linkage in the Principal Component Space.

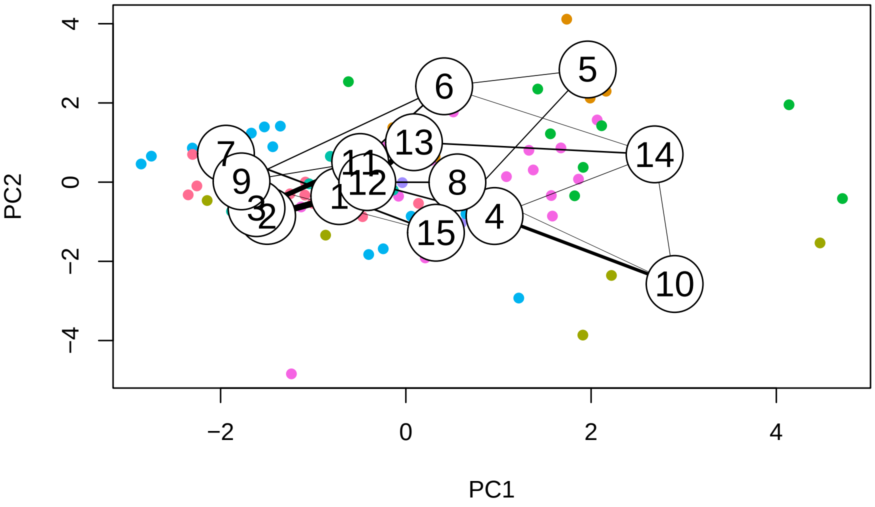

Figure 8: Additional Visualization where the Line Thickness represents the Nearness of the two Connected Clusters.

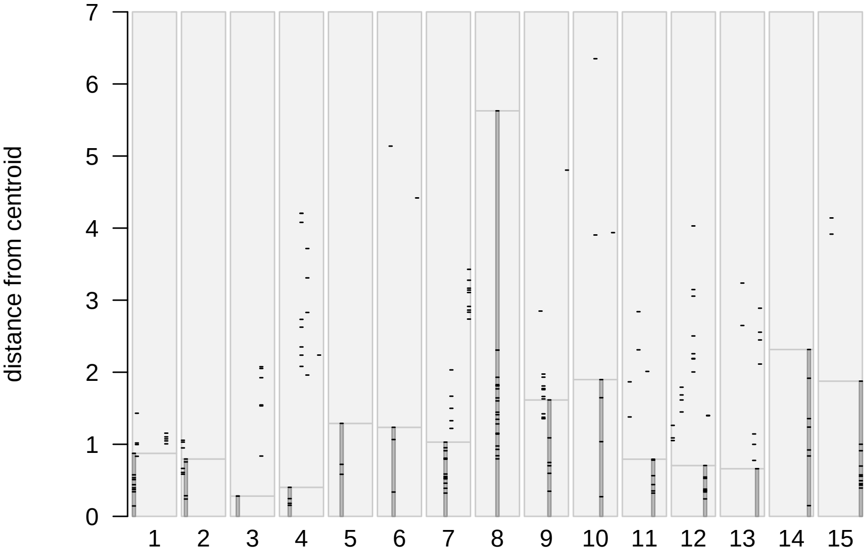

Figure 9: Stripe Plot of the Second Closest cluster for the K-Means Results.

Model-Based Approach

The model based clustering method assumes distributions for both the population and the subpopulation. While the probability distribution functions are defined for these distributions, they will not be discussed in detail in this paper.

The plot above uses Bayesian Information Criterion (BIC), which penalizes models for additional components, to determine which model-based clustering method is best. Here, we want to *maximize* the BIC. Based on the plot, the algorithm selected the "EEV" (ellipsoidal, equal volume and equal shape) model with *5* clusters as the superior model.

Based on the clusters, there is one point that stands out. Observation 44 is singled out in its own cluster; this could either indicate that this observation is an outlier or that this clustering method is faulty.

When visualizing these clusters in the PC Plot, one can see that the point in cluster 5 does appear to be isolated compared to the other clusters. The point in cluster 5 has a lower value for PC2 than the other points. While there is some overlapping from cluster 1 across the other points, in this plot is easier to see the trends of the various clusters, especially cluster 2 and and 4.

Results

While one can look at the visualizations of these cluster methods and make gut decisions about the quality of each method for this data set, we also want to look at the sums of squares of each method.

The hierarchical method has a sum of squares that is almost twice as large as the other two methods. While the model-based method does have a lower sum of squares than the k-means method, they do seem to be fairly close. Meaning that if researches felt that the clusters formed by the k-means method were more useful for their analysis, they would not be at a serious disadvantage for using k-means over model-based. However, based these values and the visualizations of the different clustering methods, we choose the model-based method to be the best for this data set, provided that there was no specific model goal for forming these clusters.

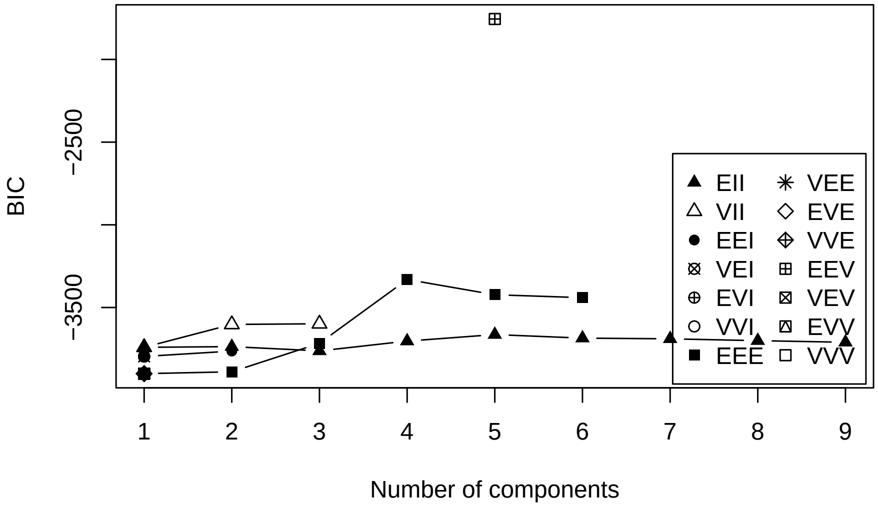

Figure 10: Model Comparisons done behind the scenes by the 'Mclust' function in R. Models are ranked by BIC.

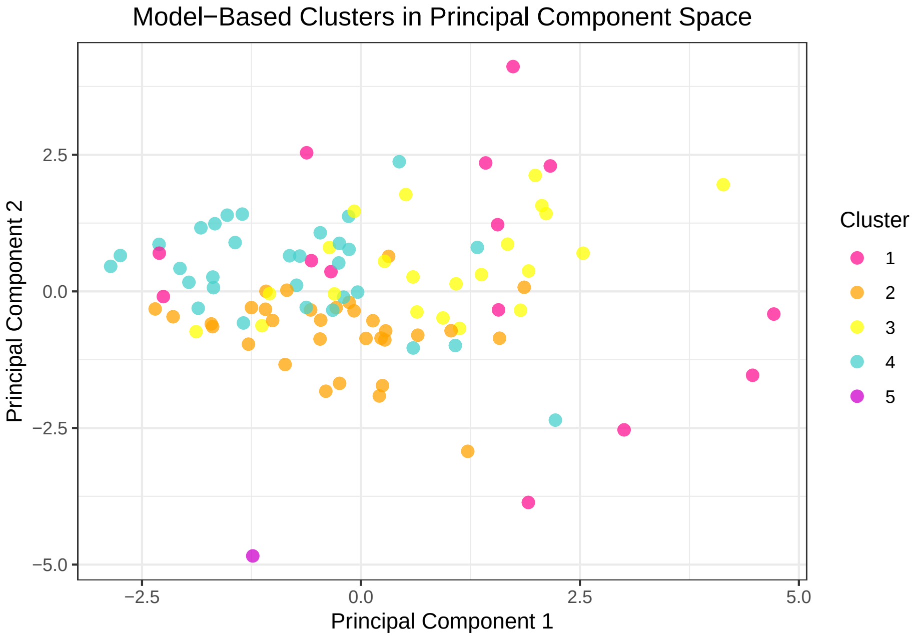

Figure 11: Scatterplot of Clusters from Model-Based Clustering with Complete Linkage in the Principal Component Space.

Algorithm	Sum of Squares
Hierarchical Clustering	27.47836
K-Means Clustering	14.39514
Model-Based Clustering	13.25232

Summary

Our analysis of the Spotify 2018 Top Songs data revealed that the clustering method that yielded the lowest sum of squares value was the model-based clustering method. We used principal component analysis on the data and for the graphs since there were 16 variables and only 100 observations. Additionally, we then explored hierarchical, k-means, and model-based clustering methods. The hierarchical analysis yielded the highest sum of squares. Similarly to the k-means graph, the hierarchical scatterplot of the clusters against the principal components has dispersed clusters and patterns that are not extremely obvious. The model-based clustering had only 5 clusters, and the pattern of the clusters appears to be much more logical. These five clusters help explain patterns in the Spotify data and provides information about the data that could not be seen in graphs such as the scatterplot of danceability vs energy. This analysis only touches on the complexity of clustering methods, but we did discover that for this dataset, model-based clustering is going to be the effective clustering choice for beneficial statistical analysis.

Citations

Tamer, N. (2018). *Top Spotify Tracks of 2018* (Version 1)[Data set]. Kaggle. https://www.kaggle.com/datasets/nadintamer/top-spotify-tracks-of-2018