Spotify Genre Analysis

Will Clatanoff, MJ Corey

[redacted]

Professor Li
May 12, 2025

1 Introduction

Spotify is the world's most popular music streaming service, with an estimated 650 million users and over 100 million songs available to all users [Leu25]. This project will look at the publicly available Spotify Dataset from Kaggle [Pan22]. This dataset contains 114,000 rows, which represent individual songs, and 15 columns, which correspond to different variables. There are 114 genres, with 1,000 songs in each genre. Spotify has provided labels for each song, including objective variables such as duration, popularity, tempo, the song's key, music genre, and even whether or not the song is explicit. In addition, the dataset also features subjective variables labeled by Spotify through some unknown algorithmic means, such as danceability, energy, acousticness, loudness, valence, and other more niche variables. This dataset provides a very high-dimensional idea of identifying songs and genres through quantitative information, making it the perfect dataset to analyze.

This dataset is particularly interesting in terms of exploratory data analysis. Through statistical machine-learning techniques, it is possible to understand what makes songs and genres similar to one another. Then, it is possible to make recommendations to a person who likes a specific genre based on genres that are quantitatively similar. Partitioning the data into certain groupings would achieve this goal, and the clustering algorithms outlined in this course allow this partitioning to be possible.

2 Methodology

2.1 Data Preprocessing

From analyzing the data, we concluded that dimensionality reduction would be needed to provide meaningful and efficient analysis. As seen in Figure 1, each variable is differently distributed across different units and scales. To ensure accurate analysis, we normalized the data because of the variation of units; we did not want to favor any variable only because it was measured on a greater scale. We then decided to omit the four categorical variables: key, mode, time signature, and explicit. This left us with 11 numerical variables to continue with our analysis. Also, since the dataset contains 114,000 songs, we decided to focus our analysis on the 114 genres. We took the average value of each variable for all the songs within each genre to give us one representative row per genre. Finally, we found no missing values present that we could have imputed, since missing or unknown values were assigned values of zero, and there was no way to tell if the variable was a true zero or a missing value. With these necessary changes to the data, we were ready to continue with our analysis.

2.2 Dimension Reduction

To lessen the impact of the high-dimensional data, we applied Principal Component Analysis (PCA) to the genre-level dataset. After applying PCA to our matrix of 114 genres by 11 numerical variables, we found that the first five principal components captured about 70% of the total variance (Table 2). This reduction retained most of the underlying structure of the data while making the data simpler to use in the clustering process.

2.3 PCA Scatter Plot

After computing PCA, we visualized the first two principal components using a scatter plot, with each point representing an individual song and colored by its associated genre (Figure 2). The goal was to assess whether genres naturally cluster in a lower-dimensional space. Also, we aimed to visually interpret how genres are distributed across the feature space.

2.4 Hierarchical Clustering

To improve interpretation and reduce visual clutter, we applied hierarchical clustering using complete linkage to the genre-level dataset. This ensures that genres within the same cluster are grouped and exhibit consistently similar audio features across all variables. To form the clusters, we decided to cut the dendrogram (Figure 3) at a height of seven, which gave us seven genre clusters.

2.5 Cluster Shapes on PC Plots

With the clusters derived from hierarchical clustering, we looked to visualize the data on the reduced PC axes to better understand their spatial separation. We plotted the seven genre clusters across all pairwise combinations of the first five PCs, resulting in ten unique, two-dimensional plots (Figure 4 and Appendix B).

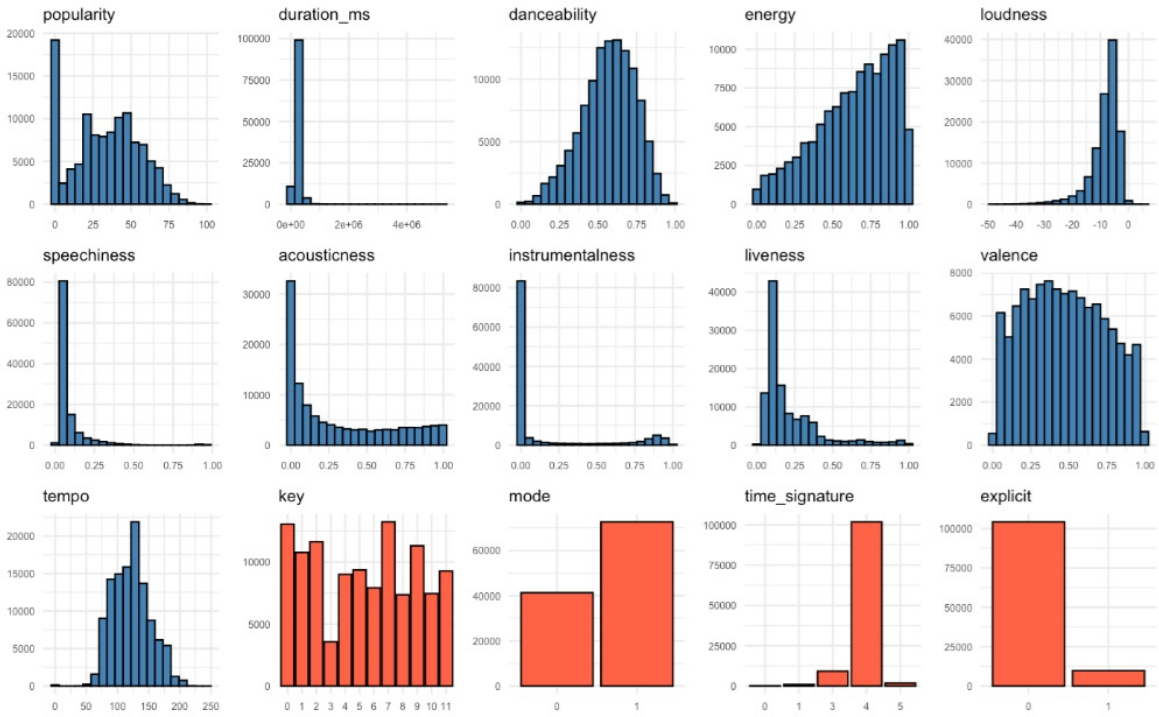


Figure 1: Distribution of data across the 15 variables.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
popularity	0.054	-0.029	-0.142	0.871	0.266
duration_ms	-0.024	0.499	-0.189	-0.157	0.559
danceability	0.419	-0.597	-0.228	-0.226	0.302
energy	0.861	0.327	0.048	-0.070	0.029
loudness	0.876	0.096	-0.048	0.078	0.010
speechiness	0.170	-0.109	0.688	-0.165	0.144
acousticness	-0.743	-0.390	0.192	0.082	-0.089
instrumentalness	-0.482	0.399	-0.146	-0.353	0.063
liveness	0.144	0.179	0.765	0.119	0.111
valence	0.501	-0.600	-0.061	-0.165	-0.051
tempo	0.320	0.297	-0.077	0.065	-0.662

Table 1: Table of the first five principal components.

comp 1	comp 2	comp 3	comp 4	comp 5	comp 6	comp 7	comp 8	comp 9	comp 10	comp 11
26.14	13.86	11.25	9.54	8.74	7.85	7.56	6.70	4.12	2.96	1.27

Table 2: Percent covered variance per principal component.

3 Results

3.1 Reduced Dimension Interpretation

One of the hardest challenges with using Principal Component analysis is interpreting what each dimension represents as a combination of the original variables. While each variable inherently contributes to each principal component to some degree, for this analysis and interpretation, we will consider only variables that offer more than a 0.3 coefficient to the PC loading to be a significant contributor to the dimension.

- PC1: High Danceability, Energy, Loudness, and low Acousticness and Instrumentalness. This dimension helps differentiate songs that you may hear played on the radio from classical or orchestral pieces.
- PC2: High Duration, Energy, Instrumentalness, and Tempo but lower Danceability, Acousticness, and Valence. This dimension separates dramatic pieces that you may find in an upbeat movie soundtrack from more acoustic pop songs.
- PC3: High Liveness and Speechiness. This PC serves to separate speechy tracks, like comedy routines or words of affirmation for sleep, from tracks with music in them.
- PC4: High Popularity and Low Instrumentalness. This dimension exhibits the highest emphasis on the popularity of the song. This dimension would help differentiate popular songs and artists from other songs in the same genre.
- PC5: High duration and Danceability, but very low tempo. This dimension offers insight into the duality of short and fast songs versus longer and intricate songs. Movements in classical pieces would sit opposite to edm songs along this axis.

3.2 Initial Visualization

After defining and interpreting our reduced dimensions, we wanted to visualize the loadings of our first two dimensions at a song level, which is shown in Figure 2. This, however, proved to be unproductive at the song level for a few reasons. The first reason that visualizing this was difficult was just due to the sheer number of data points present; with 114,000 data points and 114 defining labels, the graph becomes very uninterpretable quickly. This is further exacerbated by the fact that the first two principal components only capture around 40% of the total variance within the data. This makes visualization on two dimensions much harder, considering there is not very much variance to be captured with that number of dimensions.

3.3 Hierarchical Clustering to Improve Visualization

To further understand the differences between genres within the defined principal components, we wanted to make the data within the graphs clearer by reducing the number of labels. To achieve this, we employed hierarchical clustering on the average variable values at the genre level to form a similarity matrix. We opted to use complete linkage for our hierarchical clustering to ensure that genres were truly similar to one another when reducing the number of labels. Using this technique produced the following dendrogram in Figure 3.

After analyzing the dendrogram, we chose a height such that there were seven remaining clusters. This number is mostly arbitrary and can be increased to capture more nuance within the larger clusters, but to increase the clarity of the visualizations, we opted for the smallest number that still captured odd subcategories of genres. The full list of what genres belong in each of the seven reduced clusters can be found in Appendix A.

3.4 Final Visualizations and Commentary

With the reduced number of clusters, we began to visualize the differences in the genres again. Our final visualization once again lies on the PC axes, but to understand the range of genres, the various song data points are replaced with opaque ellipses that cover the entirety of the seven aforementioned hierarchical clusters. The resulting graph in Figure 4 shows the new structure on the first two principal components.

This visualization is much better at showing the difference between the genre categories. While the first and largest cluster centralizes around the origin of the graph, the more exotic subgenres start to show their differences, such as classical, piano, and opera, showing a lower PC1 value. This approach makes it much easier to understand the intricacies between the different PCs and where genres differ from each other. Since the PCA calls for 5 principal components, there are 10 total resulting graphs to attempt to visualize the graphs in 5 5-dimensional space. The remaining 9 graphs can be found in Appendix B to further illustrate the difference between genres.

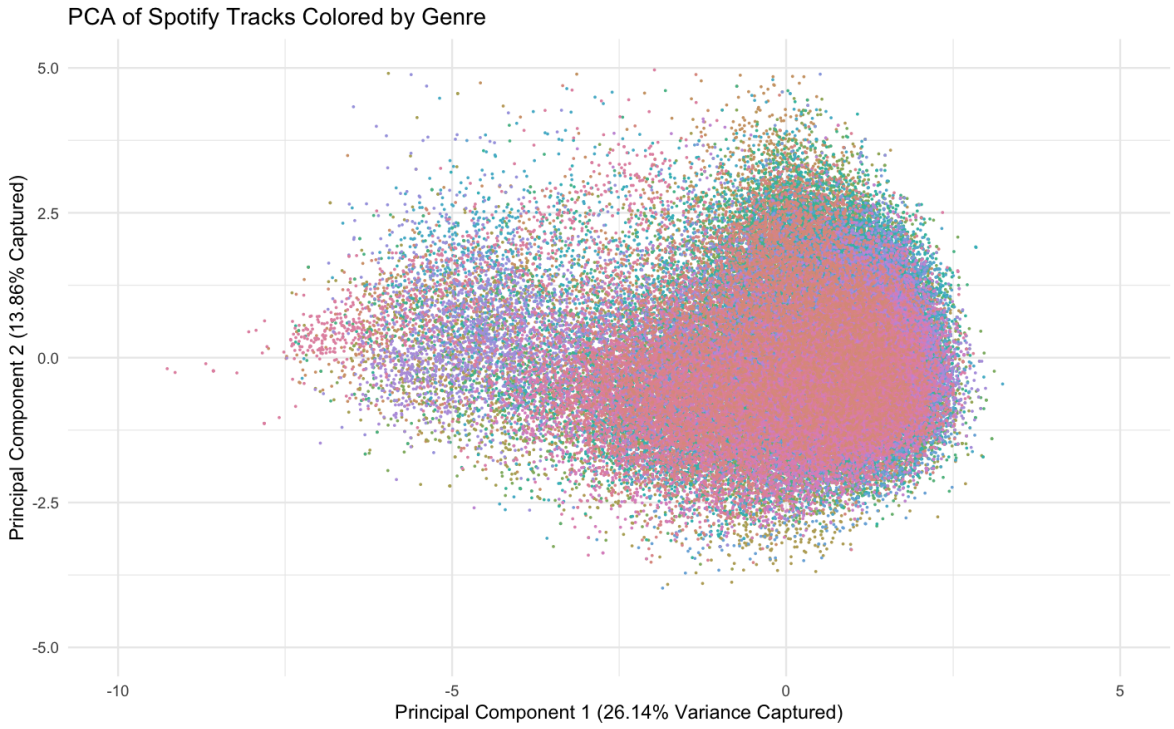


Figure 2: Song data mapped onto the first two principal components.

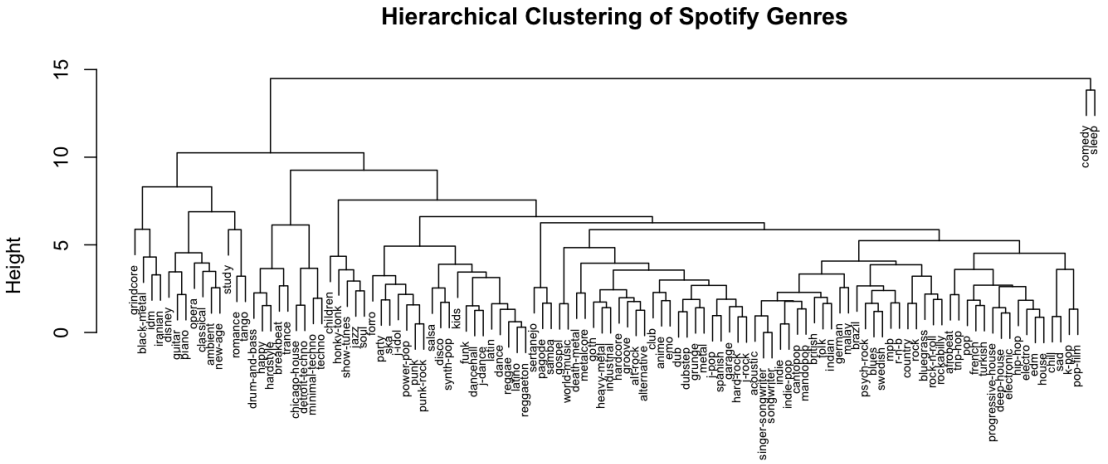


Figure 3: Dendrogram of hierarchical clustering using complete linkage on genre average variables.

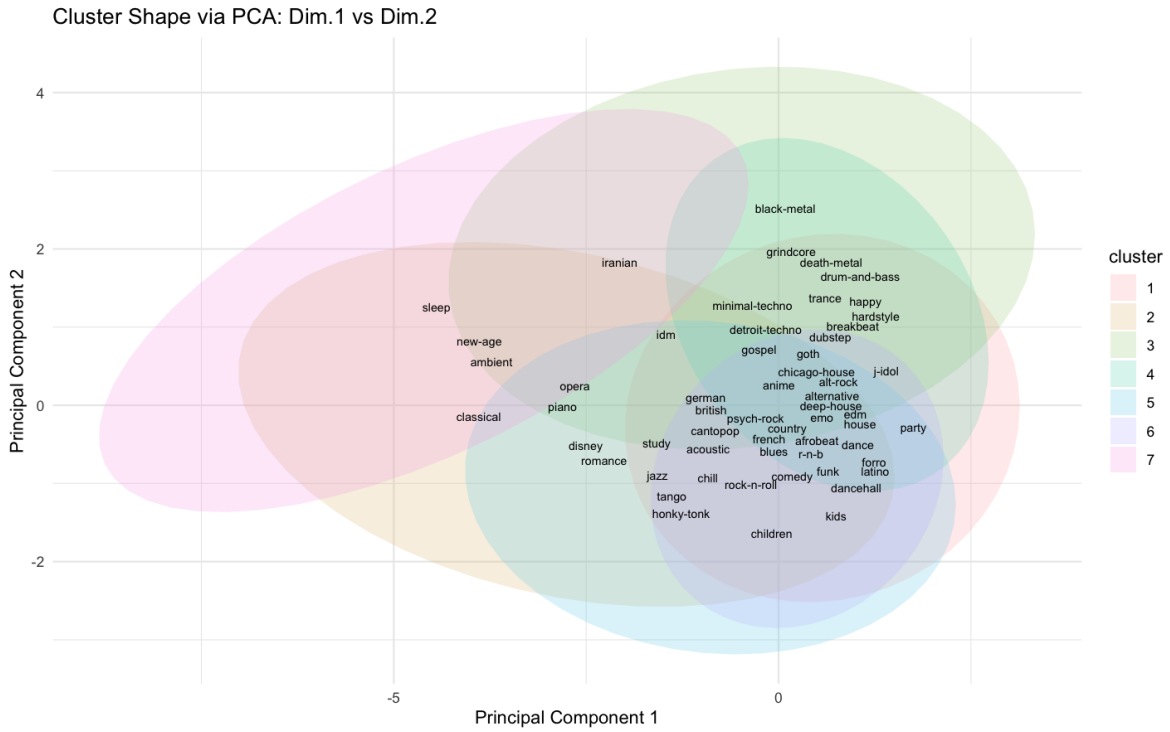


Figure 4: Graph of hierarchical clusters complete range across PC1 vs PC2

4 Conclusion

Overall, the experiment yielded success in the hypothesis that Spotify genres can be grouped and their differences can be analyzed. However, there is further research that can be taken into this study that was out of scope for the time of this project. The next step would be incorporating the categorical variables into the dimension reduction. While from an initial viewpoint, these variables don't seem to influence the differences in genres, there is still some potential variance being given up by excluding them initially, such as blues being predominantly minor key or waltzes having the characteristic of being in 3/4 time. Another step that needs to be taken for this study is to look specifically at the genres within cluster 1 further. A large number of genres ended up in this cluster, still exhibit differences from an intuitive view of the domain. It would be very interesting to redo this study on only the genres in cluster one to reduce the influence of heavy outliers like sleep and comedy. Overall, the claim that there are differences in the genres on Spotify holds, and this approach shows some of those differences.

A Genre Clusters from PCA Analysis

The following clusters were obtained using PCA and subsequent clustering on Spotify genre embeddings.

Cluster 1

- acoustic, afrobeat, alt-rock, alternative, anime, bluegrass, blues, brazil, british, cantopop
- chill, club, country, dance, dancehall, death-metal, deep-house, disco, dub, dubstep
- edm, electro, electronic, emo, folk, forro, french, funk, garage, german
- gospel, goth, groove, grunge, hard-rock, hardcore, heavy-metal, hip-hop, house, indian
- indie, indie-pop, industrial, j-dance, j-idol, j-pop, j-rock, k-pop, kids, latin
- latino, malay, mandopop, metal, metalcore, mpb, pagode, party, pop, pop-film
- power-pop, progressive-house, psych-rock, punk, punk-rock, r-n-b, reggae, reggaeton, rock, rock-n-roll
- rockabilly, sad, salsa, samba, sertanejo, singer-songwriter, ska, songwriter, spanish, swedish
- synth-pop, trip-hop, turkish, world-music

Cluster 2

- ambient, classical, disney, guitar, new-age, opera, piano, romance, study, tango

Cluster 3

- black-metal, grindcore, idm, iranian

Cluster 4

- breakbeat, chicago-house, detroit-techno, drum-and-bass, happy, hardstyle, minimal-techno, techno, trance

Cluster 5

- children, honky-tonk, jazz, show-tunes, soul

Cluster 6

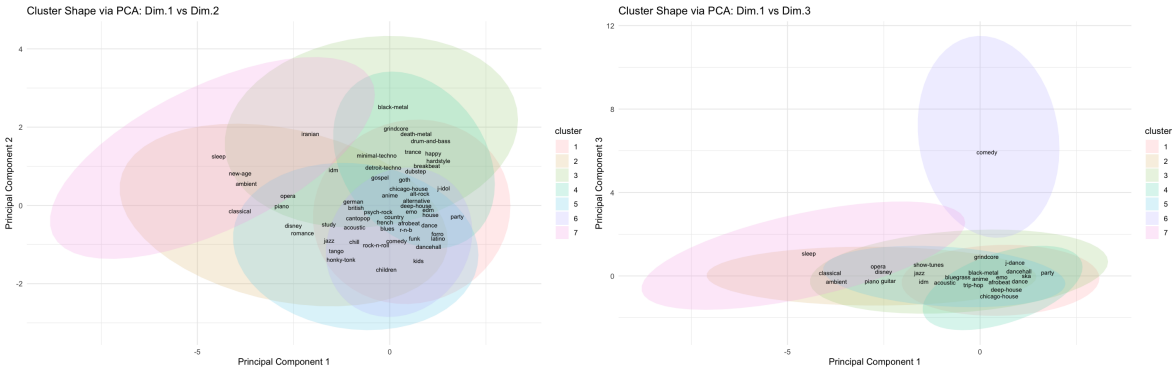
- comedy

Cluster 7

- sleep

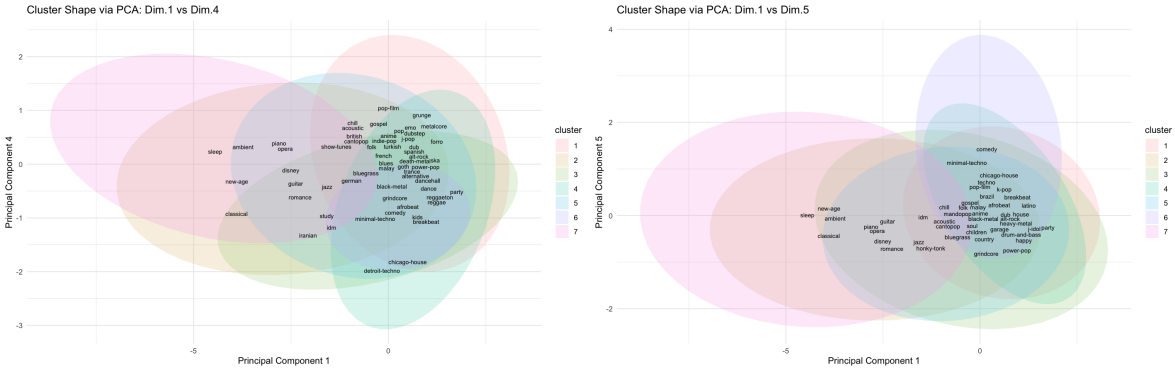
B Principal Component Combination Graphs

Below are the scatter plots of all 10 pairwise combinations of the first 5 principal components (PC1–PC5), labeled according to the form `XvY.png`, representing PCX vs. PCY.



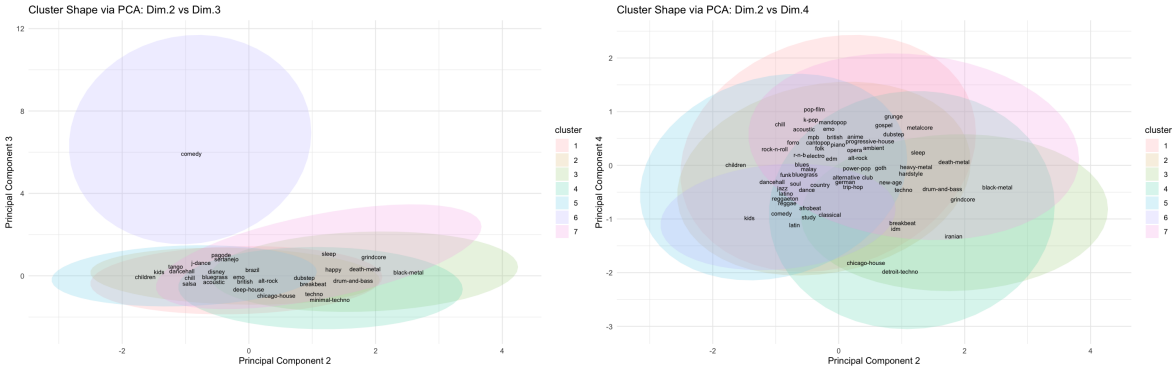
PC1 vs PC2

PC1 vs PC3



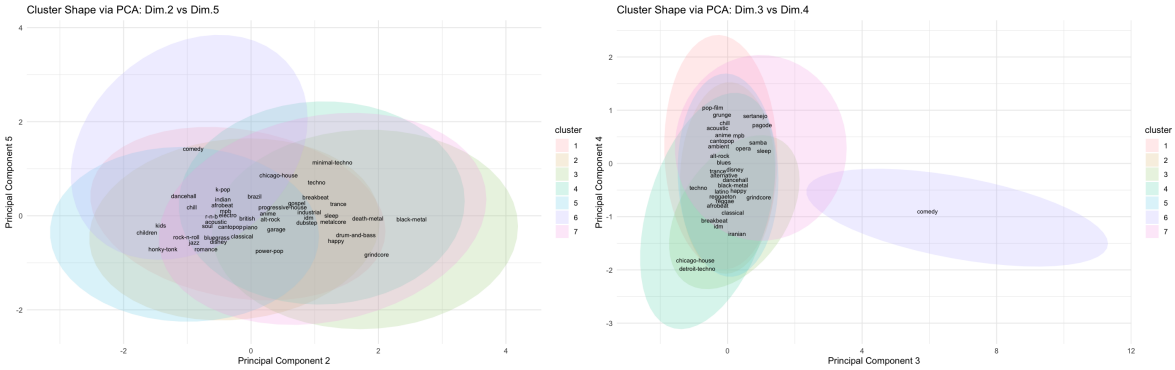
PC1 vs PC4

PC1 vs PC5



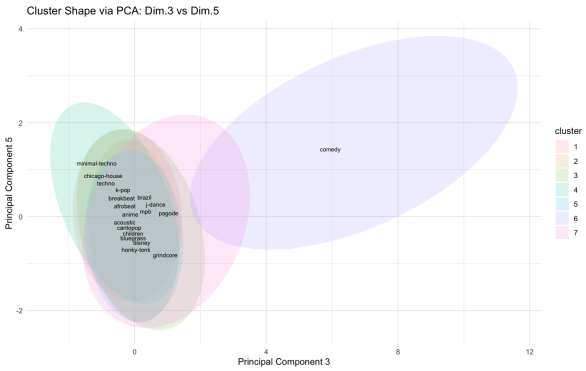
PC2 vs PC3

PC2 vs PC4

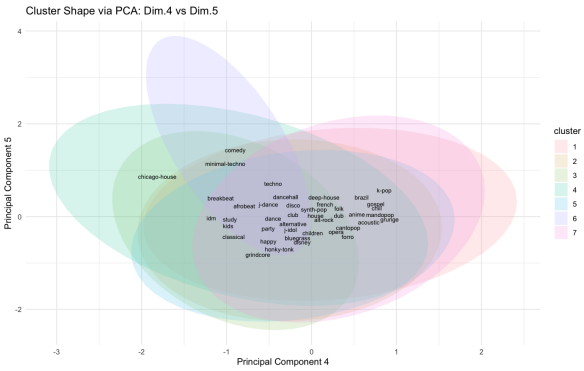


PC2 vs PC5

PC3 vs PC4



PC3 vs PC5



PC4 vs PC5

References

[Leu25] Patrick Leu. Number of spotify monthly active users (maus) worldwide from 1st quarter 2015 to 4th quarter 2024. *Statista*, 2025.

[Pan22] Maharshi Pandya. Spotify tracks dataset [data set]. *Kaggle*, 2022.