Tunes Beyond Borders:
Exploring Music Taste Similarities on Spotify Across Regions

SI 670 Project Report
Nan-Hsin Lin (nanhsin), Yi-Chun Wang (ritaycw)

1. Introduction

This project seeks to explore and compare Spotify's local and global charts to uncover patterns and differences in music preferences across regions. The motivation behind this study stems from the intriguing interplay between language and music preferences. People from different regions often prefer to songs in their native languages, reflecting the cultural and emotional connections fostered by familiar lyrics and melodies. However, music also has the potential to cross language barriers, with global hits resonating in regions far from the language of the lyrics. This duality raises fascinating questions: What drives the global appeal of certain songs? How similar or distinct are the musical preferences of listeners across regions?

By examining these questions, we aim to uncover the extent of overlap in music tastes across regions and identify opportunities for cross-cultural music recommendations. Understanding these patterns could pave the way for innovative strategies, such as recommending music from one region to listeners in another, even if they are unfamiliar with the language. Such recommendations could enrich listening experiences, foster cultural exchange, and create a more interconnected global music community.

2. Methods

2.1. Data

The dataset for this project was sourced from the Spotify Web API, focusing on 74 regional and global daily charts for November 2024. The data includes 12 audio features for each track, including acousticness, danceability, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, valence, and duration (in milliseconds). Definitions for these features are available in Spotify's documentation [1], providing a robust foundation for analyzing music characteristics across regions.

Since Spotify's API offers real-time data, we fetched and accumulated data continuously throughout the project. However, due to the API's deprecation midway through our study [2], we could not collect complete daily chart data for the entire month. To address this limitation, we aggregated the tracks for each region, identifying the most popular songs for November as a whole. The data size ended up with 4716 tracks. This

approach allowed us to compile a meaningful dataset representative of music preference for each region during the study period.

2.2. Algorithm

To analyze the similarity of audio features across regions, we began with similarity metrics, specifically cosine similarity and Euclidean distance. These methods provided a quick and intuitive way to measure how closely regions aligned in terms of their music preferences based on Spotify's audio features. Cosine similarity is particularly suited for high-dimensional data as it focuses on the orientation rather than magnitude, while Euclidean distance captures absolute differences between feature vectors. To ensure the data was suitable for these methods and subsequent analyses, we standardized all features, which is essential for maintaining the relative importance of features and avoiding biases caused by differing scales.

For clustering, we employed multiple algorithms to explore patterns in the data: K-Means, Gaussian Mixture Model (GMM), Agglomerative Hierarchical Clustering, and DBSCAN. K-Means was chosen for its simplicity and efficiency in partitioning data into distinct clusters, while GMM allowed us to model clusters as probabilistic distributions, which can capture overlapping groups. Agglomerative Hierarchical Clustering provided a hierarchical perspective, showing relationships between regions at varying levels of granularity. Lastly, DBSCAN was used to identify clusters of varying densities and detect potential outliers, making it particularly effective for unevenly distributed data. Applying a diverse set of clustering methods enabled us to robustly compare and validate regional groupings.

To handle the high dimensionality of the audio features, we applied dimensionality reduction techniques: Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and t-SNE. PCA was selected for its ability to capture maximum variance with fewer components, effectively identifying and retaining the most important features while discarding noise. MDS preserved pairwise distances between regions, providing a geometric visualization of similarities. t-SNE, known for its strength in visualizing high-dimensional data in two or three dimensions, captured non-linear relationships in the data. In addition to simplifying the dataset, these methods reduced unimportant features that could act as noise, enhancing the quality of clustering and improving computational efficiency. Standardizing the features before applying dimensionality reduction ensured a fair contribution of all attributes to the results.

2.3. Implementation

The code begins by aggregating data by region and standardizing the features to ensure consistency across analyses. Similarity metrics are then applied, with results visualized using a heatmap to reveal patterns across regions. Various clustering models with different numbers of clusters are implemented, and the outcomes are illustrated through scatterplots and dendrograms to provide insights into regional groupings. Finally, dimensionality reduction techniques are applied to simplify the data, followed by clustering on the reduced features, improving the clustering performance and interpretability by focusing on the most relevant patterns in the data. Several numbers of components for dimensionality reduction were also experimented to achieve the highest performance.

3. Evaluation and Analysis

3.1. Evaluation

The primary evaluation metric we used was the silhouette score, which measures how well data points in a cluster align with their assigned group compared to other clusters. This metric was chosen over methods like the elbow chart because it provides a clear numerical assessment of cluster cohesion and separation without requiring subjective interpretation of a graph. A silhouette score closer to 1 indicates well-separated clusters, while a score near -1 suggests overlapping or poorly defined clusters.

To further visualize and interpret the results, we employed silhouette plots and clustering scatterplots as shown in *Figure 1*. Silhouette plots offer an intuitive way to assess how individual data points align with their respective clusters. Additionally, scatterplots help visualize the spatial distribution of clusters, especially when dimensionality reduction techniques like t-SNE are applied. For benchmarking, we considered a silhouette score greater than 0.5 as indicative of good clustering performance, given its scale between -1 and 1.

3.2. Analysis

Table 1 shows the result of our experiments applying different clustering algorithms on dimensionality reduction components. Our analysis revealed that the most effective clustering model was Agglomerative Hierarchical Clustering with 8 clusters applied to t-SNE reduced features with 2 components, achieving a silhouette score of 0.64. This score indicates well-defined, distinct clusters with minimal overlap. The clusters appeared visually separate, with clear boundaries in the scatterplots as shown in *Figure 2*, affirming the clustering quality. The dendrogram in *Figure 3* represents the relationships of similarity among clustered regions with a hierarchical tree view.

Comparing clustering methods, K-Means, Gaussian Mixture Model, and Agglomerative Hierarchical Clustering performed similarly well with the silhouette score ranging between 0.63 to 0.64, while DBSCAN failed to produce meaningful results, determining all data points as noisy samples and assigning them to a single cluster labeled as -1. Among the dimensionality reduction techniques, t-SNE outperformed others by effectively preserving the non-linear structures of the data. MDS performed moderately well but struggled with high-dimensional feature nuances, while PCA underperformed due to its focus on linear variance, which did not align well with the data's distribution.

To further analyze the cluster characteristics, we computed the mean and variance of audio features for each cluster and visualized them through heatmaps as shown in *Figure 4*, which provides a clear, quantitative representation of the distinctive musical preferences across clusters. For example, Cluster 7, encompassing countries like the Czech Republic, Poland, and Slovakia, exhibited higher energy, faster tempo, and lower valence, suggesting a preference for energetic, fast-paced, and emotionally negative tracks. On the other hand, Cluster 6, including Iceland, Israel, and Sweden, was characterized by low danceability, energy, and loudness but high acousticness, reflecting a preference for acoustic, softer music.

In summary, our evaluation and analysis underscore the effectiveness of hierarchical clustering combined with t-SNE for uncovering regional music taste patterns. This approach provides actionable insights into how diverse audiences experience music across the globe, and the analysis results highlight the potential for leveraging clustering insights to guide region-specific music recommendations.

4. Related work

There have been numerous studies on clustering music to enhance recommendation systems and analyze musical styles.

P. N. et al. (2022) [3] analyzed the effectiveness of five clustering algorithms—K-Means, Mini-Batch K-Means, Agglomerative Clustering, DBSCAN, and BIRCH—for building a music recommendation system. They used a Spotify dataset with similar features our project used and applied dimensionality reduction techniques like PCA and t-SNE before clustering. Their findings highlighted that K-Means clustering, with its high Silhouette Score and Davies-Bouldin Index, provided the most distinct and reliable clusters.

Z. Liumei et al. (2021) [4] proposed a novel approach to analyze traditional Chinese folk music by textualizing MIDI data into weighted text using tf-idf and then applying K-

Means clustering. Their work explored symbolic music data instead of acoustic features, incorporating domain-specific music theory to validate clustering results. Through this method, they uncovered distinct modal characteristics in Chinese folk music and demonstrated that clustering could reveal meaningful insights into regional musical grammar. Visualization of results was performed using t-SNE, aligning their approach with dimensionality reduction techniques used in modern MIR systems.

R. Sun et al. (2018) [5] implemented a music segmentation method based on histogram clustering to analyze the structure of pop music. By extracting beat-based pitch class profile (PCP) features and applying clustering algorithms such as K-means, K-means++, and Isodata, they achieved an average segmentation accuracy of 71.34% for a dataset of 200 Chinese pop songs. Their findings demonstrated that the K-means++ algorithm offered improved accuracy and lower time complexity compared to the traditional K-means algorithm, making it a suitable choice for music segmentation tasks.

C. Guan et al. (2015) [6] introduced the Intuitionistic Fuzzy Agglomerative Hierarchical Clustering (IFAHC) algorithm for music recommendation in folksonomy systems. This method integrates the Intuitionistic Fuzzy Set (IFS) concept to address vagueness and uncertainty in user-defined tags. IFAHC clusters music items based on their IFS values, producing dendrograms for decision-making and recommendation purposes. The approach was demonstrated with a social tagging dataset, showcasing its ability to manage fuzzy data and uncover meaningful clustering patterns for music recommendation.

5. Discussion and Conclusion

Our analysis revealed that Agglomerative Hierarchical Clustering, combined with t-SNE for dimensionality reduction, was the most effective approach, achieving a silhouette score of 0.64. This highlights the strength of hierarchical methods paired with non-linear dimensionality reduction techniques for uncovering regional music preferences.

The results show that music taste indeed crosses national boundaries, with unexpected similarities between distant regions. For instance, Israel shares a similar music profile with Sweden and Iceland, despite being in the Middle East while the latter two are in Northern Europe. This finding suggests the potential for recommending Israeli music to Swedish and Icelandic audiences, fostering greater playlist diversity while aligning with their preferences. Meanwhile, regional proximity still plays a role, as seen in the similarity between Australia and New Zealand. This balance between cross-cultural influences and regional affinities underscores the complex interplay of factors shaping regional music taste.

Through this project, we navigated unexpected challenges and gained valuable insights. Initially, we aimed to predict songs likely to appear on Spotify's Daily Song Charts and analyze differences between global and local charts to guide artists and record companies in optimizing their promotional strategies. However, Spotify deprecated the API midway through our project, limiting our access to essential data [2]. Consequently, predictive analysis and supervised learning approaches were no longer feasible. This experience stressed the importance of starting data collection as early as possible in a project to mitigate unforeseen disruptions.

The shift in focus allowed us to explore unsupervised learning more deeply. Although we had limited prior exposure to these techniques which were covered later in the course, the project provided an excellent opportunity to review clustering algorithms, implement dimensionality reduction, and understand their applications. This experience enriched our understanding of unsupervised learning methods and their ability to uncover hidden patterns in high-dimensional datasets.

Looking forward, we could explore new approaches for data collection and analysis. Without Spotify's audio feature API, we would focus on extracting audio features directly from track recordings. Applying deep learning models, such as Siamese Networks, for learning embeddings and searching similarity could help capture intricate relationships within regional music profiles and improve clustering effectiveness. Incorporating lyrics as a feature would also allow us to analyze the role of language in music preferences, providing deeper insights into whether music taste truly transcends national and linguistic boundaries. These steps could further refine our understanding of regional music trends, opening up new possibilities for music recommendation systems.

6. References

[1] Spotify for Developers, "Get Audio Features," available at: https://developer.spotify.com/documentation/web-api/reference/get-audio-features, accessed December 9, 2024.

[2] Spotify Community, "Changes to Web API," available at: https://community.spotify.com/t5/Spotify-for-Developers/Changes-to-Web-API/td-p/6540414, accessed December 9, 2024.

[3] P. N, D. Khanwelkar, H. More, N. Soni, J. Rajani and C. Vaswani, "Analysis of Clustering Algorithms for Music Recommendation," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9824160.

[4] Z. Liumei, J. Fanzhi, L. Jiao, M. Gang and L. Tianshi, "K-means clustering analysis of Chinese traditional folk music based on midi music textualization," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2021, pp. 1062-1066, doi: 10.1109/ICSP51882.2021.9408762.
[5] R. Sun, J. Zhang, W. Jiang and Y. Hu, "Segmentation of Pop Music Based on Histogram Clustering," 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 2018, pp. 1-5, doi: 10.1109/CISP-BMEI.2018.8633060.
[6] C. Guan, K. K. F. Yuen and F. Coenen, "Towards an Intuitionistic Fuzzy Agglomerative Hierarchical Clustering Algorithm for Music Recommendation in Folksonomy," 2015 IEEE International Conference on Systems, Man, and Cybernetics, Hong Kong, China, 2015, pp. 2039-2042, doi: 10.1109/SMC.2015.356.

7. Code Repository

The code of this project is stored in GitHub Repository as the link below:
https://github.com/nanhsin/tunes_beyond_border

8. Appendix

Table 1. Comparison of Clustering Algorithms on Dimensionality Reduction Components

Silhouette Score (n_clusters=n)				
	Original Features	**PCA** (n_components=5)	**MDS** (n_components=2)	**t-SNE** (n_components=2)
K-Means	0.21 (n=2)	0.21 (n=6)	0.40 (n=4)	0.63 (n=7)
GMM	0.21 (n=4)	0.22 (n=6)	0.39 (n=4)	0.63 (n=7)
Agglomerative	0.22 (n=10)	0.21 (n=2)	0.39 (n=4)	0.64 (n=8)

Figure 1.

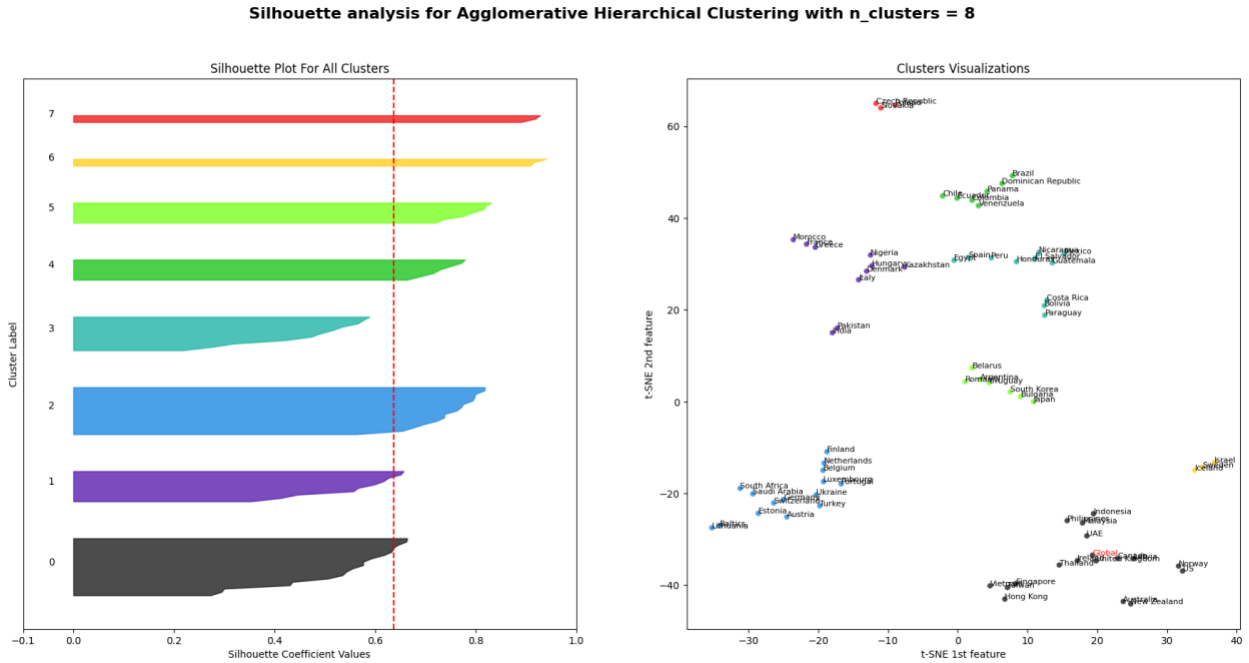

Figure 2.
Scatterplot of Agglomerative Hierarchical Clustering (n_clusters=8) with t-SNE (n_components=2)

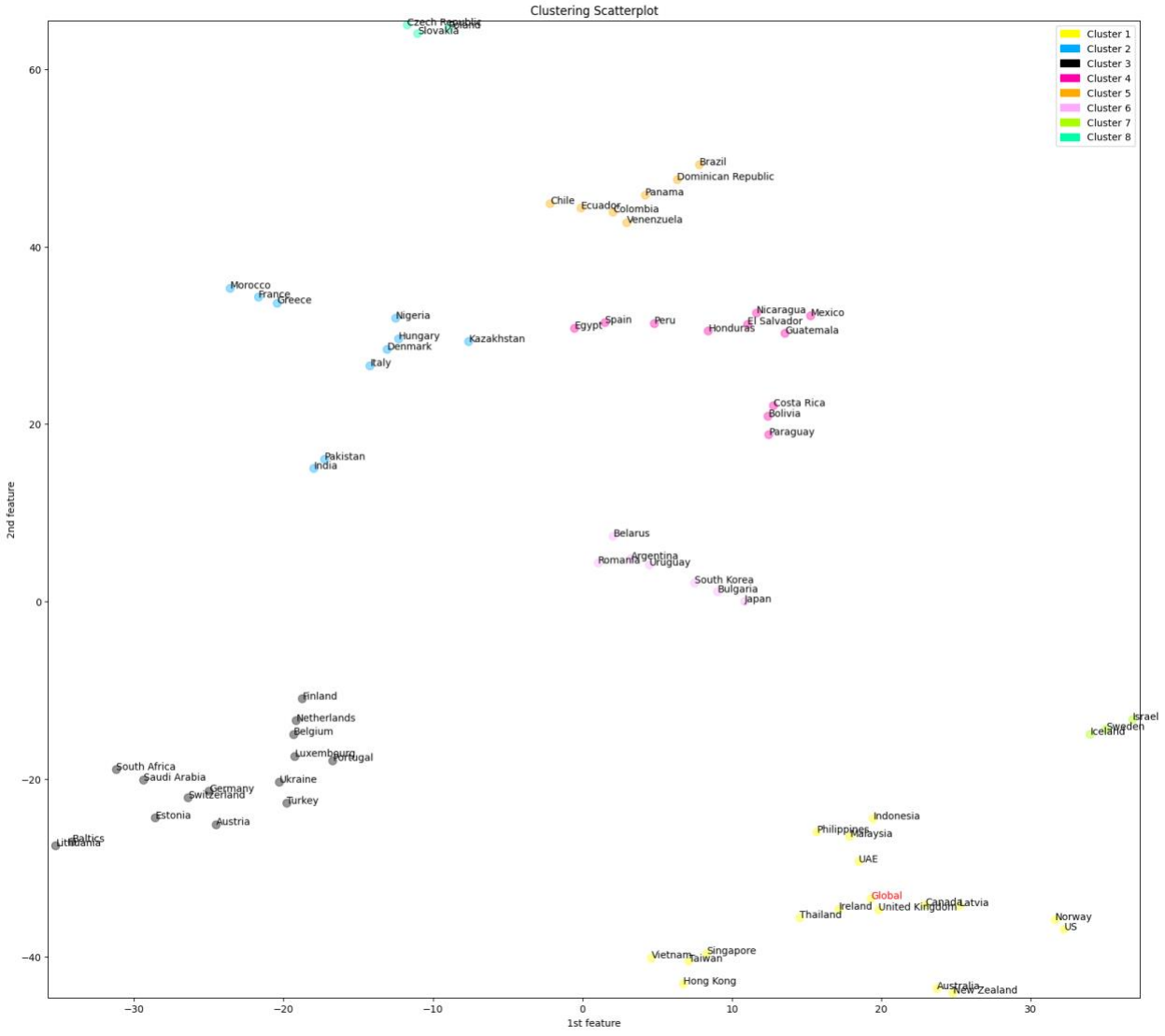

Figure 3. Dendrogram of Agglomerative Hierarchical Clustering (n_clusters=8) with t-SNE (n_components=2)

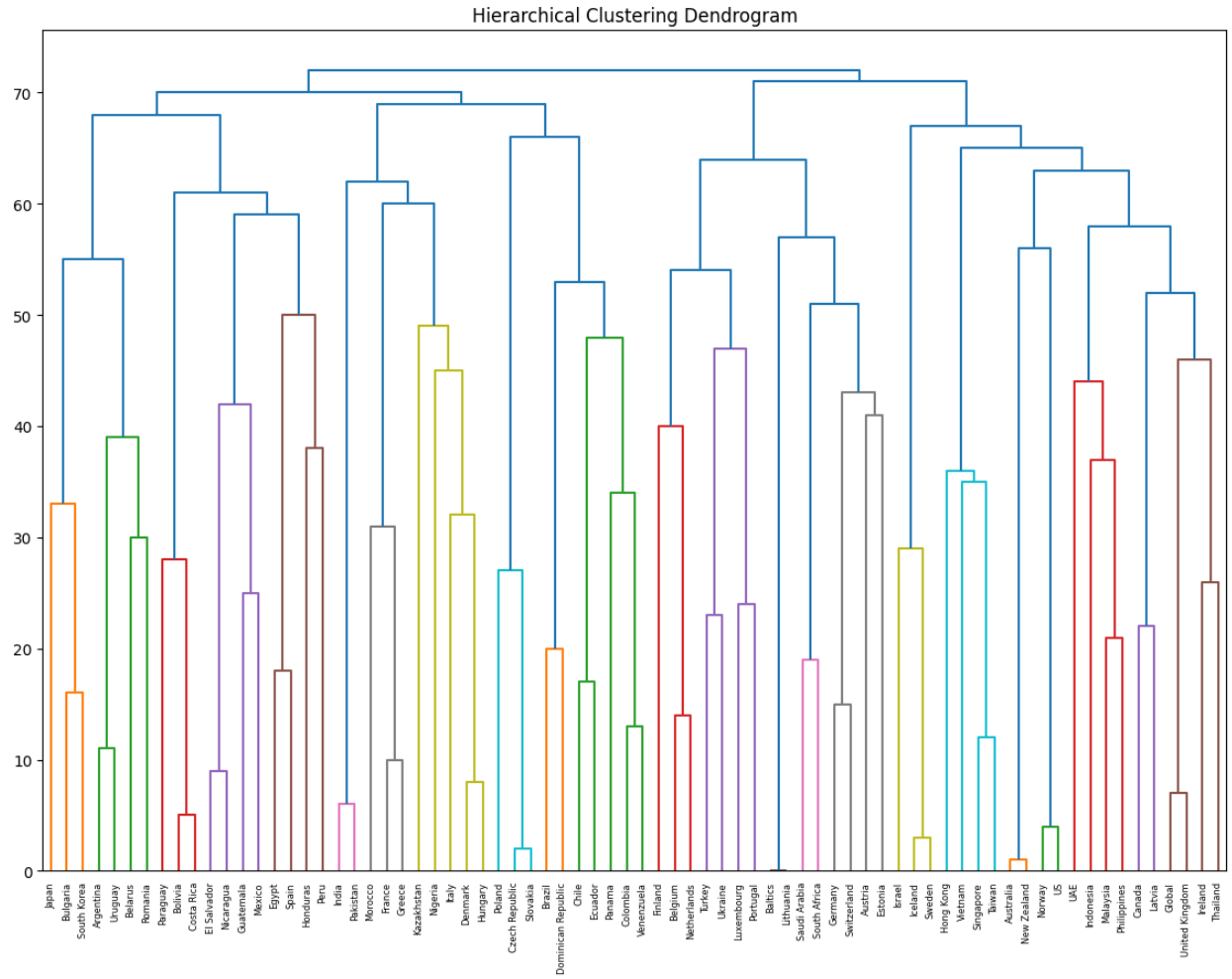

Figure 4. Heatmap of Feature Averages and Variances by Cluster