Rachael Enders
Faith Mathew
Cameron Schneider
Shannon Wasson
CIS 635
June 14, 2024

An Analysis of Spotify's Recommendation System

Introduction

Spotify is a Swedish company, known worldwide for its massive library of songs, podcasts, and videos on demand. In a world where technology is improving so swiftly, however, Spotify needs a leg up to edge out the competition. Apple Music, Pandora, and Soundcloud are all close competitors in the race, trying to steal the spotlight on music streaming that Spotify has held for over a decade. The main feature that Spotify uses to continue to lead this industry is its recommendation system, where Spotify is able to accurately deduce a user's listening preferences and recommend new songs to them. This system is so robust and accurate that it singlehandedly puts Spotify well above its competitors.

This begs the question: How can Spotify's recommendation algorithm stand so tall above its competitors? Each and every Spotify song has a series of seemingly unimportant or ambiguous data attached to it, ranging from definitive data such as tempo and loudness to nearly non-quantifiable data such as danceability. The goal of this project is to understand how these data points are set as well as how they are used to compare songs to make improvements on the recommendation system.

Related Work

Spotify is one of the most popular music streaming applications available and its algorithm is quite well known. As a result, quite a few projects have been done on this topic. Since there are only so many changes that can be made to an algorithm, they all do hold some similarities to our

project at their core. However, there are differences in the goals and manipulation of the data that make the projects different. A few examples of these include:

1. This study aimed to see how changing Spotify's interface and user interaction impacts the effectiveness of the music recommendation algorithm. They built two different interfaces, one that allowed the user to slide on how much they liked certain attributes of music (for example: danceability slid to 80) and another that was shaped like a radar that allowed the user to select how much they liked certain attributes of music on a radar graph. This is very different from our project because they are testing the algorithm by changing the user interface whereas we are just building our own algorithm and comparing it to Spotify's own. [1]
2. This study used data mining techniques similar to ours (t-tests) but the data gathering and goals of the study were different. The authors of this project sent out a survey and gathered songs from the survey taker's public Spotify playlists. From these playlists, they used similar data mining techniques as us but used them to try to predict personality traits, demographic characteristics, and random habits (for example: smoker versus nonsmoker). [2]
3. This study does not necessarily delve into the technical data but rather what makes an algorithm so important. By using data mining techniques to determine what someone will most likely like based on their gender, ethnicity, and location, the algorithm feeds the user only specific types of music, and in Spotify's case, podcasts. This can lead to users being only exposed to a very small subset of what is available just due to what the algorithm is set to. [3]

These are just a few examples of similar projects that use Spotify's algorithm as the basis of their projects. Overall, ours is different because most are trying to determine how someone's playlist is an indicator of who they are as a person. Instead, we are working to build and potentially better an existing algorithm.

Methods

We used a predefined dataset. This dataset is from Kaggle, and is titled "Spotify Tracks Dataset" (see Data and Software Availability). This dataset includes over 100,000 songs from 25 different

genres, with each song having 21 different data points that are used in the recommendation process. This is not a conclusive list of all the songs on Spotify, but it is a well-sized sample for our purposes.

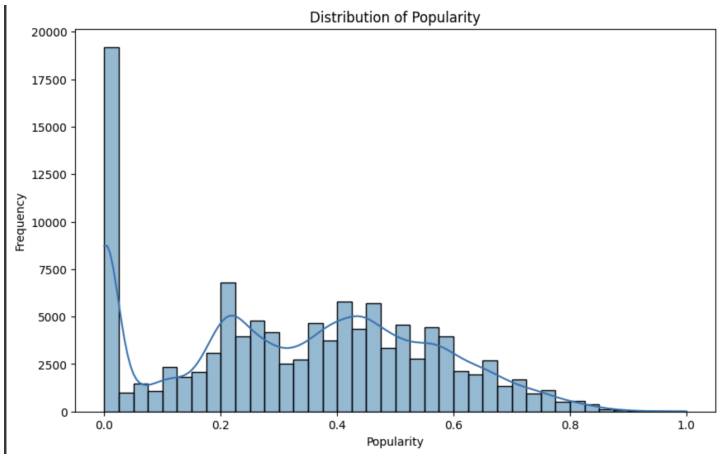
The first step that we took before beginning data analysis was to clean up the dataset. Here, we filled any empty data points with generic data, such as empty Artist columns being filled with 'Unknown Artist'. We normalized the numerical data points, using Min-Max Scaling to normalize the data. Finally, we encoded the categorical variables for easier analysis. EDA was done to gain insights into the distribution and relationships among the variables. Visualizations including histograms, correlation heatmaps, and pairplots were generated using libraries such as Matplotlib and Seaborn to understand the distribution of features and identify potential correlations. We used K-means clustering to group tracks based on their musical traits. The optimal number of clusters was determined using the elbow method, which evaluates the inertia for different numbers of clusters. The silhouette score, a metric that measures the compactness and separation of clusters, was used to evaluate the performance of the clustering algorithm. A higher silhouette score indicates better-defined clusters with minimal overlap. The entire data mining pipeline, from data loading to clustering, was documented with code snippets and explanations. The code provided can be executed in a Python environment, ensuring the reproducibility of results. The dataset used in this analysis is available for download, enabling others to replicate the analysis and explore further. The entire data mining pipeline, from data loading to clustering, was documented with code snippets and explanations. The code provided can be executed in a Python environment, ensuring the reproducibility of results. Additionally, the dataset used in this analysis is available for download, enabling others to replicate the analysis and explore further.

Results and Discussion

The analysis of the Spotify tracks dataset included data preprocessing, data cleaning, visualization, and clustering to uncover insights and enhance recommendations. Since we had missing values in the columns artists, album_name, and track_name.For effective clustering, we carried out normalization using MinMaxScaler to standardize the scale. Visualizations provided a comprehensive understanding of the dataset.

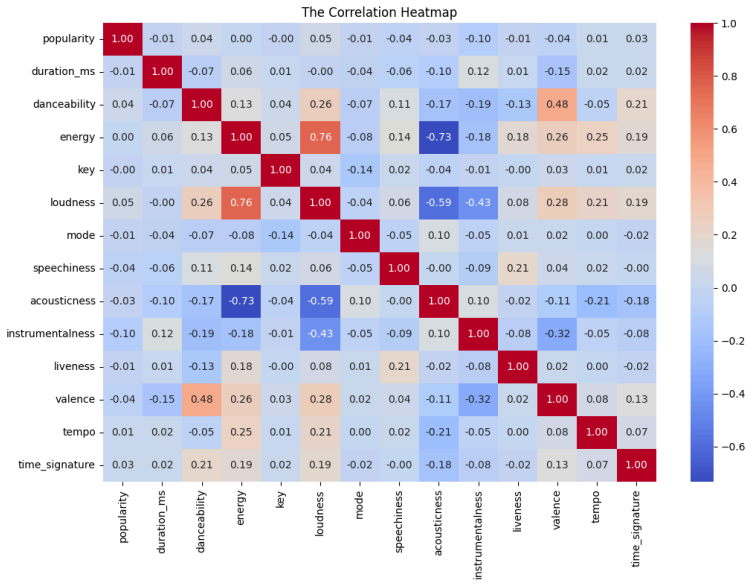
Popularity Graph

The popularity distribution showed a skew towards lower popularity tracks.



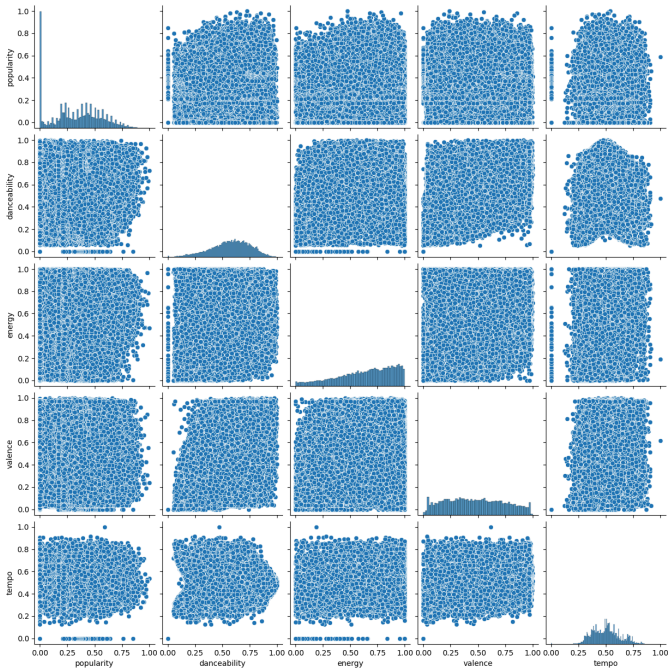
Correlation Heatmap

The correlation heatmap highlighted significant relationships among features, such as the positive correlation between danceability and energy.

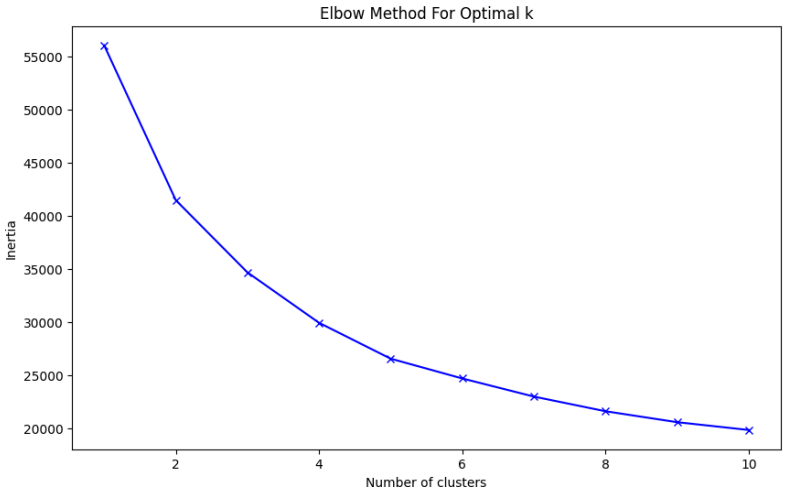


Pairpoint Plot

The pairplot further illustrated these relationships, aiding in selecting features for clustering.



Elbow Method:



Using the elbow method, we identified four as the optimal number of clusters, with the silhouette score of 0.4678 indicating moderate clustering performance. The clusters were visualized based on danceability and energy, revealing distinct groupings that could be leveraged for recommendations.

Conclusion

Through data mining techniques and analysis, we were able to determine that Spotify is able to build an algorithm that is personalized to an individual's taste by using specific musical traits. These traits independently are not enough to base an algorithm on. However, when combined and used in tests like pairpoint plots, relationships are able to be found and an algorithm is able to be built off of that idea. Time was a limit on this project since we were only able to delve so far with a short turnaround. If our group had more time to continue this project in the future, it would be interesting to add more traits to consider in the algorithm. Demographic traits, including but not limited to gender, sexuality, relationship status, ethnicity, education level, and income status could all be very intriguing algorithmic predictors to look into. Future work could include incorporating additional features, refining the clustering algorithm, and conducting user studies to evaluate recommendation effectiveness.

Data and Software Availability

GitHub Repository for this Project:

https://github.com/GVSU-CIS635/term-project-spotify-algorithm

Dataset Used:

https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

The Python code implementing our data mining pipeline and clustering algorithm is also provided in the repository for replication and further research.

References

[1] Millecamp, M., Htun, N., Jin, Y., & Verbert, K. (2018). Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces. In Proceedings of the 26th Conference on User Modeling, Adaption and Personalization, *(UMAP '18)* (101-109). Association of Computing Machinery. https://doi.org/10.1145/3209219.320922

[2] Tricomi, P. P., Pajola, L., Pasa, L., & Conti, M. (2024). "All of Me": Mining Users' Attributes from their Public Spotify Playlists. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24)* (634-636). Association for Computing Machinery. https://doi.org/10.1145/3589335.3651459 1

[3] Werner, A. (2020). Organizing music, organizing gender: algorithmic culture and Spotify recommendations. *Popular Communication*, *18*(1), 78–90. https://doi.org/10.1080/15405702.2020.1715980