Supervised and Unsupervised Song Recommendation Via K-Means Clustering and K-Nearest Neighbors

Purple Team 8

Brian Yoonjae Kim, Ben Nussbaum, Jessie Cecilya, Navi Singh, Steve Cutler

Project Goal

The goal of the project is to implement new methods of recommendation and song exploration for the consumer. We attempt to make recommendations based on how the song sounds rather than listed attributes such as genre and artist name. Our clustering only uses publicly available data about songs and does not use behavioral data.

Description of Data

Our main dataset contains information on 169,909 songs on Spotify and additional attributes based on Spotify's analysis on the songs. This dataset comes from kaggle: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks. The descriptions of the data are listed in the below table.

Attribute	**Description**
Year	Year the song was released
Acousticness	Acousticness of the track
Artists	Artists that contributed to the song
Danceability	How suitable the song is for dance. Based on tempo, rhythm stability, beat strength, and overall regularity
Valence	Positivity of the track
Duration_ms	Length of the track in milliseconds
Energy	Perceived intensity and activity. Higher energy songs feel noisy, fast, and loud
Explicit	Whether or not the track has explicit lyrics
Id	Unique id of the track
Instrumentalness	Predicts whether a track contains no vocals

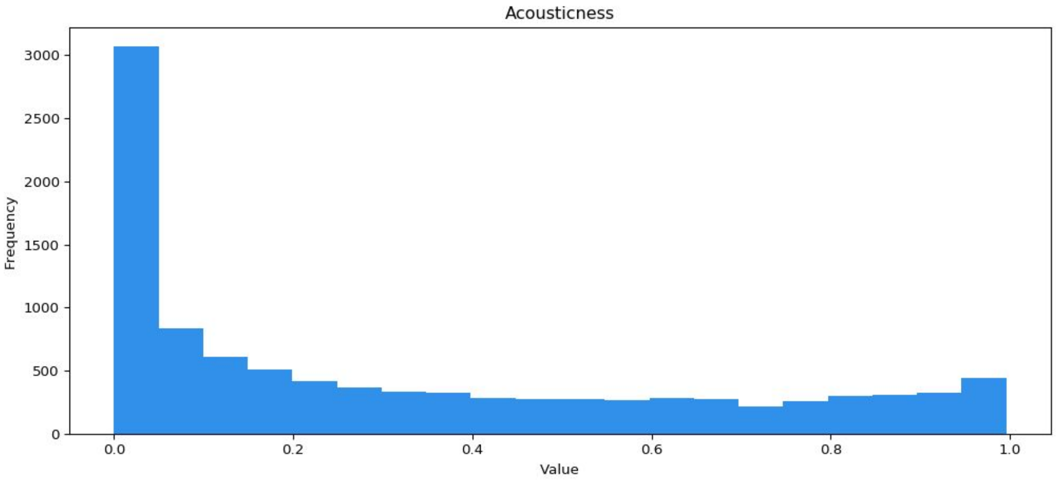
Liveness	Detects presence of a live audience in the recording. A high value indicates that the likelihood of a live audience is high
Loudness	Average loudness of the track in decibels
Mode	Musical Modality of the track (Major and Minor)
Key	Musical key of the track
Name	Name of the track
Popularity	Popularity of the track, based on total numbers of plays and how recent those plays are
Release_date	Release date of the track
Speechiness	Detects spoken words on a track. A higher score indicates that there are more spoken words
Tempo	Average beats per minute

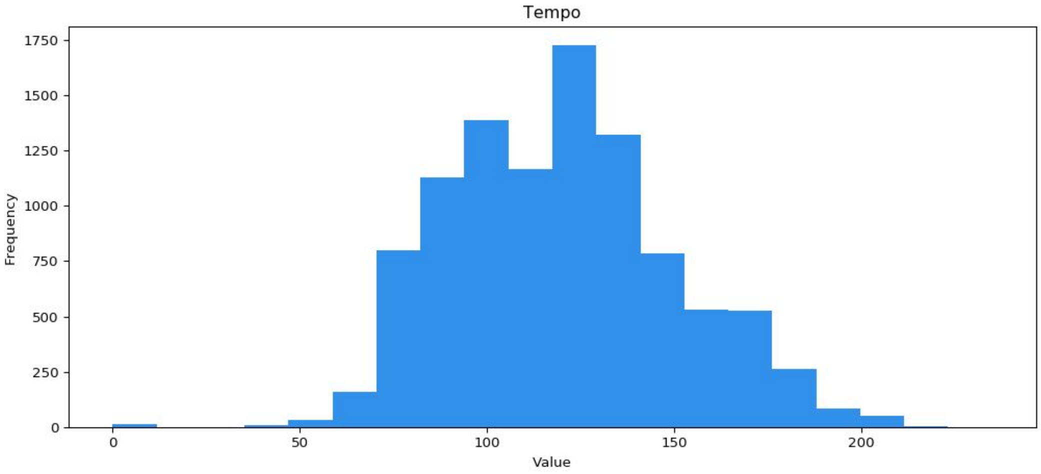
Data Preparation

Our analysis requires that we only use numerical values, so we had to exclude year, artists, explicitness, id, key, mode, name, and release date from the main clustering analysis. Tempo, loudness, and duration had to be normalized because they are on their own scales.

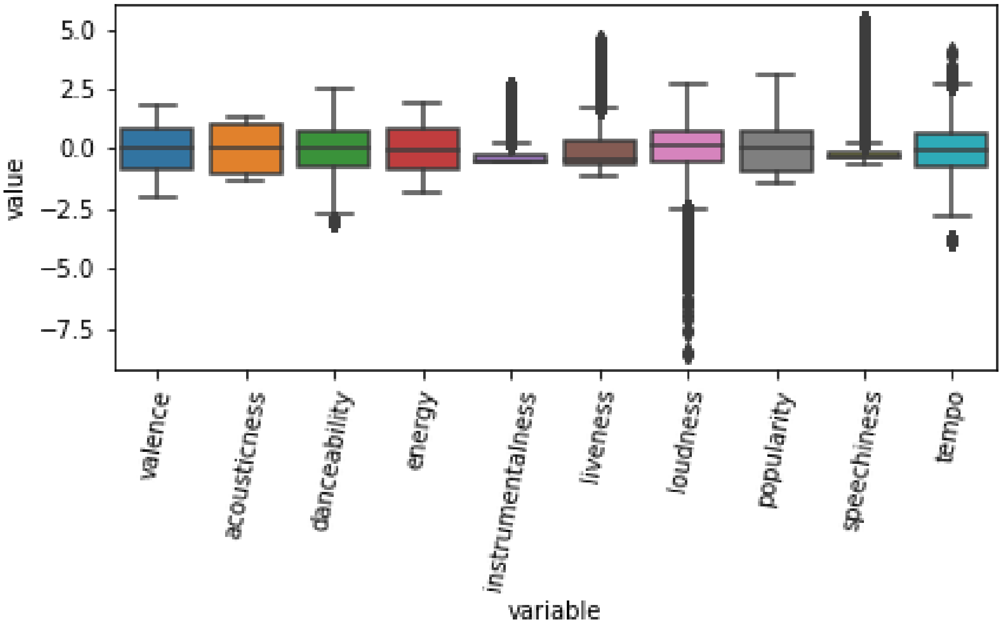
Data Exploration

Spotify API included the individual distributions of the attributes. Acousticness, instrumentalness, liveness, loudness, speechiness, are very skewed and the rest are close to normal distributions. Some examples are below and all of the distributions can be seen on the spotify api website.





We decided to plot the attributes using a box plot to compare distributions. Although the distributions are close to one another, the skewed distributions have lots of outliers indicated as black dots outside of the boxes.



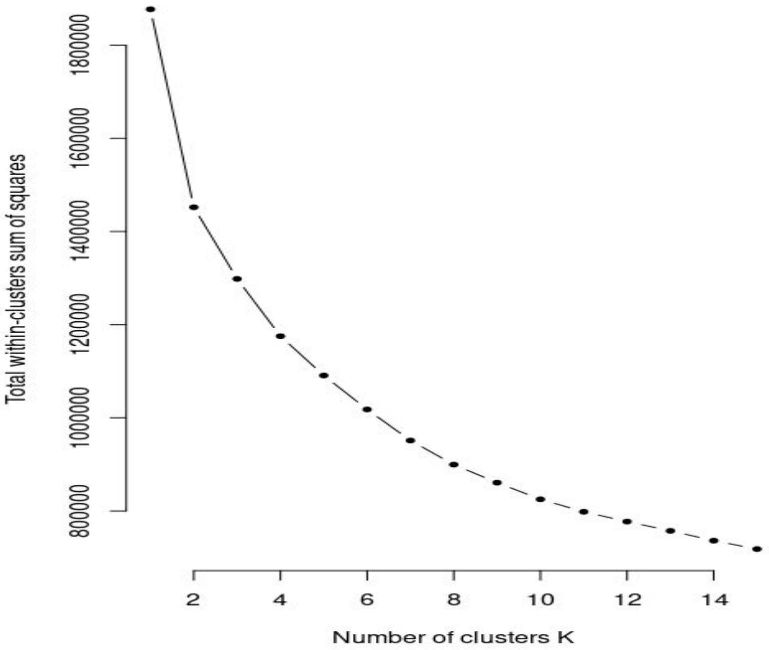
Data Modeling

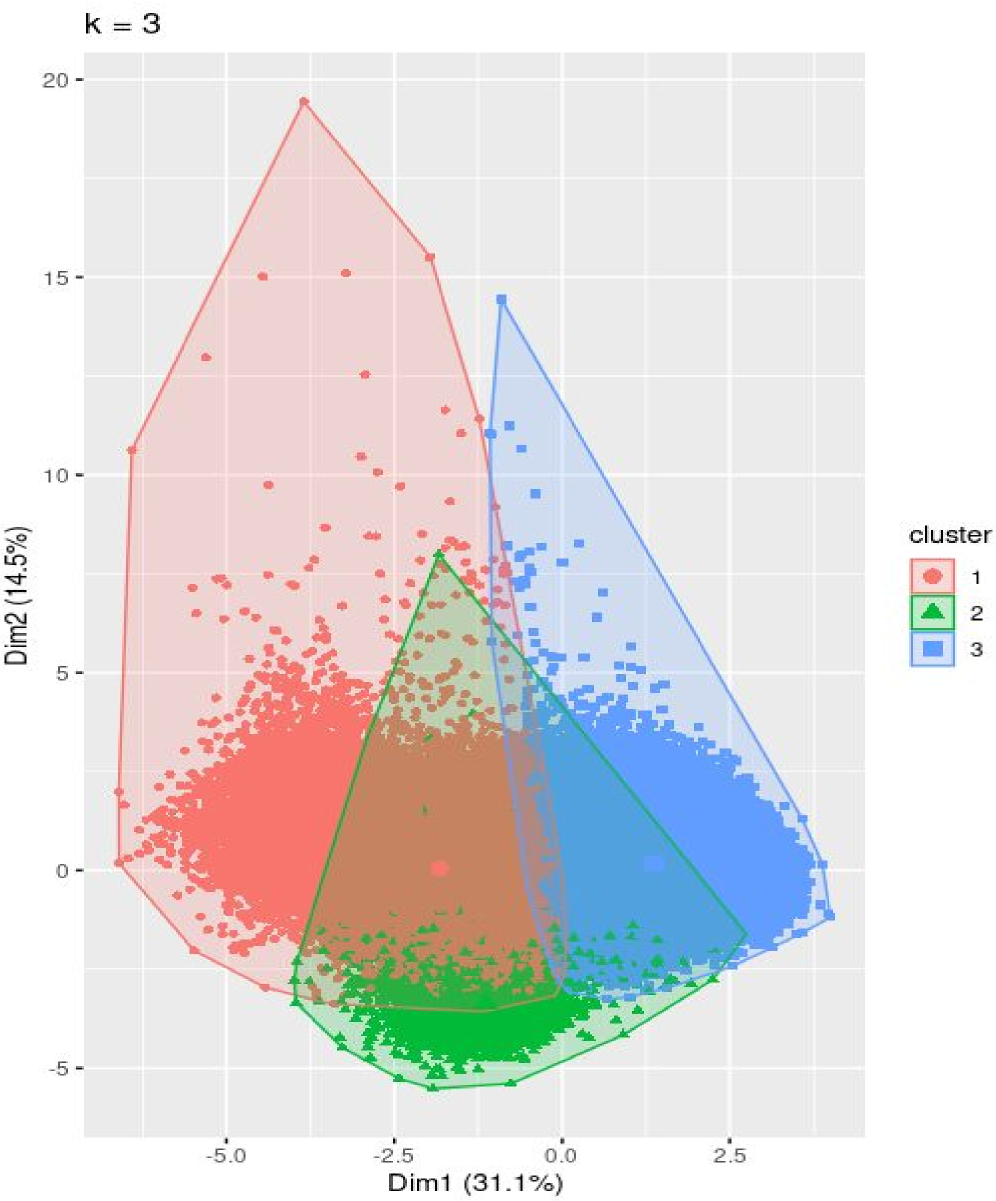
Our goal of creating a new recommendation system requires us to map how similar or different one song is from another. Other than single song recommendations, we thought we could also create prepopulated playlists that every user could see. Our first exploration of this saw us performing k-means clustering using only the numerical attributes. In order to calculate the optimal number of clusters, we used the knee/elbow method. We then utilized k-nearest neighbors to enable single song recommendations when a potential user searches, though this could be used to create playlists as well.

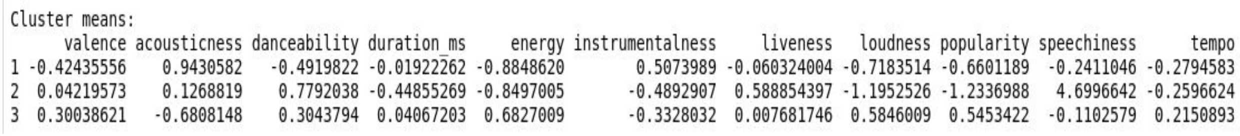
K-Means Clustering

In the plot here, our elbow bends when the number of clusters is 3.

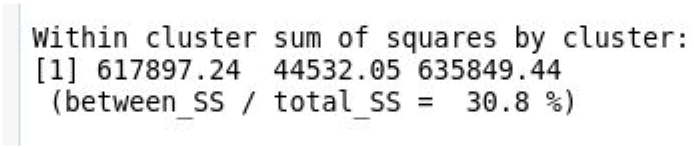
After running k-means with 3 clusters, we get the following output. (Note - the cluster package automatically performs Principal Component Analysis and reduces all the dimensions down to just 2 dimensions so it can be plotted). Following this image is the output of our clustering analysis.







```
Cluster means:
      valence acousticness danceability duration_ms      energy instrumentalness      liveness   loudness popularity speechiness       tempo
1 -0.42435556    0.9430582   -0.4919822 -0.01922262 -0.8848620        0.5073989 -0.060324004 -0.7183514 -0.6601189  -0.2411046 -0.2794583
2  0.04219573    0.1268819    0.7792038 -0.44855269 -0.8497005       -0.4892907  0.588854397 -1.1952526 -1.2336988   4.6996642 -0.2596624
3  0.30038621   -0.6808148    0.3043794  0.04067203  0.6827009       -0.3328032  0.007681746  0.5846009  0.5453422  -0.1102579  0.2150893
```



```
Within cluster sum of squares by cluster:
[1] 617897.24  44532.05 635849.44
 (between_SS / total_SS =  30.8 %)
```

Result:

For our k-means analysis our within cluster sum of square by cluster was 30.8 percent. We could have reduced this variability by removing some songs, but that would potentially mean never recommending a particular song.

Listed below are the major attributes of each of the clusters, and what we think they should be labeled.

Cluster category characteristics

#1 - Moody Instrumental Music

High: Acoustic + instrumental + danceability + energy + loudness

Low: Positivity

#2 - Groovy but Chill Music

High: Danceability + Lively + lyrical

Low: Energy

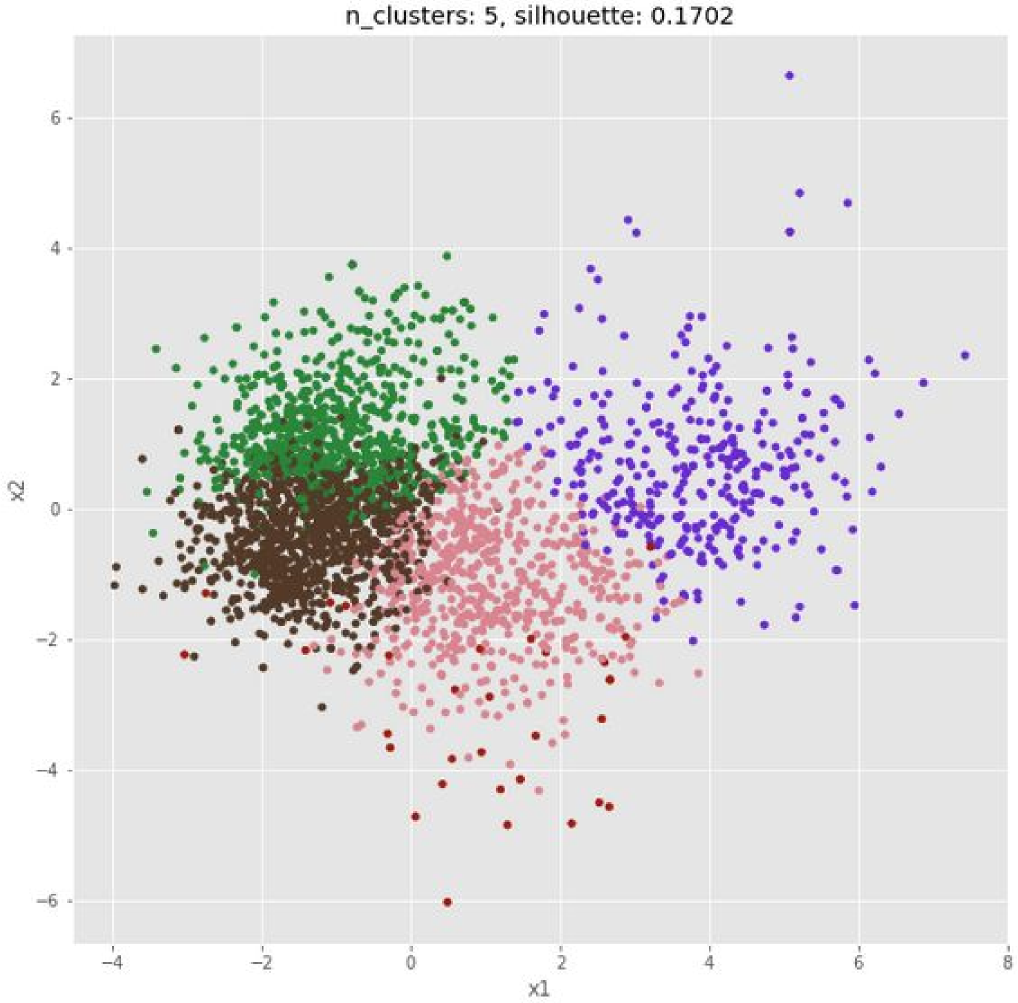
#3 - Top Dance Hits

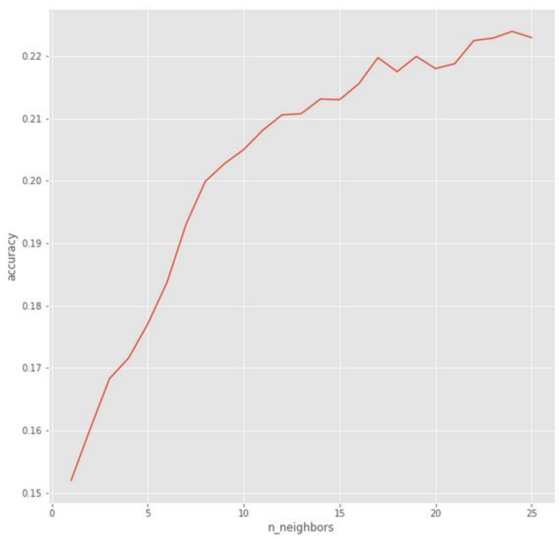
High: Positivity + danceability + high energy + loud + popularity + high tempo

Low: Acoustic + speechiness

K-Nearest Neighbors

We then performed k-nearest neighbors in using the genres as our target variable. Because there are so many unique genres, we had to reduce the number of genres to the top 25 genres. A k-nearest neighbors analysis with the same attributes as our k-means cluster analysis and 5 clusters gives us a silhouette score of 0.1702. This is the highest score when we tested 1-7 clusters. We used the value k = 26 for the analysis because it offered the best accuracy of 22% The output of the k-nearest neighbor model and it's fit are as follows:





Recap and Insights: Methods For Generating Recommendations

Playlist Recommendations:

If a streaming service had limited behavioral data but still wanted to display interesting pre-populated playlists, it could sample random songs from each cluster from the k-means analysis or the k-nearest-neighbors analysis. The clusters could be further filtered by the categorical attributes such as key, mode, album, and year. Over time as more behavioral data comes in, song picks could be informed by clustering people and recommending songs that similar behaving users have listened to. For example one person’s song search history, song repeat counter, and song skips could determine which songs to recommend or filter out. Our song analysis offers a baseline for what the users will initially see in the recommendations without the behavioral data and would be valuable to a streaming service to potentially prevent a new user from ditching the service altogether.

Single Song Recommendations:

If someone searches for a song, we can recommend songs based on the clusters in the k-nearest-neighbors analysis. Just like playlist recommendations, we could further segment the clusters by album, artist, mode, and key. Ideally we would use a hybrid method and include behavioral data such as song skips and repeated plays along with the k-nearest-neighbors analysis. Both K-Means and K nearest neighbors could be useful for classifying songs with no genre at all.

Challenges and Possible Improvements

Challenge: Lack of Behavioral data

Having behavioral data would make a huge difference in how recommendations are made. Demographic data on the users could also be useful for recommendations. Being able to also cluster people based on their behaviors would make it easier to predict single song recommendations. If we had the chance to collect the data ourselves, we would attach each genre directly to the song so that we would not have to wrangle so much with the dataset.

References

https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

https://developer.spotify.com/documentation/web-api/

https://en.wikipedia.org/wiki/K-means_clustering

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

https://www.kaggle.com/hafizamunshi/spotify-mood-prediction-and-music-recommendation