Spotify Recommender.

By, Robert Huey

Table of contents

Introduction

I wanted to build a music recommender for spotify based on metrics used in music by finding clusters in song features.

Gathering data.

Originally I was using the API, but it wouldn't connect to the live host to gather recommendation data. So I found this lovely application that would allow me to strip the info from spotify by using the URI paths to playlists. I then created a bunch of playlist URI's and ran them through the application to gather my music data and turn them into csv's.

By, plamere
Website -
http://organizeyourmusic.playlistmachinery.com/

The Track Properties

Descriptions by, plamere.

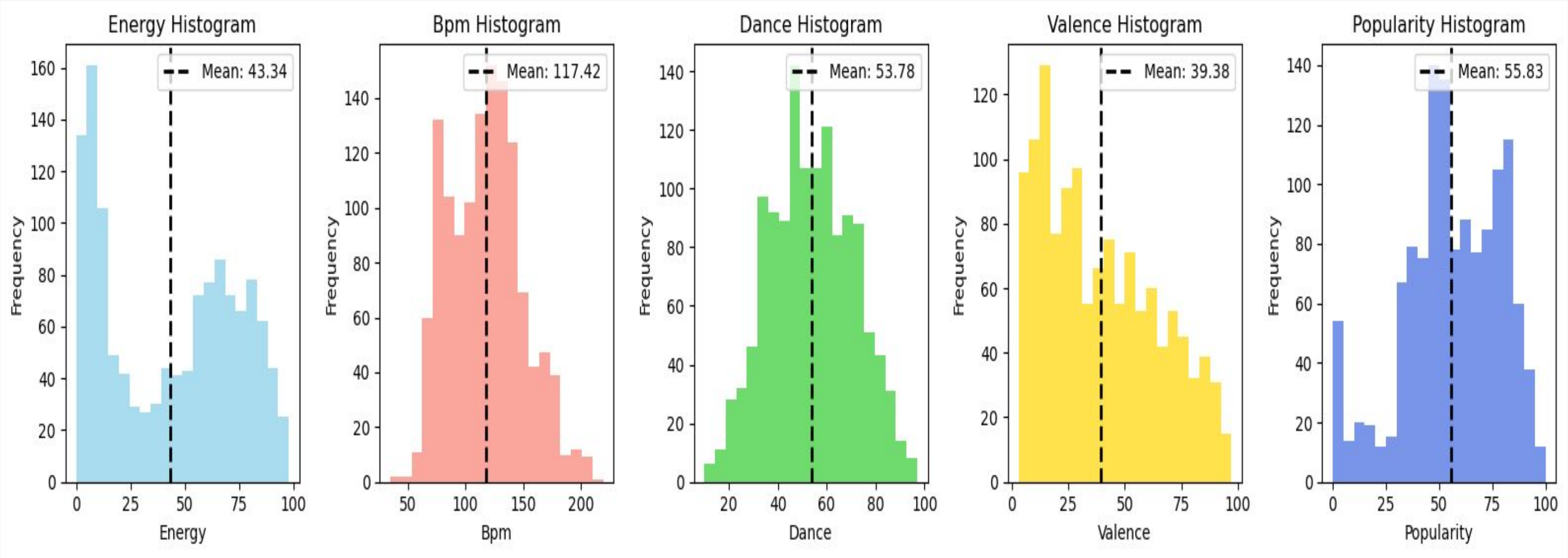
1. **Genre** - the genre of the track
2. **Year** - the release year of the recording. Note that due to vagaries of releases, re-releases, re-issues and general madness, sometimes the release years are not what you'd expect.
3. **Added** - the earliest date you added the track to your collection.
4. **Beats Per Minute (BPM)** - The tempo of the song.
5. **Energy** - The energy of a song - the higher the value, the more energtic. song
6. **Danceability** - The higher the value, the easier it is to dance to this song.
7. **Loudness (dB)** - The higher the value, the louder the song.
8. **Liveness** - The higher the value, the more likely the song is a live recording.
9. **Valence** - The higher the value, the more positive mood for the song.
10. **Length** - The duration of the song.
11. **Acousticness** - The higher the value the more acoustic the song is.
12. **Speechiness** - The higher the value the more spoken word the song contains.
13. **Popularity** - The higher the value the more popular the song is.
14. **Duration** - The length of the song.

Cleaning.

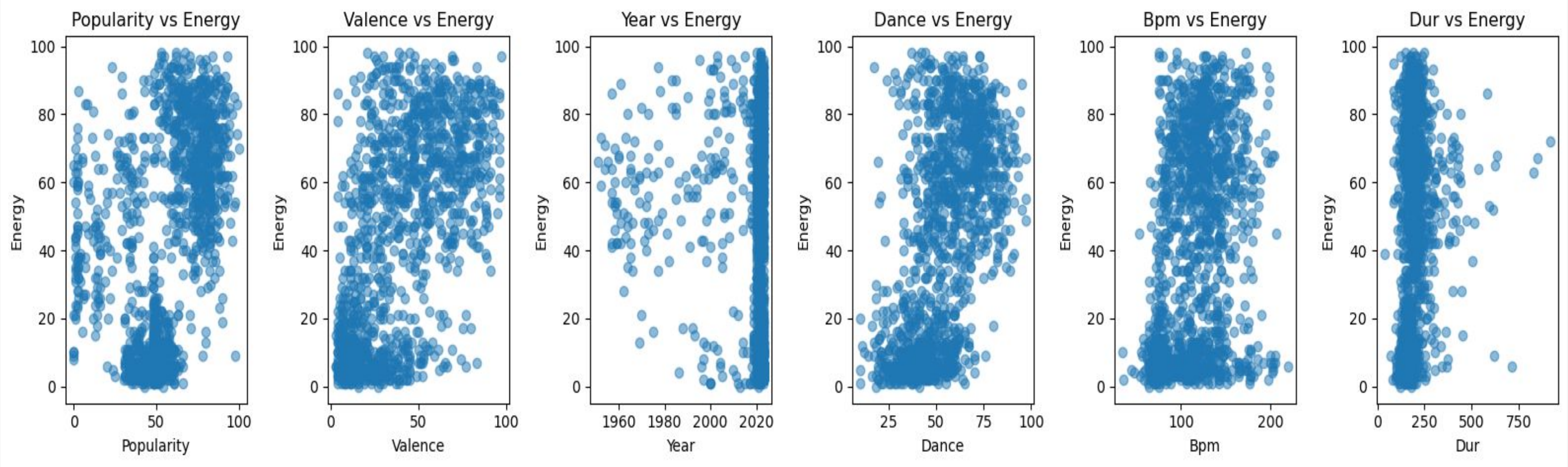
Artists that belonged to multiple genre's would show up as NaN values. So I assigned them the 'alternative' genre to not lose a large amount of data. I also re-named the columns and deleted duplicate songs.

Exploratory data analysis.

First, I wanted to take a look at the feature distributions separately to see what insights I could pull from the graphs and the playlists I concatenated.



Then I wanted to see if the data clustered anywhere.

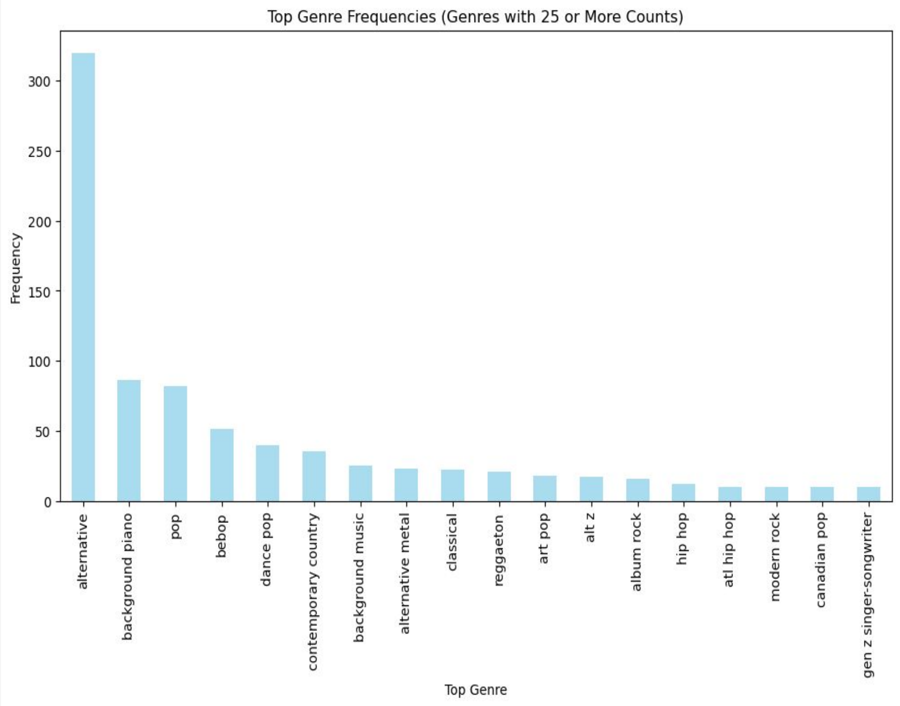


Popularity had about 3 clusters and valencer with 2.

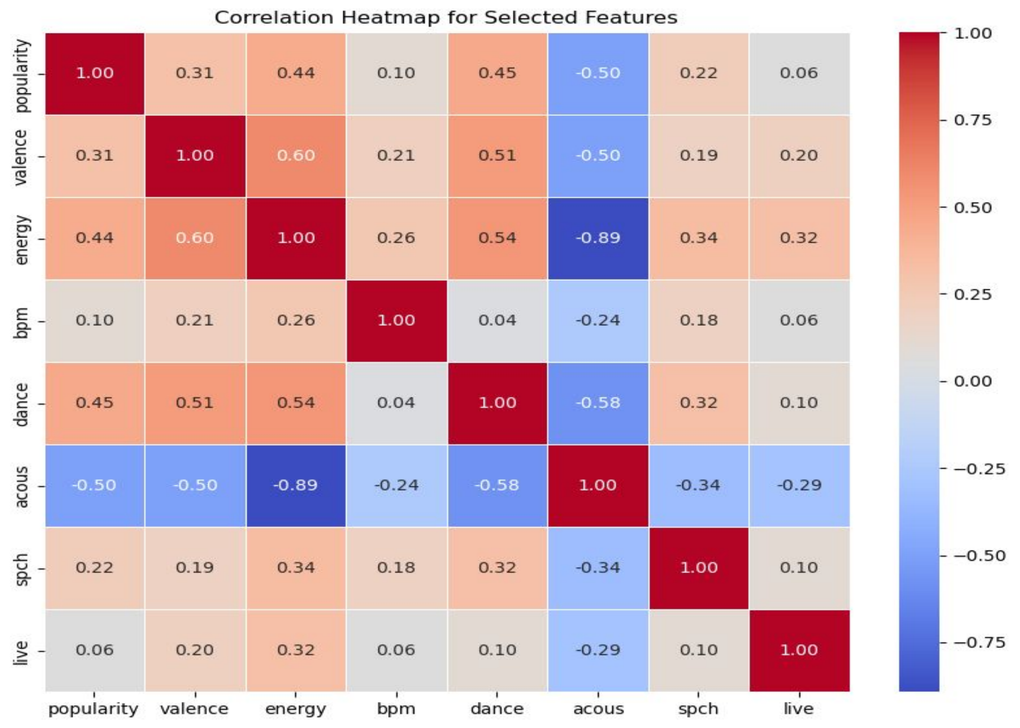
Year was just 1. Danceability had 2 clusters.

BPM semi has 2 and duration is just 1 cluster.

Lastly, I wanted to see how the top 25 genres were distributed.



Heatmap for correlations.



.60

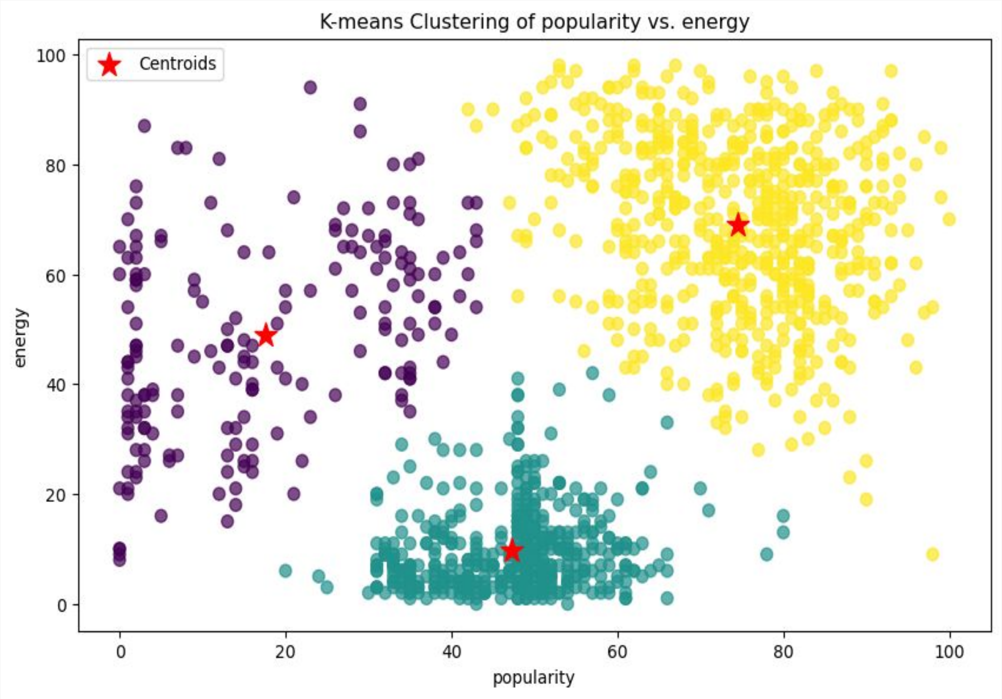
Energy and Valence.

Had the largest impacts out of all features. Acoustic had the worst negative impact by far.

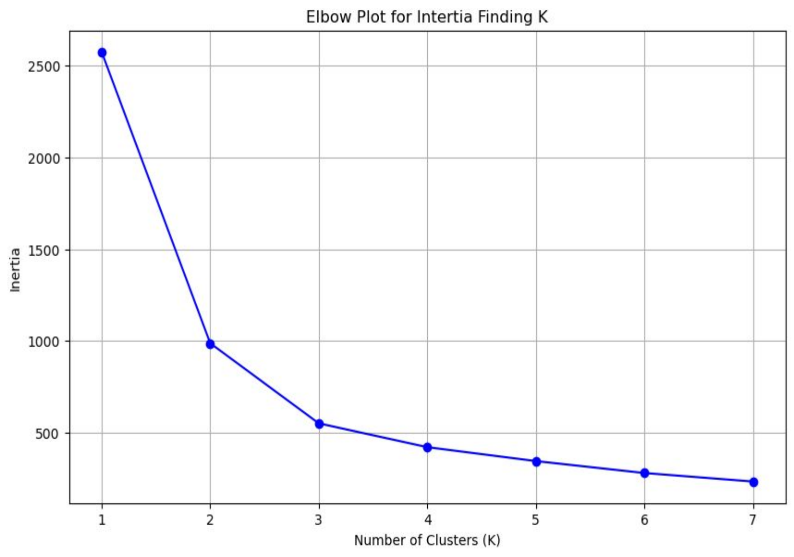
Modeling.

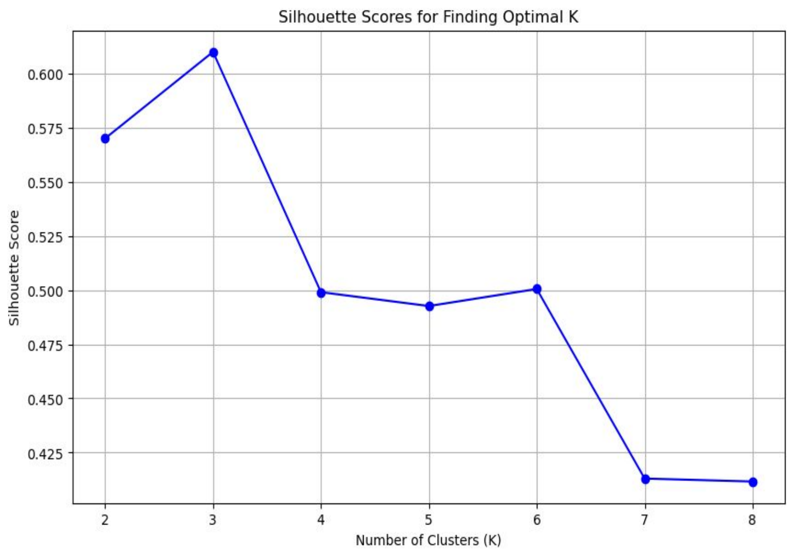
K-means Clustering.

After multiple iterations with different features the clearest clusters were with just using two. Energy and popularity. As we can see in the model we have 3 somewhat clear clusters using the combination of the two.



We can see inertia sticking at 2 clusters but our silhouette elbow aiming for 3 clusters with a 61% score.

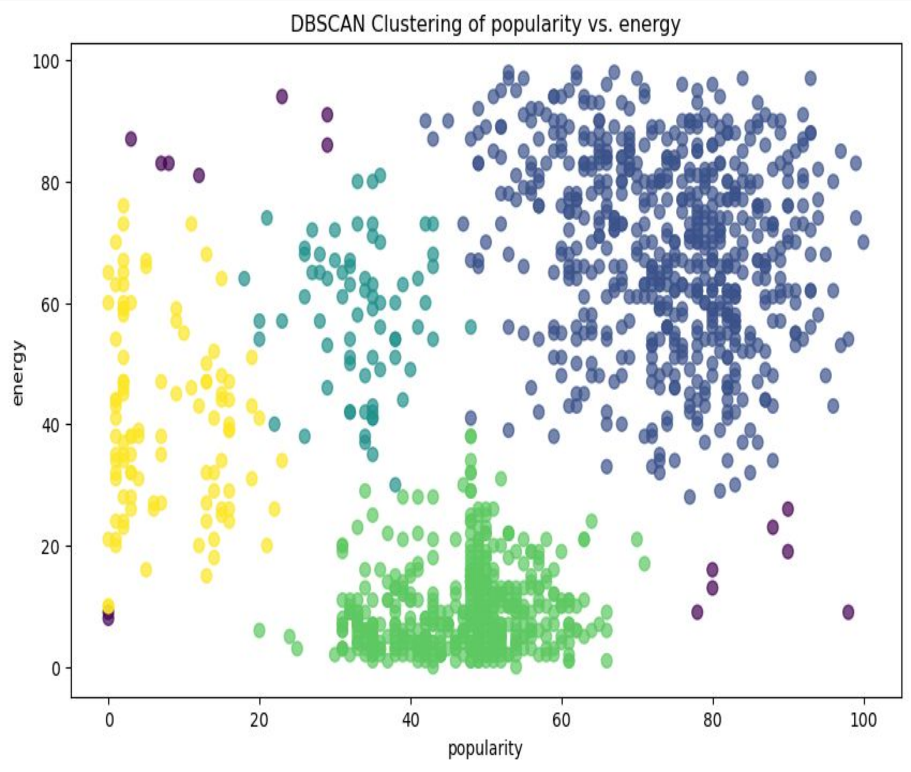




DBSCAN

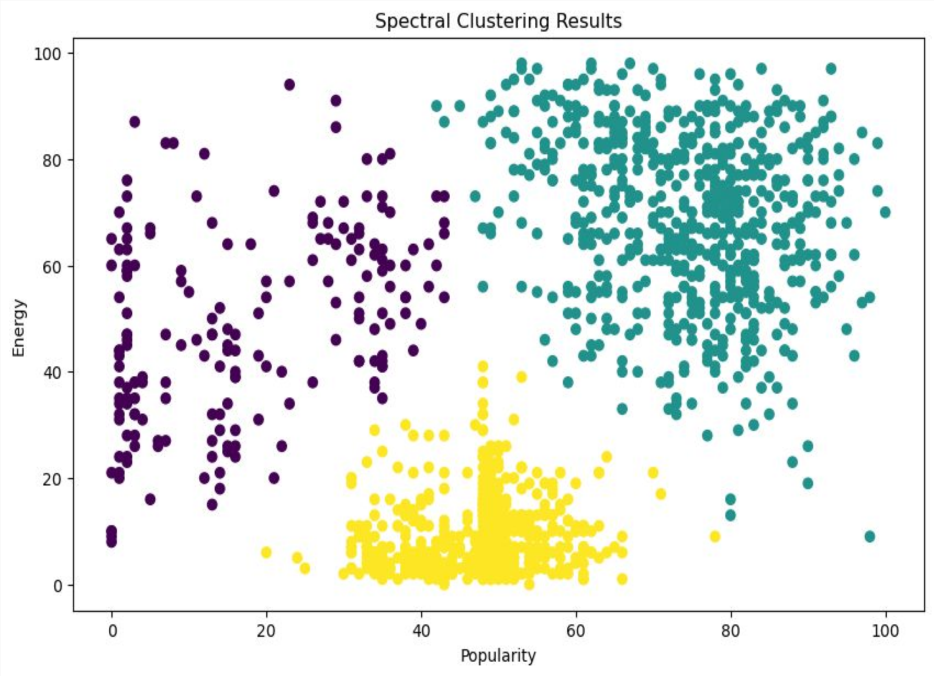
After multiple iterations the best model I could get out of DBSCAN was at eps = .39, minimum samples = 20.

The DBSCAN ended up having a silhouette score of 56%. It did however end up having more clusters than the other models. Due to the closeness of the clusters it was having trouble figuring out the clear distinctions.



Spectral Clustering.

Now that we knew the ideal amount of clusters for model performance I decided to try a model that had a different distance metric. The Spectral clustering model focuses on graph distance instead of point distance. If you rounded to the nearest whole number this tied the k-means silhouette score of 61%.



Streamlit.

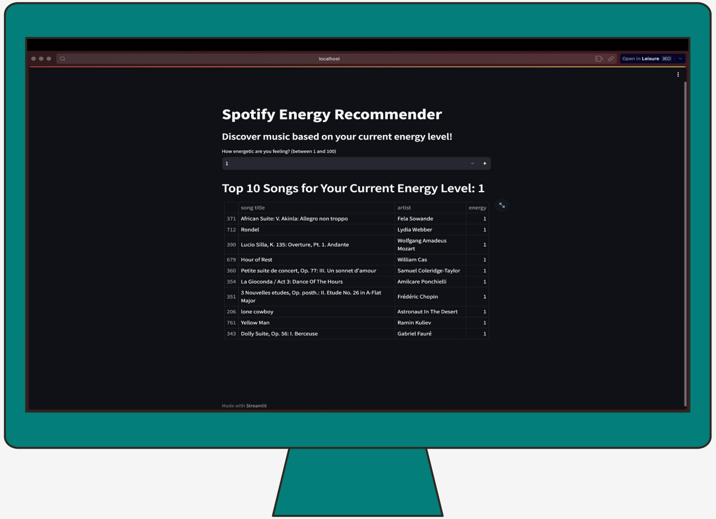


Tableau Public link below.

https://public.tableau.com/app/profile/robert.huey7416/viz/Projectfinal_16950725212800/Energy?publish=yes

Conclusion.

The recommender system worked well in assigning recommendations based off the clusters we received from energy and popularity. I would like in the future to focus more on the playlists gathered to create more distinct clusters to create a more accurate system.

Thanks!

Do you have any questions?
roberthuey94@gmail.com

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution