Spotify Playlist Clustering using K-Means Algorithm

Yunidh Rawal

INTRODUCTION

Background

In today's digital age, music streaming platforms like Spotify have become the go-to place for music listeners worldwide as they offer vast libraries of songs for users to explore and enjoy. With millions of songs available at their fingertips, users often face the challenge of discovering new music that aligns with their preferences. They can also struggle with organizing their music easily and efficiently. To address these challenges, Spotify continuously explores innovative approaches to enhance user experience and engagement.

Objective

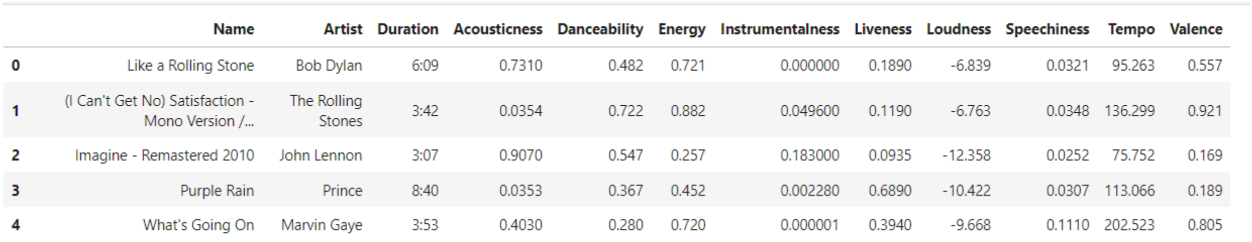
The primary objective of this study is to:

- Leverage K-means clustering algorithm to categorize songs in a Spotify playlist based on their attributes.
- Uncover patterns and similarities among songs, ultimately facilitating personalized recommendations and playlist curation.

METHODOLOGY

Overview of Data

Before diving into the clustering process, it's essential to understand the structure of the dataset. The dataset utilized in this analysis comprises attributes extracted from songs within a Spotify playlist. These attributes include numerical features such as Acousticness, Danceability and other relevant characteristics. By examining the dataset's dimensions and distributions, we gain insights into the nature of the data and can make informed decisions regarding pre-processing and analysis techniques.



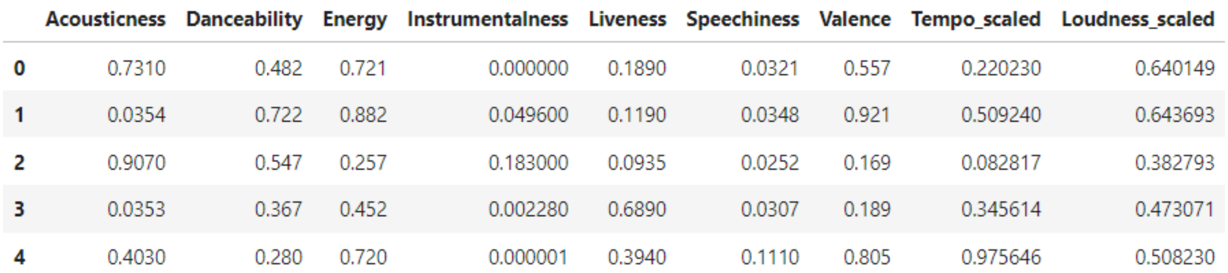
	Name	Artist	Duration	Acousticness	Danceability	Energy	Instrumentalness	Liveness	Loudness	Speechiness	Tempo	Valence
0	Like a Rolling Stone	Bob Dylan	6:09	0.7310	0.482	0.721	0.000000	0.1890	-6.839	0.0321	95.263	0.557
1	(I Can't Get No) Satisfaction - Mono Version /...	The Rolling Stones	3:42	0.0354	0.722	0.882	0.049600	0.1190	-6.763	0.0348	136.299	0.921
2	Imagine - Remastered 2010	John Lennon	3:07	0.9070	0.547	0.257	0.183000	0.0935	-12.358	0.0252	75.752	0.169
3	Purple Rain	Prince	8:40	0.0353	0.367	0.452	0.002280	0.6890	-10.422	0.0307	113.066	0.189
4	What's Going On	Marvin Gaye	3:53	0.4030	0.280	0.720	0.000001	0.3940	-9.668	0.1110	202.523	0.805

Dataframe before pre-processing

Data Pre-processing

The dataset was loaded into a **Pandas** DataFrame, and irrelevant columns such as song names and artist names were dropped to focus solely on the numerical attributes.

The tempo and loudness attributes were scaled using **MinMaxScaler( )** function from sk-learn library to ensure that all features contribute equally to the clustering process.

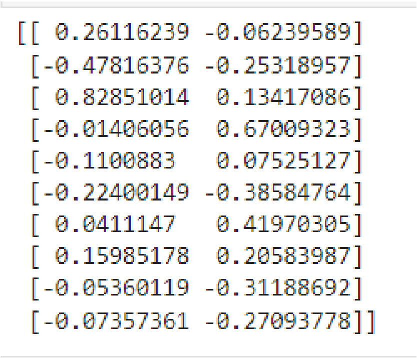


	Acousticness	Danceability	Energy	Instrumentalness	Liveness	Speechiness	Valence	Tempo_scaled	Loudness_scaled
0	0.7310	0.482	0.721	0.000000	0.1890	0.0321	0.557	0.220230	0.640149
1	0.0354	0.722	0.882	0.049600	0.1190	0.0348	0.921	0.509240	0.643693
2	0.9070	0.547	0.257	0.183000	0.0935	0.0252	0.169	0.082817	0.382793
3	0.0353	0.367	0.452	0.002280	0.6890	0.0307	0.189	0.345614	0.473071
4	0.4030	0.280	0.720	0.000001	0.3940	0.1110	0.805	0.975646	0.508230

Data after pre-processing

Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset to two dimensions while preserving the variance as much as possible.



```
[[ 0.26116239 -0.06239589]
 [-0.47816376 -0.25318957]
 [ 0.82851014  0.13417086]
 [-0.01406056  0.67009323]
 [-0.1100883   0.07525127]
 [-0.22400149 -0.38584764]
 [ 0.0411147   0.41970305]
 [ 0.15985178  0.20583987]
 [-0.05360119 -0.31188692]
 [-0.07357361 -0.27093778]]
```

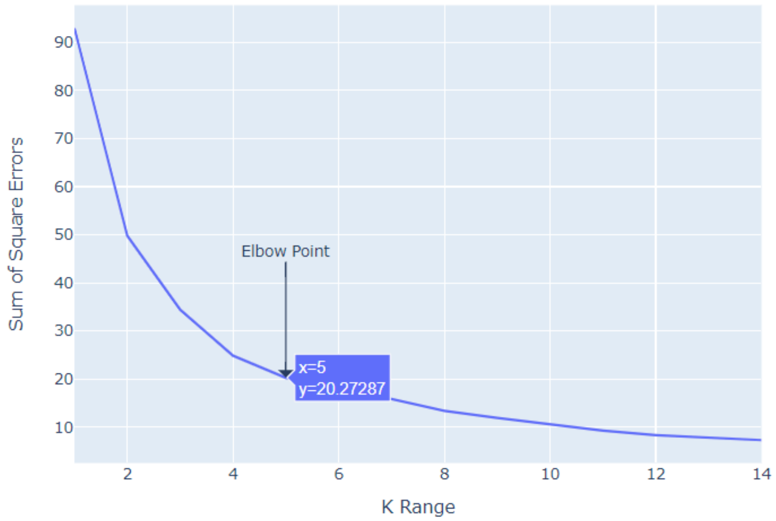
Dataframe reduced to 2 dimensions

Deciding no. of Clusters

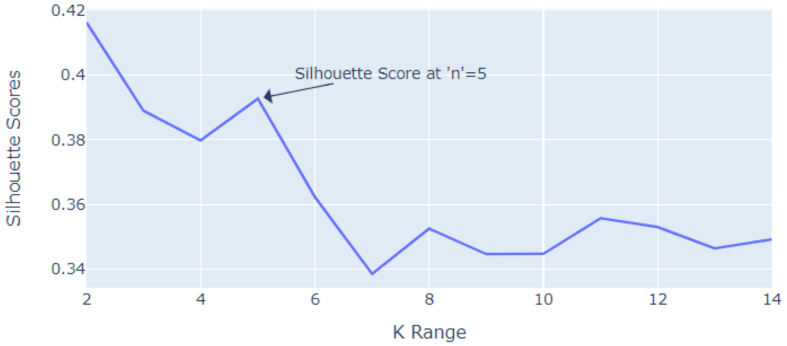
Choosing the correct no. of clusters is a vital step for accurate execution of k-means algorithm.

The two steps taken were:

1. **Plotting Elbow point with Sum of Squared Errors (SSE) :**



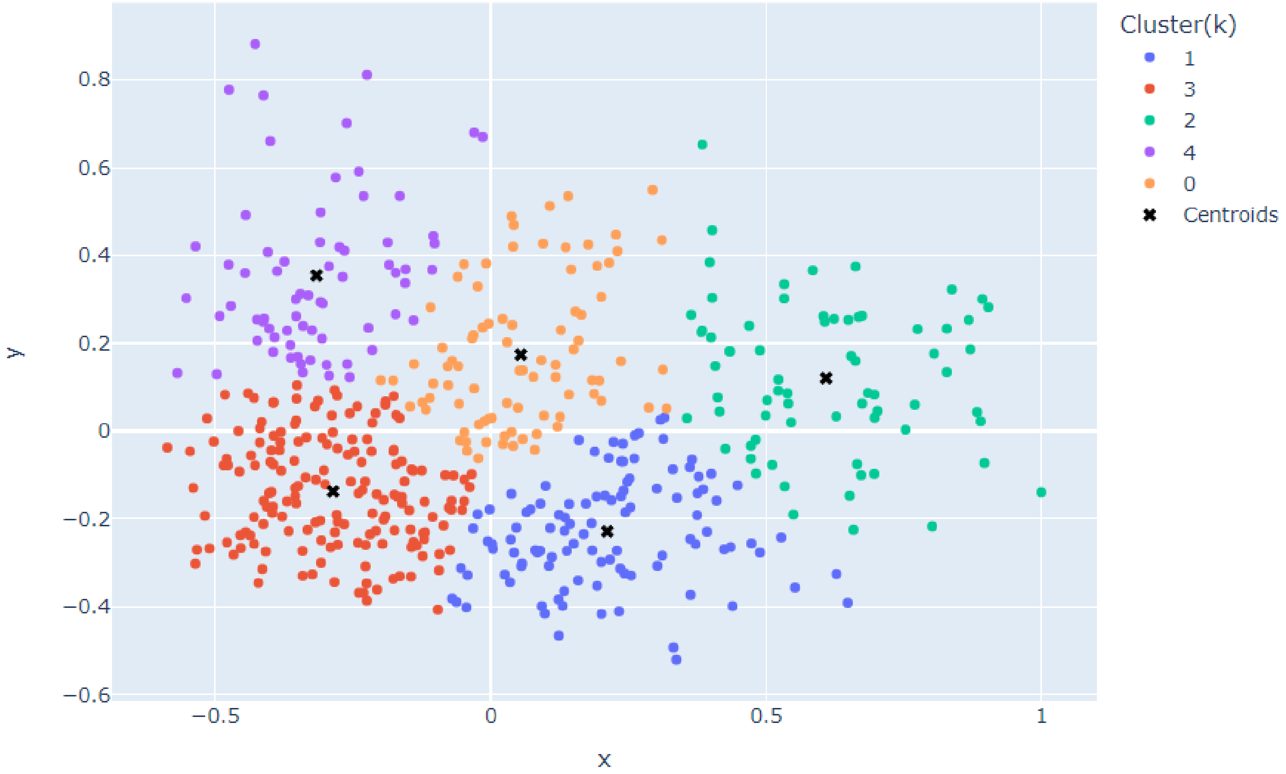
2. **Plotting Silhouette Scores to verify:**



These plots were used to determine and verify the ideal no. of clusters that playlist should be divided into, which is **5**.

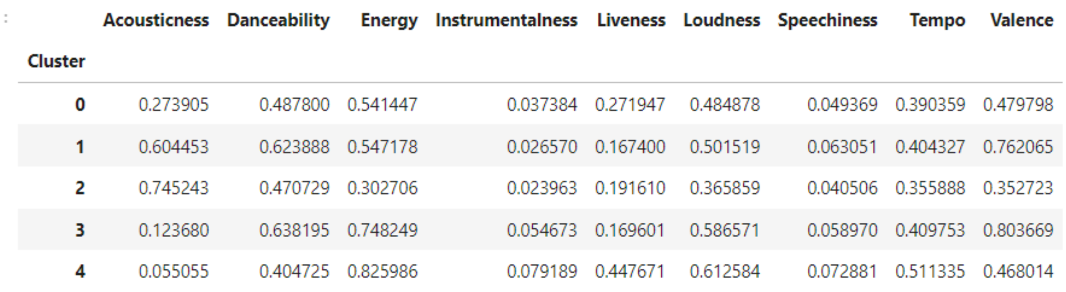
Clustering with K-Means

With the pre-processed and scaled data, the k-means clustering algorithm is applied to group the songs into clusters based on their shared attributes.



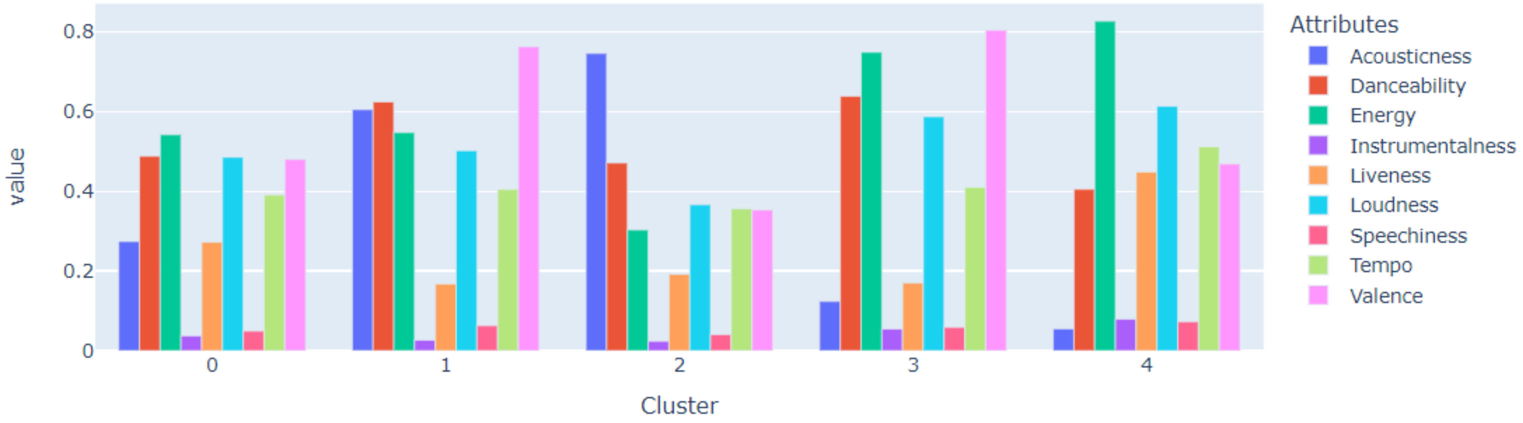
Visualization of the Clusters using Scatterplot

Finally, to visualize the characteristics of the tracks, we group the songs according to their clusters then find the mean of attributes in each cluster.



	Acousticness	Danceability	Energy	Instrumentalness	Liveness	Loudness	Speechiness	Tempo	Valence
Cluster									
0	0.273905	0.487800	0.541447	0.037384	0.271947	0.484878	0.049369	0.390359	0.479798
1	0.604453	0.623888	0.547178	0.026570	0.167400	0.501519	0.063051	0.404327	0.762065
2	0.745243	0.470729	0.302706	0.023963	0.191610	0.365859	0.040506	0.355888	0.352723
3	0.123680	0.638195	0.748249	0.054673	0.169601	0.586571	0.058970	0.409753	0.803669
4	0.055055	0.404725	0.825986	0.079189	0.447671	0.612584	0.072881	0.511335	0.468014

Table of means grouped by Cluster



Bar graph of mean attributes in each Cluster

For this specific playlist for example,

- We can label **Cluster 1** as the ‘Positive Tunes’ as its most significant attributes are high Valence (Positivity), Danceability and Acousticness.
- We can label **Cluster 2** as ‘Moody Tunes’ sub-playlist as its most significant attributes are high Acousticness but low Valence.
- We can label **Cluster 4** as ‘Hard Rock’ as its most significant attributes are high Energy and Loudness but almost very little Acousticness.

This analysis provides insights into the distinct attributes and patterns associated with each cluster.

Result

Recommendation based on findings:

Hence, the use of the K-means algorithm has many useful implications for Spotify in enhancing user experience and engagement.

By leveraging the data from clustering analysis, Spotify can:

- Improve personalized recommendations, playlist curation, and discoverability of new music for its users.
- Categorize songs to create curated playlists tailored to specific musical preferences, thereby increasing user satisfaction and retention.
- Provide a new feature to organize users' large playlists or liked songs into sub-playlists which group similar songs , making the process of accessing music according to their current mood much easier.

CONCLUSION

In summary, the findings from this study highlight the effectiveness of utilizing clustering techniques to analyze and categorize songs based on attributes in Spotify. By understanding the underlying characteristics of songs and patterns in playlists, Spotify can further optimize its music recommendation algorithms and enhance the overall user experience on the platform.