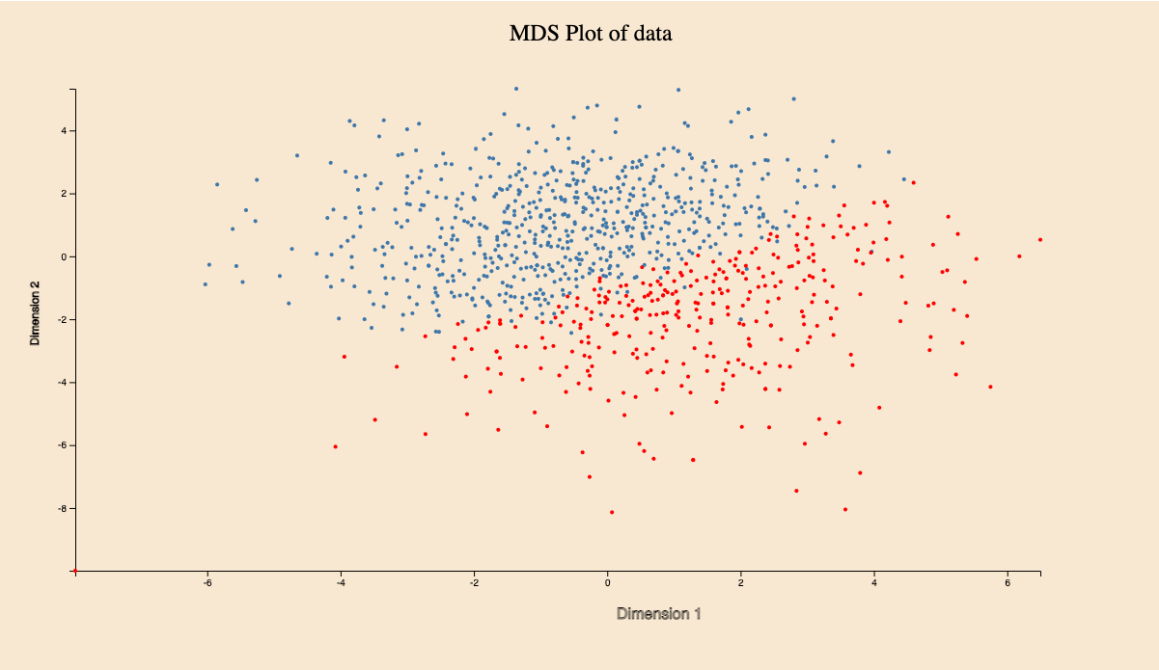
Dataset Source:
https://www.kaggle.com/datasets/muhmores/spotify-top-100-songs-of-20152019?select=Spotify+2010+-+2019+Top+100+Songs.xlsx

Attributes:

Artist	Skipped	Song's artist
Genre	Skipped	Genre of song
Released	Skipped	Year the song was released
Tempo		Beats Per Minute - The tempo of the song
Energy		How energetic the song is
Danceability		How easy it is to dance to the song
Intensity		Decibel - How loud the song is
Live Likelihood		How likely the song is a live recording
Positiveness		How positive the mood of the song is
Duration		Duration of the song
Acoustic		How acoustic the song is
Speech Focus		The more the song is focused on spoken word
Popularity		Popularity of the song (not a ranking)
Top Year	Skipped	Year the song was a top hit
Artist Type	Skipped	Tells if artist is solo, duo, trio, or a band

MDS Plot of Data

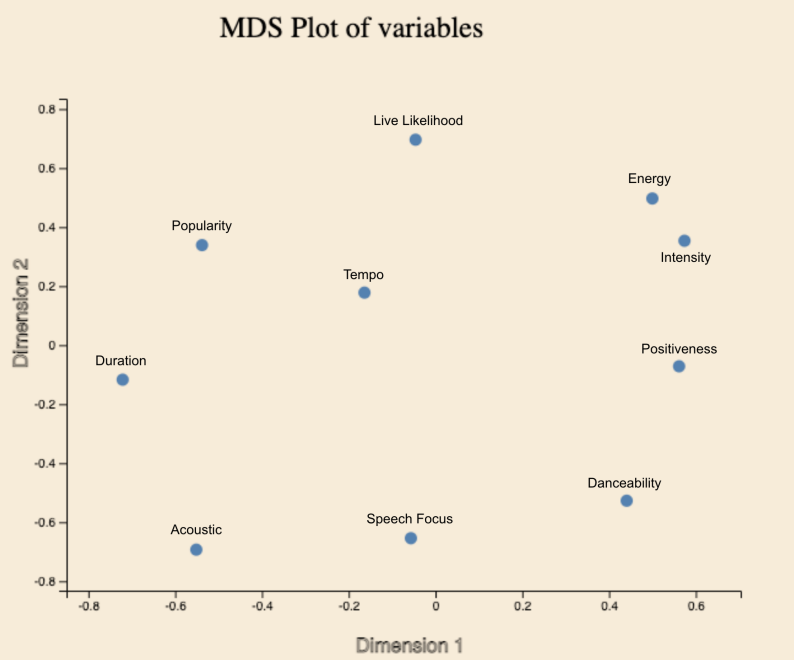
The MDS plot for data attempts to maintain the original distances between individual data points, while plotting them in 2D. The most interesting thing here is when the MDS plot is



colored using the k-means clustering labels, it does not really demonstrate any clear delineation between the two clusters. The distances between the data points are all over the place, and that makes sense. Because the ten numerical attributes used in these plots all contribute more or less uniformly to the popularity of a song. There are no predominantly consistent differences in the values. We can have a low tempo, low energy, not very danceable, live performance that became a popular track. And we can also have a high tempo, high energy, not very danceable, live performance that also became popular. And then we can also have a high tempo, but low energy, not very danceable, studio track that became popular. The point, there are no predominantly overwhelming correlations between any pair of attributes that governs the clustering.

MDS Plot of Variables

The MDS plot for variables visualizes the attributes or dimensions of the dataset in 2D by attempting to maintain the relative distances between the dimensions. The distance metric used in this case is the correlation distance which is computed as (1 - |correlation|). Therefore the distances between the attributes are representative of how commonly they function. There are a few interesting observations in this graph.



Clicking on a point will move the corresponding variable to the end of the parallel coordinates plot.

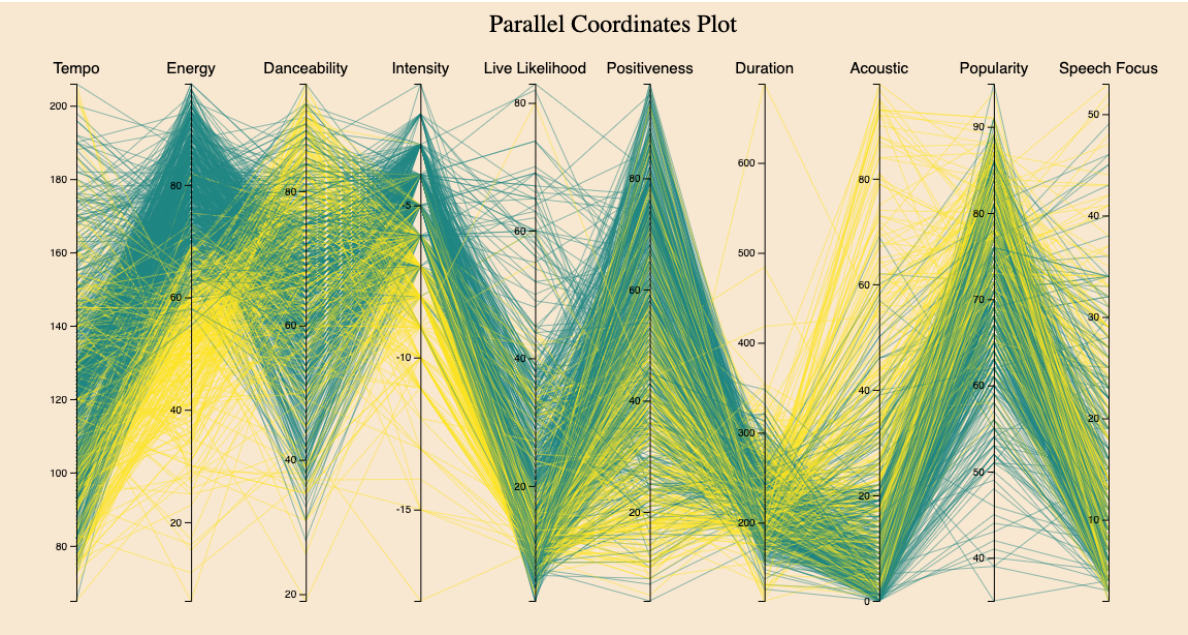
Firstly, as expected, we can see that energy and intensity have a negligible distance between them and therefore function hand-in-hand in terms of value. So a track that is intense is also perceived as high energy, and vice versa.

Secondly, contrary to what we might normally expect, energy does not go hand-in-hand with danceability. You would trivially expect a danceable track to be high-energy almost all the time. Yet, we can see that this is not the case. There are obviously many low-energy but danceable tracks, and also high-energy but not-very-danceable tracks. This is in contradiction with what one might normally expect to find.

Thirdly, we can see that tempo of a song is not in high correlation with its danceability. This is also contrarian to what you would expect to find from this dataset. That means popular danceable songs are not necessarily always high-tempo.

Parallel Coordinates Plot

This plot presents the ten numerical attributes of the dataset in parallel coordinate axes, with polylines connecting the attribute value of data across all ten axes. The polylines are colored using the same k-means clustering labels as used in the data MDS plot. That the clustering more or less accurately captures the differences in values for the attributes is evident from the



colors of the polylines. We can see that for most attributes, the cluster of yellow lines is sparsely overlapped with the blue lines. This means that the clustering presents a more or less reasonable picture of which data points are likely to be grouped together. We can see the very strong delineation between the clusters for energy, intensity, and acoustic measure. The other dimensions also have a less but more or less visible delineation between the clusters.

Another very interesting observation is that the clusters flip positions when you compare energy and intensity with acoustic measurements. The acoustic dimension essentially tells us how much more the use of electric instruments was present in the song as opposed to acoustic instruments. A high acoustic value indicates modest use of electric instruments. You can see that the vast majority of tracks that are high energy and high intensity have low acoustic measures, indicating that the use of electric instruments is directly related to a track being high energy and high intensity in nature.