Origins of Music: Looking at links between music around the globe using clustering

Shushruth Kallutla

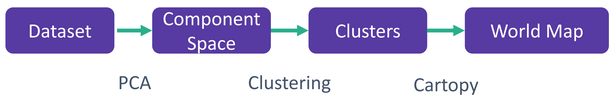
Introduction

The Geographical Original of Music Dataset consists of audio features as well as location data of 1059 music tracks sourced from 33 countries. Originally Linear Regression and Classification techniques were applied to use the audio features to predict geographic origin. In this study, we will use clustering techniques to gauge similarities in features of music from different regions of the world. Clustering may allow us to see trends and links that may not seem apparent at first. These trends could be starting point for researchers studying the evolution of music in these regions. We will use Cartopy, a cartography python module, to better visualize the geographic data which will help us spot relationships between the datapoints.

Methods

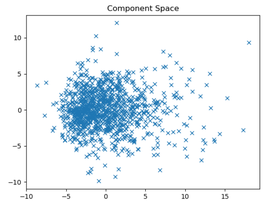
The Dataset was sourced from UCI ML Repository.
It has 1059 instances and 70 attributes (68 audio features along with latitude and longitude).

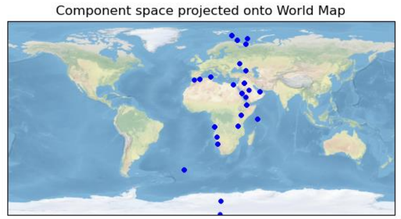
PCA was first used to simply the data.
K means clustering using different cluster numbers (2,4 and 8) was performed on the simplified dataset. The k means algorithm was iterated 200 times.
Clustered data was then projected on to a world map using the Cartopy module.



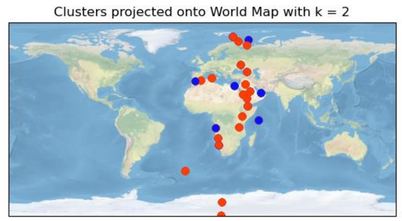
Results

Component Space

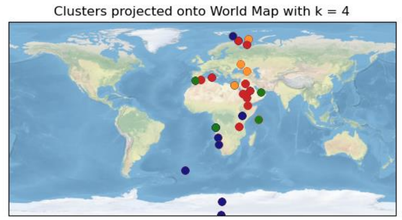




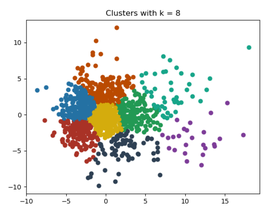
K = 2

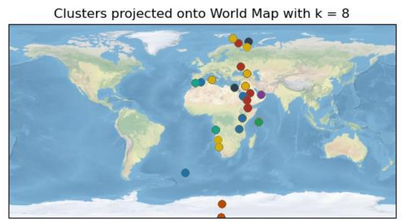


K = 4



K = 8





Discussion

We can see that the component space does not have any apparent clusters and hence does not have an apparent k value that would best fit the data. When we plot the space on a map we can see that the music in the dataset comes primarily from African and Middle Eastern countries.

We must note that as the data is sourced from only a few countries and the dataset only records the location of the country's capital and not the exact location of the music source. This may lead to points overlapping . Hence the cluster maps we see may not be a perfect representation of the geographic distribution of the clusters. Future analysis could find a better way of representing the overlapping data.

In the cluster map with two clusters, the map does not seem to have an obvious trend. The red cluster seems to be spread all over the map while the blue cluster seems to cover Northern Africa. This could imply that North African Music that make up this cluster share some musical features. This could further imply a common origin and a history of cultural contact between these countries.

In the cluster map with four clusters, we see that some clusters seem to be localized to certain regions. Such as the yellow cluster seems to be localized around Europe and the green cluster seems to be localized around Africa. The Blue cluster is generally localized towards the south with one exception. The red cluster seems to be localized in the north. These cluster suggesting relationships in the music of those regions mapped by each cluster. This map seems to have the most apparent trends amongst the maps we've plotted and hence k = 4 clusters could be a candidate for best fit.

In the cluster map with 8 clusters, there doesn't seem to be any identifiable trend. Some clusters such seem to be localized such as brown in the south. But others seem more spread out, such as yellow.

The clusters that seem to be spread over the globe indicate that our analysis method seems to group together vastly different datapoints. It is more likely that our analysis method is flawed rather than there being a relationship between the music of these regions.