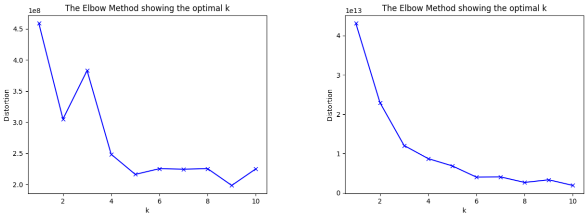
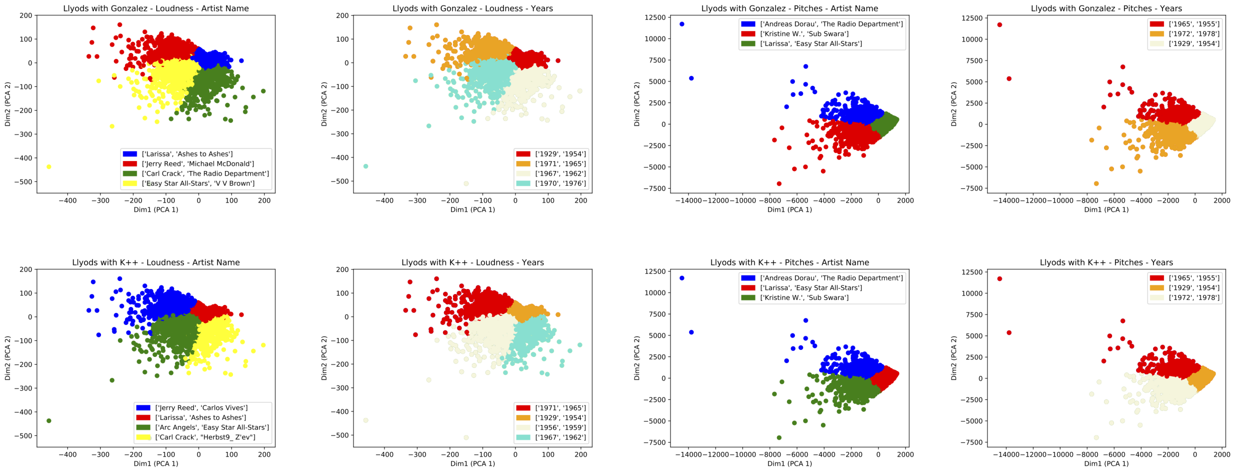
Intermediate Report

Angel Dhungana

For the Intermediate report, I ran the Assignment Clustering to cluster the songs based on their loudness and pitches dataset. Then, using the cluster I took out top 2 most common artist names and years just to see if there exists a relationship between the artists and the year.
For clustering, every row represented a track, and each column represented an audio feature. For this report, I just ran Llyod's Algorithm with K++ and Gonzalez algorithm serving as the the provider for the initial centers.



To get the optimal k for K-Means Clustering, I calculated the sum squared error for each k on the dataset, and plotted it. The k where the plot makes an elbow arc is the optimal k. For loudness dataset, the optimal k was 4 and the for pitches dataset, the optimal k was 3 respectively.



- The loudness and pitches dataset had a large range of columns. Hence, for the visualization purpose I used the PCA to reduce the dataset dimension to 2, and plotted the graph based on it.
- From the graph, we can see that the Gonzalez and K++ don't have any significant differences. In Pitches, there doesn't seem to be any difference but in Loudness, we can see some of the outlier tracks are being classified into different clusters.
- Pitches vs Loudness
  - Only one artist name came up as the most common in both dataset clusters. And its 'Larissa'. Based on the Loudness cluster, "Larissa' is most similar to 'Ashes to Ashes' and based on Pitches cluster, 'Larissa' is most similar to 'Easy Star All-Stars'.
  - We can see that '1929' and '1954' appear both on Loudness as well as Pitches graph as the top 2 most common. Hence, we can conclude that the songs of the year 1929 and 1954 were the most similar based both on their loudness and pitches.
  - We can also see that based on the pitches, the year '1965' was most similar to '1955' while with loudness, the year '1971' was most similar to '1965'.