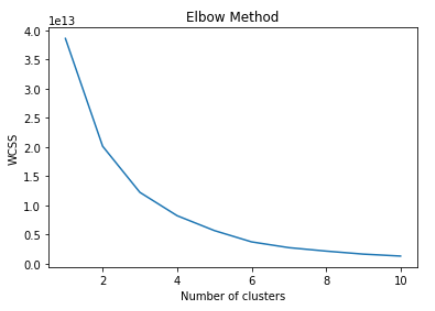
Clustering Analysis

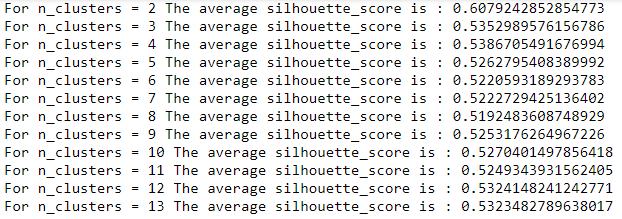
The purpose of implementing a clustering models is to assist with exploratory data analysis and to give clues about the structure of the data. For this project, the team used clustering to give an idea of how many genre should be used in our analysis. The original Spotify data set has over a hundred different genres/sub-genres, which was much too broad for this project. We had to find a way to narrow our focus to a few, select genres. How might the genres be selected and how many should be used? We looked to K-Means Clustering Analysis for some insights.

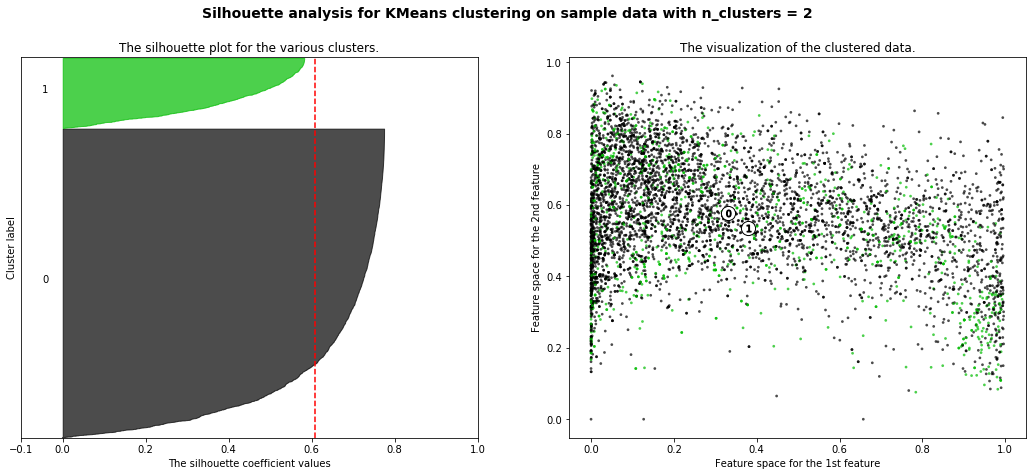
Methodology

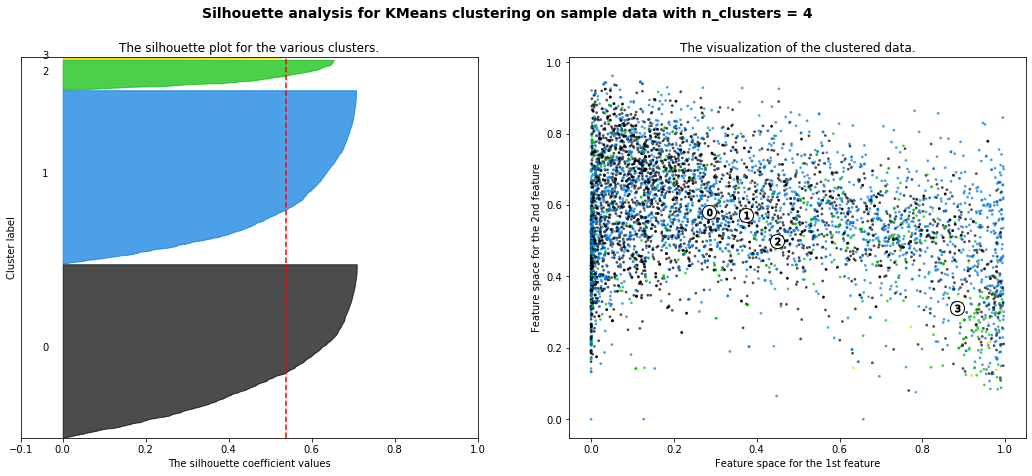
The dataset was parred down to just the 10 musical elements provided by Spotify to accommodate the K-Means algorithm. The K-Means algorithm only accepts numerical information. Next the dataset was fed into the K-Means analysis algorithm and then the following elbow plot was generated. The elbow plot is used to determine the optimal number of clusters in k-means clustering. The bend in the “elbow” is the most optimal cluster.

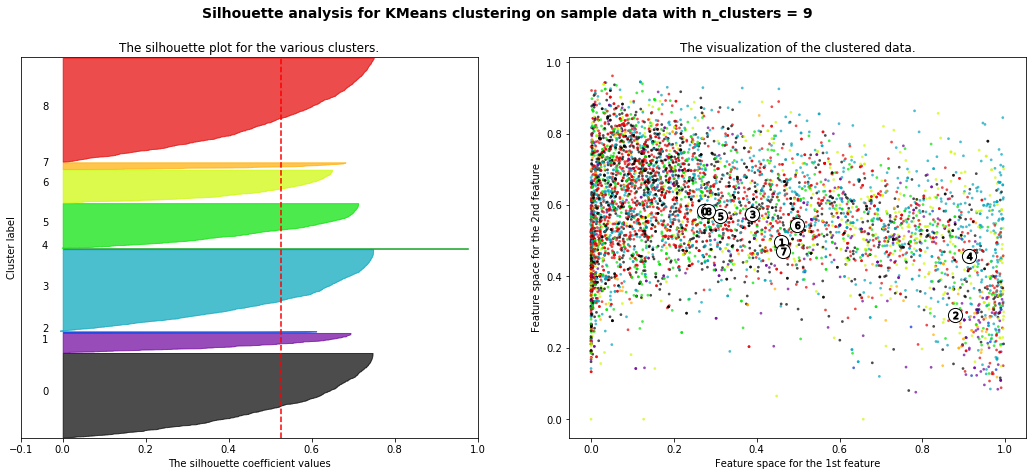


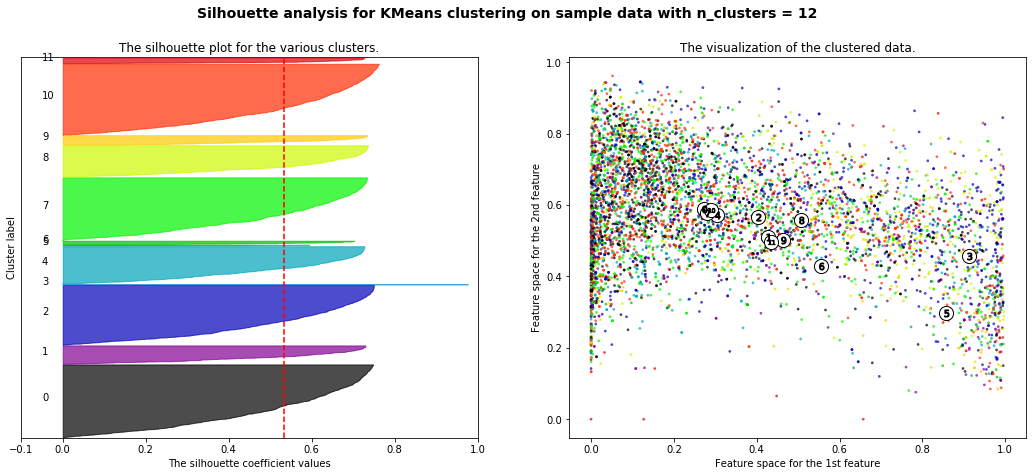
According to the K-Means analysis/elbow chart, 4 to 6 genres should have been selected. After review, the team felt that 6 genres was too small and that perhaps additional analysis was in order. Next a Silhouette Analysis was done. A Silhouette analysis is method that used to determine the separation distance between K-Means clusters. Silhouette analysis can be used to choose an optimal number of clusters. A set of Silhouette score and Silhouette plots were generated. A Silhouette score is between (-1 and 1). The closer to 1, the better the score. Negative values indicate that values are in the wrong cluster. For the plots, similar sized plots with clear separation, and no negative values are most desirable. Of the plots with the higher numbers of genres, 12 clusters was the most desirable set.











Future work.

Our K-Means analysis could used to make predictions in tandem with other Machine Learning methods as a verification other methods finding. One short coming of this implementation is that while it tells that we have 12 clusters, it does not tell which songs go into which category. A future implementation would allow predictions to be made to which genres a song might fit in.