- The insights I gained from my EDA were plenty. I got a better understanding of the distribution of the features present in the data. Such as the fact that most of the genres were pop and the fact that sadness was a very common topic in the data. I also found how the distribution of genre and topic varied due to their length and age. The ‘age’ and ‘released_date’ columns were negatively correlated with a value of -1, which makes sense since the older the release date the older the song is. I also found that obscene and len columns were correlated with a value of 0.44.
- Determining which columns to drop or keep was fairly straightforward in this project due to the fact that the dataset was clean. I removed columns that kept track of artist name, track name and lyrics simply because this type of data was not going to be needed for clustering. Since release_date was extremely correlated with age I decided to remove it as well.
- In order to determine the optimal number of clusters I used the elbow method as well as the silhouette plot method. Both these methods show me that the optimal number of clusters was 6.
- I noticed that Jerry Lee Lewis’ “Your Cheating Heart” and Taste’s “Railway and Gun” were assigned the same cluster. This was surprising at first since they don't sound like similar songs but overall noticed that the themes within the lyrics were very similar. So I think the algorithm is picking up song themes and topics rather than the sound of the songs. This makes sense because we only have data on the themes of the songs rather than how they sound like.
- Based on what clusters were assigned to the test sample, for this user I would recommend songs that fall on the 3 and 4 clusters of the KMeans algorithm. These are songs that are usually newer and have a violence topic.