For the final project, I chose to work with the Million Song Dataset. This dataset consisted of several hundred thousand songs. Each song had 90 attributes and a release data. The goal was to classify songs by their release year. In order to do this, I used a KMeans unsupervised learning algorithm. I ran three experiments with the dataset.

The first experiment involved classifying songs based on their release year (as the original problem instructed). I ran the training data through the KMeans model to fit, and then predicted the test attributes. I converted the test attribute clusters to years and calculated the accuracy of the predictions. I got an accuracy of 8.52%. There were 89 total classes, meaning a completely random assignment should result in a roughly 1.1% accuracy, so the KMeans model clearly improved classification, but not to anything very impressive. Using the same predictions, however, I added some code to see whether the predicted year was within 3 years of the actual year. The model correctly predicted the year to within 3 years 41.99% of the time, which indicates to me that the model is at the very least on the right track.

For the second experiment, I reclassified all of the training and testing data into decades instead of years. Music is often associated with particular decades, so I reasoned the model might be able to better predict the decade of a song than the exact release year. I made a new KMeans model with only 10 clusters this time and repeated the experiment using decade classifications. The model correctly predicted the correct release decade of a song 58% of the time.

The final experiment involved running 12 different KMeans models on the data, each with a different number of means. See the Jupyter Notebook Pdf for more details on the means and the graph. Accuracy decreased as the number of means increased. This was to be expected, as it is far easier to predict which half of a century a song was released (2 means) than the year (89 means).