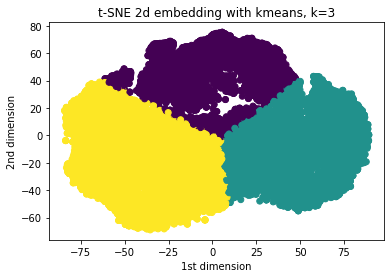
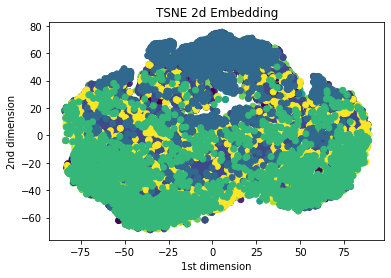
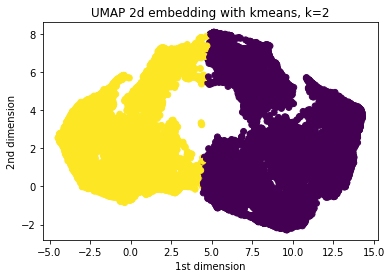
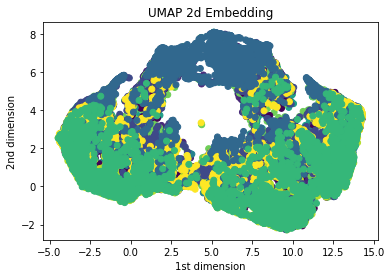
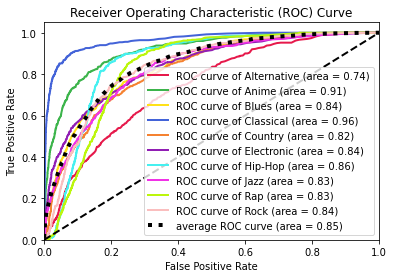
In order to solve many of the challenges, I needed to do some preprocessing after loading the data into a data frame. Since the cases of missing data were relatively few, I decided to impute with the overall mean of the column. I briefly considered imputing with the mean of the column only within the same genre, but thought it might artificially make classification better. For the categorical variables like “key,” “mode,” and “genre” I encoded with numbers. Before dimension reduction, I normalized the numerical variables, except for the categorical ones I encoded previously.

For dimension reduction I tried three different methods: t-SNE with 2 dimensions, with 3 dimensions, and UMAP with 2 dimensions. I used the 2 dimensional embeddings for visualization in the graphs below. The graphs on the left show the embedding with the real labels, while the graphs on the right show the embedding after k-means clustering using the ideal number of clusters as found by the silhouette method. In both 2D embeddings, the real labels don’t show very good separation at all, with maybe the UMAP embedding being slightly better. This is reflected in the clustering step, as k-means cannot find the right number of clusters or even create very separated or clear cut clusters.





I used the t-SNE 3D embedding for the classification step, as I found it outperformed the other two embedding methods. From the t-SNE 3D embedding, I sampled 500 random songs from each genre for the test set, and kept the remaining data for training. I classified using a random forest classifier with n\_estimators = 1000, and scored it using multiclass One vs Rest ROC. This method compares the probability of classifying to a specific class to the probability of classifying to the rest of the classes, creating 10 different ROC scores, as there are 10 possible classes. The average of these ROC scores is then the overall ROC score, which solves the challenge of multiclass classification and ensures the classifier is scored positively only if the predicted class matches the actual class. The results of the individual class classifications and the average are plotted in the graph below. I attribute the success of the classification to the dimension reduction method of t-SNE with 3 dimensions, as it greatly boosted the AUROC compared to the others.



My final AUROC from the test set was 0.85. But for some reason, my random seed did not result in consistent sampling for the train test split, or consistent behavior from the models. But I was consistently getting results between 0.84-0.85.