Project Group 3 - Milestone Report:

For a preliminary exploration of the data, we focused on the role of the features when classifying the data. This dataset has 57 features for each observation (excluding filename, number of samples, and label). We wanted to know which of these features were most important in classifying the data and where they are derived. Most of the features from our dataset are taken from the spectrum of the audio files. An example is shown in Figure 4 of the Appendix, where each vertical slice shows the intensity of the spectral frequencies at that particular moment in time. Features were derived in this dataset from the spectrum information, as well as some Time-Domain characteristics to provide a robust representation for each genre. In order to determine which features carried the most importance for classification, we started by using a simple Naive Bayes classifier, and verified the results using K-folds cross-validation with 5 splits. While Naive Bayes is not the most robust classifier, it should still allow us to gain a baseline understanding of the data. We ran this classifier between all genres, using all features. Doing this, we obtained an accuracy of 53.4%.

Next, we decided to observe how well the classifier worked when using only a single feature to make its decision. Running the classifier for each feature, we found that the classifier only obtained an average accuracy of 19.4% when using only one feature. While this is a significant drop in accuracy, it is an expected one, as we are using much less data than before. The feature that resulted in the best classifier performance was the Spectral Bandwidth mean, which directly corresponds to the range of frequencies in the song, while the worst-performing feature was the Perceptrual mean, which corresponds to the rhythmic/percussive components of the audio.

Another topic of our preliminary investigation was how well the classifier could distinguish between pairs of genres. In order to investigate this, we compared the 10 different genres pairwise using the classifier, allowing us to get an accuracy score for each pair of genres. The higher the accuracy score, the better the classifier was at distinguishing between them. We first ran this pairwise classifier using all features. On average, the classifier had an accuracy of 88.1%. It performed best when classifying between Classical and Metal (99.5% accuracy), and worst when trying to distinguish between Blues and Rock (74.0% accuracy).

We then decided to see whether the classifier could match this performance while only using one feature. Using only the spectral bandwidth mean (the feature that we had previously determined led to the best performance), the classifier had an average accuracy of 73.5%, and was able to reach an accuracy of 98.5% when distinguishing between Classical and Pop. (Between Classical and Metal, it reached an accuracy of 95.0%). While this accuracy is slightly lower than the previous example, it is still higher than expected, especially since we are only using one feature to classify the data. However, the worst-case scenarios for this classifier performed much worse than the previous classifier, as it was only able to distinguish between Hip-Hop and Disco with an accuracy of 44.0%. (Between Blues and Rock, it reached an accuracy of 65.0%.) Data for these classifiers is included in the appendix below as table 1.

A boxplot of the spectral bandwidth mean values for each genre (Figure 2) is also included below as illustration of these results. Observing this figure, we can see that some genres (such as Classical and Pop) do not overlap much, and so are relatively easy to distinguish between. Other genres (such as Disco and Hip-Hop, or Reggae and Rock), display large amounts of overlap, and so are difficult for this classifier to handle. These results indicate that while the spectral bandwidth mean does lead to good separability between certain genres, this feature on its own is not enough to build a consistent classifier.

To gain some simplicity to our data exploration methods, we used Principal Component Analysis to reduce the dimensionality of our features from 57 down to the 2 primary components. The feature values were first normalized using a MinMax scaling method to set them between 0 and 1, and then the principal components were extracted for each labeled genre entry and plotted in Figure 3 in the Appendix. We observed that genres such as Classical and Metal have their data points clustered in separate areas from one another, while the genre of Rock is scattered across the whole figure. The cluttering of the principal components provided us with more visual evidence that some genres from this dataset are more similar to some and different to others. To further our data examination, we plan to examine the K-Means Clustering of the data points to obtain the parameters for the maximum-likelihood estimates for each genre's cluster.

For the remainder of the project, we plan to utilize and experiment on our dataset with more techniques we've covered in class. From the experiments we've performed thus far, we've seen that some genres are easier to distinguish from one-another and others are more difficult (e.g., Classical music and Metal have clear separability from their features, and genres such as Disco and Hip-Hop). So, to further support these findings, we plan to visualize these scenarios with some linear discriminant analysis models showing either our best case of distinguishability between two genres, like Classical and Metal, and then the worst case, where LDA methods are not sufficient (i.e., Disco vs Hip-Hop). To complement our simplest case of classification we've performed using Naive Bayes, we'd also like to implement some Decision Tree classification methods. We've also explored some of the available research papers on the topic of Music Information Retrieval, where Music Genre Classification is a sub-category, and have found some resources that mention some features like the Mel-Frequency Cepstral Coefficients (MFCCs), are typically used for characterization of speech, and a limited number of these coefficients serve best for music genre classification. Our dataset was included with 40 values for mean and variance estimates, which we suspect is hurting our classification accuracy scores. We then plan to re-evaluate which MFCCs are the most crucial towards music genre classification, and reduce the number of MFCC features used. After doing so, we plan to draw some comparisons on the performance of our classification models, to see if our cited research findings will match up with improved accuracy. Lastly, to include one more linear classification model, we plan to examine the performance of a Linear SVN model on the dataset.

Milestone Report Appendix:

Classifier:	**Accuracy**		
	Average	**Best**	**Worst**
Classifying all genres, using all features	53.4%	-	-
Classifying all genres, using one feature	19.4%	27.9% (Spectral Bandwidth mean)	13.5% (Perceptrual mean)
Classifying pairs of genres, Using all features	88.1%	99.5% (Between Classical and Metal)	74.0% (Between Blues and Rock)
Classifying pairs of genres, using Spectral Bandwidth mean	73.5%	98.5% (Between Classical and Pop)	42.5% (Between Disco and Hip-Hop)
Classifying pairs of genres, using Perceptrual mean	56.7%	69.5% (Between Classical and Rock)	44.0% (Between Blues and Rock)

Table 1: Naive Bayes Classification Estimates

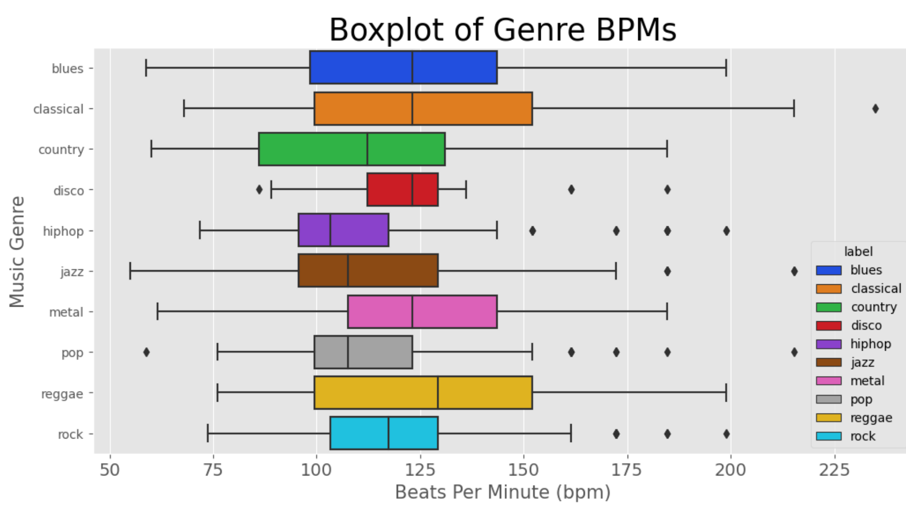

Figure 1: Box Plot of Tempo Distribution of Music Genres

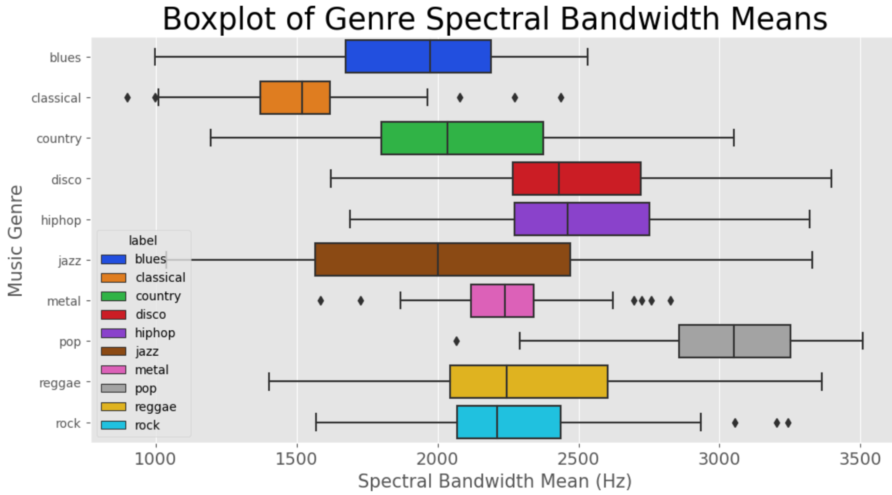

Figure 2: Box Plot of Spectral Bandwidth Mean Distribution of Music Genres

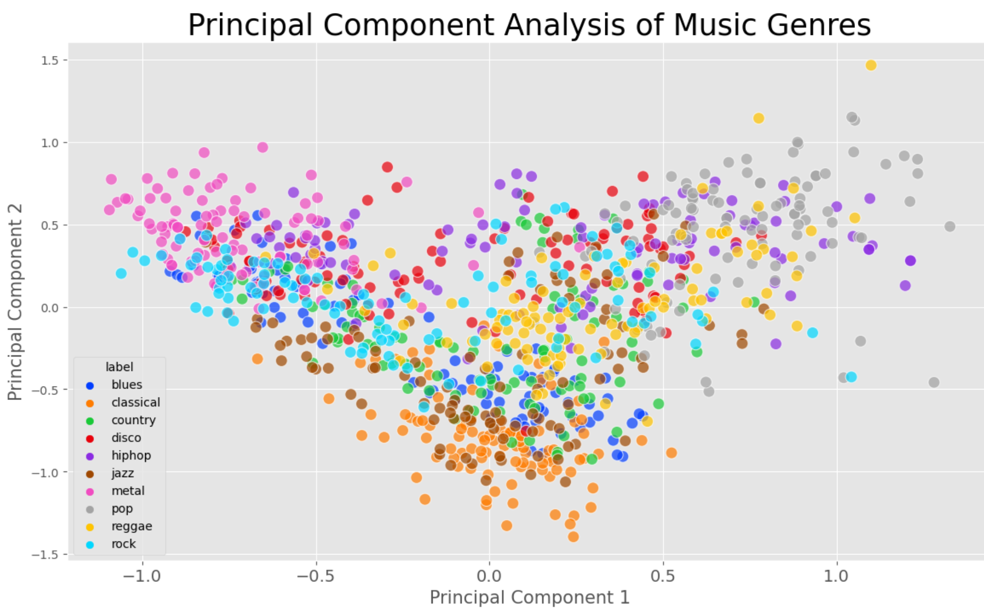

Figure 3: Principal Components of Music Genres

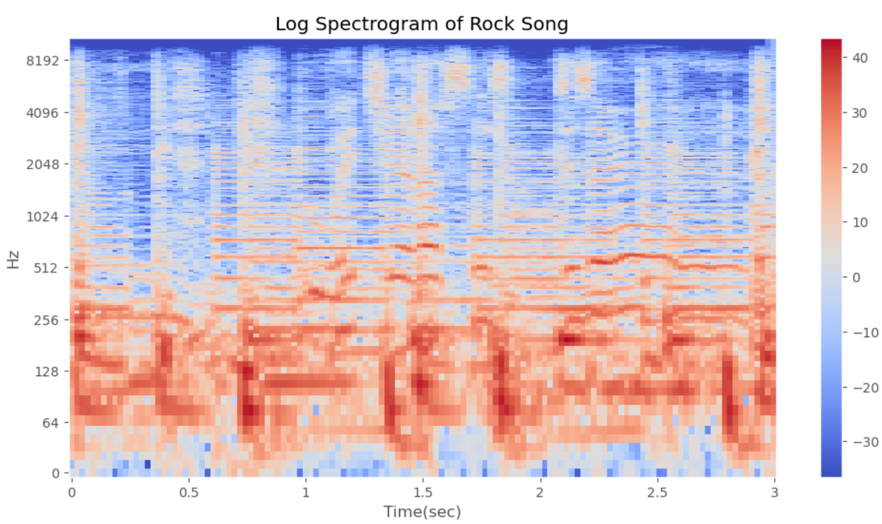

Figure 4: Log Spectrum - Intensity of Spectral Frequencies over Time