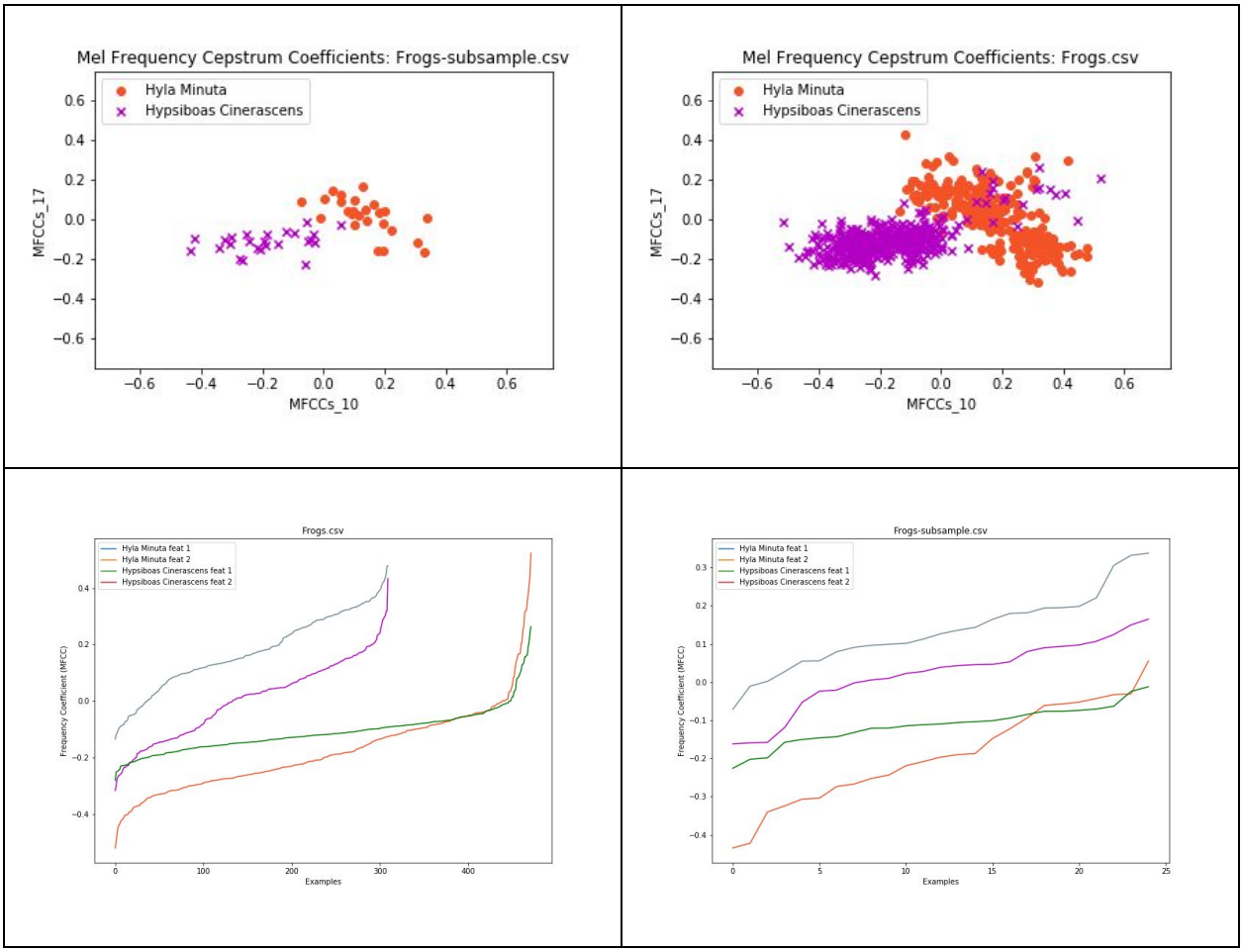
Mel Frequency Cepstrum Coefficients from recordings of Frog calls from UCI ML repository: https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29

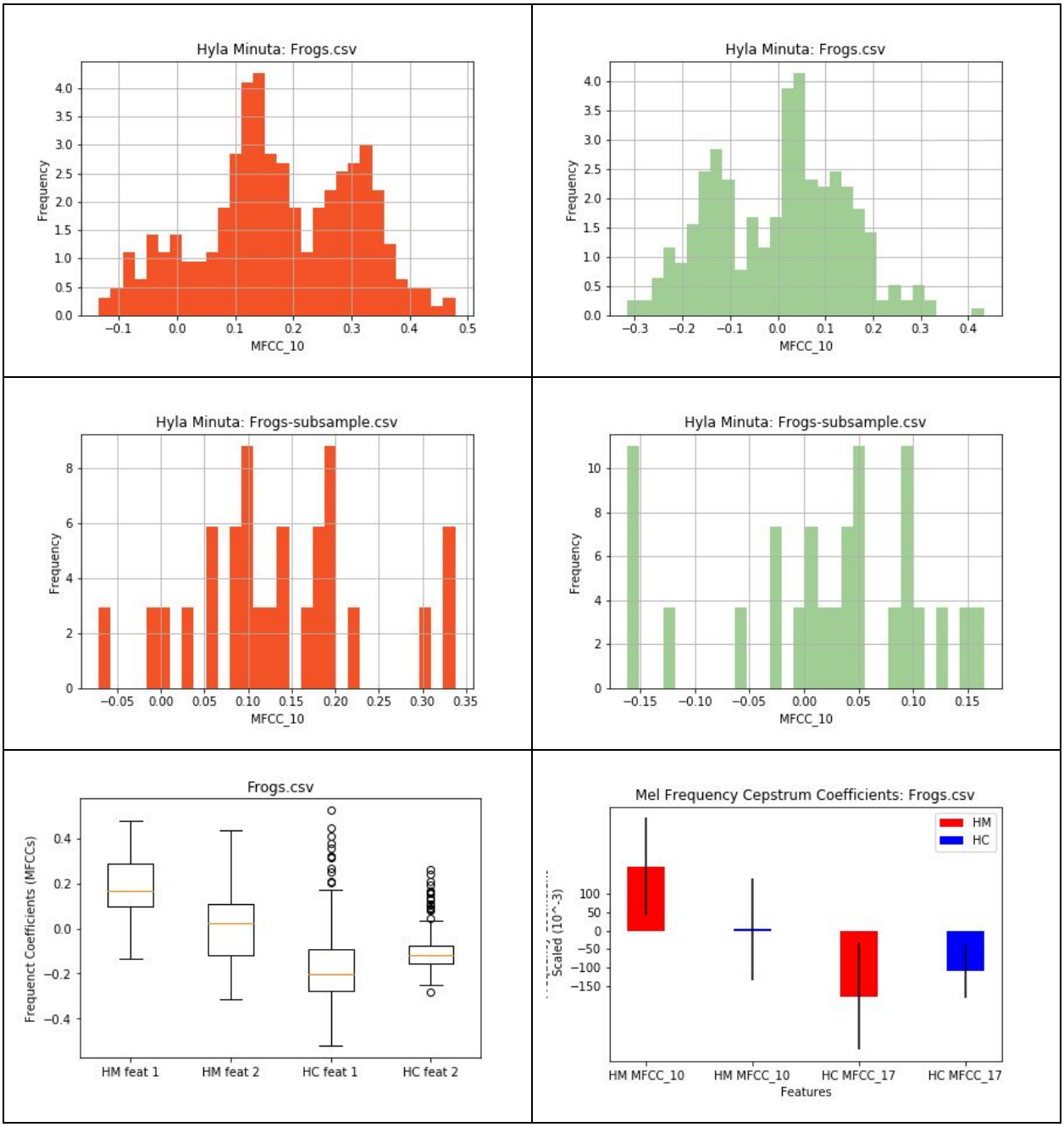


Above (row 1): Mel Frequency Cepstrum Coefficient scatter plots from the two datasets provided.
Above(row 2): Line graphs representing the range of frequencies observed in the dataset.
Below: Statistical information regarding the Frogs.csv and Frogs-subsample.csv datasets.

Dataset:	Frogs.csv	Frogs-subsample.csv
Mean: Class A	x1: 175.102, x2: 4.275	x1: 134.064, x2: 20.093
Mean: Class B	x1: 178.493, x2: -109.695	x1: -190.087, x2: -112.753
Standard Deviation Class A:	x1: 131.494, x2: 138.069	x1: 99.519, x2: 90.617
Standard Deviation Class B:	x1: 145.468, x2: 74.565	x1: 126.789, x2: 49.717

Covariance Matrices:	Covariance matrix of Hypsiboas Cinerascens: [[0.021206 0.00635566] [0.00635566 0.00557183]] Covariance matrix of Hyla Minuta: [[0.01734669 -0.01235759] [-0.01235759 0.01912491]]	Covariance matrix of Hypsiboas Cinerascens: [[0.0167455 0.00236565] [0.00236565 0.00257479]] Covariance matrix of Hyla Minuta: [[0.01031672 -0.00597868] [-0.00597868 0.0085536]]

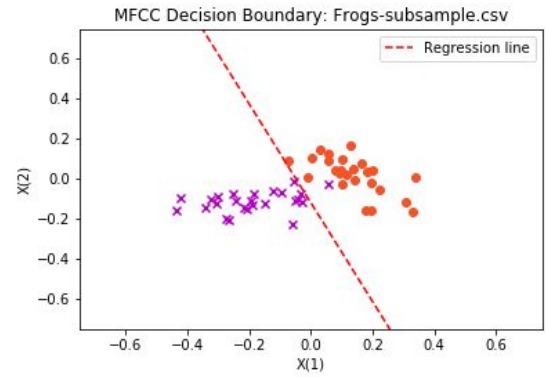
Below (Row 1): Histograms of the MFCC_10 feature variable for both classes of Frogs.csv
Below (Row 2): Histograms of the MFCC_10 feature variable for both classes of Frogs-subsample.csv

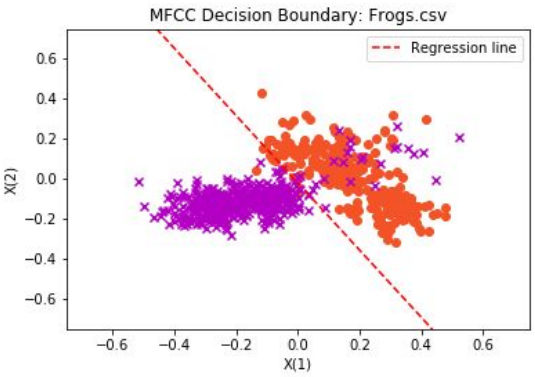


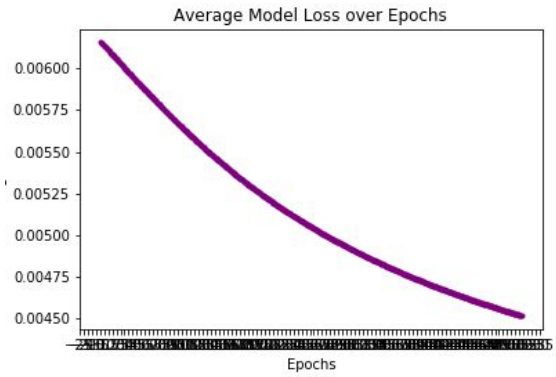
Above (Row 3): Box plot with error bars (Left) and Bar graph with whiskers (Right) of Frogs.csv

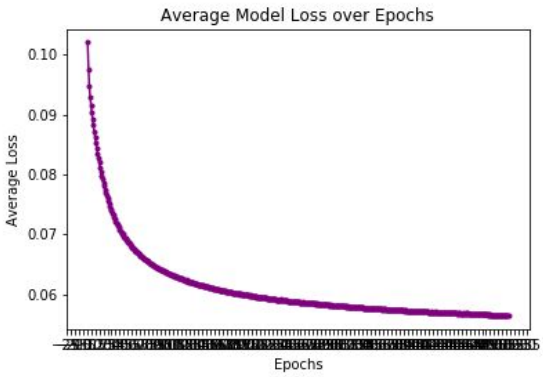
Looking at the scatter plots in table 1, row 1, we find that the distributions of both classes are generally similar between both the full dataset and the subsample with HM roughly centered at (0.2, 0) and HC centered at (-0.2, -0.1), however there is some overlap. Observing the box plot above (row 3), there are many more outliers to the distribution in HC, whereas there are none in HM.

Question 2:









Frogs-subsample.csv

Frogs.csv

Above (Row 1): Scatter plot with decision boundary for both datasets.
Above(Row 2): Model loss plotted as a function of epochs for each dataset.

After training the model on both datasets, I observed that the functions produced by the model were similar in their intercept, but deviated slightly in slope, suggesting that there are more data points for class HM further to the left of 0.0 than in the subsample dataset. The regressor seems to attempt to accommodate this mass by having a slightly less steep slope and also suggests that this is closer to the true distribution on the two random variables (features).

There are also more HC points spread throughout the center of the mass of HM, suggesting that there are some overlap of features. After training significantly (1000+) epochs, I saw no improvement in the loss; perhaps adding more features will lead to better hyperplane separation with a more complex model.