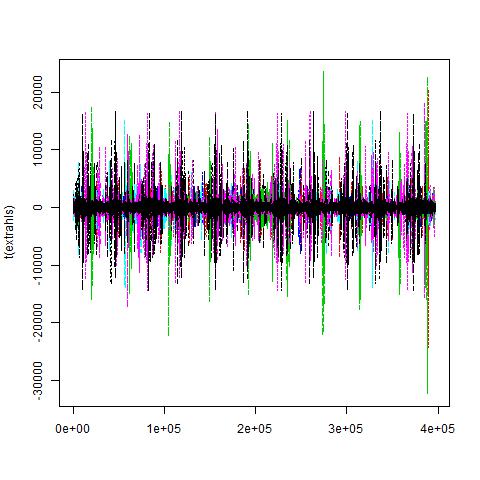
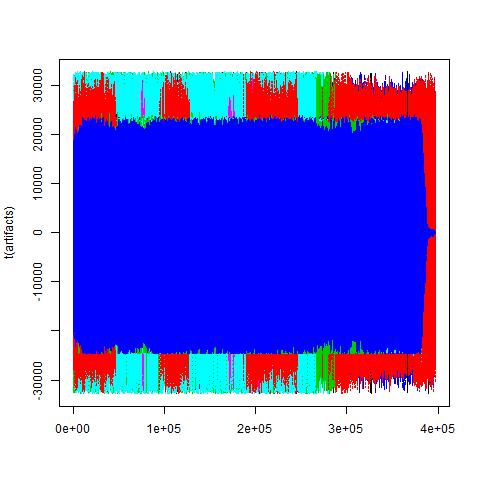
Update 4: Exploratory Data Analysis

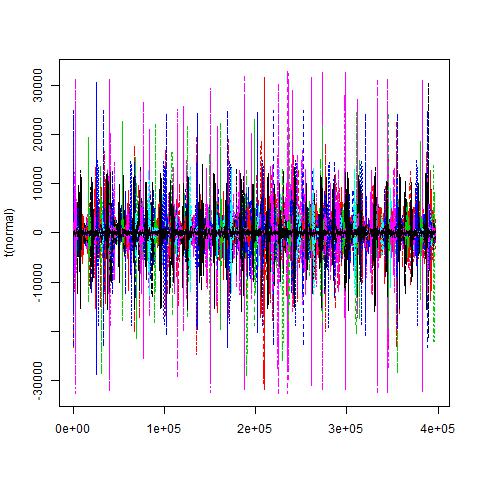
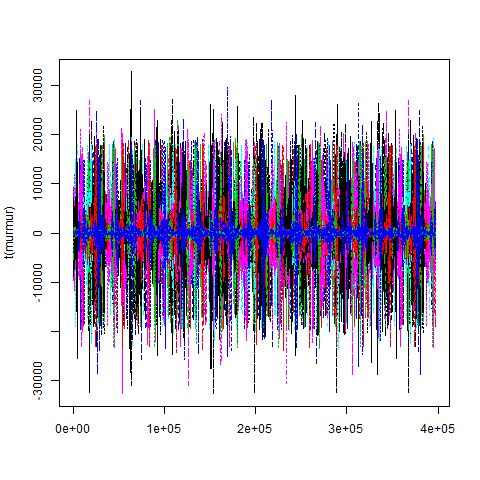
Dataset: Heartbeat sounds from <https://www.kaggle.com/kinguistics/heartbeat-sounds>

Goal: To classify unlabeled sounds into murmurs, extrahls, artifacts, or normal heartbeat sounds

Preparing the data: Sound files were imported into R vectors using a library called tuneR.

Plotting the data from labelled sound files yielded some messy plots (see below):



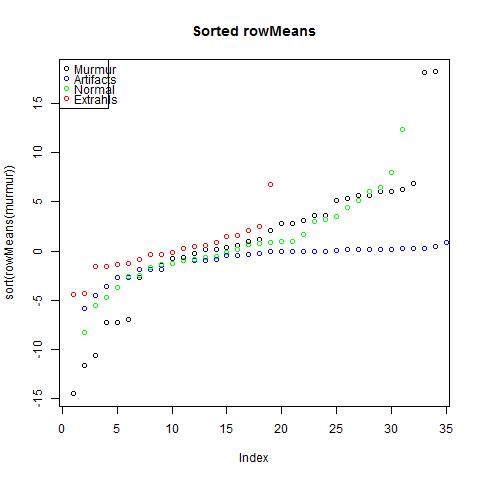


The data is in 396900 dimensions, and I was trying to use Principal Components Analysis to reduce the dimensionality to 2 so that the data would be easier to visualize. However, computing the covariance matrix took up too much memory and my PC was unable to handle it.

Instead, I reduced data into one dimension using some properties of the 1x396900 vector to see if there are any trends:

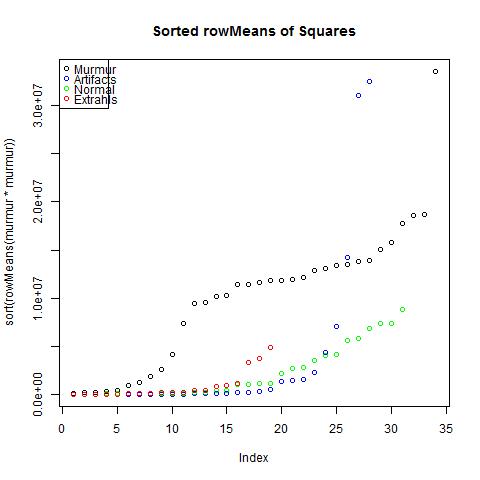
* Mean of each vector
* Mean of each vector squared
* Maximum of each vector
* Minimum of each vector
* Index of maximum
* Index of minimum
* Median
* Mode

***Mean***



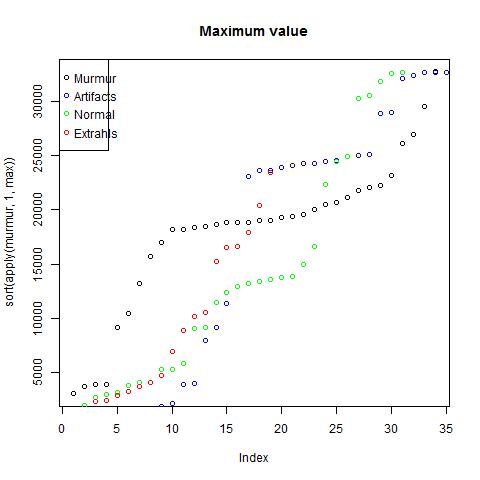
Observations: The data is not completely separable since we see some overlaps between index=5:10 but means of extrahl heartbeat sounds stayed consistently over the other three categories.

***Mean of each vector squared***



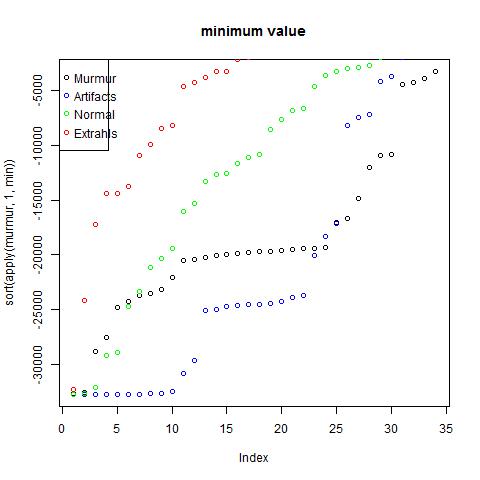
Observations: There is still no clear separation, but means for murmur heartbeat sounds squared have a clear pattern that we may be able to model.

***Maximum of each vector***



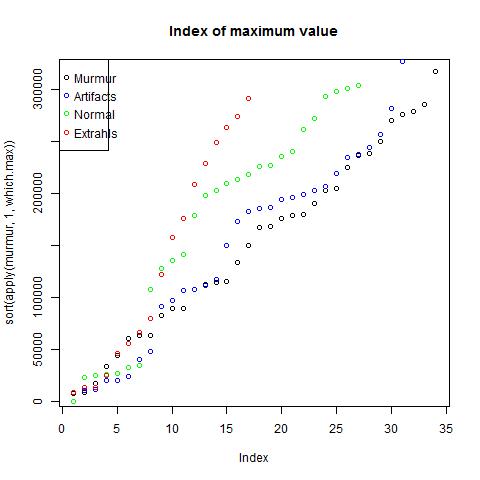
Observations: There is no clear pattern here.

***Minimum of each vector***



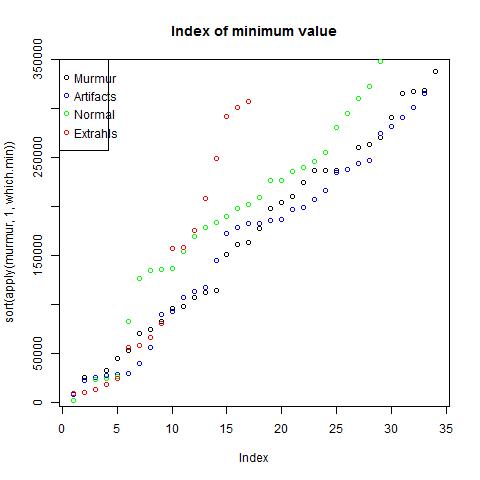
Observations: The minimum values seem to be separated in tiers, with extrahls having the highest minimum values.

***Index of maximum***



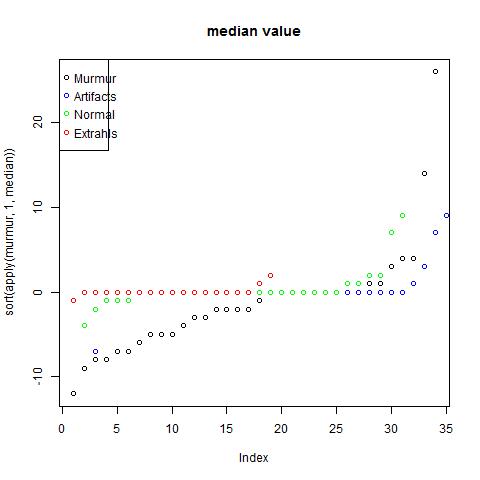
Observations: There is no clear patter from index=0:15, but for index>15, there is a tiered system that is linearly separable.

***Index of minimum***



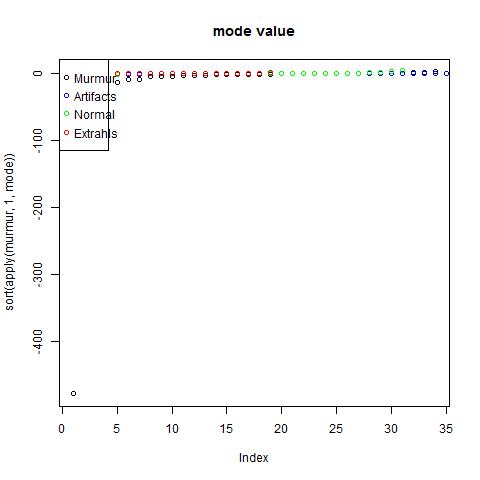
Observations: Similar to maximum index, there is no clear distinct pattern in the beginning, but we can see tiers for index>15.

***Median***



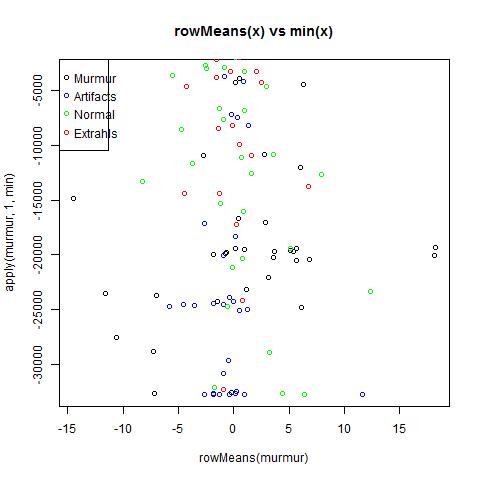
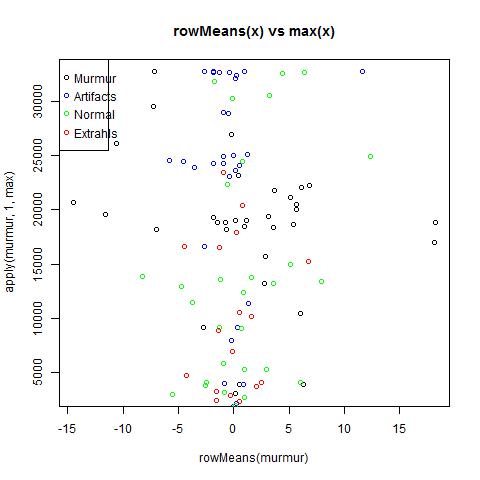
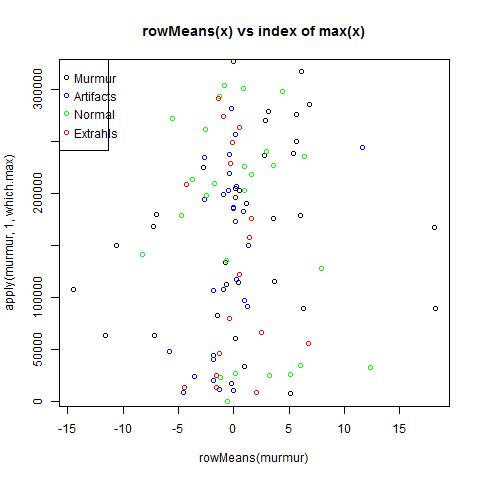
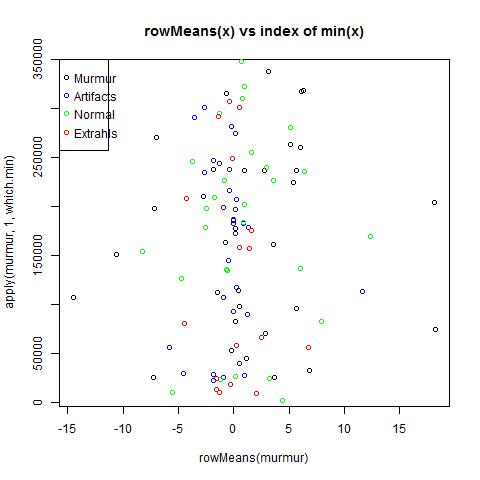
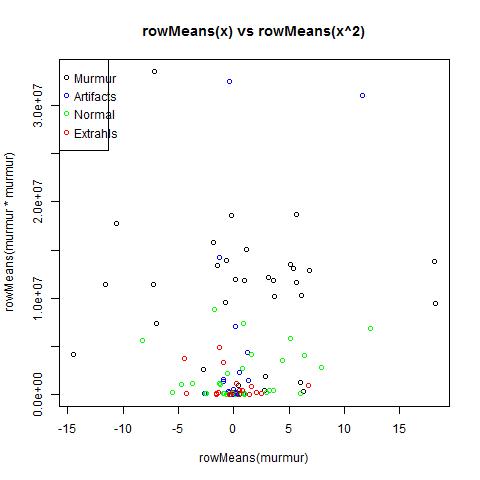
Observations: Many of the median values fell onto 0, most likely because these are audio signals. The only distinction here is that murmur heartbeat sounds are the least likely to have a median of 0.

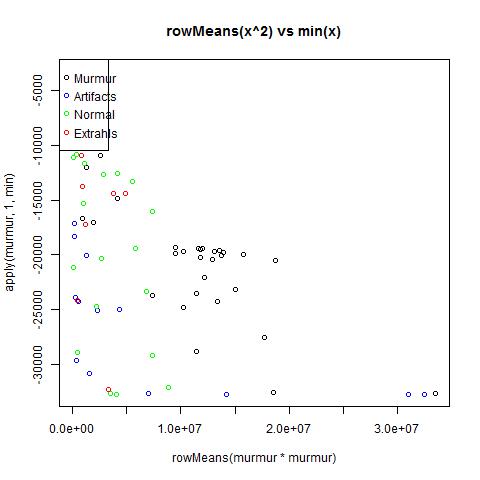
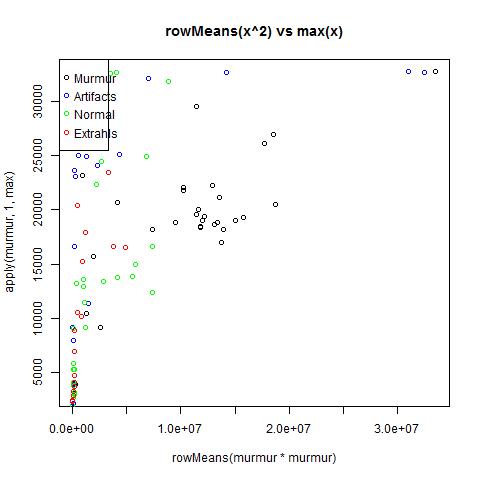
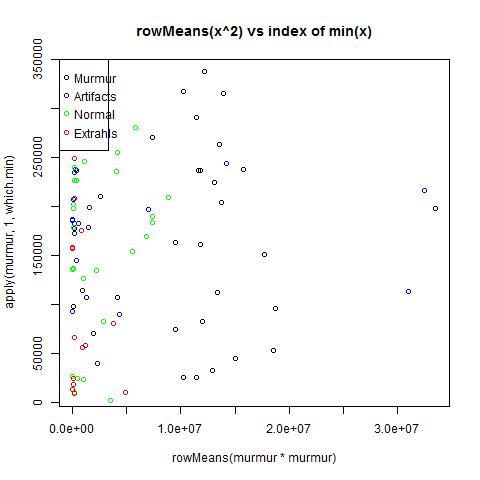
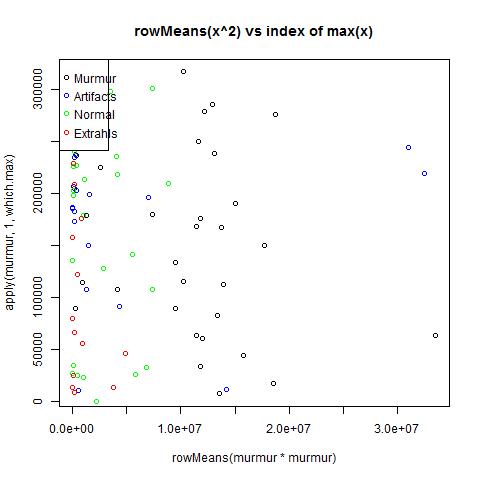
***Mode***

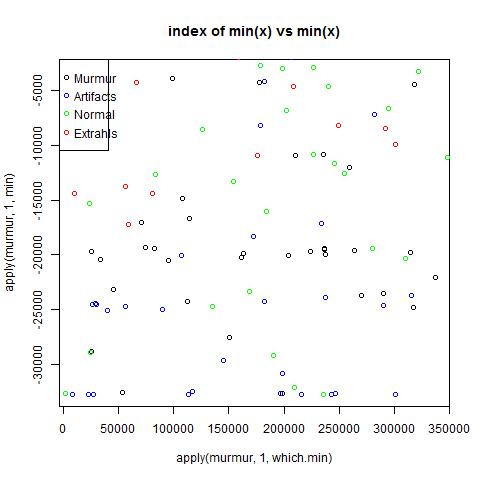
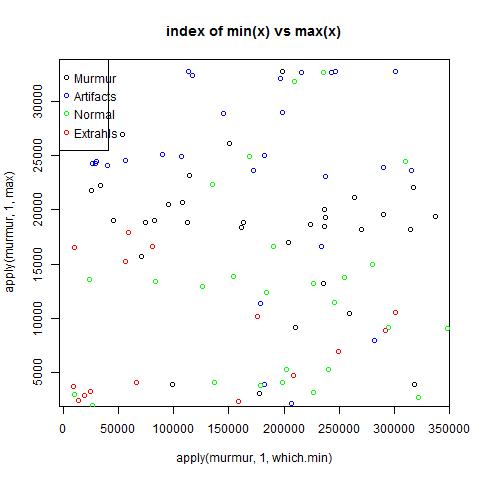
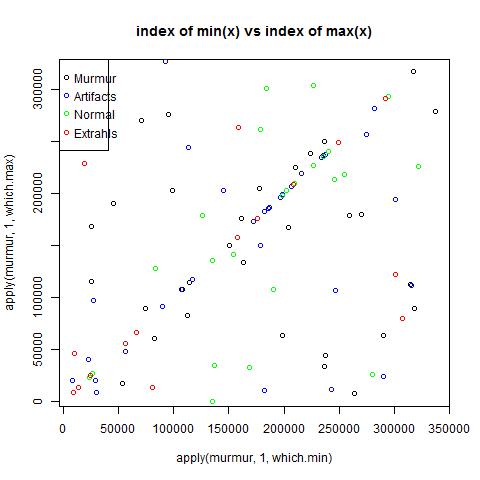


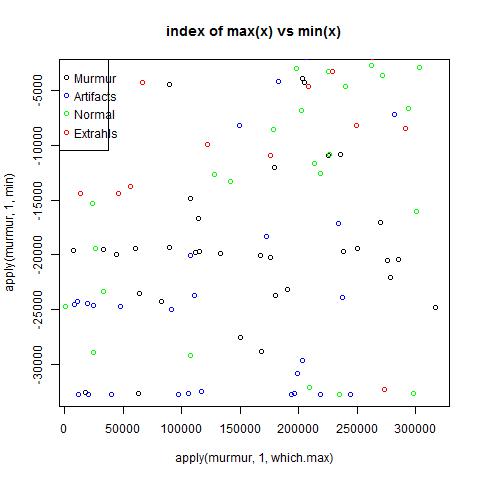
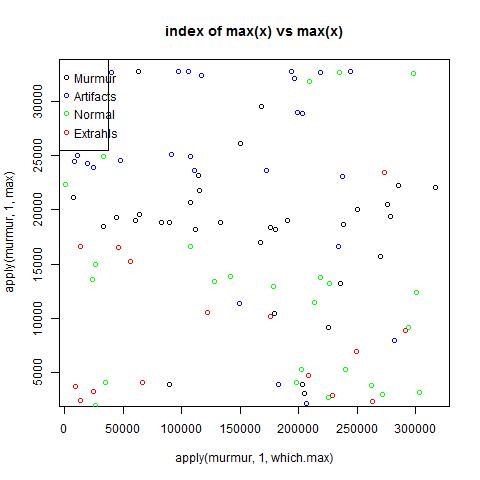
Observations: Mode values are mostly 0.

I also attempted to plot some characteristics together to see if there were any correlations:









Observations: There are no significant correlations between these characteristic.