Update 5: Reduction in dimensions

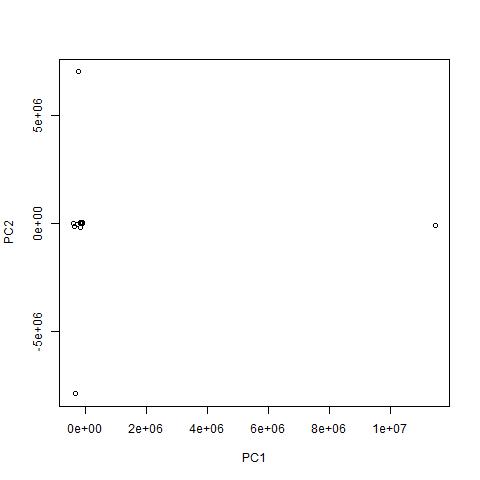
Dataset: Heartbeat sounds from <https://www.kaggle.com/kinguistics/heartbeat-sounds>

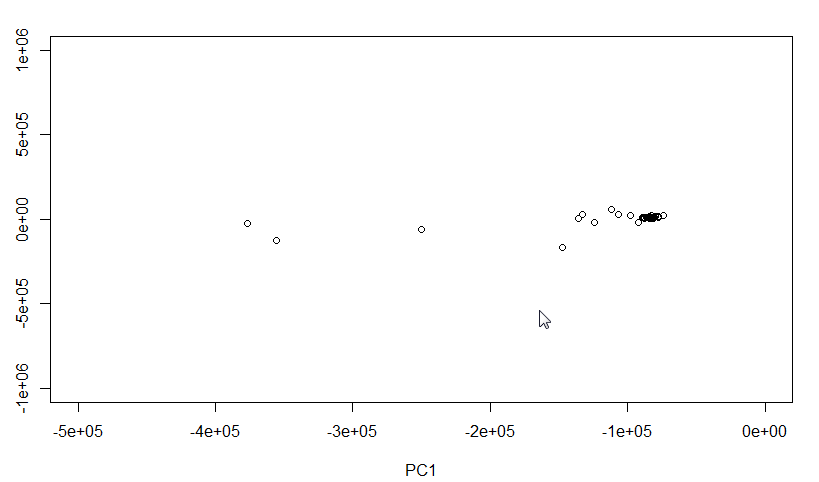
Goal: To classify unlabeled sounds into murmurs, extrahls, artifacts, or normal heartbeat sounds

Since there are about 400000 independent variables for each dataset, I would like to reduce dimensionality to 2 or 3. In my last update, I reduced dimensionality to 1 because I ran into an issue while attempting PCA manually, while creating the covariant matrix, a stronger computer was necessary (strong enough to store or compress a matrix of size 1170 GB). Therefore, this week, I attempted PCA using the built-in function prcomp.

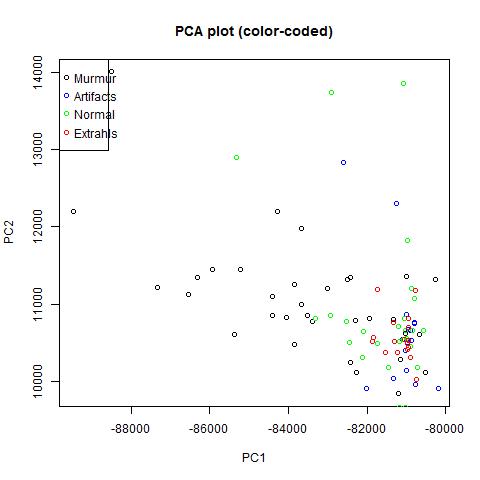
First, I used prcomp to find the model and then I used predict to convert all the data to PC1-2 or PC1-3.

For PC1-2, the data was not visually separable (See below). There were a few outliers, but even examining closely, the group of points in the middle form an inseparable blob:

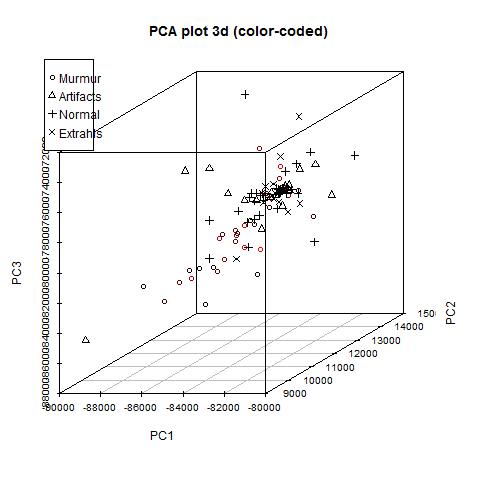




Then, I tried to separate out the Principal Components by classes (See below). This provided a better visual of the separation, but there is still a blob in the bottom right of the graph where several values are stuck together.



Using scatterplot3d, I plotted a Principal Components 1-3 (See below). The separation here was not as obvious as well.



Using PRCOMP we can also plot the variances (See below). The variance is quite large for 3 components (8e11) so results from PCA will not be as accurate.

