Report of Applied Data Science Project 4

Firstly, I selected four feathers as my input. They are Loudness\_max, Loudness\_max\_time, Segments\_Pitches and Segments\_Timbre. Because the dimension of each variables are from 4 to 3990, I repeat and cut them into 2000. Then, I did k-means cluster to these four feathers. For each feather, the number of cluster is 10. So, I have the data set called “input\_data.RData”. And then, I did topic modeling for the outcome. There are 19 cluster in it. I name the label data set “label\_outcome”. Then, I used xgboost to make a model and then selected parameters. Then, I will apply this model to the test data set and give each songs a cluster. After that, I will calculate the mean ranking of the songs in different clusters. Then use the mean ranking of it’s cluster as it’s predicted ranking.