Shuk Ying Leung

Muu Nkhoma

FINAL PROJECT: AUDIO TRACK

**Abstract**

Research question:

1. What is the performance of the clusters we form to identify the region where all the observations were from?
2. How well an audio tracks can be traced to the particular region that they belong?
3. What variables are good at describing audio tracks according to particular regions or clusters?

Objective and Method:

All the variables in the dataset are continuous, where 116 variables are soundtracks contains performance with basic timbal information and the last two variables are the longitude and latitude of the countries or place the sound track from. We created two 2 categorical variable, country and region to locate where the sound tracks from, then, we started our analysis. Firstly, we decided to divide the 1059 audio tracks according to the 6 regions of the world. Thus, we decide to use K mean clusters to classify the sound tracks into 6 clusters. Then, we will use classification model to distinguish 6 regions from the total observations and choose the model which has the lowest misclassification rate as a criterion to be the best model. The classification model we will use are different Kernel method with k from 1 to 10, Monte Carlo cross validation, Naive Bayes and discriminant method (LDA, QDA, RDA).

We used both JMP and R/RStudio in our analysis.

**Description of the dataset**

The data set we use in this project was built from a personal collection of 1059 tracks covering 33 countries or area. The music track is either traditional, ethnic or ‘world’ only, which were classified by the publishers of the product. However, it does not include Western music. The information of geographic location of origin are from the last two variables longitude and latitude and it was collected from the CD sleeve notes. The country of origin was determined by the artists or artists’ main country or area. Two extra categorical variables “country” and “region” are created from longitude and latitude. There are 120 variables in total, where consists 116 quantitative audio features of the tracks, longitude, latitude and 2 additional categorical variables. The program MARSYAS, which is a software framework for rapid prototyping of audio applications was used to extract audio features from the wave files. The default MARSYAS settings in single vector format (68 features) to estimate the performance with basic timbal information covering the entire length of each track. No feature weighting or pre-filtering was applied. All features were transformed to have a mean of 0, and a standard deviation of 1.

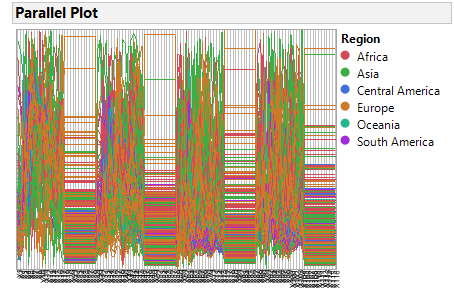
**Analysis**

*Table: The distribution of number of regions and countries.*

|  |  |  |
| --- | --- | --- |
| Region | Number of Countries | Numbers of Observations |
| Africa | 9 | 305 |
| Asia | 13 | 440 |
| Central America | 2 | 33 |
| Europe | 7 | 231 |
| Oceania | 1 | 14 |
| South America | 1 | 36 |
| **Total** | **33** | **1059** |

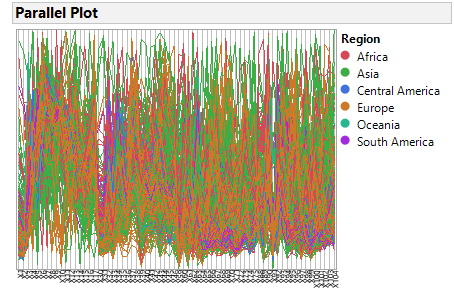
**EXPLORATORY MULTIVARIATE ANALYSIS**

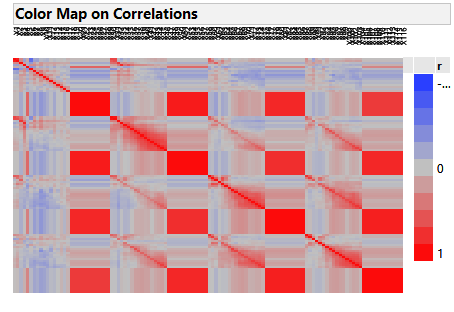
**Parallel Coordinate Plots**

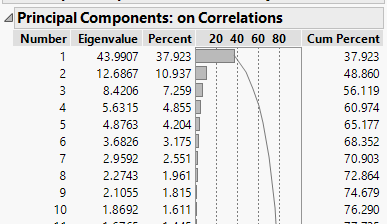
****

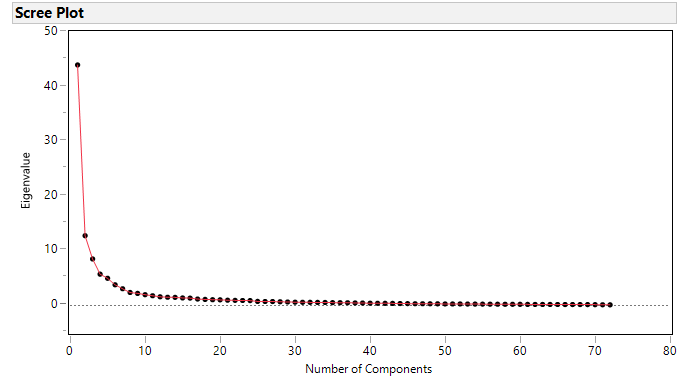
We performed an exploratory multivariate analysis on the data and found the following:

The data had variables that formed distinct bands of highly correlated if not perfectly correlated variables after almost a specific interval. The bands can be seen in both the Parallel Coordinate Plots and the Color Map on correlations.

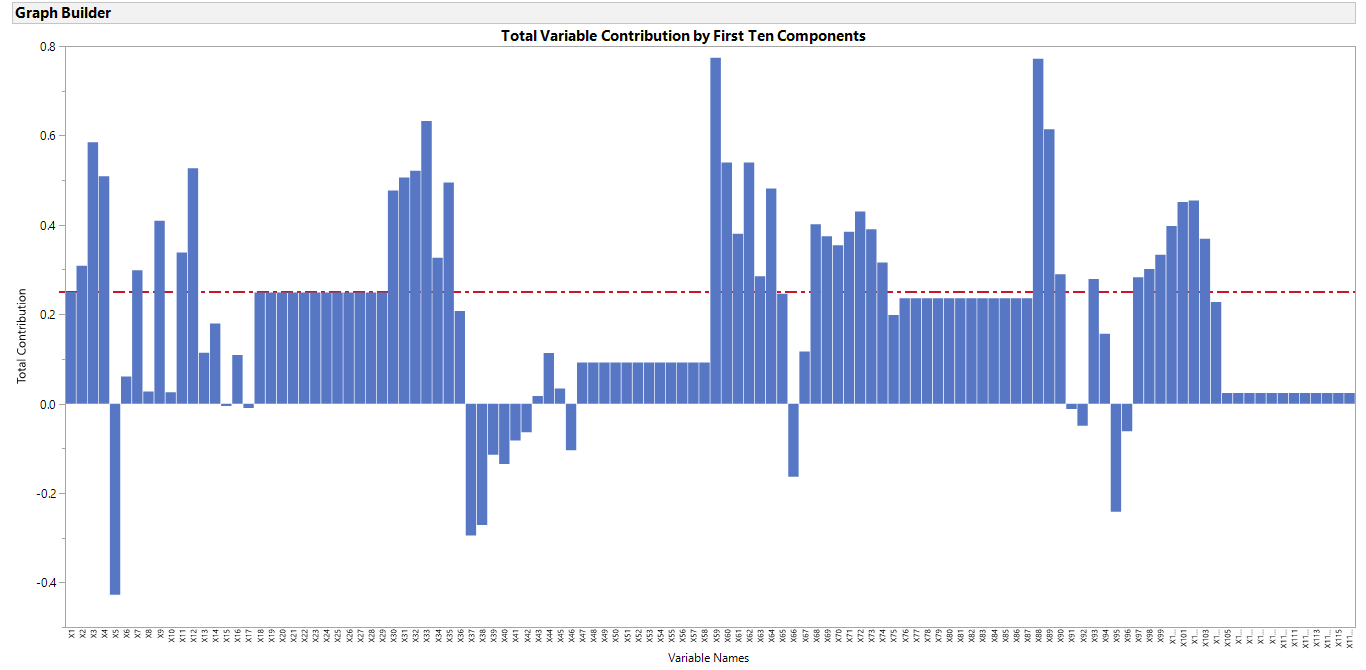
****

****

****

****

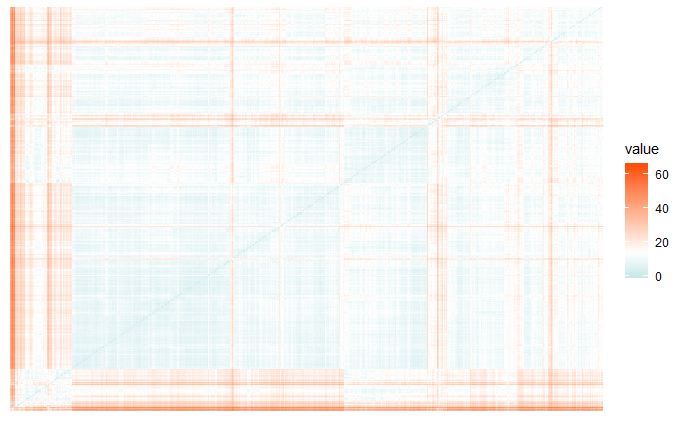
Based on the pattern of the parallel plot and result of the color map, they give evidence that some of the variables with high correlations causes horizontal lines in the parallel and red squares zone in the color map. We suspect some of those variables might not be important in the dataset. We generated PCA analysis and look over the scree plot. We chose ten principal components based on the elbow angle from the above curvature and the cumulative percentage of ten PCAs provided, which is 76.29%. Then, evaluate the total variability of those principal component provided and the amount of variability each predictors contributed.

****

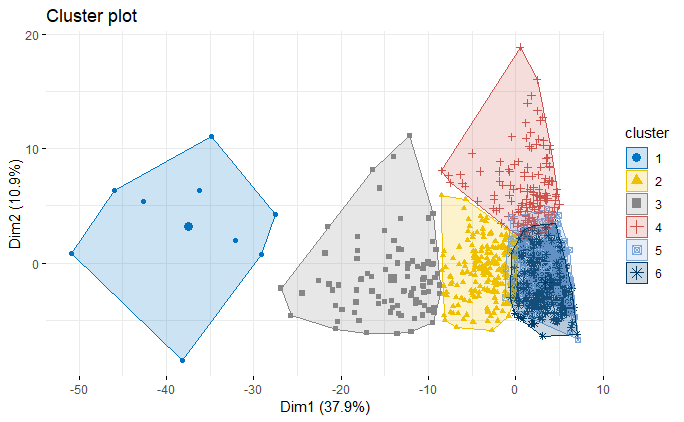
This plot shows the total variability each variable provided for the first ten principal components. The red dash line is at 0.25 which is the cut off line to the total contribution for the variability. We chose 0.25 as cut off line because it is able to separate the correlated variables, which contributes the same amount of variabilities. For instance, variables from x18 to x29 have the identical variability 0.25 of total contribution that you can observed from their identical length of bars. Moreover, the cut off line also helps ignore those variables contributed too little amount of variability and those variables have negative total contribution.

**CLUSTERING METHODS**

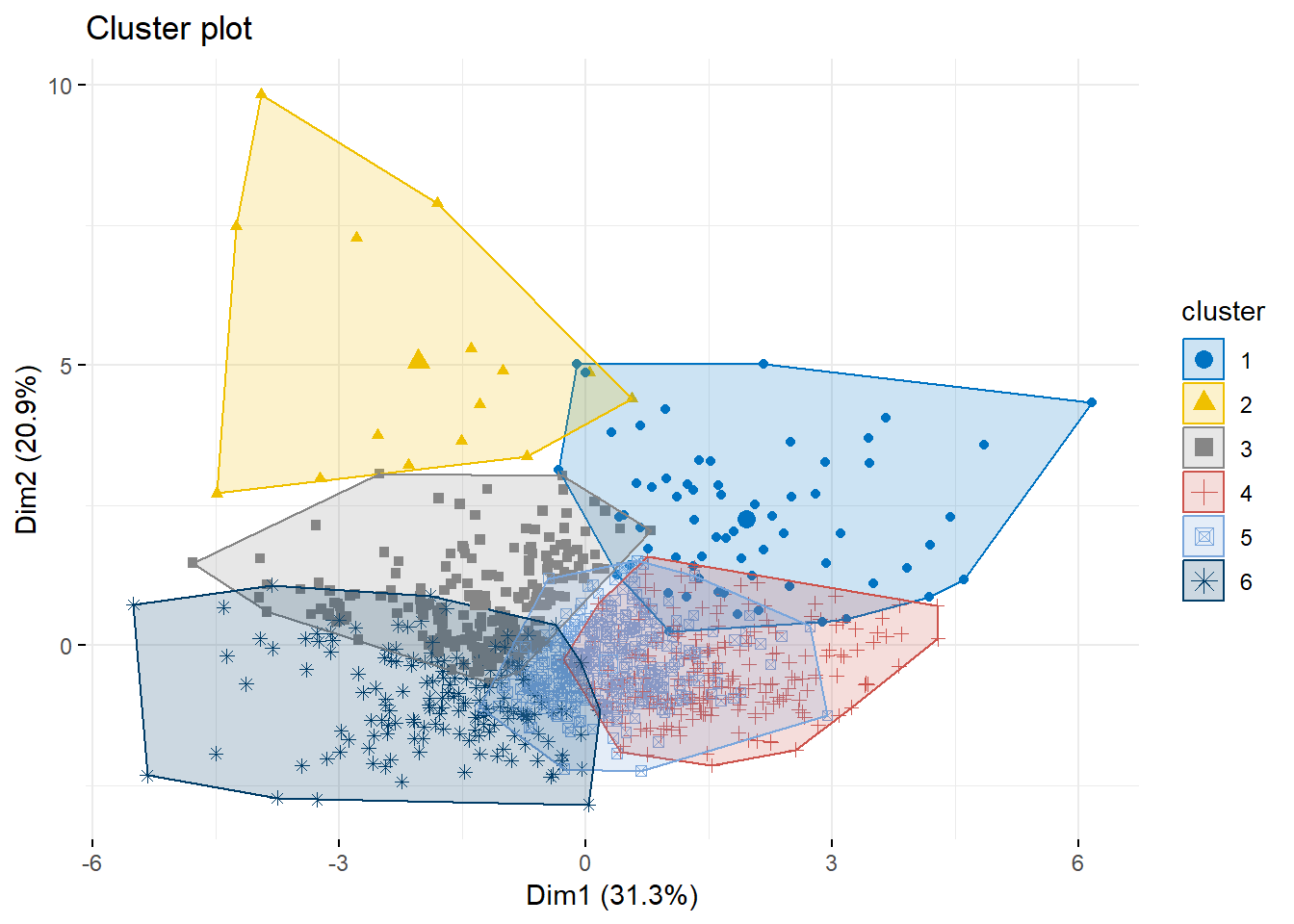
Distance Matrix



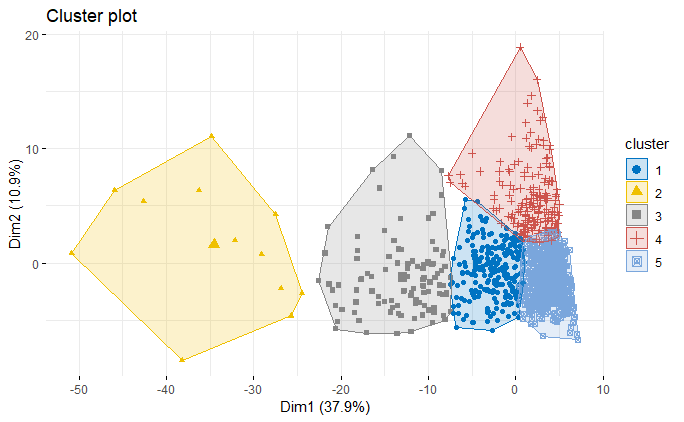
We first try to use Hierarchical Clustering methods and we had clusters that did not fit well as the number of clusters was large and there was less distinction between clusters. We then decided to use K-means clustering.



This is a 6 mean cluster plot. The reason we input k = 6 because we were attempting to have 6 clusters to predict 6 regions we created. However, there is overlap between cluster no.5 and cluster no.6. We assume 5 clusters is better than 6 clusters because besides the cluster no. 5 and no.6, other 4 plots are approximately distinct and separate from each other.



The above Cluster Plot is from the most important variables as identified from the top Ten Principal Components. Even though the clusters are spread out, they still do overlap a lot. Therefore, we did not use those variables in the final clustering model.

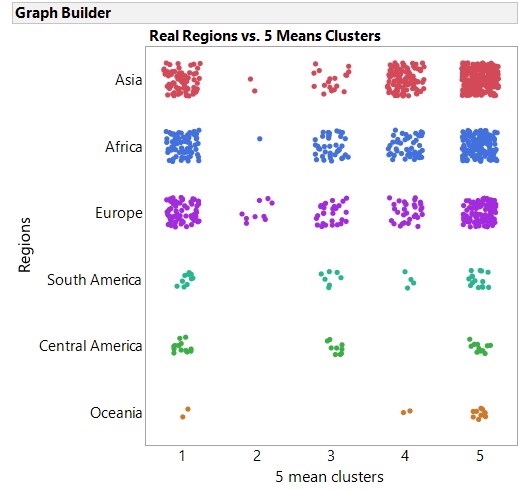


This is a 5 mean cluster plot. After observing the last 6 mean clusters from all the variables and the most important variables, we decided to generate a 5 mean clusters plot to see if it fit better, having all clusters separating from each other with little overlapping. As you can see the result above, 5 clusters are approximately separated from each other. We conclude that k = five forms better clusters k = 6. We saved 5 clusters for further analysis.

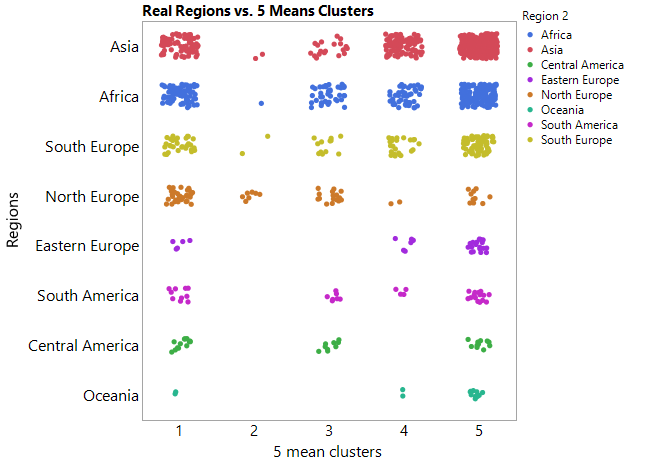
*Table: the distribution of number of observations in each cluster*

|  |  |
| --- | --- |
| 5 means clusters | Numbers of obs. |
| 1 | 232 |
| 2 | 12 |
| 3 | 102 |
| 4 | 183 |
| 5 | 530 |

Different area of Europe could validate the reason why there are Europe soundtracks are misclassified as other regions. As a result, we decide to look deeper different geological area of Europe by dividing Europe into South and North region to investigate if there are relationships within them.

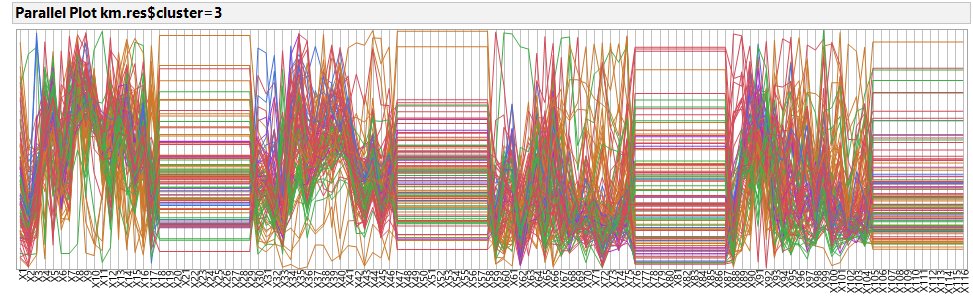


This graph we have compare the five mean clusters we formed and the real regions. We were looking for having clusters formed correctly to the real identified regions. The different colors of the dots indicate their different regions and the x axis is the 5 mean clusters from the research above. The distribution of the dots (total observations) shows the dots from one region are spread in a few different clusters. If there are dots from one region is totally belong to one single cluster. Then, the clusters will be a good identifier for the region. The result above illustrates the cluster #2 has only 14 observations, which indicates the cluster #4 is very distinctive than the other 4 groups. 9 out of 14 are from the Europe, where 2 belongs to Asia and 1 belongs to Africa, which means cluster 2 is a good identifier for distinguishing Europe from the total observations. However, the overall performance of the clusters does not be able to identify all the regions precisely.

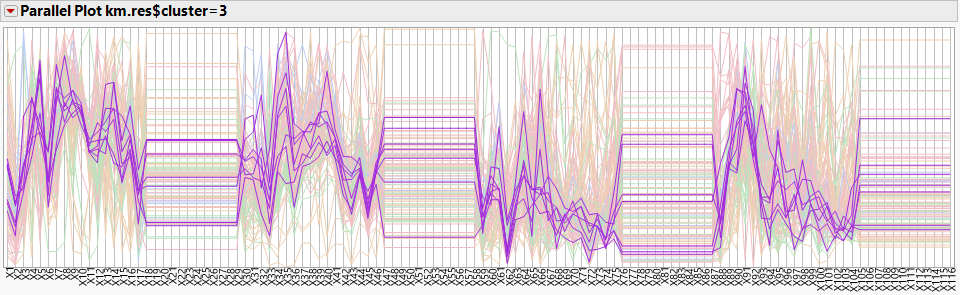


Then, we were wondering which part of Europe the observations identified in cluster #4 from. Thus, we have the region Europe distinguished apart from North, Eastern and South from the countries they belong. We found out 7 of 9 observation from Europe is from North, 2 of them from South and none of them from Africa.

Then, we intend to investigate the variabilities of each variables from clusters to see which variables is special to a certain region than to another, different variabilities from variable from each region. In this case, we picked cluster #3 as an example for this analysis.

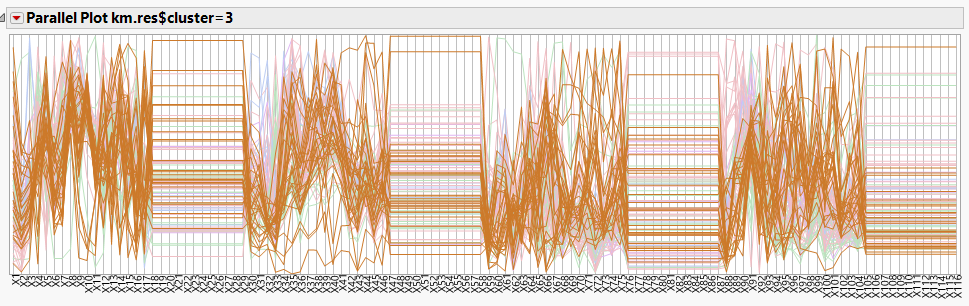


South America:



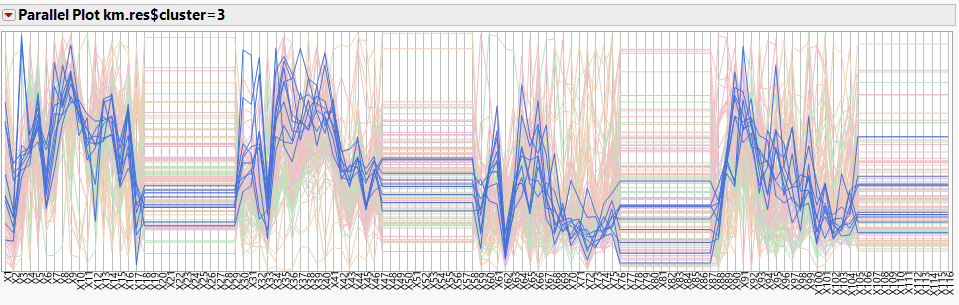
Important Variables: X2, X5, X6, X44, X45, X59, X62, X88.

Europe:



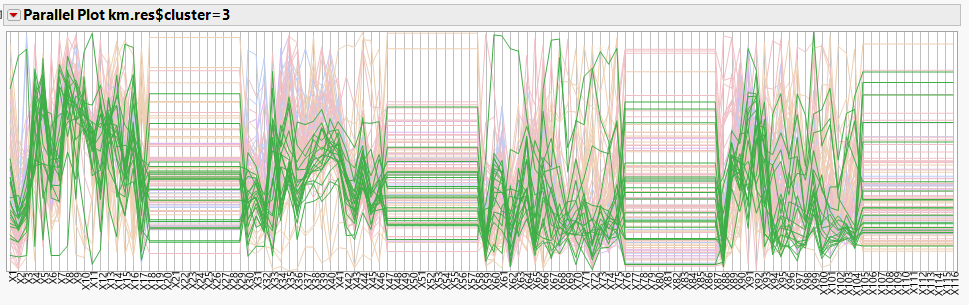
Important Variables: X8, X9, X10, X91.

Central America:



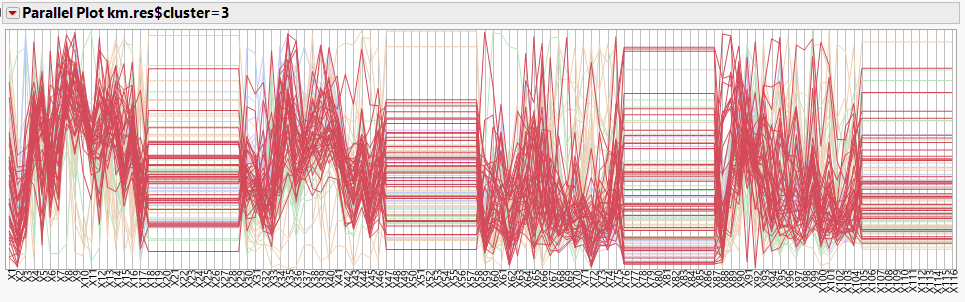
Important Variables: X5, X6, X15, X16, X59, X62

Asia:



Important Variables: X2, X6, X43, X59, X62, X88.

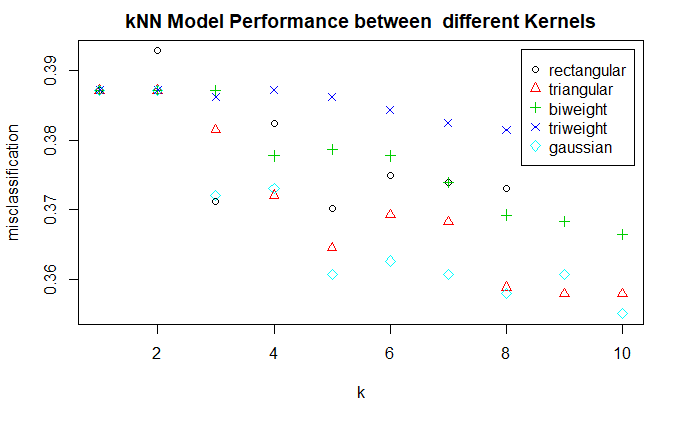
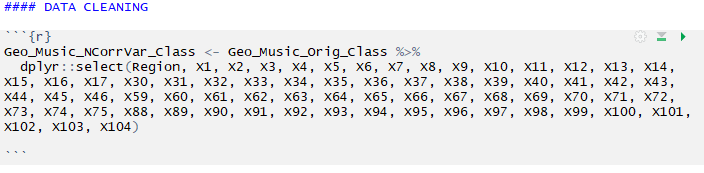
Africa:

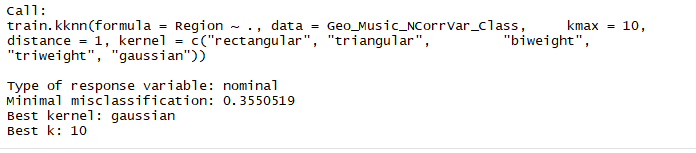


Important Variables: X7, X8, X9, X43, X45.

**CLASSIFICATION METHODS**

For classification we used a cleaned data set without the highly/ perfectly correlated variables. The total number of variables we use for classification is 69, including class variable, Region.

****

****

This plot shows the misclassification rates for various knn model with 1 to 10 k neighbors. We define the best model if it has the lowest misclassification rate. Gaussian has the lowest misclassification 0.3550 among all Kernel models at k=10.

****

****

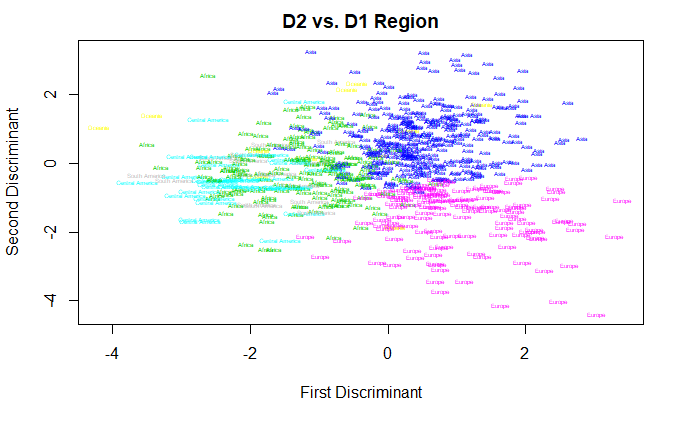
The knn monte Carlo has misclassification rate 0.3942.

****

****

**DISCRIMINANT ANALYSIS**

**Linear Discriminant Analysis**

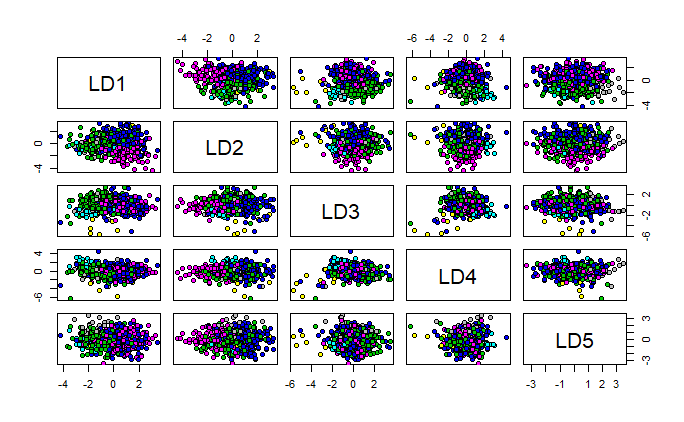
****

**Europe, Asia, Africa, Central America, South America, Oceania**

The Second Linear Discriminant seems to allocate the audio tracks from Europe better by packing at values below 0, while the First and Second Linear Discriminants have grouped audio tracks from Asia in the first Quadrant. This can also be evidenced by looking at the matrix below between the LD1 and LD2.

The train.kknn() function in R which uses the leave-out cross-validation method compares various kernel and gave us the best model as the k=10 Gaussian model with a misclassification rate of 0.3550, while the in class Monte-Carlo Cross-Validation method using the cv.kknn () of the same parameters gave us a mean misclassification rate of 0.3942.

These two discrepancies made us not use any of the models as none of the Gaussian kernel models from k = 1 to k= 10 produced a misclassification as low as the train.knn.

****

**Europe, Asia, Africa, Central America, South America, Oceania**

*Table: The misclassification rates for multiple classification comparison models*

|  |  |
| --- | --- |
| Model | Misclassification rate from validation set |
| K nearest neighbor K=1, Rectangular | 0.3900 |
| K nearest neighbor K=7, Biweight | 0.3750 |
| K nearest neighbor K=9, Triweight | 0.3750 |
| K nearest neighbor K=9, Triangular | 0.3670 |
| K nearest neighbor K=1, Gaussian | 0.3900 |
| KNN Monte Carlo CV, Gaussian | 0.3942 |
| Naive Bayes | 0.6440 |
| Naive Bayes, Monte Carlo CV | 0.6557 |
| LDA | 0.4320 |
| RDA with gamma = 0.004316288 and lambda = 0.865518083 | 0.3900 |

**Issues Encountered.**

We did not run the Quadratic Discriminant Analysis in R as we were having an error on “some groups are small to be used in the analysis” This was a result of having some clusters with fewer observations than the total number of predictors. (n < p) problem.

**CONCLUSION**

The number of observations from each region are not equally spread as 6. Some of them has highly more observations than another. The audio tracks in the data set do not have sufficient distinct qualities that can be grouped together according to the area of region the audio track originated from. However, we did find some variables which were important in grouping audio tracks from a particular region together if they were in the same cluster.

We have looked over the variability of variables by the PCA analysis and multivariate correlation, we found out that there are four groups of variables that are highly correlated. Since the variables did not distinctly group the audio tracks well, most of the classification methods also did have problems with correctly classifying the audio track to a particular region, and with the Naive Bayes algorithm doing the worst. The model selected for classification is the 9-Nearest Neighbors with Triangular kernel is the model that performed best in classifying the audio tracks.