**Question 2**

**a**

Whistle Duration Center\_Freq Low\_Freq Delta\_Freq Max\_Freq Range\_50

Concave :300 Min. :0.0640 Min. : 2756 Min. : 0 Min. : 2734 Min. : 689.1 Min. :-4747

Constant :300 1st Qu.:0.4040 1st Qu.:11025 1st Qu.: 5649 1st Qu.: 9840 1st Qu.: 8785.5 1st Qu.: 3466

Convex :300 Median :0.6525 Median :12231 Median : 7512 Median :12392 Median :11369.5 Median : 4536

Downsweep:300 Mean :0.7092 Mean :12307 Mean : 7617 Mean :12333 Mean :11213.1 Mean : 4690

Sine :300 3rd Qu.:0.8790 3rd Qu.:13437 3rd Qu.: 9476 3rd Qu.:14949 3rd Qu.:13436.7 3rd Qu.: 5766

Upsweep :300 Max. :7.2980 Max. :20327 Max. :16800 Max. :22050 Max. :21533.2 Max. :11899

Range\_100 Inflections

Min. :-8785.6 Min. :0.000

1st Qu.:-2584.0 1st Qu.:1.000

Median : -689.1 Median :1.000

Mean :-1094.4 Mean :1.055

3rd Qu.: 172.3 3rd Qu.:1.000

Max. : 7924.2 Max. :3.000

The above R-output shows the summary of the dataset, and it will be seen that it is required to standardize the numeric variables before performing principal component analysis which makes thye mean = 0 and variance = 1 to satisfy the PCA assumptions. Also, it can be seen that each whistle group have the same number of observed dolphins (300 each)

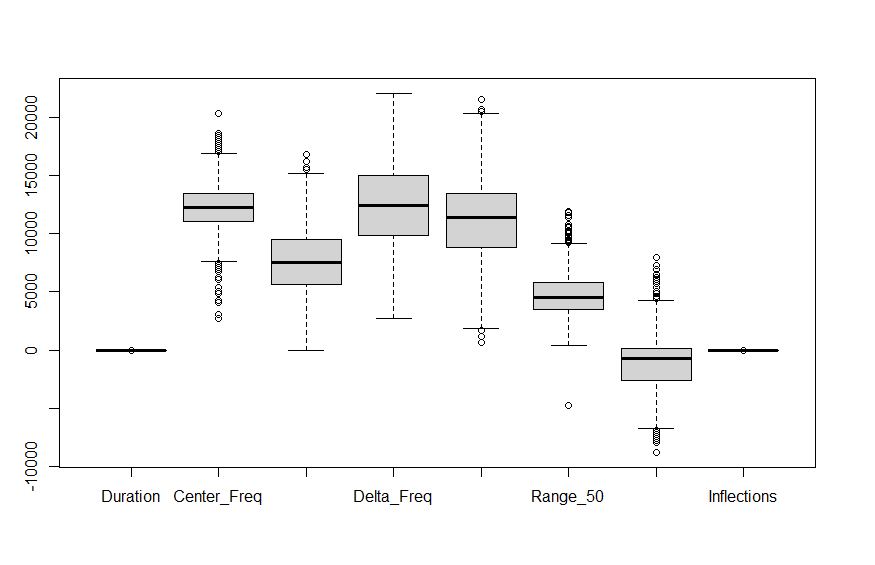


Figure 1: Boxplot of numeric variables

The boxplot above confirms the presence of outliers in the variables of the dataset.

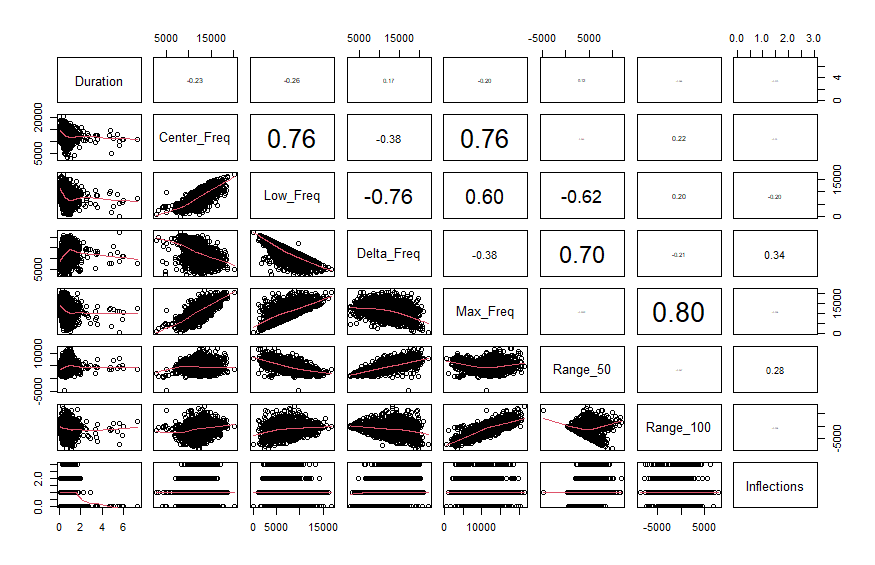


Figure 2: Correlation plot across variables

The pairs plot above shows that there are some frequency variables with a significant high correlation with each other, and some of the pairs with high correlation include; (center\_freq V low\_freq with 76%), (low\_freq V delta\_freq with -76%), (center\_freq V max\_frew with 76%), (max\_freq V range\_100 with 80%) and (Delta\_freq V Range\_50 with 70%). And from the plot of variables against each other , it can be seen that majority are linearly related to each other.

**b**

Standard deviations (1, .., p=8):

[1] 1.842 1.315 1.031 0.923 0.859 0.472 0.000 0.000

Rotation (n x k) = (8 x 8):

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8

Duration -0.180 -0.075 0.670 -0.463 -0.546 -0.016 -0.000 -0.000

Center\_Freq 0.405 0.290 -0.197 -0.546 0.050 -0.041 -0.621 -0.158

Low\_Freq 0.502 -0.136 -0.167 -0.223 -0.153 -0.325 0.618 -0.376

Delta\_Freq -0.432 0.340 -0.024 -0.056 0.137 -0.822 -0.000 -0.000

Max\_Freq 0.433 0.433 0.186 0.027 0.005 -0.055 0.221 0.733

Range\_50 -0.282 0.557 0.020 -0.315 0.295 0.448 0.403 -0.245

Range\_100 0.277 0.383 0.463 0.548 -0.040 -0.045 -0.146 -0.486

Inflections -0.146 0.363 -0.484 0.182 -0.754 0.102 0.000 -0.000

The table above gives the principal component table approximated to 3 decimal places. The result shows the standard deviation of each component. The eigen vectors for each principal components for which each variable is factored by to get its respective component score.

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8

Standard deviation 1.842 1.315 1.031 0.923 0.859 0.472 0.000 0.000

Proportion of Variance 0.424 0.216 0.133 0.107 0.092 0.028 0.000 0.000

Cumulative Proportion 0.424 0.641 0.773 0.880 0.972 1.000 1.000 1.000

The above table shows the importance of each principal component approximated to 3 decimal places. It shows the standard deviation of the data along each component as well as their individual and cumulative contribution to the model. From the table, it can be seen that Component 1 accounts for 42.4% of the model variance, PC2 accounts for 21.6%, PC3 accounts for 13.3%, PC4 accounts for 10.7%, PC5 accounts for 9.2%, PC6 accounts for 2.8%, but PC7 and PC7 do not account for any significant amount of the variability in the model. Therefore, it can be concluded that all the variability of the model can be explained by the Components 1 through Component 6.

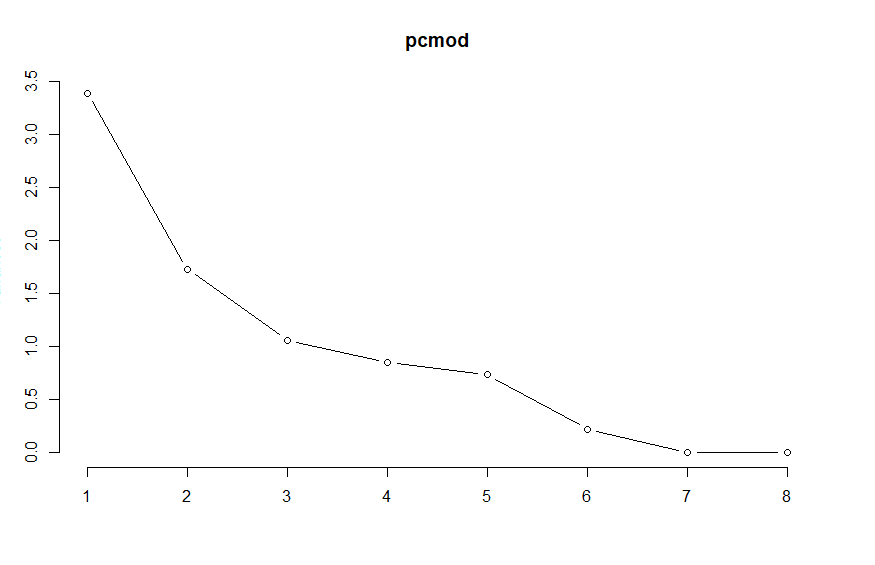


Figure 3: Variance Across each Principal Component

The plot above shows the variance explained across each principal component, and it can be seen that with the addition of each component moving from left to right, the variance decreases and the entire variability is already accounted for at the 6th Component.

|  |  |
| --- | --- |
|  | **Correlation between Principal components and variables** |
| Figure 4: Biplot of the PC1 vs PC2 | PC1 PC2  Duration -0.33 -0.10  Center\_Freq 0.75 0.38  Low\_Freq 0.92 -0.18  Delta\_Freq -0.80 0.45  Max\_Freq 0.80 0.57  Range\_50 -0.52 0.73  Range\_100 0.51 0.50  Inflections -0.27 0.48 |

The above plot and correlation table helps interpret the principal component across each variable; the principal component 1 and 2 are selected for this analysis as the combination of both of accounts for the highest variability in the model with 64.1% variability accounted for as seen in the table of importance described earlier. The relative length of the vectors (red line) gives the relative variability of the variable. The biplot shows that the maximum and center frequency and the difference between both of them are relatively highly correlated with each other

The plot above can be explained as follows:

* As an observation increases in value on the PC1 axis, its Center, Low, and Maximum Frequency along with its difference between Maximum and Center frequency (Range\_100) all increases. At the same time, its Duration, Delta frequency, Inflection and difference between Center and Minimum Frequency (Range\_50) all decreases.
* For that same observation, its increase in value along the PC2 axis will cause all variables to increase except for its whistle duration and low frequency which decreases.
* On PC2, much weight is placed on Range\_50 variable and the least weight is on the whistle duration, therefore, while on PC1, the most weight is placed on low frequency and the least is on inflection.

|  |  |
| --- | --- |
| Figure 5: Scree Plot | Figure 6: Cumulative PVE plot |

The scree and cumulative plots above simply give a representation of the amount of variability captured by each principal component relative to each other. The cumulative PVE plot already shows that the entire model is captured after the 6th loading vector (principal component). And the scree plot shows the level of variability is reduced on addition on a new principal component.

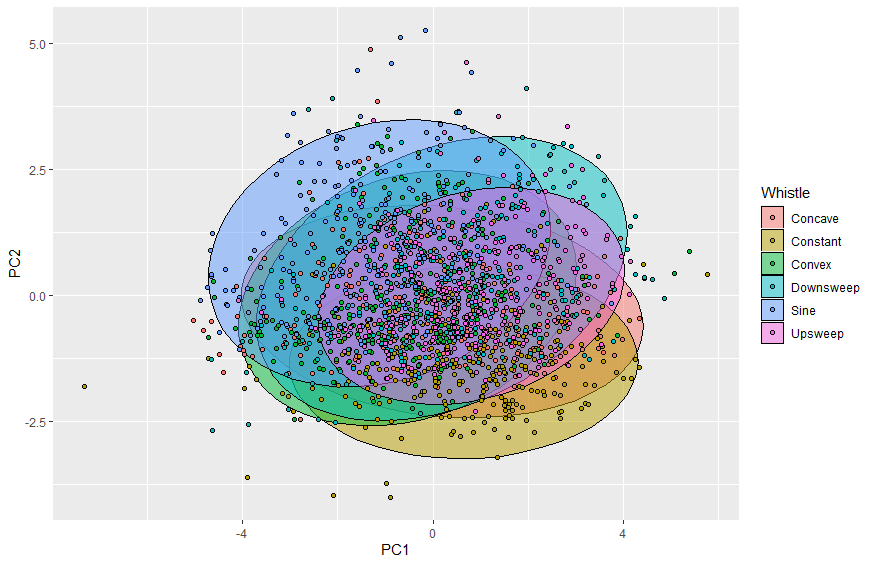


Figure 7: Regions of each group explained by low-dimension representation

The plot above shows the region covered by each group of observation with respect to the variability explained by PCA1 and PCA2. It will be observed that there is still much overlapping between the respective groups which is a result of the 35.9% of the variability unaccounted for from the low-dimension representation. A higher dimension can provide a better structure description of the classification groups thereby explaining more variability in the data.

**Question 3**

**a. Kmeans Cluster**

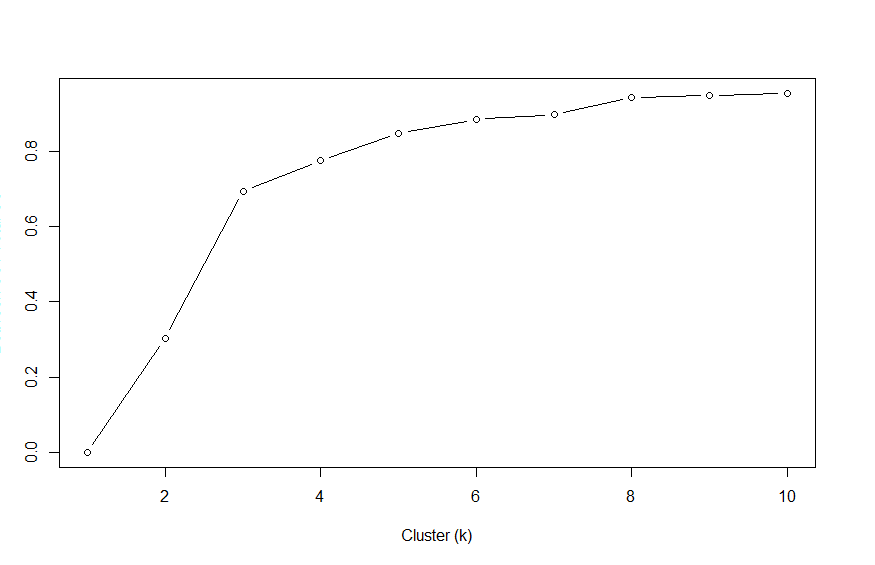


Figure 8: K-Optimization

In the attempt to find the optimal k, the model loops through different values of k, gets the ratio of the between sum of squares to total sum of squares for each model. The result of this ratio is plotted against the number of clusters. The shoulder/kink of the plot is assumed to be the optimal k. The value of the optimal k is 3.

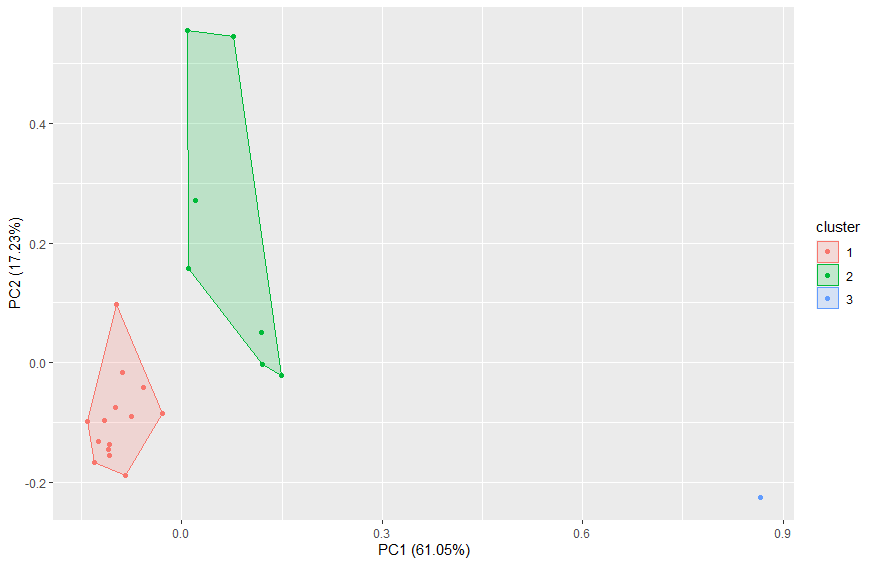
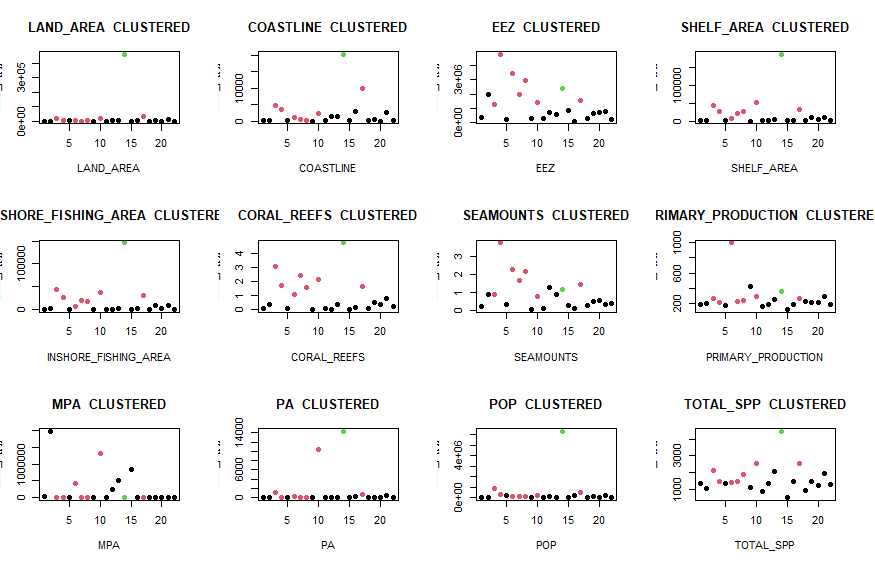


Figure 9: Cluster Plot

The cluster plot above shows the 3 clusters formed by the model from the data, and it can be said that there is a significant distinction between these three clusters. And there is no overlapping between the clusters; with the PCA component plotted with collectively accounting for over 75% of the variability involved.



The above scatter plot simply shows the result of using the Kmeans cluster model to group the observations of the different variables.

|  |  |
| --- | --- |
| **PICT prediction actual**  1 American Samoa 1 Polynesia  2 Cook Islands 1 Polynesia  3 Fiji 2 Melanesia  4 French Polynesia 2 Polynesia  5 Guam 1 Micronesia  6 Kiribati 2 Micronesia  7 Marshall Islands 2 Micronesia  8 Micronesia (FSM) 2 Micronesia  9 Nauru 1 Micronesia  10 New Caledonia 2 Melanesia  11 Niue 1 Polynesia  12 Northern Mariana Islands 1 Micronesia  13 Palau 1 Micronesia  14 Papua New Guinea 3 Melanesia  15 Pitcairn 1 Polynesia  16 Samoa 1 Polynesia  17 Solomon Islands 2 Melanesia  18 Tokelau 1 Polynesia  19 Tonga 1 Polynesia  20 Tuvalu 1 Polynesia  21 Vanuatu 1 Melanesia  22 Wallis and Futuna 1 Polynesia | **1 2 3**  **Melanesia** 1 3 1  **Micronesia** 4 3 0  **Polynesia** 9 1 0 |

The above output compares the cluster output from the kmeans method to the actual output. It would be noticed that the method performed better in clustering the region “Polynesia” into cluster 1 and was not so accurate in clustering the other regions (“Micronesia” and “Melanesia”). This could imply that the model does not account for some variabilities in the data.

**b. Hierarchical Cluster**

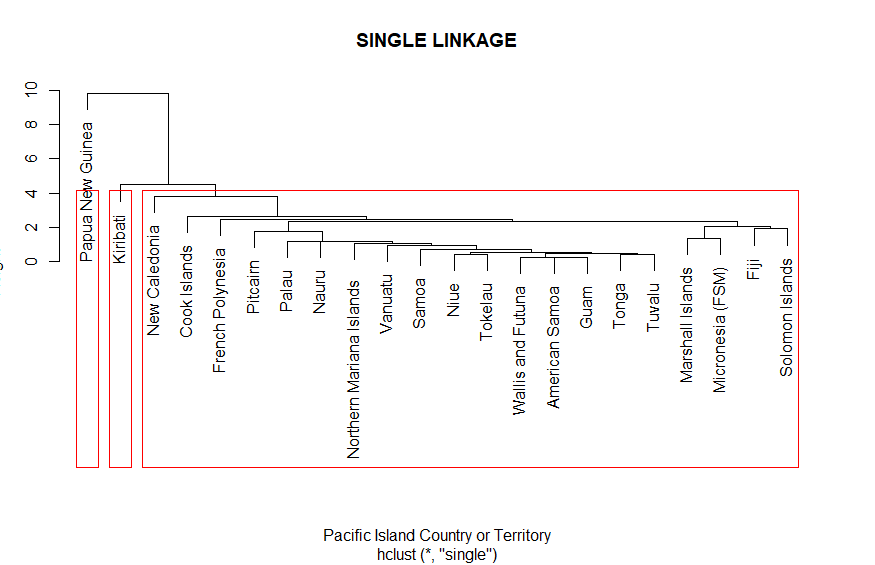


Figure 10: SINGLE LINKAGE DENDROGRAM

The dendrogram above is created using the Single Linkage Method, and it will be observed that this method of cluster formation minimises the distance between unit groups to form a cluster. With 3 cluster groups selected to classify the data, 2 clusters consisting of a unit observation each and the third cluster consisting of the remaining 20 observations were formed. This signifies that the 20 observations in the third cluster have the smallest distance among each unit compared to the other 2 clusters that are significantly farther from the cluster.

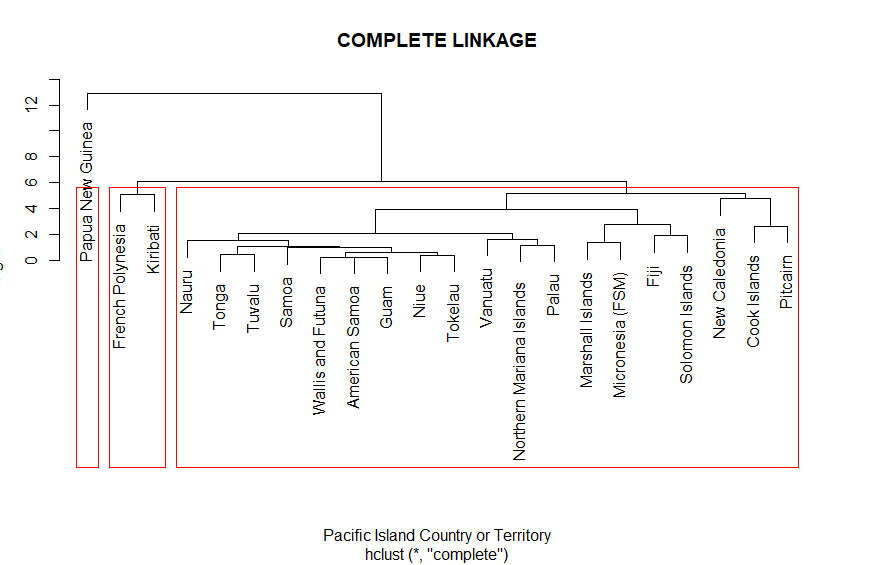


Figure 11: COMPLETE LINKAGE DENDOGRAM

The complete linkage dendrogram is formed by maximizing the distance between points when computing new distance points. It can be seen that the clusters are tighter than the single linkage method. From the above figure, it can be seen that clusters are mostly formed in pairs and this indicates that those pairs are at the farthest distance from each other, and this helps to create a balance in the dendrogram.