MATH38161 Coursework

Rutwik Mudholkar 10327919

Section 1

In this report, we will perform clustering analysis on the Palmer station penguin data set, found here, based on its intrinsic categories. We will use a variety of clustering algorithms and compare their efficacy using the known categorizations. Our dataset consists of 333 penguin data points, with 4 measured variables for each: bill length (mm), bill depth (mm), flipper length (mm) and weight (g). Each penguin is categorized into their sex $\in$ {male, female}

```
table(L.sex)
```

```
## L.sex
## female   male
##    165    168
```

, their species $\in$ {Adelie, Chinstrap, Gentoo}

```
table(L.species)
```

```
## L.species
##    Adelie Chinstrap    Gentoo
##       146        68       119
```

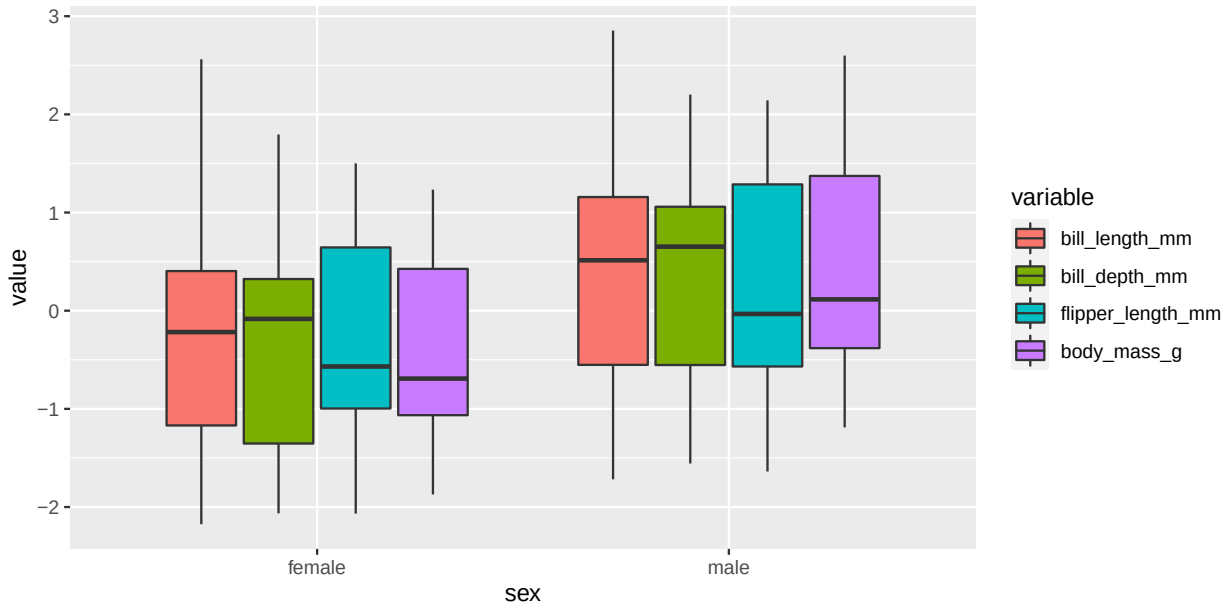
, and the island they live on $\in$ {Biscoe, Dream, Torgersen}.

```
table(L.islands)
```

```
## L.islands
##    Biscoe     Dream Torgersen
##       163       123        47
```

These categorizations are known beforehand; our clustering analysis obective will be to identify the correct species and island categories for each penguin. Whether we analyze the sexes separately depends on their respective standardized variable distributions.

```
t <- cbind(normalize(as.data.frame(X.penguins), method="standardize"), sex=L.sex)
tsplit <- split(t, t$sex)
m.df <- melt(rbind(tsplit$male, tsplit$female), id.var="sex")
ggplot(m.df, aes(x=sex, y=value)) + geom_boxplot(aes(fill=variable))
```



Comparing variables pairwise between sexes, the distributions are fairly similar with respect to the skew and interquartile range. However, the variables means are significantly larger for the males. Grouping the two sexes into one dataset may result in high uncertainty due to homogenizing the data, and removing key distinguishing features present when comparing variables for just one sex. We will therefore split the dataset by sex, and perform species cluster analysis separately on both. We will combine the sexes into one dataset for island clustering however, since it will be more dependant on the species itself.

Section 2

We will use Principal Component Analysis (PCA) for species clustering. PCA is closely linked to PCA whitening, and works by reducing the dimensionality of datasets, whilst boosting interpretability and minimizing information loss. This is done by creating new uncorrelated variables from the original, likely correlated variables (flipper length, bill length etc.) that maximize the variance. These new variables are called the Principal Components (PCs), where each PC carries a succesively decreasing proportion of the total variation. We select only the first few PCs that achieve close to 100% of the total variation as our axes to recast our data along, and hence identify clusters based on the data point labels.

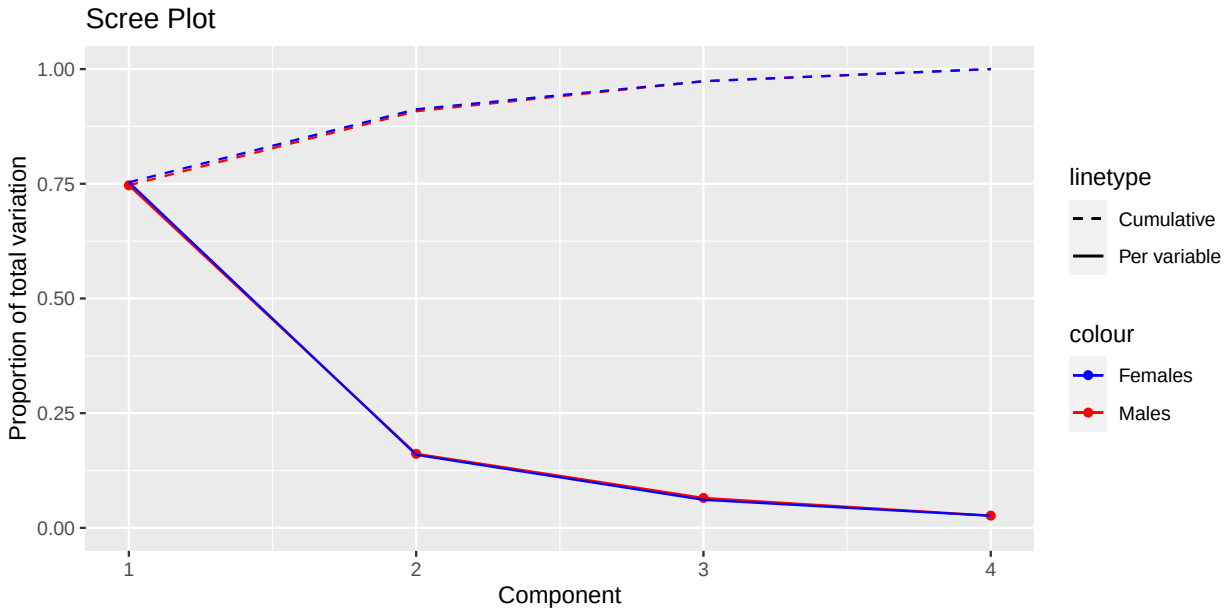
Island clustering will be more difficult, since the original variables are naturally intrinsic to the penguin species rather than the favoured location of the species. We can assume there will be significant overlap of clusters, since the species is likely the primary differential. We will therefore use a finite Gaussian Mixture model (GMM) fitted via the EM algorithm, a probabilistic Model-based clustering method that utilizes the sizes of clusters (how many penguins per island). GMMs work by modelling our multimodal distribution with multiple unimodal distributions and then learning their parameters, which are the component weights, means, and covariances. GMMs are ideal for our island clustering since they produce optimal clusters for complex, non-linear decision boundaries, and handle uncertainty well when a data point is close to more than one cluster centre. It is more computationally expensive than an algorithmic method, but our dataset is small, so this is negligible.

Section 3

3.1 Species

We will now perform our clustering analysis, starting with PCA for species clustering. We first pre-process our data by scaling and standardizing, then compute PCA components and plot the proportion of total variance contributed by each component.

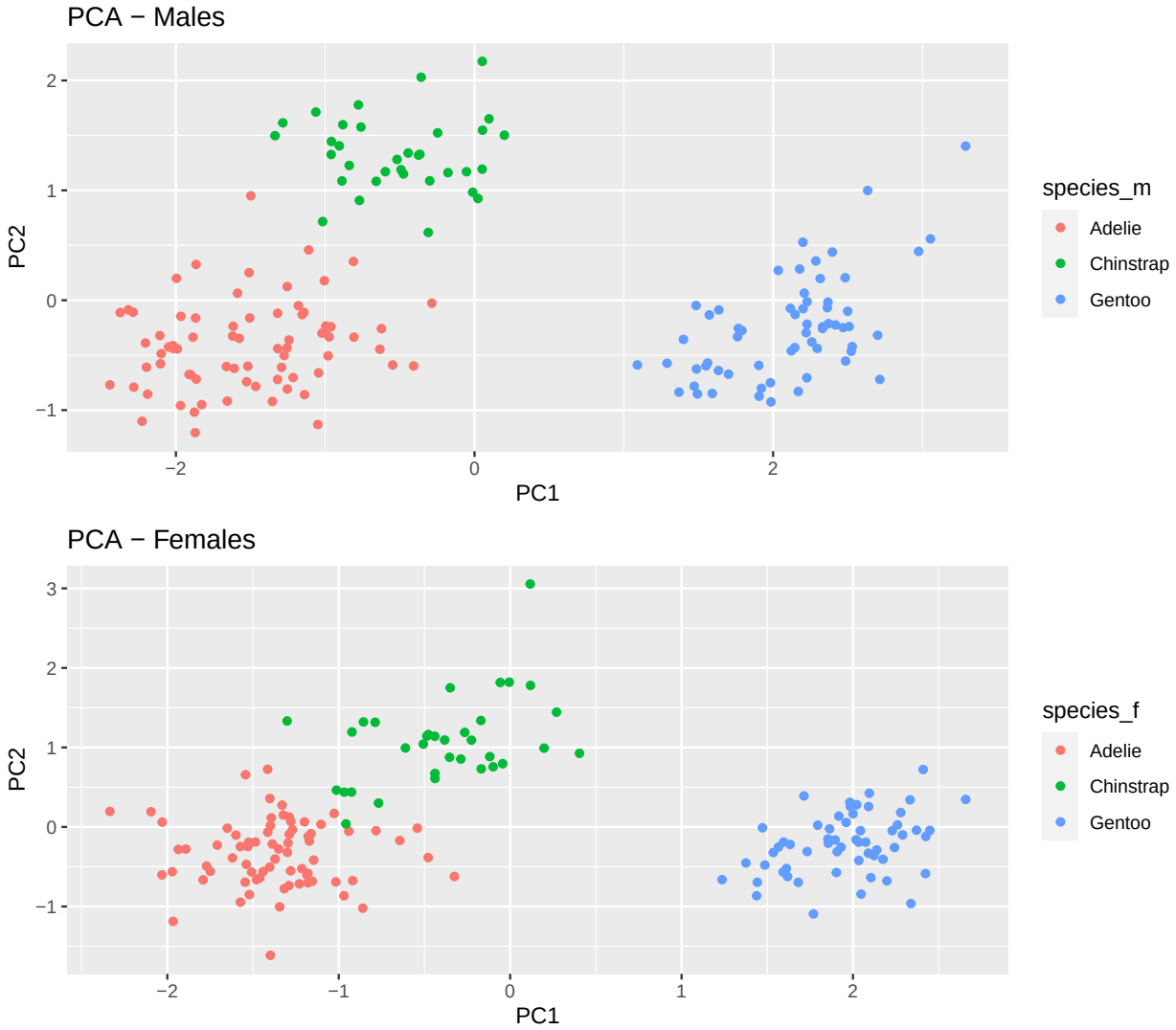
```
X = scale((X.penguins[, 1:4]), scale=TRUE)
processX <- function(sex) {
  Xt =  split(as.data.frame(X), L.sex)[[sex]]
  pca = prcomp(Xt)
  V = (pca$sdev^2)/(sum(pca$sdev^2))
  return(list(V = V, pca = pca))
}
rf = processX("female"); rm = processX("male")
qplot(1:4, rm$V, geom=c("point", "line"), xlab="Component", ylim=c(0,1),
      ylab="Proportion of total variation", color="red", linetype = "solid") +
  ggtitle("Scree Plot") +
  geom_line(aes(y=cumsum(rm$V), color="red", linetype = "dashed")) +
  geom_line(aes(y = rf$V, color="blue", linetype = "solid")) +
  geom_line(aes(y=cumsum(rf$V), color="blue", linetype = "dashed")) +
  scale_color_manual(labels = c("Females", "Males"), values = c("blue", "red")) +
  scale_linetype_manual(labels = c("Cumulative", "Per variable"),
                        values = c("dashed", "solid"))
```



We can see that only the first two PCA components are needed to achieve ~90% of the total variation for both male and female penguins, which is satisfactory.

```
vplayout <- function(x, y) viewport(layout.pos.row = x, layout.pos.col = y)
species_m <- split(L.species, L.sex)$male
species_f <- split(L.species, L.sex)$female
p1 <- qplot(data=as.data.frame((rm$pca)$x),PC1,PC2,color=species_m,main="PCA - Males")
p2 <- qplot(data=as.data.frame((rf$pca)$x),PC1,PC2,color=species_f,main="PCA - Females")
```

```
grid.newpage()
pushViewport(viewport(layout = grid.layout(2, 1)))
print(p1, vp = vplayout(1, 1)); print(p2, vp = vplayout(2, 1))
```



From our PCA scatter plots for out first two components, we see 2 to 3 clusters. There is a distinct cluster for high PC1 for both males and females, which represents the Gentoo species. For low PC1, we see 2 semi-distinct clusters categorised by either high or low PC2, representing the Chinstrap and Adelie species respectively. There is a clearer cluster boundary between them for males, but slightly more overlap and uncertainty for females. Overall, we can conclude that our PCA clustering can categorize the dataset by species reasonably well.

3.2 Islands

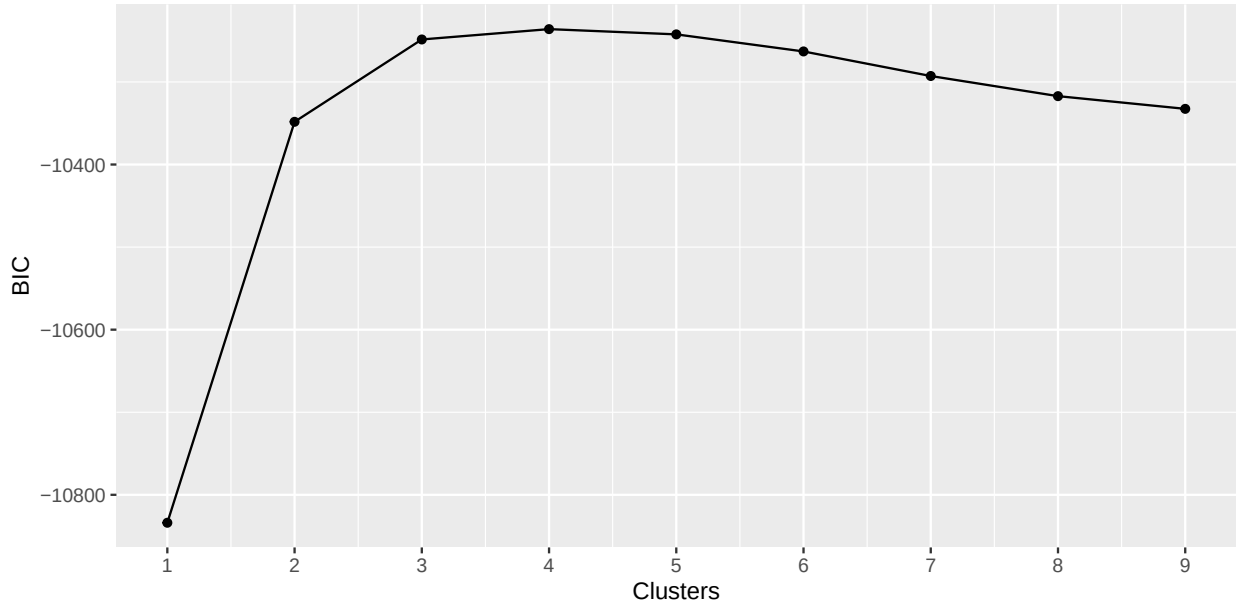
We now perform our GMM analysis for island clustering.

```
gmm = Mclust(X.penguins)
summary(gmm)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEE (ellipsoidal, equal shape and orientation) model with 4 components:
##
##  log-likelihood   n df       BIC       ICL
##       -5025.089 333 32 -10236.04 -10276.59
##
## Clustering table:
##   1   2   3   4
## 150  64  59  60
```

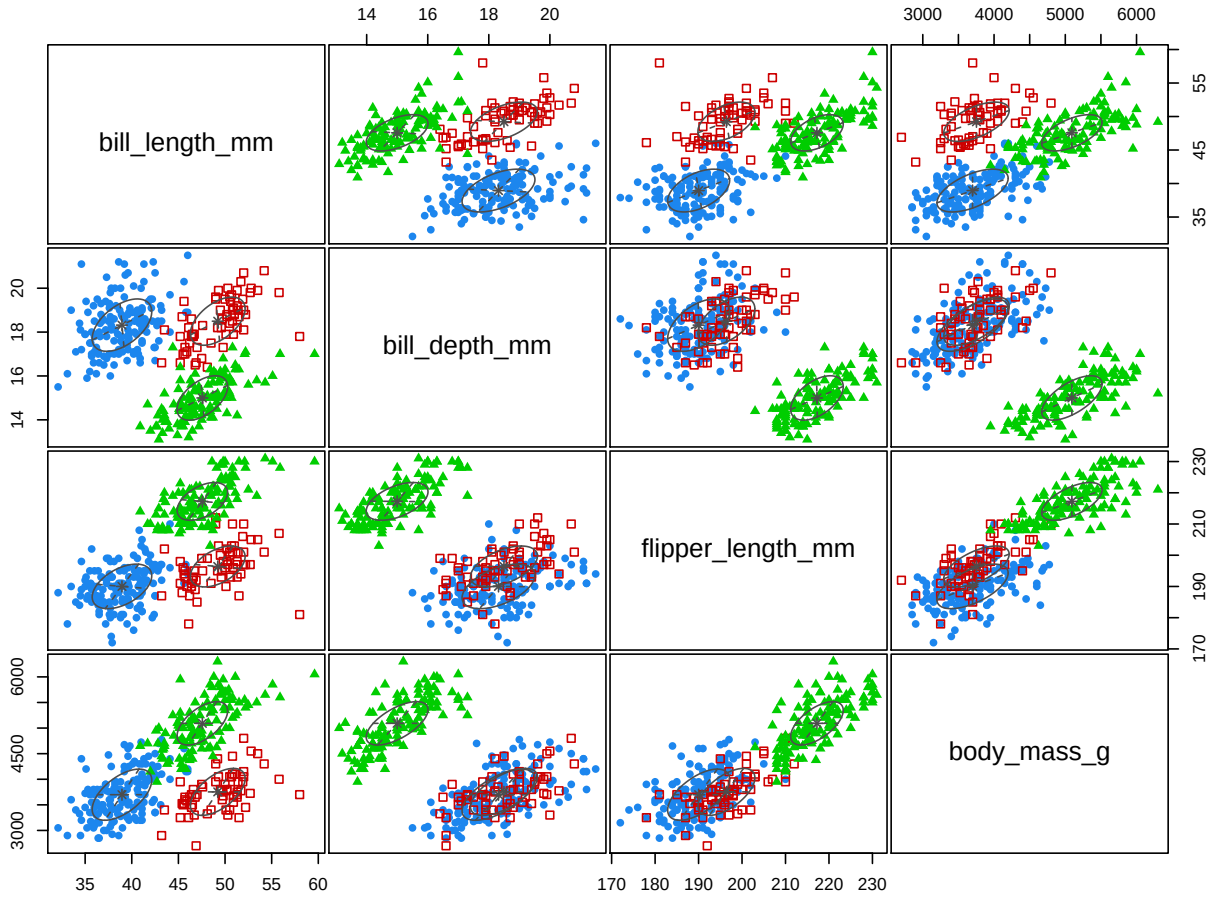
From our mixture model we identify four clusters, and produce the likelihood and the Bayesian Information criterion (BIC) score. Maximising the BIC score balances model complexity and goodness of fit to determine how many clusters to use. Using the 'VEE' covariance parameterisation as given, we can further plot BIC against a range of cluster numbers.

```
BIC <- mclustBIC(X.penguins)
qplot(1:9, value, data=melt(BIC[,"VEE"]), geom=c("point", "line"),
      xlab="Clusters",ylab="BIC") +
      scale_x_continuous(labels = 1:9, breaks = 1:9)
```



We have 3 islands so want 3 clusters. We can see that 3 clusters produces an almost negligible difference in BIC from 4 clusters, meaning our GMM model is still usable. We can now create a new non-optimized but practical mixture model by enforcing 3 clusters.

```
gmm3 = Mclust(X.penguins, G=3, verbose=FALSE)
plot(gmm3, what="classification")
```



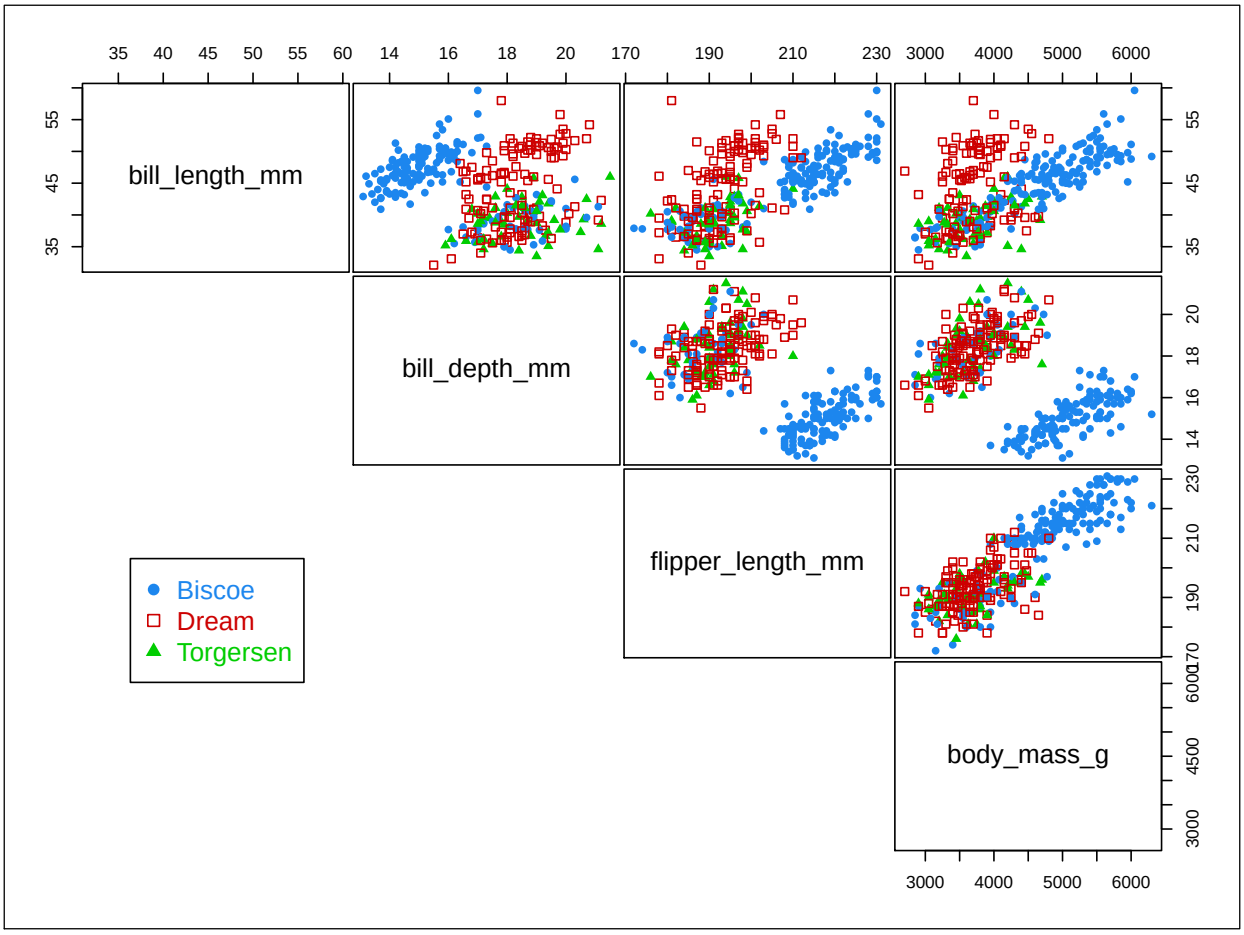
Plotting pairwise scatter plots with points marked according to their model classifications, we can see distinct separation of the green cluster in all cases. We see separation of the red and blue clusters in some cases, but significant overlap between them in a few other cases.

```
table(gmm3$classification, L.islands)
```

```
##    L.islands
##     Biscoe Dream Torgersen
##  1     44    60        47
##  2      0    63         0
##  3    119     0         0
```

Comparing the model clustering with the island data given in Section 1, it appears to have a superficially high misclassification error. Cluster 1 contains penguins from all 3 islands in roughly equal proportions. Clusters 2 and 3 clearly represent Dream and Biscoe with no island overlap, but are underpopulated due to large numbers of their penguins falling into cluster 1. By plotting the same graphs but colour coding by island labels instead,

```
clp <- clPairs(X.penguins, L.islands, lower.panel = NULL)
clPairsLegend(0.1, 0.4, class = clp$class, col = clp$col, pch = clp$pch)
```



we can see that Biscoe Island corresponds with the distinct green cluster (cluster 3) from the previous plot, and Dream Island with the red cluster (cluster 2) from the previous plot. As expected, the blue cluster (cluster 1) from the previous plot contains all the Torgerson Penguins and chunks from the other two islands. If we instead see the number of each species on each island from our data:

```
table(L.species, L.islands)
```

```
##            L.islands
## L.species   Biscoe Dream Torgersen
##   Adelie        44    55        47
##   Chinstrap      0    68         0
##   Gentoo       119     0         0
```

we can see that while our GMM fails to produce distinct clusters for each island, it models the underlying species distribution extremely well. This tells us that our GMM has inadvertently performed species clustering again instead.

We propose that distinct island clustering is only feasible if species and islands are correlated. Since Adelie penguins decided to inhabit every island, there is no theoretical way to cluster islands distinctly with our given variables, since the variables only correspond to penguin specific features like their species and sex. This could also tell us that while the islands may be separate, there hasn't been a significant amount of time passed for evolution to yield different penguin features based purely on their island, regardless of species.