Unsupervised learning report

Student ID: 10275181
Name: Luka Ross

Section 1

In this section we introduce the data and present a basic plot and summary statistics.
First we load in the data and labels.

```
data = load("penguins.rda")
data
```

```
## [1] "X.penguins" "L.species"  "L.islands"  "L.sex"
```

Checking the dimensions of the data we see that there are 333 observations in 4 variables.

```
dim(X.penguins)
```

```
## [1] 333   4
```

```
colnames(X.penguins)
```

```
## [1] "bill_length_mm"    "bill_depth_mm"     "flipper_length_mm"
## [4] "body_mass_g"
```

check for any missing values

```
apply(X.penguins, 2, function(x) any(is.na(x)))
```

```
##    bill_length_mm     bill_depth_mm flipper_length_mm       body_mass_g
##             FALSE             FALSE             FALSE             FALSE
```

Summary statistics of the data including the standard deviation

```
summary(X.penguins)
```

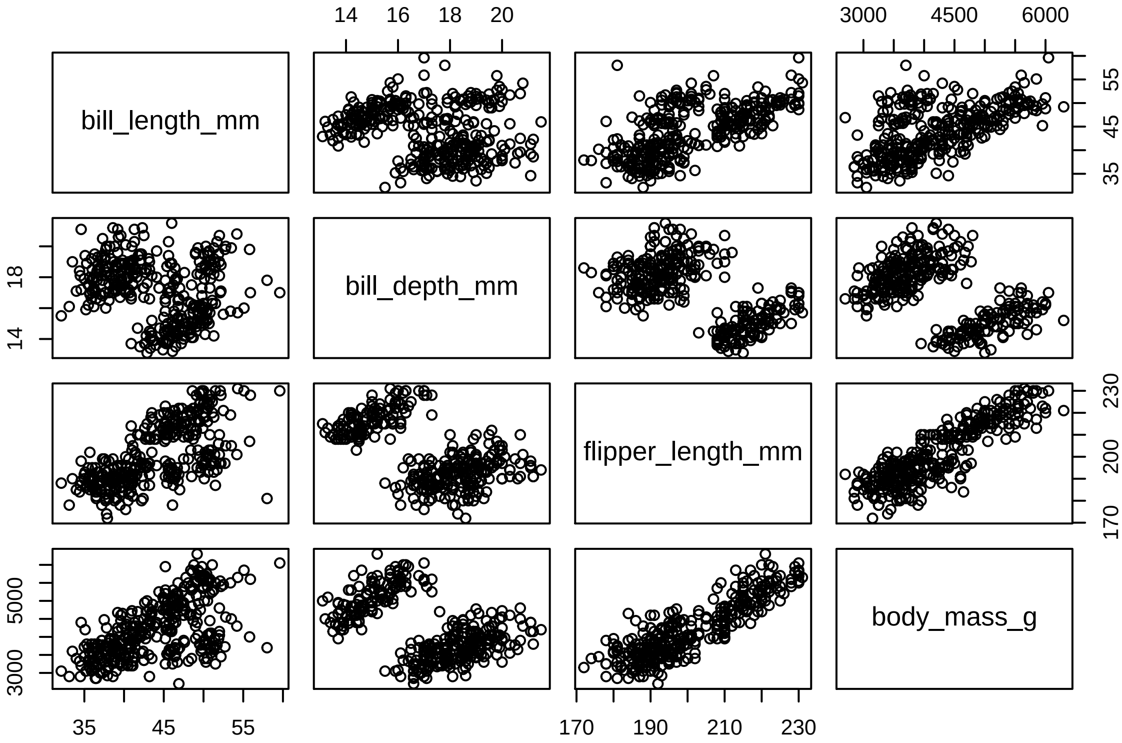
```
##  bill_length_mm  bill_depth_mm   flipper_length_mm  body_mass_g
##  Min.   :32.10   Min.   :13.10   Min.   :172       Min.   :2700
##  1st Qu.:39.50   1st Qu.:15.60   1st Qu.:190       1st Qu.:3550
##  Median :44.50   Median :17.30   Median :197       Median :4050
##  Mean   :43.99   Mean   :17.16   Mean   :201       Mean   :4207
##  3rd Qu.:48.60   3rd Qu.:18.70   3rd Qu.:213       3rd Qu.:4775
##  Max.   :59.60   Max.   :21.50   Max.   :231       Max.   :6300
```

```
apply(X.penguins, 2,sd)
```

```
##    bill_length_mm     bill_depth_mm flipper_length_mm       body_mass_g
##          5.468668          1.969235         14.015765        805.215802
```

from this plot of each of the variables you can see at least two clusters in most cases

```
table = as.data.frame(X.penguins)
plot(table)
```



For this analysis we will not be splitting the data into male and female birds.

Section 2

In this section we outline the clustering methods to be used in section 3.

K Means Clustering

The first method we will use to cluster the data is K means clustering. The K means algorithm requires a given number of clusters so in this analysis we will run the algorithm multiple times and assess the resulting within group and between group variation to determine the optimum number of clusters to run the algorithm on. Here is an outline of the general k means method.
Initially the data points are randomly allocated to one the K groups. The function $c(x_i) \in \{1, ..., K\}$ assigns cluster labels to each data point (group allocations). The following steps are then iterated until convergence.

- Compute the group means for each group
- Update the group allocations for each data point, based on the smallest euclidean distance from the point to each group mean.

Hierarchical Clustering

The next method we will use to cluster the data is agglomerative hierarchical clustering, using euclidean distance for the individual points and Wards minimum variance approach to calculate the distance between groups.
The general outline for the method is as follows. Initially a matrix containing all the pair-wise distances is computed using our chosen distance metric. The algorithm then iterates the following steps.

- First the closest pairs of objects are grouped together into a set. In the tree this is represented by an internal node.
- Then the distance matrix is updated by computing the distance between the new set and all other objects.

The algorithm terminates when the new set contains all points in the input data.
In this report we use wards minimum variance approach. When the algorithm is comparing the distance of sets it merges the two sets that give the smallest increase in within-group variation.

Section 3

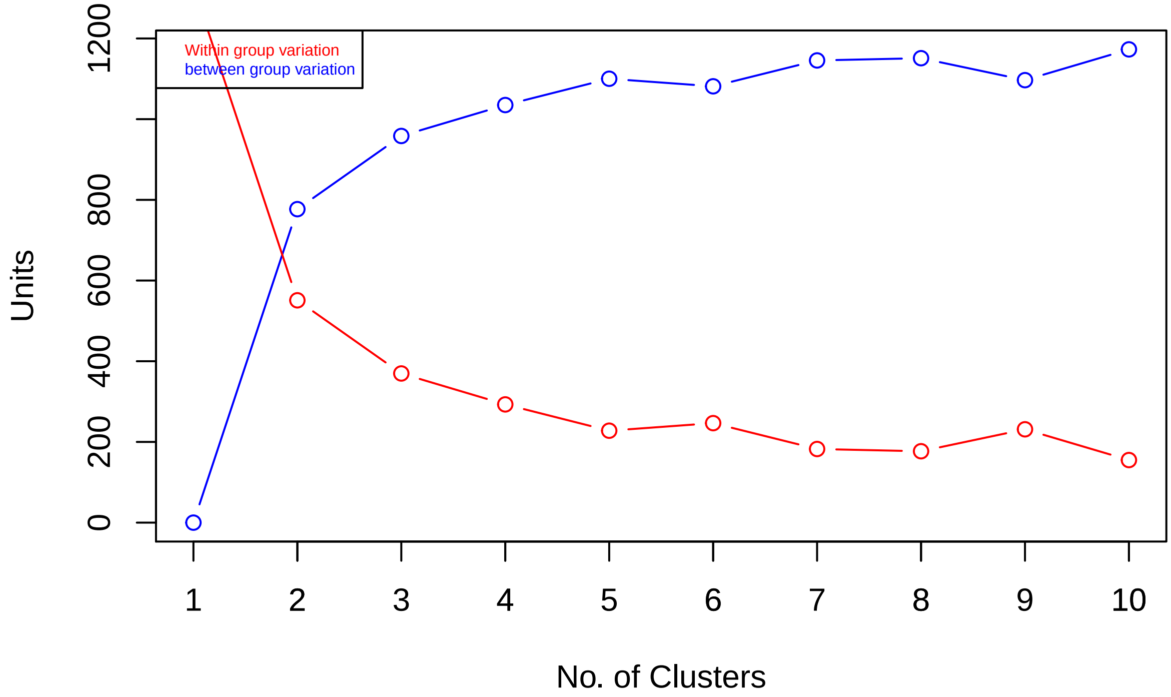
In this section we will run the previously discussed methods and interpret the results.
First standardize the data.

```
std.table = scale(table)
```

K Means Clustering

```
set.seed(53) # ensure reproducibility

b = numeric(10)
w = numeric(10)

for (i in 1:10){ #run k-means 10 times and store btween and within variation
  k.means =  kmeans(std.table,i)
  b[i] = k.means$betweenss
  w[i] = k.means$tot.withinss
}
```

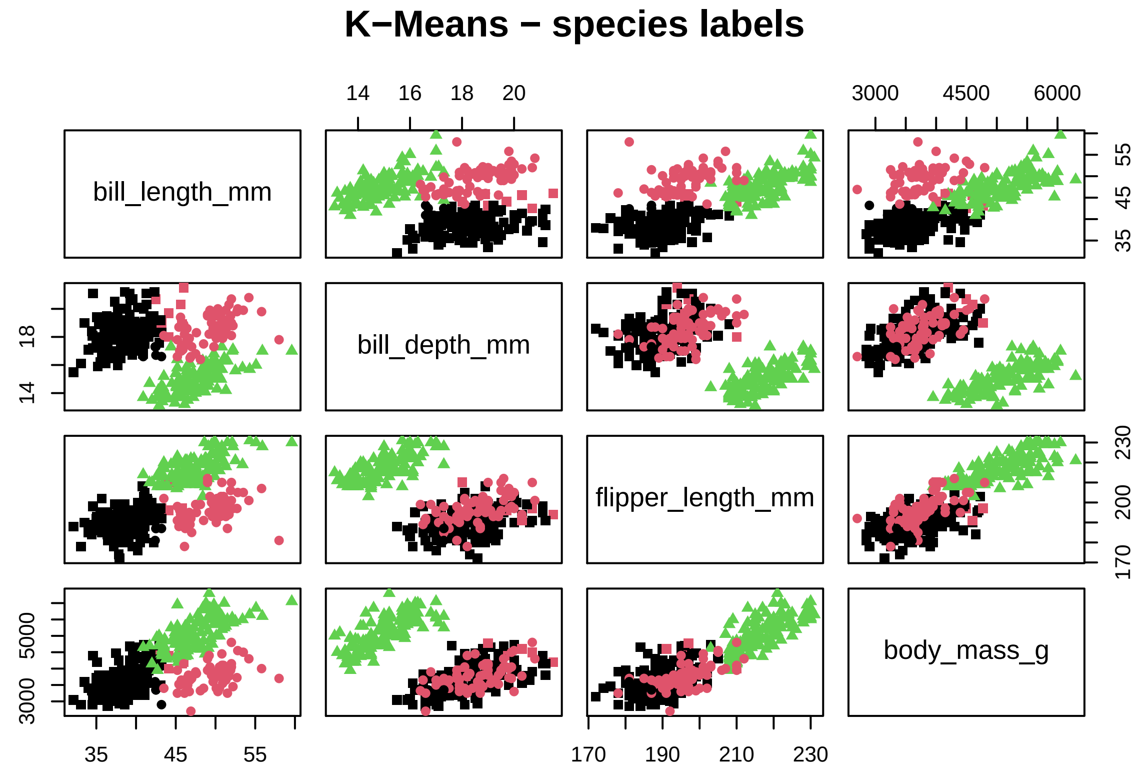
```
plot(c(b),type = "b",xlab = "No. of Clusters",ylab = "Units",col="blue") # plot variation
points(w,col="red",type = "b")
legend("topleft",legend = c("Within group variation","between group variation"),
       text.col=c("red","blue"),cex = 0.5)
axis(1,at=seq(1,10,1))
```



This plot tells us that after 3 clusters there is no substantial increase in between group variation and no substantial decrease in within group variation. Thus the optimum amount of clusters for this analysis is 3.

```
set.seed(53)
k.means.cluster = kmeans(std.table,3) #run k-means for K=3
```

```
plot(table, col=k.means.cluster$cluster, pch=as.integer(L.species)+14,
     main="K-Means - species labels")
```



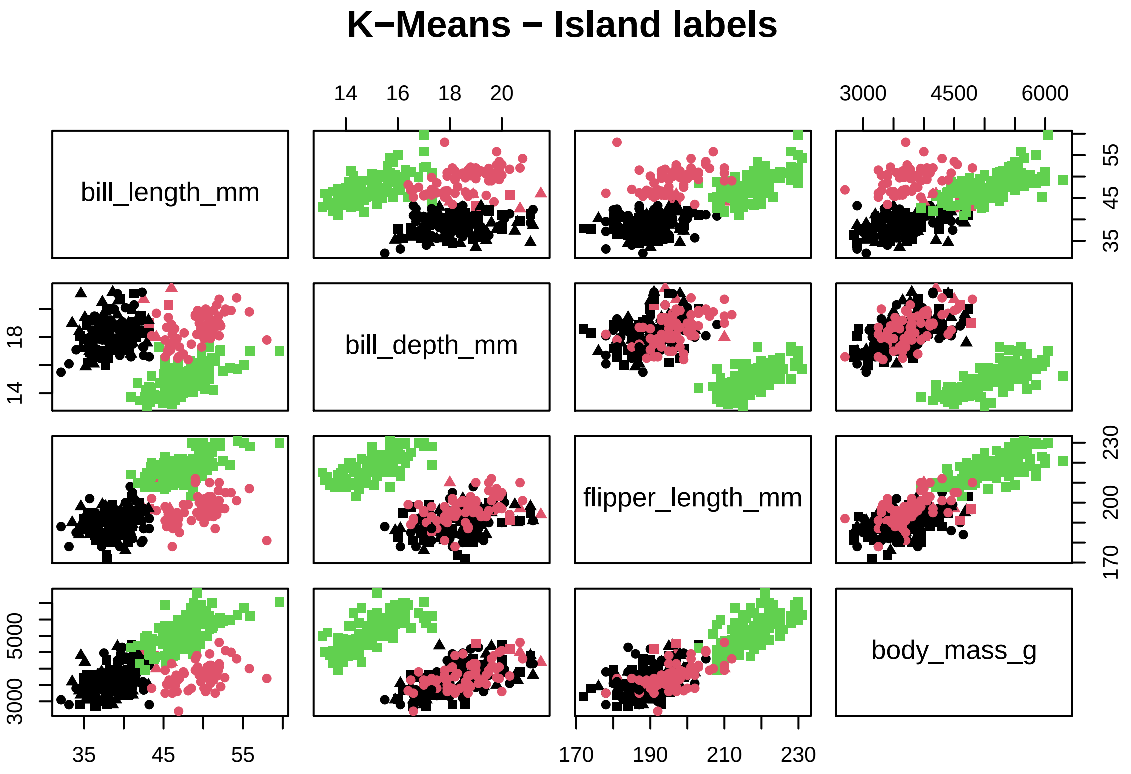
In the above plot the colors represent the grouping by k-means and shapes represent the true grouping.

```
table(L.species,k.means.cluster$cluster) #misclassification table
```

```
## 
## L.species     1   2   3
##   Adelie    139   7   0
##   Chinstrap   5  63   0
##   Gentoo      0   0 119
```

K-means preforms well when classifying the species. A total of 12 birds were incorrectly classified, 7 Adelie and 5 Chinstrap. All Gentoo penguins were correctly grouped.

```
plot(table, col=k.means.cluster$cluster, pch=as.integer(L.islands)+14,
     main="K-Means - Island labels")
```



```
table(L.islands,k.means.cluster$cluster) #misclassification table
```

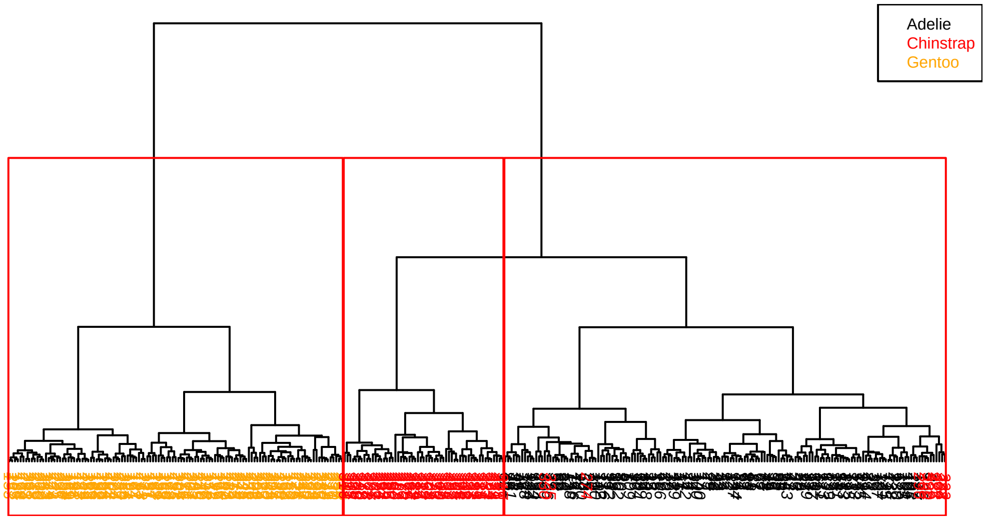
```
##
## L.islands     1   2   3
##   Biscoe     42   2 119
##   Dream      59  64   0
##   Torgersen  43   4   0
```

The island labels proved more problematic to classify. All islands have misclassifications with penguins from Torgersen island being the most correctly classified with only 4 incorrect. Birds from Biscoe and Dream island have large numbers of data points incorrectly classified. 47% misclassification in brids from dream island and 26% misclassification in brids from Biscoe island.

Hierarchical Clustering

```
d = dist(std.table) #compute distance matrix with euclidean distance

hc1 = hclust(d,method = "ward.D2") #preform HC with wards method

hc1_groups = cutree(hc1,3) #cut the tree into 3 clusters
```

```r
library("ape")

col.species = c("black", "red", "orange") #initialize color map
colMap = col.species[as.integer(L.species)]
names(colMap) = rownames(X.penguins)

plot(as.phylo(hc1),  tip.color = colMap, #plot dendrogram
  label.offset = .5, cex = 0.5, direction="downwards")
rect.hclust(hc1, k=3, border="red")
legend("topright", legend=levels(L.species), text.col=col.species, cex=0.5)
```

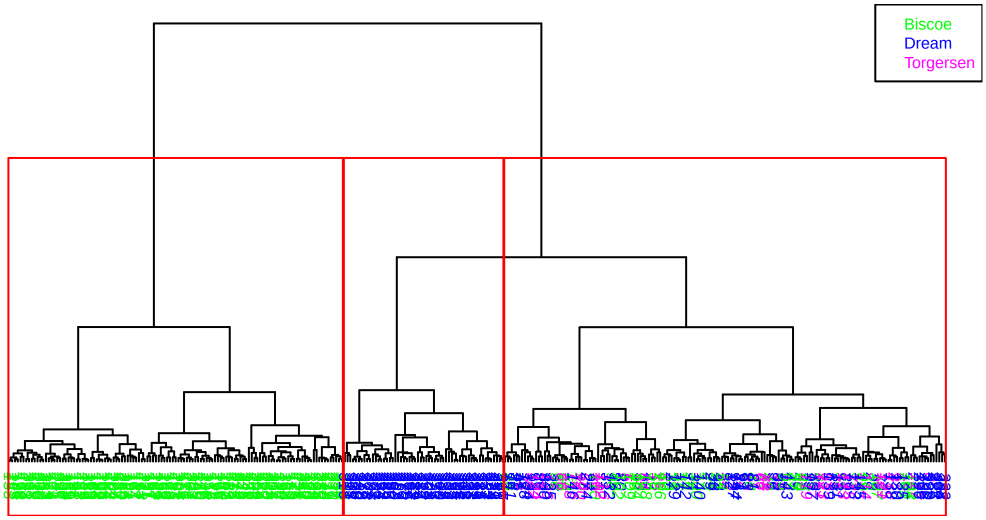


```r
table(L.species,hc1_groups) #misclassification table
```

```
##            hc1_groups
## L.species     1   2   3
##   Adelie    146   0   0
##   Chinstrap  11   0  57
##   Gentoo      0 119   0
```

As we can see from dendrogram and table all Adelie and Gentoo penguins were correctly grouped with no misclassifications. 11 Chinstrap penguins were incorreclty grouped with the Adelie penguins.

```
col.islands = c("green", "blue", "magenta") #initialize color map
colMap.2 = col.islands[as.integer(L.islands)]
names(colMap) = rownames(X.penguins)


plot(as.phylo(hc1),  tip.color = colMap.2, #plot dendrogram
  label.offset = .5, cex = 0.5, direction="downwards")
rect.hclust(hc1, k=3, border="red")
legend("topright", legend=levels(L.islands), text.col=col.islands, cex=0.5)
```



```
table(L.islands,hc1_groups) #misclassification table
```

```
##              hc1_groups
## L.islands      1   2   3
##   Biscoe      44 119   0
##   Dream       66   0  57
##   Torgersen   47   0   0
```

The island labels are again less successfully classified. While the correct number of penguins from Torgersen island are grouped together, the cluster also contains a large number of incorrectly classified data points of penguins from both Dream and Biscode island. Brids from Dream island have the highest misclassification at 53% and birds from Biscoe have a misclassifciation of around 26%.
To conclude both K means and Hierarchical Clustering preform fairly similarly. Both methods cluster the species well with low misclassifications and have trouble classifying the island, displaying similar misclassifaction errors.