# Cluster Analysis - 19BCE1460

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
require("datasets")
penguinsdata <- read.csv("S:/WIN SEM 21-22/Data Visualization/Lab/penguins_lter.csv")
penguinsdata <- na.omit(penguinsdata)
glimpse(penguinsdata)
```

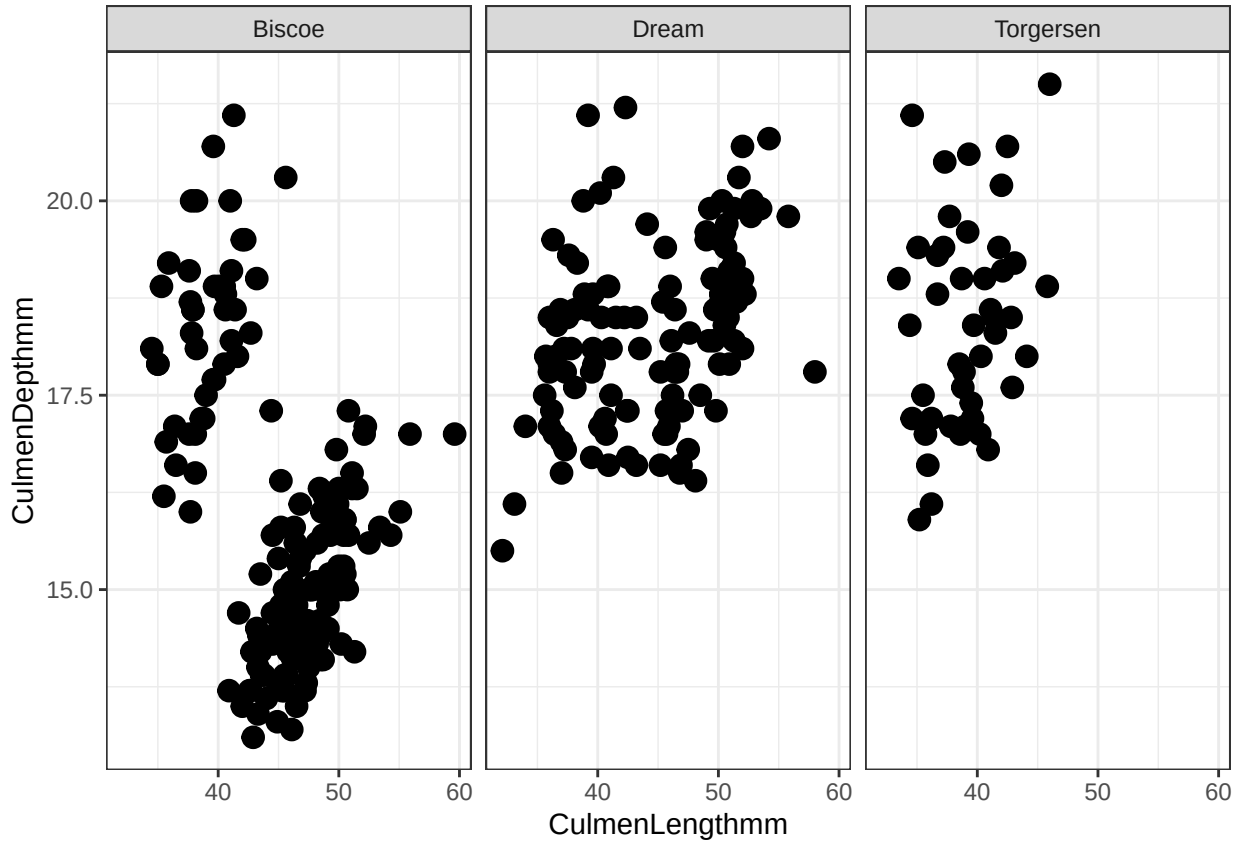
```
## Rows: 330
## Columns: 17
## $ studyName         <chr> "PAL0708", "PAL0708", "PAL0708", "PAL0708", "PAL0708~
## $ Sample.Number     <int> 2, 3, 5, 6, 7, 8, 10, 11, 15, 17, 18, 19, 20, 21, 22~
## $ Species           <chr> "Adelie Penguin (Pygoscelis adeliae)", "Adelie Pengu~
## $ Region            <chr> "Anvers", "Anvers", "Anvers", "Anvers", "Anvers", "A~
## $ Island            <chr> "Torgersen", "Torgersen", "Torgersen", "Torgersen", ~
## $ Stage             <chr> "Adult, 1 Egg Stage", "Adult, 1 Egg Stage", "Adult, ~
## $ Individual.ID     <chr> "N1A2", "N2A1", "N3A1", "N3A2", "N4A1", "N4A2", "N5A~
## $ Clutch.Completion <chr> "Yes", "Yes", "Yes", "Yes", "No", "No", "Yes", "Yes"~
## $ Date.Egg          <chr> "11-11-2007", "11/16/07", "11/16/07", "11/16/07", "1~
## $ CulmenLengthmm    <dbl> 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 42.0, 37.8, 34.6~
## $ CulmenDepthmm     <dbl> 17.4, 18.0, 19.3, 20.6, 17.8, 19.6, 20.2, 17.1, 21.1~
## $ FlipperLengthmm   <int> 186, 195, 193, 190, 181, 195, 190, 186, 198, 195, 19~
## $ BodyMassg         <int> 3800, 3250, 3450, 3650, 3625, 4675, 4250, 3300, 4400~
## $ Sex               <chr> "FEMALE", "FEMALE", "FEMALE", "MALE", "FEMALE", "MAL~
## $ Delta.15.N..o.oo. <dbl> 8.94956, 8.36821, 8.76651, 8.66496, 9.18718, 9.46060~
## $ Delta.13.C..o.oo. <dbl> -24.69454, -25.33302, -25.32426, -25.29805, -25.2179~
## $ Comments          <chr> "", "", "", "", "Nest never observed with full clutc~
```

```
head(penguinsdata)
```

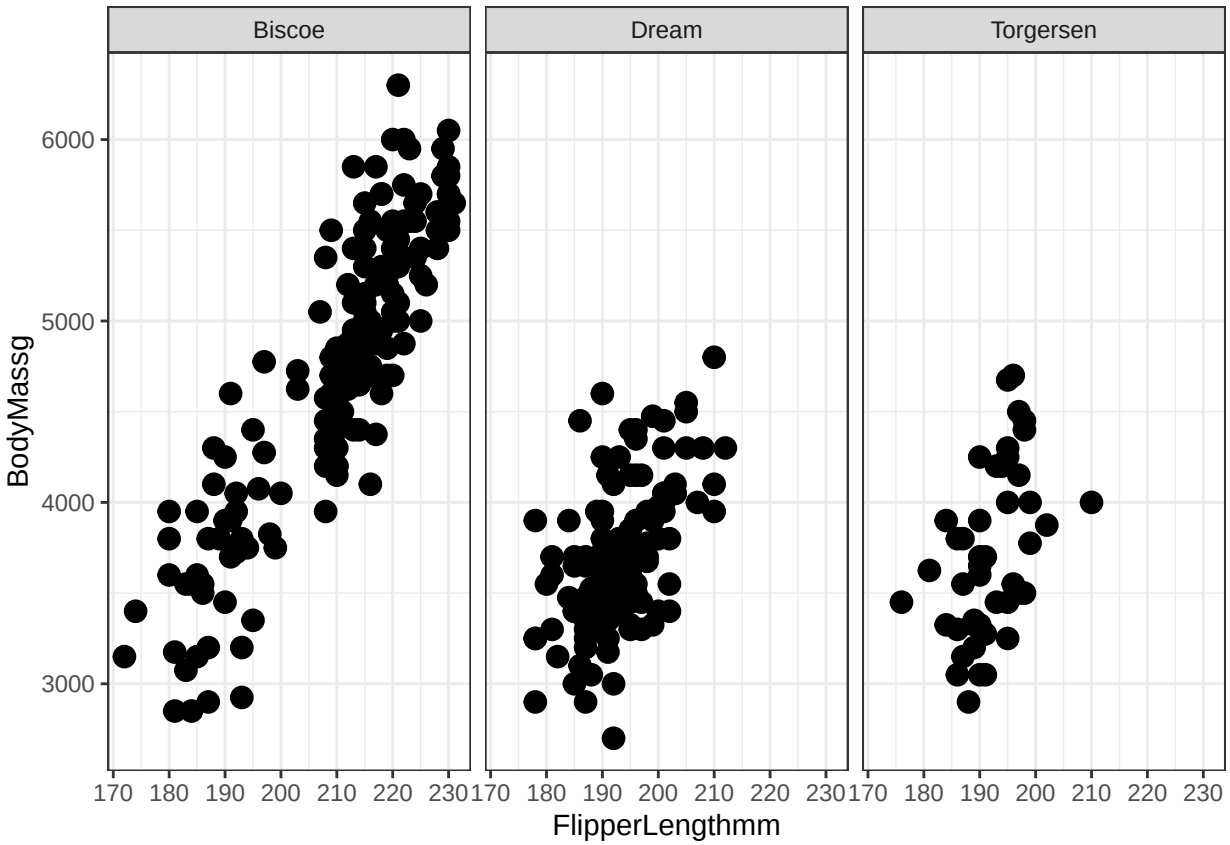
```
##   studyName Sample.Number                              Species Region    Island
## 2   PAL0708             2 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 3   PAL0708             3 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 5   PAL0708             5 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 6   PAL0708             6 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 7   PAL0708             7 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
## 8   PAL0708             8 Adelie Penguin (Pygoscelis adeliae) Anvers Torgersen
##                Stage Individual.ID Clutch.Completion   Date.Egg CulmenLengthmm
## 2 Adult, 1 Egg Stage          N1A2               Yes 11-11-2007           39.5
## 3 Adult, 1 Egg Stage          N2A1               Yes   11/16/07           40.3
## 5 Adult, 1 Egg Stage          N3A1               Yes   11/16/07           36.7
## 6 Adult, 1 Egg Stage          N3A2               Yes   11/16/07           39.3
## 7 Adult, 1 Egg Stage          N4A1                No   11/15/07           38.9
## 8 Adult, 1 Egg Stage          N4A2                No   11/15/07           39.2
##   CulmenDepthmm FlipperLengthmm BodyMassg    Sex Delta.15.N..o.oo.
## 2          17.4             186      3800 FEMALE           8.94956
## 3          18.0             195      3250 FEMALE           8.36821
## 5          19.3             193      3450 FEMALE           8.76651
## 6          20.6             190      3650   MALE           8.66496
## 7          17.8             181      3625 FEMALE           9.18718
## 8          19.6             195      4675   MALE           9.46060
##   Delta.13.C..o.oo.                              Comments
## 2         -24.69454
## 3         -25.33302
## 5         -25.32426
## 6         -25.29805
## 7         -25.21799 Nest never observed with full clutch.
## 8         -24.89958 Nest never observed with full clutch.
```

```
ggplot(penguinsdata)+
  geom_point(aes(x = CulmenLengthmm, y = CulmenDepthmm), stroke = 2)+
```

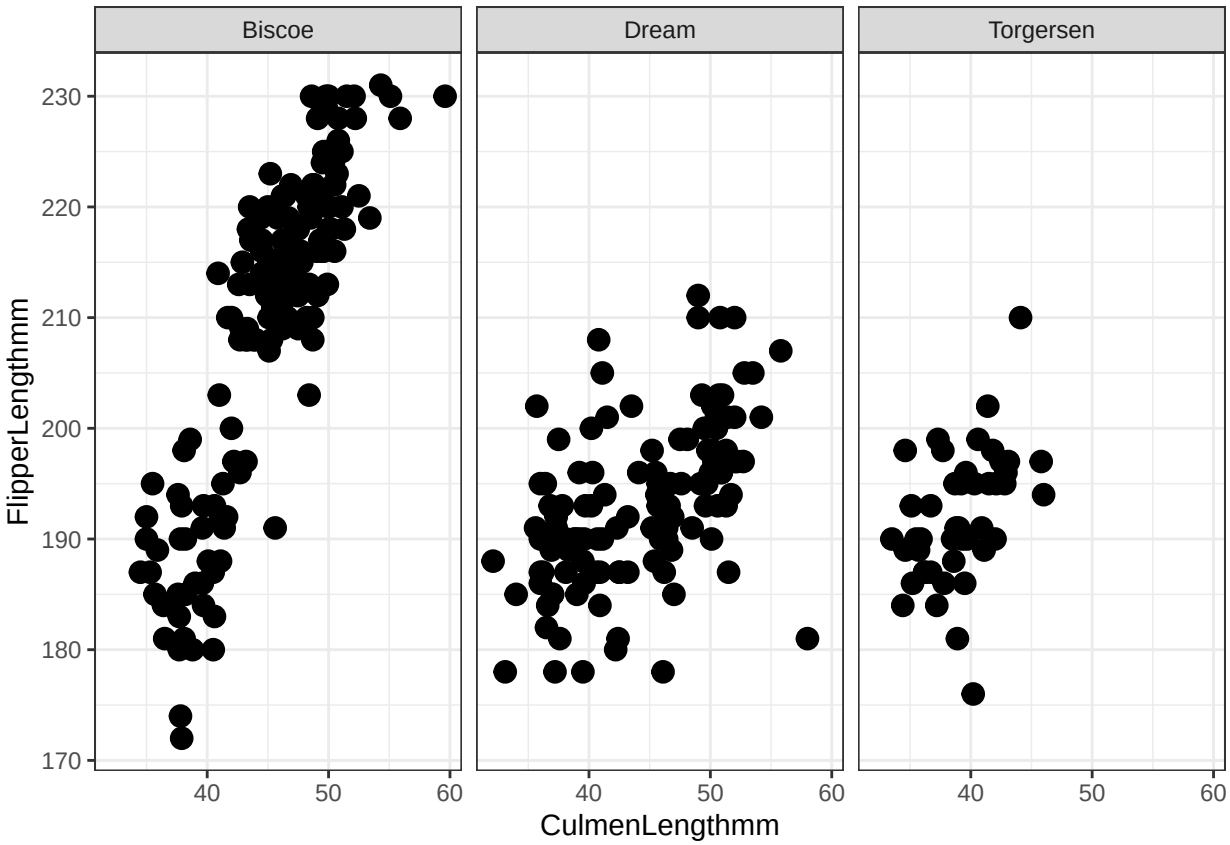
```
  facet_wrap(~ Island)+
  labs(x = "CulmenLengthmm", y = "CulmenDepthmm")+
  theme_bw()
```



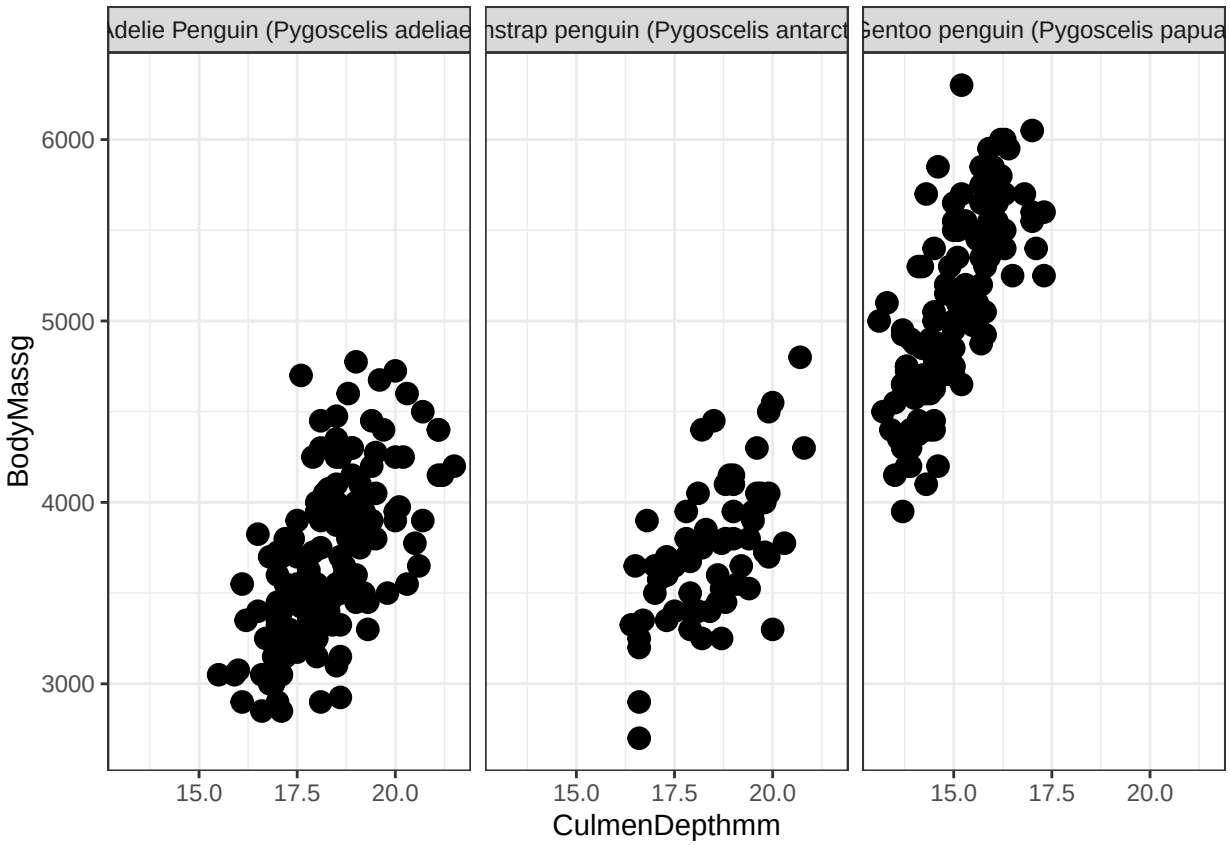

```
ggplot(penguinsdata)+
  geom_point(aes(x = FlipperLengthmm, y = BodyMassg), stroke = 2)+
  facet_wrap(~ Island)+
  labs(x = "FlipperLengthmm", y = "BodyMassg")+
  theme_bw()
```



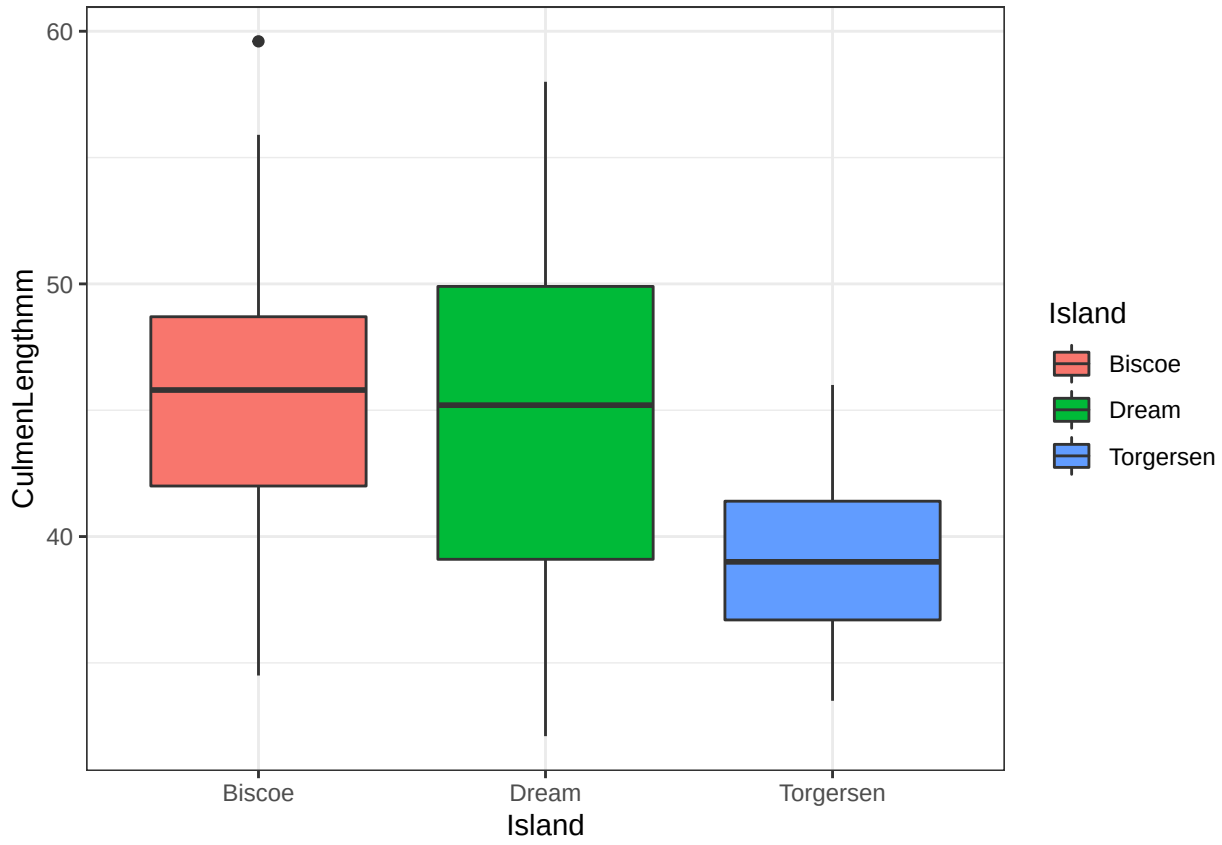

```
ggplot(penguinsdata)+
  geom_point(aes(x = CulmenLengthmm, y = FlipperLengthmm), stroke = 2)+
  facet_wrap(~ Island)+
  labs(x = "CulmenLengthmm", y = "FlipperLengthmm")+
  theme_bw()
```



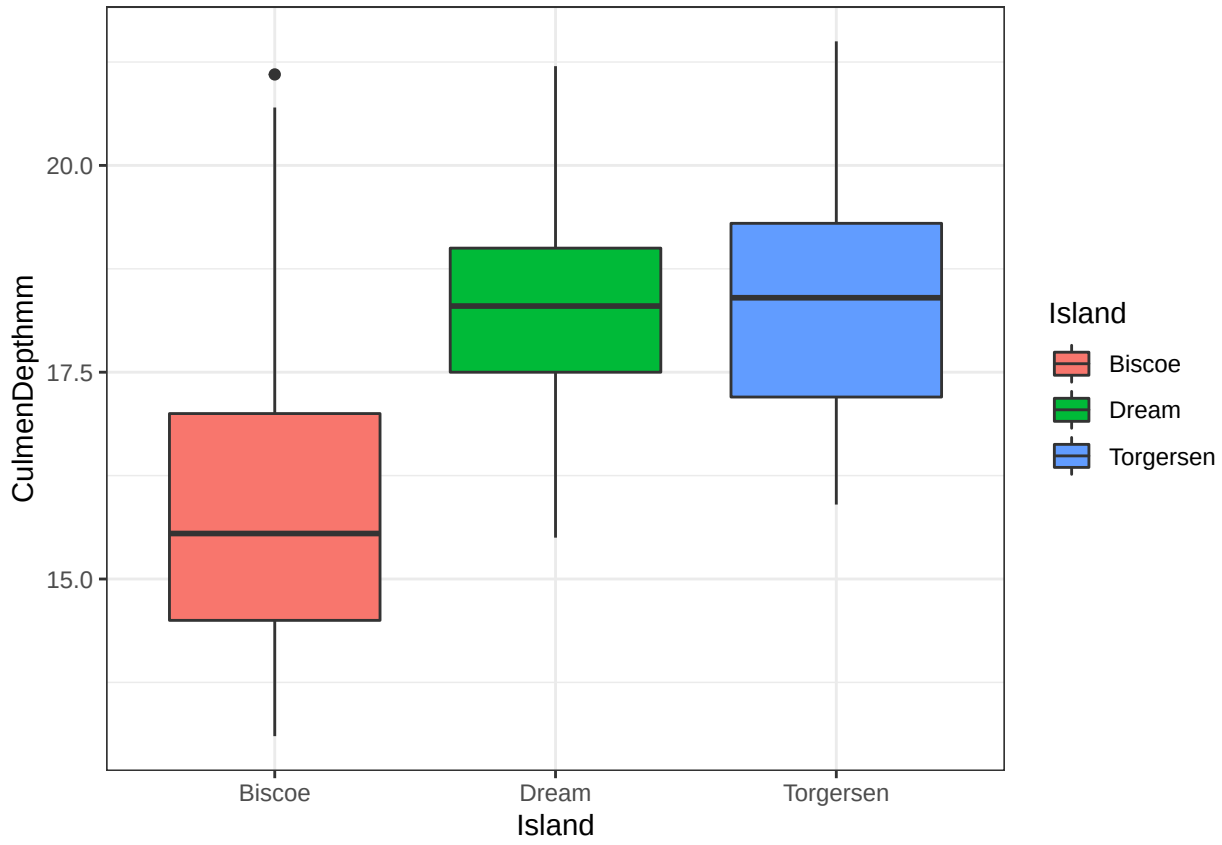

```
ggplot(penguinsdata)+
  geom_point(aes(x = CulmenDepthmm, y = BodyMassg), stroke = 2)+
  facet_wrap(~ Species)+
  labs(x = "CulmenDepthmm", y = "BodyMassg")+
  theme_bw()
```



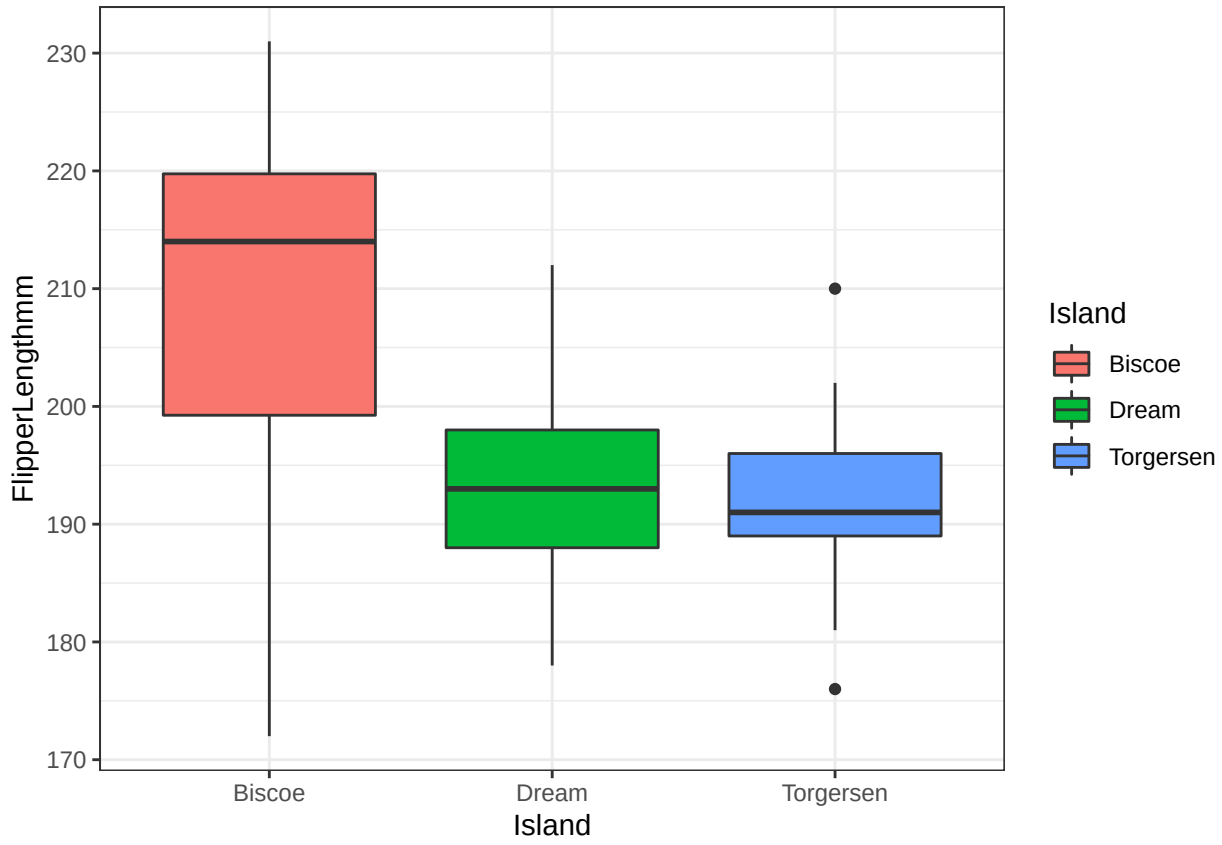

```
ggplot(penguinsdata)+
  geom_boxplot(aes(x = Island, y = CulmenLengthmm, fill = Island))+
  theme_bw()
```



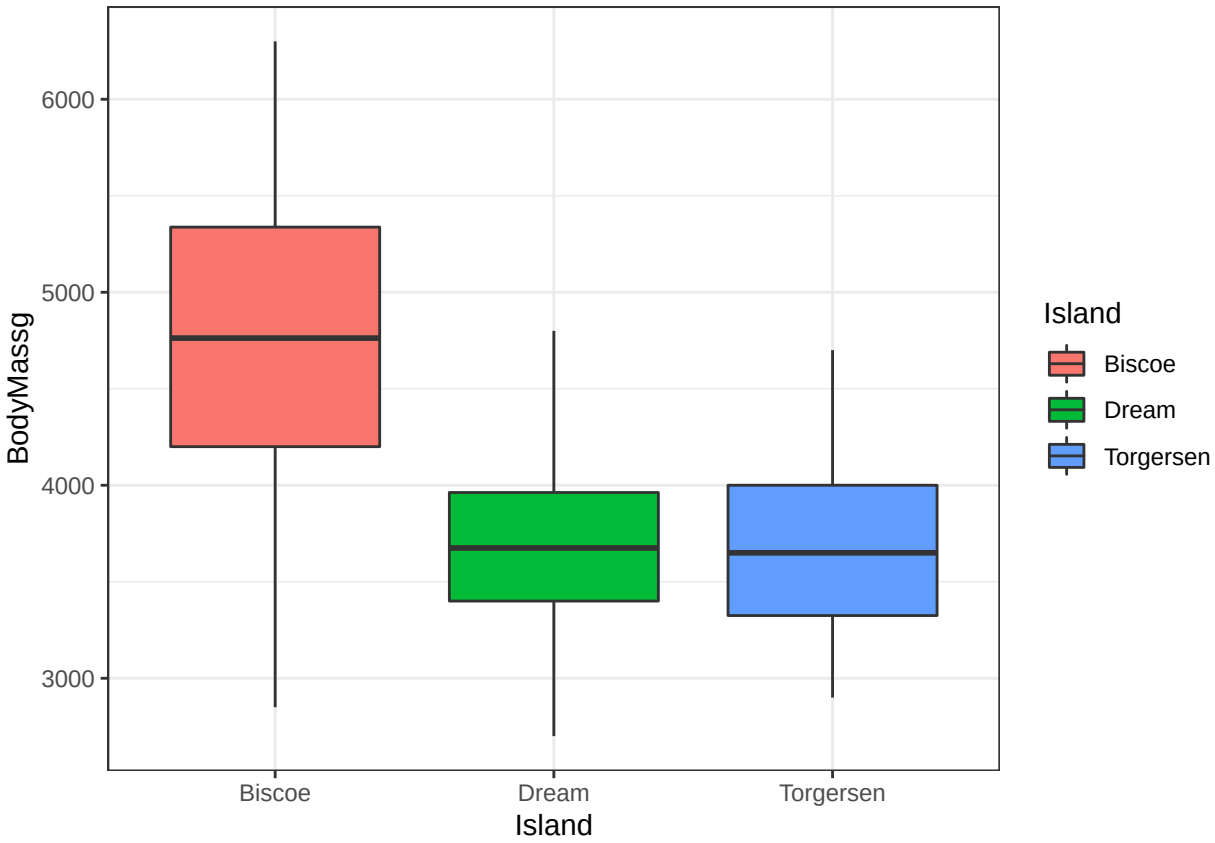

```
ggplot(penguinsdata)+
  geom_boxplot(aes(x = Island, y = CulmenDepthmm, fill = Island))+
  theme_bw()
```




```
ggplot(penguinsdata)+
  geom_boxplot(aes(x = Island, y = FlipperLengthmm, fill = Island))+
  theme_bw()
```




```
ggplot(penguinsdata)+
  geom_boxplot(aes(x = Island, y = BodyMassg, fill = Island))+
  theme_bw()
```




```
#-------------------------------------------------------
#K-means Clustering
penguins.new<- penguinsdata[,c(10,11,12,13)]
penguins.class<- penguinsdata[,"Region"]
penguins.class = as.factor(penguins.class)
head(penguins.new)
```

```
##   CulmenLengthmm CulmenDepthmm FlipperLengthmm BodyMassg
## 2           39.5          17.4             186      3800
## 3           40.3          18.0             195      3250
## 5           36.7          19.3             193      3450
## 6           39.3          20.6             190      3650
## 7           38.9          17.8             181      3625
## 8           39.2          19.6             195      4675
```

```
head(penguins.class)
```

```
## [1] Anvers Anvers Anvers Anvers Anvers Anvers
## Levels: Anvers
```

```
#4. Create a function to normalize the data before clustering
# Normalization
normalize <- function(x){
  return ((x-min(x))/(max(x)-min(x)))
```

```
}
penguins.new$CulmenLengthmm<- normalize(penguins.new$CulmenLengthmm)
penguins.new$CulmenDepthmm<- normalize(penguins.new$CulmenDepthmm)
penguins.new$FlipperLengthmm<- normalize(penguins.new$FlipperLengthmm)
penguins.new$BodyMassg<- normalize(penguins.new$BodyMassg)
head(penguins.new)
```

```
##   CulmenLengthmm CulmenDepthmm FlipperLengthmm BodyMassg
## 2      0.2690909     0.5119048       0.2372881 0.3055556
## 3      0.2981818     0.5833333       0.3898305 0.1527778
## 5      0.1672727     0.7380952       0.3559322 0.2083333
## 6      0.2618182     0.8928571       0.3050847 0.2638889
## 7      0.2472727     0.5595238       0.1525424 0.2569444
## 8      0.2581818     0.7738095       0.3898305 0.5486111
```

```
#5. Apply k-means clustering algorithm with k = 3
result<- kmeans(penguins.new,3) #aplly k-means algorithm with no. of centroids(k)=3
#6. Find the number of records in each cluster
result$size # gives no. of records in each cluster
```

```
## [1]  46 208  76
```

```
#7. Display the cluster center data point values
result$centers # gives value of cluster center datapoint value(3 centers for k=3)
```

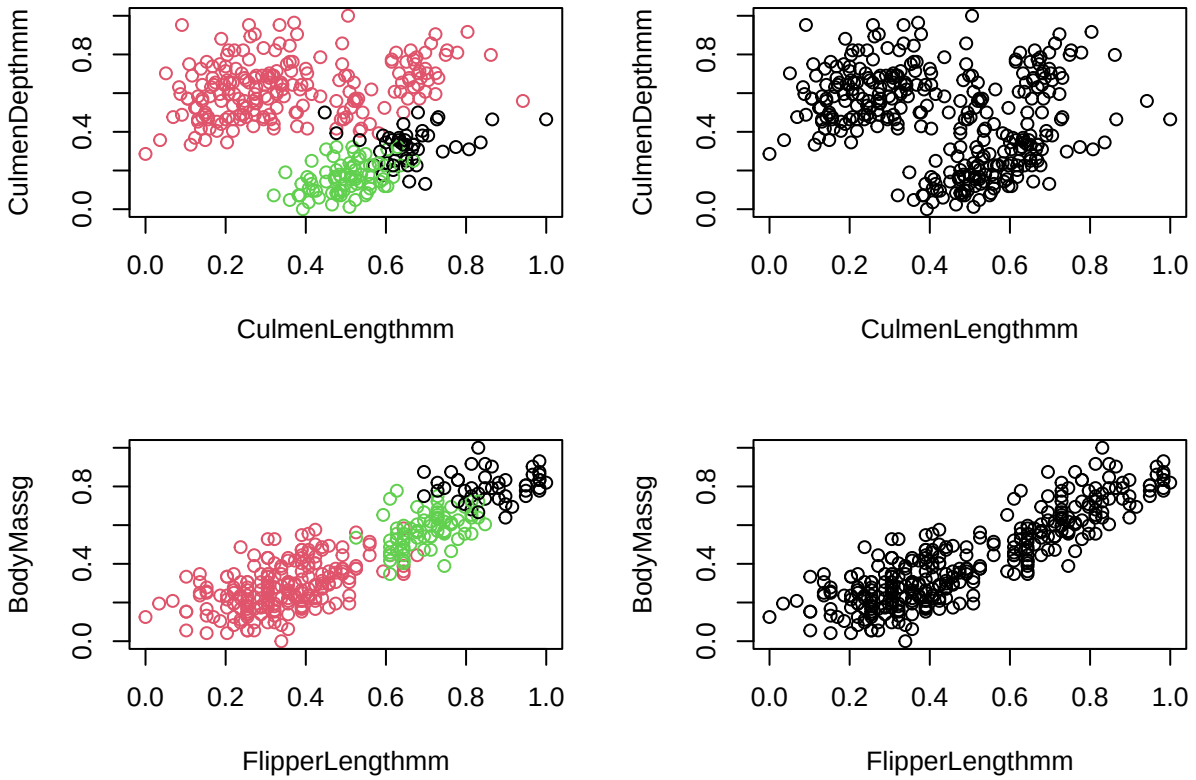
```
##   CulmenLengthmm CulmenDepthmm FlipperLengthmm BodyMassg
## 1      0.6616601     0.3291925       0.8677229 0.8025362
## 2      0.3606643     0.6252862       0.3393905 0.2811498
## 3      0.4988517     0.1599311       0.7031668 0.5730994
```

```
#8. Display the cluster vector showing the cluster where each record falls
result$cluster #gives cluster vector showing the cluster where each record falls
```

```
##   2   3   5   6   7   8  10  11  15  17  18  19  20  21  22  23  24  25  26  27
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  28  29  30  31  32  33  34  35  36  37  38  39  41  43  44  45  46  49  50  51
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
```

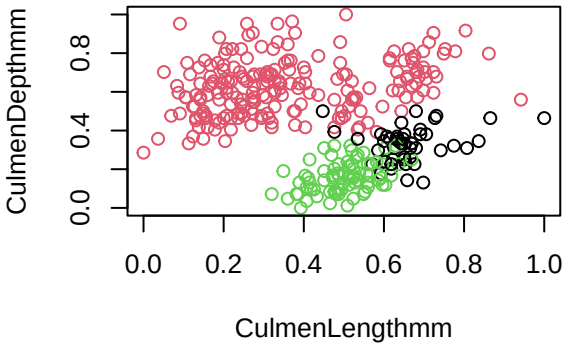
```
## 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 212 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232
##   2   2   2   2   2   2   2   2   3   1   3   1   3   3   3   3   3   3   3   1
## 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 252 253
##   3   1   3   1   3   1   3   1   1   3   3   3   3   3   3   1   3   1   3   3
## 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273
##   1   1   1   3   1   3   3   3   1   3   3   1   3   3   1   3   3   3   3   3
## 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293
##   1   3   3   3   3   3   1   3   3   3   1   3   1   3   1   3   1   3   3   1
## 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313
##   3   3   1   3   1   3   1   3   1   3   1   3   1   3   1   3   1   3   1   3
## 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333
##   1   3   1   3   3   3   1   3   1   3   1   3   1   3   1   3   3   3   1   3
## 334 335 336 337 338 339 341 342 343 344
##   1   3   1   3   1   3   3   1   3   1
```

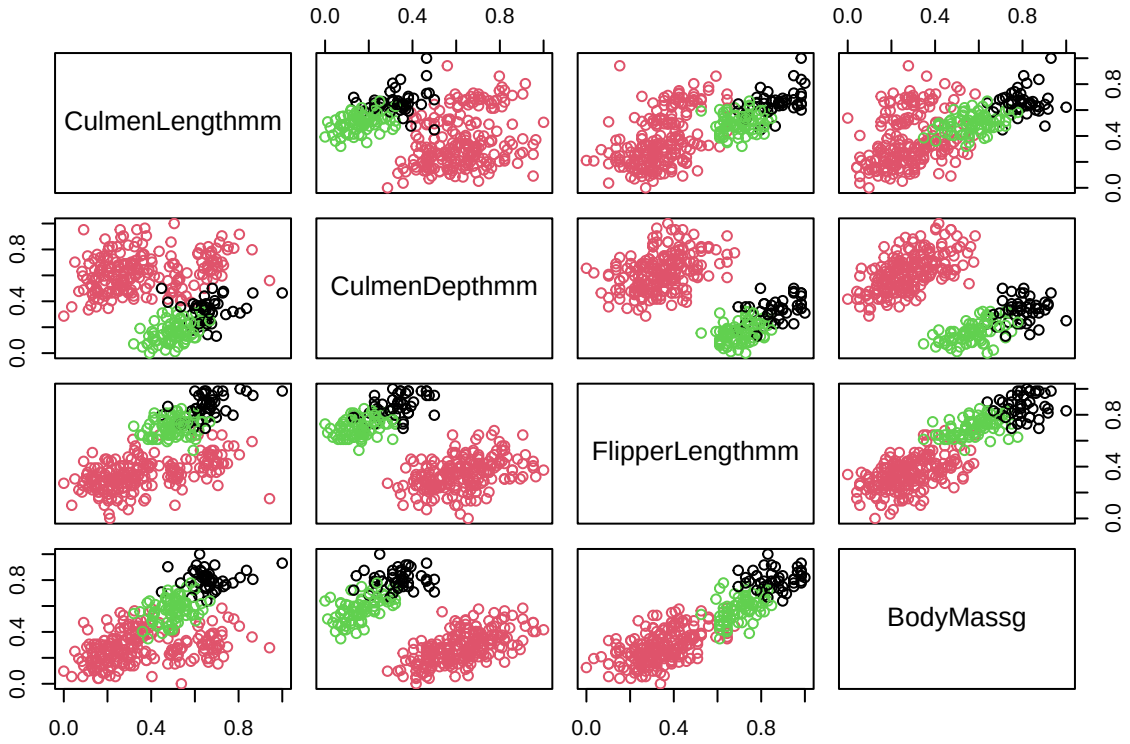
```r
# Verify results of clustering
par(mfrow=c(2,2), mar=c(5,4,2,2))
plot(penguins.new[c(1,2)], col=result$cluster)
plot(penguins.new[c(1,2)], col=penguins.class)
plot(penguins.new[c(3,4)], col=result$cluster)
plot(penguins.new[c(3,4)], col=penguins.class)
```



```
result$cluster <- as.factor(result$cluster)
#13.  Install the package ggplot2 and import it.
library(ggplot2)
#14. Plot the clusterresults using ggplot
ggplot(penguins.new, aes(FlipperLengthmm, BodyMassg, color = result$cluster)) + geom_point()
```


```
plot(penguins.new[c("CulmenLengthmm", "CulmenDepthmm")], col=result$cluster)
#15. Display the clustering results with all parameters
plot(penguins.new[,], col=result$cluster)
```

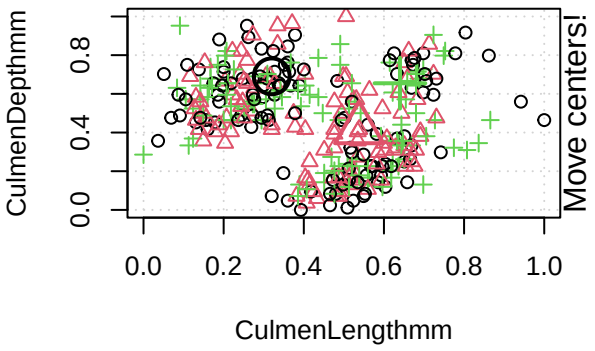

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm

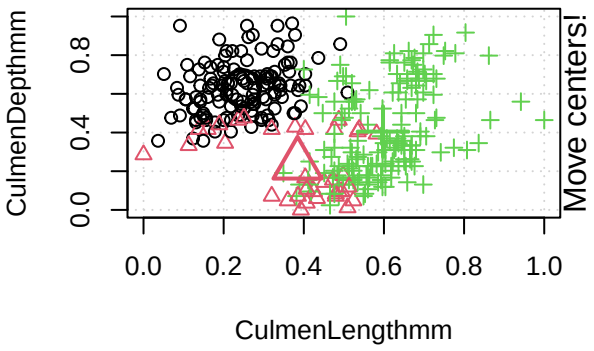


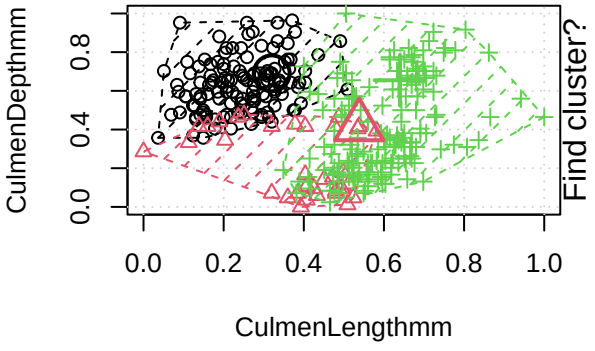

```
#16. Display the results in table
table(result$cluster,penguins.class) # Result of table shows that Cluster 1 corresponds to Virginica, C
```

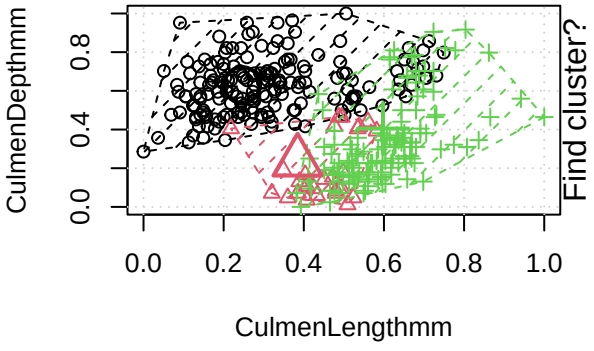
```
##    penguins.class
##     Anvers
##   1     46
##   2    208
##   3     76
```

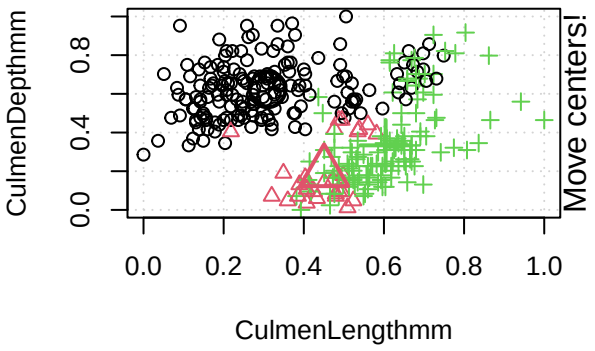
```
#Total number of correctly classified instances are: 36 + 47 + 50= 133
#Total number of incorrectly classified instances are: 3 + 14= 17
#Accuracy = 133/(133+17) = 0.88 i.e our model has achieved 88% accuracy!
#In order to improve this accuracy further, we may try different values of "k".
#=========================================================================
# K means algorithms with Animation
#=========================================================================
#17. Display the K Means Algorithm with Animation and visualize the changes in the cluster center
library(animation)
km1<-kmeans.ani(penguins.new,3)
```

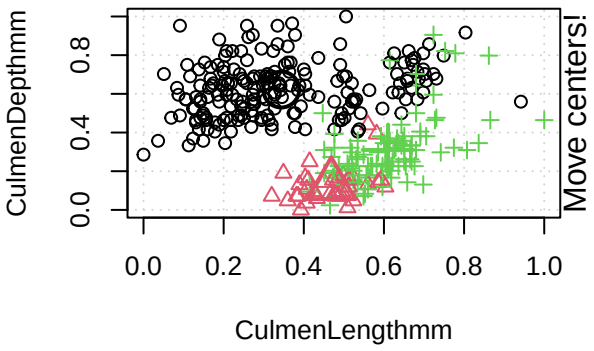

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Move centers!

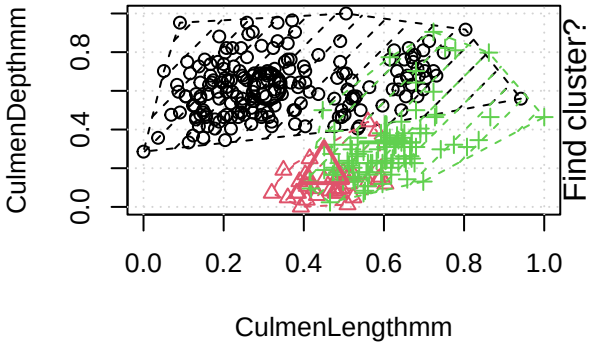

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Move centers!

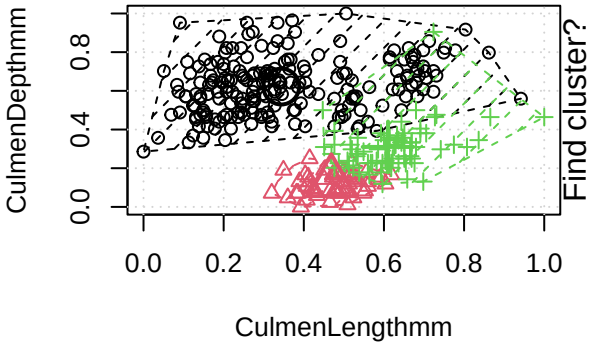

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Find cluster?

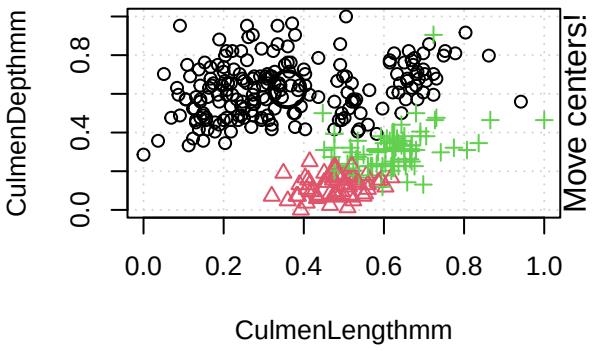

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Find cluster?

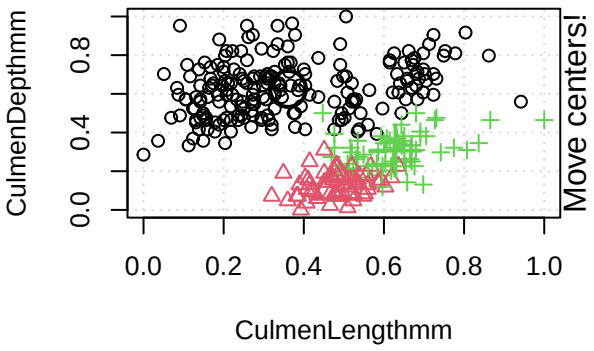

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Move centers!

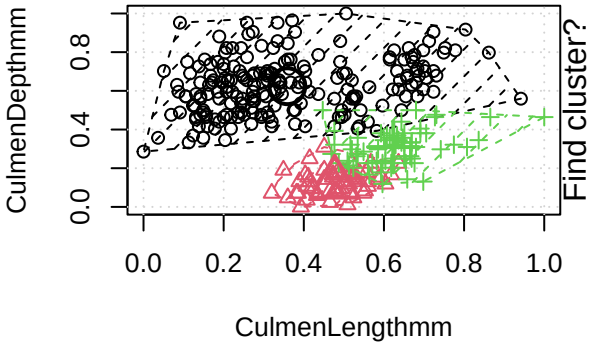

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Move centers!

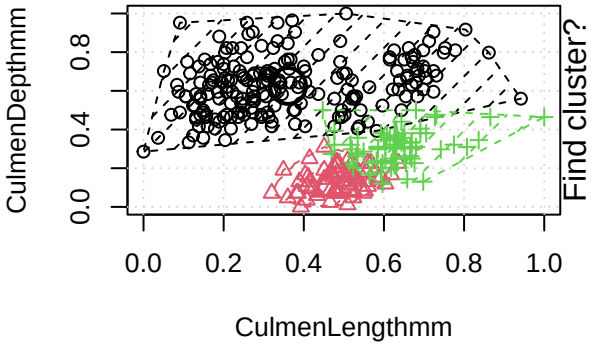

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Find cluster?

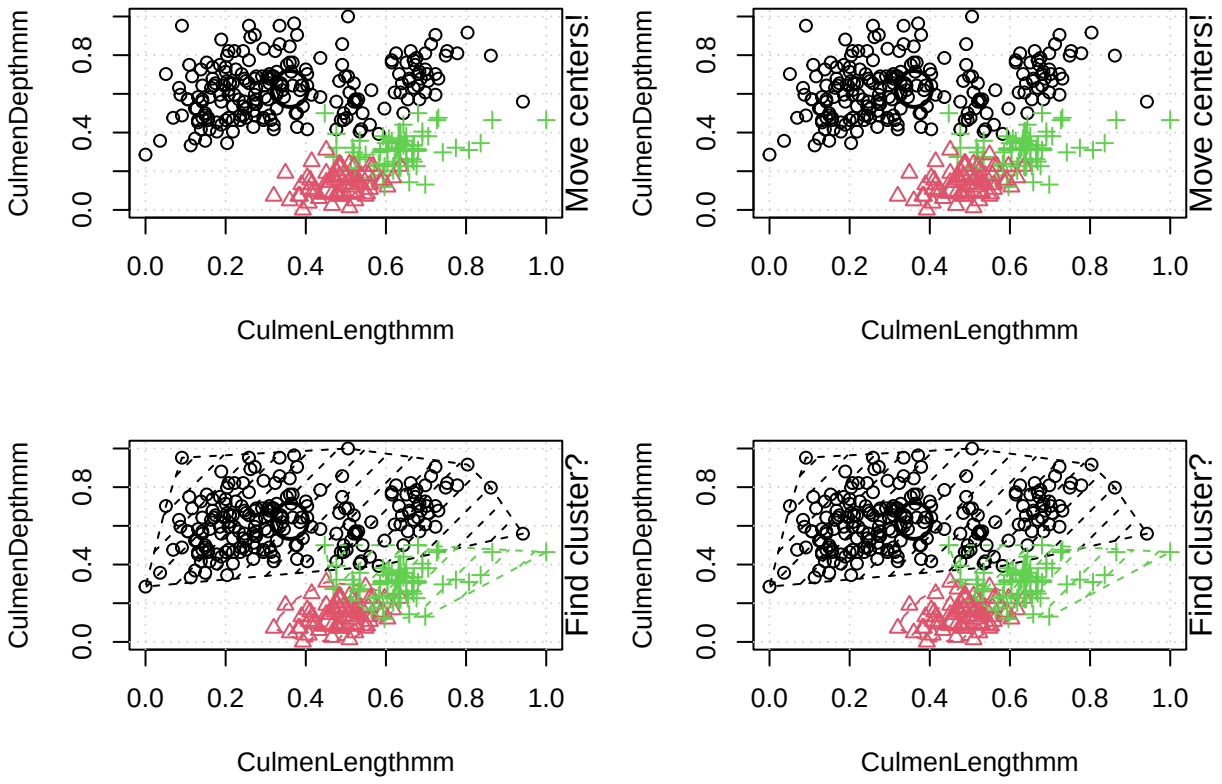

CulmenDepthmm
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm
Find cluster?


CulmenDepthmm
0.0
0.4
0.8
Move centers!
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm


CulmenDepthmm
0.0
0.4
0.8
Move centers!
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm


CulmenDepthmm
0.0
0.4
0.8
Find cluster?
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm


CulmenDepthmm
0.0
0.4
0.8
Find cluster?
0.0
0.2
0.4
0.6
0.8
1.0
CulmenLengthmm

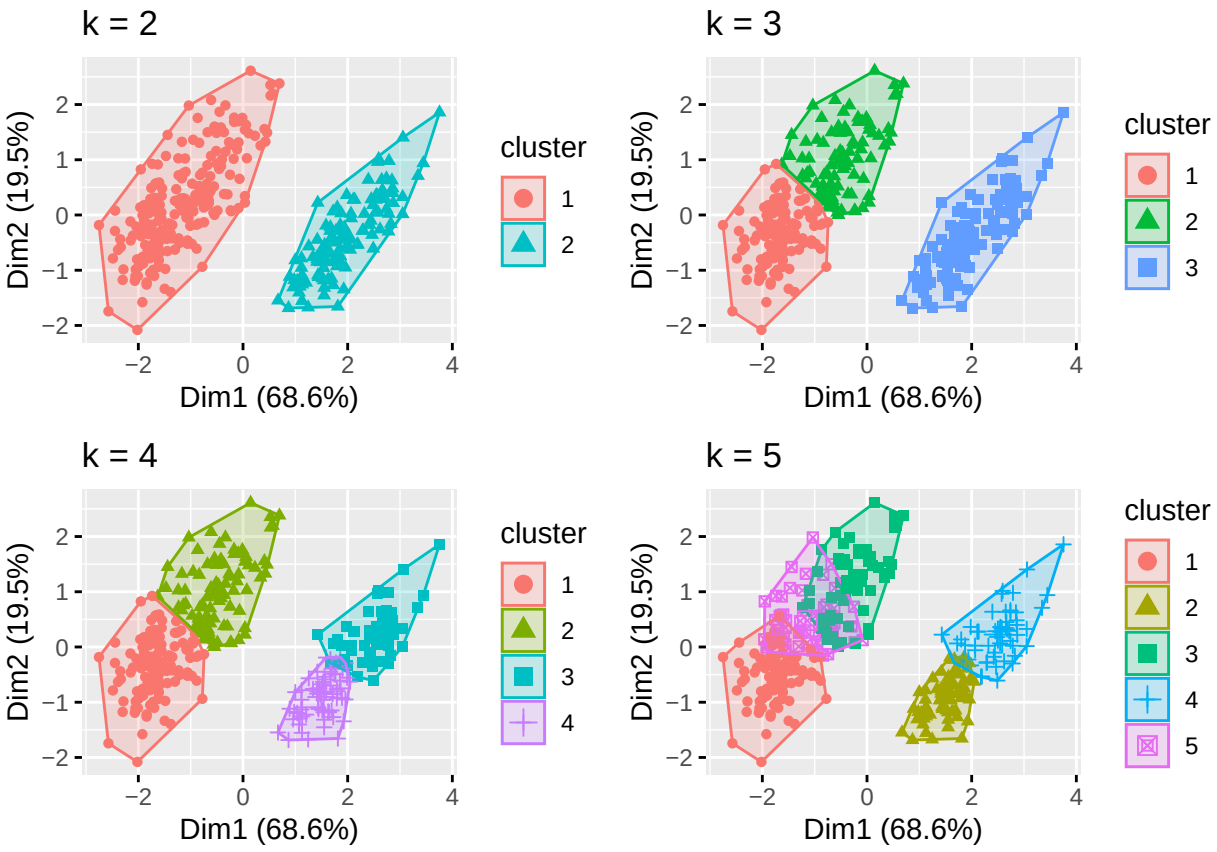



```
#18. Import factoextra package and visualize the cluster result
library(factoextra) # clustering algorithms & visualization
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_cluster(result, data = penguins.new)
#19. Explore the cluster analysis result with various value of k like 3,4,5
k2 <- kmeans(penguins.new, centers = 2, nstart = 25)
k3 <- kmeans(penguins.new, centers = 3, nstart = 25)
k4 <- kmeans(penguins.new, centers = 4, nstart = 25)
k5 <- kmeans(penguins.new, centers = 5, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data = penguins.new) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = penguins.new) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = penguins.new) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = penguins.new) + ggtitle("k = 5")
grid.arrange(p1, p2, p3, p4, nrow = 2)
```


k = 2
k = 3
k = 4
k = 5
Dim1 (68.6%)
Dim2 (19.5%)
cluster
1
2
3
4
5