Final Assignment

Shashidhar Reddy Boreddy

2022-12-16

R Markdown

Data is from biologists collecting data on penguins:

Prepare the data of penguins

#importing the required packages

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ggplot2)
library(lattice)
library(class)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gmodels)
library(knitr)
library(rmarkdown)
library(tidyverse)
```

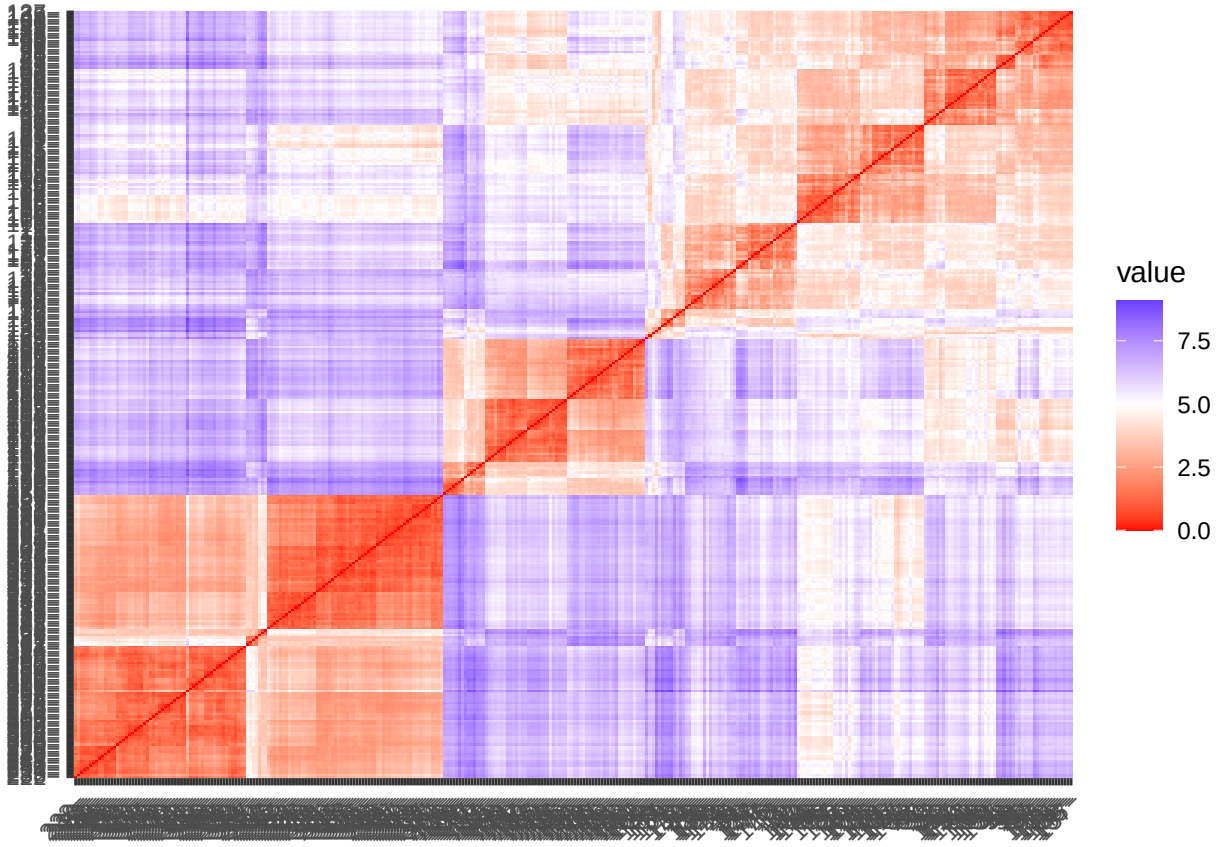
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8     v purrr   0.3.5
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(dplyr)
library(factoextra)
```

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

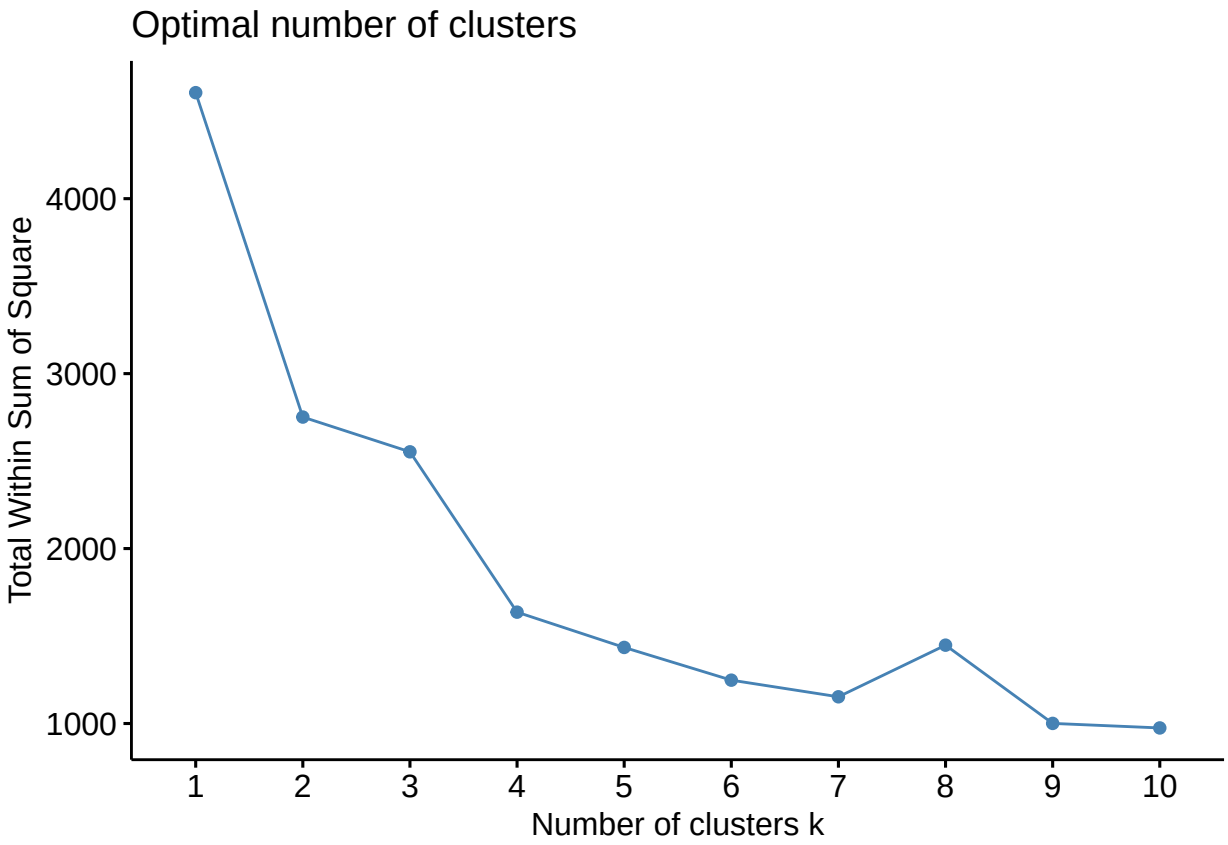
```
#importing a dataset
SB_data <- read.csv("C:/Users/shash/Dropbox/PC/Downloads/penguins_lter.csv")
SB_data <- na.omit(SB_data)
Island <- dummyVars(~Island,SB_data)
IslDV <- predict(Island, SB_data)
#using appropriate predict function
Species <- dummyVars(~Species,SB_data)
SpecDV <- predict(Species, SB_data)

#creating sub sets
SB_data <- subset(SB_data, select = -c(Island))
SB_data$Clutch.Completion <- ifelse(SB_data$Clutch.Completion == "Yes",1,0)
SB_data$Sex <- ifelse(SB_data$Sex == "MALE",1,0)

dvSB_data <- cbind(SB_data,IslDV,SpecDV)


clust_constraint <- dvSB_data %>% select_if(is.numeric)
clust_constraint$Sample.Number = NULL



set.seed(2)
#using the dist function
clust_constraint <- scale(clust_constraint)
distance <- get_dist(clust_constraint)
fviz_dist(distance)
```



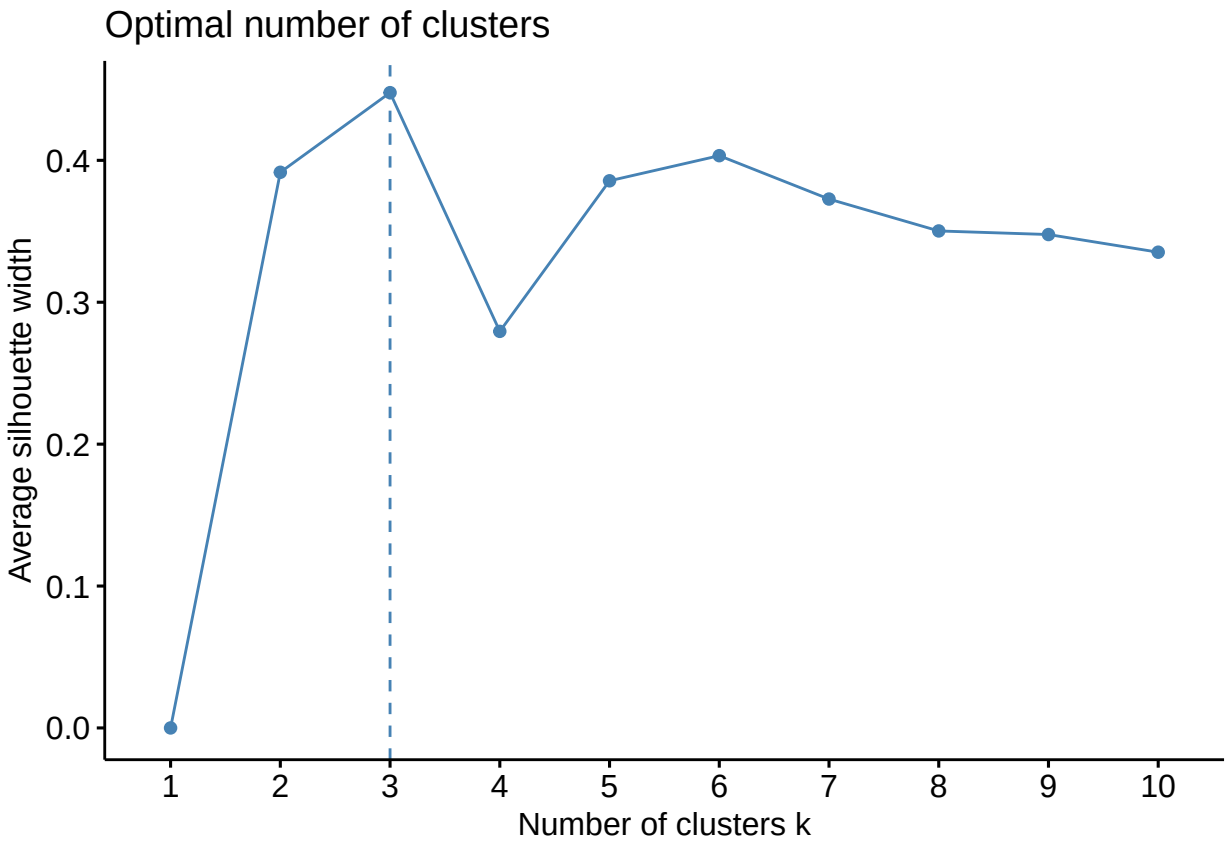
Finding ideal K

Given that we are seperationg by, islands, k=3 is obvious, though the code below confirms this.

```
clust_constraint <- scale(clust_constraint)
fviz_nbclust(clust_constraint,kmeans, method = "wss")
```




```
fviz_nbclust(clust_constraint,kmeans, method = "silhouette")
```



Visualizing the k means clustering

#The distinct groups are visually shown here, along with how they differ statistically.

```
k2 <- kmeans(clust_constraint, centers = 3, nstart = 25)

k2$centers
```

```
##   Clutch.Completion Culmen.Length..mm. Culmen.Depth..mm. Flipper.Length..mm.
## 1        0.14994823          0.6350952        -1.0744525            1.134141
## 2        0.03557682         -0.9630941         0.6177020           -0.789354
## 3       -0.34791067          0.8703681         0.6565258           -0.403974
##   Body.Mass..g.          Sex Delta.15.N..o.oo. Delta.13.C..o.oo. IslandBiscoe
## 1     1.0665813  0.007839384        -0.8845046        -0.6257757    0.9924505
## 2    -0.6360437 -0.011204851         0.2289921        -0.1422800   -0.3813750
## 3    -0.6035934  0.009305659         1.1286817         1.4388972   -1.0045536
##   IslandDream IslandTorgersen SpeciesAdelie Penguin (Pygoscelis adeliae)
## 1 -0.74984799      -0.3967572                                  -0.8624216
## 2  0.01702676       0.5318235                                   1.1560120
## 3  1.32956240      -0.3967572                                  -0.8624216
##   SpeciesChinstrap penguin (Pygoscelis antarctica)
## 1                                       -0.5039652
## 2                                       -0.5039652
## 3                                        1.9782513
##   SpeciesGentoo penguin (Pygoscelis papua)
```

```
## 1                                    1.3037452
## 2                                   -0.7646967
## 3                                   -0.7646967
```

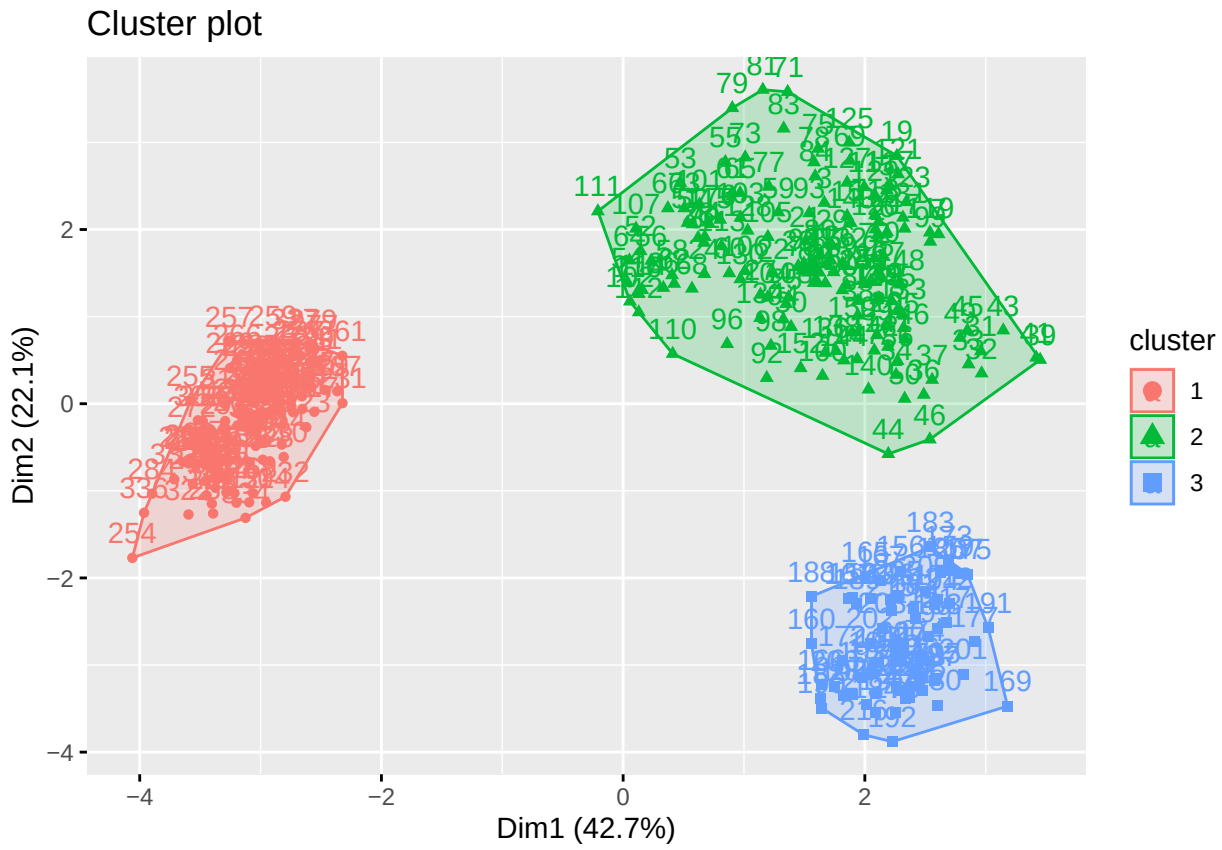
```
k2$size
```

```
## [1] 122 141  67
```

```
k2$cluster[11]
```

```
## 18
##  2
```

```
fviz_cluster(k2, data = clust_constraint)
```



Identifiying the detailed clusters

Prepare again

#Clustering is used to restore the clusters shown below.

```
library(gmodels)
library(knitr)
library(rmarkdown)
library(readr)
library(tidyverse)
library(caret)
library(cluster)
library(factoextra)
library(RColorBrewer)
library(dplyr)
library(ggraph)
library(igraph)
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tidyr':
##
##     crossing

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following object is masked from 'package:class':
##
##     knn

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
SB_data <- read.csv("C:/Users/shash/Dropbox/PC/Downloads/penguins_lter.csv")
SB_data <- na.omit(SB_data)

Island <- dummyVars(~Island,SB_data)
IslDV <- predict(Island, SB_data)

Species <- dummyVars(~Species,SB_data)
```

```
SpecDV <- predict(Species, SB_data)



SB_data$Clutch.Completion <- ifelse(SB_data$Clutch.Completion == "Yes",1,0)
SB_data$Sex <- ifelse(SB_data$Sex == "MALE",1,0)

dvSB_data <- cbind(SB_data,IslDV,SpecDV)




numeric_Penguins <- dvSB_data %>% select_if(is.numeric)
numeric_Penguins$Sample.Number = NULL


SB_data_norm <- as.data.frame(scale(numeric_Penguins))

d <- dist(SB_data_norm, method = "euclidean")
```

Select Method

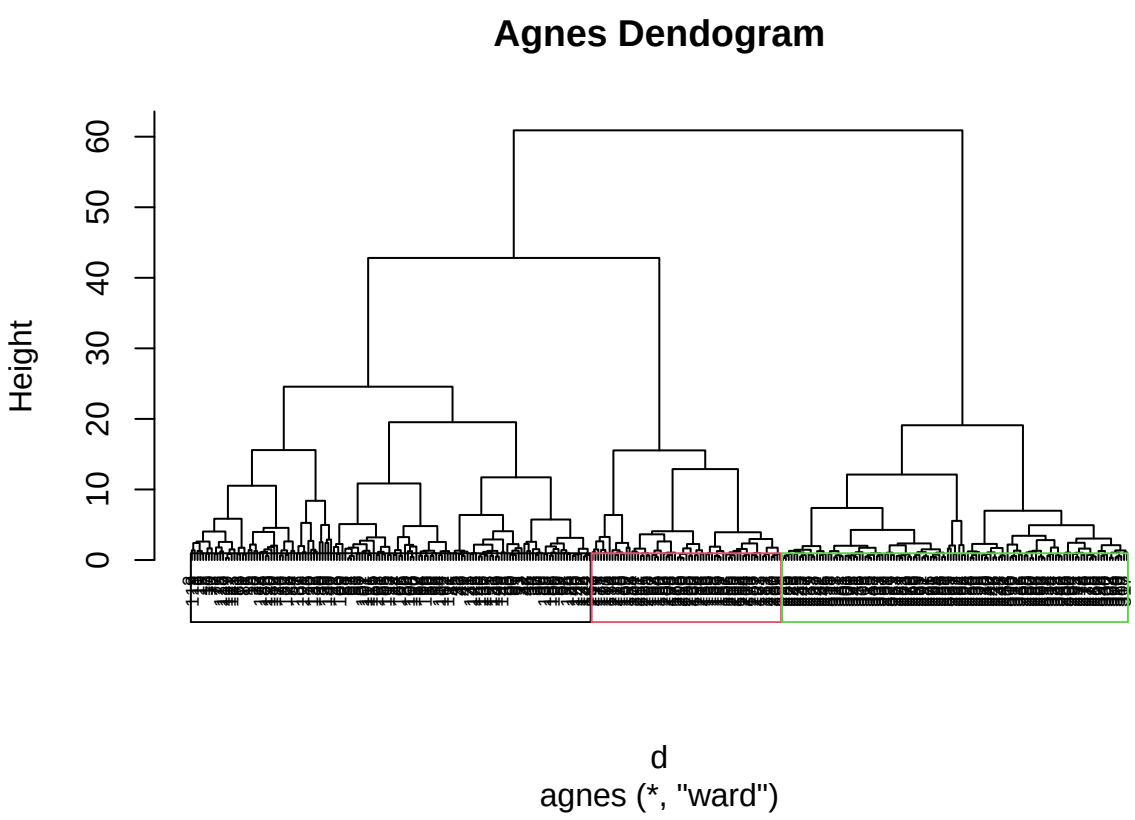
#Now we'll see which clustering approach performs the best.

```
##   average    single  complete      ward
## 0.8639980 0.7866488 0.9050510 0.9858434
```

Agnes of visualization

##Now we'll create an Agnes Dendogram to show how the clusters divide.

```
hc <- agnes(d, method = "ward")
pltree(hc, cex = 0.6, hang = -1, main = "Agnes Dendogram")
rect.hclust(hc, k = 3, border = 1:5)
```



Create cluster partitions

#In the two parts that follow, we will first create the cluster divisions and then generate the centroids for each group.

```
cluster_part <- cutree(hc, k = 3)
Penguins_clustered <- mutate(SB_data_norm, cluster = cluster_part)
set.seed(23)

part_index <- createDataPartition(Penguins_clustered$cluster, p = 0.7, list = FALSE)
Part_A <- Penguins_clustered[part_index,]
Part_B <- Penguins_clustered[-part_index,]
```

```
Part_A_centroid <- Part_A %>% gather("features", "values", -cluster) %>% group_by(cluster,features) %>%
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

```
cluster_B <- data.frame(data = seq(1,nrow(Part_B), 1), Cluster_B_Part = rep(0,nrow(Part_B)))

for (x in 1:nrow(Part_B)) {
  cluster_B$Cluster_B_Part[x] <- which.min(as.matrix(get_dist(as.data.frame(rbind(Part_A_centroid[-1], I
}
```

```
cluster_B <- cluster_B %>% mutate(original_clusters = Part_B$cluster)
mean(cluster_B$Cluster_B_Part) == cluster_B$original_clusters
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [97] FALSE FALSE
```

```
split_clusters <- split(Penguins_clustered, Penguins_clustered$cluster)
mean_split <- lapply(split_clusters,colMeans)
mean_split
```

```
## $`1`
##                                    Clutch.Completion
##                                           0.03557682
##                                  Culmen.Length..mm.
##                                          -0.96309414
##                                   Culmen.Depth..mm.
##                                           0.61770198
##                                 Flipper.Length..mm.
##                                          -0.78935398
##                                       Body.Mass..g.
##                                          -0.63604369
##                                                Sex
##                                          -0.01120485
##                                    Delta.15.N..o.oo.
##                                           0.22899214
##                                    Delta.13.C..o.oo.
##                                          -0.14228000
##                                       IslandBiscoe
##                                          -0.38137499
##                                        IslandDream
##                                           0.01702676
##                                    IslandTorgersen
##                                           0.53182347
##       SpeciesAdelie Penguin (Pygoscelis adeliae)
##                                           1.15601195
## SpeciesChinstrap penguin (Pygoscelis antarctica)
##                                          -0.50396515
##         SpeciesGentoo penguin (Pygoscelis papua)
##                                          -0.76469672
##                                            cluster
##                                           1.00000000
##
## $`2`
##                                    Clutch.Completion
##                                         -0.347910674
```

```
##                                 Culmen.Length..mm.
##                                        0.870368083
##                                  Culmen.Depth..mm.
##                                        0.656525791
##                                Flipper.Length..mm.
##                                       -0.403973966
##                                     Body.Mass..g.
##                                       -0.603593371
##                                                Sex
##                                        0.009305659
##                                  Delta.15.N..o.oo.
##                                        1.128681695
##                                  Delta.13.C..o.oo.
##                                        1.438897192
##                                     IslandBiscoe
##                                       -1.004553565
##                                      IslandDream
##                                        1.329562404
##                                   IslandTorgersen
##                                       -0.396757190
##       SpeciesAdelie Penguin (Pygoscelis adeliae)
##                                       -0.862421616
## SpeciesChinstrap penguin (Pygoscelis antarctica)
##                                        1.978251264
##         SpeciesGentoo penguin (Pygoscelis papua)
##                                       -0.764696719
##                                            cluster
##                                        2.000000000
##
## $`3`
##                                 Clutch.Completion
##                                        0.149948229
##                                Culmen.Length..mm.
##                                        0.635095178
##                                 Culmen.Depth..mm.
##                                       -1.074452521
##                                Flipper.Length..mm.
##                                        1.134140710
##                                     Body.Mass..g.
##                                        1.066581280
##                                                Sex
##                                        0.007839384
##                                  Delta.15.N..o.oo.
##                                       -0.884504631
##                                  Delta.13.C..o.oo.
##                                       -0.625775672
##                                     IslandBiscoe
##                                        0.992450510
##                                      IslandDream
##                                       -0.749847991
##                                   IslandTorgersen
##                                       -0.396757190
##       SpeciesAdelie Penguin (Pygoscelis adeliae)
##                                       -0.862421616
```

```
## SpeciesChinstrap penguin (Pygoscelis antarctica)
##                                     -0.503965151
##         SpeciesGentoo penguin (Pygoscelis papua)
##                                      1.303745226
##                                          cluster
##                                      3.000000000
```

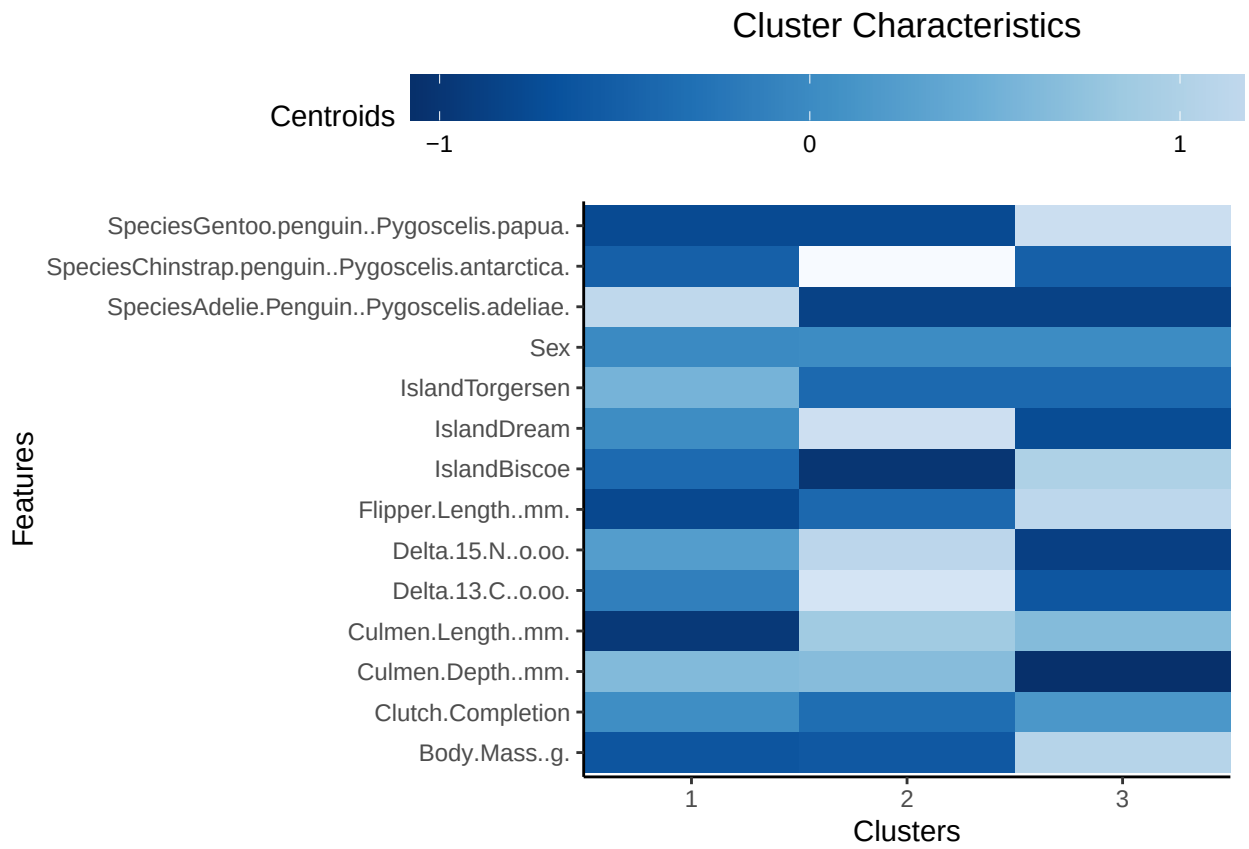
```
(centroids <- do.call(rbind, mean_split))
```

```
##   Clutch.Completion Culmen.Length..mm. Culmen.Depth..mm. Flipper.Length..mm.
## 1        0.03557682         -0.9630941         0.6177020           -0.789354
## 2       -0.34791067          0.8703681         0.6565258           -0.403974
## 3        0.14994823          0.6350952        -1.0744525            1.134141
##   Body.Mass..g.          Sex Delta.15.N..o.oo. Delta.13.C..o.oo. IslandBiscoe
## 1    -0.6360437 -0.011204851         0.2289921        -0.1422800   -0.3813750
## 2    -0.6035934  0.009305659         1.1286817         1.4388972   -1.0045536
## 3     1.0665813  0.007839384        -0.8845046        -0.6257757    0.9924505
##   IslandDream IslandTorgersen SpeciesAdelie Penguin (Pygoscelis adeliae)
## 1  0.01702676       0.5318235                                  1.1560120
## 2  1.32956240      -0.3967572                                 -0.8624216
## 3 -0.74984799      -0.3967572                                 -0.8624216
##   SpeciesChinstrap penguin (Pygoscelis antarctica)
## 1                                       -0.5039652
## 2                                        1.9782513
## 3                                       -0.5039652
##   SpeciesGentoo penguin (Pygoscelis papua) cluster
## 1                               -0.7646967       1
## 2                               -0.7646967       2
## 3                                1.3037452       3
```

#details of cluster

#Finally, we are plotting the clusters in order to determine the specifics of each cluster.

```
hc.graph <-
  colorRampPalette(rev(brewer.pal(9, 'Blues')), space = 'Lab')
data.frame(centroids) %>% gather("features", "values",-cluster) %>%
  ggplot(aes(
    x = factor(cluster),
    y = features,
    fill = values
  )) +
  geom_tile() + theme_classic() +
  theme(
    legend.position = "top",
    plot.title = element_text(hjust = 0.5),
    legend.key.width = unit(3, "cm"),
  ) +
  scale_x_discrete(expand = c(0, 0)) +
  scale_fill_gradientn(colours = hc.graph(100)) +
  labs(title = "Cluster Characteristics",
       x = "Clusters",
       y = "Features",
       fill = "Centroids")
```



#Individuals are being screened. Below We evaluate k2 so we can Identify which individual Penguin goes in Which cluster.

```
k2
```

```
## K-means clustering with 3 clusters of sizes 122, 141, 67
##
## Cluster means:
##   Clutch.Completion Culmen.Length..mm. Culmen.Depth..mm. Flipper.Length..mm.
## 1        0.14994823          0.6350952        -1.0744525            1.134141
## 2        0.03557682         -0.9630941         0.6177020           -0.789354
## 3       -0.34791067          0.8703681         0.6565258           -0.403974
##   Body.Mass..g.          Sex Delta.15.N..o.oo. Delta.13.C..o.oo. IslandBiscoe
## 1     1.0665813  0.007839384        -0.8845046        -0.6257757    0.9924505
## 2    -0.6360437 -0.011204851         0.2289921        -0.1422800   -0.3813750
## 3    -0.6035934  0.009305659         1.1286817         1.4388972   -1.0045536
##   IslandDream IslandTorgersen SpeciesAdelie Penguin (Pygoscelis adeliae)
## 1 -0.74984799      -0.3967572                                  -0.8624216
## 2  0.01702676       0.5318235                                   1.1560120
## 3  1.32956240      -0.3967572                                  -0.8624216
##   SpeciesChinstrap penguin (Pygoscelis antarctica)
## 1                                       -0.5039652
## 2                                       -0.5039652
## 3                                        1.9782513
##   SpeciesGentoo penguin (Pygoscelis papua)
## 1                                1.3037452
```

```
## 2                                -0.7646967
## 3                                -0.7646967
##
## Clustering vector:
##   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   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 212 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232
##   3   3   3   3   3   3   3   3   1   1   1   1   1   1   1   1   1   1   1   1
## 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 252 253
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 334 335 336 337 338 339 341 342 343 344
##   1   1   1   1   1   1   1   1   1   1
##
## Within cluster sum of squares by cluster:
## [1]  420.1459 1115.3081  299.7299
##  (between_SS / total_SS =  60.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

#The higher the body mass, the larger the beack (culmen) and flipper; however, the culmen is shallower.

The 'echo = FALSE' argument was added to the code chunk to prevent the R code that created the plot from being printed.