Palmer Penguins Classification

Devin DeLeon-Dowd

```
library(Hmisc)
library(ISLR)
library(knitr)
library(KernSmooth, lib.loc = "C:/Program Files/R/R-4.2.1/library")
library(mgcv, lib.loc = "C:/Program Files/R/R-4.2.1/library")
library(randomForest)
library(SDAResources)
library(tinytex)
library(tidyr)
library(dplyr)
library(e1071)
library(ggplot2)
library(car)
library(lme4)
library(lmtest)
library(tree)
library(glmnet)
library(xtable)
library(ggridges)
library(gridExtra)
library(nnet)
```

```
#importing the data
library(palmerpenguins)
```

```
## Warning: package 'palmerpenguins' was built under R version 4.2.2
```

```
penguins <- palmerpenguins::penguins
penguins_raw <- palmerpenguins::penguins_raw
```

```
#ridding full data set of identifier variables for classification and clustering
str(penguins_raw)
```

```
## tibble [344 x 17] (S3: tbl_df/tbl/data.frame)
##  $ studyName          : chr [1:344] "PAL0708" "PAL0708" "PAL0708" "PAL0708" ...
##  $ Sample Number      : num [1:344] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Species            : chr [1:344] "Adelie Penguin (Pygoscelis adeliae)" "Adelie Penguin (Pygoscelis
##  $ Region             : chr [1:344] "Anvers" "Anvers" "Anvers" "Anvers" ...
##  $ Island             : chr [1:344] "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
##  $ Stage              : chr [1:344] "Adult, 1 Egg Stage" "Adult, 1 Egg Stage" "Adult, 1 Egg Stage" "A
##  $ Individual ID      : chr [1:344] "N1A1" "N1A2" "N2A1" "N2A2" ...
##  $ Clutch Completion  : chr [1:344] "Yes" "Yes" "Yes" "Yes" ...
##  $ Date Egg           : Date[1:344], format: "2007-11-11" "2007-11-11" ...
##  $ Culmen Length (mm) : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ Culmen Depth (mm)  : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ Flipper Length (mm): num [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
```

```
##  $ Body Mass (g)      : num [1:344] 3750 3800 3250 NA 3450 ...
##  $ Sex                : chr [1:344] "MALE" "FEMALE" "FEMALE" NA ...
##  $ Delta 15 N (o/oo)  : num [1:344] NA 8.95 8.37 NA 8.77 ...
##  $ Delta 13 C (o/oo)  : num [1:344] NA -24.7 -25.3 NA -25.3 ...
##  $ Comments           : chr [1:344] "Not enough blood for isotopes." NA NA "Adult not sampled." ...
##  - attr(*, "spec")=List of 3
##   ..$ cols   :List of 17
##   .. ..$ studyName          : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Sample Number      : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
##   .. ..$ Species            : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Region             : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Island             : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Stage              : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Individual ID      : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Clutch Completion  : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Date Egg           :List of 1
##   .. .. ..$ format: chr ""
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_date" "collector"
##   .. ..$ Culmen Length (mm) : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
##   .. ..$ Culmen Depth (mm)  : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
##   .. ..$ Flipper Length (mm): list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
##   .. ..$ Body Mass (g)      : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
##   .. ..$ Sex                : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   .. ..$ Delta 15 N (o/oo)  : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
##   .. ..$ Delta 13 C (o/oo)  : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_double" "collector"
##   .. ..$ Comments           : list()
##   .. .. ..- attr(*, "class")= chr [1:2] "collector_character" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr [1:2] "collector_guess" "collector"
##   ..$ skip   : num 1
##   ..- attr(*, "class")= chr "col_spec"
```

```
penguins <- penguins_raw %>% select('Culmen Length (mm)','Culmen Depth (mm)', 'Flipper Length (mm)', 'B
colSums(is.na(penguins))
```

```
##  Culmen Length (mm)   Culmen Depth (mm) Flipper Length (mm)       Body Mass (g)
##                   2                   2                   2                   2
```

```
#option 1 to complete data set, drop NAs
penguins_1 <- penguins %>% drop_na()
```

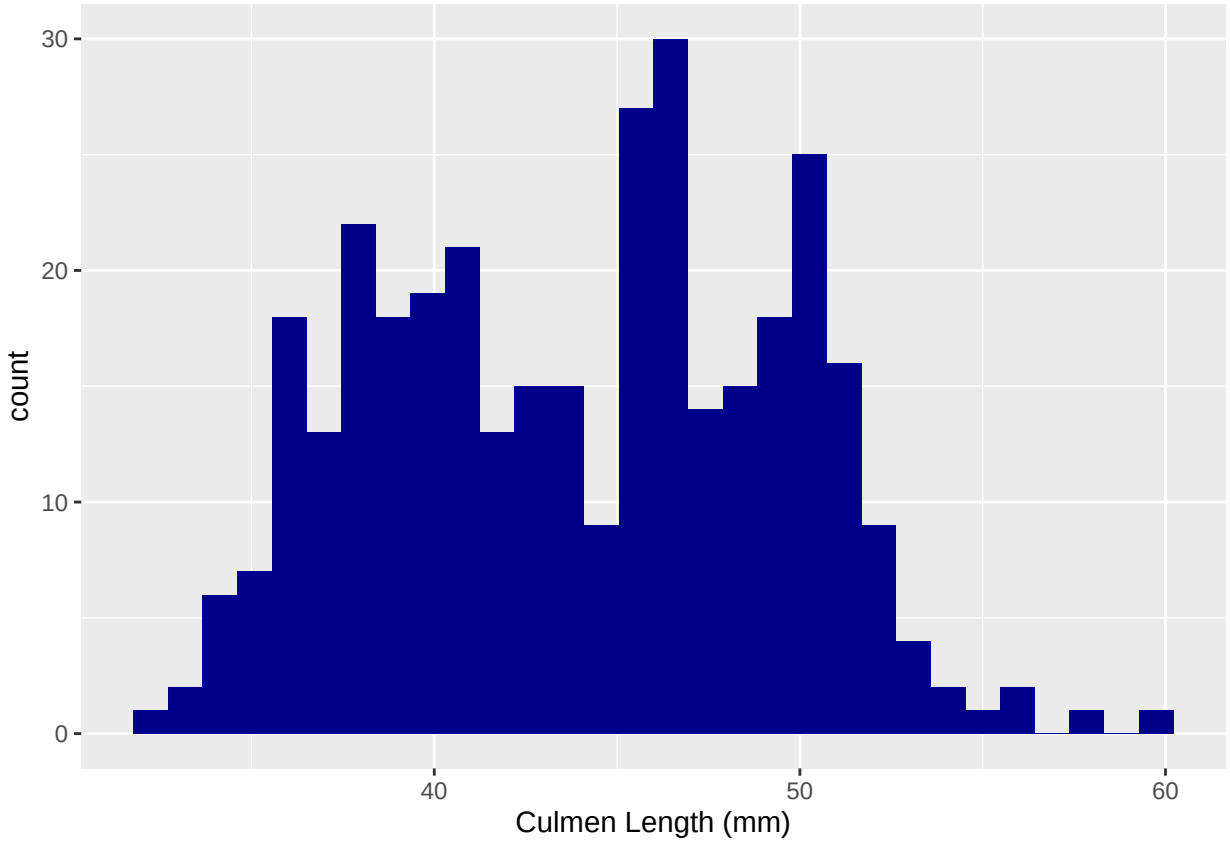
```
colSums(is.na(penguins_1))
```

```
##  Culmen Length (mm)   Culmen Depth (mm) Flipper Length (mm)       Body Mass (g)
##                   0                   0                   0                   0
```

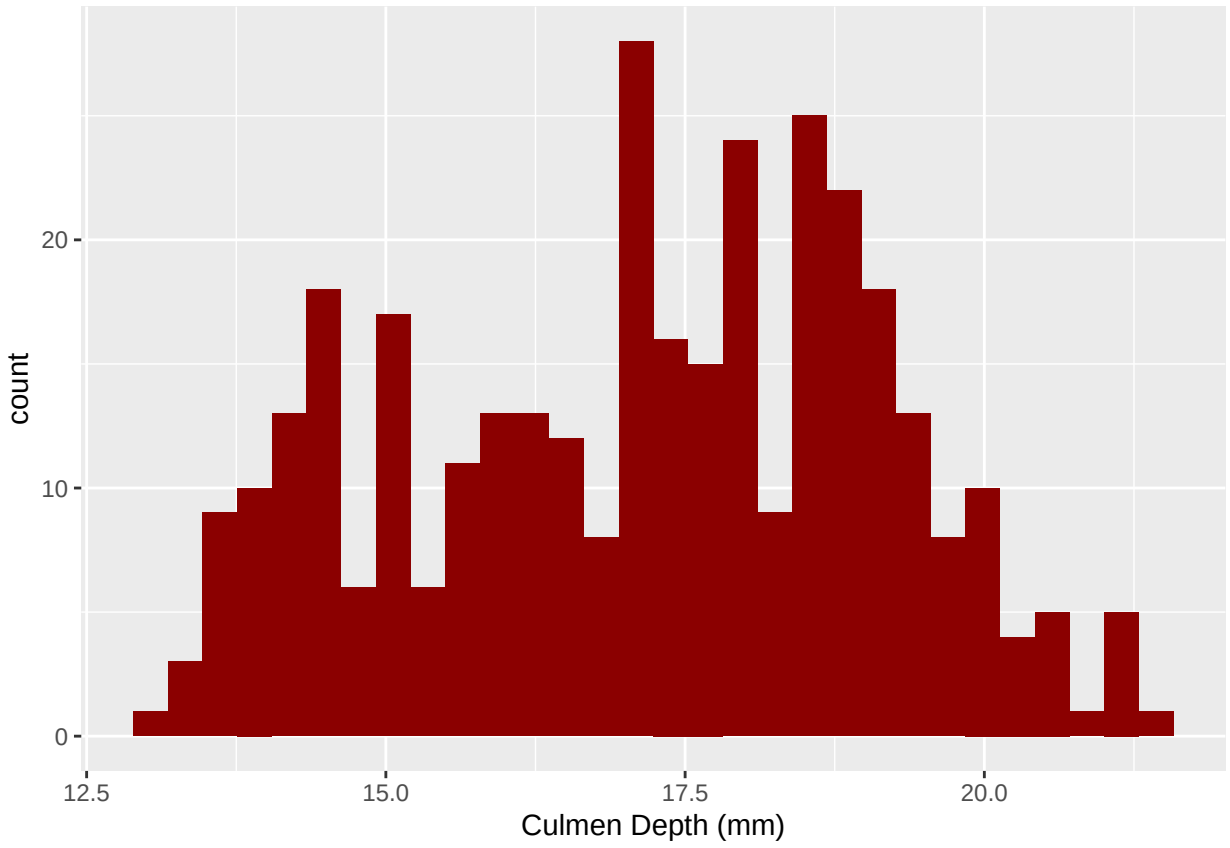
```
#option 2 impute NAs with mean values
penguins_2 <- penguins
penguins_2$`Culmen Length (mm)`[is.na(penguins_2$`Culmen Length (mm)`)] <- mean(penguins_2$`Culmen Leng
penguins_2$`Culmen Depth (mm)`[is.na(penguins_2$`Culmen Depth (mm)`)] <- mean(penguins_2$`Culmen Depth
penguins_2$`Flipper Length (mm)`[is.na(penguins_2$`Flipper Length (mm)`)] <- mean(penguins_2$`Flipper L
penguins_2$`Body Mass (g)`[is.na(penguins_2$`Body Mass (g)`)] <- mean(penguins_2$`Body Mass (g)`, na.rm
colSums(is.na(penguins_2))
```

```
##  Culmen Length (mm)   Culmen Depth (mm) Flipper Length (mm)       Body Mass (g)
##                   0                   0                   0                   0
```

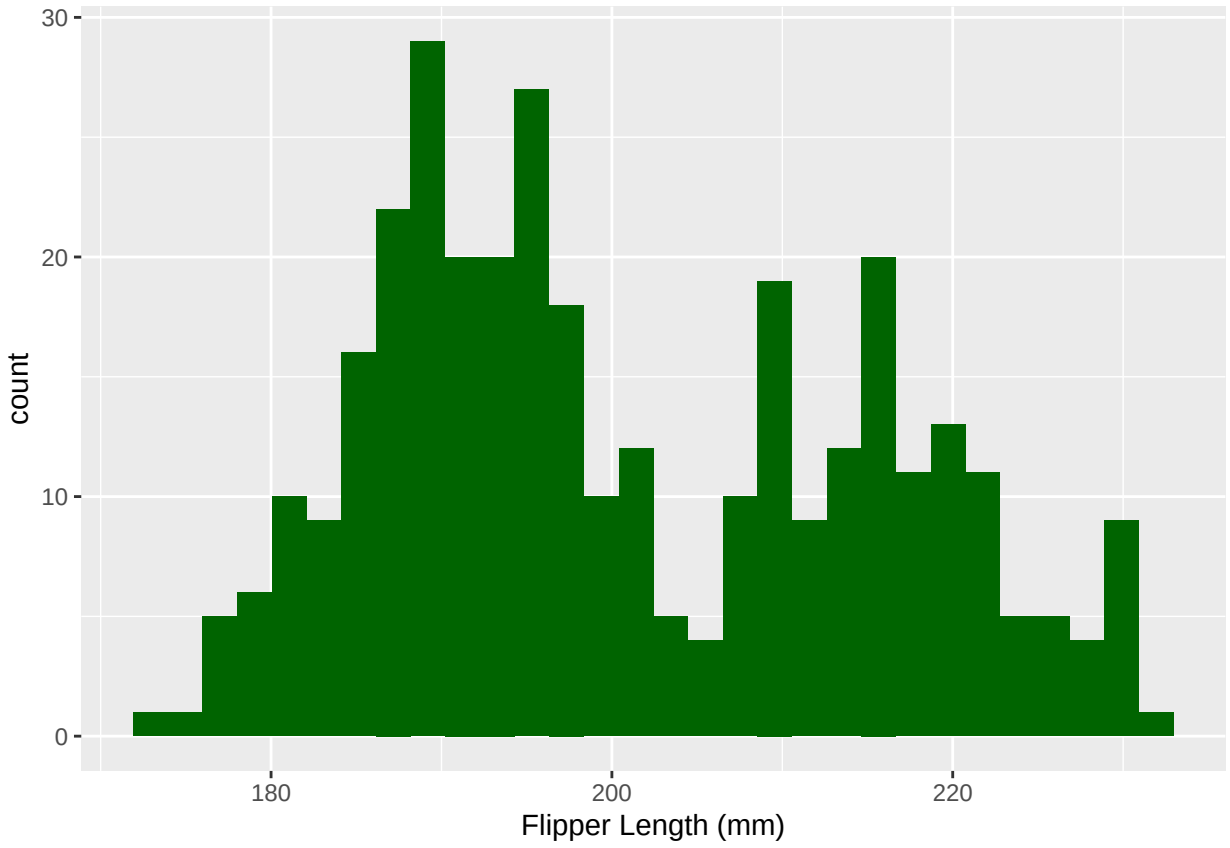
```
#check distributions of the measurements
ggplot(penguins_2, aes(x = `Culmen Length (mm)`)) + geom_histogram(bins = 30, fill = "dark blue")
```



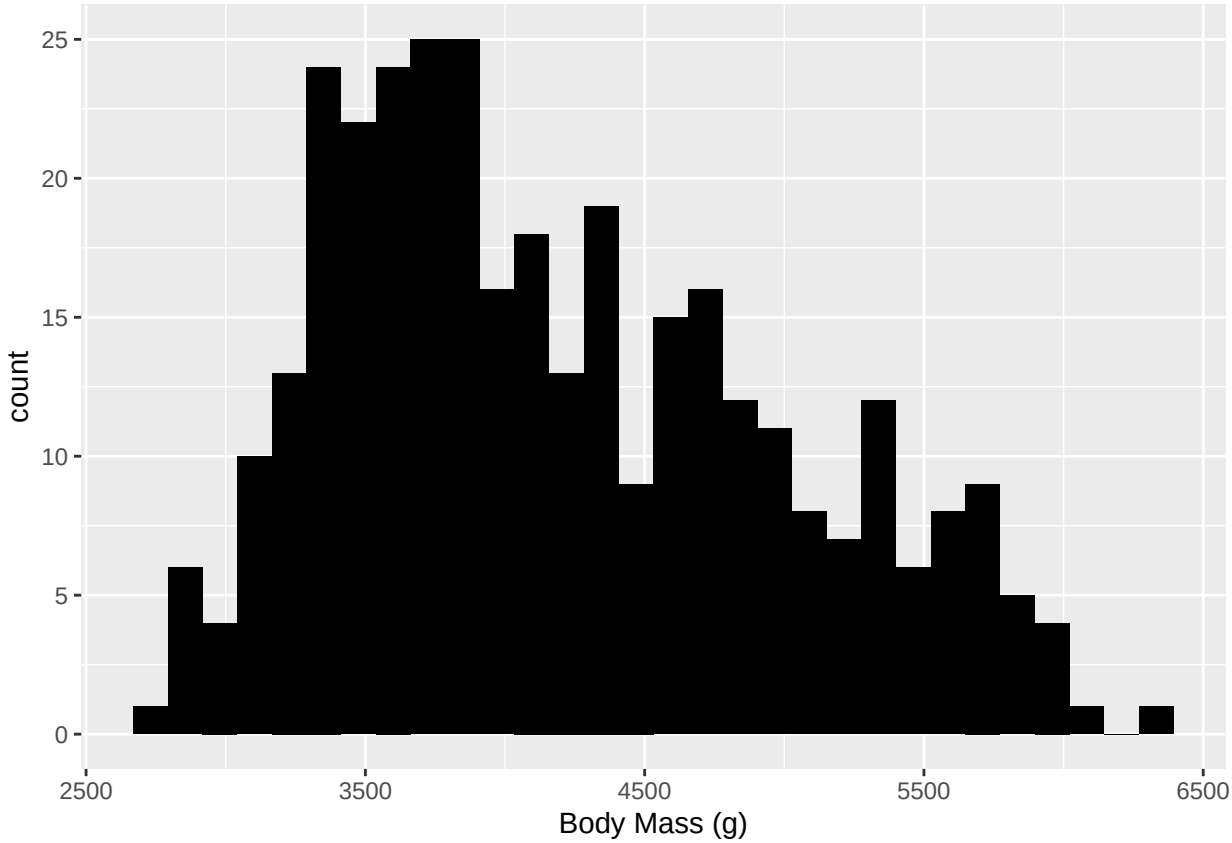

```
ggplot(penguins_2, aes(x = `Culmen Depth (mm)`)) + geom_histogram(bins = 30, fill = "dark red")
```




```
ggplot(penguins_2, aes(x = `Flipper Length (mm)`)) + geom_histogram(bins = 30, fill = "dark green")
```




```
ggplot(penguins_2, aes(x = `Body Mass (g)`)) + geom_histogram(bins = 30, fill = "black")
```



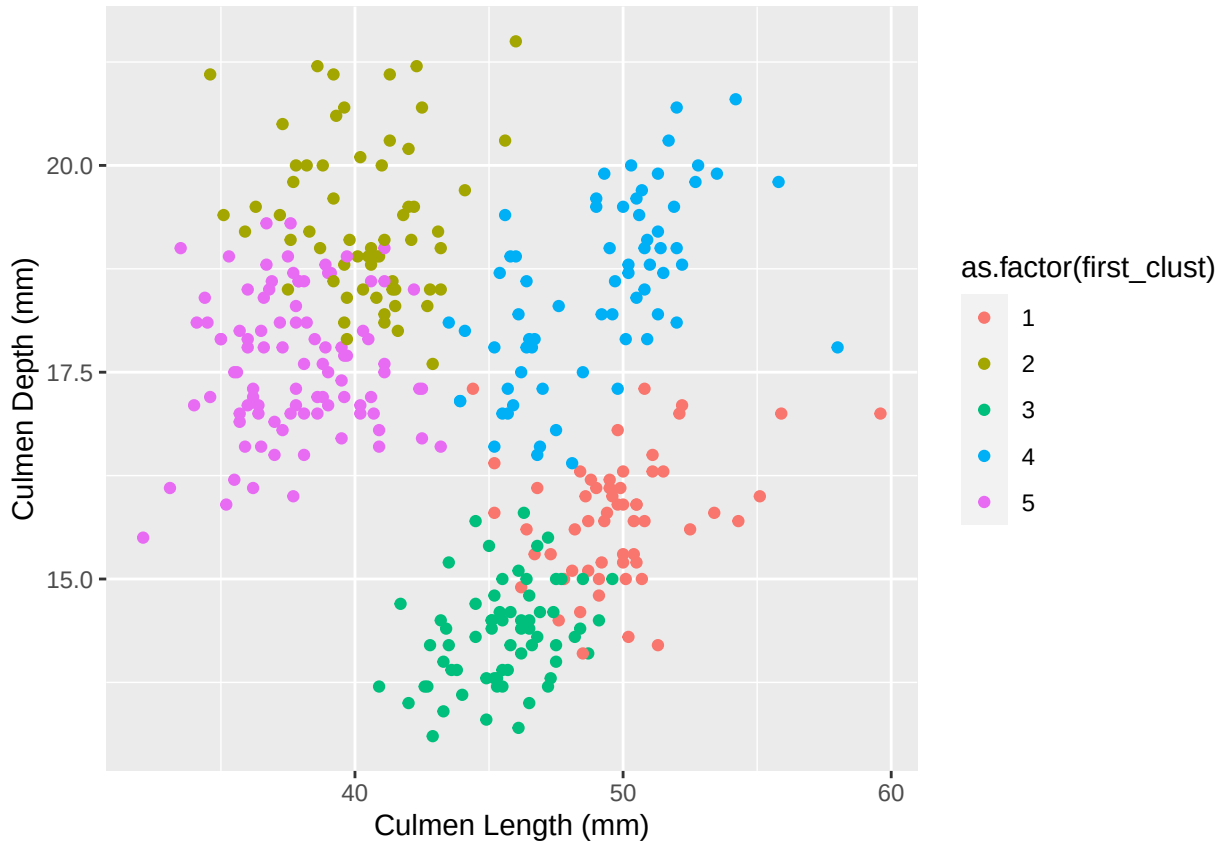

```
#scaling/standardizing the data to get better results for clustering
scale_peng <- data.frame(scale(penguins_2))
#create k means formula for cluster 1
set.seed(123) #for reproducible results
first_clust <- kmeans(scale_peng, centers = 5, nstart = 1)
first_clust$size
```

```
## [1] 57 61 66 67 93
```

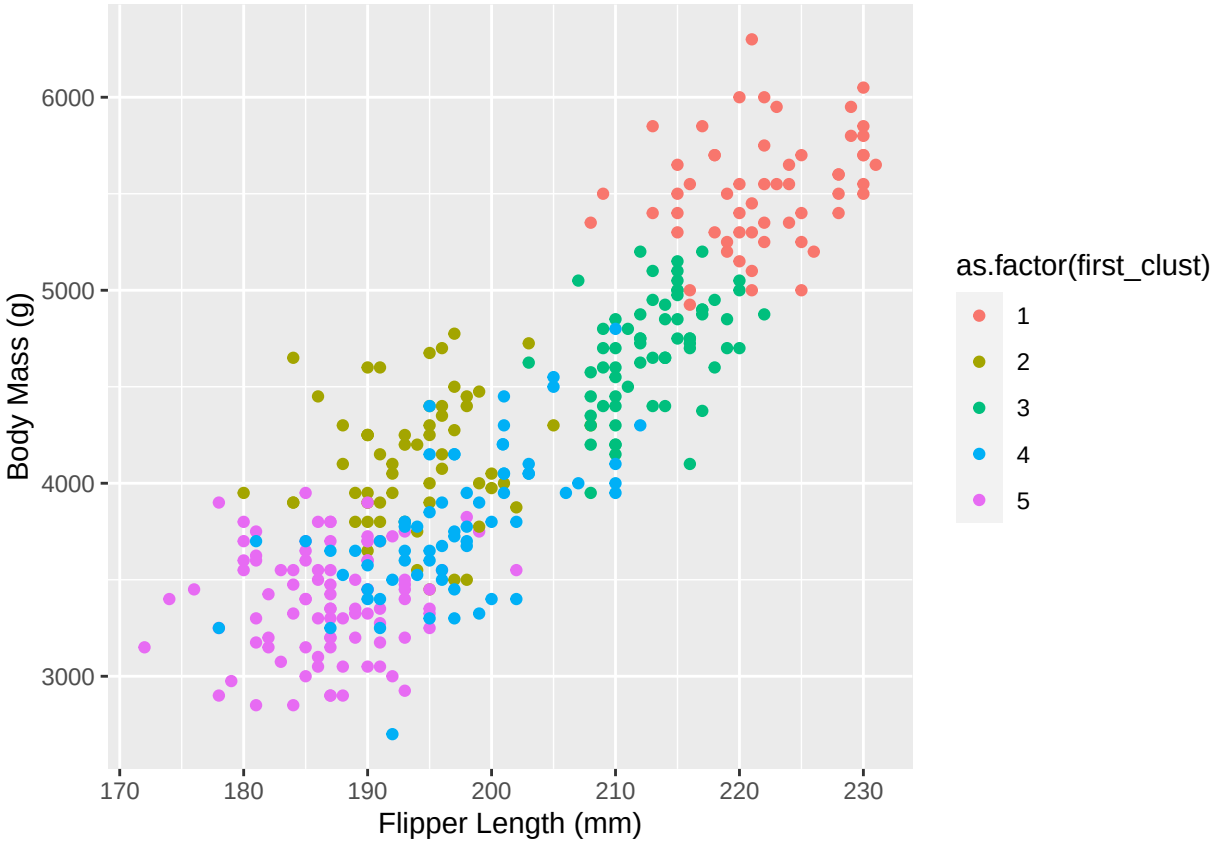
```
#create k means formula for cluster 2
set.seed(321)
second_clust <- kmeans(scale_peng, centers = 5, nstart = 1)
second_clust$size
```

```
## [1] 57 66 67 93 61
```

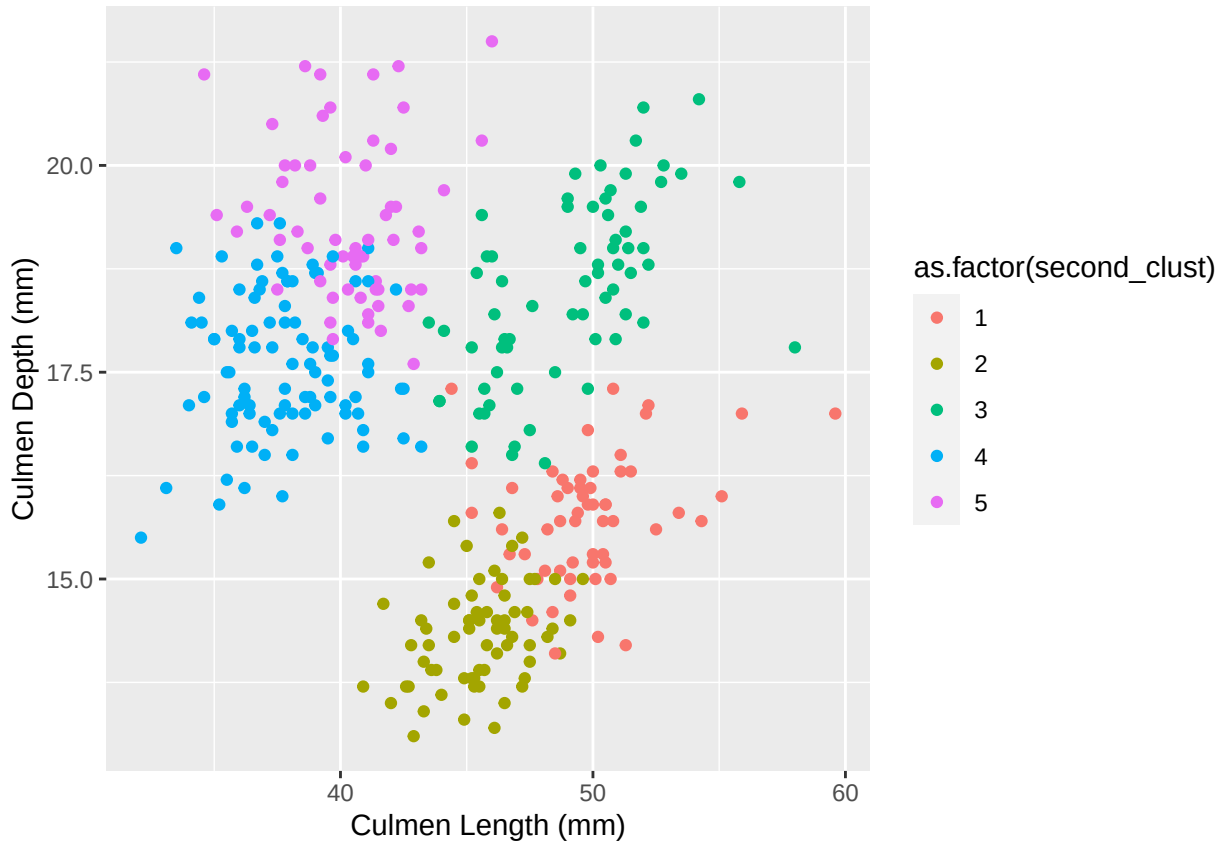
```
#add clusters to data frame unscaled
penguins_2$first_clust <- first_clust$cluster
penguins_2$second_clust <- second_clust$cluster

#plot variables colored by first cluster
ggplot(penguins_2, aes(x = `Culmen Length (mm)`,y = `Culmen Depth (mm)`, color = as.factor(first_clust))
```



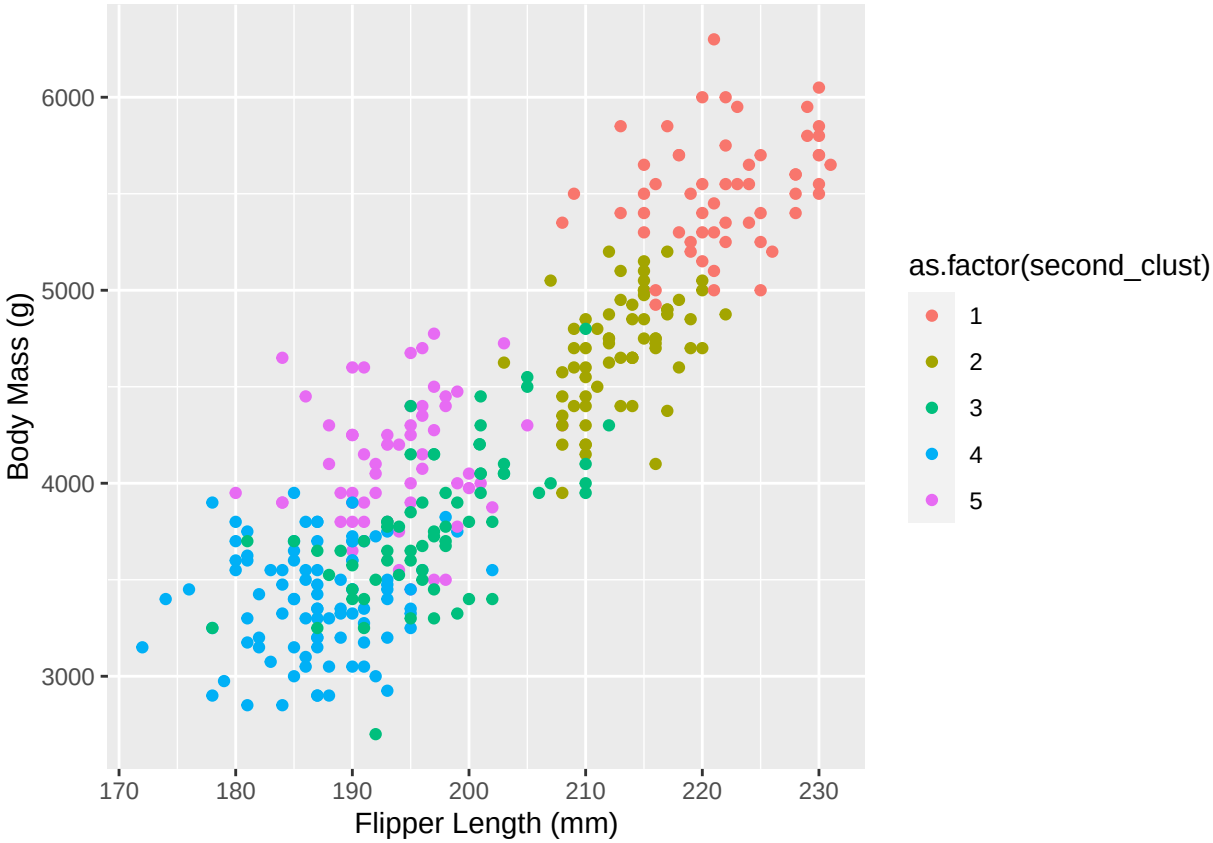

```
ggplot(penguins_2, aes(x = `Flipper Length (mm)`,y = `Body Mass (g)`, color = as.factor(first_clust))) +
```



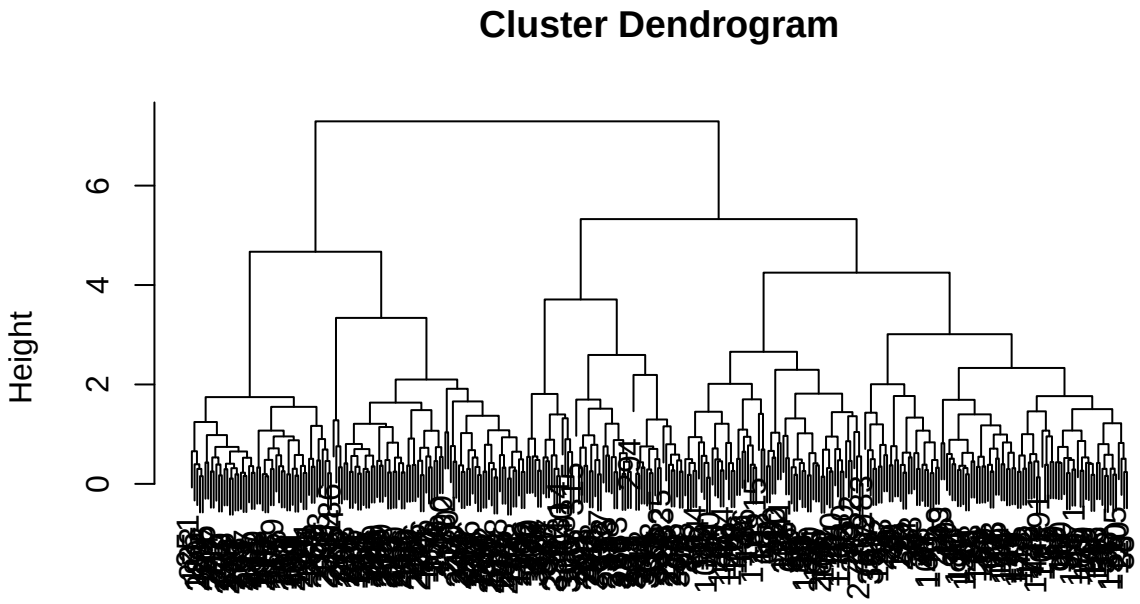

```
#plot variables colored by first cluster
ggplot(penguins_2, aes(x = `Culmen Length (mm)`,y = `Culmen Depth (mm)`, color = as.factor(second_clust
```




```
ggplot(penguins_2, aes(x = `Flipper Length (mm)`,y = `Body Mass (g)`, color = as.factor(second_clust)))
```



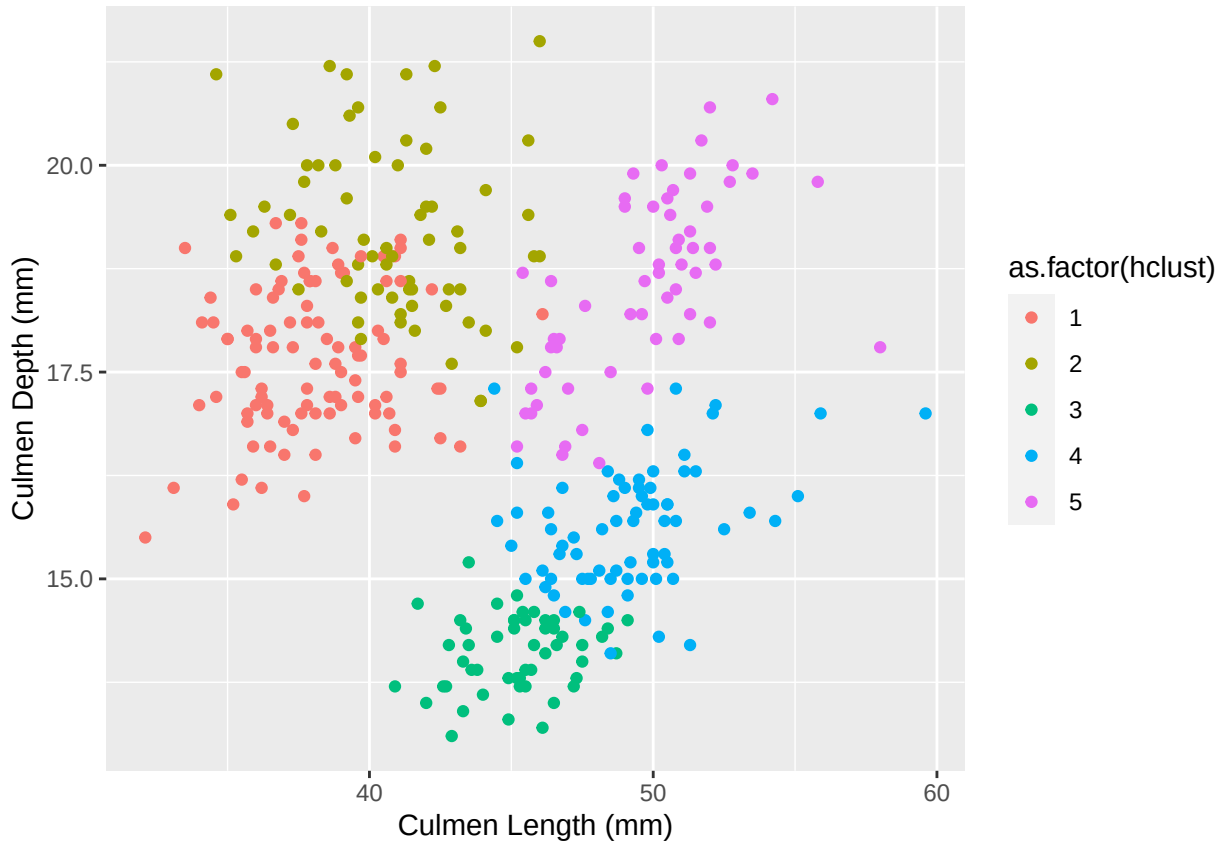

```
#now use hierarchical clustering with complete linkage
hclust_comp <- hclust(dist(scale_peng), method = "complete")
plot(hclust_comp)
```



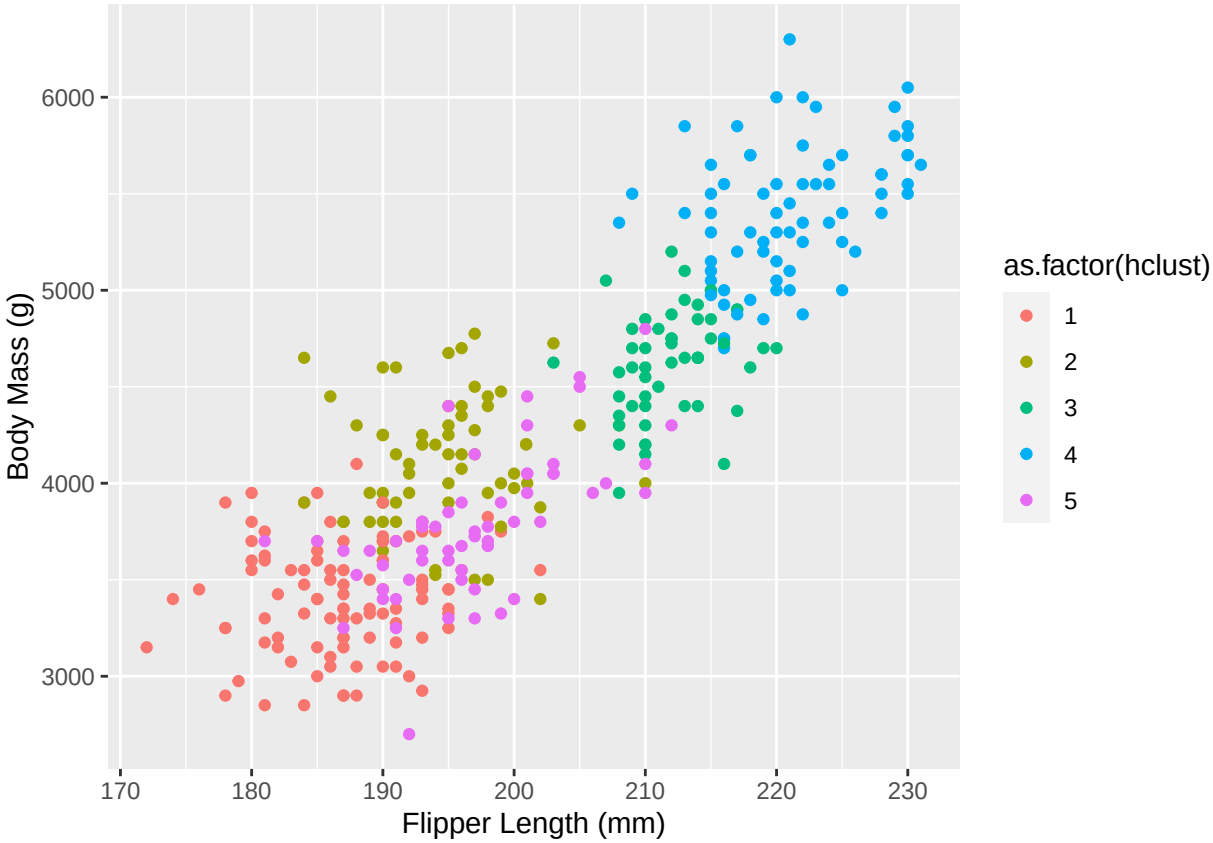

```
hc_comp_assign <- cutree(hclust_comp, k = 5)
#add clustering to data frame
penguins_2$hclust <- hc_comp_assign
pen_simple <- penguins_2 %>% select(-first_clust, -second_clust)
#get summary statistics for hclust
hclust_summary <- do.call(data.frame, aggregate(. ~ hclust, data = pen_simple, function(x) c(avg = mean
hclust_summary
```

```
##   hclust Culmen.Length..mm..avg Culmen.Length..mm..sd Culmen.Depth..mm..avg
## 1      1               38.04021              2.452875               17.66289
## 2      2               40.80824              2.760430               19.21519
## 3      3               45.20000              1.866422               14.10385
## 4      4               49.19296              2.679303               15.62535
## 5      5               49.70345              2.712218               18.54655
##   Culmen.Depth..mm..sd Flipper.Length..mm..avg Flipper.Length..mm..sd
## 1            0.8696646                186.9381                5.576764
## 2            1.0373798                194.6641                5.309690
## 3            0.4498198                212.0000                3.429972
## 4            0.7334300                220.9859                5.486328
## 5            1.1209669                196.8966                6.549915
##   Body.Mass..g..avg Body.Mass..g..sd
## 1          3410.052         285.7205
## 2          4107.629         323.9330
## 3          4624.038         275.4658
## 4          5407.042         353.1790
## 5          3778.879         371.7491
```

```
#plot variables colored by hierarchichal cluster
ggplot(penguins_2, aes(x = `Culmen Length (mm)`,y = `Culmen Depth (mm)`, color = as.factor(hclust))) + g
```




```
ggplot(penguins_2, aes(x = `Flipper Length (mm)`,y = `Body Mass (g)`, color = as.factor(hclust))) + geo
```


Body Mass (g)
6000
5000
4000
3000
170
180
190
200
210
220
230
Flipper Length (mm)
as.factor(hclust)
1
2
3
4
5