Homework #1

Matt Quintiere and Rohit Gunda

2023-10-16

Setup

```
library(tidyverse)
library(tidymodels)
library(dplyr)
library(knitr)
library(palmerpenguins)
library(GGally)
library(factoextra)
library(psych)
library(patchwork)
library(factoextra)
library(cluster)
library(mclust)
library(kernlab)
```

Introduction

```
penguins_raw <- penguins_raw
names(penguins_raw)
```

```
##  [1] "studyName"           "Sample Number"       "Species"
##  [4] "Region"              "Island"              "Stage"
##  [7] "Individual ID"       "Clutch Completion"   "Date Egg"
## [10] "Culmen Length (mm)"  "Culmen Depth (mm)"   "Flipper Length (mm)"
## [13] "Body Mass (g)"       "Sex"                 "Delta 15 N (o/oo)"
## [16] "Delta 13 C (o/oo)"   "Comments"
```

Data Introduction: The dataset we aim to analyze contains data on 344 penguins. This dataset contains data on 3 species of penguins (Adelie, Chinstrap, and Gentoo) from 3 islands in the Palmer Archipelago (Torgerson, Biscoe, and Dream). There are 17 variables in the dataset, and they consist of the following:

- **studyName:** Sampling expedition from which data were collected, generated, etc.
- **Sample Number:** an integer denoting the continuous numbering sequence for each sample
- **Species:** a character string denoting the penguin species
- **Region:** a character string denoting the region of Palmer LTER sampling grid
- **Island:** a character string denoting the island near Palmer Station where samples were collected
- **Stage:** a character string denoting reproductive stage at sampling
- **Individual ID:** a character string denoting the unique ID for each individual in dataset

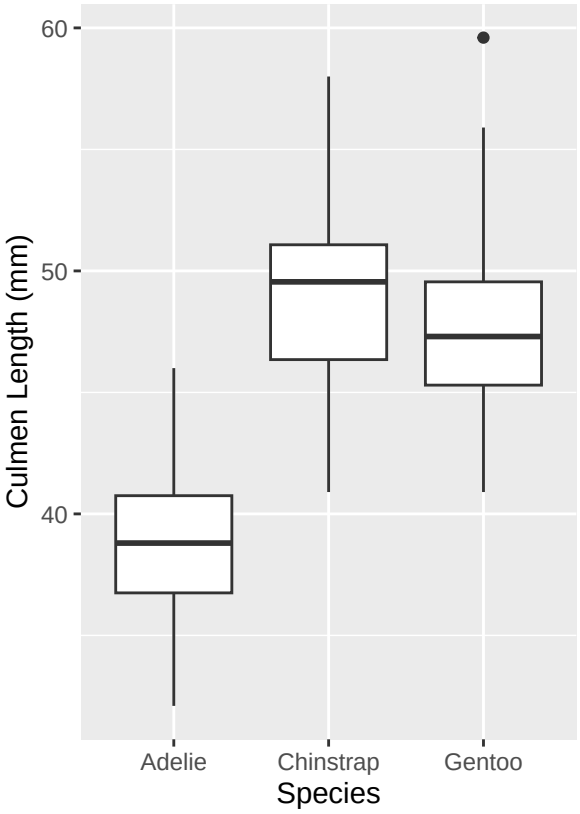
- **Clutch Completion:** a character string denoting if the study nest observed with a full clutch, i.e., 2 eggs
- **Date Egg:** a date denoting the date study nest observed with 1 egg (sampled)
- **Culmen Length (mm):** a number denoting the length of the dorsal ridge of a bird's bill (millimeters)
- **Culmen Depth (mm):** a number denoting the depth of the dorsal ridge of a bird's bill (millimeters)
- **Flipper Length (mm):** an integer denoting the length penguin flipper (millimeters)
- **Body Mass (g):** an integer denoting the penguin body mass (grams)
- **Sex:** a character string denoting the sex of an animal
- **Delta 15 N:** a number denoting the measure of the ratio of stable isotopes 15N:14N
- **Delta 13 C:** a number denoting the measure of the ratio of stable isotopes 13C:12C
- **Comments:** a character string with text providing additional relevant information for data

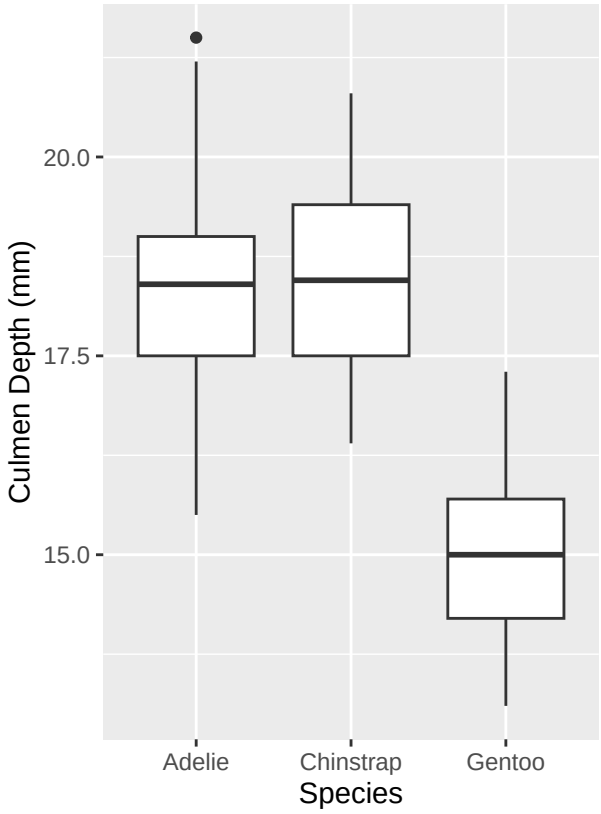
Motivation: We decided to use this data set for our analysis because we were interested to see if there was any variation in some of the key metrics between penguins of different species or penguins of different islands. To gain a better understanding of these key metrics, we will be using unsupervised learning techniques of Principal Component Analysis (PCA), Factor Analysis, and Cluster Analysis to get a better understanding of how the metrics are related and can be useful in predicting the species of a penguin on these islands. The metrics we will be using are the numeric variables of `Culmen Length (mm)`, `Culmen Depth (mm)`, `Flipper Length (mm)`, and `Body Mass (g)`. We have also kept `Species`, `Island`, and `Sex`. While we will be focusing on analyzing the numeric variables to get a better understanding of how they relate to the penguins' species, we may be interested in the future in conducting a similar analysis to get better understanding of how these numeric variables relate to a penguins' sex or island. Therefore, we will keep the `sex` and `island` variables. We have removed `Region`, `studyName`, `Sample Number`, `Stage`, `Individual ID`, `Clutch Completion`, `Data Egg`, `Delta 15 N`, `Delta 13 C`, and `Comments` as we believe that they are not relevant for our analysis. We will now go ahead and clean the dataset to get it to the version that we want use for our analysis.

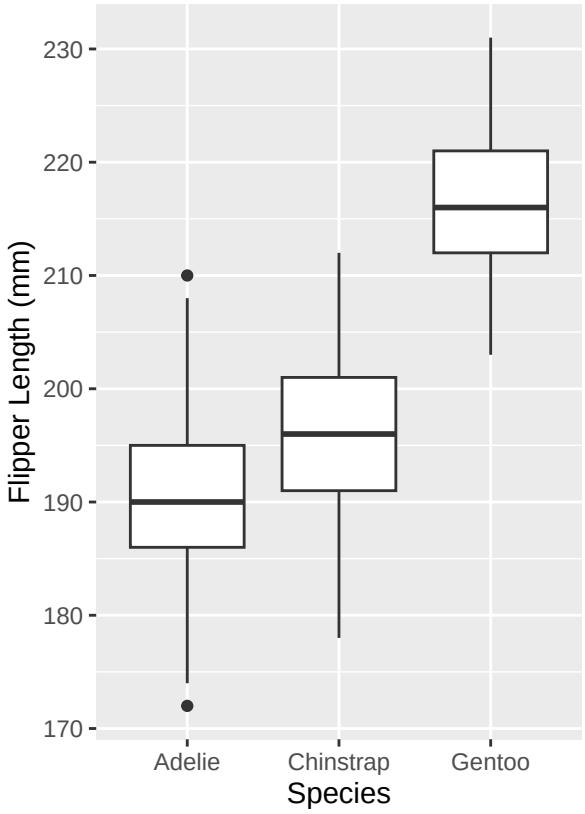
```
penguins_mod <- penguins_raw |>
  select(-studyName, -`Sample Number`, -Region, -`Individual ID`, -Stage,
         -`Clutch Completion`, -`Date Egg`, -`Delta 15 N (o/oo)`,
         -`Delta 13 C (o/oo)`, -Comments)
```

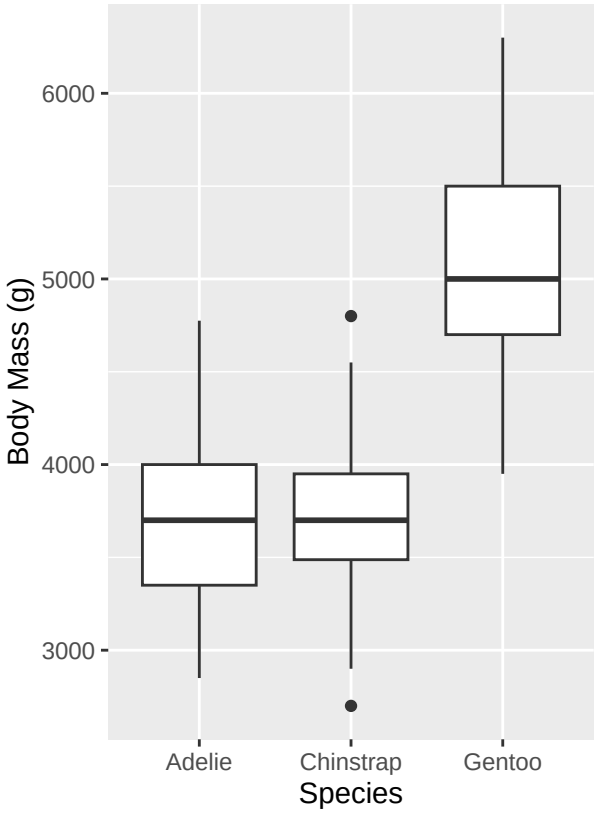
```
penguins2 <- penguins_mod[rowSums(is.na(penguins)) < 2, ]
penguins2 <- penguins2 |>
  mutate(
    Species = case_when(
      Species == "Adelie Penguin (Pygoscelis adeliae)" ~ "Adelie",
      Species == "Gentoo penguin (Pygoscelis papua)" ~ "Gentoo",
      Species == "Chinstrap penguin (Pygoscelis antarctica)" ~ "Chinstrap"
    )
  )
duplicated(penguins2)
```

Here we did some initial data cleaning by removing rows that had 2 or more NA values, which removed all the NA values in our numeric variable columns. We also shortened the names of the species to make it easier when writing our report. We then checked for duplicates in our data and found none to remove. Now we will check for extreme outliers to remove in our dataset:


60
50
40
Culmen Length (mm)
Adelie
Chinstrap
Gentoo
Species


20.0
17.5
15.0
Culmen Depth (mm)
Adelie
Chinstrap
Gentoo
Species





Through viewing these boxplots of our numeric variables and how they relate to the species of the penguins, we can tell that there are no extreme outliers, which means that we can use this data set for our unsupervised learning analysis:

Species	Island	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex
Adelie	Torgersen	39.1	18.7	181	3750	MALE
Adelie	Torgersen	39.5	17.4	186	3800	FEMALE
Adelie	Torgersen	40.3	18.0	195	3250	FEMALE
Adelie	Torgersen	36.7	19.3	193	3450	FEMALE
Adelie	Torgersen	39.3	20.6	190	3650	MALE
Adelie	Torgersen	38.9	17.8	181	3625	FEMALE

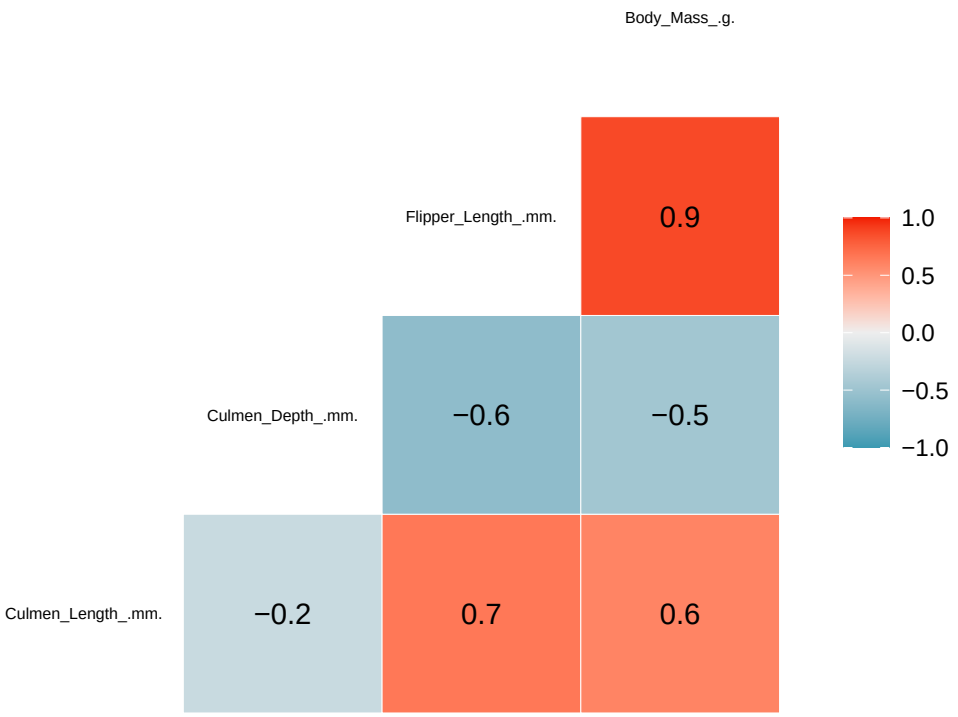
PCA

Our first unsupervised learning technique to learn more about our dataset will be the PCA. To explore the principal components of our data set to explain the variability, we have to isolate the numeric variables in the data set, which are `Culmen Length (mm)`, `Culmen Depth (mm)`, `Flipper Length (mm)length_mm`, `Body Mass (g)`. The reason we can only use numeric variables is because the PCA relies linear algebra calculations to find a linear combination that explains majority of the total variance, which can only be used with numeric data.

Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)
39.1	18.7	181	3750
39.5	17.4	186	3800

Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)
40.3	18.0	195	3250
36.7	19.3	193	3450
39.3	20.6	190	3650
38.9	17.8	181	3625

Let's visualize how correlated these variables are by creating a correlation plot:

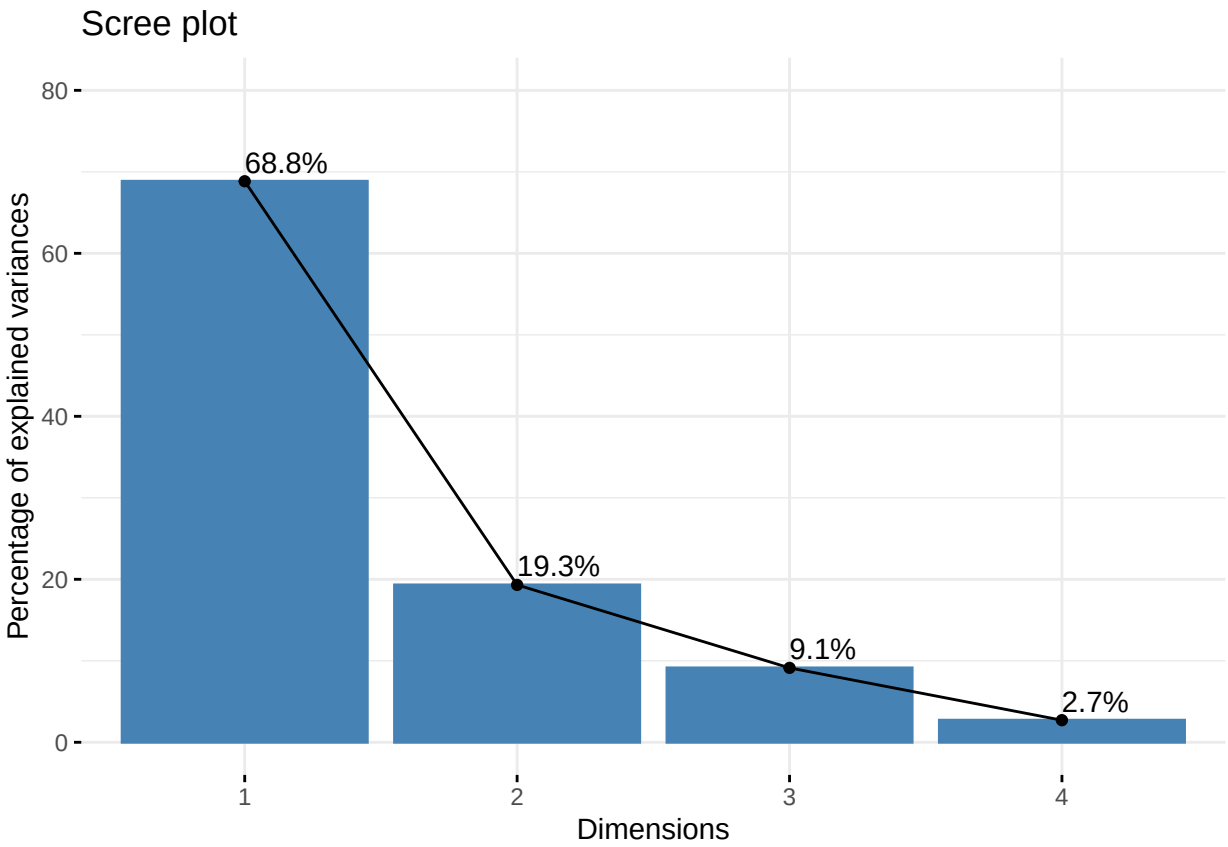


From the correlation matrix, we can see that `Body Mass (g)` is very positively correlated with `Flipper Length (mm)`. Also `Culmen Length (mm)` has a somewhat strong correlation with `Flipper Length (mm)`, while `Culmen Depth (mm)` has a negative correlation with all of the variables and does not have a strong correlation with any of the variables.

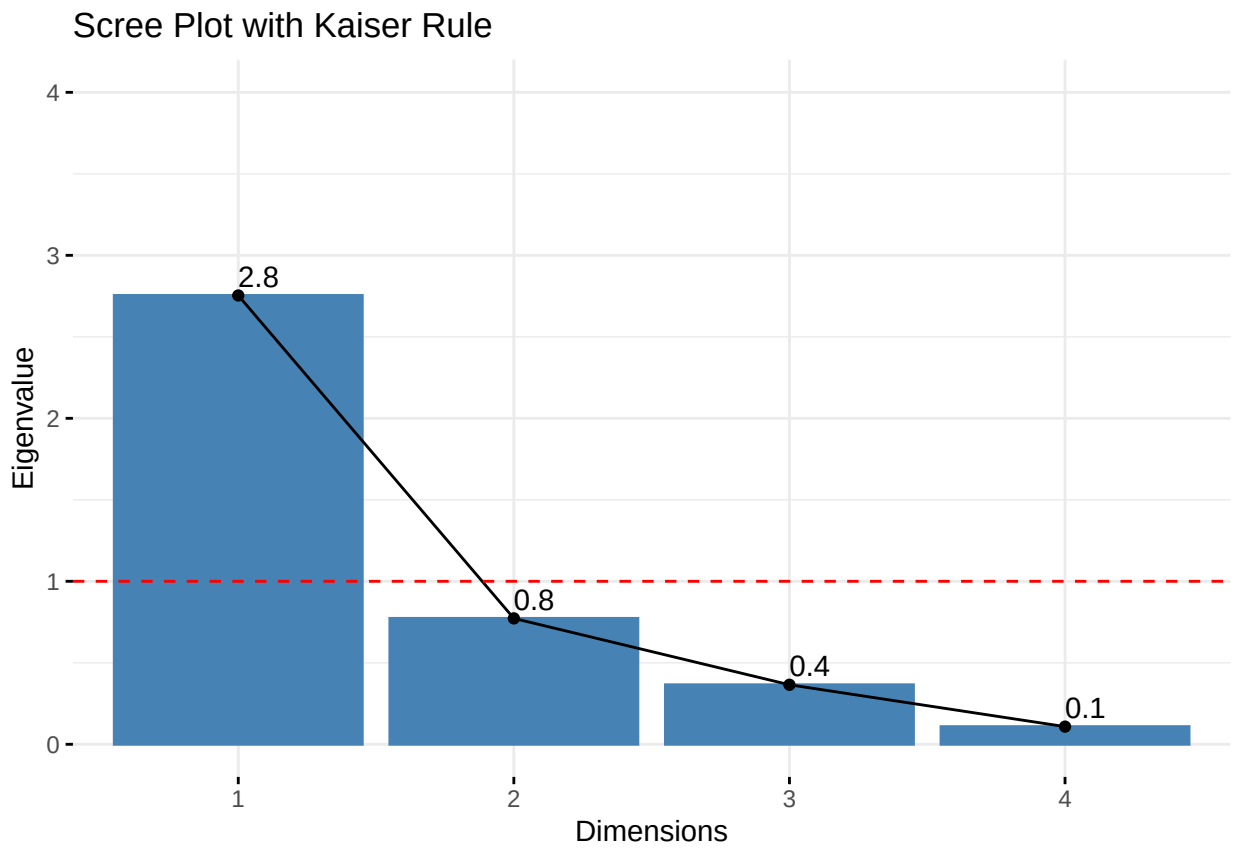
```
pca = prcomp(pca_penguins2, scale = TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.6594 0.8789 0.60435 0.32938
## Proportion of Variance 0.6884 0.1931 0.09131 0.02712
## Cumulative Proportion  0.6884 0.8816 0.97288 1.00000
```

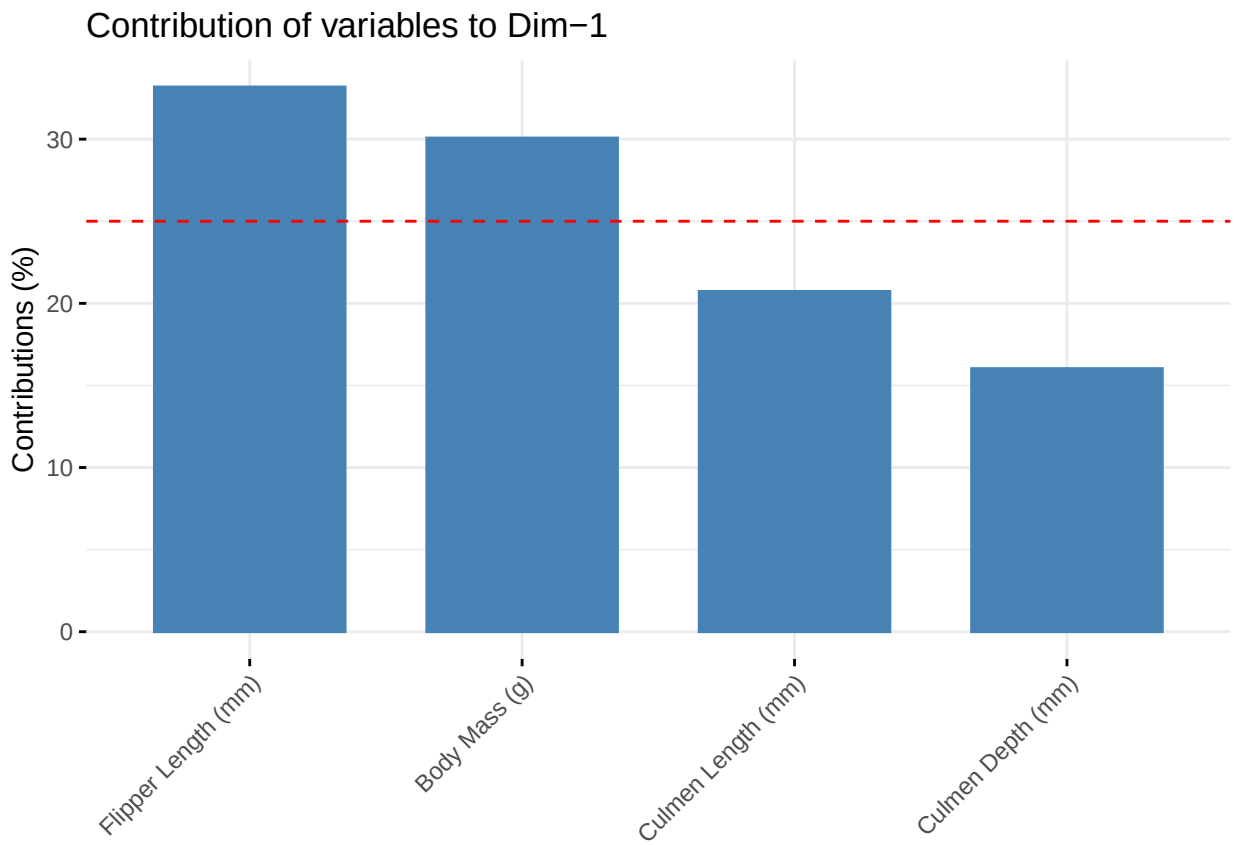
Running the PCA calculations out using the prcomp tool, we see that the first two principal components are responsible for about 88% percent of the data. Let's create a screeplot to visualize this:



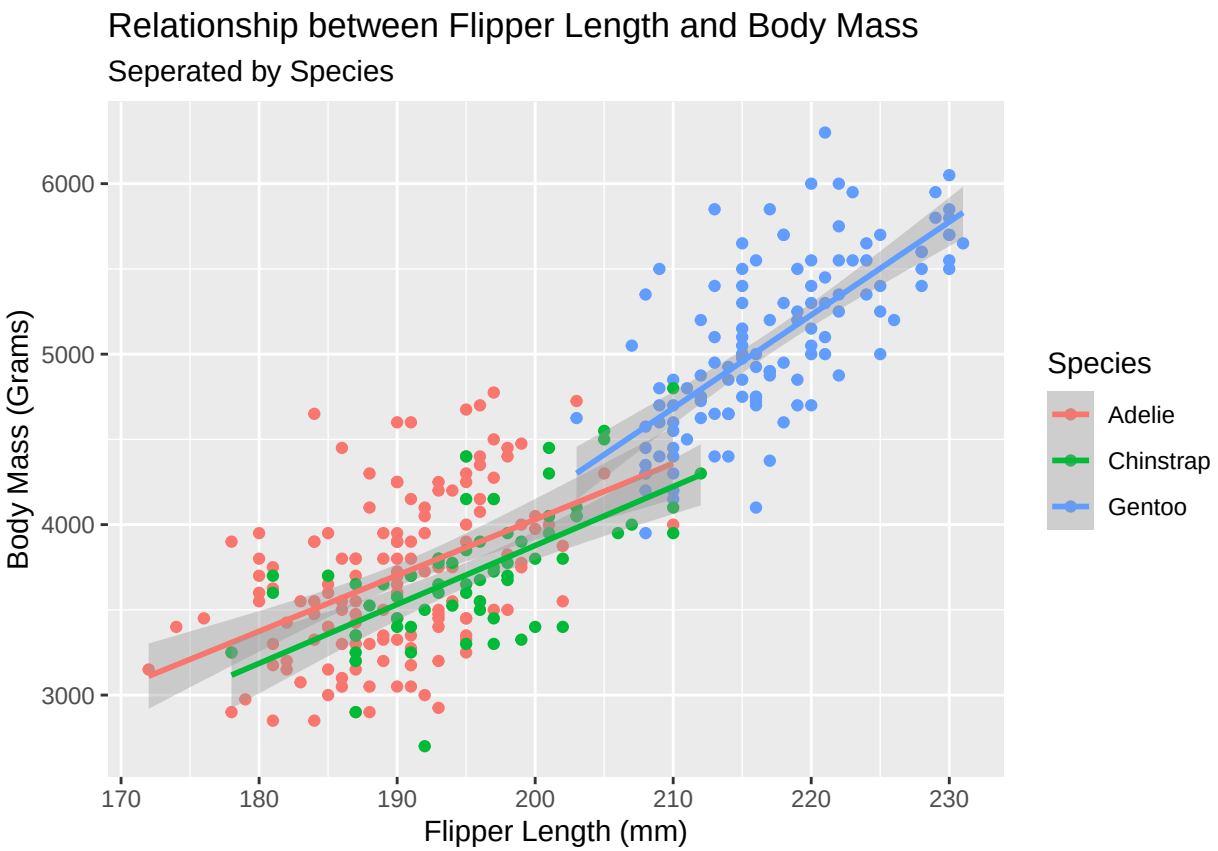
Now let's apply the Kaiser rule to select the number of components, we will use in the PCA:



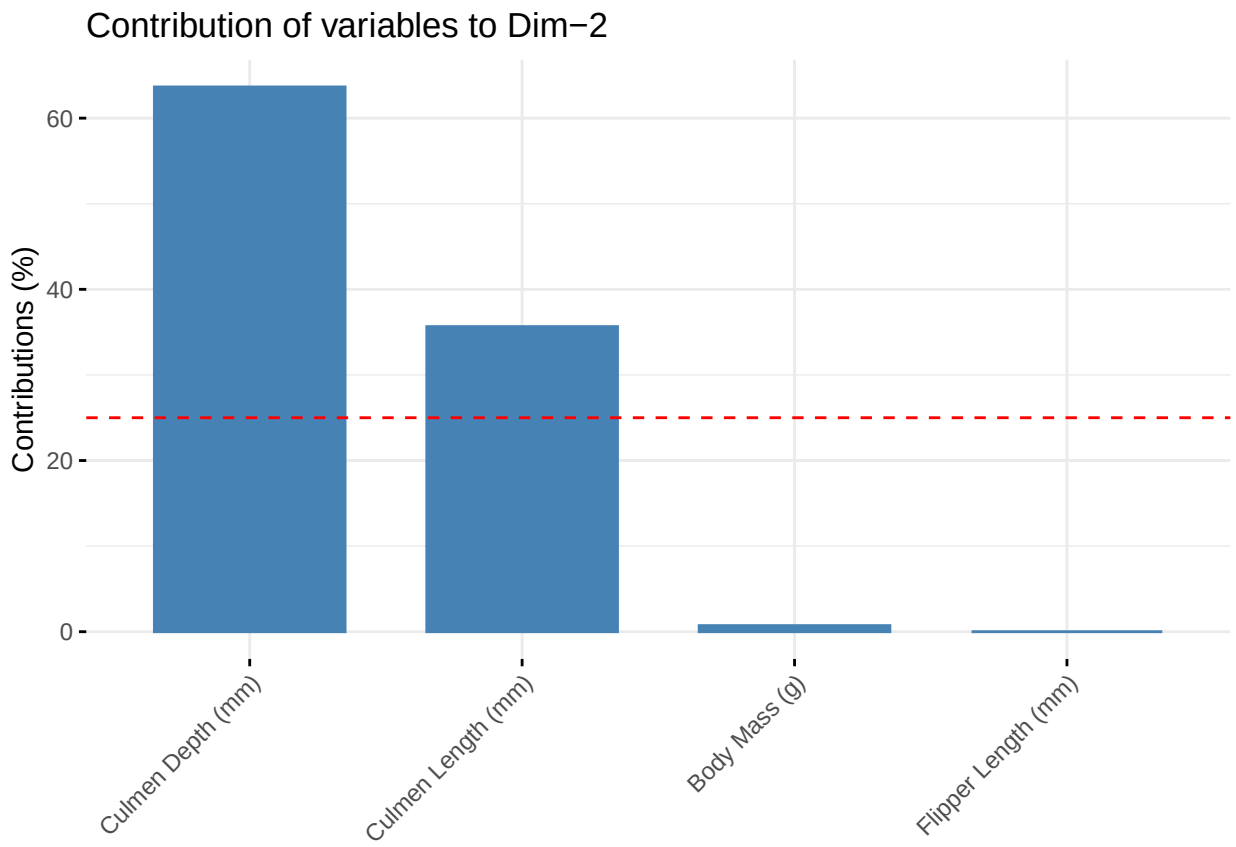
Only our first principal component has an eigenvalue greater than one, so our analysis will focus on the first principal component to start, which explains about 69% of the variability and is the maximum variance direction in the data. However, we may still use the second principal component if necessary. Now let's look at what variables contribute the most to our first principal component:



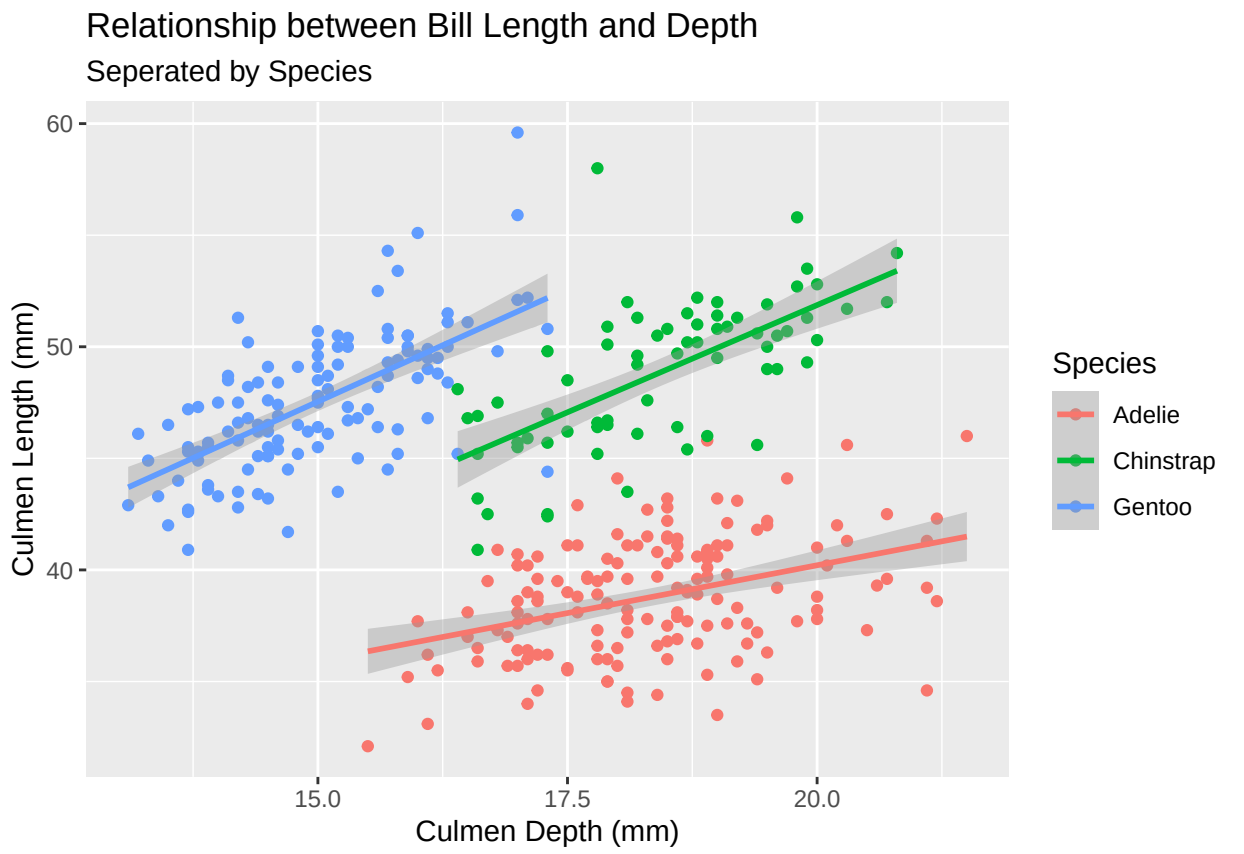
On average, each variable is expected to contributed 25% to the first principal component. However, only two of those variables `Flipper Length (mm)` and `Body Mass (g)` contribute over 25% to the first principal component. We should note that a reason that this could occur is that `Flipper Length (mm)` and `Body Mass (g)` are highly correlated. Let's visualize this correlated relationship with respect to `Species`:



From this chart, we can see the differentiation of the Gentoo species from the Adelie and Chinstrap species of penguins, as the Gentoo species tends to have a greater body mass and flipper length. However, Adelie and Chinstrap are not differentiable based on their flipper length and body mass relationship. Therefore, the first principal component is an overall measure for the size of the penguins which differentiates the Gentoos. Another thing to note is that `Body Mass (g)` and `Flipper Length (mm)` display a strong positive correlation for each of the species as predicted in the correlation plot. Let's circle back to the second principal component, so that we can find a way to differentiate the Adelie and Chinstrap species as they are similar sizes:



In the second principal component, the two other variables contribute more than expected. Let's visualize the relationship of `Culmen Length (mm)` and `Culmen Depth (mm)` with respect to species:



The second principal component is explained primarily by `Culmen Length (mm)` and `Culmen Depth (mm)`. From looking at the scatter plot above, we can determine that Chinstrap penguins have similar culmen depths as Adelie penguins, but much larger culmen lengths. Also, despite being smaller penguins in size, the Chinstrap and Adelie penguins have a larger culmen depths than Gentoo penguins. Gentoo penguins; however, have a much larger culmen length than Adelie penguins. Also, it is interesting to note that `Culmen Depth (mm)` and `Culmen Length (mm)` were not strongly correlated based on our correlation matrix and were also negatively correlated. However, when the relationship of these variables are seperated by species, there is at least a somewhat strong positive correlation.

Factor Analysis

After determining the meaning of our first two principal components and using the PCA to attempt to explain the total variance, we will move on to completing a factor analysis in order to better explain the variance and covariance of the observed variables in our data set by a set of fewer latent variables. Let's test if our factor analysis will run:

```
det(cor(pca_penguins2))
```

```
## [1] 0.08429567
```

Our determinant is positive, so that means our factor analysis will most likely run.

We will be using the principal axis factor analysis and maximum likelihood factor analysis methods. We will only be using a total of two factors based on the reasoning from the PCA above. We will also be using the varimax rotational method which will minimize the number of variables that have high loadings on each factor and simplify the interpretation of each factor. We will use the method that has the highest cumulative variance on the second factor.

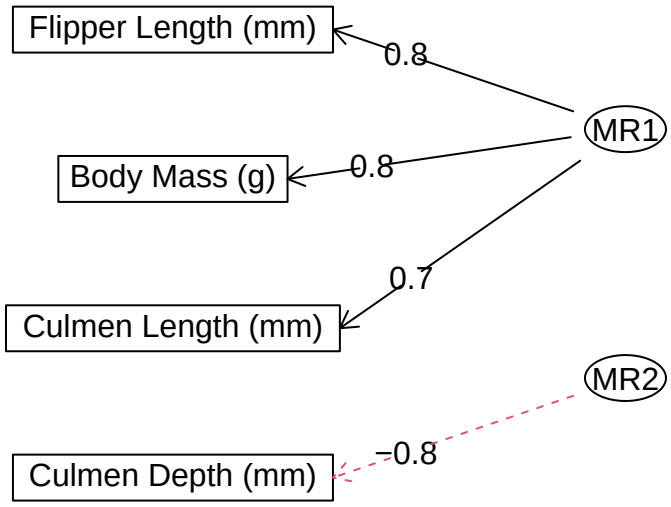
```
pa <- fa(r = pca_penguins2,
         nfactors =2,
         rotate = "varimax",
         fa = "pa",
         resdiuals = TRUE)
ml <- fa(r = pca_penguins2,
         nfactors =2,
         rotate = "varimax",
         fa = "ml",
         resdiuals = TRUE)
```

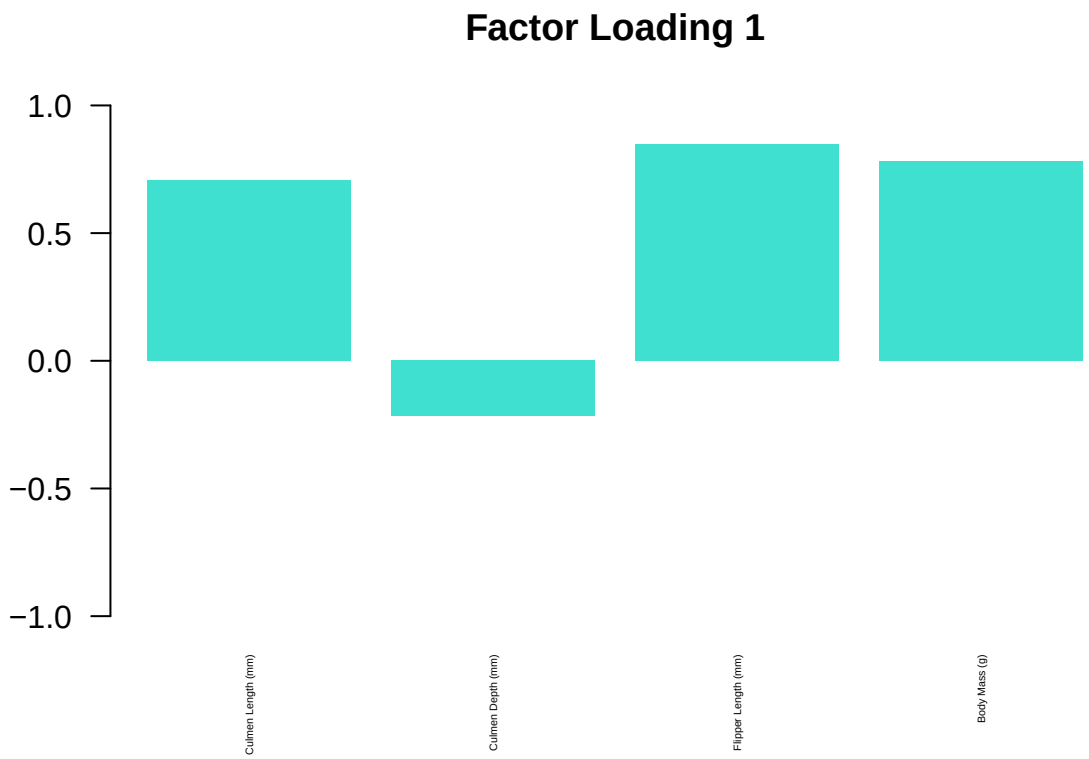
The PA and ML methods both produce a cumulative variance of 0.72, so either method will work in this case. We will use the ML method going ahead. So let's view and interpret the results:

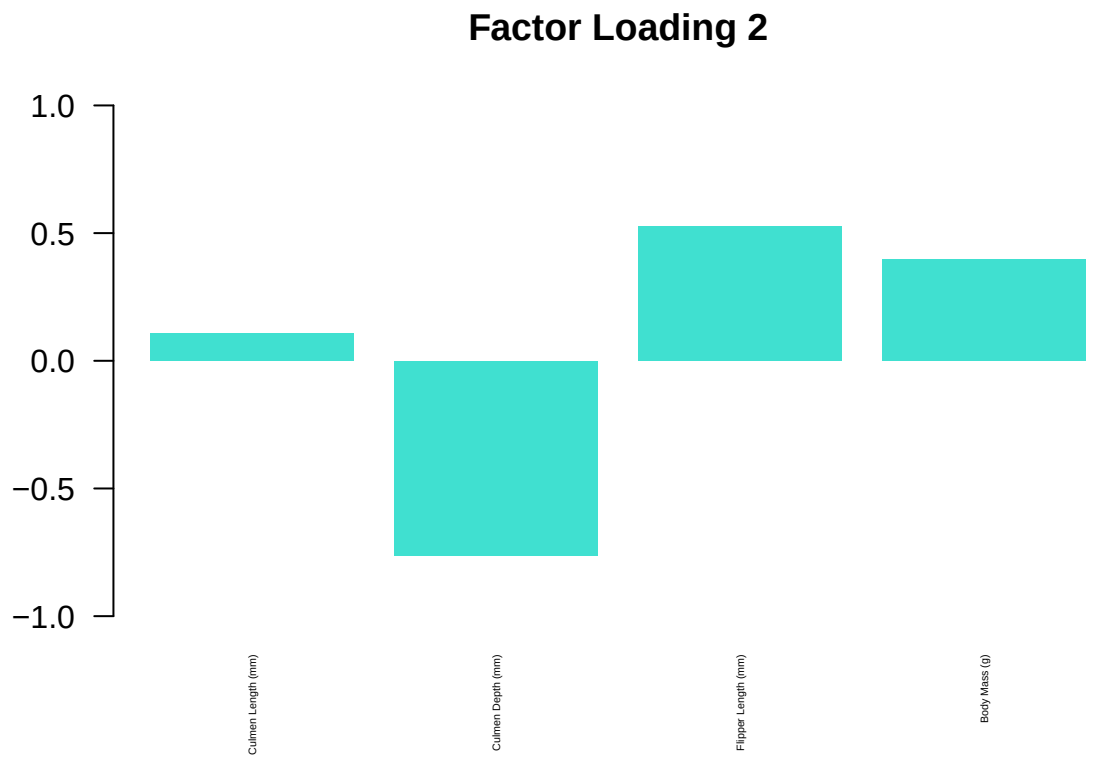
```
ml
```

```
## Factor Analysis using method =  minres
## Call: fa(r = pca_penguins2, nfactors = 2, rotate = "varimax", fa = "ml",
##     resdiuals = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                       MR1   MR2   h2    u2 com
## Culmen Length (mm)   0.71  0.11 0.51 0.489 1.0
## Culmen Depth (mm)   -0.22 -0.76 0.63 0.371 1.2
## Flipper Length (mm)  0.85  0.52 0.99 0.005 1.7
## Body Mass (g)        0.78  0.40 0.77 0.232 1.5
##
##                        MR1  MR2
## SS loadings           1.88 1.03
## Proportion Var        0.47 0.26
## Cumulative Var        0.47 0.73
## Proportion Explained  0.65 0.35
## Cumulative Proportion 0.65 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  6  with the objective function =  2.47 with Chi Square =  838.08
## df of  the model are -1  and the objective function was  0
##
## The root mean square of the residuals (RMSR) is  0
## The df corrected root mean square of the residuals is  NA
##
## The harmonic n.obs is  342 with the empirical chi square  0  with prob <  NA
## The total n.obs was  342  with Likelihood Chi Square =  0  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  1.007
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                MR1  MR2
## Correlation of (regression) scores with factors   0.92 0.79
## Multiple R square of scores with factors          0.85 0.63
## Minimum correlation of possible factor scores     0.71 0.26
```

Factor Analysis Charted




Factor Loading 1
1.0
0.5
0.0
−0.5
−1.0
Culmen Length (mm)
Culmen Depth (mm)
Flipper Length (mm)
Body Mass (g)



From these tables and charts, we can determine that the first factor (MR1) is related to `Flipper Length (mm)`, `Body Mass (g)`, and `Culmen Length (mm)`. However, `Culmen Length (mm)` is a feature that is poorly explained by factor analysis because less 50% of its variance is explain by the two given factor loadings. `Flipper Length (mm)` is the best explained variable by factor analysis as over 99% of its variance is explained by the two factor loadings. Therefore, the first factor loading represents the size of the penguins as it did in the PCA and it causes the flipper lengths and masses of the penguins. The second loading factor is most related to `Culmen Depth (mm)` as `Culmen Depth (mm)` has the highest MR2 value in absolute value. The second factor loading once again represents the shape of the beak, but is not as well explained as the second principal component from the PCA as `Culmen Length (mm)` is not well explained by the second factor loading. The second factor loading is also somewhat related to `Flipper Length (mm)` and `Body Mass (g)`, which also have the highest complexity variables. This means that these two variables do most of the explanation of the variance and covariance in the data set. According to our factor analysis, `Culmen Length (mm)` does not explain a large amount of the variance in the data set and should be considered to be taken out of models trying to predict the species of a given penguin based on the Factor Analysis.

Similar to the PCA, we were able reduce our data into two-dimensions with the size of the penguin and the shape of the beak. It is interesting to note that the first loading factor did a much better job at explaining the variances and covariances of the variables than the second loading factor. This may have occurred because `Culmen Depth (mm)` is not that strongly correlated with the other numeric variables and that only the first principal component has an eigenvalue greater than 1, which does not include `Culmen Depth (mm)` and `Culmen Depth (mm)` is the main variable in the second loading factor.

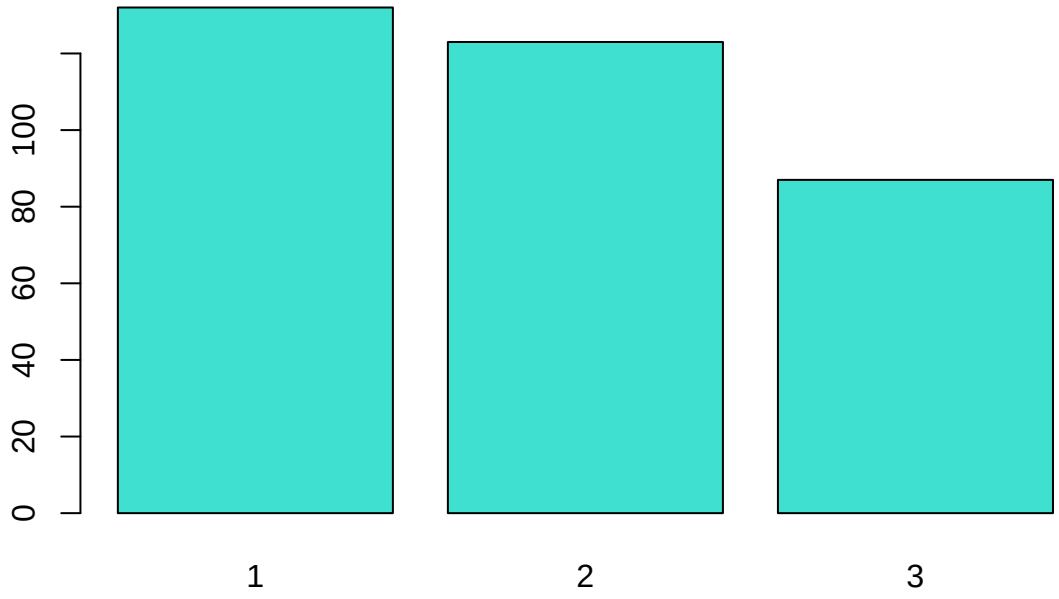
Also, our PCA and factor analysis may have produced slightly different results because all of our variables are not in the same units. The reason for this is because it is impossible to convert grams to millimeters, so the `Body Mass (g)` variable has different units than the rest.

Let's move onto our final unsupervised learning method to further explore our data before we make final conclusions.

Cluster Analysis

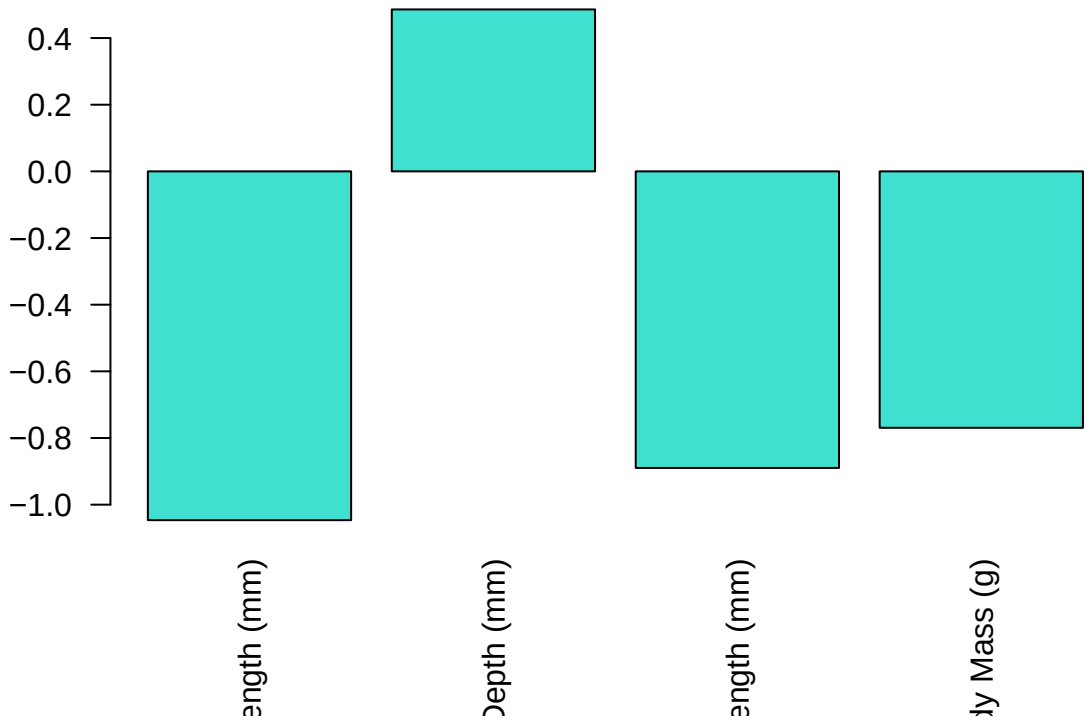
Initial guess for the number of clusters: 3 (signifies the number of species, as well as the number of islands that our data comes from). However, we believe the clusters are the species that each penguin is, as there is a higher variation between species of penguin versus the island that the penguin resides on. The islands could all have an even distribution of species of penguins, so it makes more sense that species is the cluster we should be focusing on.

```
k = 3
fit = kmeans(scale(pca_penguins2), centers=k, nstart=1000)
groups = fit$cluster
barplot(table(groups), col="turquoise")
```

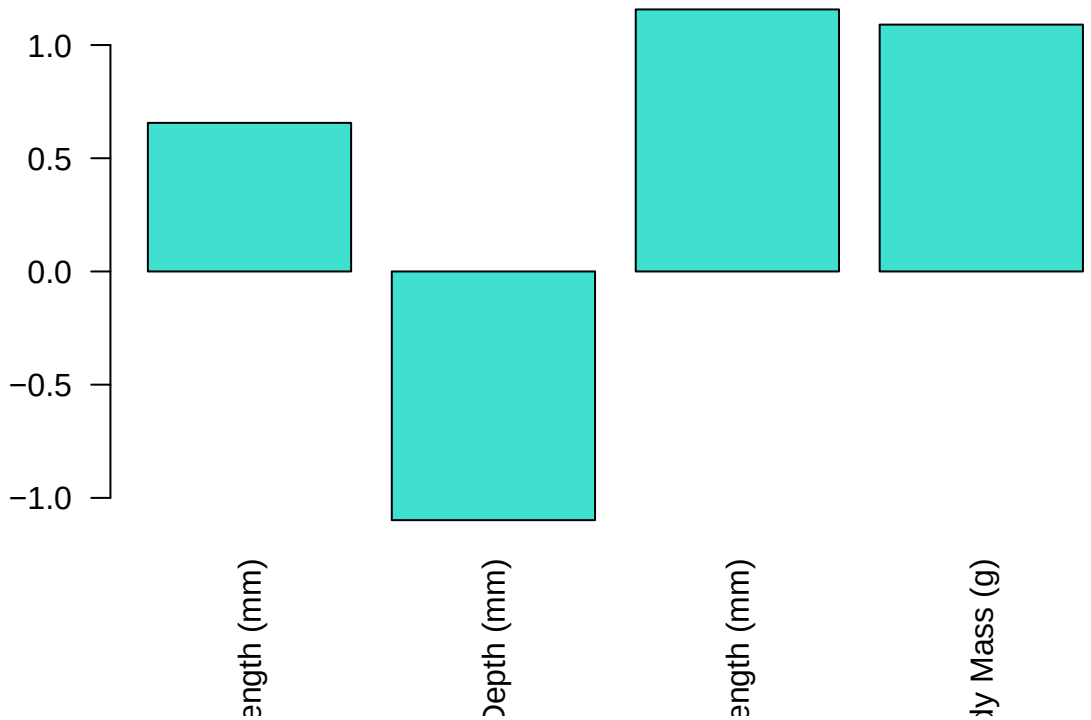


Now to interpret the centers:

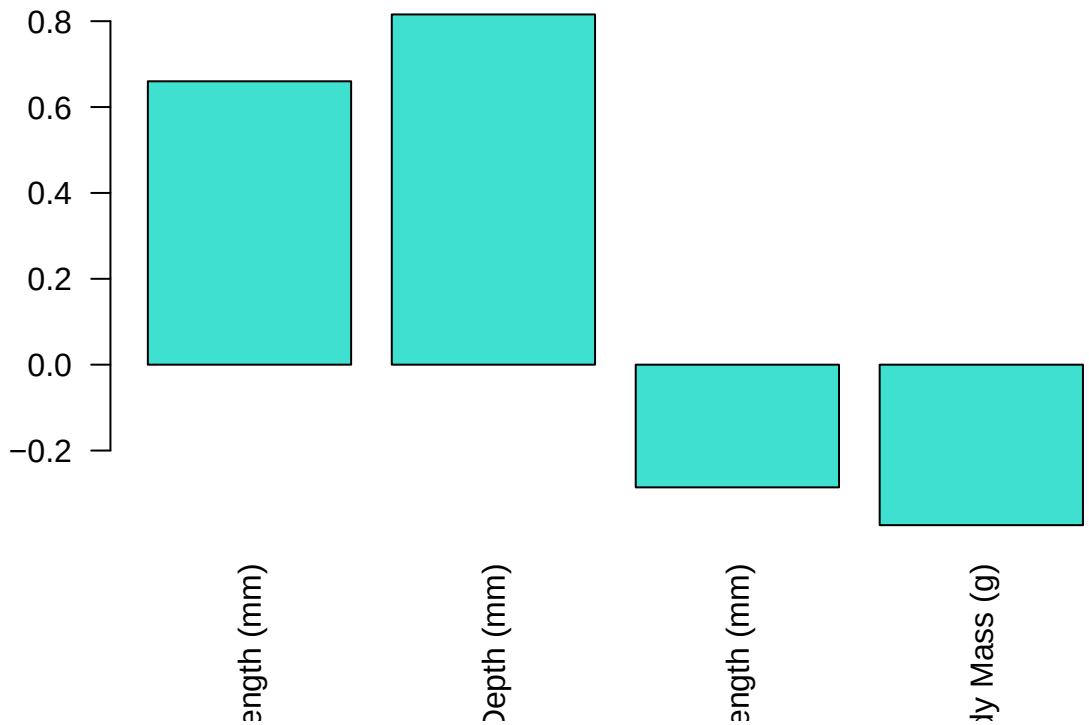
```
centers=fit$centers

barplot(centers[1,], las=2, col="turquoise")
```




```
barplot(centers[2,], las=2, col="turquoise")
```




```
barplot(centers[3,], las=2, col="turquoise")
```



One cluster seems to have the characteristics of having a longer `Culmen depth (mm)`, as seen by the analysis of the centers of our data.

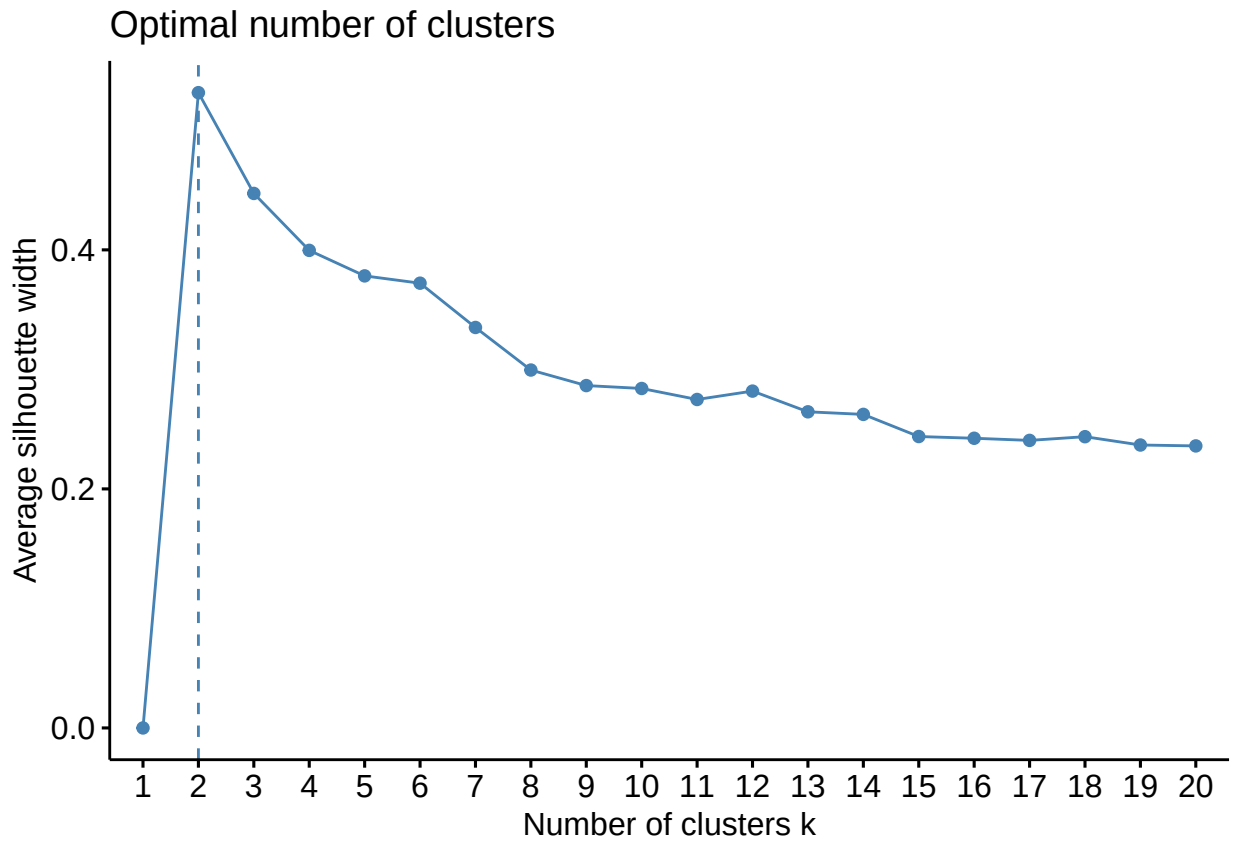
Another cluster seems to have the characteristics of having a much lower `Culmen length (mm)`.

The final cluster seems to have the characteristics of having a longer `Flipper Length (mm)` and `Body Mass (g)`.

Based on this, we can see some of the major differences between each cluster. Specifically, we see these cluster-specific differences in characteristics such as `Culmen length/depth (mm)`, `Body Mass (g)`, and `Flipper Length (mm)`.

Now, let us confirm the most optimal amount of clusters, per the silhouette method.

```
fviz_nbclust(scale(pca_penguins2), kmeans, method = 'silhouette', k.max = 20, nstart = 1000) # With thi
```




```
fit.ker <- kkmeans(as.matrix(pca_penguins2), centers=3, kernel="rbfdot") # Radial Basis kernel (Gaussia
```

```
## Using automatic sigma estimation (sigest) for RBF or laplace kernel
```

```
# By default, Gaussian kernel is used
# By default, sigma parameter is estimated

centers(fit.ker)
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,] 45.52479 17.15726 205.9573 4515.385
## [2,] 41.51870 18.08293 189.6748 3520.528
## [3,] 44.98137 16.02059 208.6863 4663.480
```

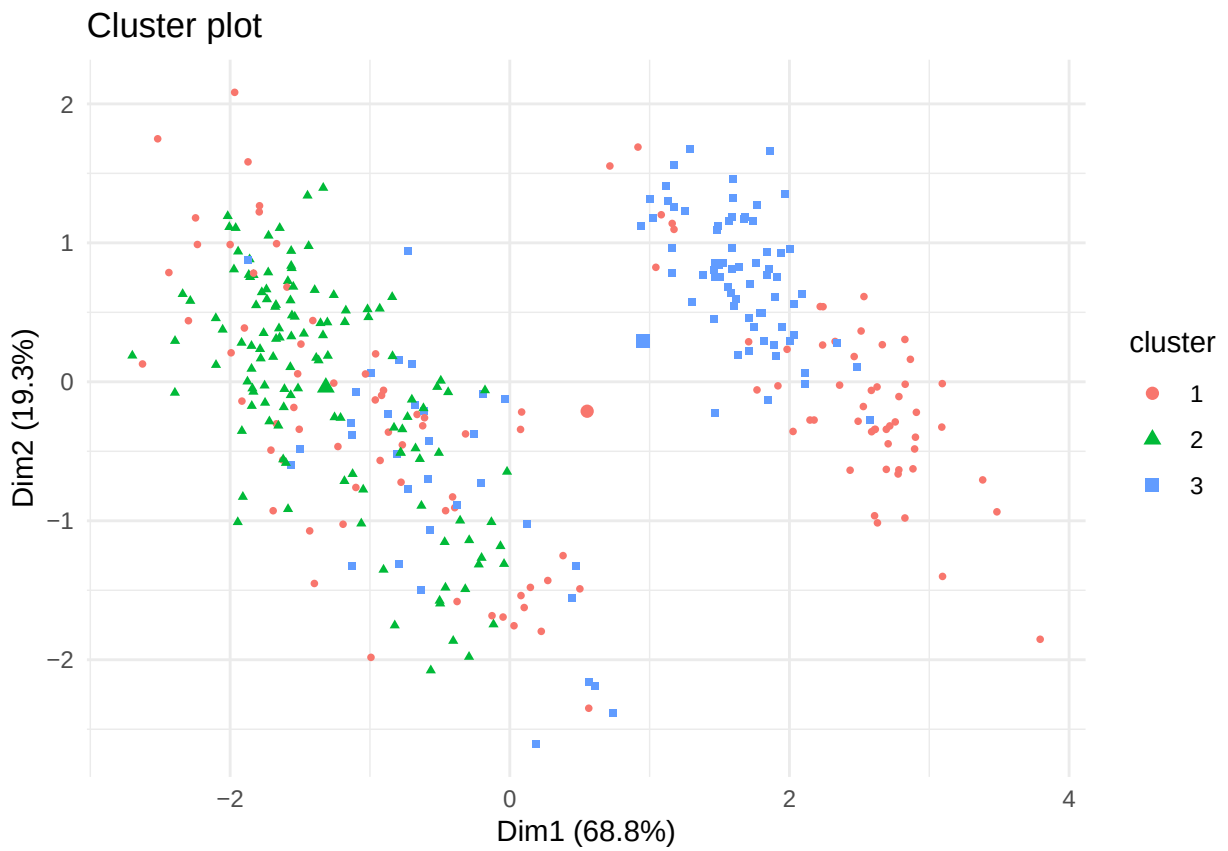
```
size(fit.ker)
```

```
## [1] 117 123 102
```

```
withinss(fit.ker)
```

```
## [1] 3532156687 2182831135 3212666301
```

```
object.ker = list(data = pca_penguins2, cluster = fit.ker@.Data)

fviz_cluster(object.ker, geom = c("point"), ellipse=F,pointsize=1) +
  theme_minimal() #+
```

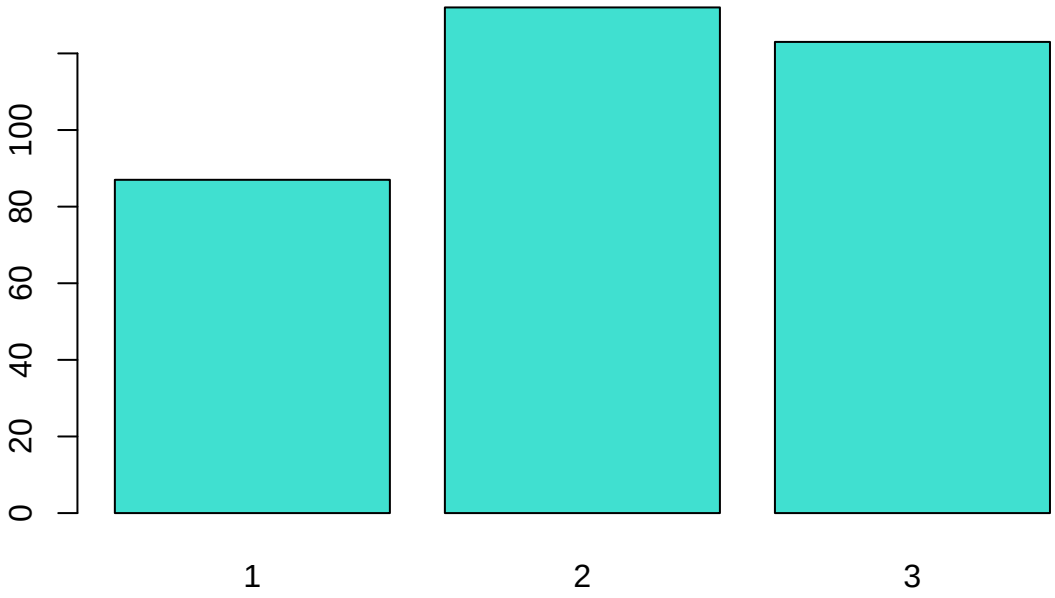



```
  #geom_text(label=names,hjust=0, vjust=0,size=2,check_overlap = T) +
  #scale_fill_brewer(palette="Paired")
```

As seen by the above Kernel k-means analysis, we are using 2 of the principal components to visualize the data. Again, the first 2 principal components make up 88% of the variability. In the above plot, we see 2 very clear groups of points. In the group on the left, it is mainly dominated by clusters 1 and 3, but the one on the right is dominated mainly by points from cluster 2. We found this very interesting as we had expected to see 3 clear clusters, one for each species. However, we only see two, and in one of the groups, points from two clusters are mixed quite evenly. However, k-means are mainly used for highly non-linear data, so maybe another approach to identify clusters may be beneficial to use.

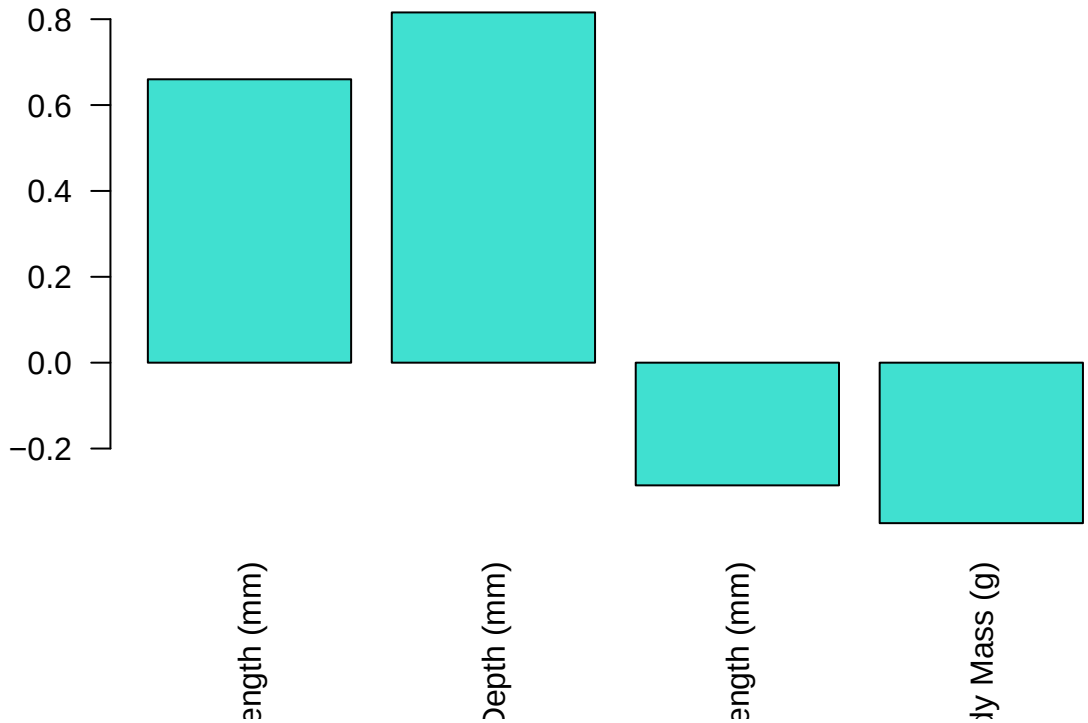
Per our above analysis on the optimal number of clusters, as well as the plot showing the relationship between `Flipper Length (mm)` and `Body Mass (g)`, we will now use 2 clusters as the "guess" when analyzing our centers.

```
y = 3
fit = kmeans(scale(pca_penguins2), centers=y, nstart=1000)
groups = fit$cluster
barplot(table(groups), col="turquoise")
```

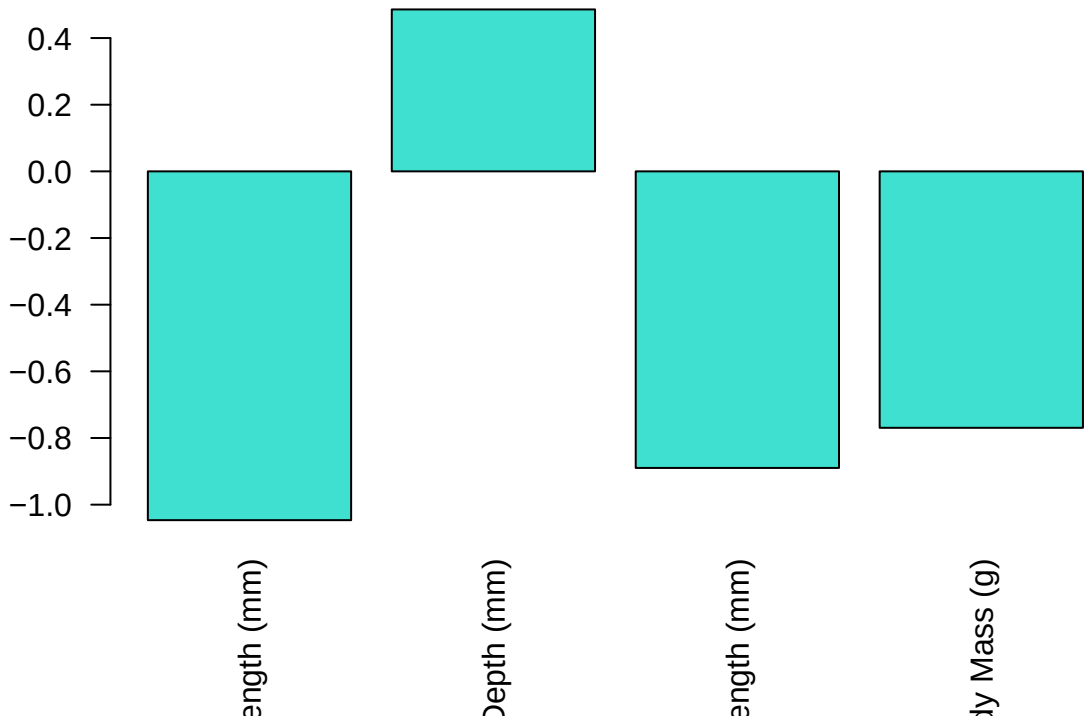


Interpret the centers:

```
centers=fit$centers

barplot(centers[1,], las=2, col="turquoise")
```




```
barplot(centers[2,], las=2, col="turquoise")
```



Now, the differences between both clusters are much more distinct. `Culmen Depth (mm)` is notably higher in the first cluster, whereas the other 3 key continuous variables (`Culmen Length (mm)`, `Flipper Length (mm)`, and `Body Mass (g)`), are noticeably higher in the second one.

To conclude, we originally thought that there would be three clusters, as species was a clear defining categorical variable that can be used to distinguish between penguins. However, after looking at the plots of the principal components, as well as the optimal cluster analysis, we can conclude that there are 2 clusters in our data.

Conclusion

It is essential to outline some of the main distinguishing factors between the 3 species of penguins that we are analyzing. Upon looking at the plot in the PCA which displays the relationship between `Flipper Length (mm)` and `Body Mass (g)` with respect to species, we can see that the Gentoo penguins are in its own group in the top right, while the Adelie and Chinstrap penguins are clumped together, as they are similiar in size. In fact, Gentoo penguins are the third largest species of penguins, while the Chinstrap and Adelie penguins are 5th and 6th largest, respectively.

Adelie and Chinstrap penguins are differentiated by the shape of their culmens. Chinstraps have similiar culmen depths as Adelies, but longer culmen lengths. However, Chinstraps have similiar culmen lengths as Gentoos, but longer culmen depths. According to the Factor Analysis, the `Culmen Length (mm)` variable was not useful, but that was proven false in PCA. Therefore, we can conclude that all four of the numeric variables that we explored are useful in predicting the species of a penguin. Our unsupervised learning techniques proved that we can reduced these four variables into the two-dimensions of penguin size and culmen shape to complete the analysis.

Further directions would include creating a model to predict a penguins species using these numeric variables, as well as exploring the `Island` and `Sex` variables of our data set in relation to the numeric variables.

Sources

1) penguins_raw data set
2) Kaiser Rule for PCA
3) PA and ML Methods for Factor Analysis