Homework #1

Matt Quintiere and Rohit Gunda

2023-10-16

Setup

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.1.1 --
## v broom        1.0.5     v rsample      1.2.0
## v dials        1.2.0     v tune         1.1.2
## v infer        1.0.5     v workflows    1.1.3
## v modeldata    1.2.0     v workflowsets 1.0.1
## v parsnip      1.1.1     v yardstick    1.2.0
## v recipes      1.0.8
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(dplyr)
library(knitr)
library(palmerpenguins)
```

```
##
## Attaching package: 'palmerpenguins'
##
## The following object is masked from 'package:modeldata':
##
##     penguins
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(factoextra)
```

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

```
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:scales':
##
##     alpha, rescale
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
view(penguins_raw)
```

Introduction

```
penguins_raw <- penguins_raw
```

Introduction of data: 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 (bill) Length:** a number denoting the length of the dorsal ridge of a bird's bill (millimeters)
- **Culmen (bill) Depth:** a number denoting the depth of the dorsal ridge of a bird's bill (millimeters)
- **Flipper Length:** an integer denoting the length penguin flipper (millimeters)
- **Body Mass:** 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 dataset for 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. As a result, our main research question is: Is there a relationship between the certain categorical variables (such as species, island, and sex) and the size of certain aspects of the penguin (such as bill length/depth, body mass, and flipper length)?

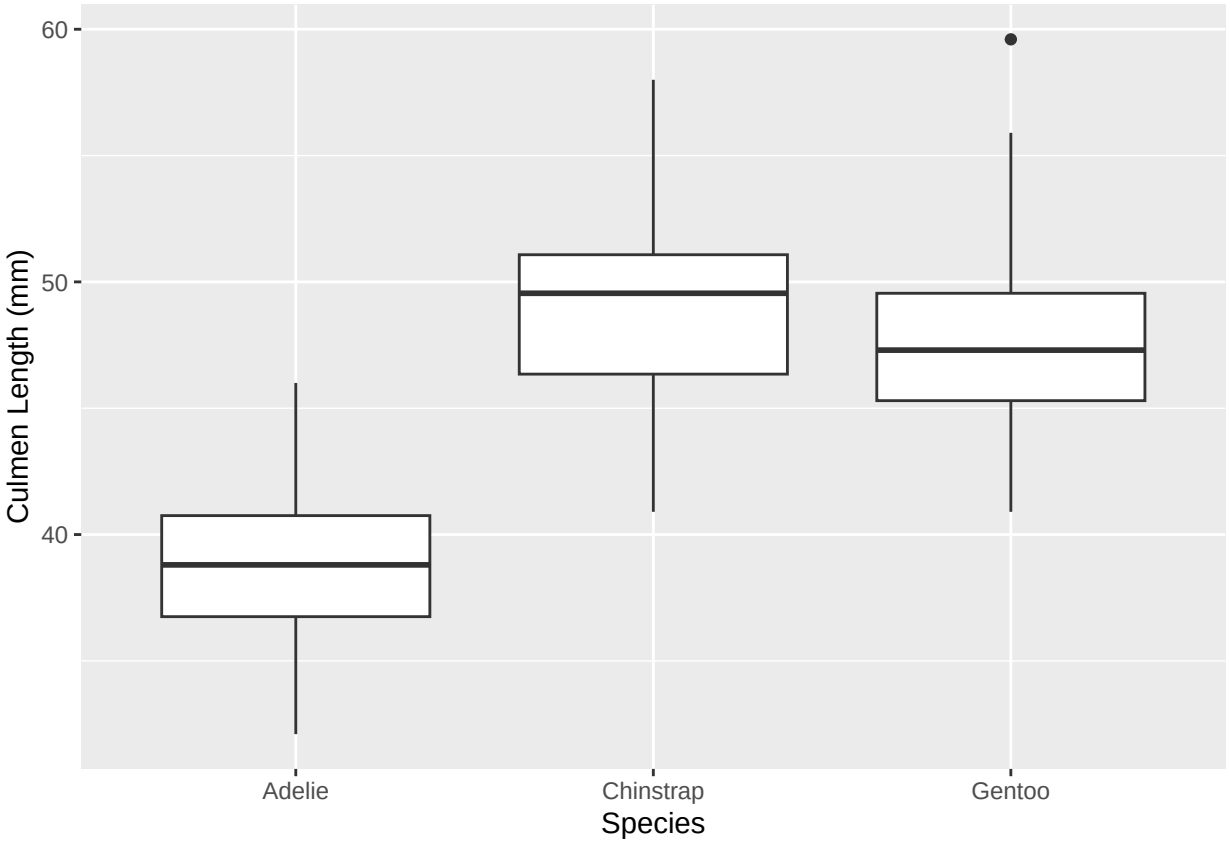
```
penguins_mod <- penguins_raw |>
  select(-studyName, -`Sample Number`, -Region, -`Individual ID`, -Stage,
         -`Clutch Completion`, -`Date Egg`, -`Delta 15 N (o/oo)`,
         -`Delta 13 C (o/oo)`, -Comments)
```

Here, we have only kept the variables that would be useful in our analysis.

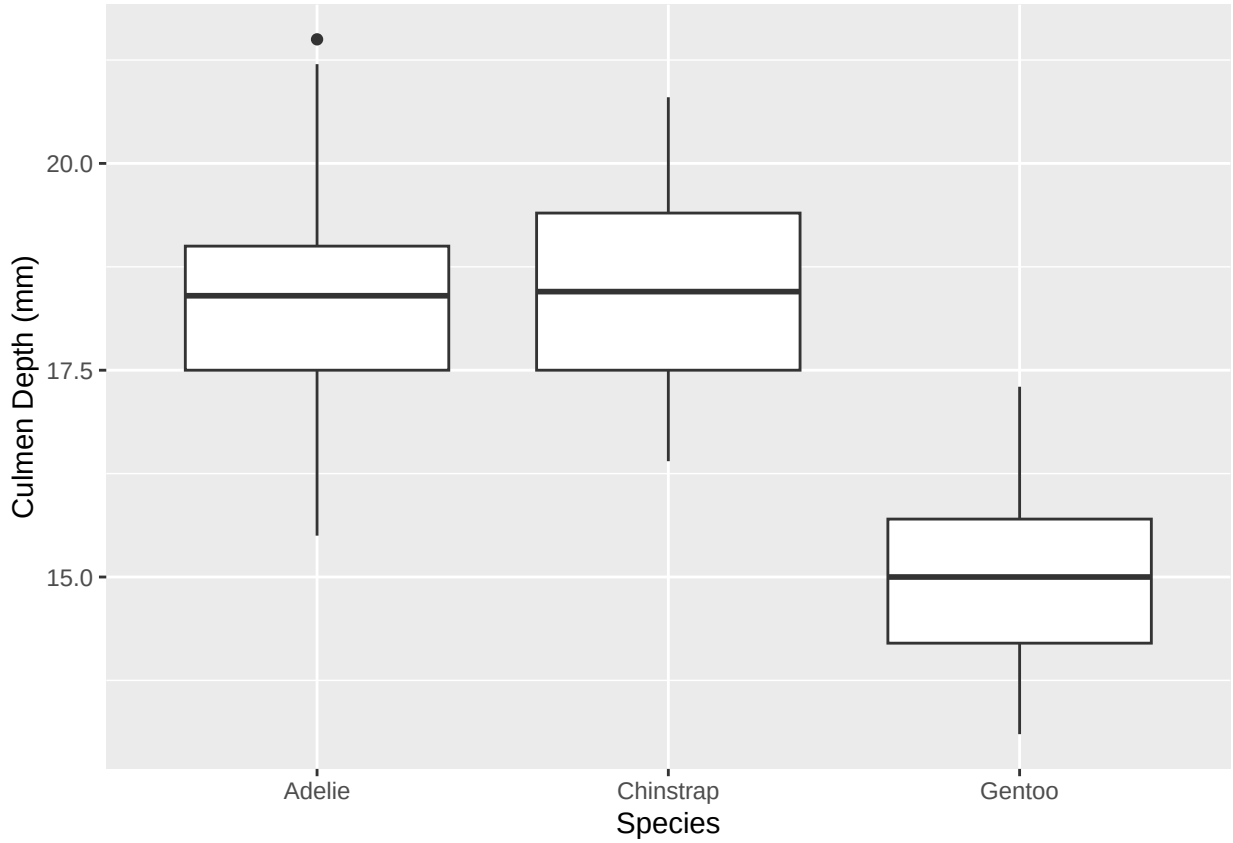
```
#Removing NAs in the dataset
penguins2 <- penguins_mod[rowSums(is.na(penguins)) < 2, ]
```

```
penguins2 <- penguins2 |>
  mutate(
    Species = case_when(
      Species == "Adelie Penguin (Pygoscelis adeliae)" ~ "Adelie",
      Species == "Gentoo penguin (Pygoscelis papua)" ~ "Gentoo",
      Species == "Chinstrap penguin (Pygoscelis antarctica)" ~ "Chinstrap"
    )
  )

ggplot(penguins2,aes(x=Species, y=`Culmen Length (mm)`)) +
  geom_boxplot()
```




```
ggplot(penguins2,aes(x=Species, y=`Culmen Depth (mm)`)) +
  geom_boxplot()
```




```
penguins2 <- penguins2 %>%
  mutate(outlier = ifelse(`Culmen Length (mm)` > 58, FALSE, TRUE))
penguins2 <- penguins2 %>%
  filter(outlier == TRUE)
penguins2 <- penguins2 %>%
  mutate(outlier2 = ifelse(`Culmen Depth (mm)` > 21.2, FALSE, TRUE))
penguins2 <- penguins2 %>%
  filter(outlier2 == TRUE)
penguins2 <- penguins2 %>%
  select(Species, Island, `Culmen Length (mm)`, `Culmen Depth (mm)`,
         `Flipper Length (mm)`, `Body Mass (g)`, Sex)
```

PCA

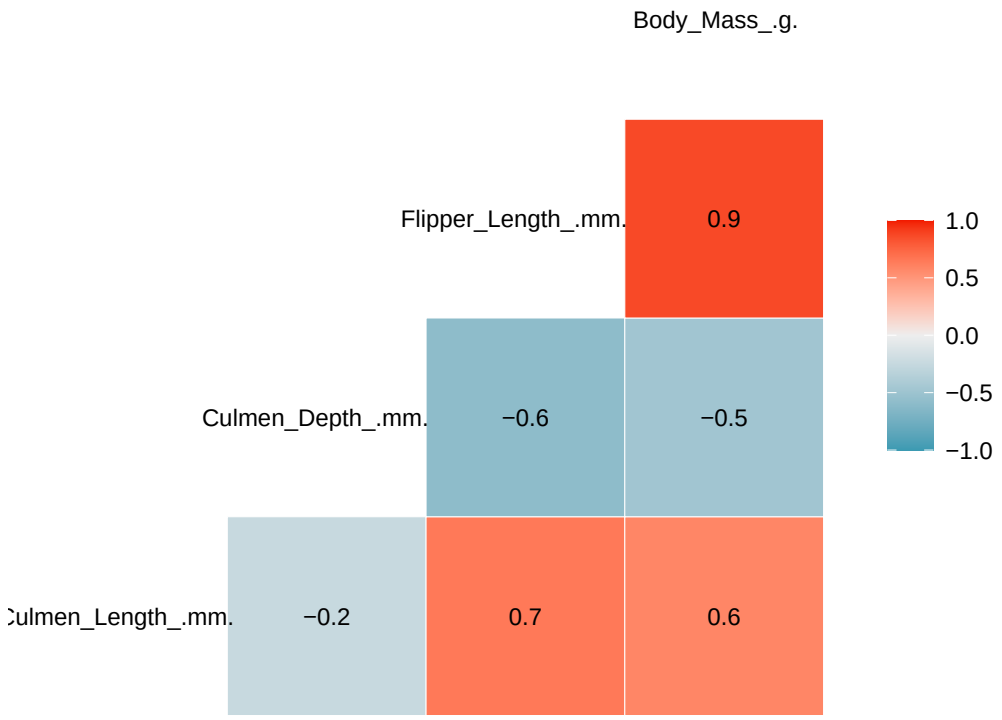
To explore the principal components of our data set to explain the variability, we have to isolate the numeric variables in the dataset, which are `Culmen Length (mm)`, `Culmen Depth (mm)`, `Flipper Length (mm)length_mm`, `Body Mass (g)`. The reason we can only use numeric variables is because the PCA relies linear algebra calculations which can only be used with numeric data.

```
pca_penguins2 <- penguins2[,3:6]

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

Let's visualize how correlated these variables are by doing bivariate analysis:

```
ggcorr(penguins2[,3:6], label = TRUE, label_size = 3, hjust = 0.55, size =3)
```



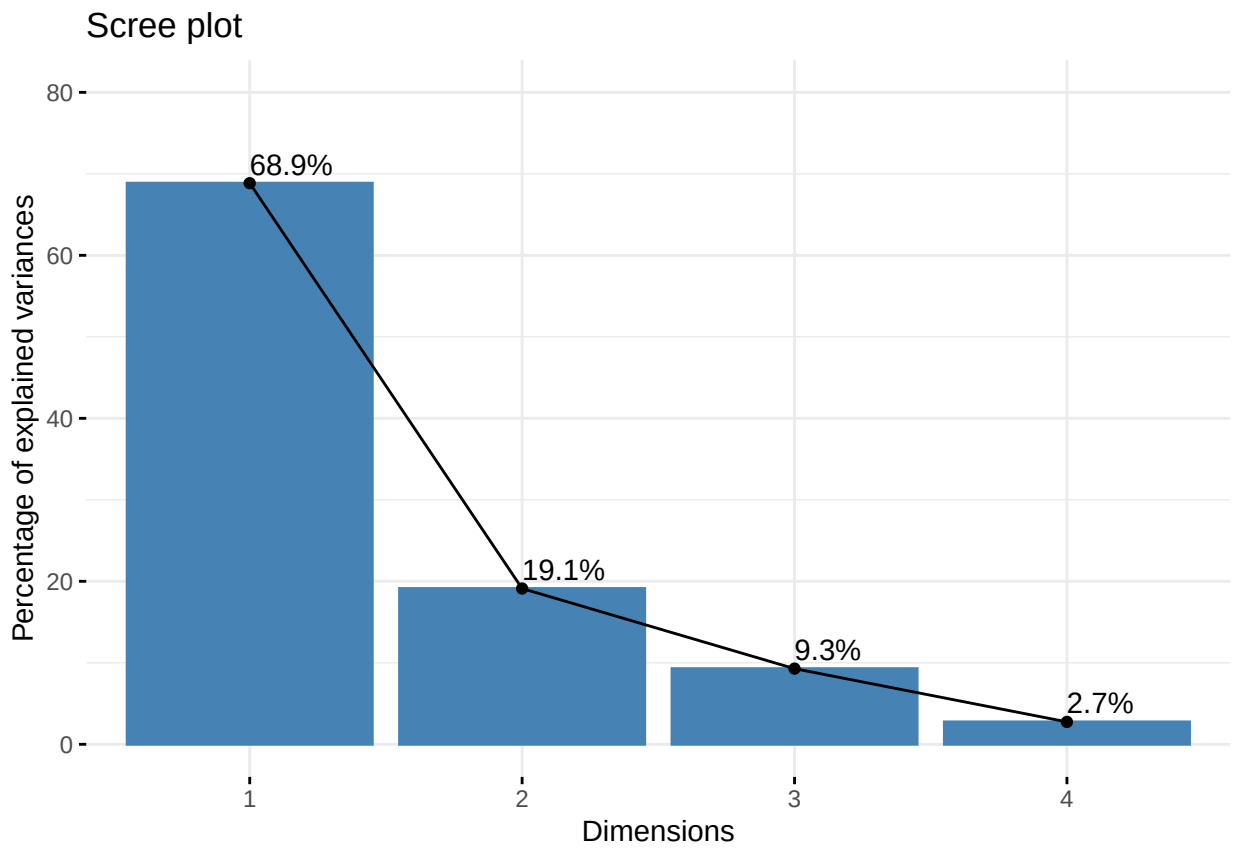
From the correlation matrix, we can see that `Body Mass (g)` is very correlated with `Flipper Length (mm)`. Also `Culment Length (mm)` has a somewhat strong correlation with `Flipper Length (mm)` and `Body Mass (g)`, while `Bill Depth (mm)` has a negative correlation with all of the variables and does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

```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4
## Standard deviation     1.6596 0.8744 0.60936 0.3316
## Proportion of Variance 0.6885 0.1911 0.09283 0.0275
## Cumulative Proportion  0.6885 0.8797 0.97250 1.0000
```

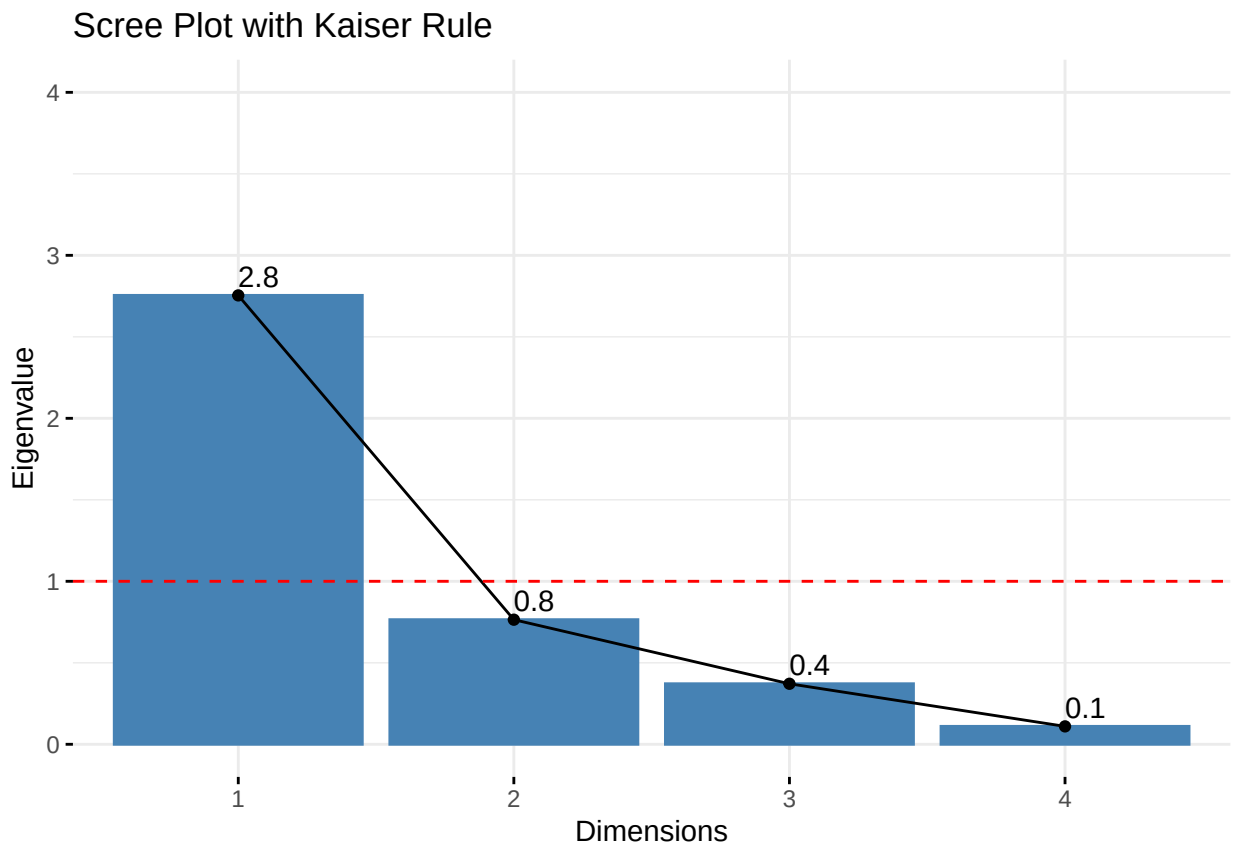
Running the PCA calculations out using the prcomp tool, we see that the first two principal components are responsible for about 88% percent of the data. Let's create a screeplot to visualize this:

```
fviz_eig(pca, addlabels = TRUE,
         ylim = c(0,80))
```



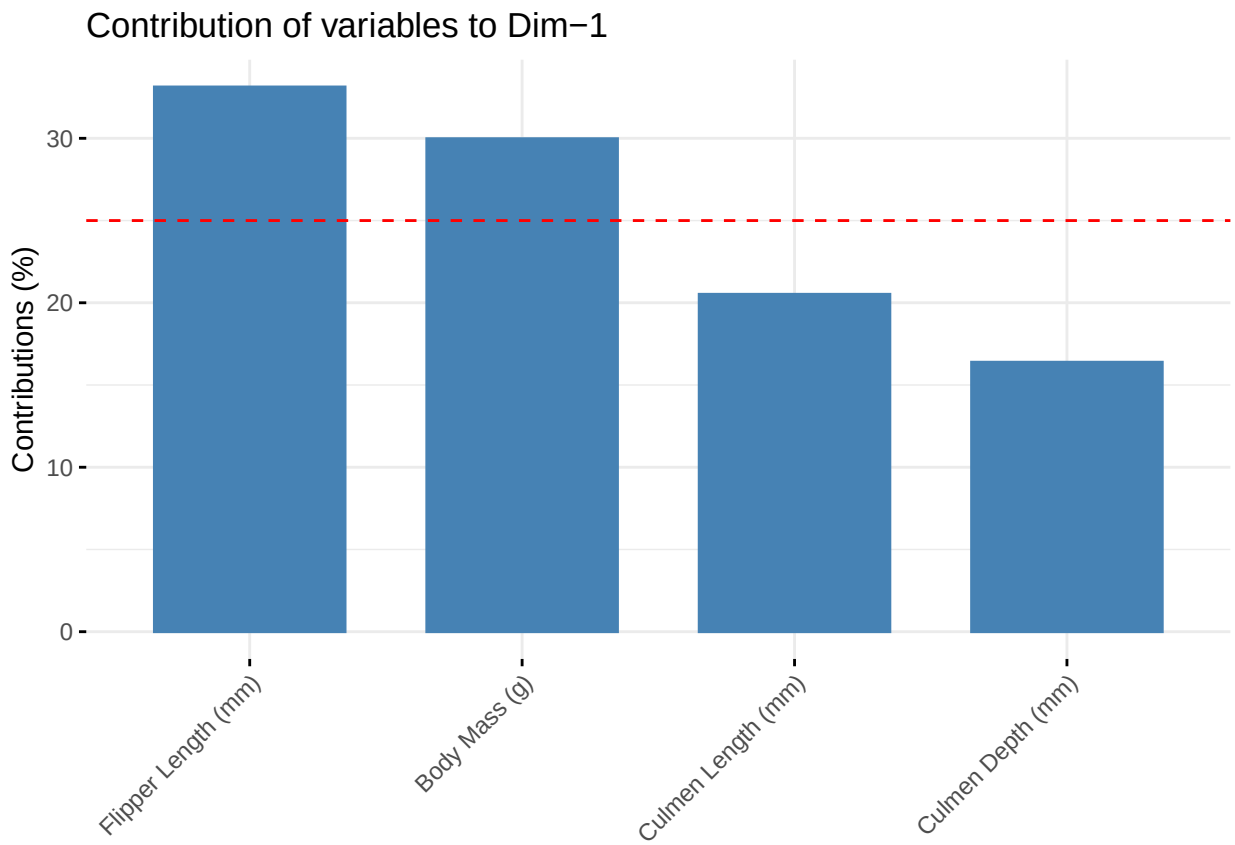
Now let's apply the Kaiser rule to select the number of components, we will use in the PCA:

```
fviz_eig(pca,
         addlabels = TRUE,
         ylim = c(0,4),
         choice="eigenvalue",
         main="Scree Plot with Kaiser Rule") +
         geom_hline(yintercept=1,
         linetype="dashed",
         color = "red")
```



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. Now let's look at what variables contribute the most to our first principal component:

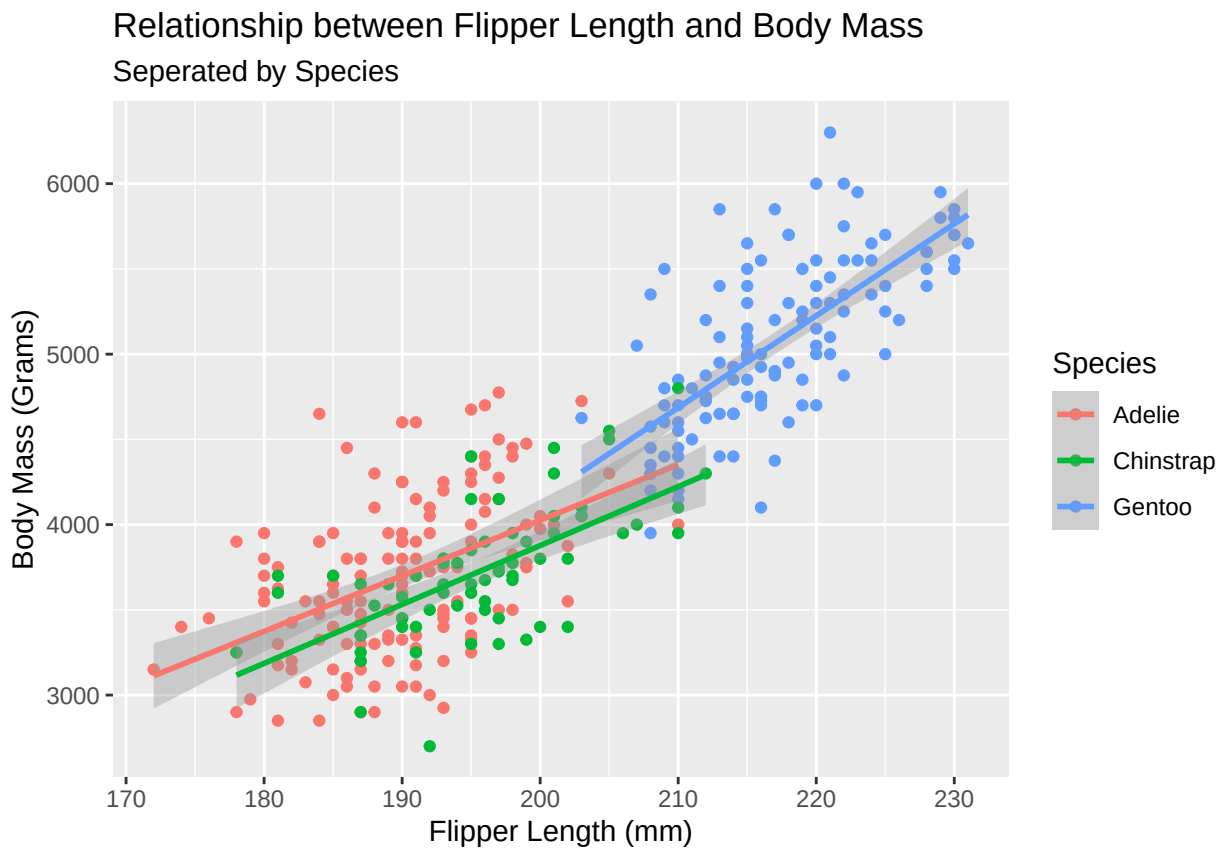
```
fviz_contrib(pca, choice = "var", axes = 1)
```



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`:

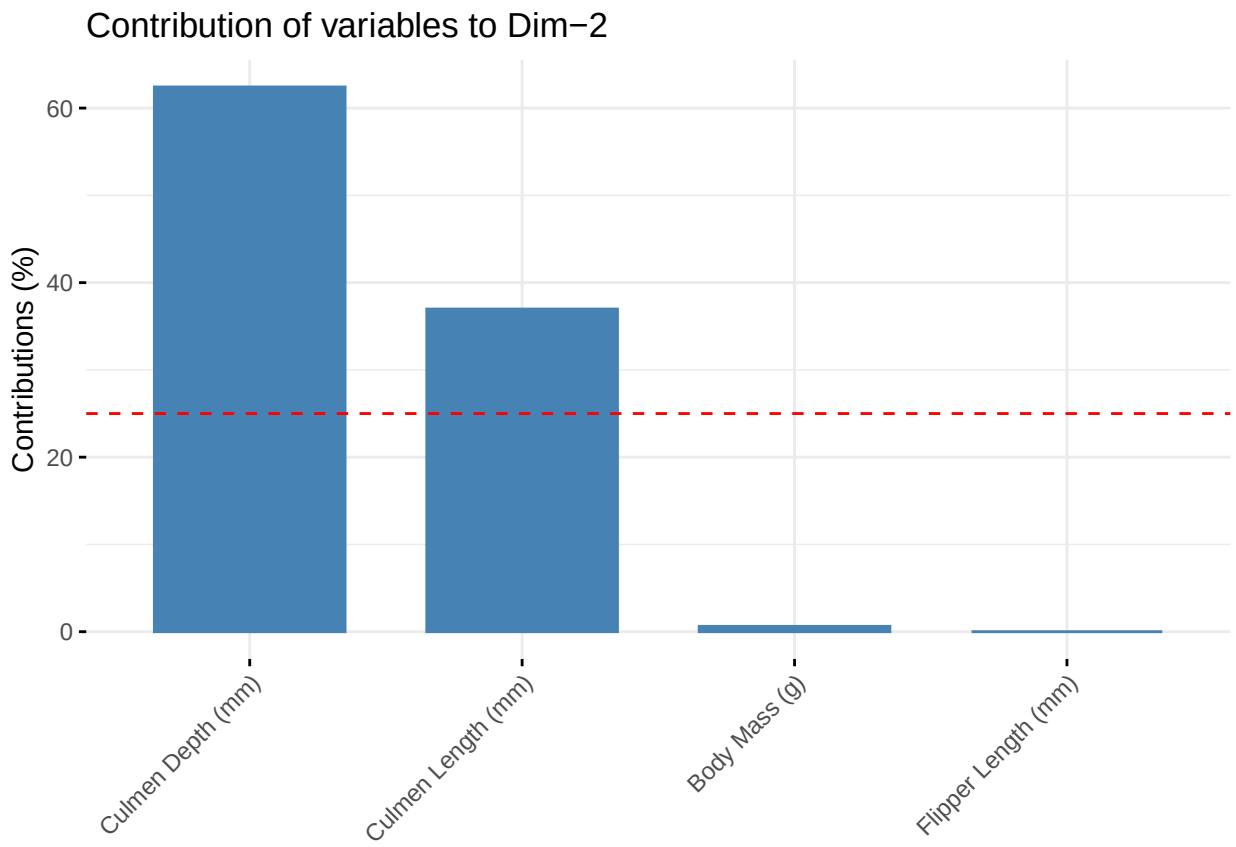
```
penguins2 %>%
  ggplot(aes(x = `Flipper Length (mm)`, y = `Body Mass (g)`, color = Species)) +
  geom_point() +
  labs(y = "Body Mass (Grams)",
       x = "Flipper Length (mm)",
       title = "Relationship between Flipper Length and Body Mass",
       subtitle = "Seperated by Species",
       color = "Species") +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



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 the strong postive correlation trend that `Body Mass (g)` and `Flipper Length (mm)` have as shown in the correlation plot and for each of the species. Let's circle back to principal component two, so we can find a way to differentiate the Adelie and Chinstrap species:

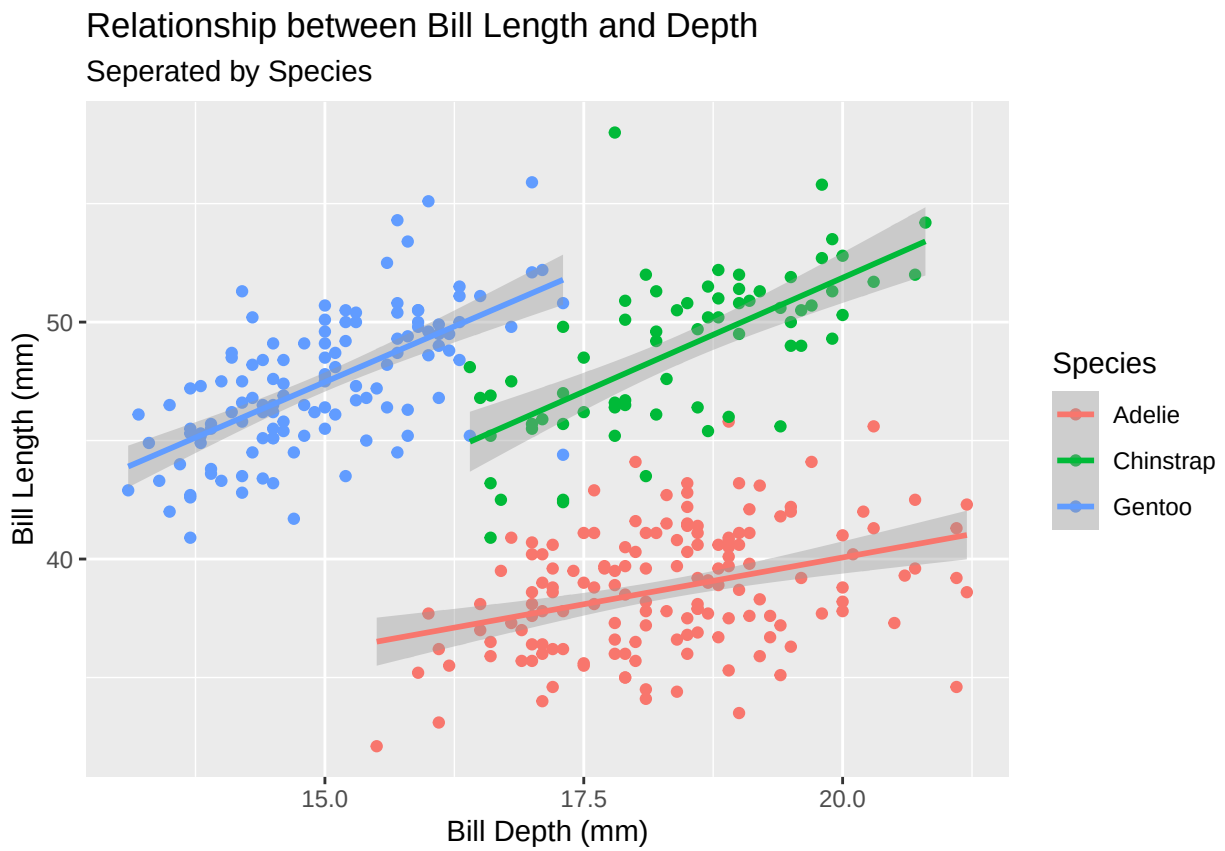
```
fviz_contrib(pca, choice = "var", axes = 2)
```



In principal component two, 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:

```
penguins2 %>%
  ggplot(aes(x = `Culmen Depth (mm)`, y = `Culmen Length (mm)`, color = Species)) +
  geom_point() +
  labs(x = "Bill Depth (mm)",
       y = "Bill Length (mm)",
       title = "Relationship between Bill Length and Depth",
       subtitle = "Seperated by Species",
       color = "Species") +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The Principal Component 2 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 bill depths as Adelie penguins, but much larger lengths. Also, despite being smaller penguins in size, the Chinstrap and Adelie penguins have a larger bill depths than Gentoo Penguins. Gentoo Penguins; however, have a much larger bill length than Adelie penguins. Also, it is interesting to note that `Culmen Depth (mm)` and `Culmen Length (mm)` are not strongly correlated based on our correlation matrix, but their relationship with respect species allows us to further differentiate the penguins species.

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] 0.08600057
```

Our determinant is postive, 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 and only be using a total of 2 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 culmative 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)
```

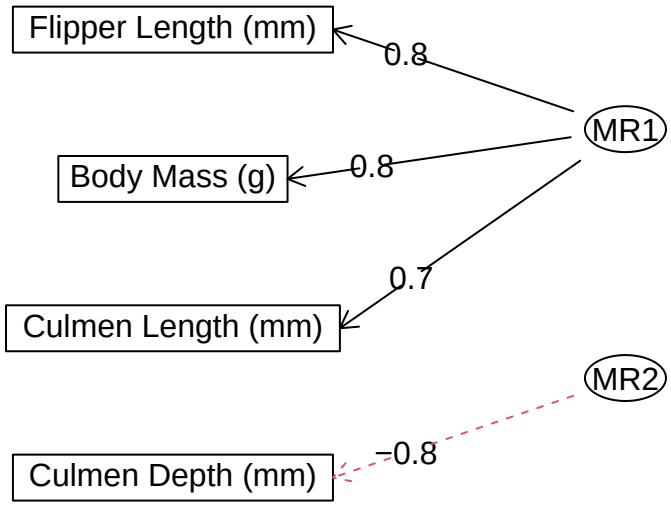
The PA and ML methods both produce a cumulative variance of 0.72, so either method will work in this case. We will use the ML method going ahead. So let's view and interpret the results:

```
ml
```

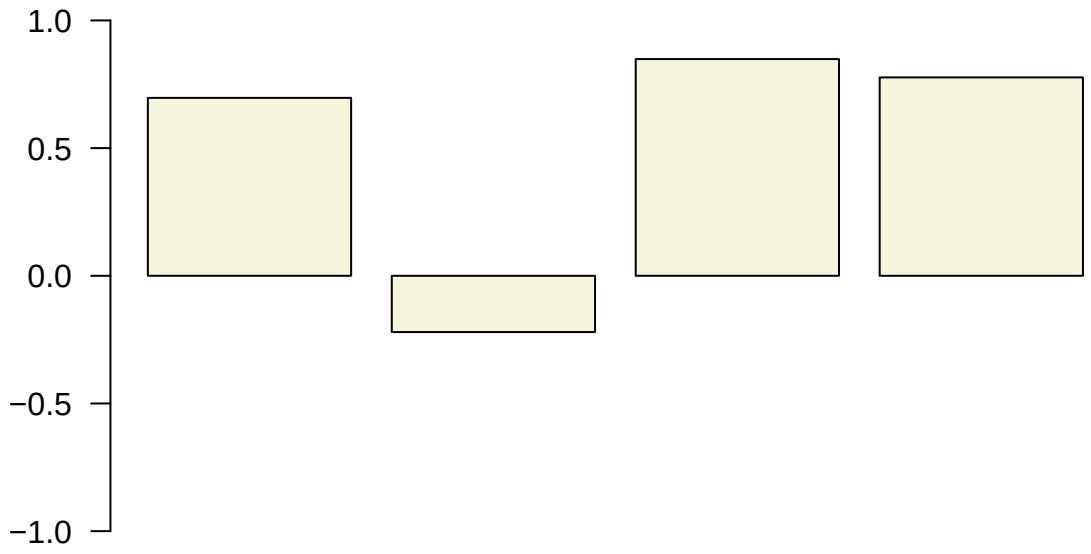
```
## Factor Analysis using method =  minres
## Call: fa(r = pca_penguins2, nfactors = 2, rotate = "varimax", fa = "ml",
##     resdiuals = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                        MR1   MR2   h2    u2 com
## Culmen Length (mm)    0.70  0.12 0.50 0.501 1.1
## Culmen Depth (mm)    -0.22 -0.77 0.64 0.365 1.2
## Flipper Length (mm)   0.85  0.52 0.99 0.005 1.7
## Body Mass (g)         0.78  0.40 0.76 0.235 1.5
##
##                         MR1  MR2
## SS loadings            1.86 1.04
## Proportion Var         0.46 0.26
## Cumulative Var         0.46 0.72
## Proportion Explained   0.64 0.36
## Cumulative Proportion 0.64 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  6  with the objective function =  2.45 with Chi Square =  826.39
## 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  340 with the empirical chi square  0  with prob <  NA
## The total n.obs was  340  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
```

```
fa.diagram(ml)
```

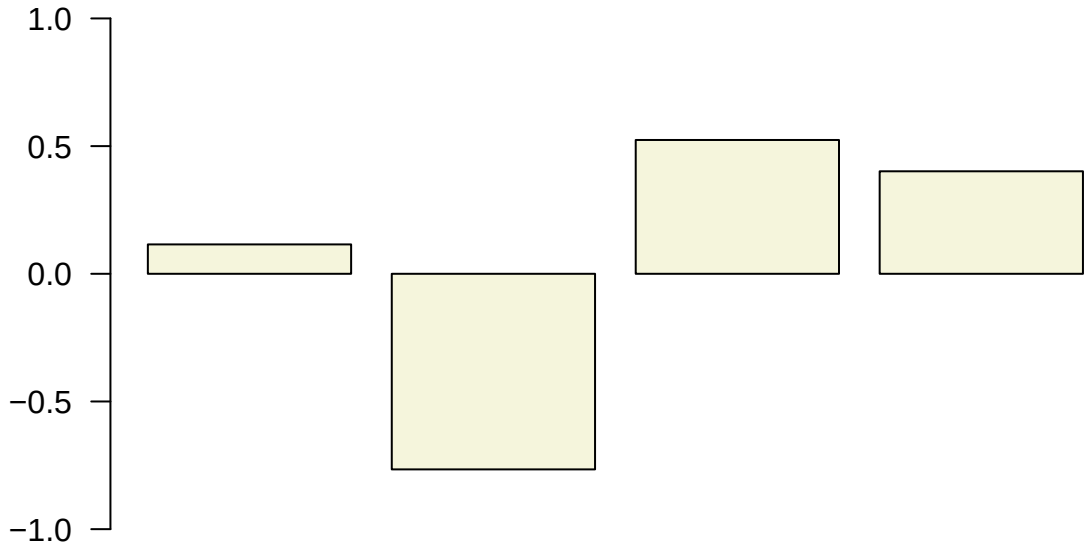
Factor Analysis




```
barplot(ml$loadings[,1], las=2, names =F,col="beige", ylim = c(-1, 1))
```




```
barplot(ml$loadings[,2], las=2, names = F, col="beige", ylim = c(-1, 1))
```



From these charts and tables, 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 loading. 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 analysis 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 or island of a given penguin.

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 occured because `Culmen Depth (mm)` is not that strongly correlated with the other numeric variables and the only prinicpal component that has an eigenvalue greater than 1 is the first one, which does not include `Culmen Depth (mm)` and `Culmen Depth (mm)` is the main variable in the second loading factor.

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

Cluster Analysis

Load important clustering libraries:

```r
library(factoextra)
library(cluster)
library(mclust)
```

```
## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.
```

```
##
## Attaching package: 'mclust'
```

```
## The following object is masked from 'package:psych':
##
##     sim
```

```
## The following object is masked from 'package:purrr':
##
##     map
```