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        1.0.4     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()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
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
```

```
view(penguins)
```

Introduction

```
penguins_raw
```

```
## # A tibble: 344 x 17
##    studyName `Sample Number` Species         Region Island Stage `Individual ID`
##    <chr>               <dbl> <chr>           <chr>  <chr>  <chr> <chr>
##  1 PAL0708                 1 Adelie Penguin~ Anvers Torge~ Adul~ N1A1
##  2 PAL0708                 2 Adelie Penguin~ Anvers Torge~ Adul~ N1A2
##  3 PAL0708                 3 Adelie Penguin~ Anvers Torge~ Adul~ N2A1
##  4 PAL0708                 4 Adelie Penguin~ Anvers Torge~ Adul~ N2A2
##  5 PAL0708                 5 Adelie Penguin~ Anvers Torge~ Adul~ N3A1
##  6 PAL0708                 6 Adelie Penguin~ Anvers Torge~ Adul~ N3A2
##  7 PAL0708                 7 Adelie Penguin~ Anvers Torge~ Adul~ N4A1
##  8 PAL0708                 8 Adelie Penguin~ Anvers Torge~ Adul~ N4A2
##  9 PAL0708                 9 Adelie Penguin~ Anvers Torge~ Adul~ N5A1
## 10 PAL0708                10 Adelie Penguin~ Anvers Torge~ Adul~ N5A2
## # i 334 more rows
## # i 10 more variables: `Clutch Completion` <chr>, `Date Egg` <date>,
## #   `Culmen Length (mm)` <dbl>, `Culmen Depth (mm)` <dbl>,
## #   `Flipper Length (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>,
## #   `Delta 15 N (o/oo)` <dbl>, `Delta 13 C (o/oo)` <dbl>, Comments <chr>
```

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:

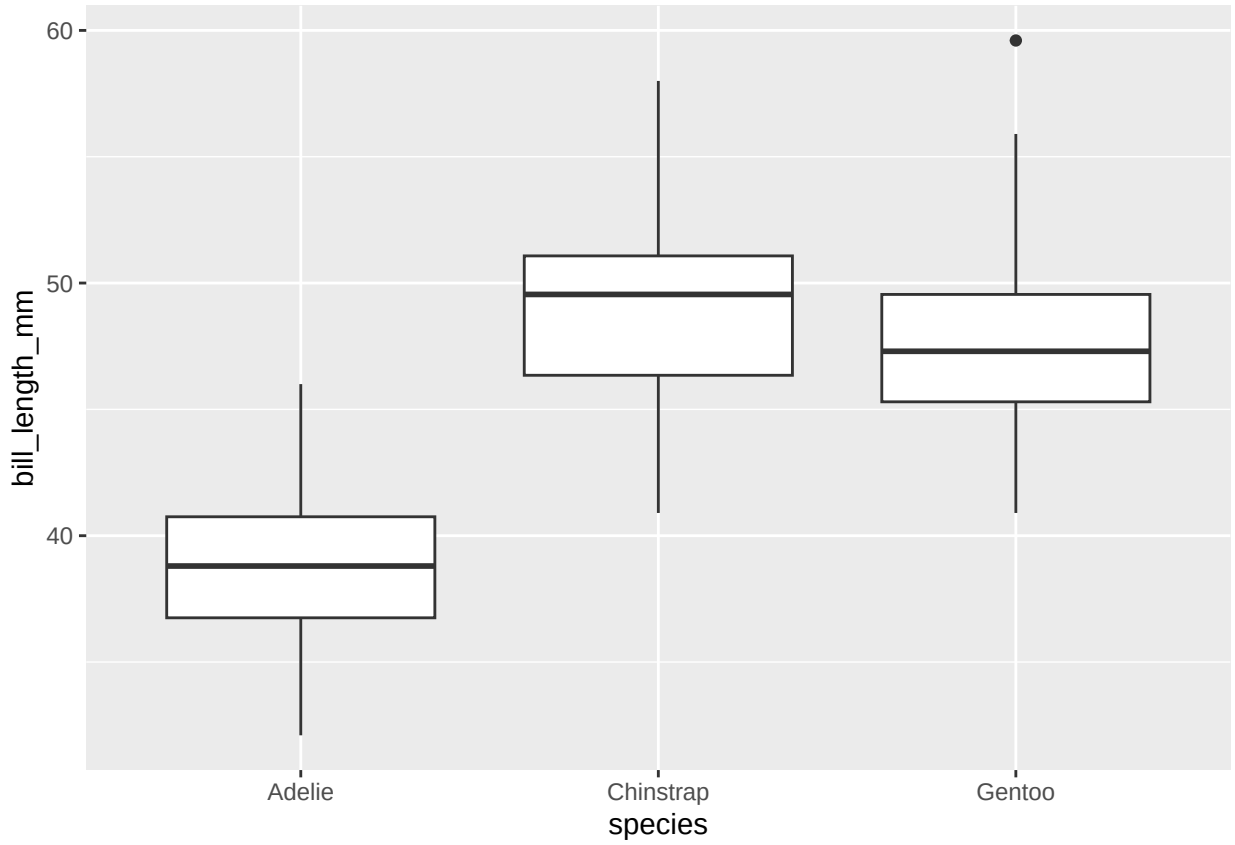
- **studyName:** Sampling expedition from which data were collected, generated, etc.
- **Sample Number:** an integer denoting the continuous numbering sequence for each sample
- **Species:** a character string denoting the penguin species
- **Region:** a character string denoting the region of Palmer 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

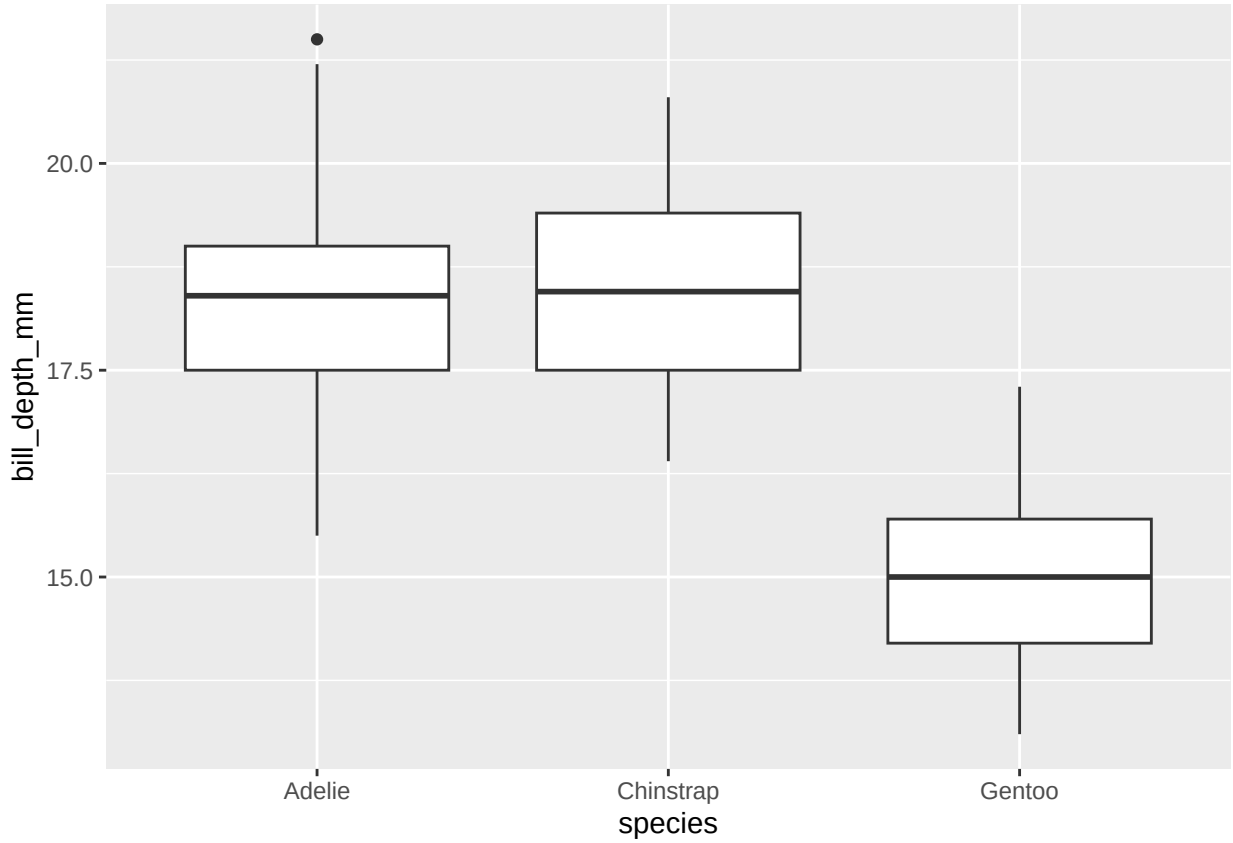
Motivation: We decided to use this dataset for analysis because we were interested to see if there was any variation in some of the key metrics between penguins of different species or penguins of different islands. As a result, our main research question is: Is there a relationship between the certain 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)?

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

```
ggplot(penguins2,aes(x=species, y=bill_length_mm)) +
  geom_boxplot()
```




```
ggplot(penguins2,aes(x=species, y=bill_depth_mm)) +
  geom_boxplot()
```




```
penguins2 <- penguins2 %>%
  mutate(outlier = ifelse(bill_length_mm > 58, FALSE, TRUE))
penguins2 <- penguins2 %>%
  filter(outlier == TRUE)
penguins2 <- penguins2 %>%
  mutate(outlier2 = ifelse(bill_depth_mm > 21.2, FALSE, TRUE))
penguins2 <- penguins2 %>%
  filter(outlier2 == TRUE)
penguins2 <- penguins2 %>%
  select(species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, sex, year)
```

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 `bill_length_mm`, `bill_depth_mm`, `flipper_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.

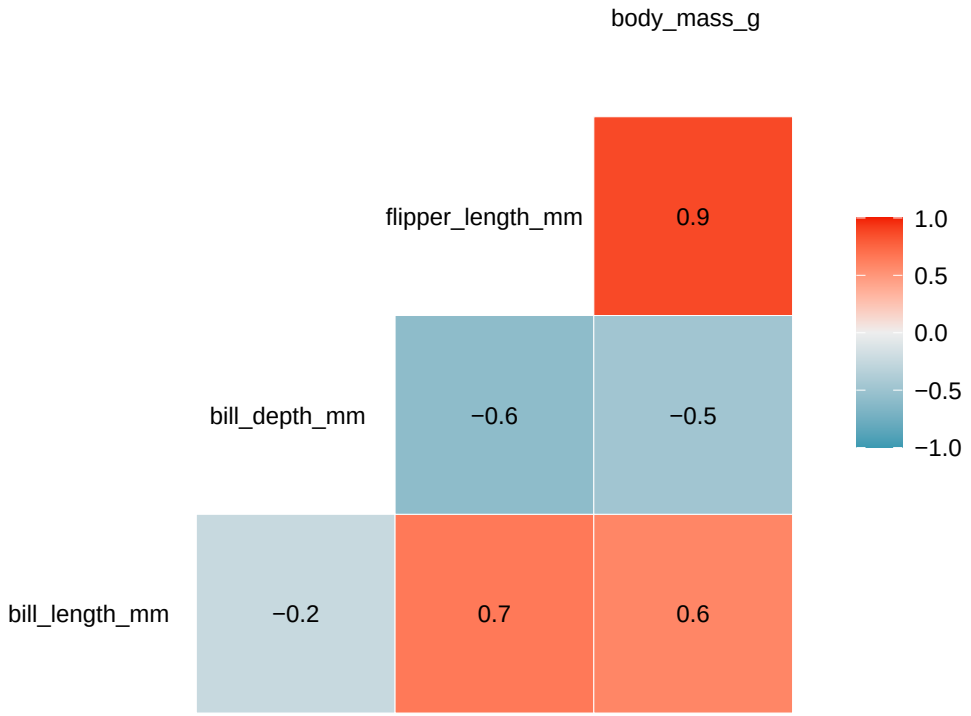
```
pca_penguins2 <- penguins2[,3:6]

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

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

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)
```



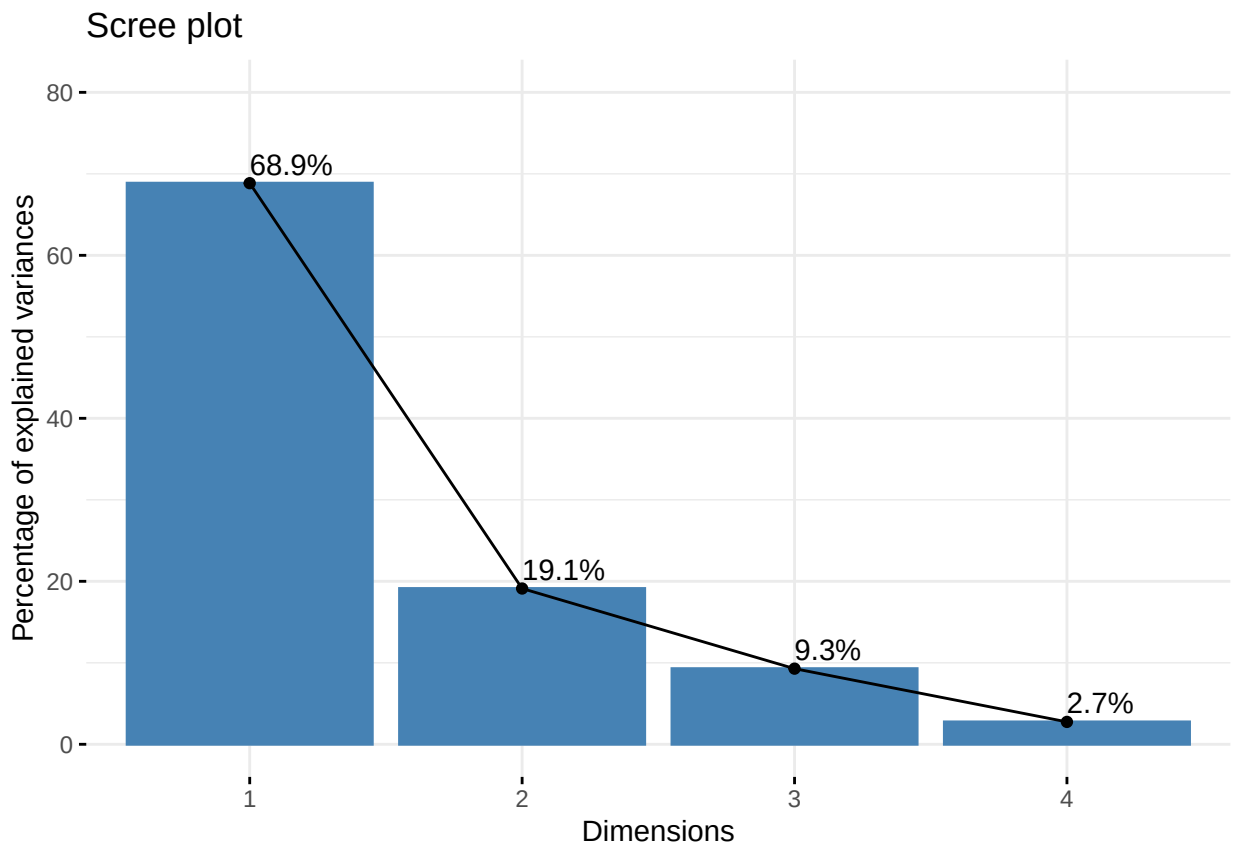
From the correlation matrix, we can see that `body_mass_g` is very correlated with `flipper_length_mm`. Also `bill_length_mm` has a somewhat strong correlation with `flipper_length_mm` and `body_mass_g`, while `bill_depth_mm` has a negative correlation with all of the variables and does not have a strong 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
```

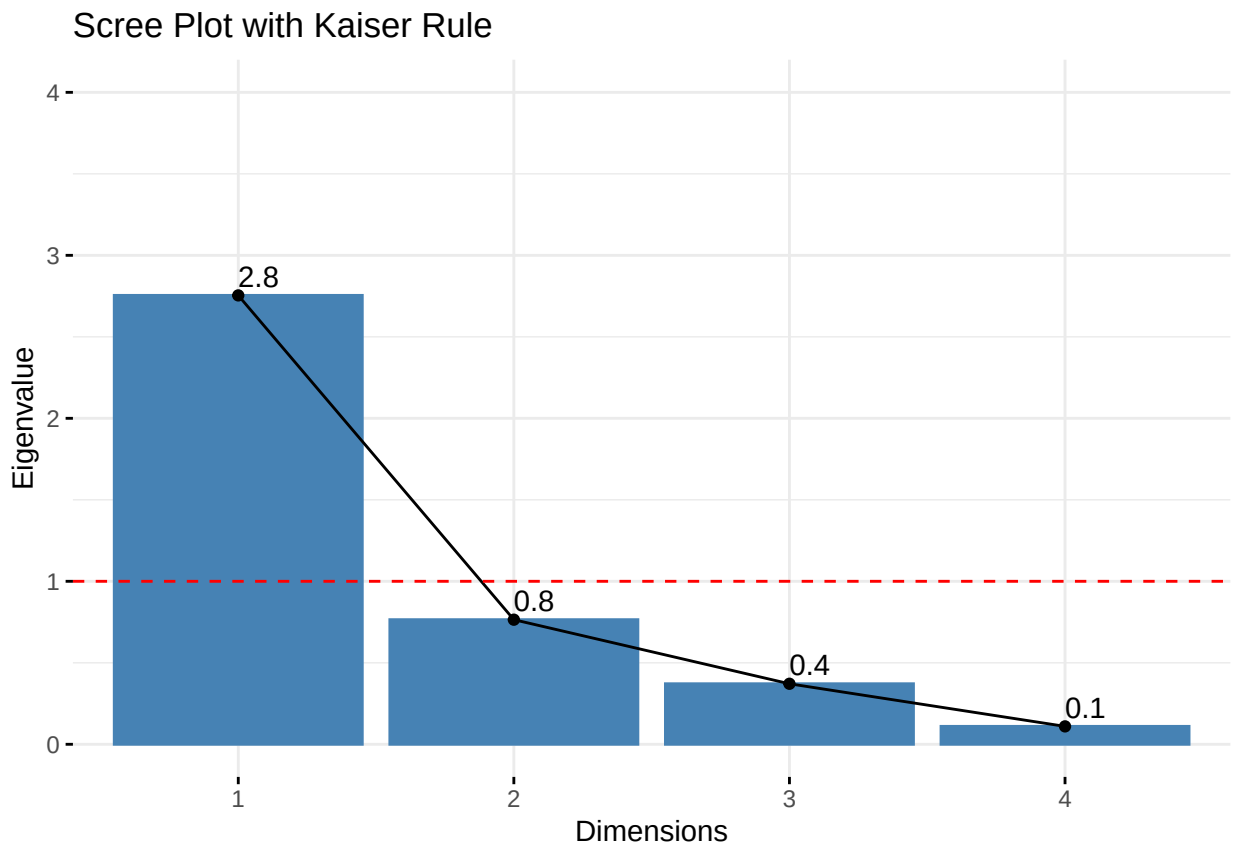
Running the PCA calculations out using the prcomp tool, we see that the first two principal components are responsible for about 88% percent of the data. Let's create a screeplot to visualize this:

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



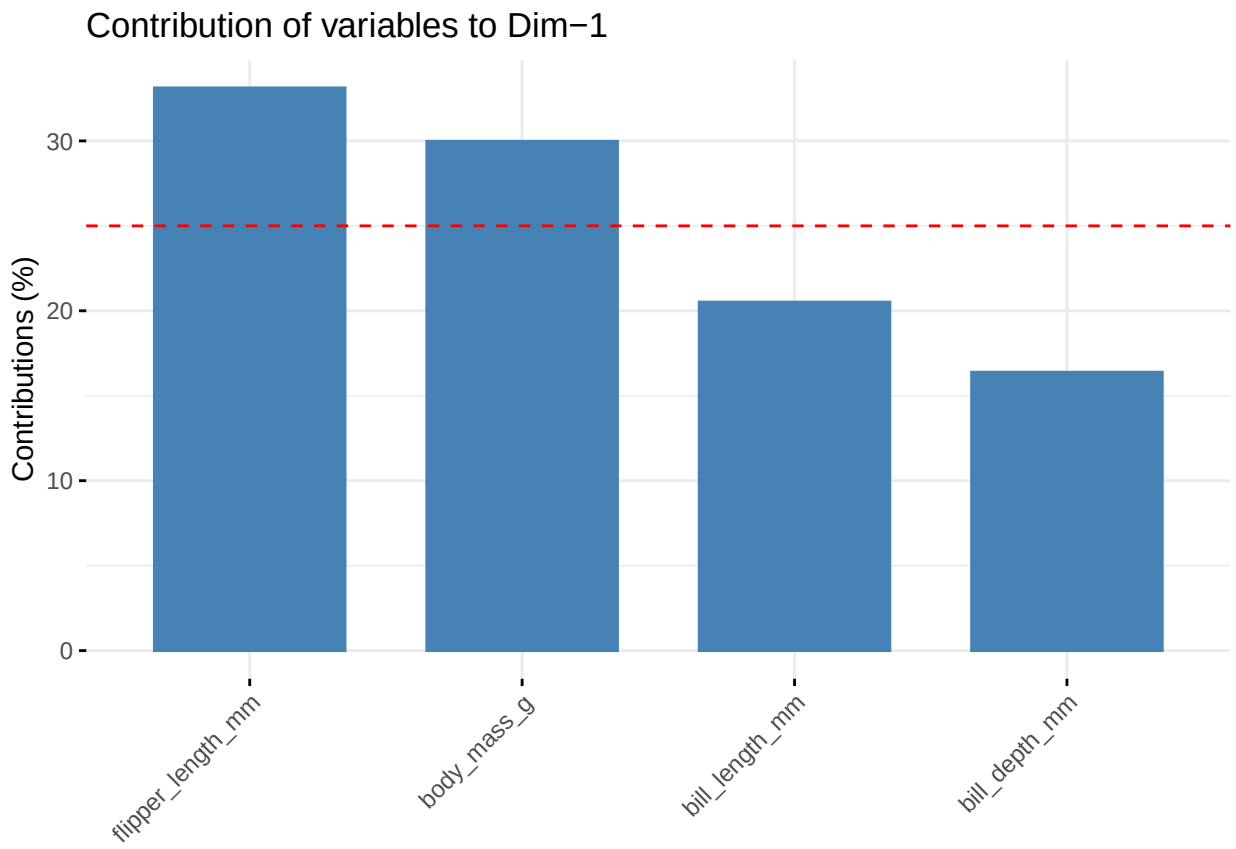
Now let's apply the Kaiser rule to select the number of components, we will use in the PCA:

```
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:

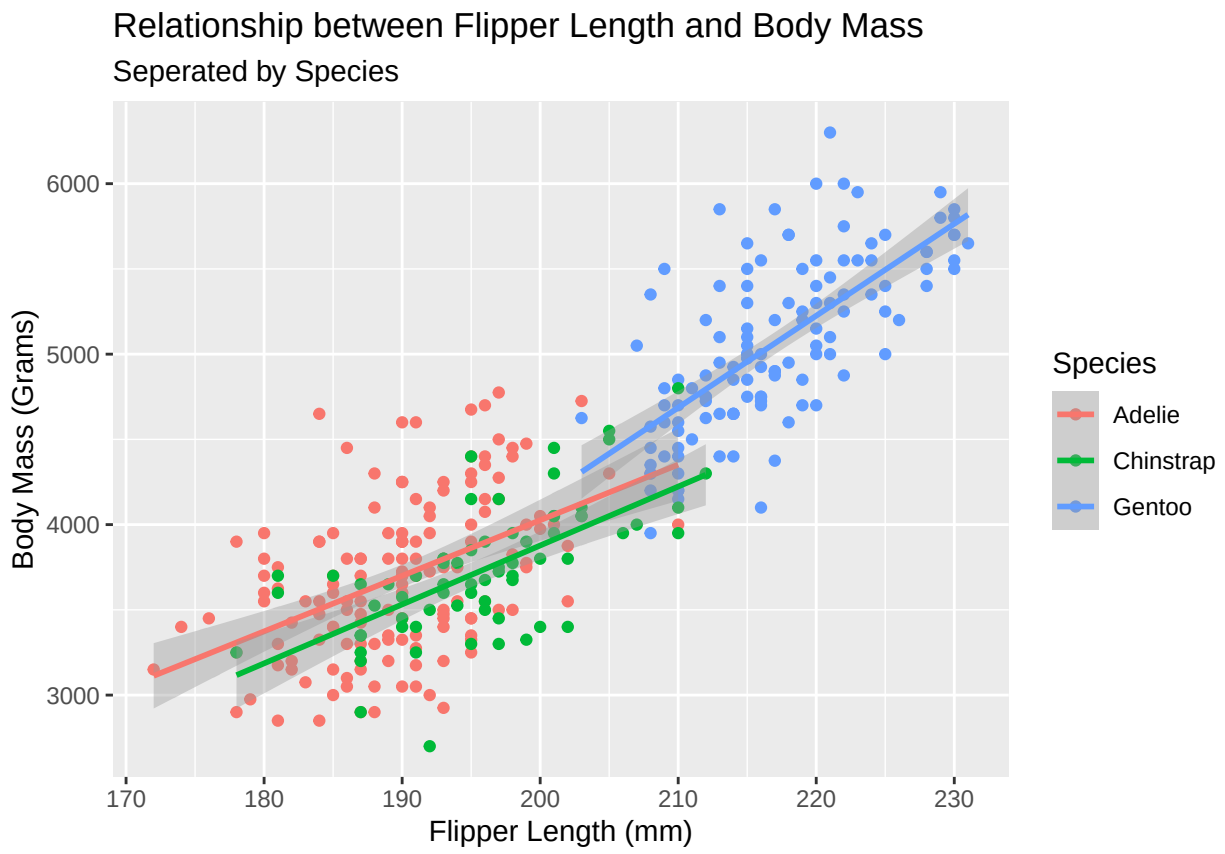
```
fviz_contrib(pca, choice = "var", axes = 1)
```



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`:

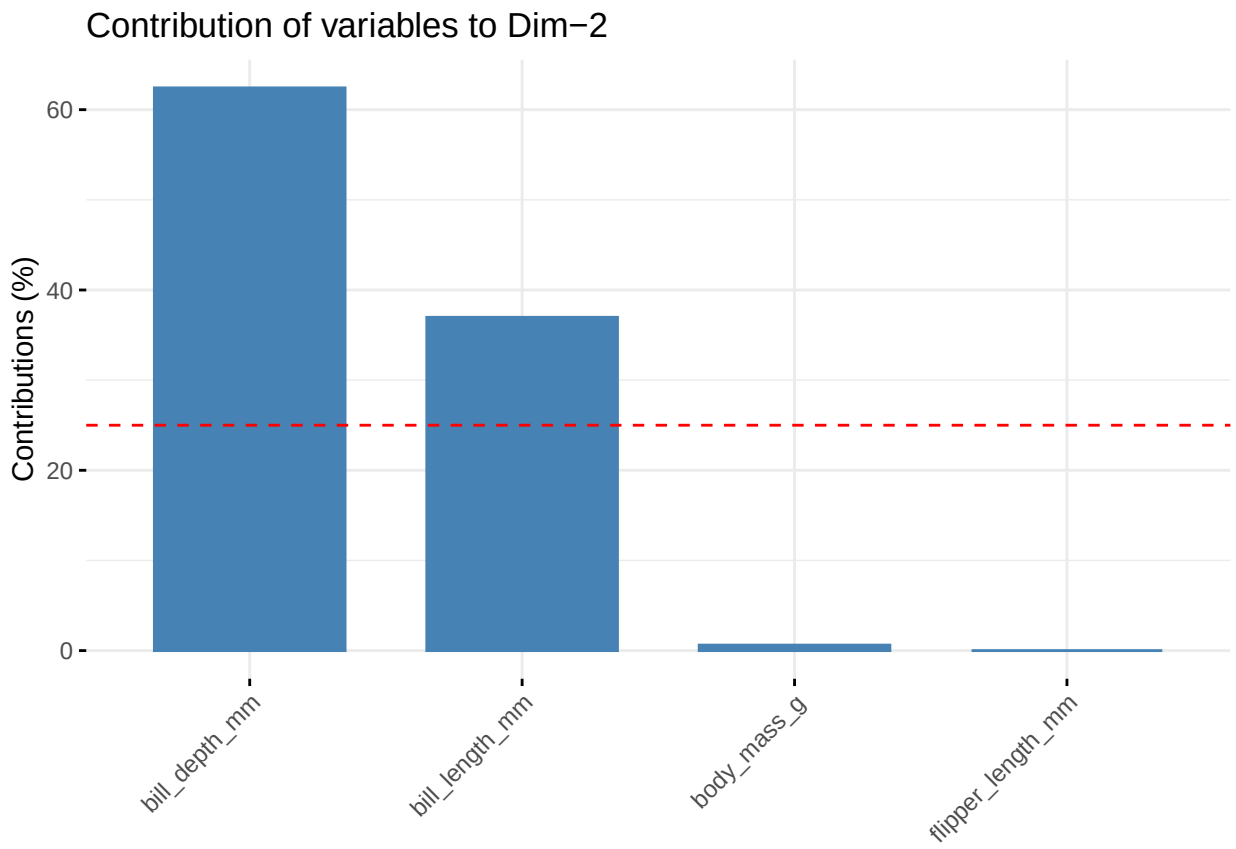
```
penguins2 %>%
  ggplot(aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point() +
  labs(y = "Body Mass (Grams)",
       x = "Flipper Length (mm)",
       title = "Relationship between Flipper Length and Body Mass",
       subtitle = "Seperated by Species",
       color = "Species") +
  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:

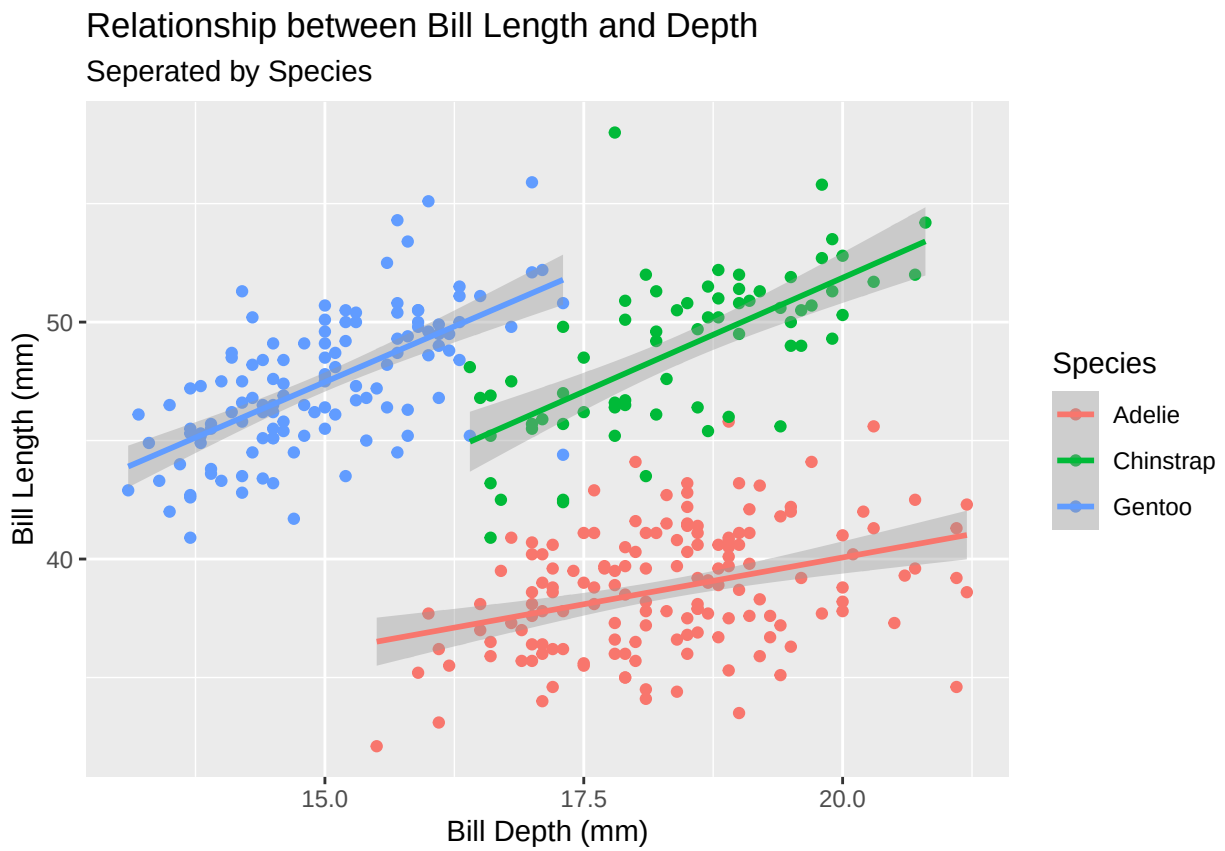
```
fviz_contrib(pca, choice = "var", axes = 2)
```



In principal component two, the two other variables contribute more than expected. Let's visualize the relationship of `bill_length_mm` and `bill_depth_mm` with respect to species:

```
penguins2 %>%
  ggplot(aes(x = bill_depth_mm, y = bill_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 `bill_length_mm` and `bill_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 `bill_depth_mm` and `bill_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.

Cluster Analysis

Factor Analysis