Homework #1

Matt Quintiere and Rohit Gunda

2023-10-16

#Setup

```r
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
```

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

```r
library(knitr)
library(palmerpenguins)
```

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

```r
library(GGally)
```

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

```r
library(factoextra)
```

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

```r
view(penguins)
```

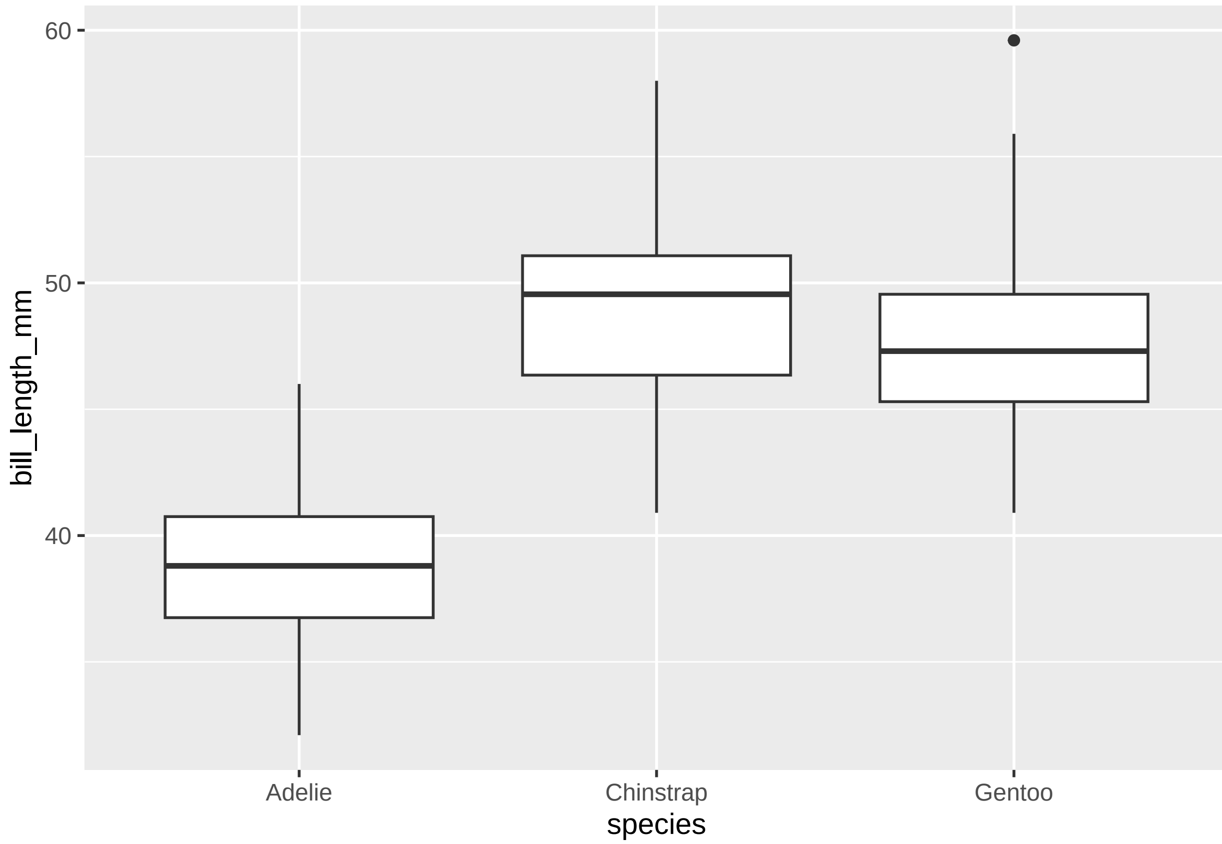
#Introduction

```r
penguins
```

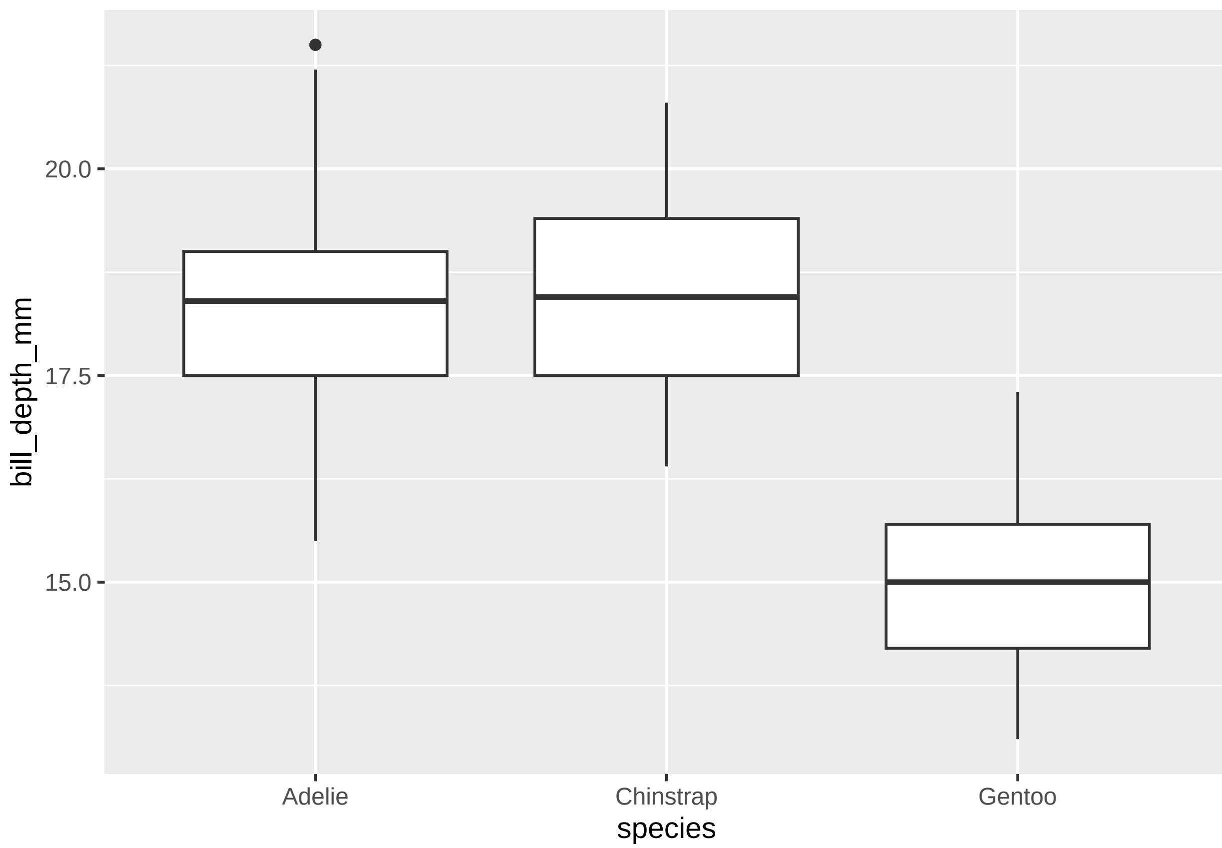
```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.1          18.7               181        3750
##  2 Adelie  Torgersen           39.5          17.4               186        3800
##  3 Adelie  Torgersen           40.3          18                 195        3250
##  4 Adelie  Torgersen           NA            NA                  NA          NA
##  5 Adelie  Torgersen           36.7          19.3               193        3450
##  6 Adelie  Torgersen           39.3          20.6               190        3650
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # i 334 more rows
## # i 2 more variables: sex <fct>, year <int>
```

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

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




```r
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.

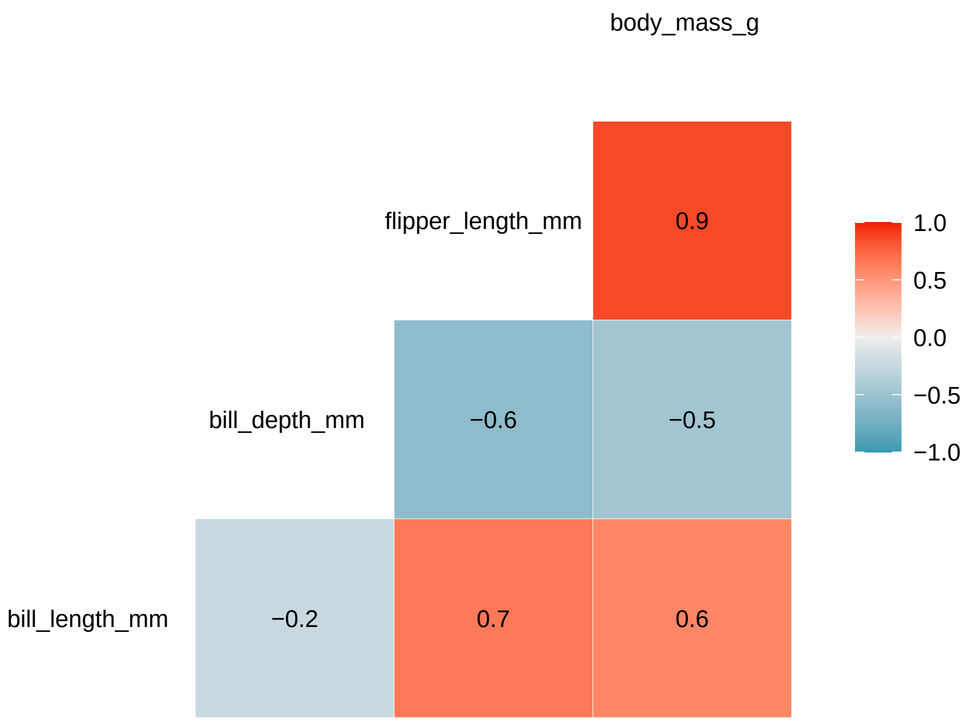
```
pca_penguins2 <- penguins2[,3:6]

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:

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



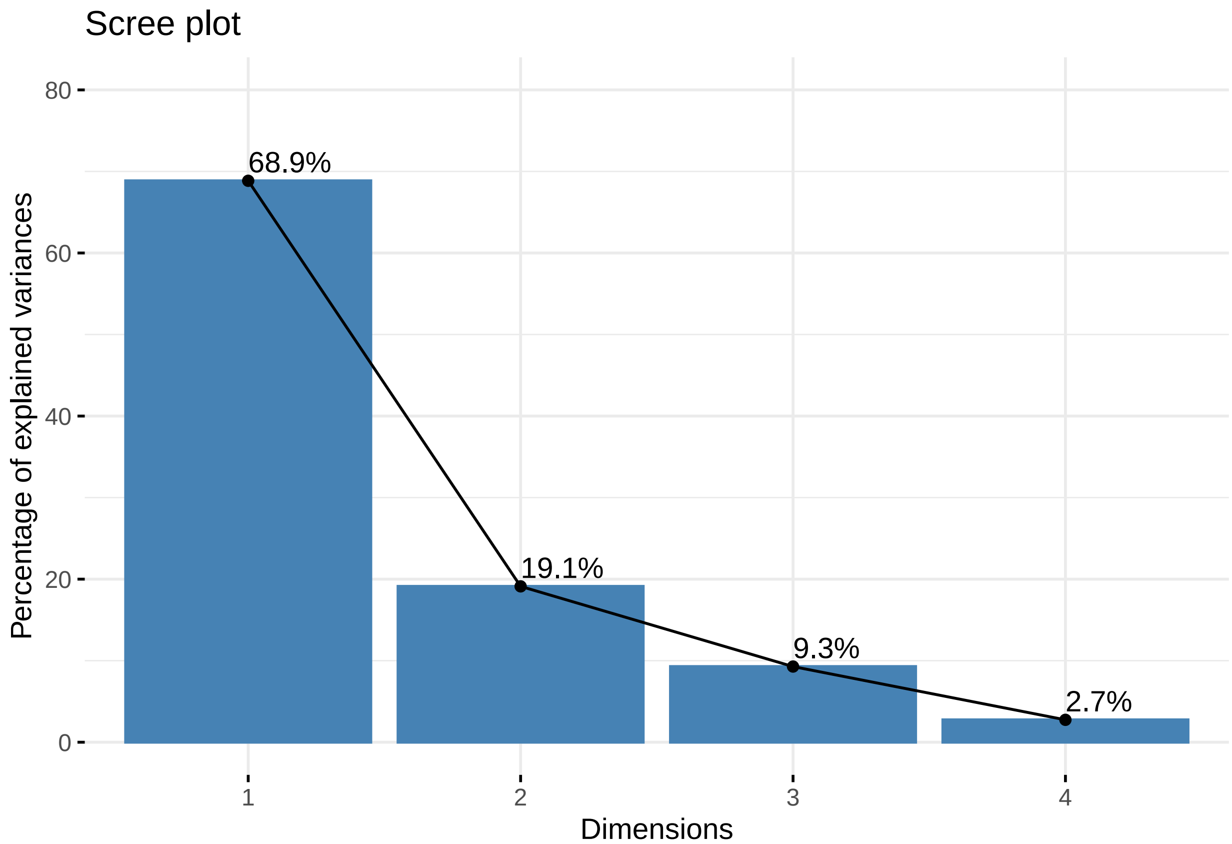
From the correlation matrix, we can see that `body_mass_g` is very correlated with `flipper_length_mm`. Also `bill_length_mm` has a somewhat strong correlation with `flipper_length_mm` and `body_mass_g`, while `bill_depth_mm` has a negative correlation with all of the variables and does not have a strong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

```r
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
```

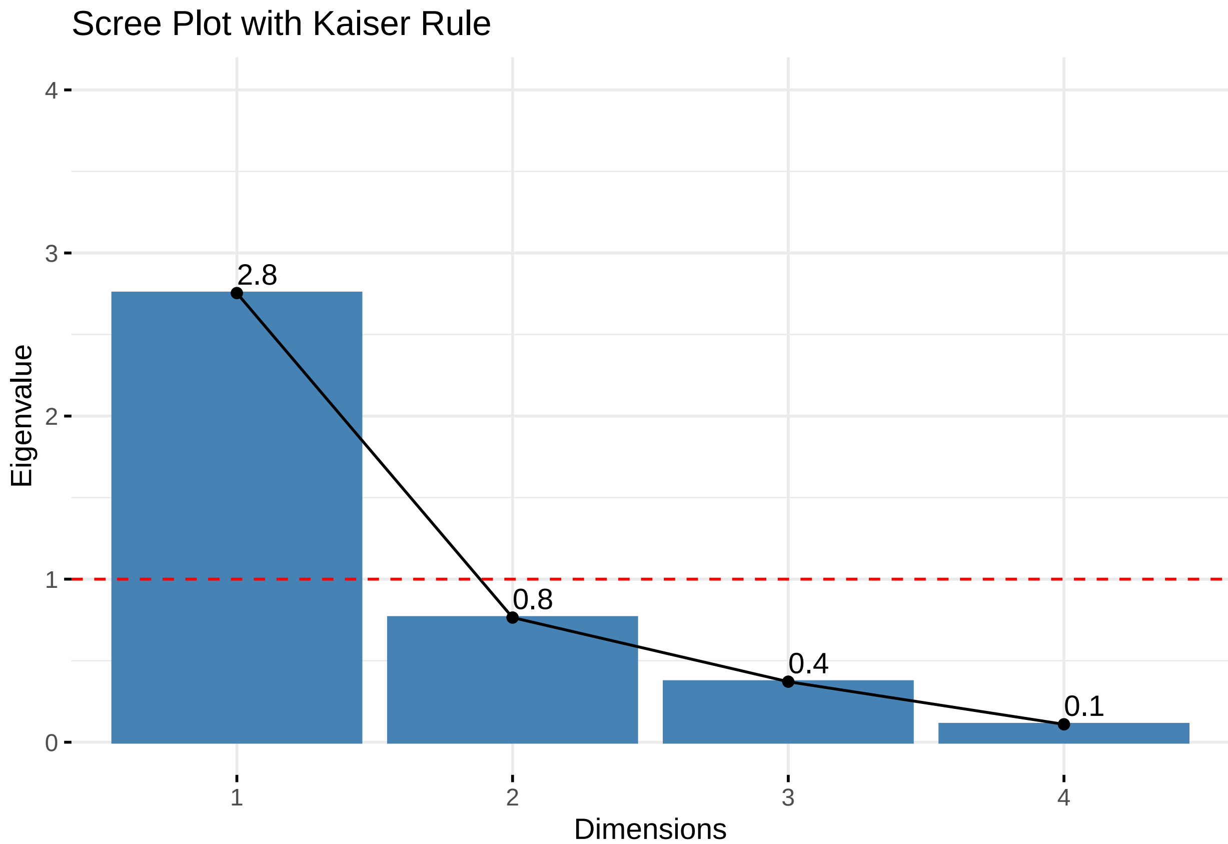
Running the PCA calculations out using the prcomp tool, we see that the first two principal components are responsible for about 88% percent of the data. Let's create a screeplot to visualize this:

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



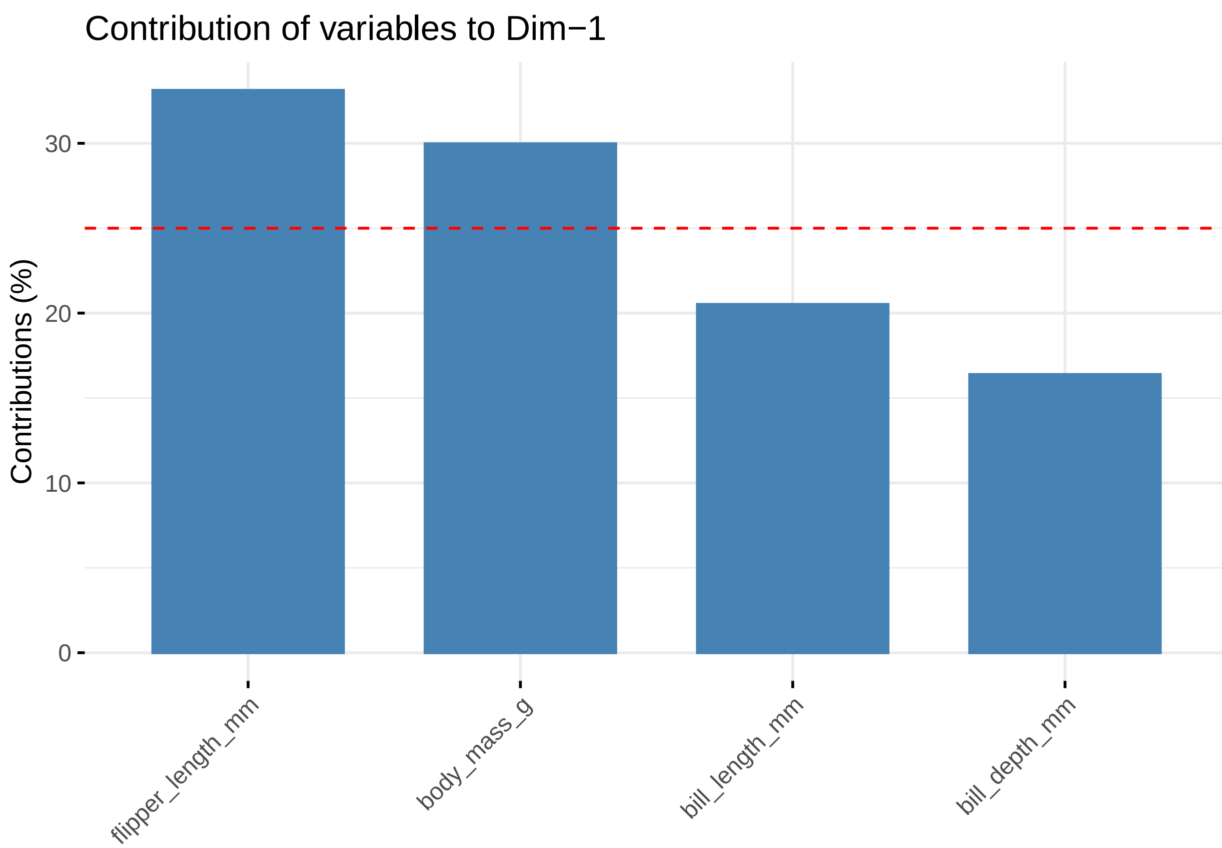
Now let's apply the Kaiser rule to select the number of components, we will use in the PCA:

```
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:

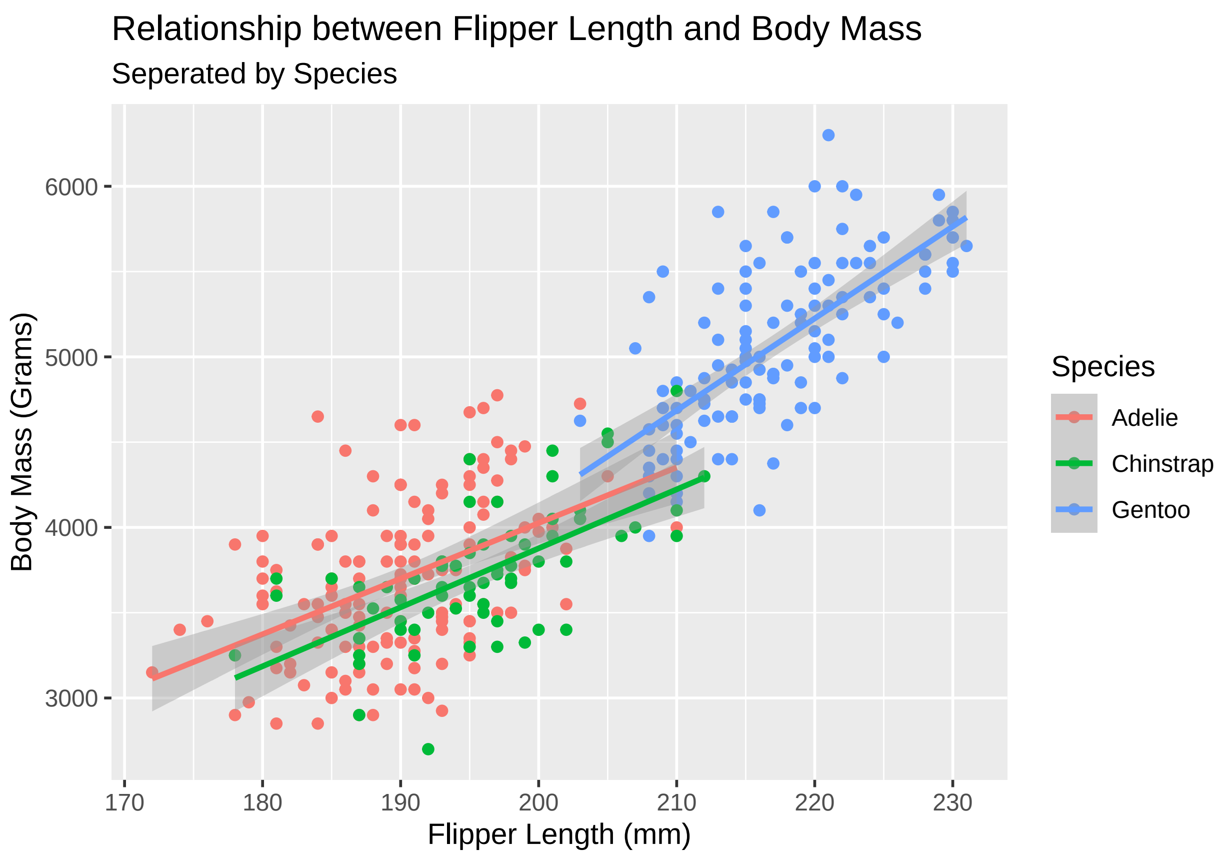
```
fviz_contrib(pca, choice = "var", axes = 1)
```



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`:

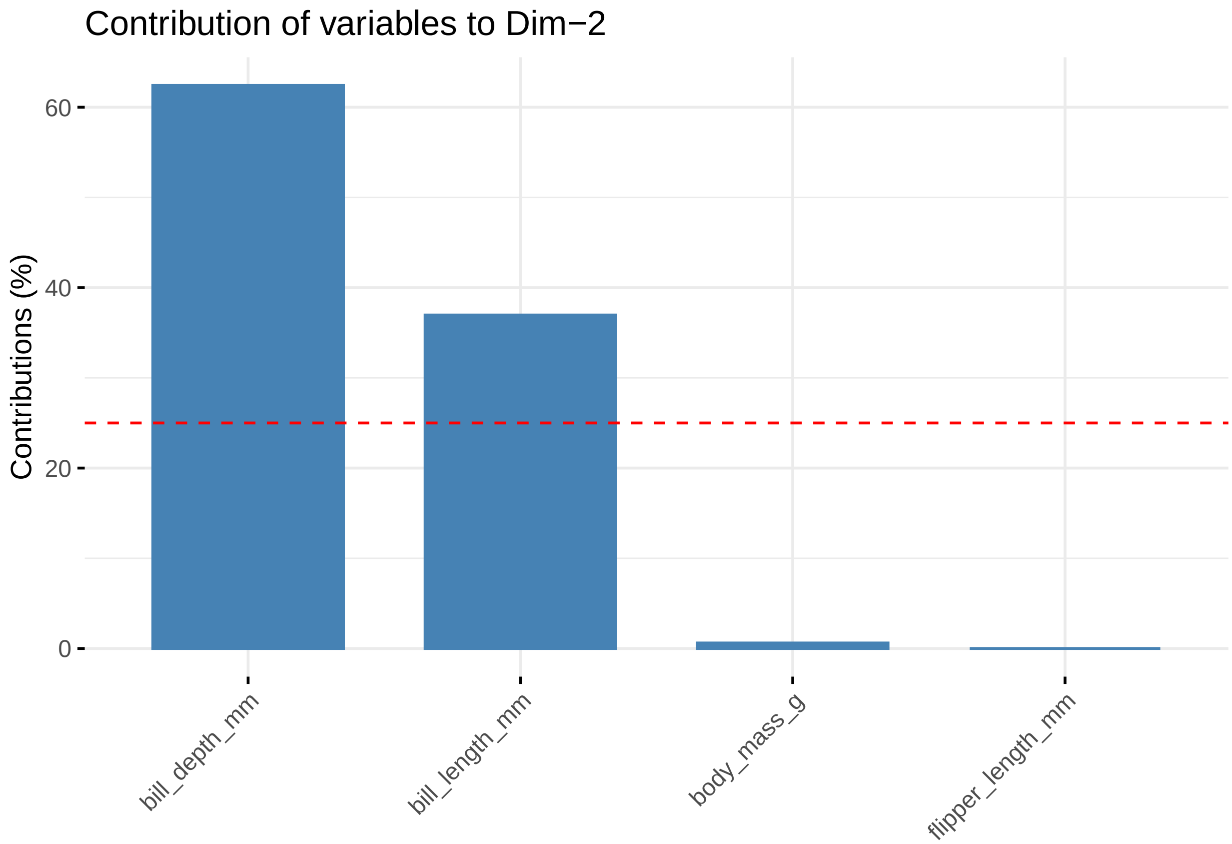
```
penguins2 %>%
  ggplot(aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point() +
  labs(y = "Body Mass (Grams)",
       x = "Flipper Length (mm)",
       title = "Relationship between Flipper Length and Body Mass",
       subtitle = "Seperated by Species",
       color = "Species")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

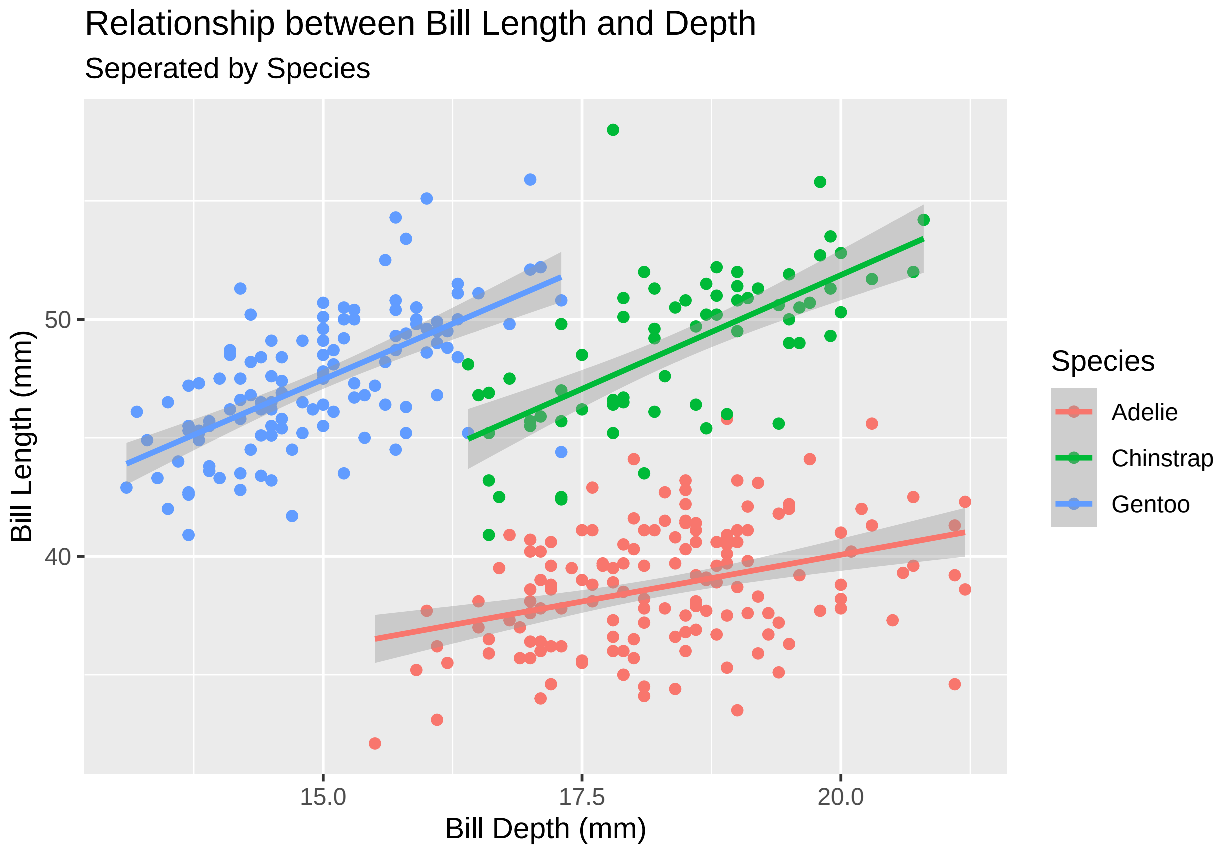
```
fviz_contrib(pca, choice = "var", axes = 2)
```



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