Analyse en Composantes Principales

Marie VAUGOYEAU

2023-01-10

Contents

Analyse factorielle

Si l'analyse factorielle dont l'ACP fait partie commence son histoire avec la parution de l'article de Karl Pearson *On lines and planes of closest fit to systems of points in space* dans *Philosophical Magazine*, c'est le développement de l'informatique qui lui permet de prendre son essor.
C'est une équipe française menée par Jean-Paul Benzécri qui a mis au point l'analyse factorielle des correspondances dans les années 1960.
Incontournable dans de nombreux domaines, elle permet de **réduire le nombre de variables**, de **connaître les liens entre les variables et/ou les individus**, de **qualifier des groupes d'individus**.
Elle est particulièrement utilisée en France.
Les analyses factorielles se réalisent toujours sur un jeu de données rectangulaire avec les individus en lignes (*k lignes*) et les mesures en colonnes (*n colonnes*).
Attention : Une seule ligne par individus !
Les k individus sont vus dans n dimensions.

3 grands types d'analyses factorielles :
_ ACP, Analyse en Composantes Principales : que des variables **quantitatives**
_ AF(D)M, Analyse Factorielle (des Données) Mixtes : variables **quantitatives** et **qualitatives**
_ A(F)CM, Analyse (Factorielle) des Correspondances Multiples : variables **qualitatives** uniquement

Les données : les pingouins de Palmer

Jeu de données extrait du package `{palmerpenguins}` qui est une alternative au jdd `iris`

```
library(palmerpenguins)
library(tidyverse)
library(tourr)

penguins
```

```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipper_~1 body_~2 sex    year
##    <fct>   <fct>              <dbl>         <dbl>      <int>   <int> <fct> <int>
##  1 Adelie  Torgersen           39.1          18.7        181    3750 male   2007
##  2 Adelie  Torgersen           39.5          17.4        186    3800 fema~  2007
##  3 Adelie  Torgersen           40.3          18          195    3250 fema~  2007
##  4 Adelie  Torgersen           NA            NA           NA      NA <NA>   2007
##  5 Adelie  Torgersen           36.7          19.3        193    3450 fema~  2007
##  6 Adelie  Torgersen           39.3          20.6        190    3650 male   2007
##  7 Adelie  Torgersen           38.9          17.8        181    3625 fema~  2007
##  8 Adelie  Torgersen           39.2          19.6        195    4675 male   2007
##  9 Adelie  Torgersen           34.1          18.1        193    3475 <NA>   2007
## 10 Adelie  Torgersen           42            20.2        190    4250 <NA>   2007
## # ... with 334 more rows, and abbreviated variable names 1: flipper_length_mm,
## #   2: body_mass_g
```

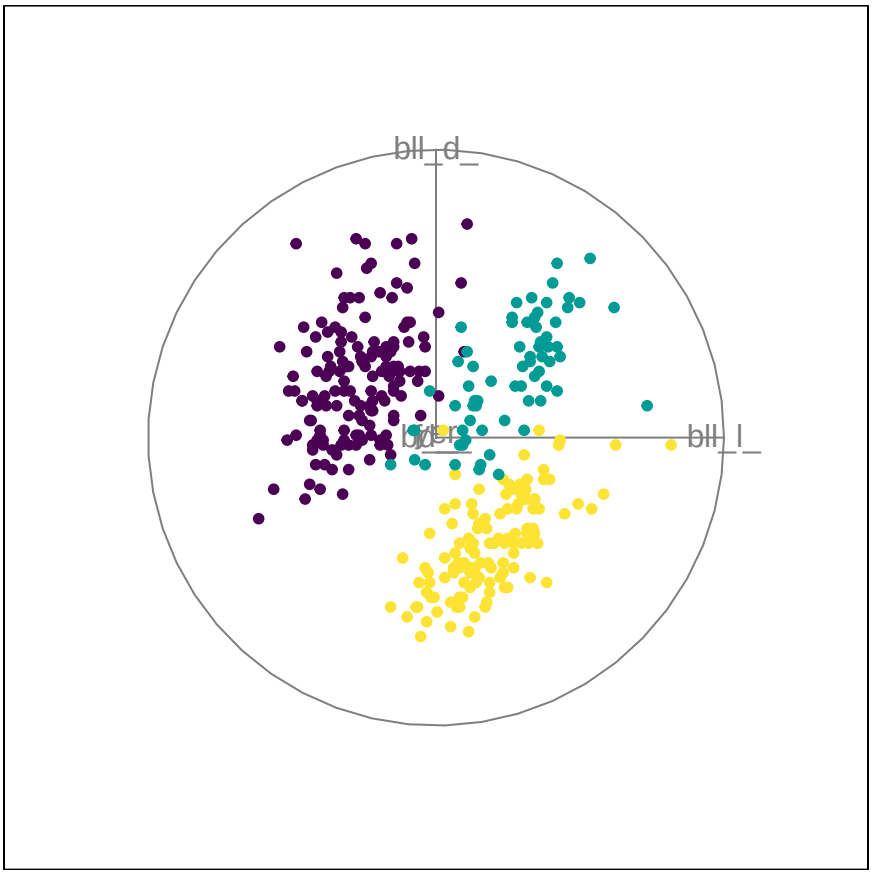
```
pingouin <- penguins %>%
  select(- sex) %>%
  drop_na()

pingouin %>%
  count(species, island)
```

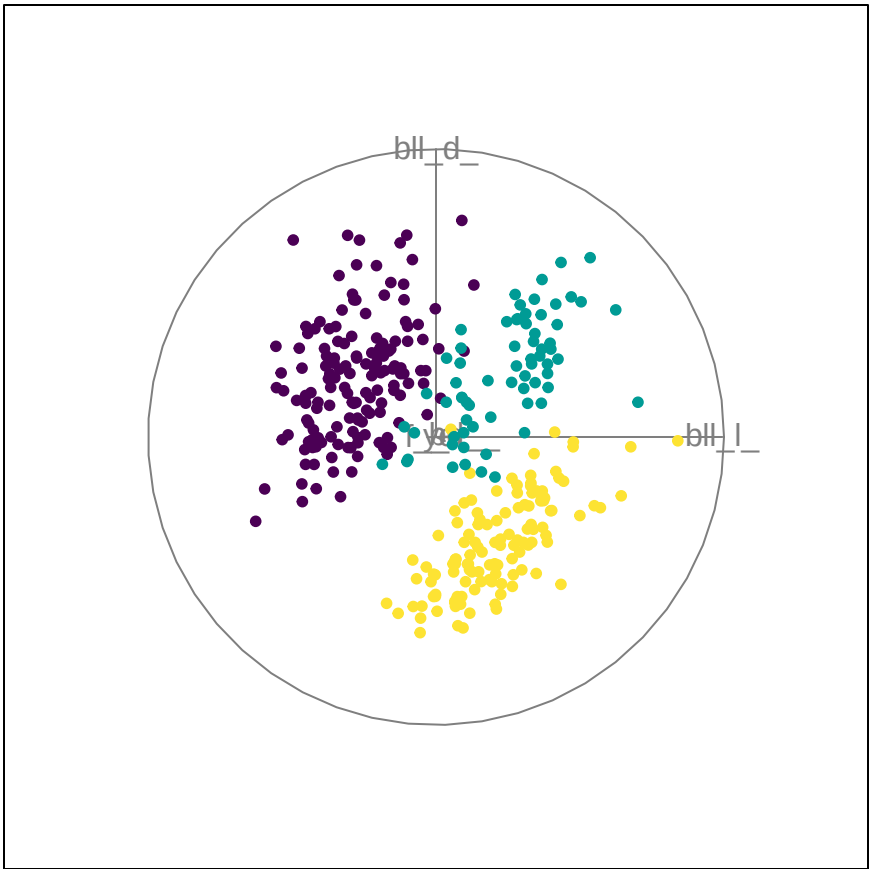
```
## # A tibble: 5 x 3
##   species   island        n
##   <fct>     <fct>     <int>
## 1 Adelie    Biscoe       44
## 2 Adelie    Dream        56
## 3 Adelie    Torgersen    51
## 4 Chinstrap Dream        68
## 5 Gentoo    Biscoe      123
```

```
pingouin %>%
  group_by(species) %>%
  summarize(
    across(where(is.numeric),
           mean
    )
  ) %>%
  ungroup()
```

```
## # A tibble: 3 x 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g  year
##   <fct>              <dbl>         <dbl>             <dbl>       <dbl> <dbl>
## 1 Adelie              38.8          18.3              190.       3701. 2008.
## 2 Chinstrap           48.8          18.4              196.       3733. 2008.
## 3 Gentoo              47.5          15.0              217.       5076. 2008.
```

```
animate(
  data = pingouin %>%
    select(-island, -species),
  display = display_xy(col = pingouin$species)
)
```



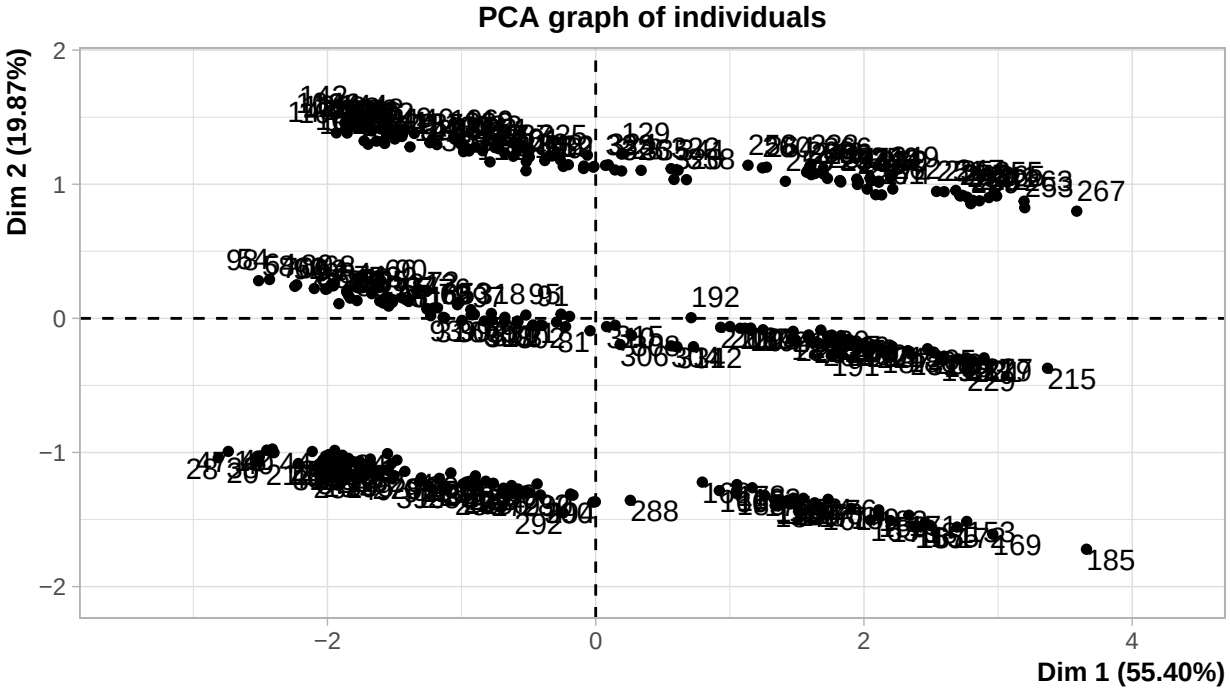


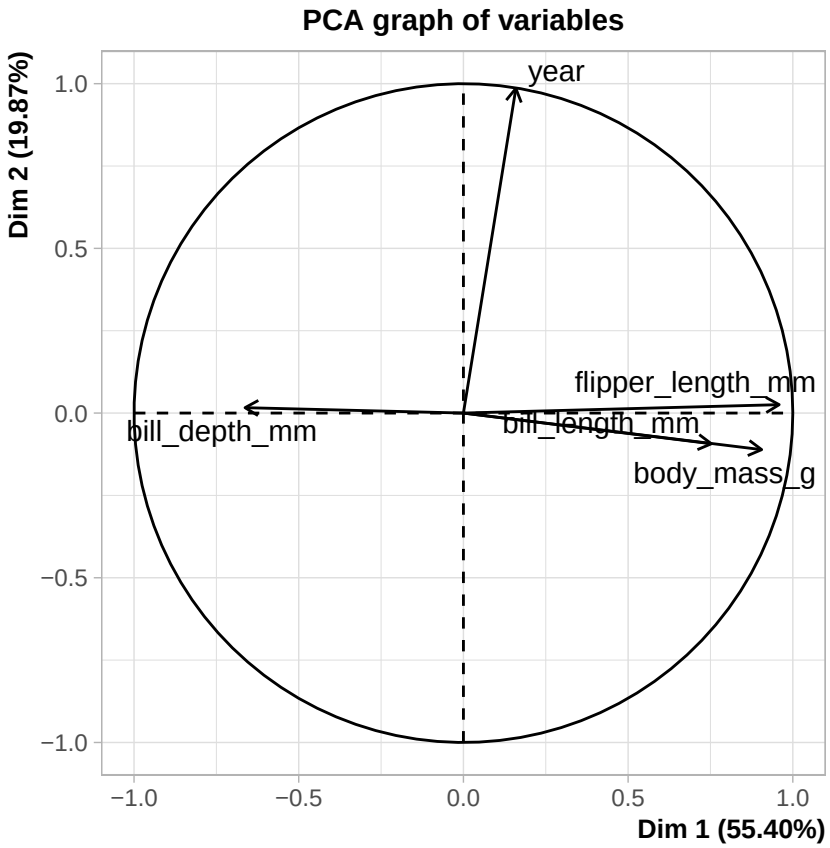
Visualisation en utilisant le package `{tourr}` qui permet de visuliser le nuage des k individus dans les n dimensions, ici 342 pingouins dans 5 dimensions.

ACP sans variables supplémentaires

Utilisation du package `{FactoMineR}` pour réaliser les analyses factorielles.

```
library(FactoMineR)

acp_simple <- pingouin %>%
  select(- species, - island) %>%
  PCA()
```


PCA graph of individuals
Dim 2 (19.87%)
Dim 1 (55.40%)
2
1
0
−1
−2
−2
0
2
4
129
267
192
215
288
185
469




```
acp_simple$eig
```

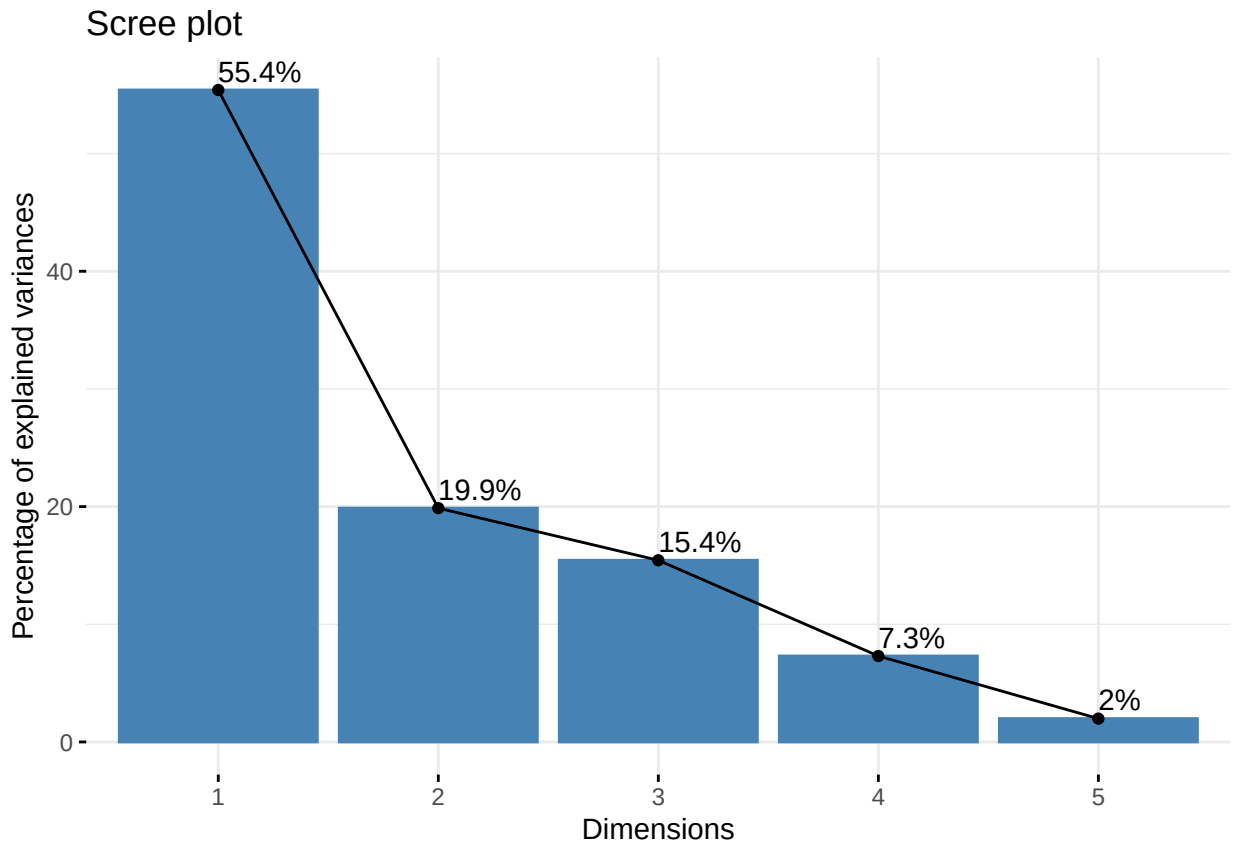
```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 2.77008681              55.401736                           55.40174
## comp 2 0.99348866              19.869773                           75.27151
## comp 3 0.77191746              15.438349                           90.70986
## comp 4 0.36520940               7.304188                           98.01405
## comp 5 0.09929767               1.985953                          100.00000
```

```
dimdesc(acp_simple)
```

```
## $Dim.1
##
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##                   correlation       p.value
## flipper_length_mm   0.9585908 1.552310e-187
## body_mass_g         0.9054051 1.493199e-128
## bill_length_mm      0.7526986  1.089066e-63
## year                0.1592763  3.140610e-03
## bill_depth_mm      -0.6629540  1.173808e-44
##
## $Dim.2
##
```

```
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##              correlation       p.value
## year           0.9856765 6.531669e-265
## body_mass_g   -0.1110509  4.011874e-02
##
## $Dim.3
##
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##                 correlation      p.value
## bill_depth_mm     0.7027357 3.289288e-52
## bill_length_mm    0.5208740 3.532738e-25
```

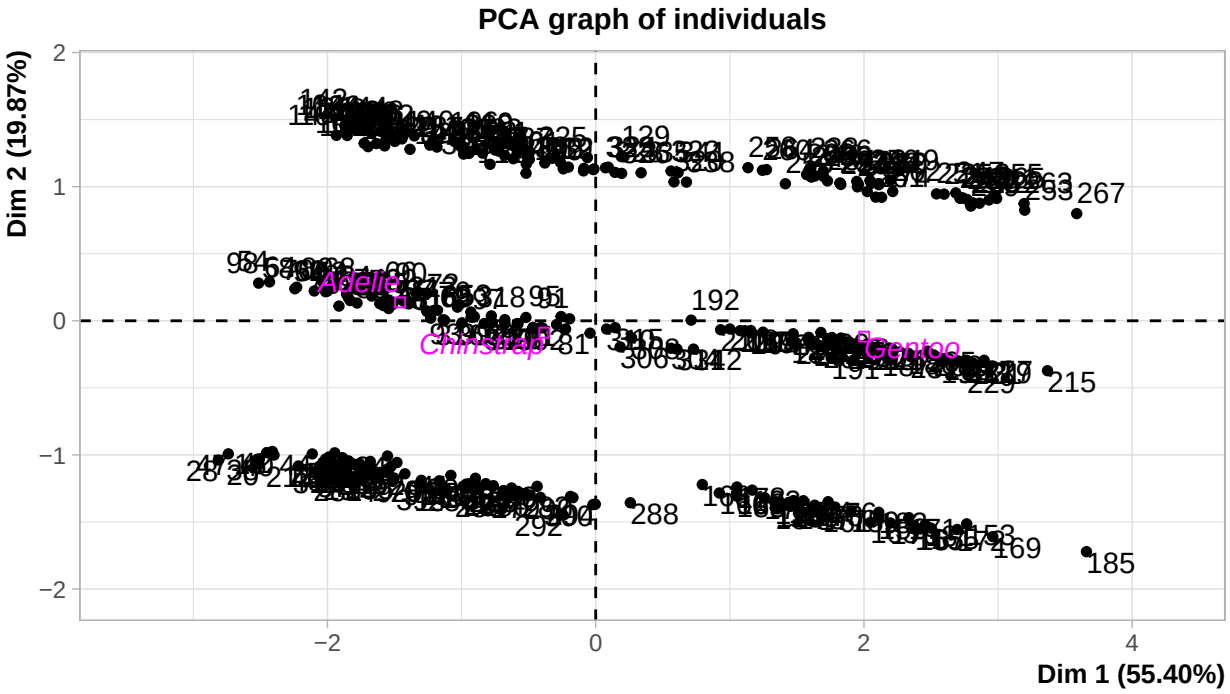
```
library(factoextra)
fviz_screeplot(acp_simple, addlabels = TRUE)
```

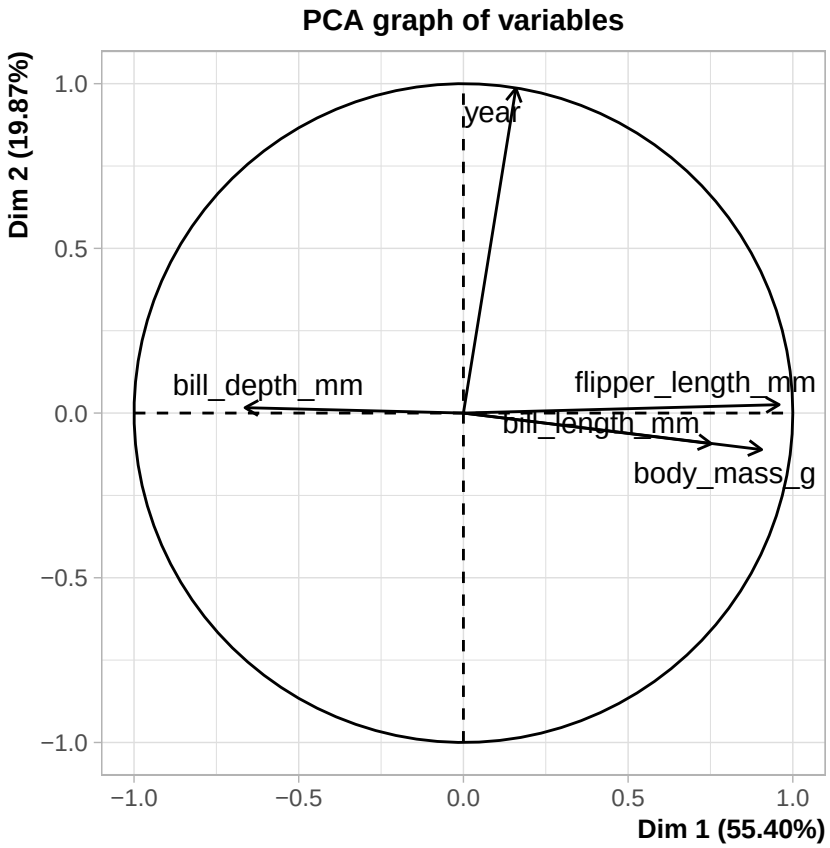


ACP avec `species` et `island` comme variables supplémentaires

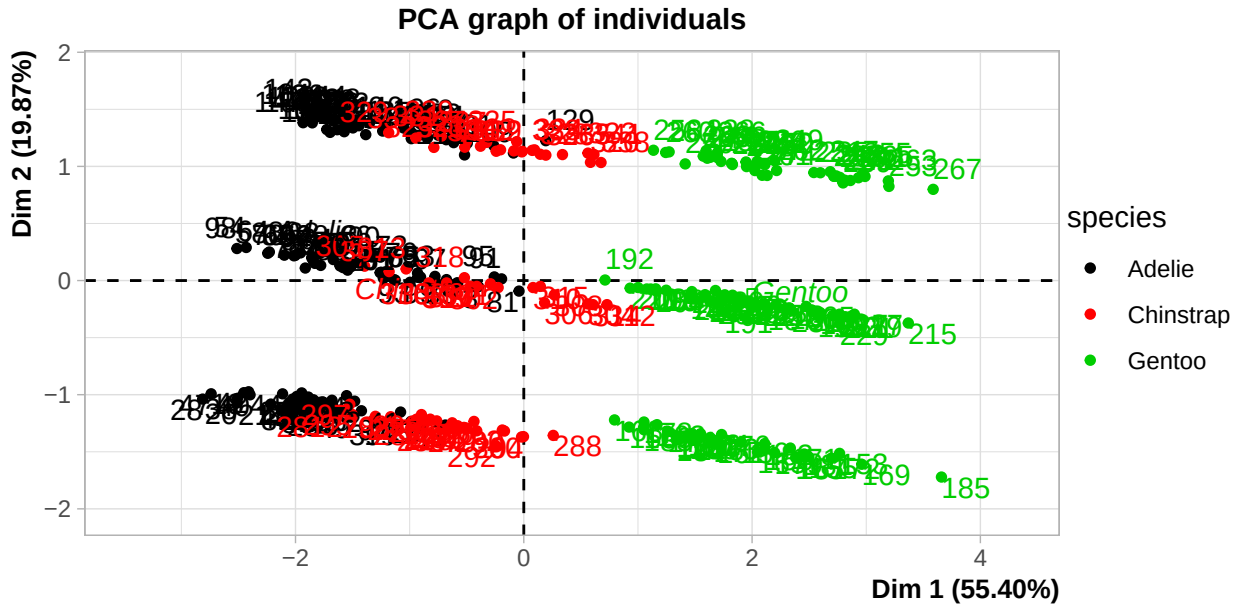
```
acp <- PCA(
  X = pingouin %>%
    select(-island),
```

```
  quali.sup = 1
)
```

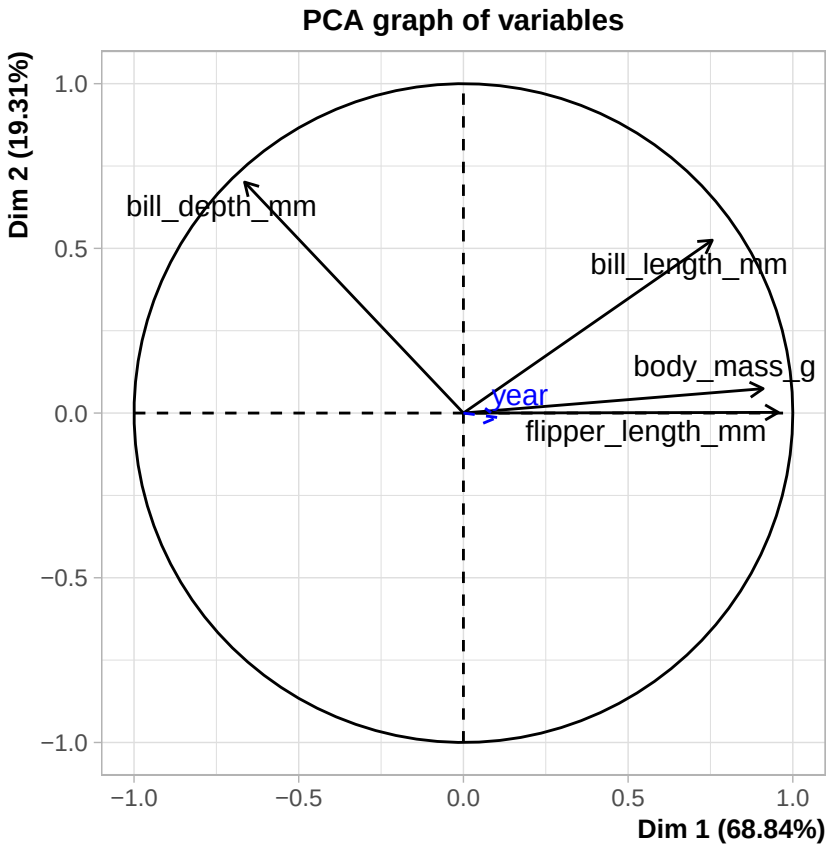

PCA graph of individuals
Dim 2 (19.87%)
Dim 1 (55.40%)
Adelie
Chinstrap
Gentoo
2
1
0
−1
−2
−2
0
2
4
192
215
267
288
185
169




```
plot.PCA(acp, choix = "ind", habillage = 1)
```



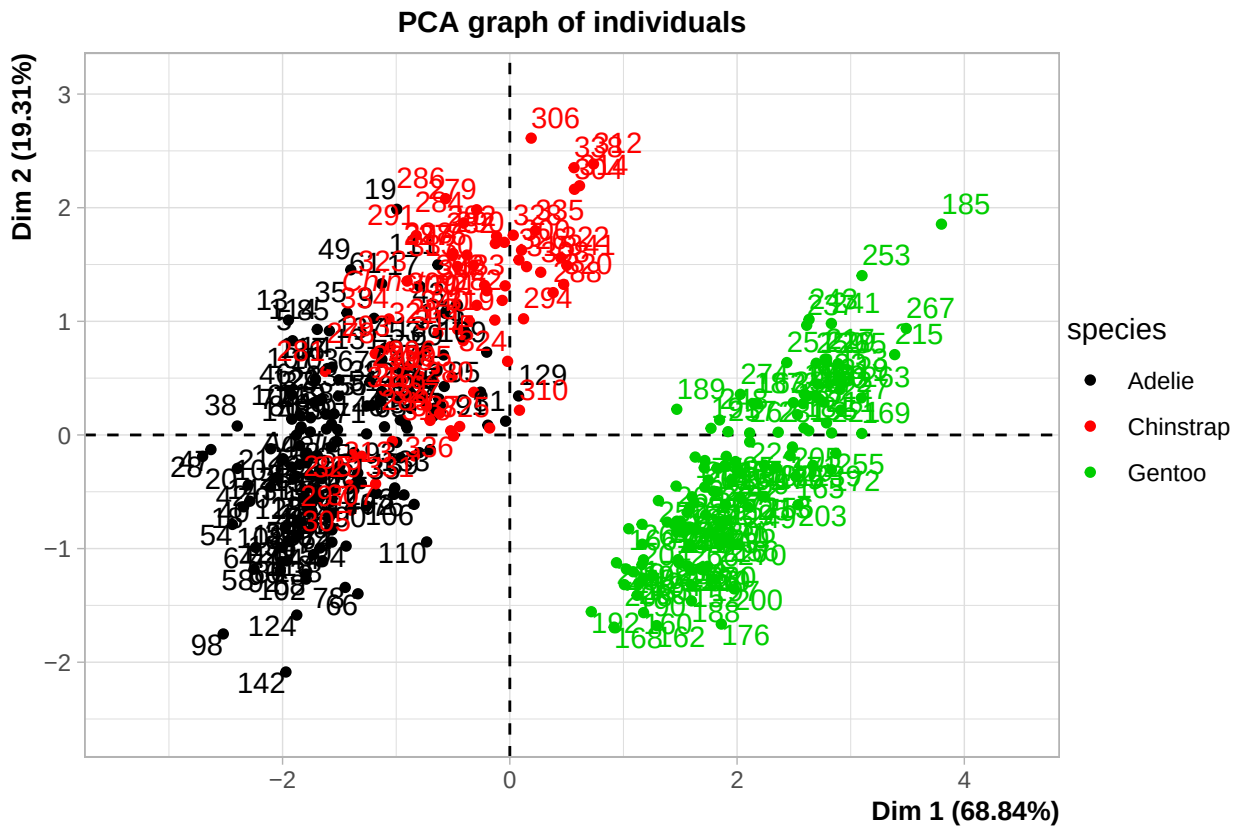

```
acp_annee <-
  PCA(
    X = pingouin %>% select(-island),
    quali.sup = 1,
    quanti.sup = 6
  )
```



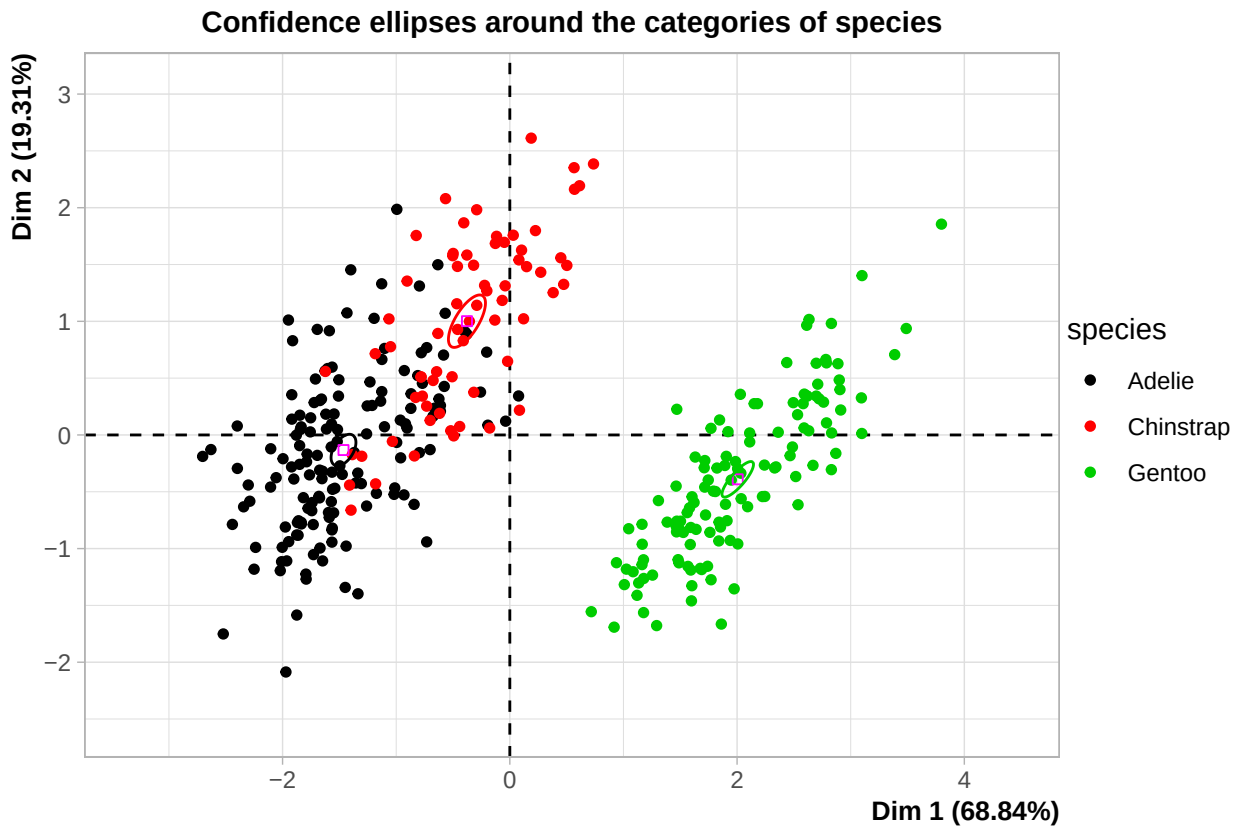

```
acp_annee$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1  2.7537551              68.843878                          68.84388
## comp 2  0.7725168              19.312919                          88.15680
## comp 3  0.3652359               9.130898                          97.28769
## comp 4  0.1084922               2.712305                         100.00000
```

```
plot.PCA(acp_annee, choix = "ind", habillage = 1)
```




```
plotellipses(acp_annee, keepvar = "species", label = "none")
```



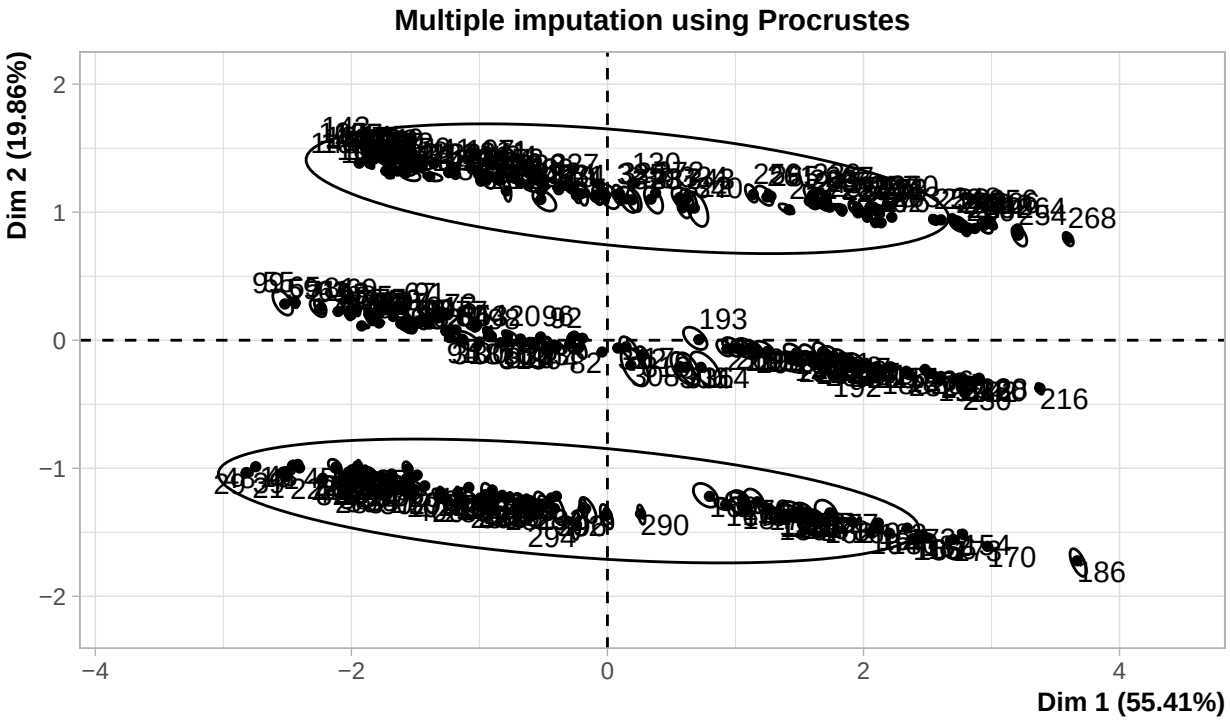
valeurs manquantes

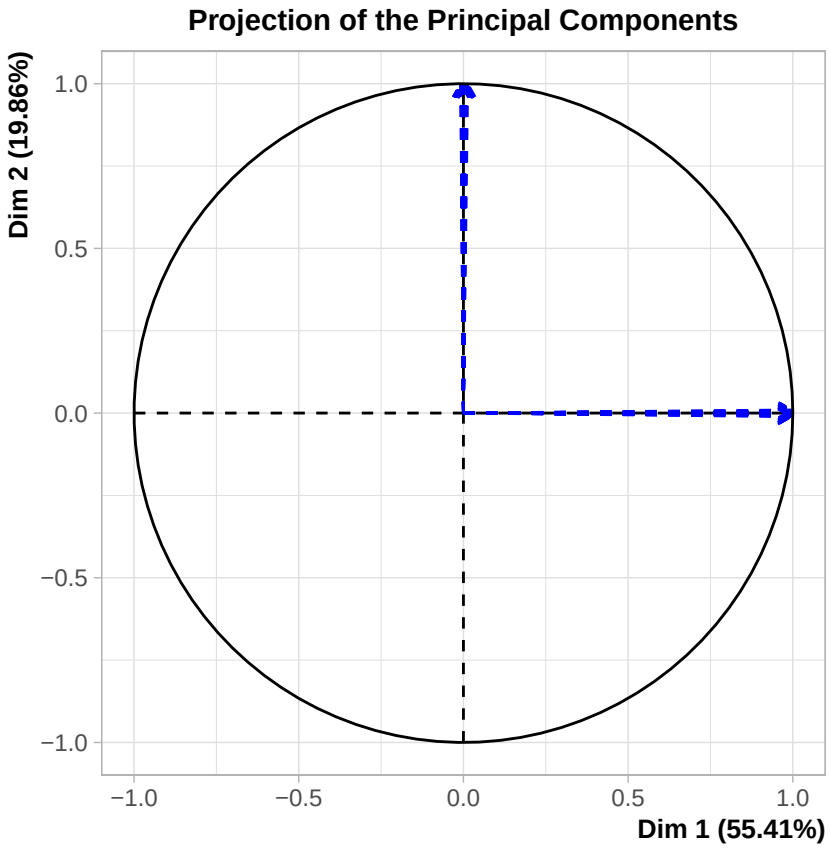
missMDA

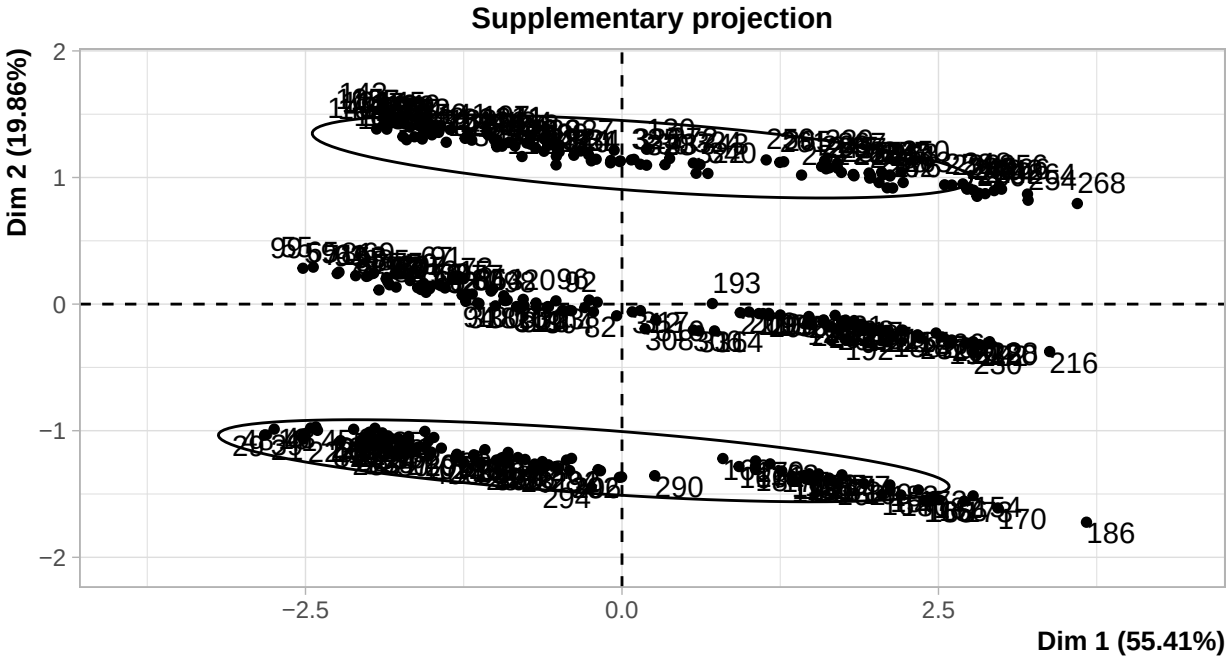
```
penguins %>%
  filter(is.na(bill_depth_mm))
```

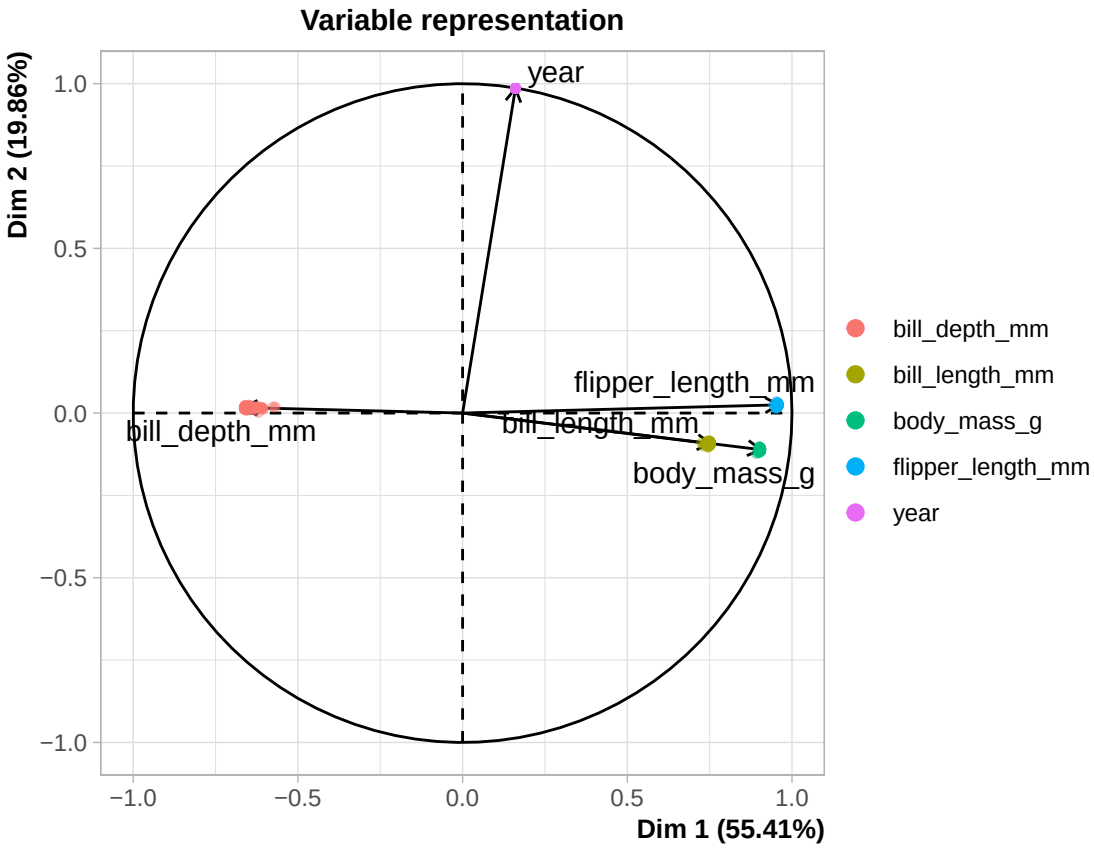
```
## # A tibble: 2 x 8
##   species island    bill_length_mm bill_depth_mm flipper_l~1 body_~2 sex    year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
## 1 Adelie  Torgersen             NA            NA          NA      NA <NA>   2007
## 2 Gentoo  Biscoe                NA            NA          NA      NA <NA>   2009
## # ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

```
pingouin_vm <- penguins %>%
  select(bill_length_mm:body_mass_g, year)

library(missMDA)
n <- estim_ncpPCA(pingouin_vm)

pingouin_vm_complete <- MIPCA(pingouin_vm, ncp = n$ncp)

plot.MIPCA(pingouin_vm_complete)
```

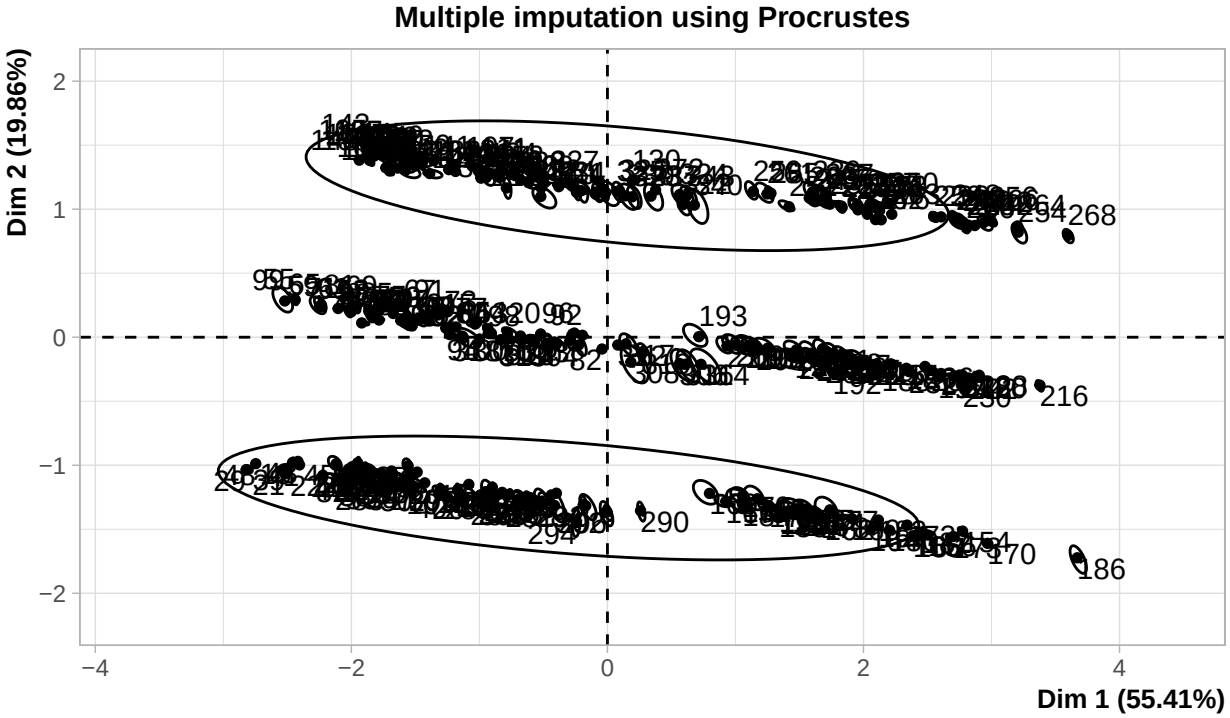

Multiple imputation using Procrustes
Dim 2 (19.86%)
Dim 1 (55.41%)
2
1
0
−1
−2
−4
−2
0
2
4
268
193
216
290
170
186


Projection of the Principal Components
Dim 2 (19.86%)
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Dim 1 (55.41%)

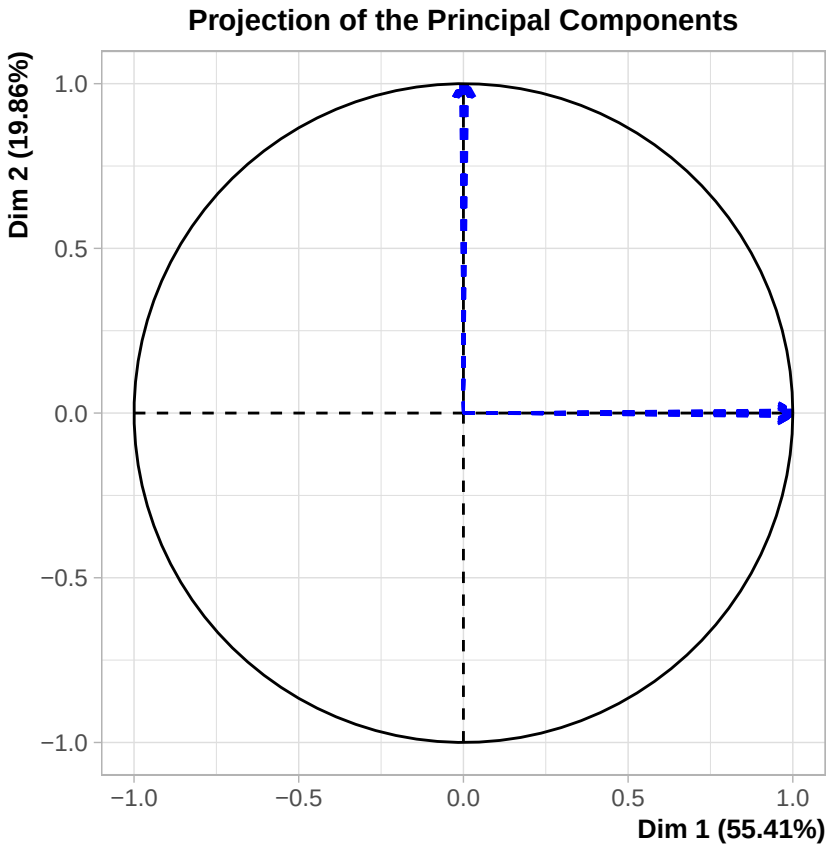

Supplementary projection
Dim 2 (19.86%)
Dim 1 (55.41%)
2
1
0
−1
−2
−2.5
0.0
2.5
193
216
290
186
170
268



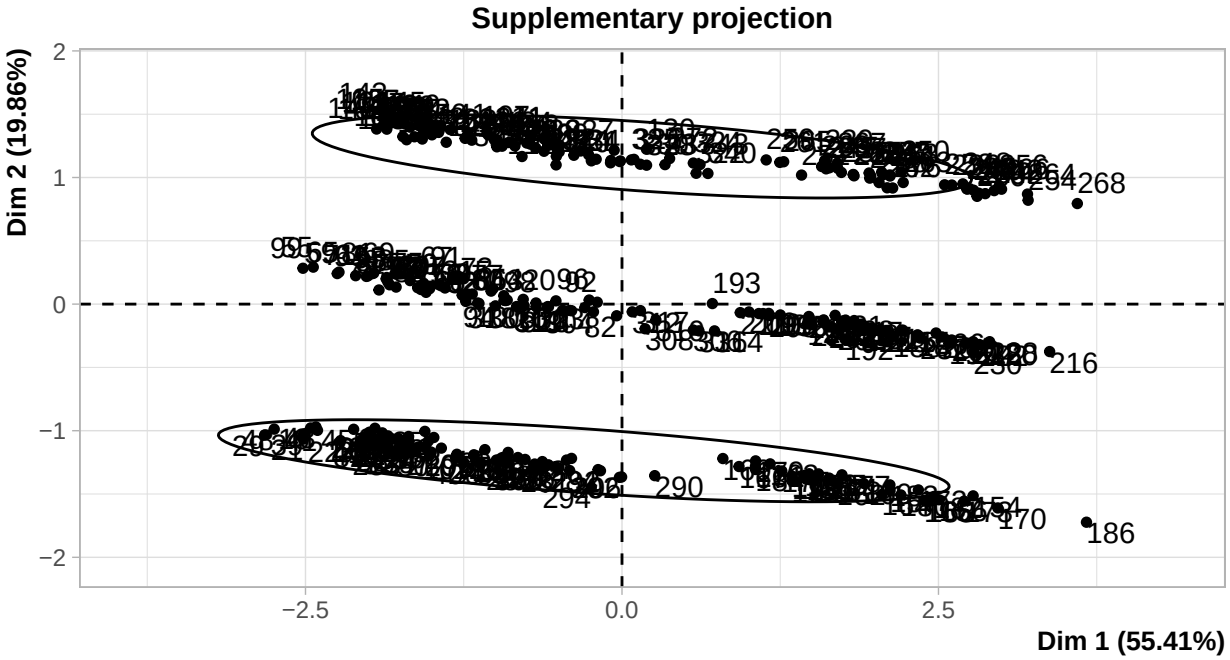

```
## $PlotIndProc
```


Multiple imputation using Procrustes
Dim 2 (19.86%)
Dim 1 (55.41%)
2
1
0
−1
−2
−4
−2
0
2
4
193
216
290
170
186
268

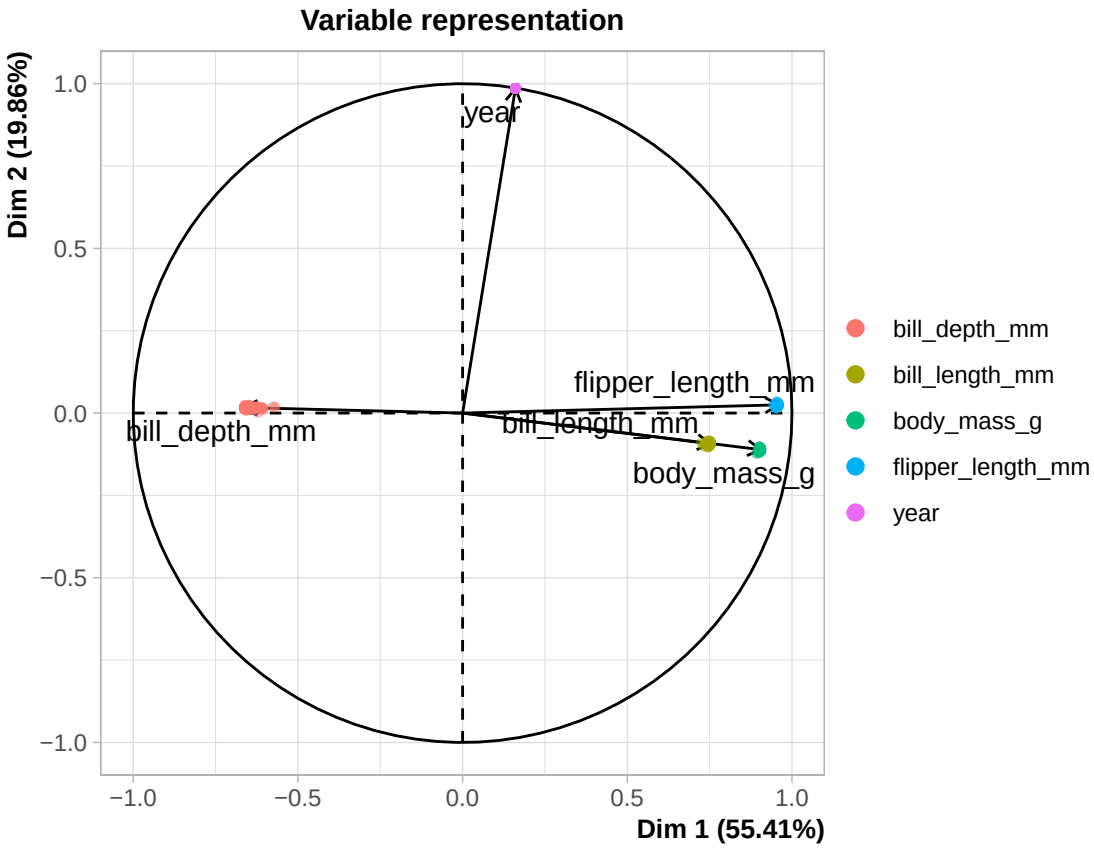
```
##
## $PlotDim
```




```
##
## $PlotIndSupp
```


Supplementary projection
Dim 2 (19.86%)
Dim 1 (55.41%)
2
1
0
−1
−2
−2.5
0.0
2.5
193
216
290
186

```
##
## $PlotVar
```




```
pingouin_complete <-
  bind_cols(
    penguins %>% select(species, island),
    pingouin_vm_complete[["res.imputePCA"]]
  )
```