Unsupervised learning in R

Using the penguins data set

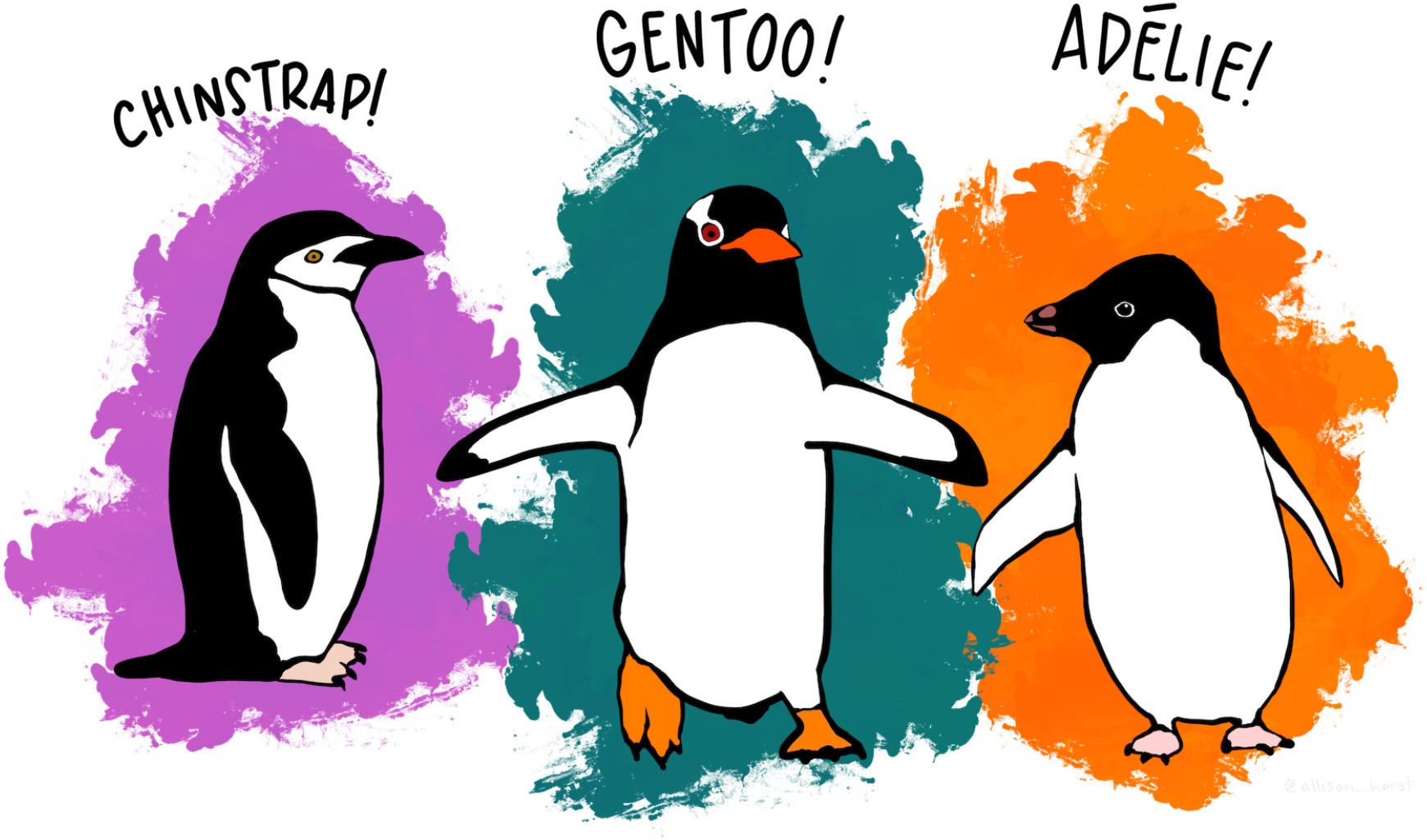
Kaspar Martens & Kasia Kedzierska

Taught on 17 Sep, 2022

Last compiled on 17 Sep, 2022

palmerpenguins data set

LTER Penguins



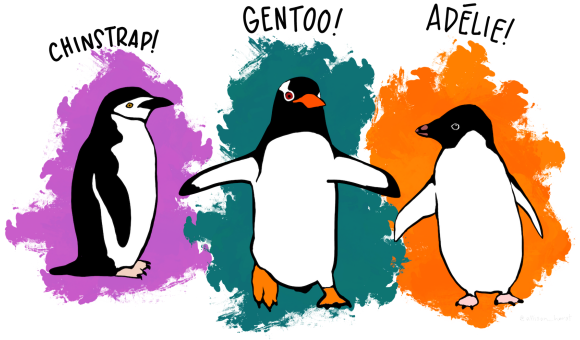
Data set

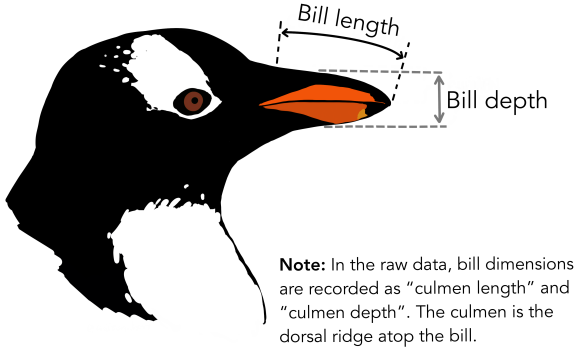
The palmerpenguins data set = 344 penguins

- specie name
- island
- sex
- year

Measurements:

- bill length,
- bill depth,
- flipper length and
- body mass.



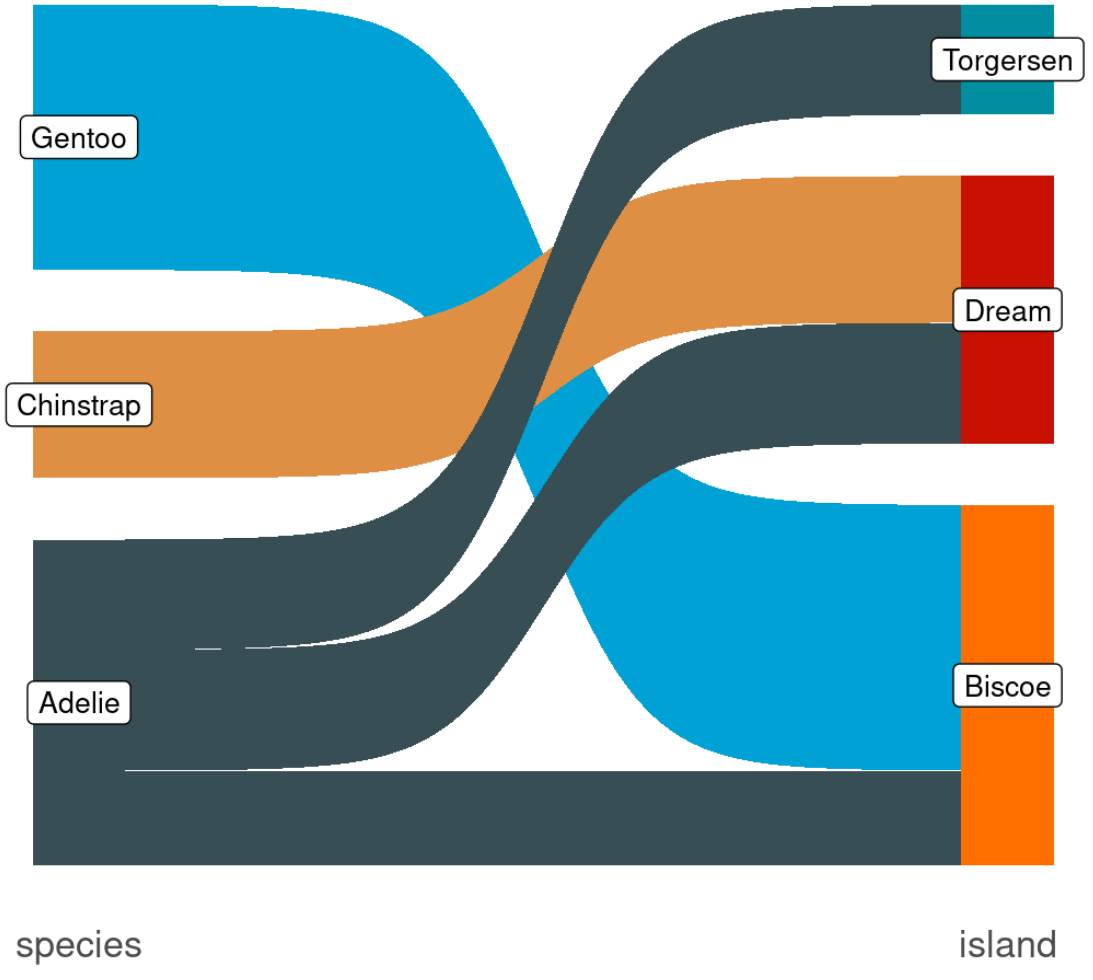


penguins data.frame

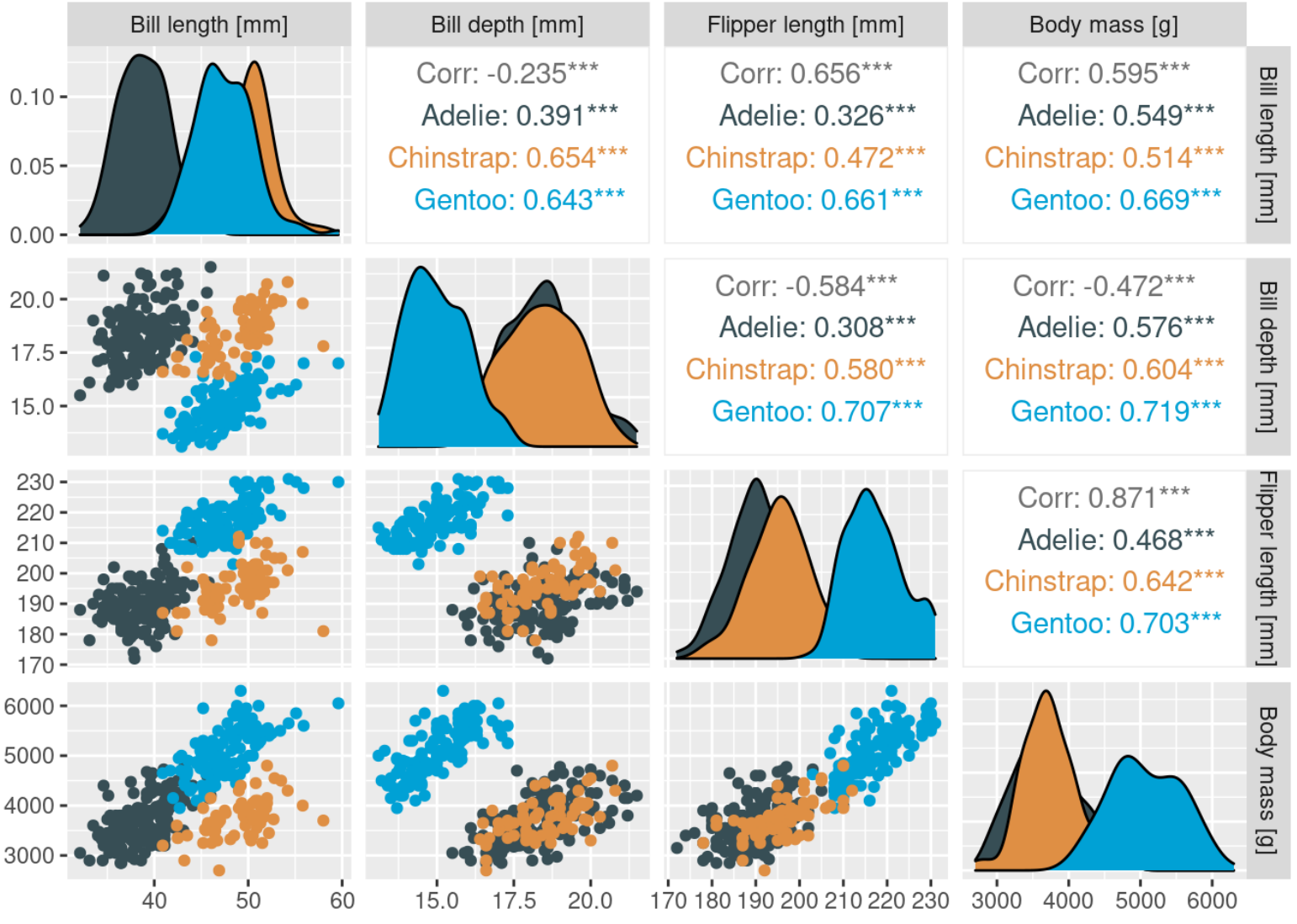
species <fct>	island <fct>	bill_length_mm <dbl>	bill_depth_mm <dbl>
Adelie	Biscoe	39.6	17.7
Gentoo	Biscoe	50.4	15.3
Adelie	Biscoe	42.2	19.5
Chinstrap	Dream	55.8	19.8
Gentoo	Biscoe	45.2	13.8

5 rows | 1-4 of 9 columns

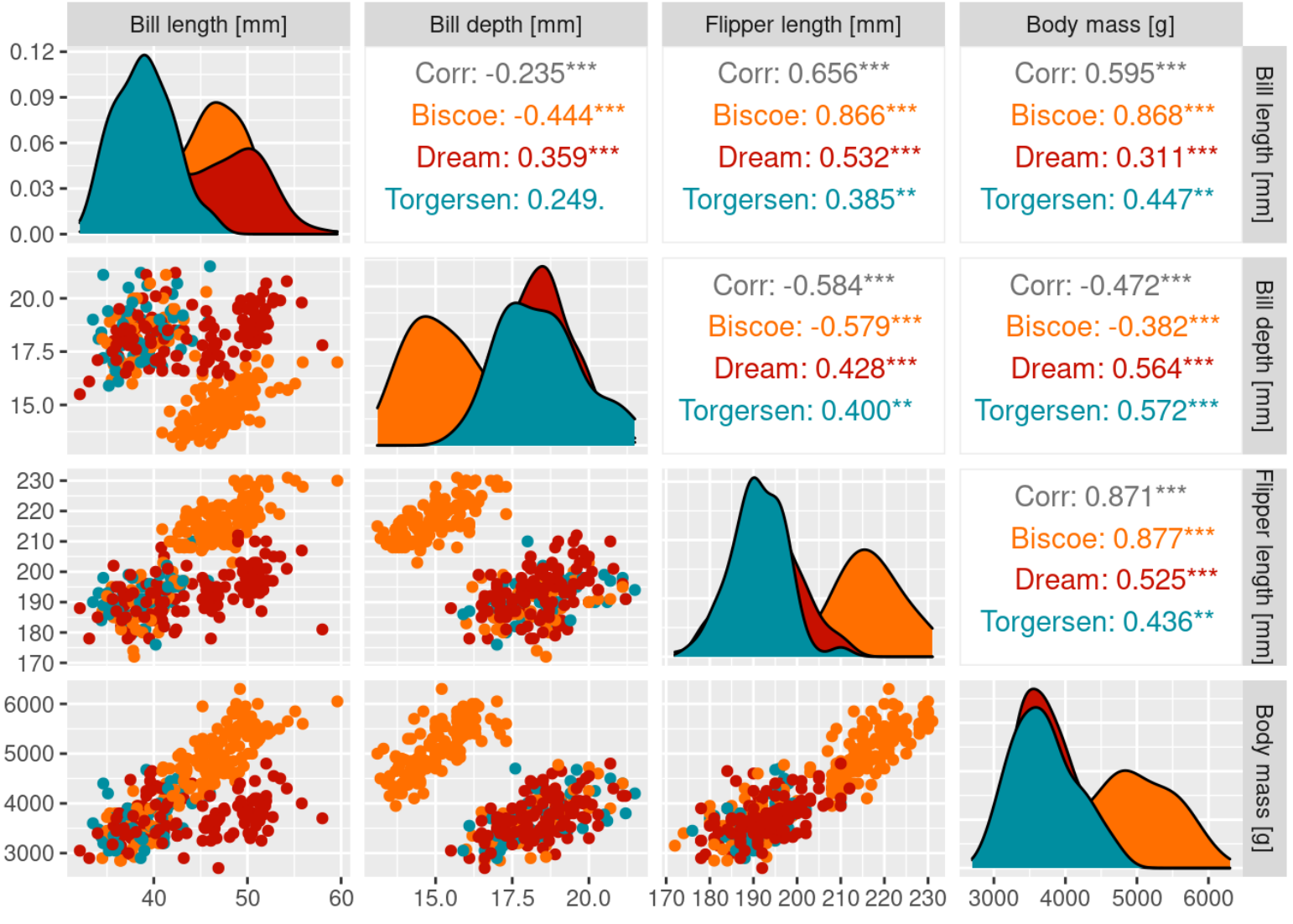
penguins data



penguins data - exploration



penguins data - exploration



PCA

Input matrix

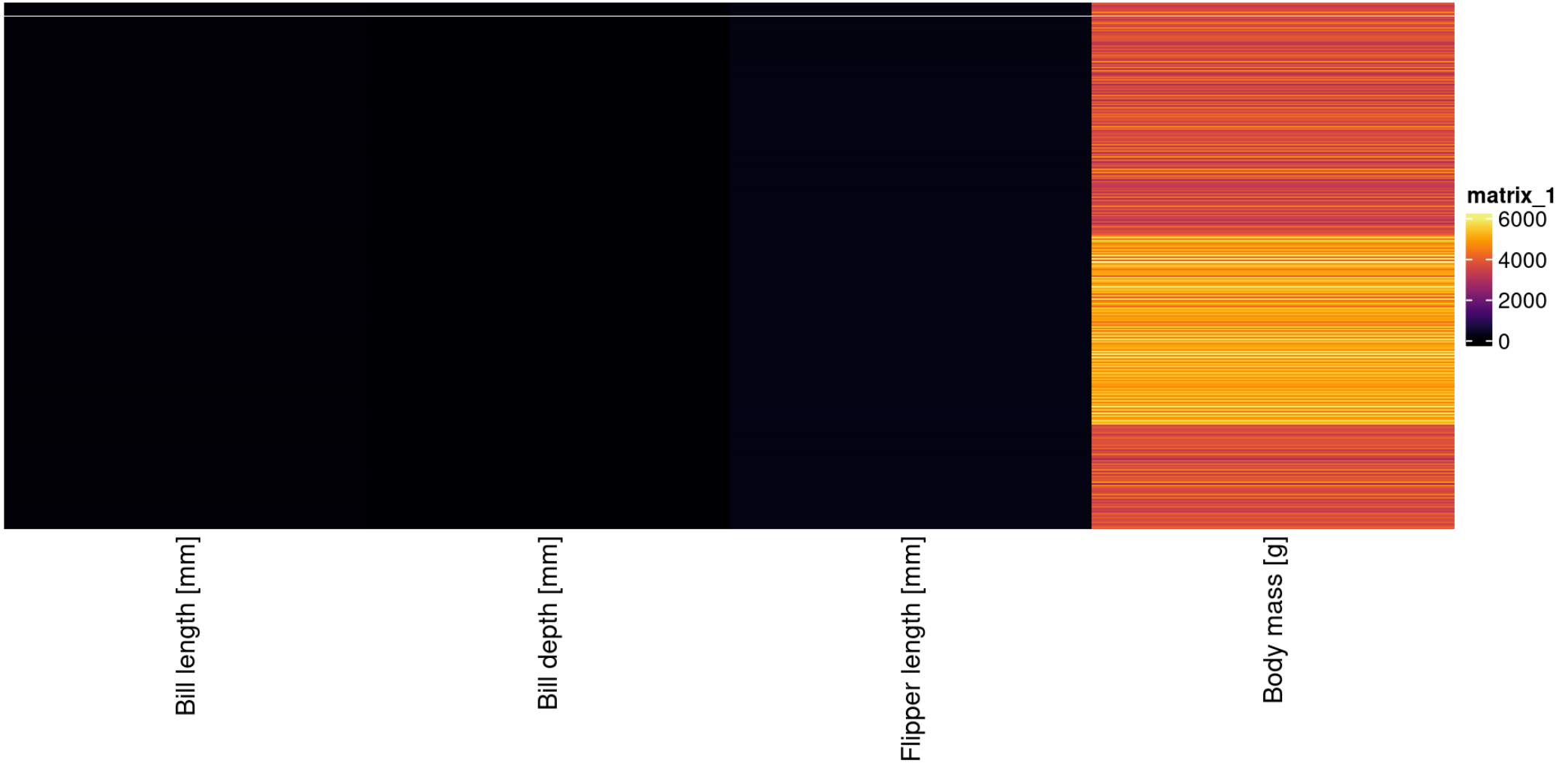
```
penguins_mat <-
  penguins_df %>%
  mutate(row_names = penguin_id) %>%
  select(penguin_id,
         bill_length_mm, bill_depth_mm,
         flipper_length_mm, body_mass_g) %>%
  column_to_rownames("penguin_id") %>%
  as.matrix()
```

Input matrix

```
1 penguins_mat[1:5,]
```

```
                        bill_length_mm bill_depth_mm flipper_length_mm
sympathetic_emu                   39.1          18.7                181
lethargic_horseshoecrab           39.5          17.4                186
enarthrodial_hogget               40.3          18.0                195
postindustrial_bunting            36.7          19.3                193
complexioned_hagfish              39.3          20.6                190
                        body_mass_g
sympathetic_emu                3750
lethargic_horseshoecrab        3800
enarthrodial_hogget            3250
postindustrial_bunting         3450
complexioned_hagfish           3650
```

Input matrix - raw

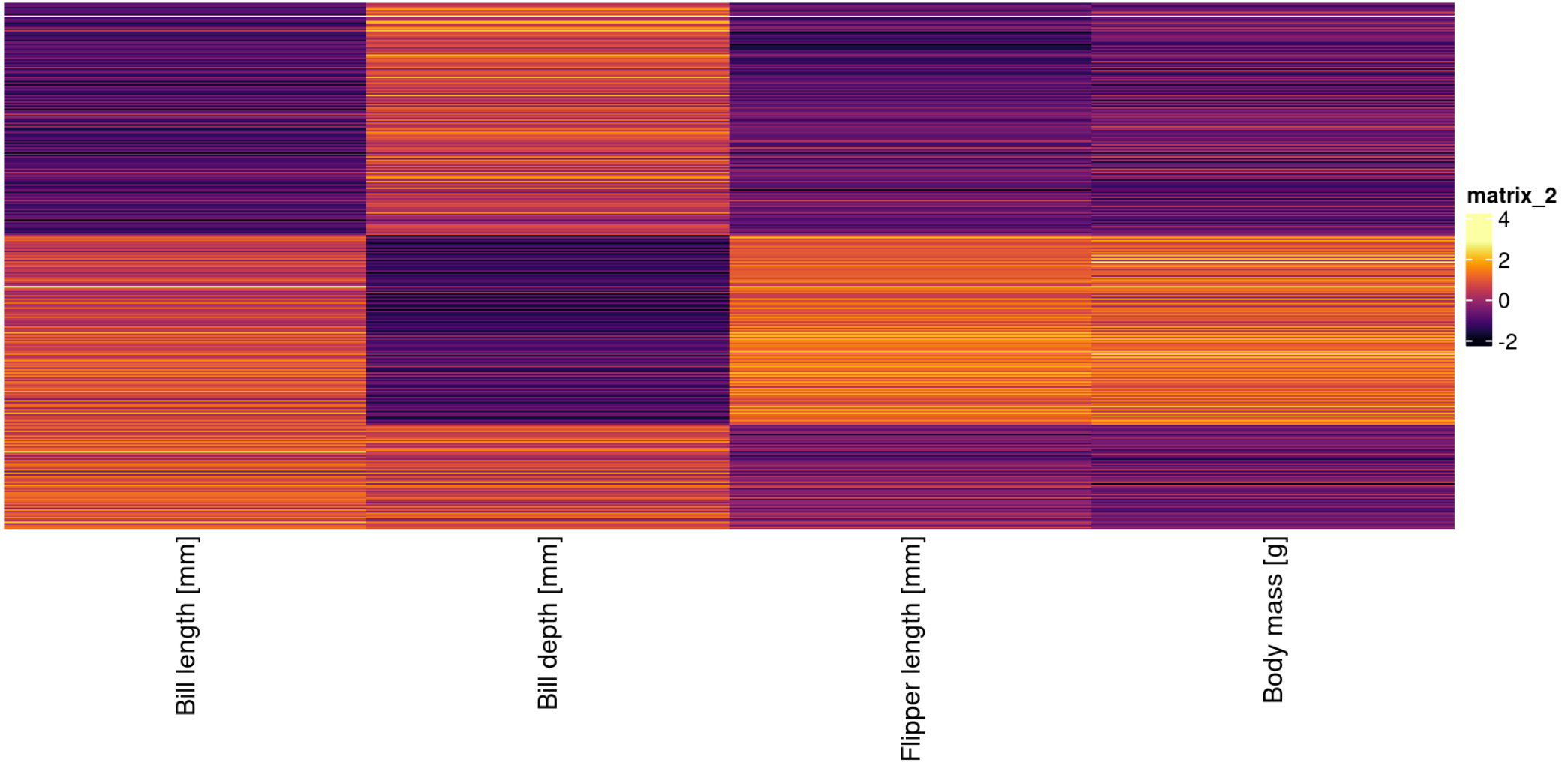


Scaled matrix

```
penguins_scaled_mat <-
  penguins_mat %>%
  scale

penguins_scaled_mat[1:3, 1:3]
```

	bill_length_mm	bill_depth_mm	flipper_length_mm
sympathetic_emu	-0.8832047	0.7843001	-1.4162715
lethargic_horseshoecrab	-0.8099390	0.1260033	-1.0606961
enarthrodial_hogget	-0.6634077	0.4298326	-0.4206603

Scaled matrix



Running PCA

```
1 penguins_pca <- prcomp(penguins_scaled_mat)
```

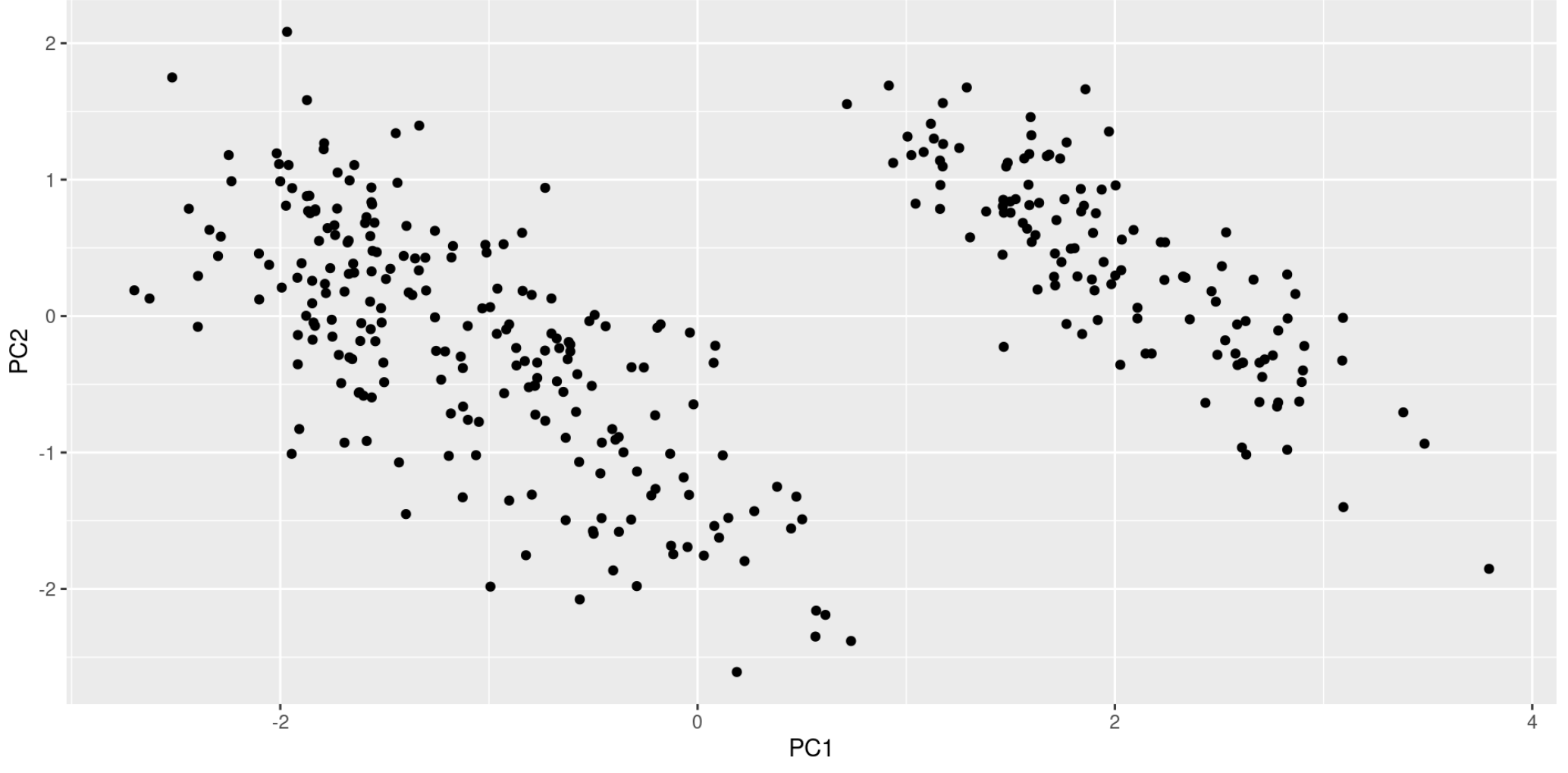
Running PCA

```
1 penguins_pca <- prcomp(penguins_scaled_mat)
2
3 str(penguins_pca)
```

```
List of 5
 $ sdev    : num [1:4] 1.659 0.879 0.604 0.329
 $ rotation: num [1:4, 1:4] 0.455 -0.4 0.576 0.548 -0.597 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "bill_length_mm" "bill_depth_mm" "flipper_length_mm"
"body_mass_g"
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
 $ center  : Named num [1:4] 1.82e-16 3.95e-16 -8.36e-16 8.27e-17
  ..- attr(*, "names")= chr [1:4] "bill_length_mm" "bill_depth_mm"
"flipper_length_mm" "body_mass_g"
 $ scale   : Named num [1:4] 5.46 1.97 14.06 801.95
  ..- attr(*, "names")= chr [1:4] "bill_length_mm" "bill_depth_mm"
"flipper_length_mm" "body_mass_g"
 $ x       : num [1:342, 1:4] -1.84 -1.3 -1.37 -1.88 -1.91 ...
  ..- attr(*, "dimnames")=List of 2
```

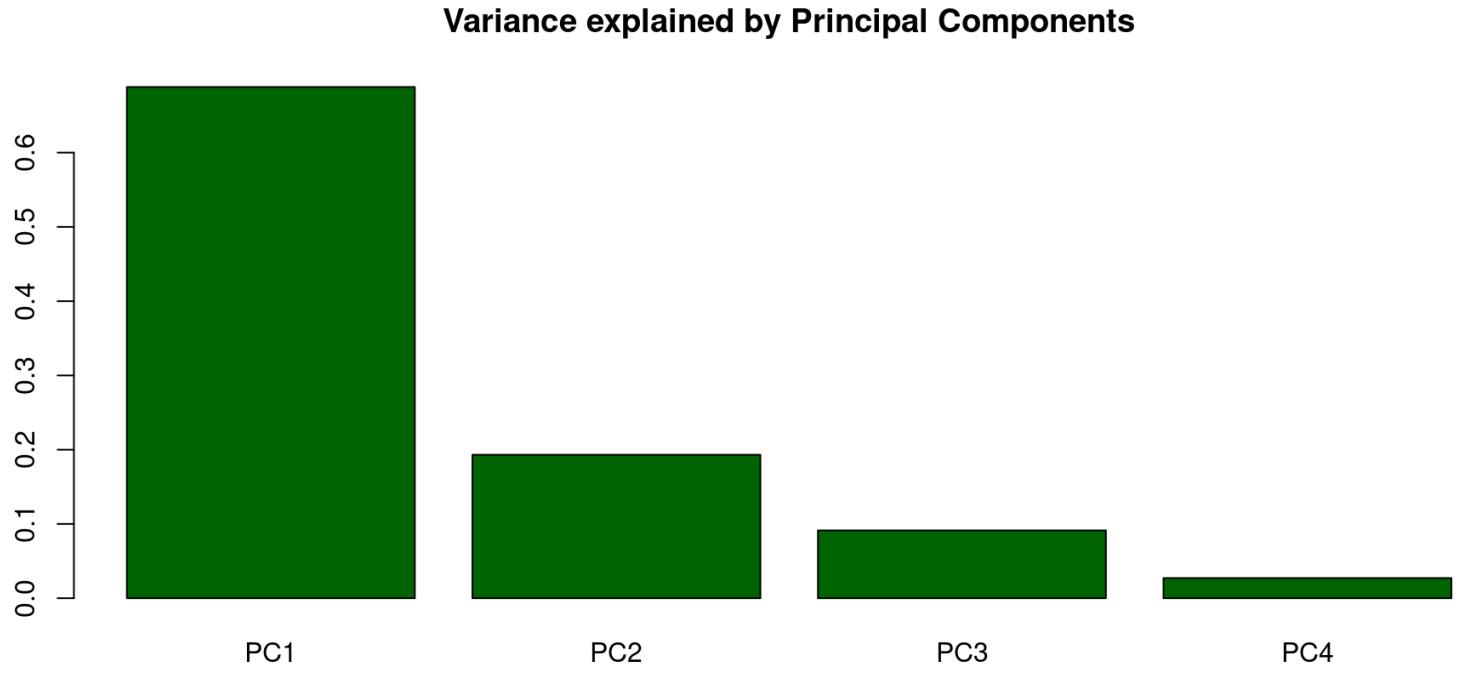
Running PCA

```
1 penguins_pca <- prcomp(penguins_scaled_mat)
```

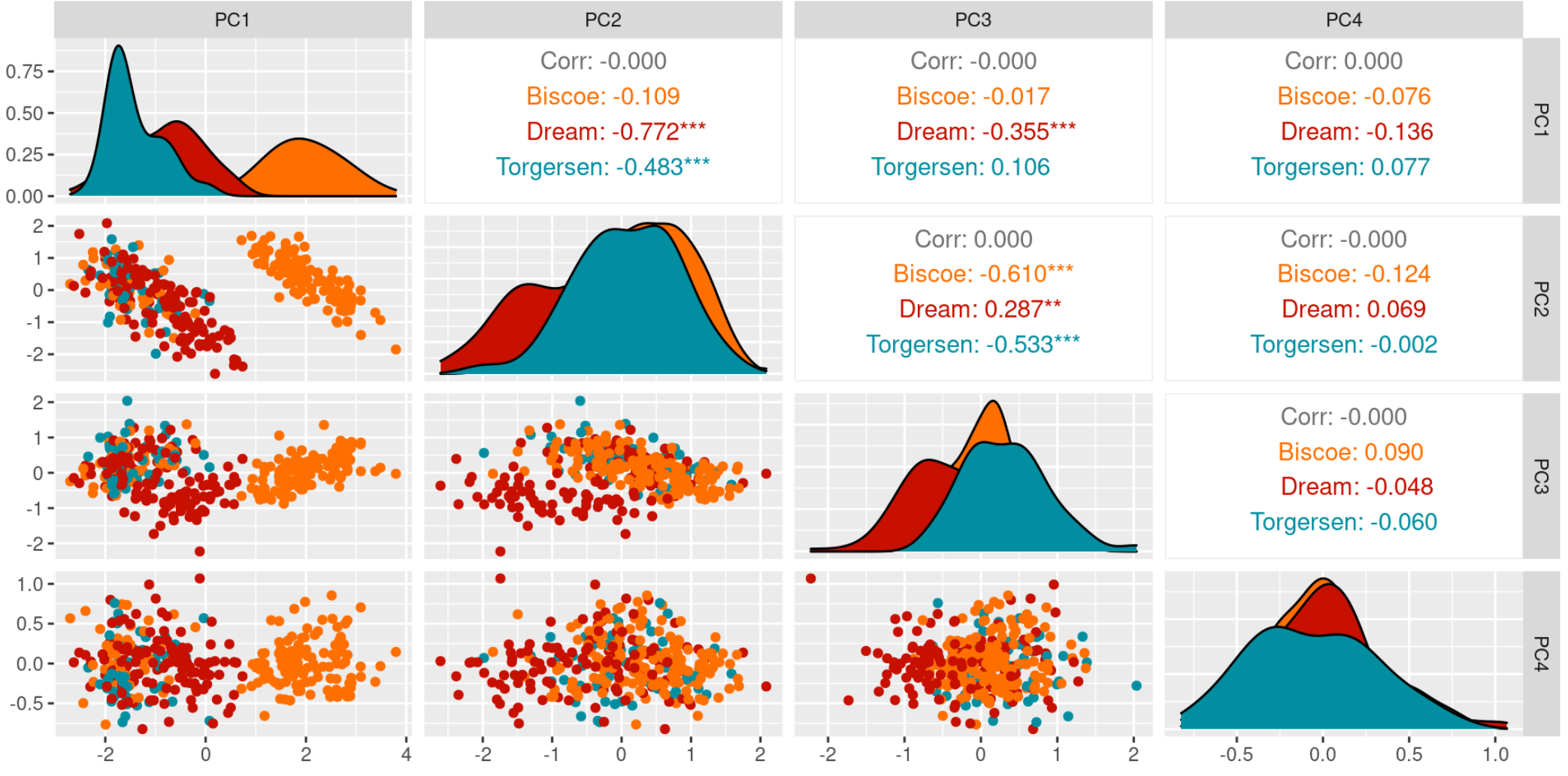


PCA - variance explained

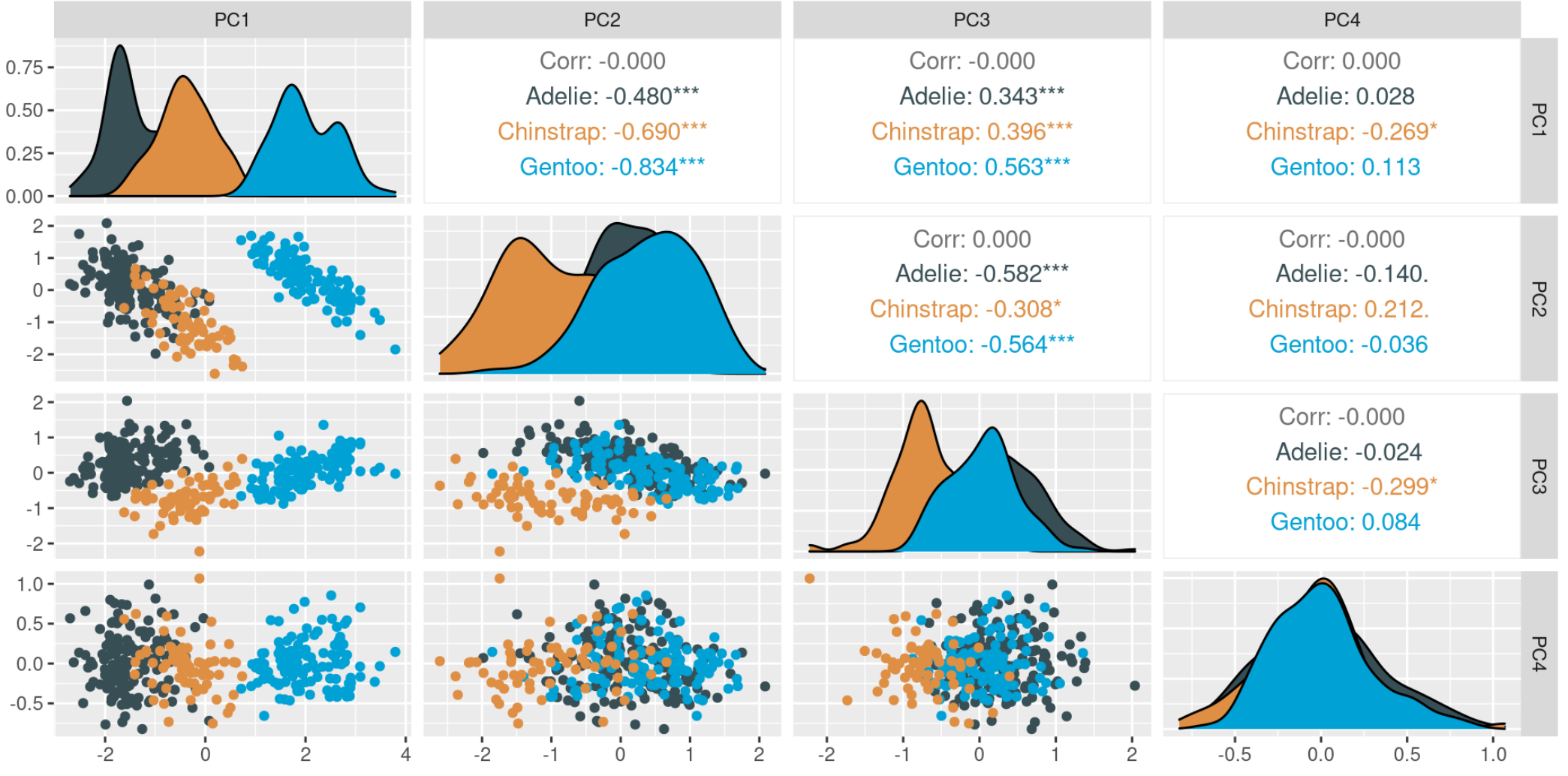
```
1 var_expl <- penguins_pca$sdev^2 / sum(penguins_pca$sdev^2)
2 names(var_expl) <- paste0("PC", 1:length(var_expl))
3
4 barplot(var_expl, col = "darkgreen",
5         main = "Variance explained by Principal Components")
```



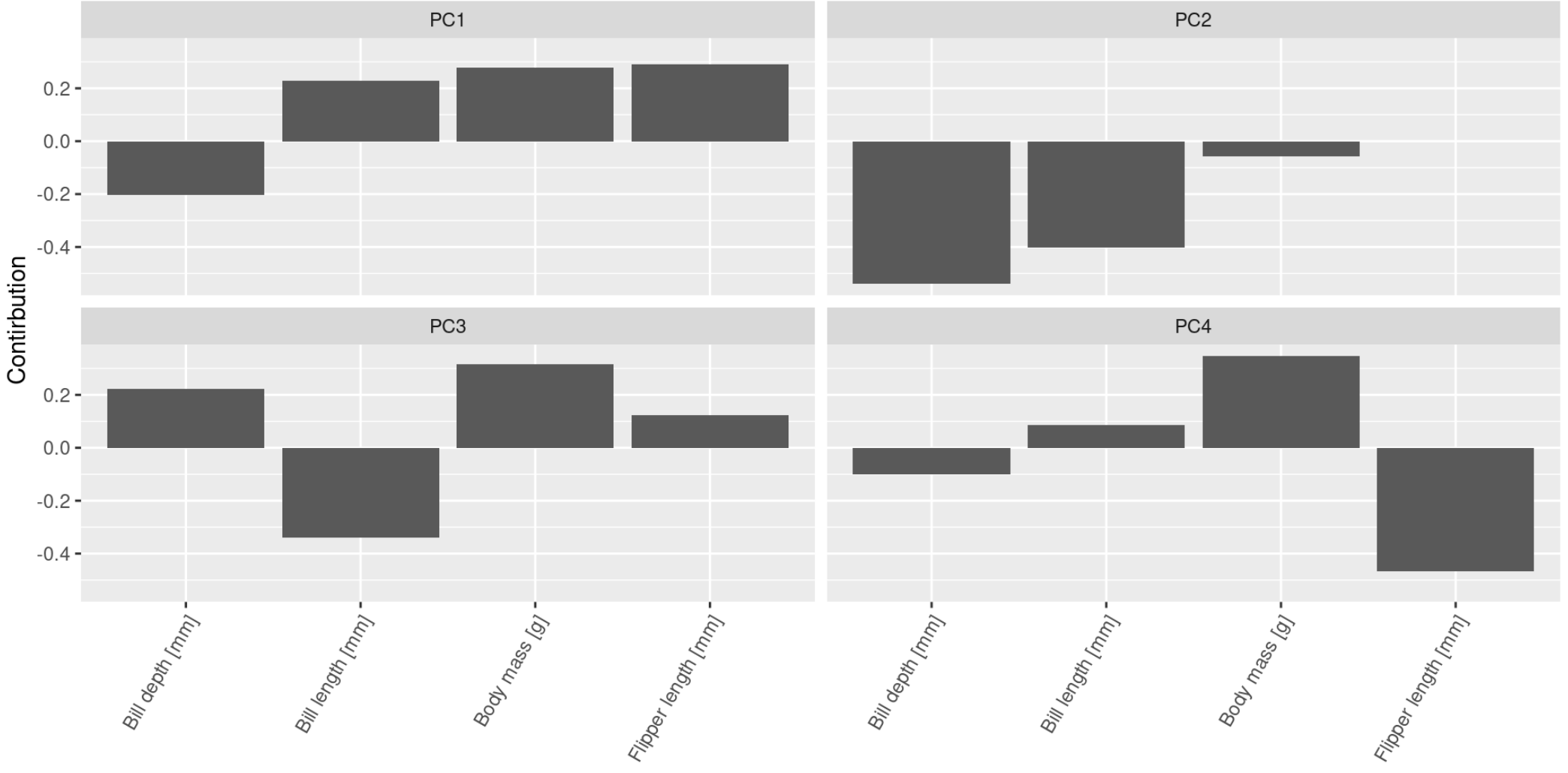
PCA representation



PCA representation

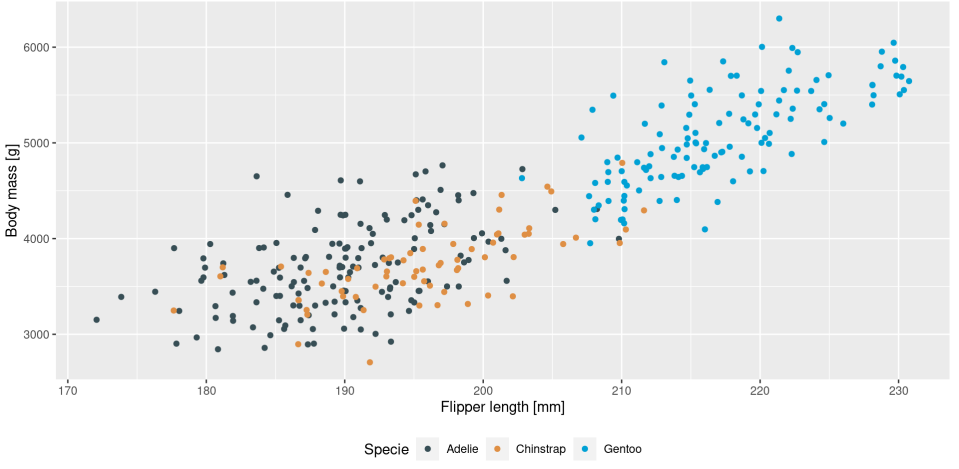


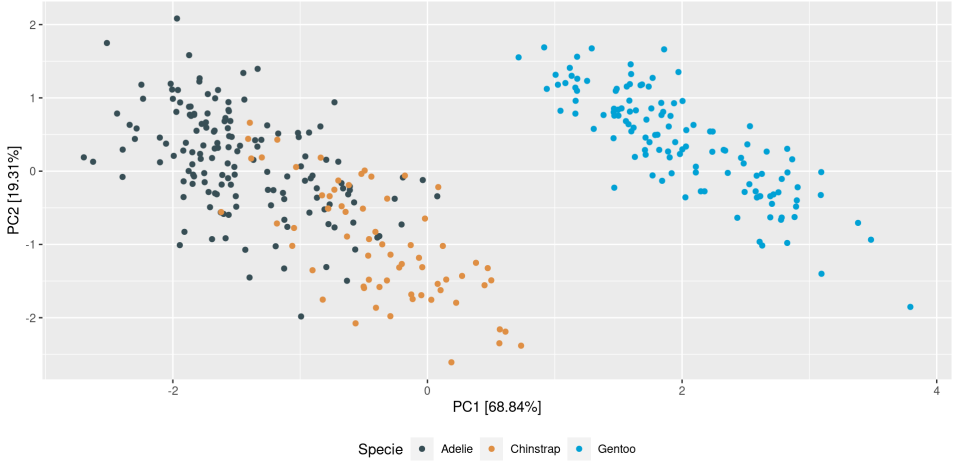
PCA - exploration



PCA - exploration

PCA vs selected variables





UMAP

UMAP

```
penguins_umap <-
  umap::umap(penguins_scaled_mat)
```

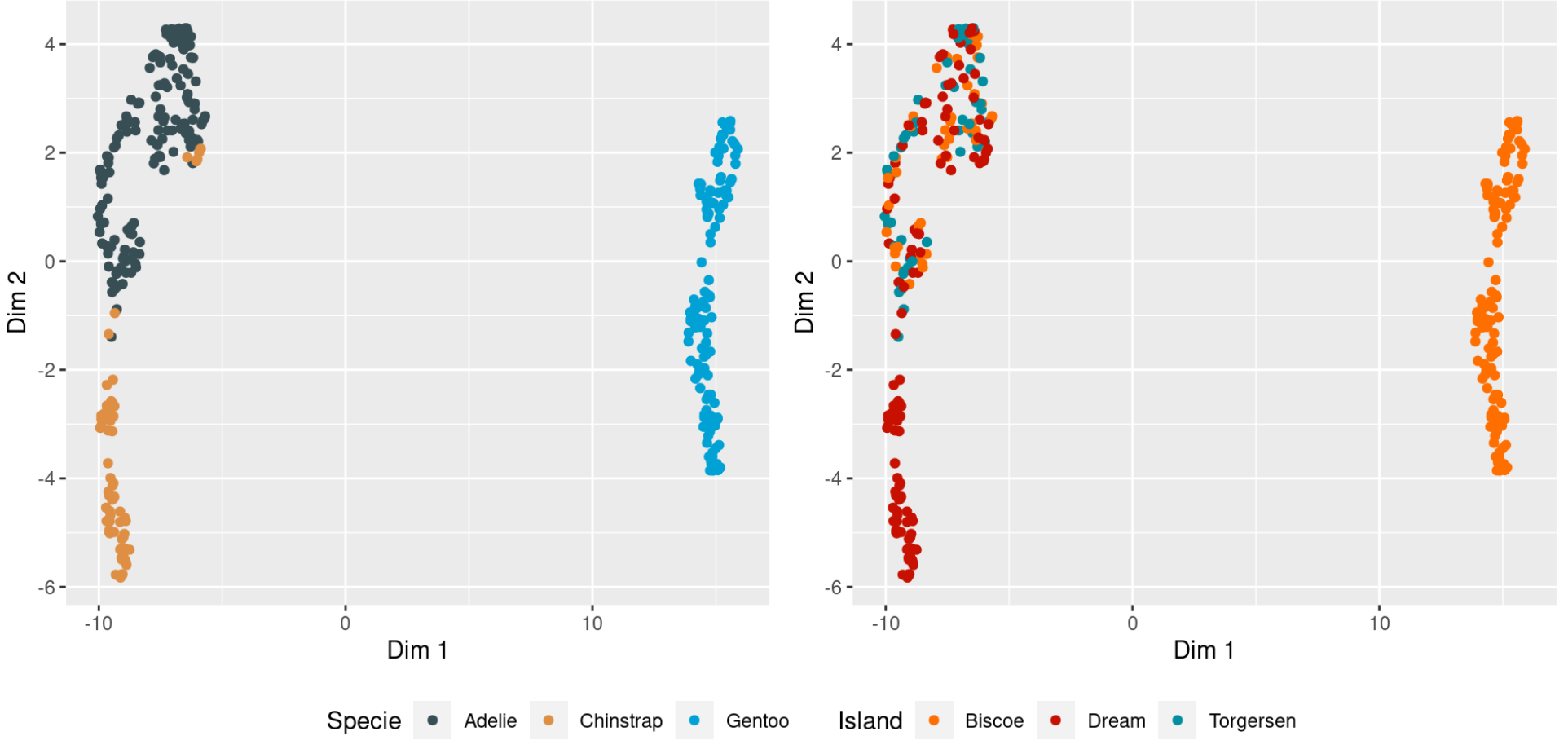
UMAP

```r
penguins_umap <-
  umap::umap(penguins_scaled_mat)

str(penguins_umap)
```

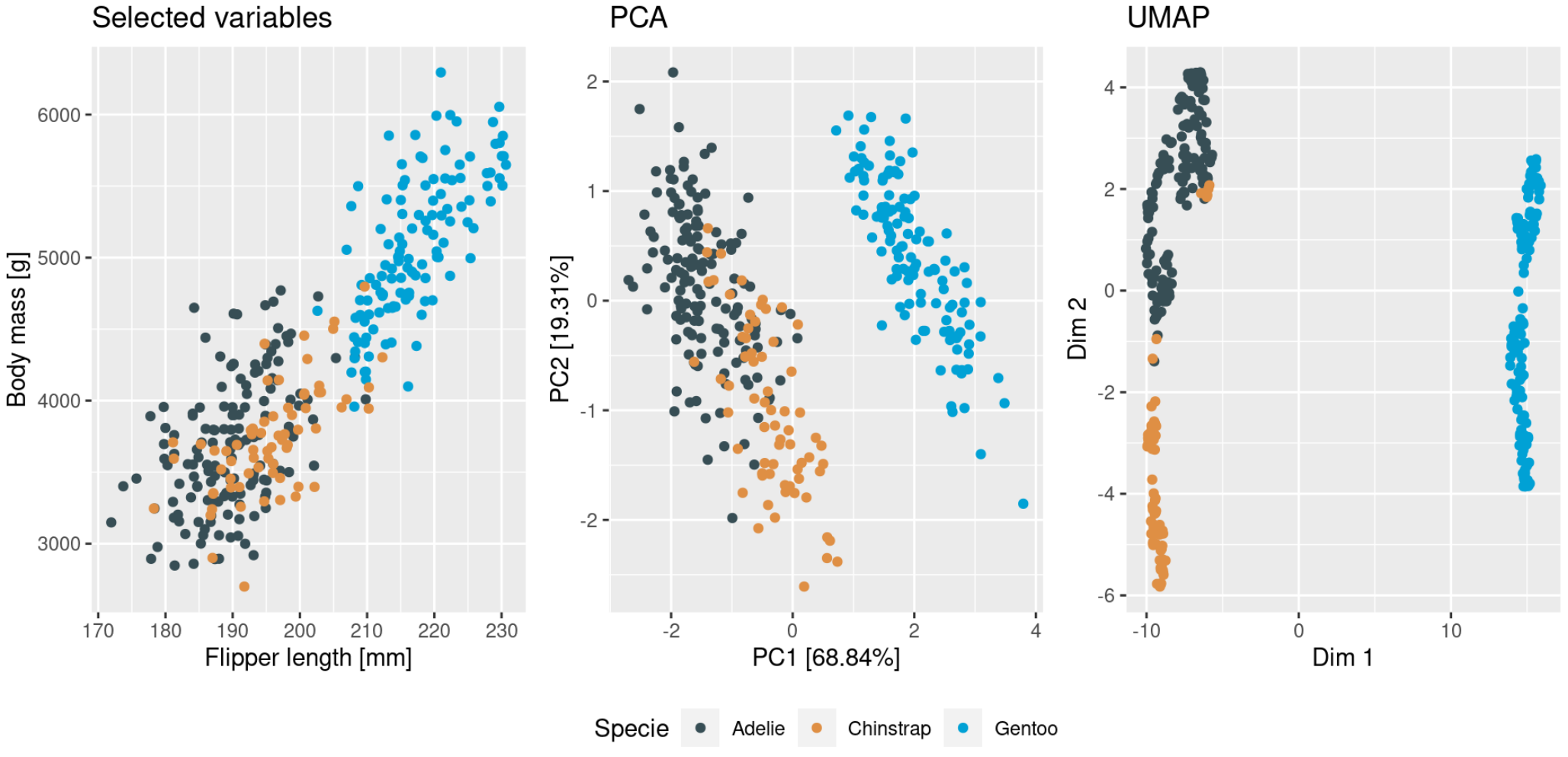
```
List of 4
 $ layout: num [1:342, 1:2] -7.64 -6.48 -6.12 -8.86 -9.95 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:342] "sympathetic_emu" "lethargic_horseshoecrab"
"enarthrodial_hogget" "postindustrial_bunting" ...
  .. ..$ : NULL
 $ data  : num [1:342, 1:4] -0.883 -0.81 -0.663 -1.323 -0.847 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:342] "sympathetic_emu" "lethargic_horseshoecrab"
"enarthrodial_hogget" "postindustrial_bunting" ...
  .. ..$ : chr [1:4] "bill_length_mm" "bill_depth_mm" "flipper_length_mm"
"body_mass_g"
  ..- attr(*, "scaled:center")= Named num [1:4] 43.9 17.2 200.9 4201.8
  .. ..- attr(*, "names")= chr [1:4] "bill_length_mm" "bill_depth_mm"
"flipper_length_mm" "body_mass_g"
```

UMAP

```
penguins_umap <-
  umap::umap(penguins_scaled_mat)
```



Comparison


Selected variables
Body mass [g]
6000
5000
4000
3000
170
180
190
200
210
220
230
Flipper length [mm]
PCA
PC2 [19.31%]
2
1
0
-1
-2
-2
0
2
4
PC1 [68.84%]
UMAP
Dim 2
4
2
0
-2
-4
-6
-10
0
10
Dim 1
Specie
Adelie
Chinstrap
Gentoo

K-means

K-means

```
penguins_kmeans <-
  kmeans(penguins_scaled_mat, centers = 3)
```

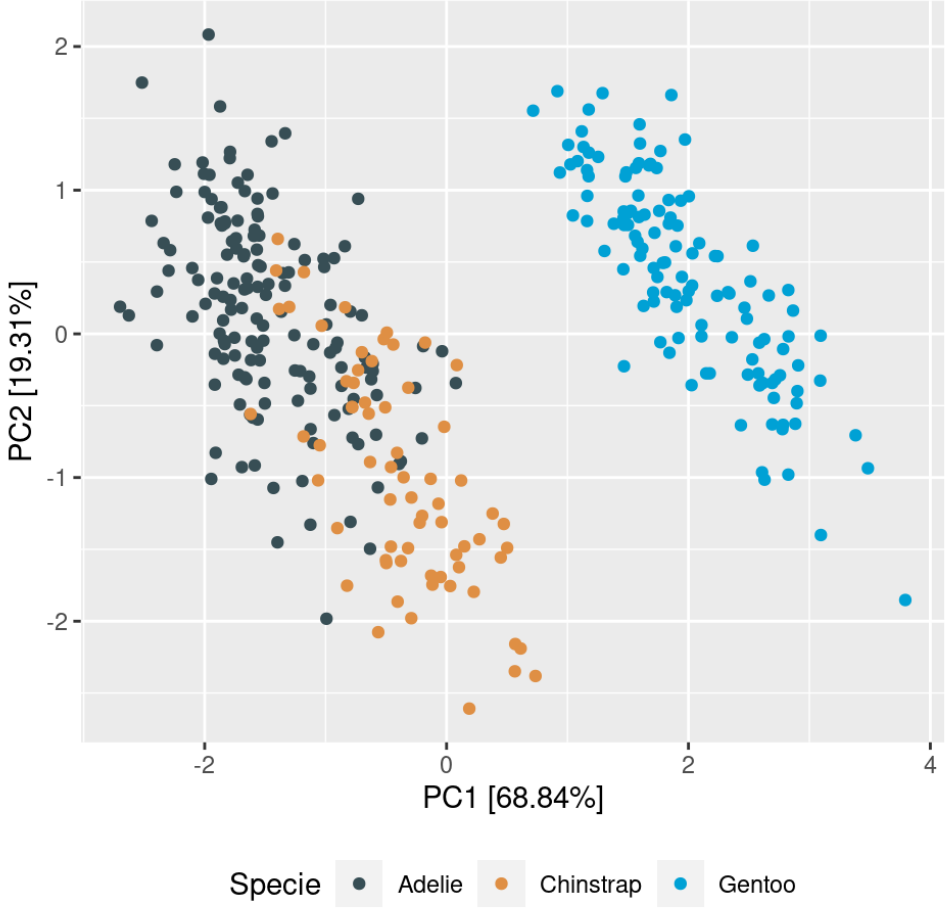
K-means

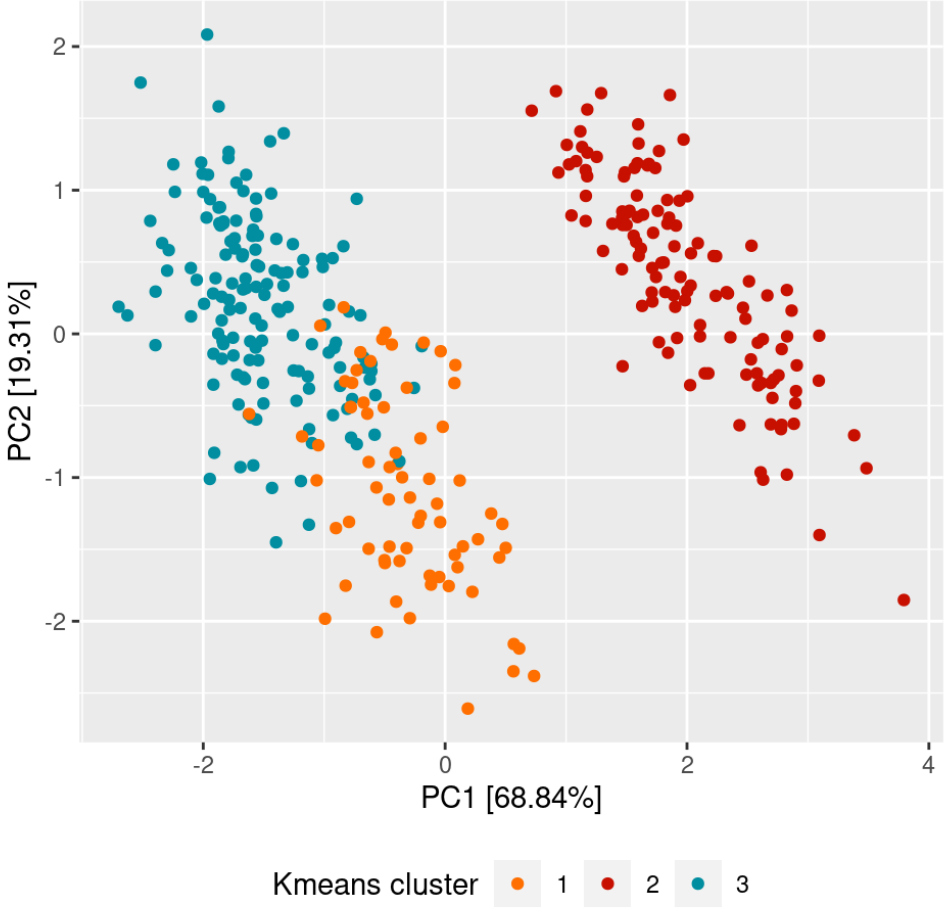
```r
penguins_kmeans <-
  kmeans(penguins_scaled_mat, centers = 3)

str(penguins_kmeans)
```

```
List of 9
 $ cluster     : Named int [1:342] 3 3 3 3 3 3 3 3 3 3 ...
  ..- attr(*, "names")= chr [1:342] "sympathetic_emu"
"lethargic_horseshoecrab" "enarthrodial_hogget" "postindustrial_bunting" ...
 $ centers     : num [1:3, 1:4] 0.89 0.656 -0.972 0.756 -1.098 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : chr [1:4] "bill_length_mm" "bill_depth_mm" "flipper_length_mm"
"body_mass_g"
 $ totss       : num 1364
 $ withinss    : num [1:3] 81.6 143.2 155.3
 $ tot.withinss: num 380
 $ betweenss   : num 984
 $ size        : int [1:3] 71 123 148
 $ iter        : int 2
 $ ifault      : int 0
```

K-means

```
penguins_kmeans <-
  kmeans(penguins_scaled_mat, centers = 3)
```





Hierarchical clustering

Hierarchical clustering

```
dist_penguins <- dist(penguins_scaled_mat)
hclust_penguins <- hclust(dist_penguins)
hclusters_penguins <- cutree(hclust_penguins, k = 3)
```

Hierarchical clustering

```r
dist_penguins <- dist(penguins_scaled_mat)
as.matrix(dist_penguins)[1:3, 1:3]
```

```
                        sympathetic_emu lethargic_horseshoecrab
sympathetic_emu               0.0000000                0.7543498
lethargic_horseshoecrab       0.7543498                0.0000000
enarthrodial_hogget           1.2465644                0.9968875
                        enarthrodial_hogget
sympathetic_emu                   1.2465644
lethargic_horseshoecrab           0.9968875
enarthrodial_hogget               0.0000000
```

Hierarchical clustering

```
dist_penguins <- dist(penguins_scaled_mat)
hclust_penguins <- hclust(dist_penguins)
hclusters_penguins <- cutree(hclust_penguins, k = 3)
```

Hierarchical clustering

```
1 hclust_penguins <- hclust(dist_penguins)
2 str(hclust_penguins)
```

```
List of 7
 $ merge      : int [1:341, 1:2] -158 -193 -36 -199 -76 -73 -304 -180 -290
-198 ...
 $ height     : num [1:341] 0.109 0.109 0.126 0.132 0.133 ...
 $ order      : int [1:342] 250 232 154 180 228 184 273 173 246 216 ...
 $ labels     : chr [1:342] "sympathetic_emu" "lethargic_horseshoecrab"
"enarthrodial_hogget" "postindustrial_bunting" ...
 $ method     : chr "complete"
 $ call       : language hclust(d = dist_penguins)
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

Hierarchical clustering

```
dist_penguins <- dist(penguins_scaled_mat)
hclust_penguins <- hclust(dist_penguins)
hclusters_penguins <- cutree(hclust_penguins, k = 3)
```

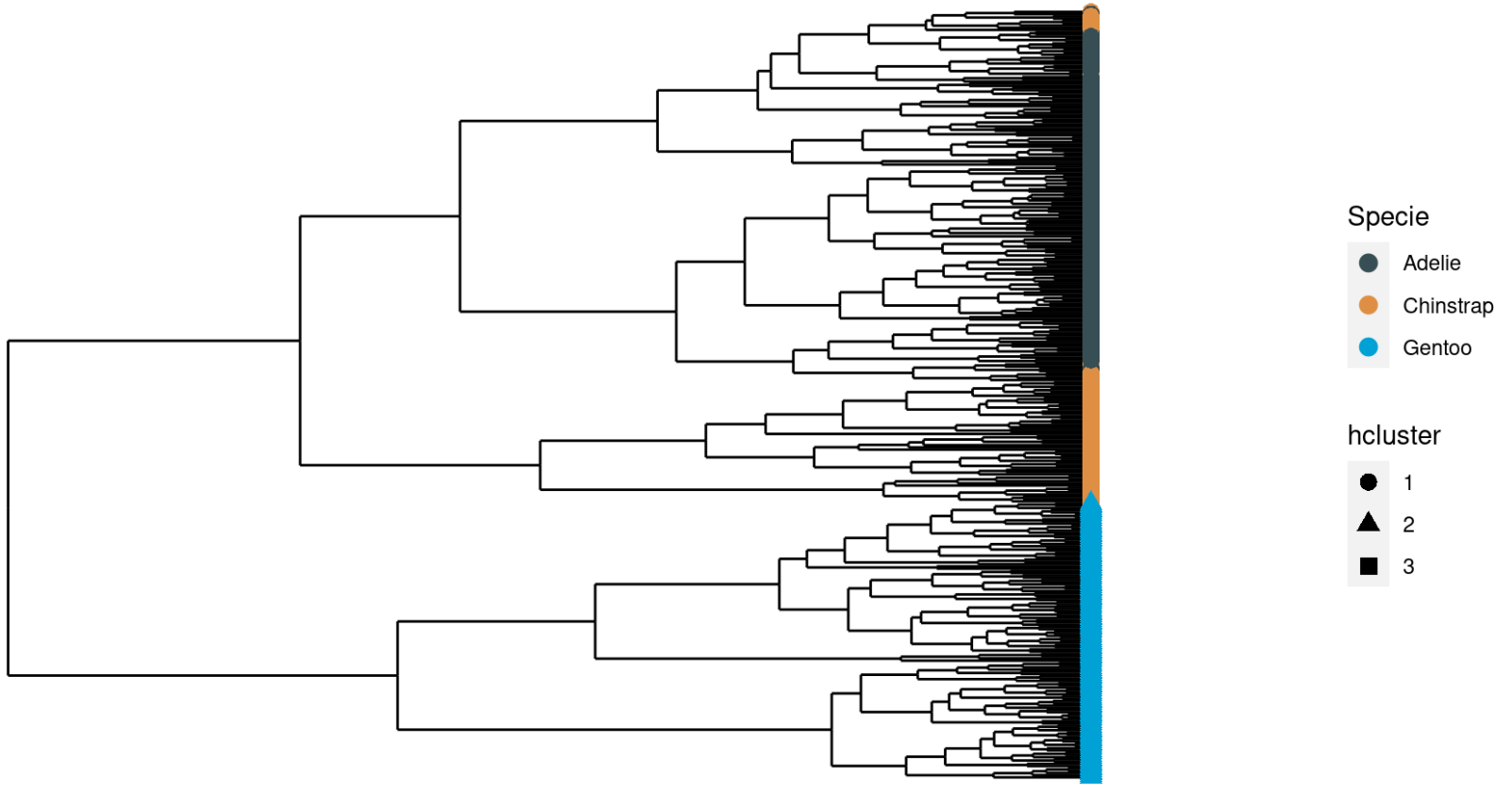
Hierarchical clustering

```r
hclusters_penguins <- cutree(hclust_penguins, k = 3)
str(hclusters_penguins)
```

```
 Named int [1:342] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "names")= chr [1:342] "sympathetic_emu" "lethargic_horseshoecrab"
"enarthrodial_hogget" "postindustrial_bunting" ...
```

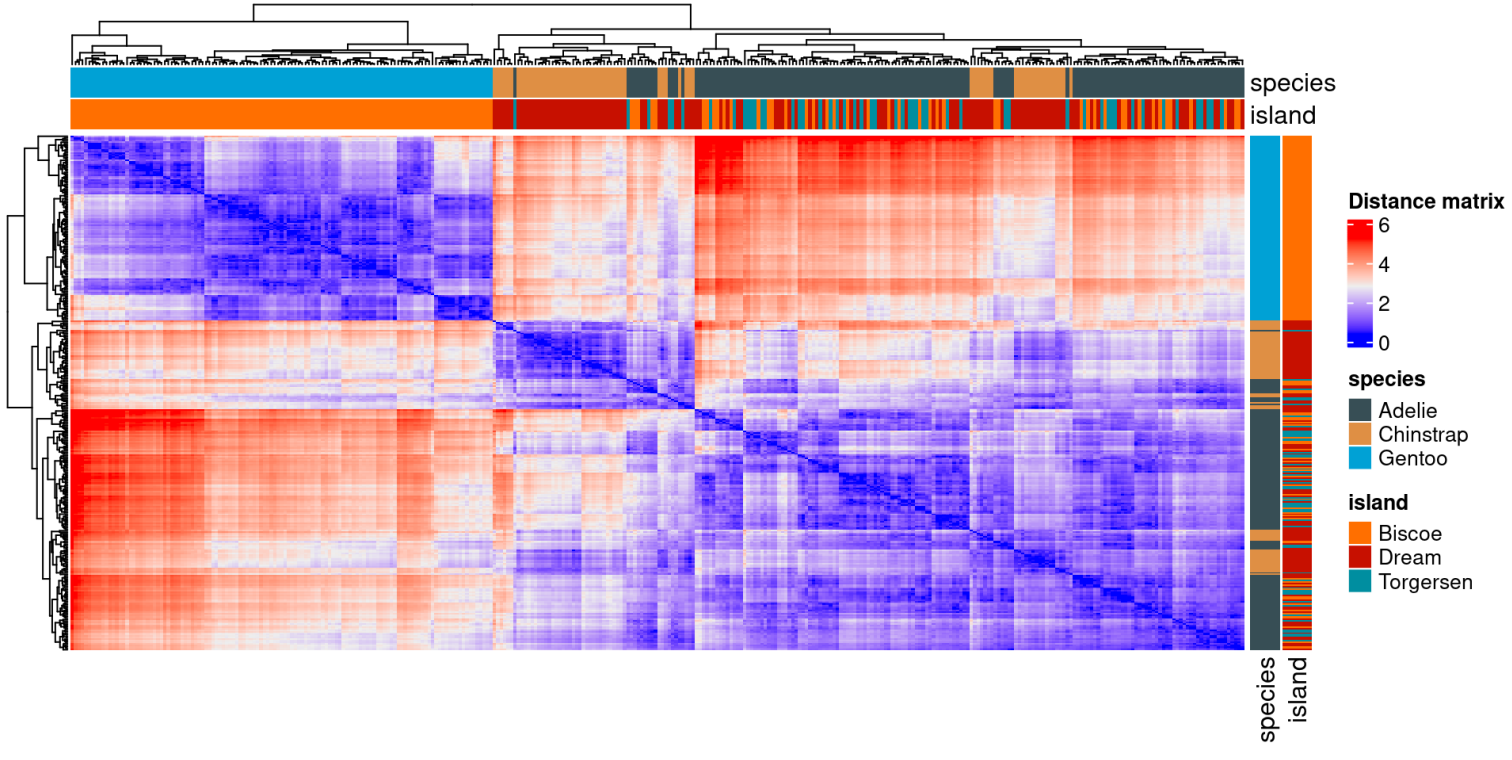
Hierarchical clustering

```
1 dist_penguins <- dist(penguins_scaled_mat)
2 hclust_penguins <- hclust(dist_penguins)
3 hclusters_penguins <- cutree(hclust_penguins, k = 3)
```



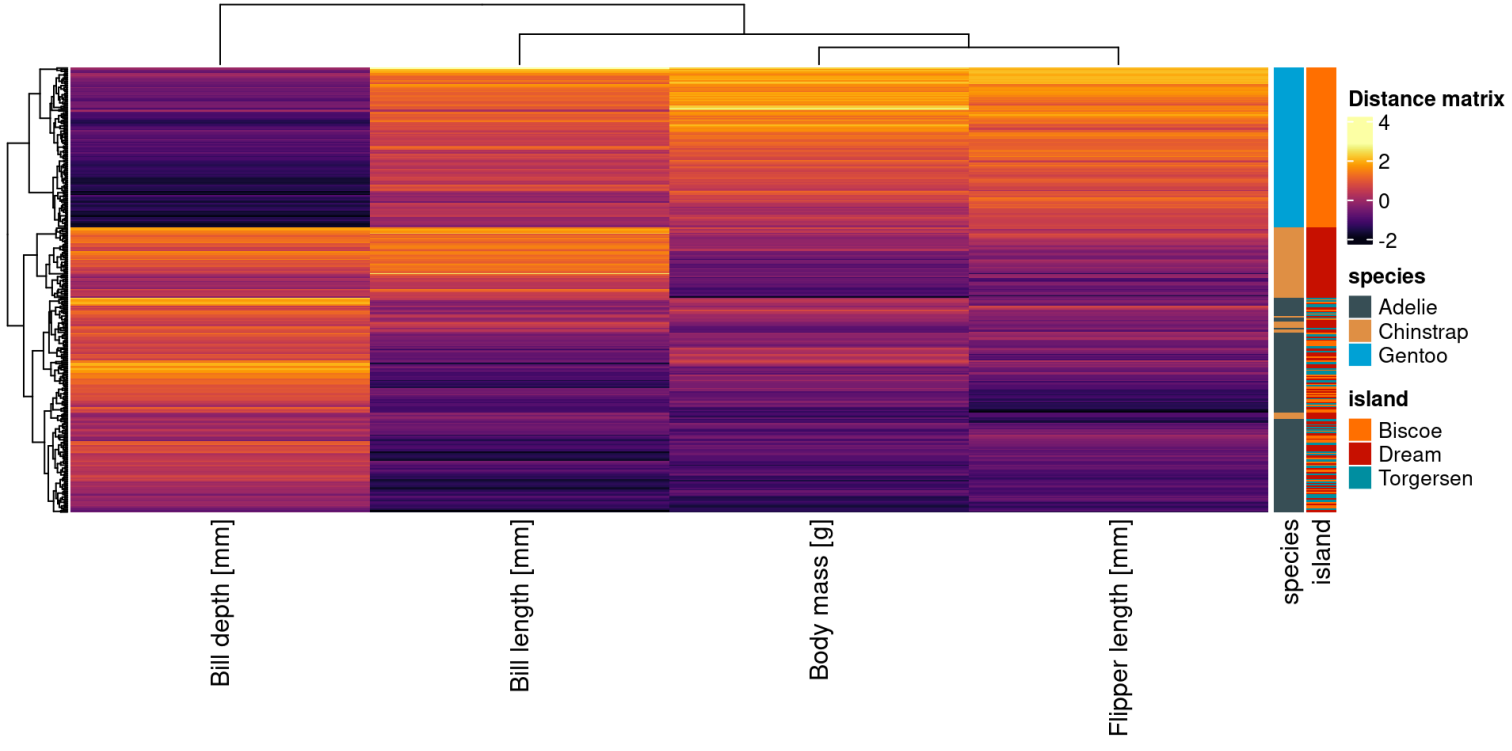
Heatmap

```
ComplexHeatmap::Heatmap(as.matrix(dist_penguins),
                        show_row_names = FALSE,
                        show_column_names = FALSE,
                        name = "Distance matrix",
                        right_annotation = ha_row,
                        top_annotation = ha_column)
```



Heatmap

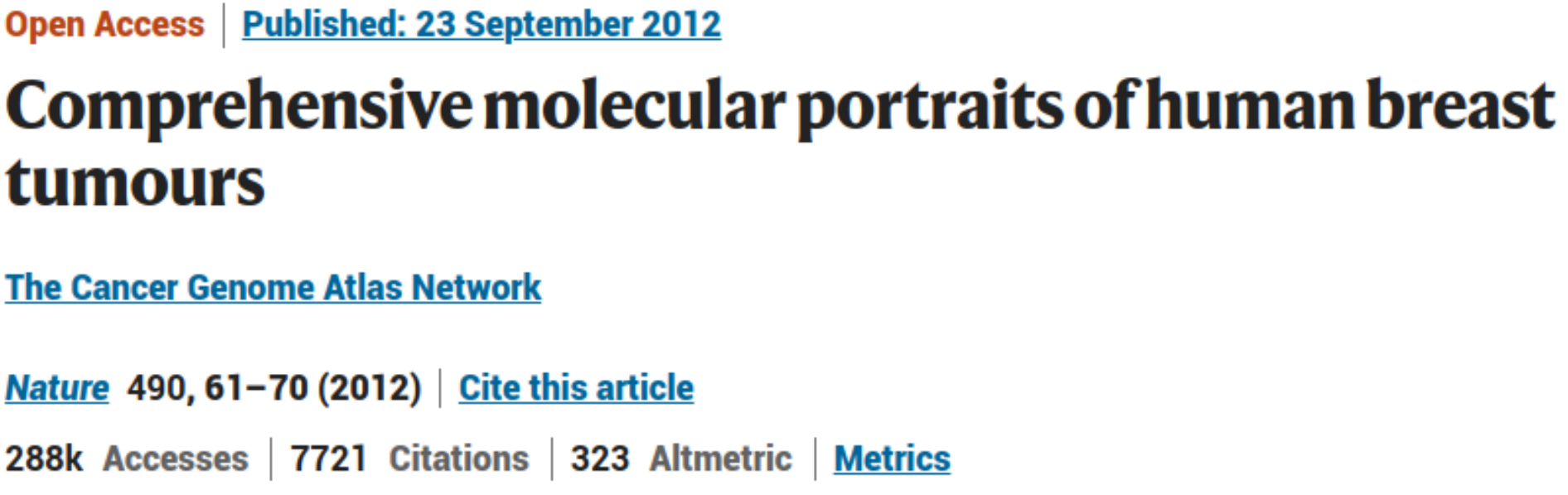
```
ComplexHeatmap::Heatmap(pen_scaled_mat,
                        show_row_names = FALSE,
                        show_column_names = TRUE,
                        col = viridis::inferno(50),
                        name = "Distance matrix",
                        right_annotation = ha_row)
```



Tutorial

Tutorial

In this tutorial we will look at the BRCA dataset from Comprehensive molecular portraits of human breast tumours



Open Access | Published: 23 September 2012

Comprehensive molecular portraits of human breast tumours

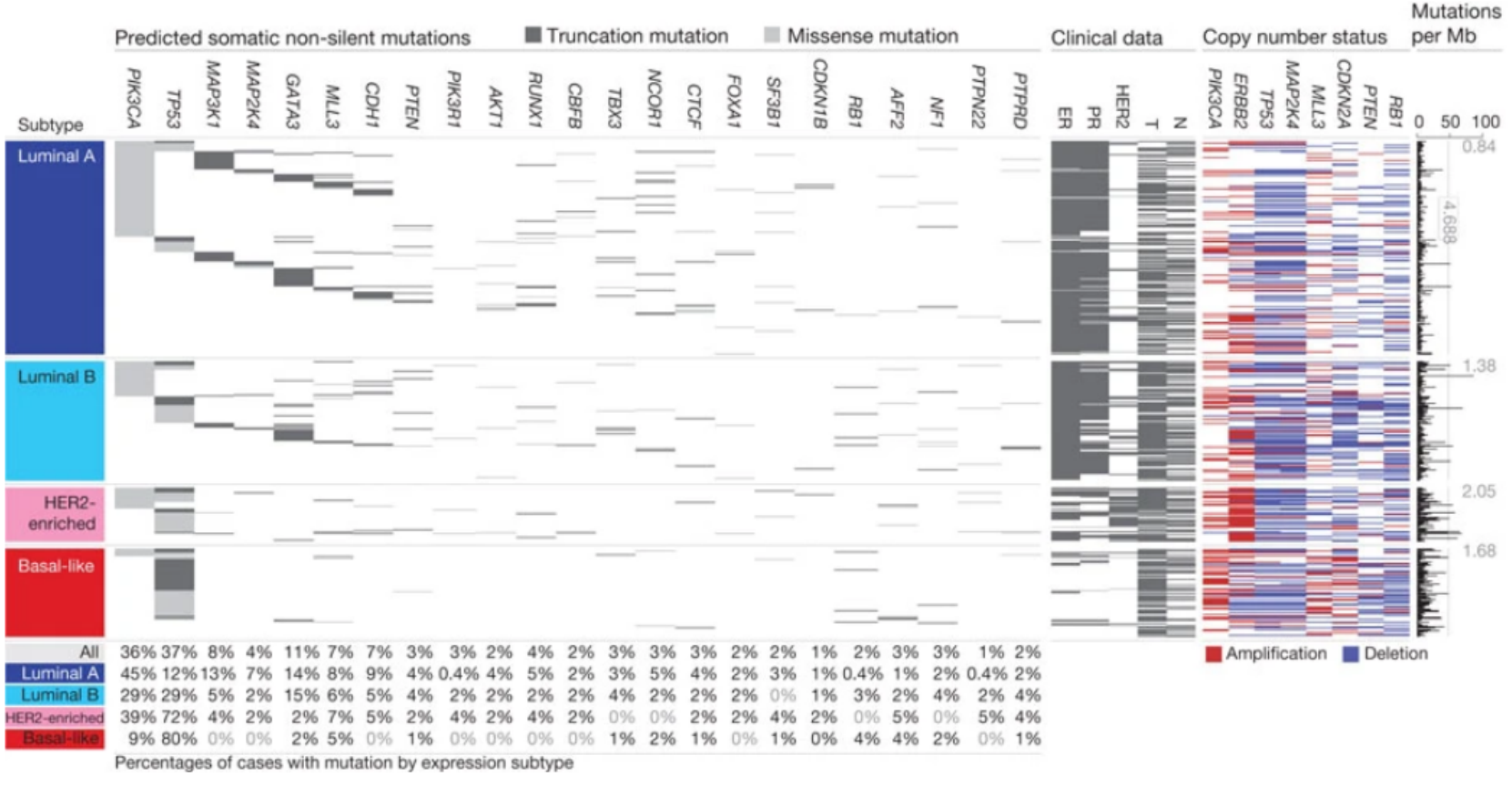
The Cancer Genome Atlas Network

Nature 490, 61–70 (2012) | Cite this article

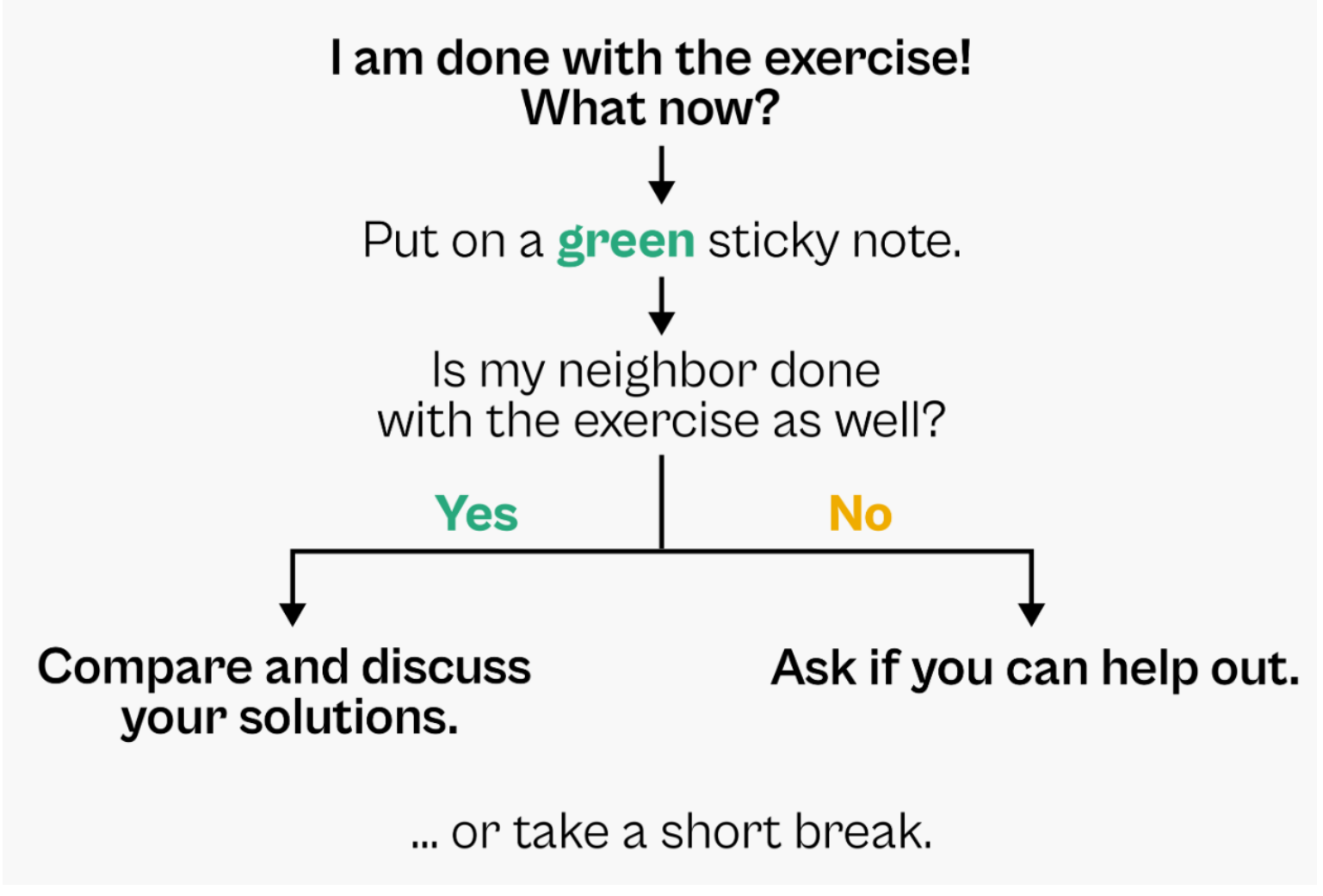
288k Accesses | 7721 Citations | 323 Altmetric | Metrics

Tutorial

In this tutorial we will look at the BRCA dataset from Comprehensive molecular portraits of human breast tumours



Tutorial



Borrowed from Cédric Scherer's RStudio Conf talk on Graphic Design with ggplot2