

Tutorial 3

Principal Component Analysis (PCA)

Clarke van Steenderen
Department of Zoology and Entomology
Rhodes University
South Africa

Clarke.vansteenderen@ru.ac.za



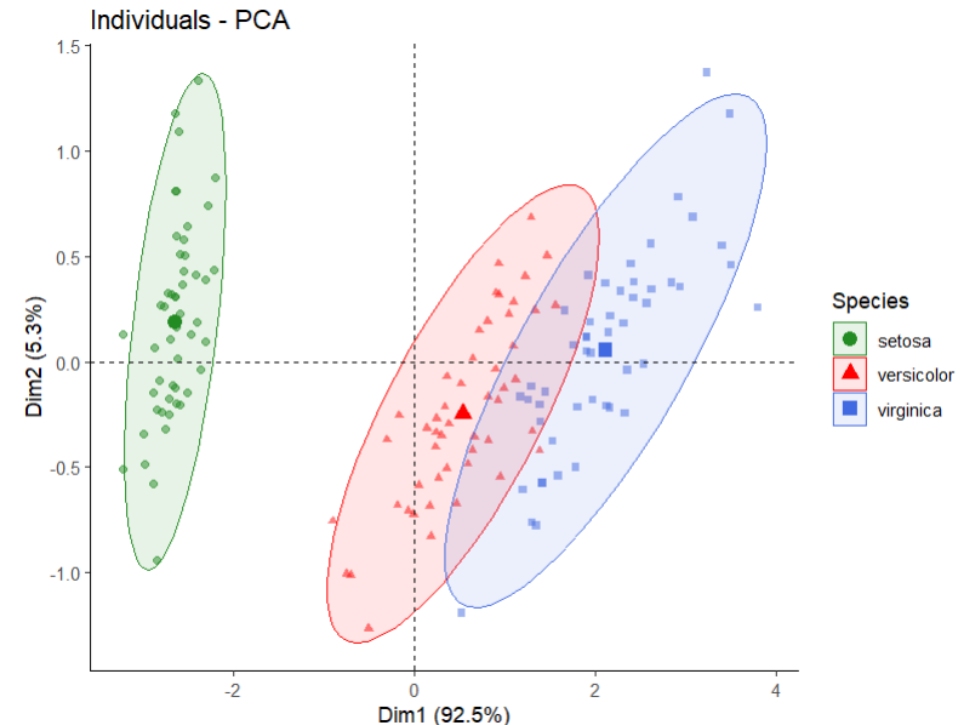
Data sets are often complex, with multiple variables to consider

Making multiple comparisons between all these variables can reach a point of impossibility!

What if there was a way to reduce this complexity, and look at overall patterns in the data?

This is what a **PCA does**

PCAs transform data by means of “dimension reductions”



Let's have a look at the Iris dataset

```
head(iris_data)
```

	sepal_length	sepal_width	petal_length	petal_width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

We have three species and four morphological measurements



Iris Versicolor

Iris Setosa

Iris Virginica

After removing the non-numeric column (species), and checking for NA values, we can run a PCA on the data frame:

```
pca_iris = prcomp(iris_data_clean, scale = TRUE)
```

```
> summary(pca_iris)
```

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	1.7084	0.9560	0.38309	0.14393
Proportion of Variance	0.7296	0.2285	0.03669	0.00518
Cumulative Proportion	0.7296	0.9581	0.99482	1.00000

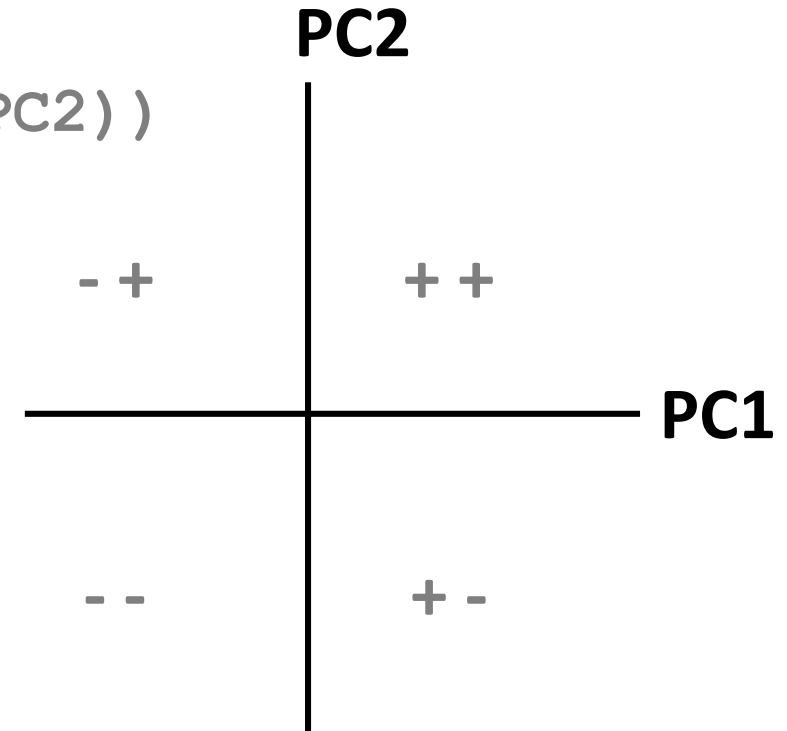
PC1 accounts for 73% of the variation in the data, and PC2 accounts for 23%

Let's look at the loadings:

```
pca_loadings_iris = pca_iris$rotation %>%  
as.data.frame() %>% dplyr::select(c(PC1, PC2))
```

```
> pca_loadings_iris
```

	PC1	PC2
sepal_length	0.5210659	-0.37741762
sepal_width	-0.2693474	-0.92329566
petal_length	0.5804131	-0.02449161
petal_width	0.5648565	-0.06694199



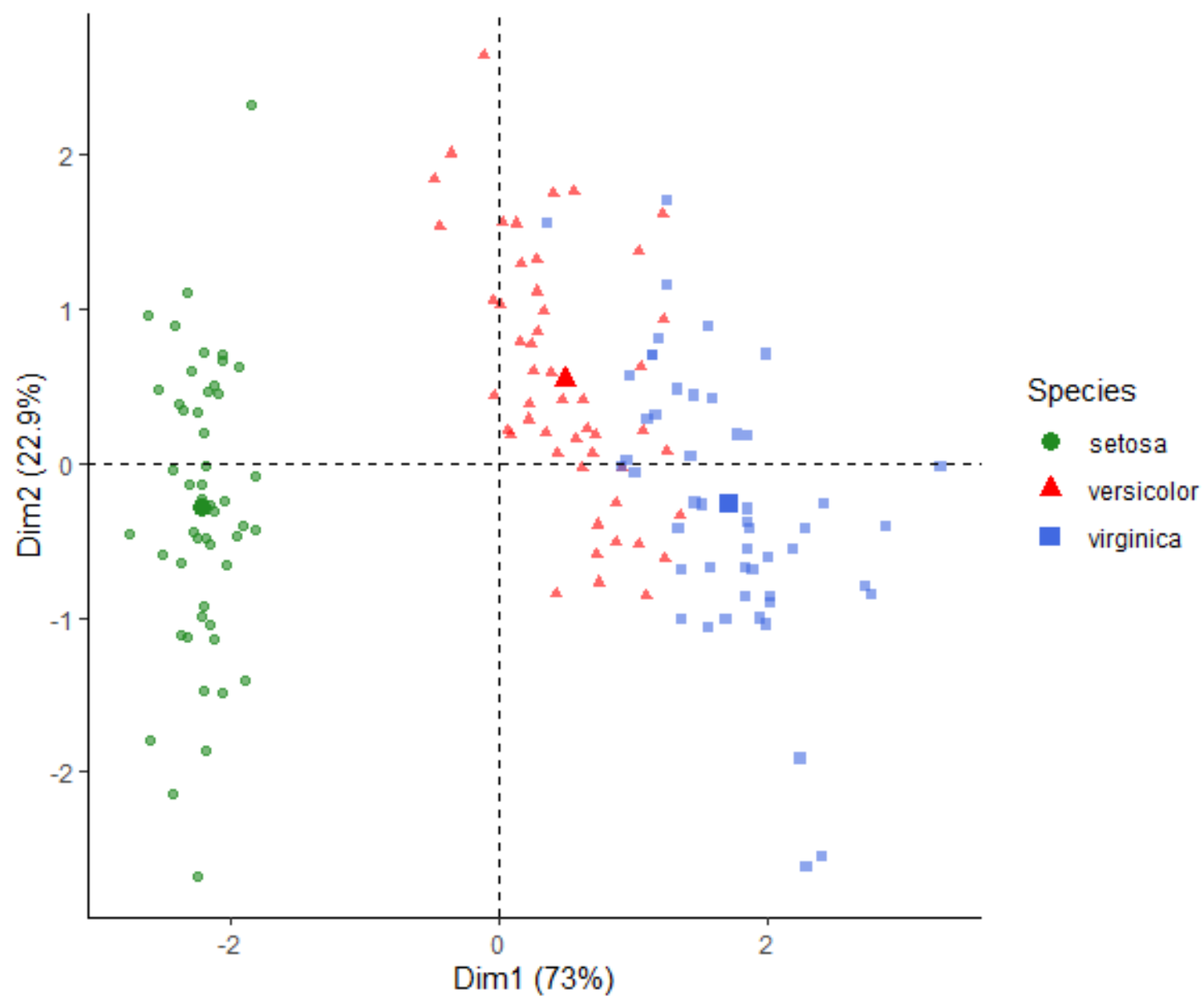
Sepal length, petal length, and petal width all increase along the **+** side of the PC1 axis, while sepal width increases on the **-** side of the PC1 axis

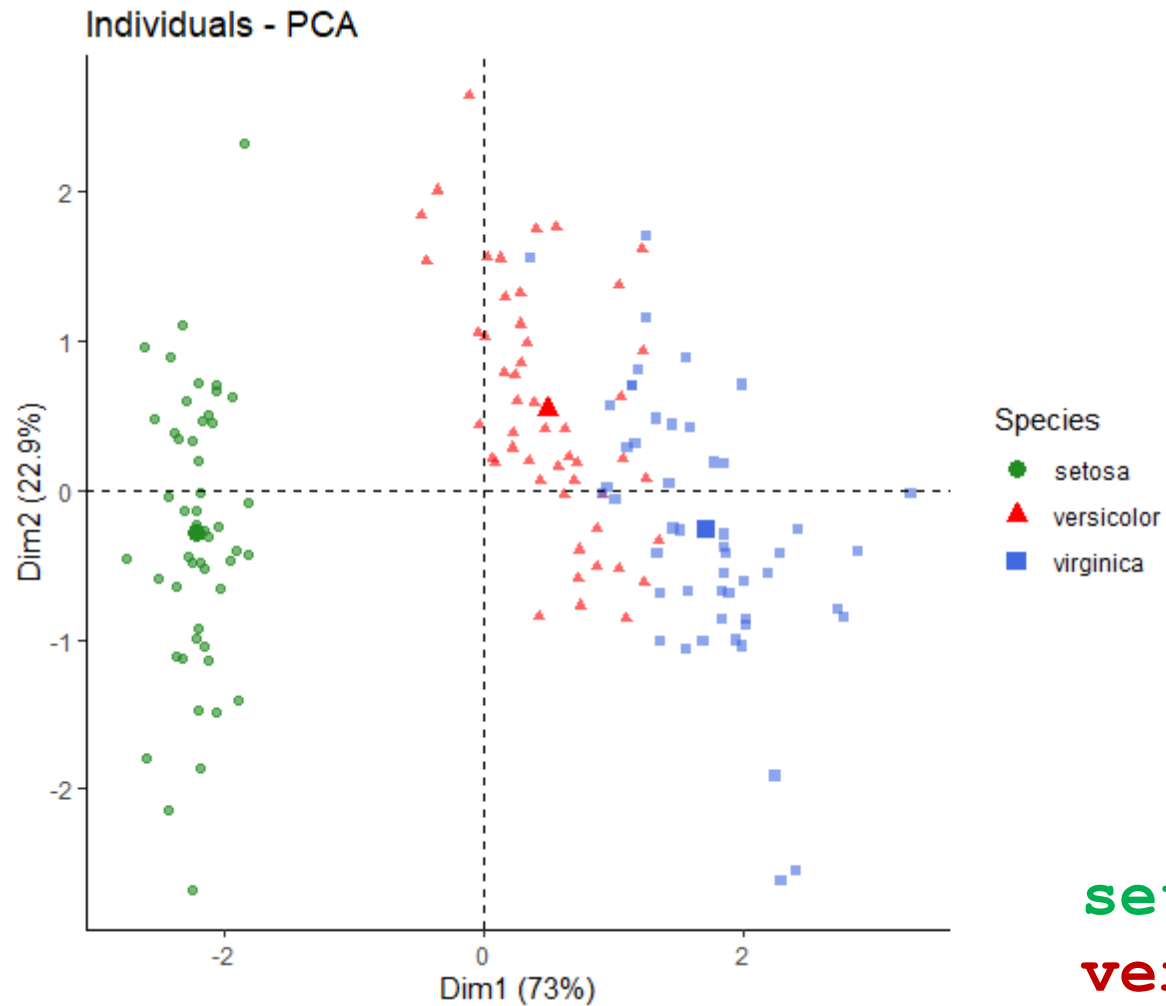
Plot using ggplot:

```
factoextra::fviz_pca_ind(pca_iris, geom = "point",  
  col.ind = iris_data$species,  
  palette = c("forestgreen", "red", "royalblue"),  
  legend.title = "Species", alpha.ind = 0.6) +  
  scale_shape_manual(values=c(19, 17, 15)) +  
  theme_classic()
```

We're specifying three colours and shape values for each of the three Iris species, and we'll use the `fviz_pca_ind()` function in the `factoextra` package

Individuals - PCA



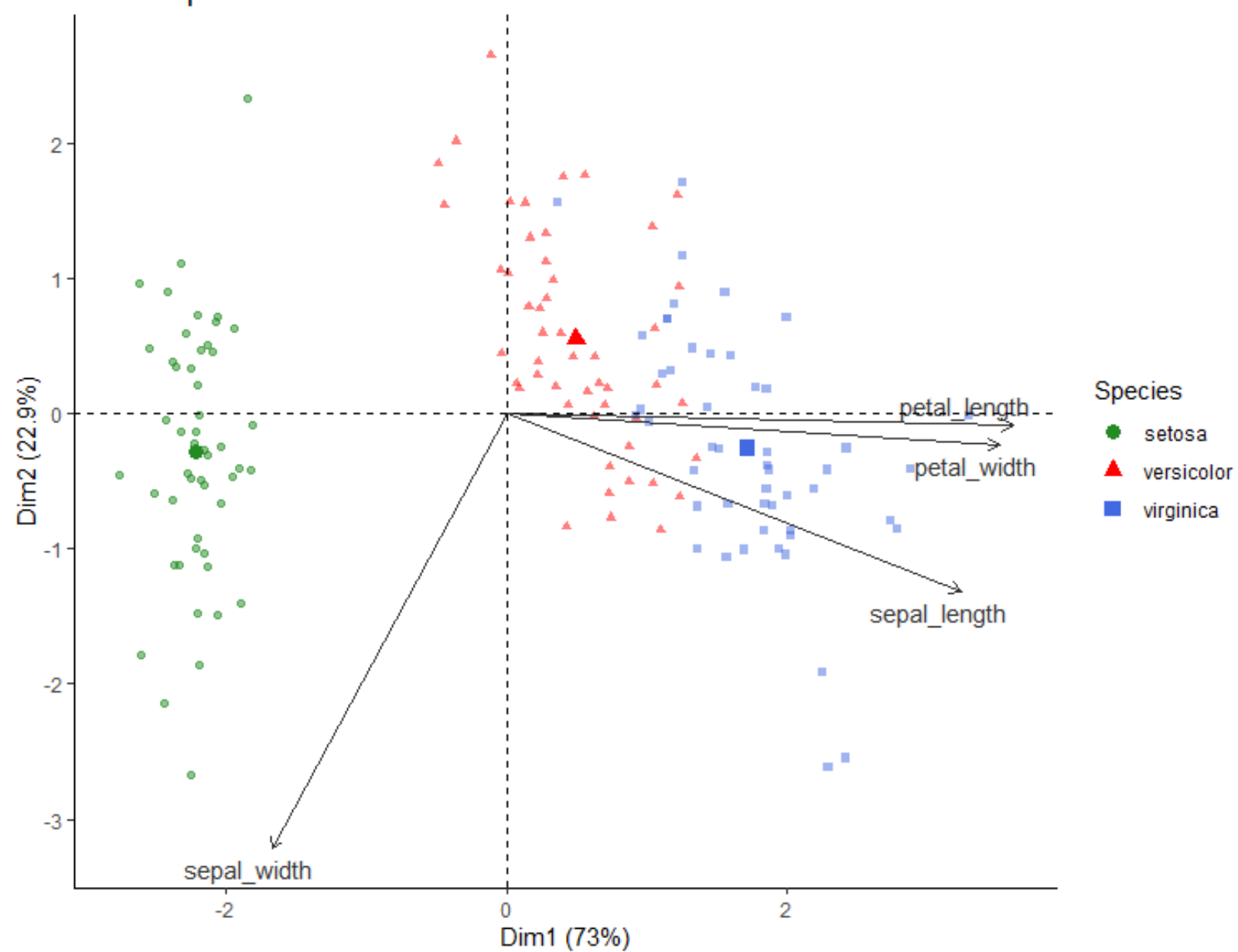


```
> pca_loadings_iris
```

	PC1	PC2
sepal_length	0.5210659	-0.37741762
sepal_width	-0.2693474	-0.92329566
petal_length	0.5804131	-0.02449161
petal_width	0.5648565	-0.06694199

setosa has larger sepal widths than **versicolor** and **virginica**, but **versicolor** and **virginica** have larger sepal lengths, and petal lengths and widths than **setosa**

PCA - Biplot



The `rgl` package offers 3D plots to explore your data

