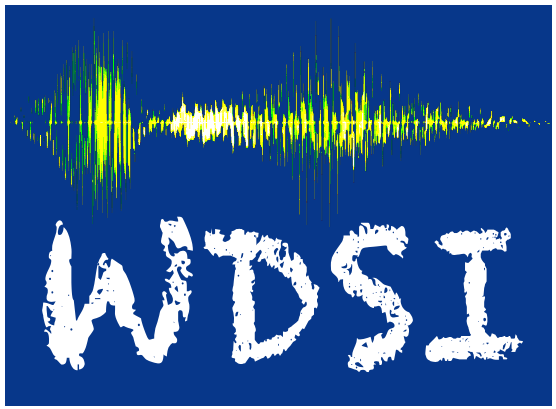
Exploratory Analysis of Multivariate Data

Matt Moores

Department of Statistics, University of Warwick



Warwick Data Science Institute

WDSI Vacation School 2017

1. Multivariate Data
2. Principal Components Analysis (PCA)
3. Clustering

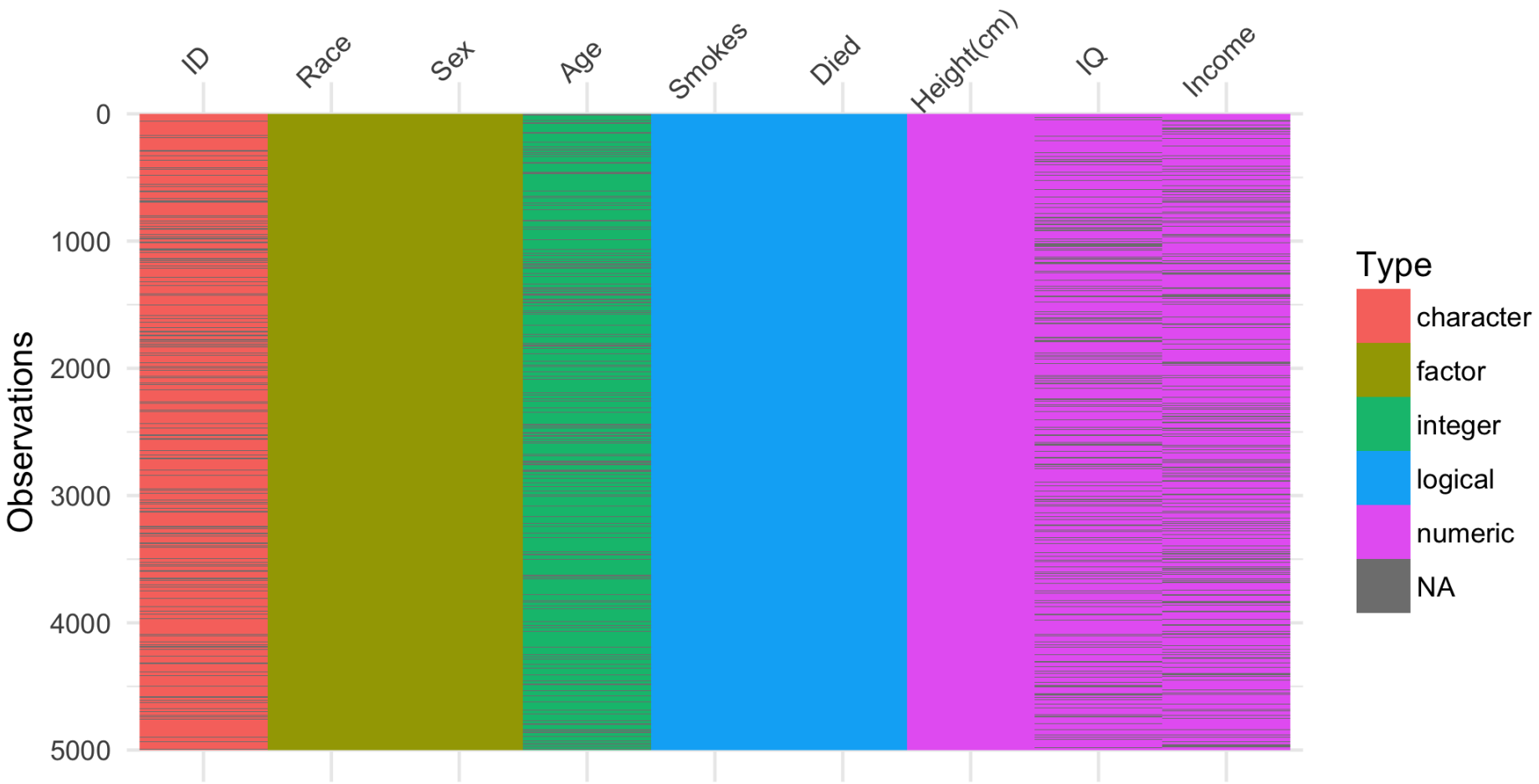
Multivariate Data

Multivariate Data

- High-dimensional (more than 2D)
- Complexity increases combinatorially
- Correlation between columns
- Mixed data types
- Missing data

Visualisation

```
vis_dat(typical_data)
```



Missing Data

- Missing completely at random (MCAR)
  - "the dog ate my homework"
- Missing at random (MAR)
- Missing not at random (MNAR)
  - nonignorable nonresponse

Examples

- Surveys
- Consumer data
- Medical records
- Morphological measurements

Structured data with millions of dimensions:

- *-omics
- Images
- Text

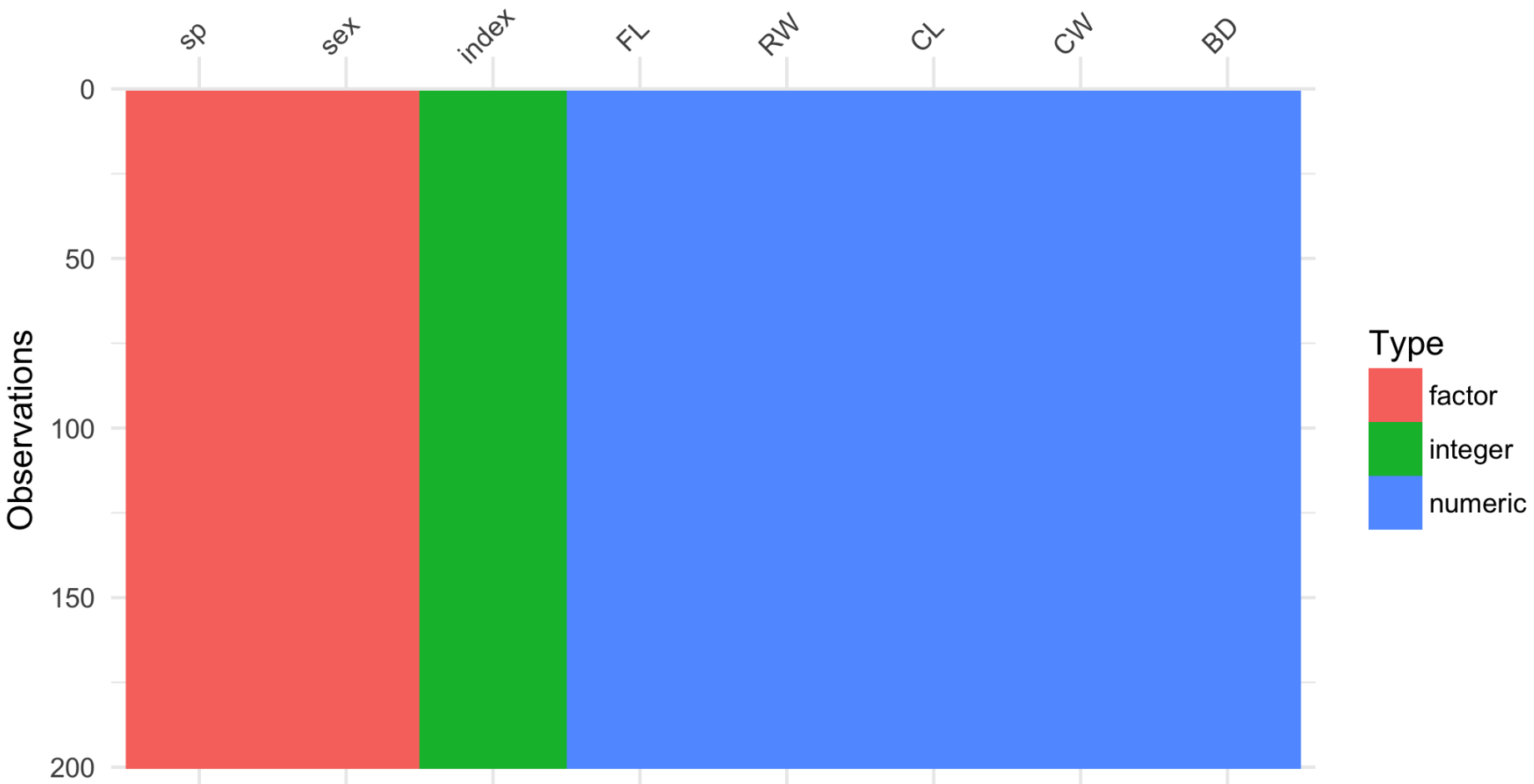
Purple Rock Crab

Figure 1: Leptograpsus variegatus

Image courtesy Wikimedia Commons, user Benjamint444

Morphology

```
library(MASS)
vis_dat(crabs)
```



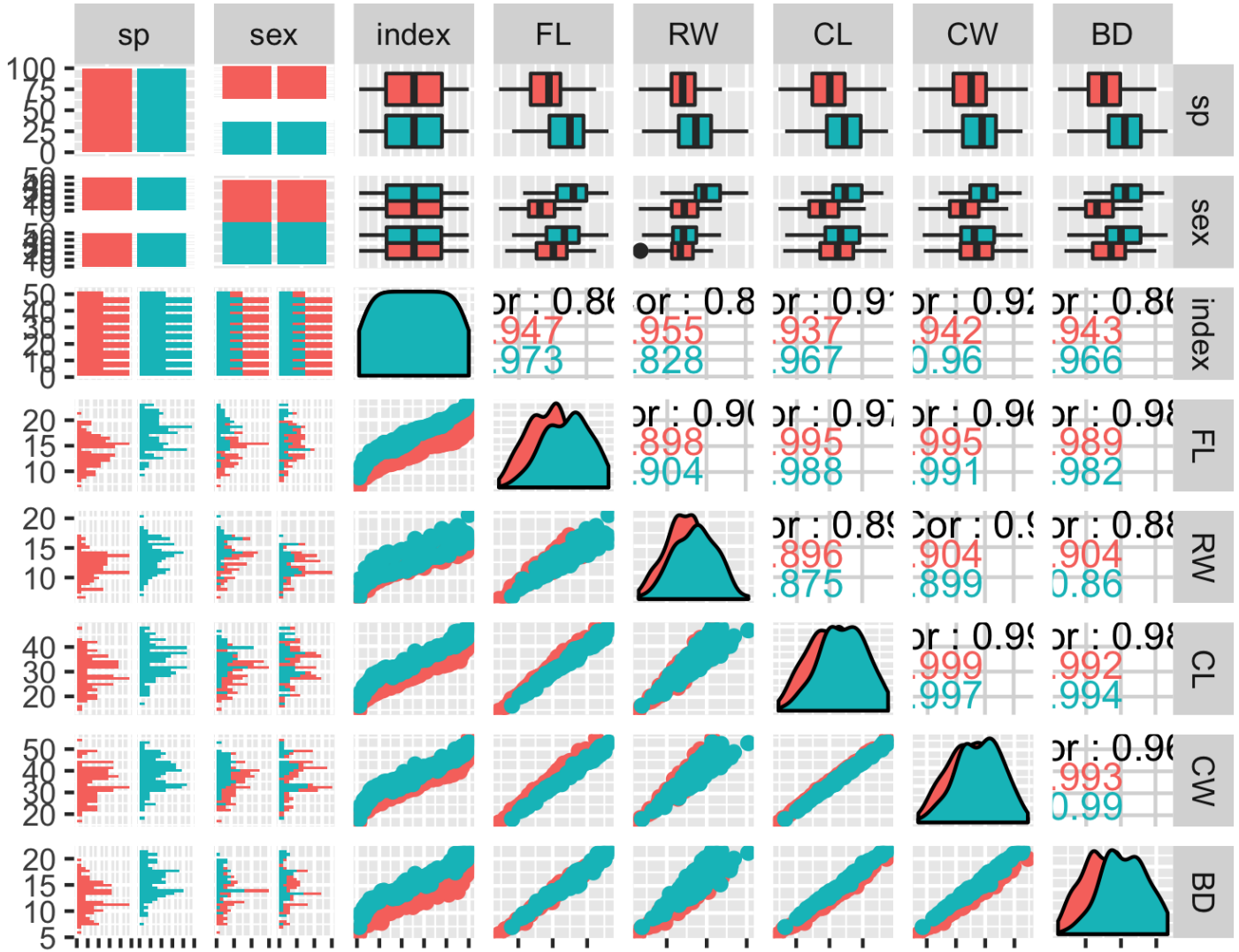
Data Dictionary

- **sp**: species ("B" or "O" for blue or orange)
- **sex**: "F" or "M"
- **index**: `1:50` within each group
- **FL**: frontal lobe size (mm)
- **RW**: rear width (mm)
- **CL**: carapace length (mm)
- **CW**: carapace width (mm)
- **BD**: body depth (mm)

Campbell & Mahon (1974) *Aust. J. Zoology* **22**: 417–425.

Exploratory Data Analysis

```
ggpairs(crabs, ggplot2::aes(colour=sp))
```



Principal Components Analysis (PCA)

PCA

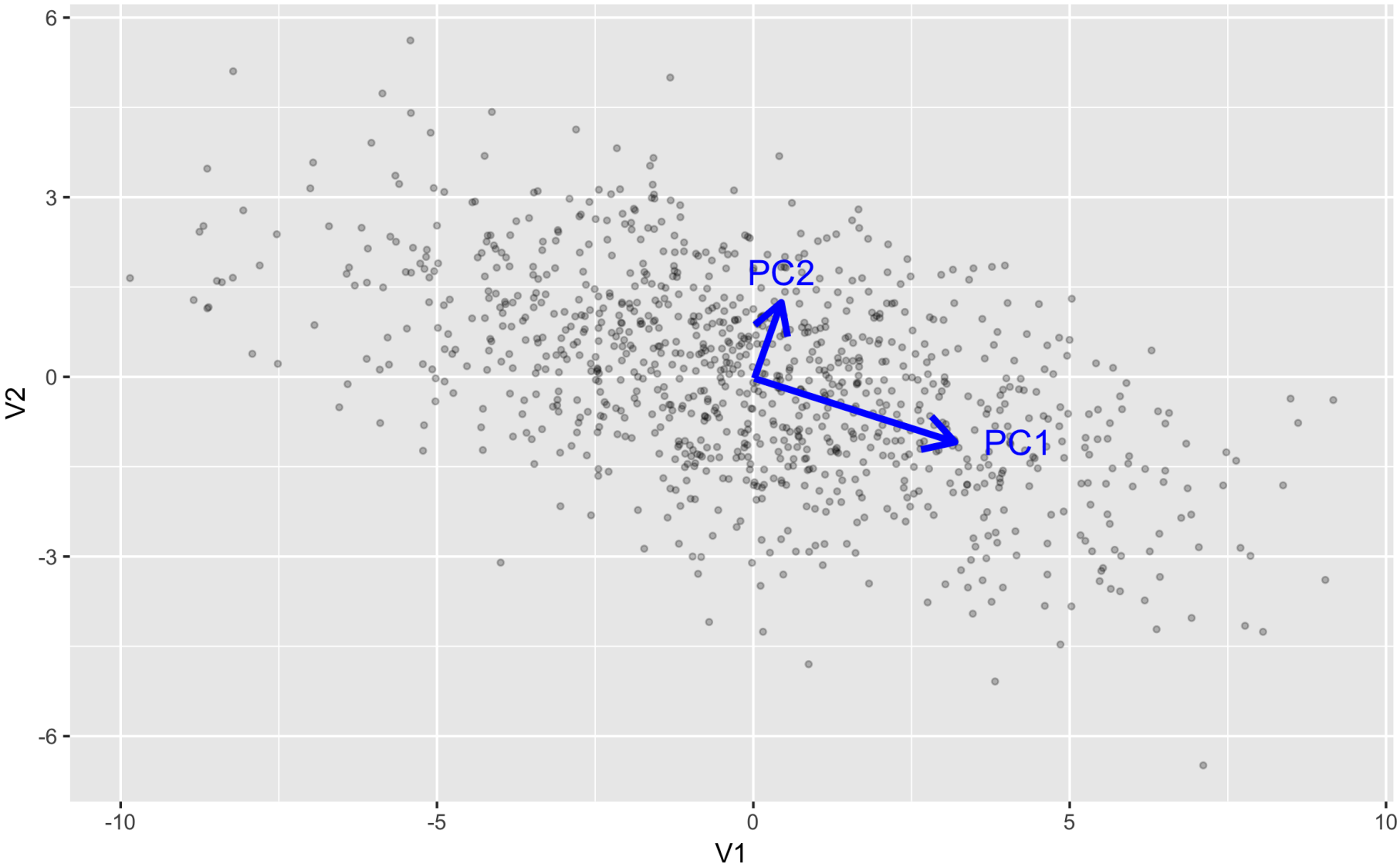
PCA is a transformation of the original variables:

- Useful for high-dimensional, multicollinear data
- Principal components are uncorrelated
  - orthogonal vectors
- Decreasing order of variance

An example of *unsupervised* machine learning

Pearson (1901) *Philosophical Magazine* **2**(11): 559–572.

PCA for 2D Gaussian



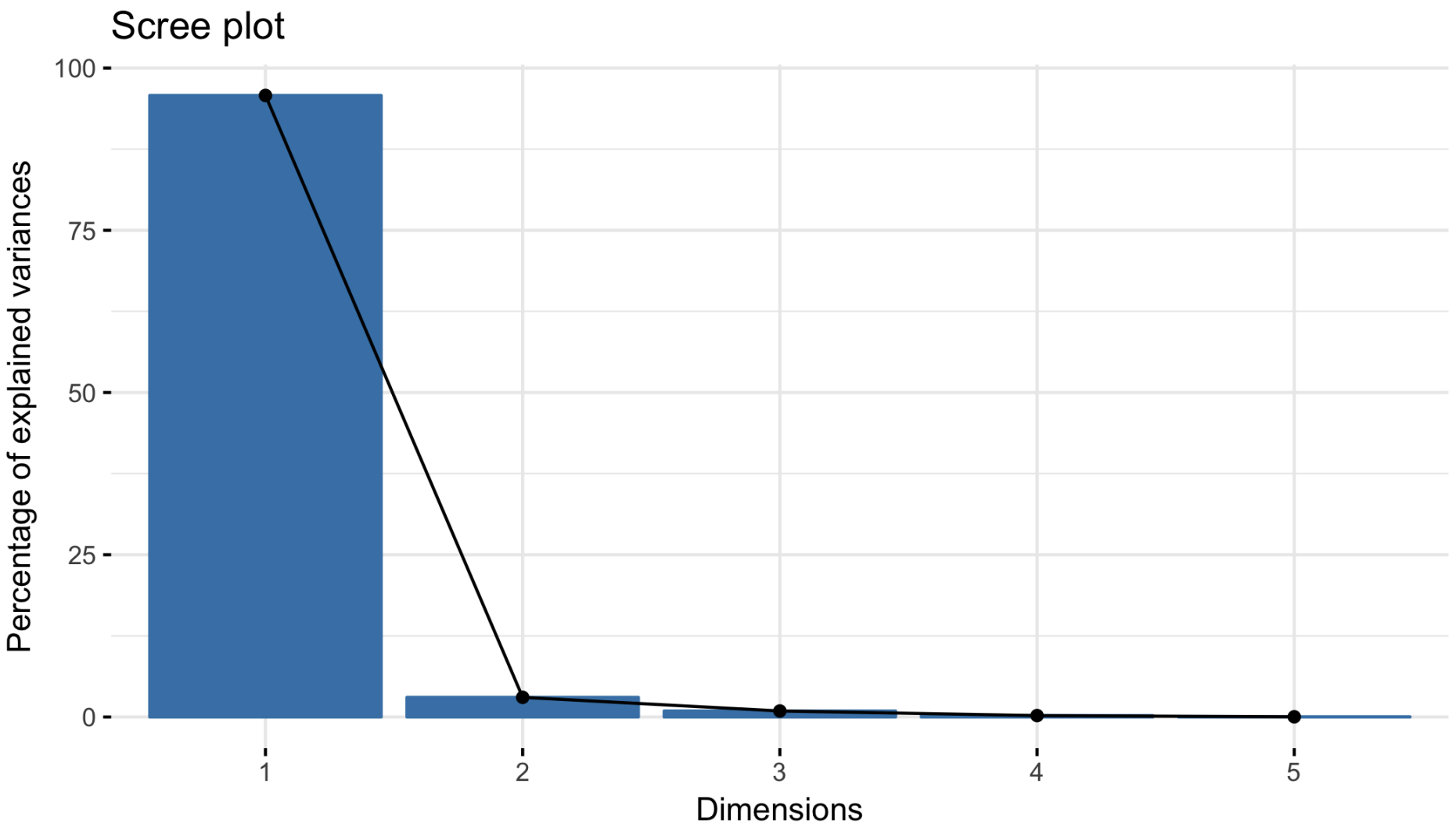
PCA for crabs

```
crab_pca <- PCA(crabs, quanti.sup=3, quali.sup=1:2,
                graph=FALSE)
knitr::kable(crab_pca$eig, digits=4,
    col.names = c("eigenvalue","% variance","cumulative %"))
```

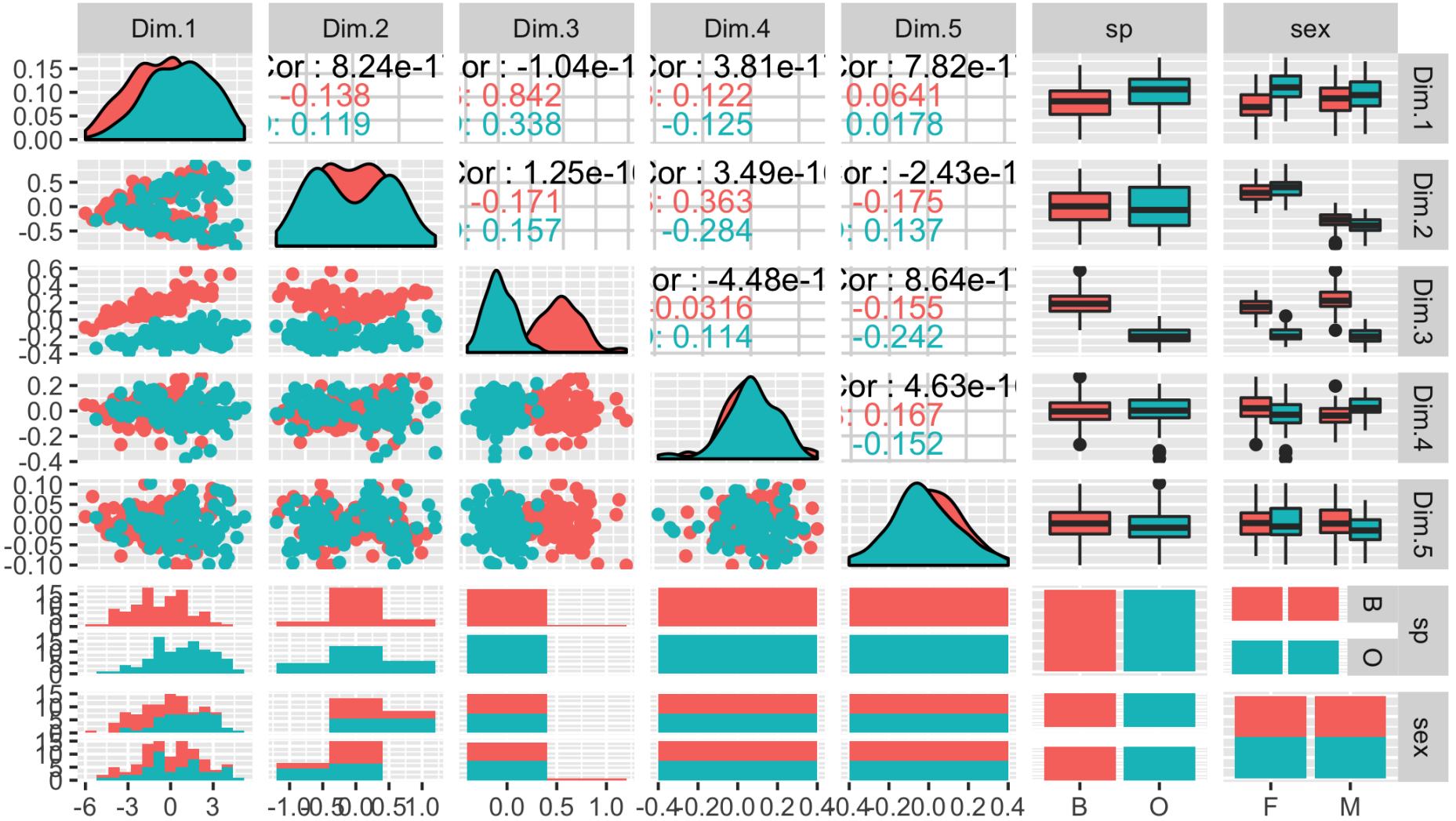
	eigenvalue	% variance	cumulative %
comp 1	4.7888	95.7767	95.7767
comp 2	0.1517	3.0337	98.8104
comp 3	0.0466	0.9327	99.7431
comp 4	0.0111	0.2227	99.9658
comp 5	0.0017	0.0342	100.0000

Scree Plot

```
fviz_screeplot(crab_pca)
```

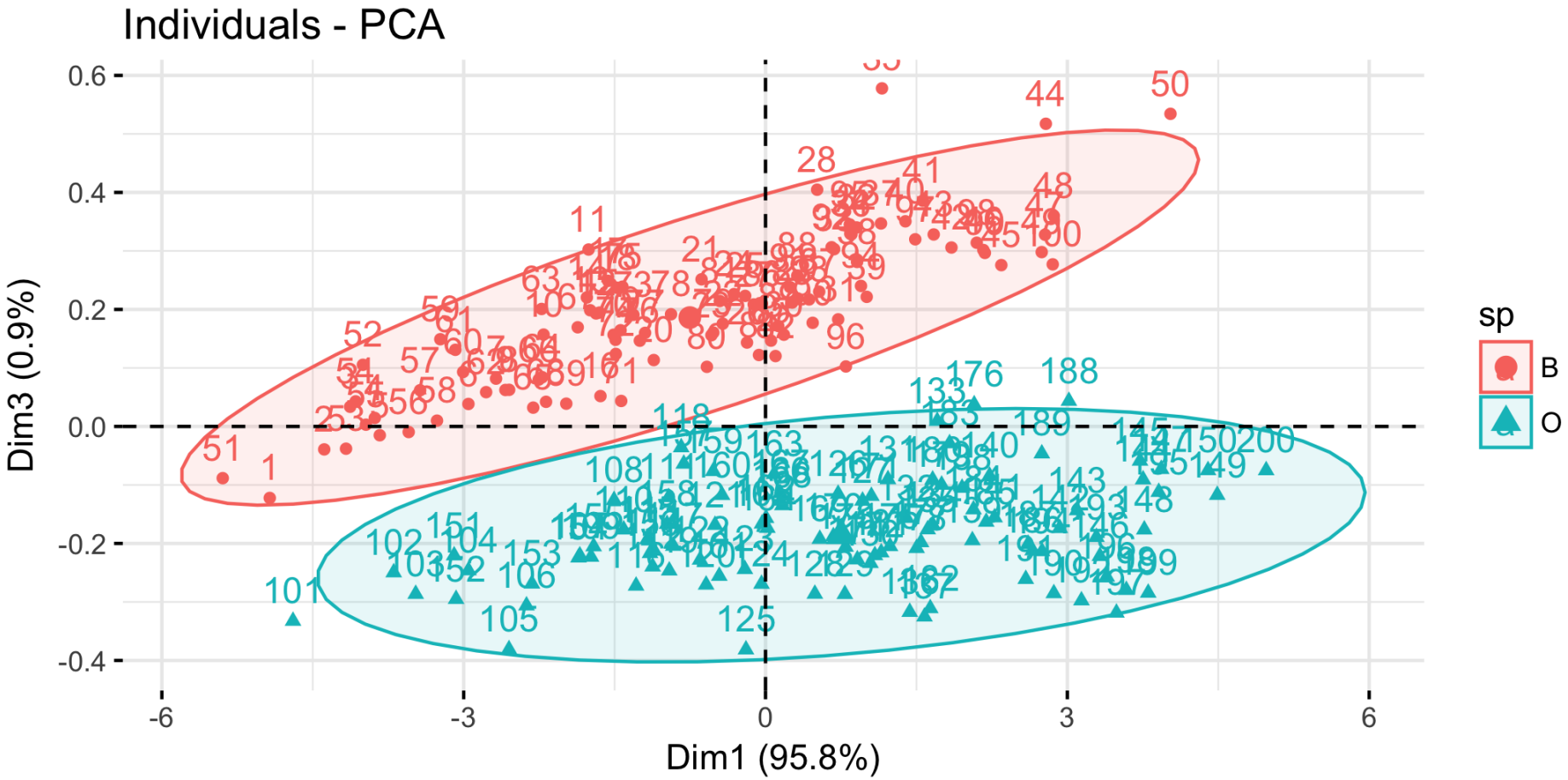


Pairs Plot



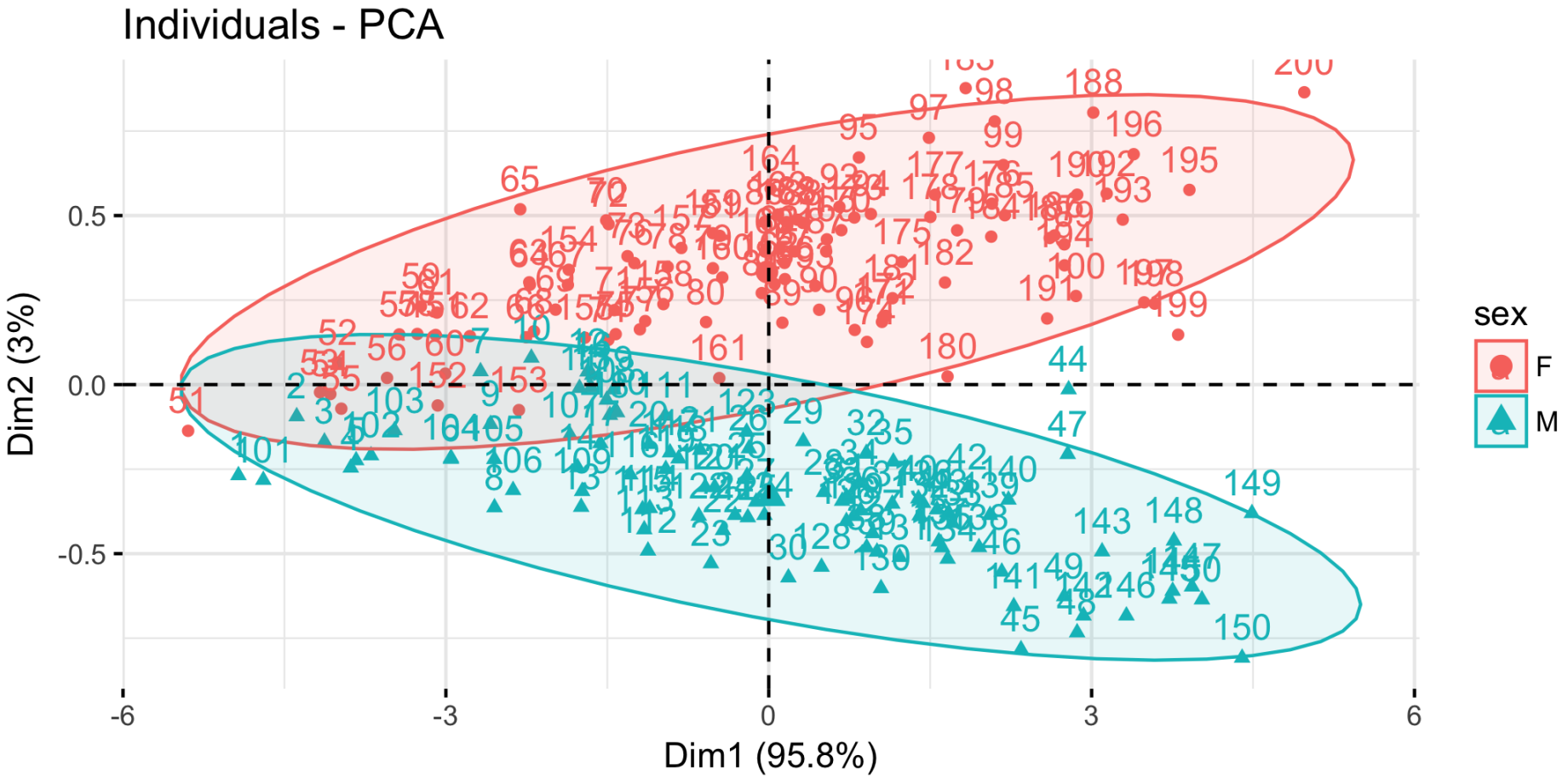
Species

```
fviz_pca_ind(crab_pca, axes = c(1,3), habillage=1,
             addEllipses=TRUE, ellipse.level=0.95)
```



Sex

```
fviz_pca_ind(crab_pca, axes = c(1,2), habillage=2,
             addEllipses=TRUE, ellipse.level=0.95)
```



Another example: genetic variation

Novembre et al. (2008) *Nature* **456**(7218): 98–101.

- $n = 1,387$ European individuals from POPRES project
- Affymetrix 500K single nucleotide polymorphism (SNP) chip
  - $n \ll p$
- Country of origin of each individual's grandparents

Data publicly available from the database of Genotypes and Phenotypes (dbGaP), http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap

PCA analysis of 197,146 genetic loci

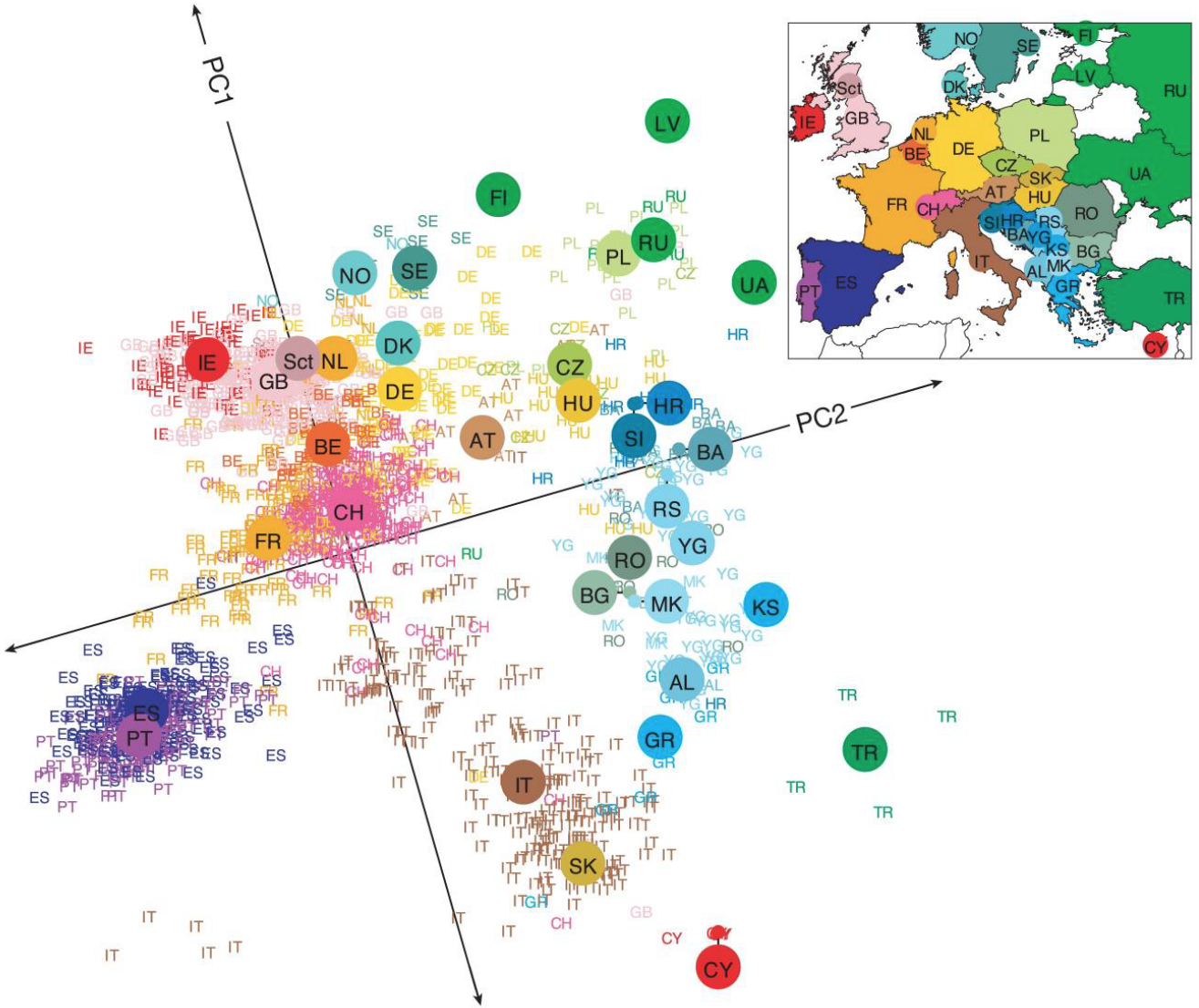


Figure 2: PCA

Clustering

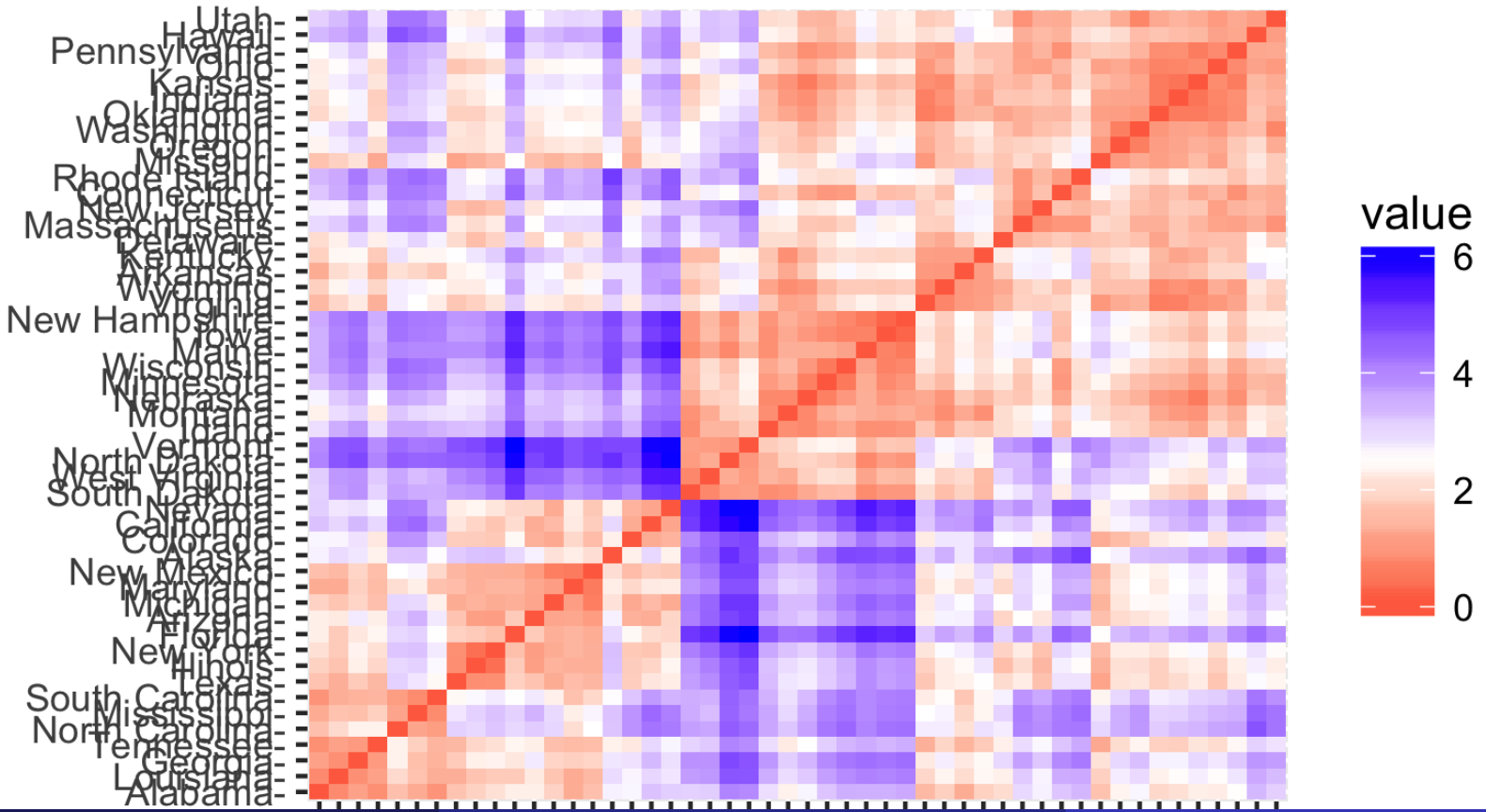
Clustering

Another unsupervised approach for finding patterns in data:

- Grouping observations based on *distance*
- Many different distance metrics (Euclidean, Manhattan, Gower, etc.)
- Also many different clustering algorithms

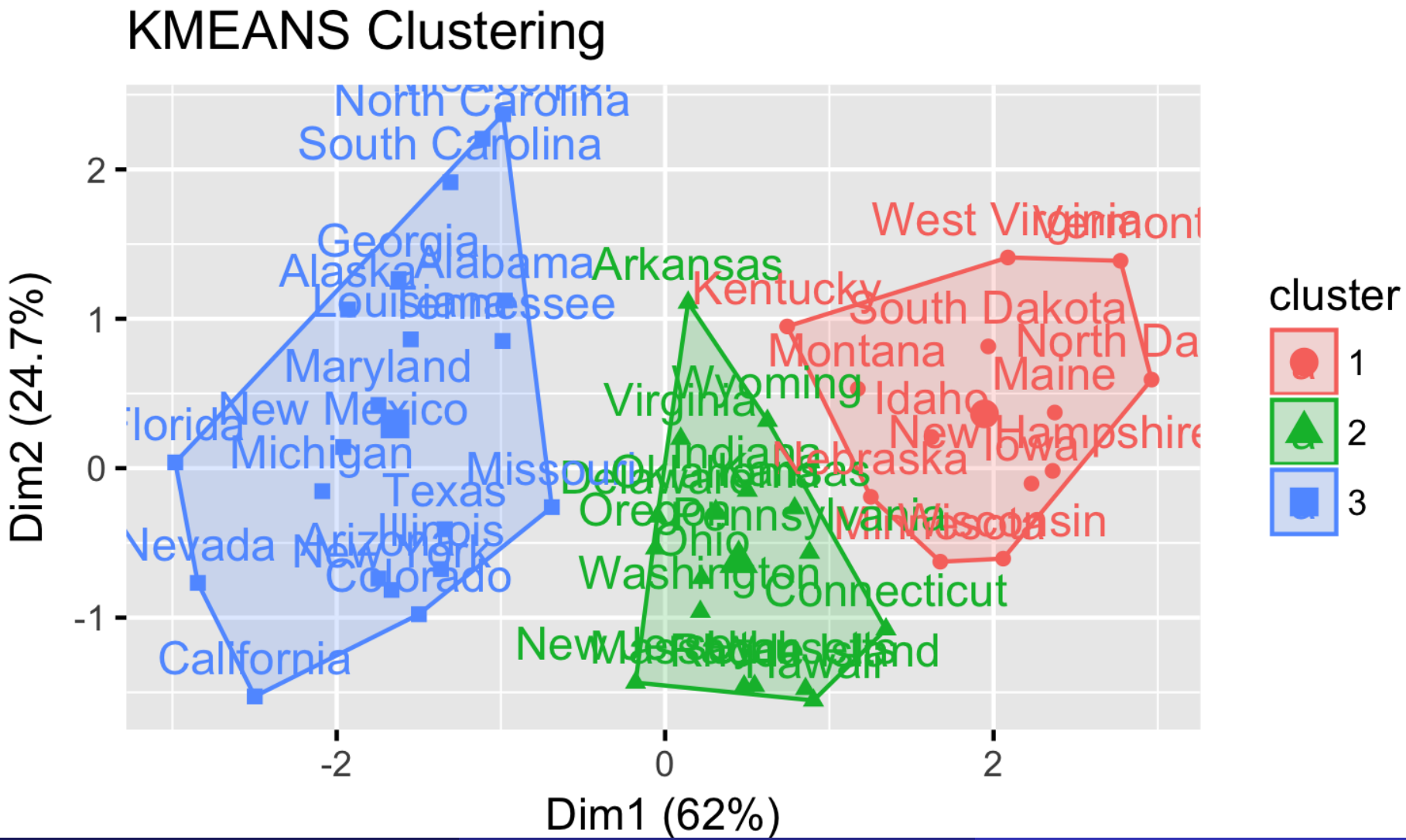
Heatmap

```
data("USArrests")
USdf <- scale(USArrests)
US.dist <- get_dist(USdf, method = "euclidean")
fviz_dist(US.dist)
```



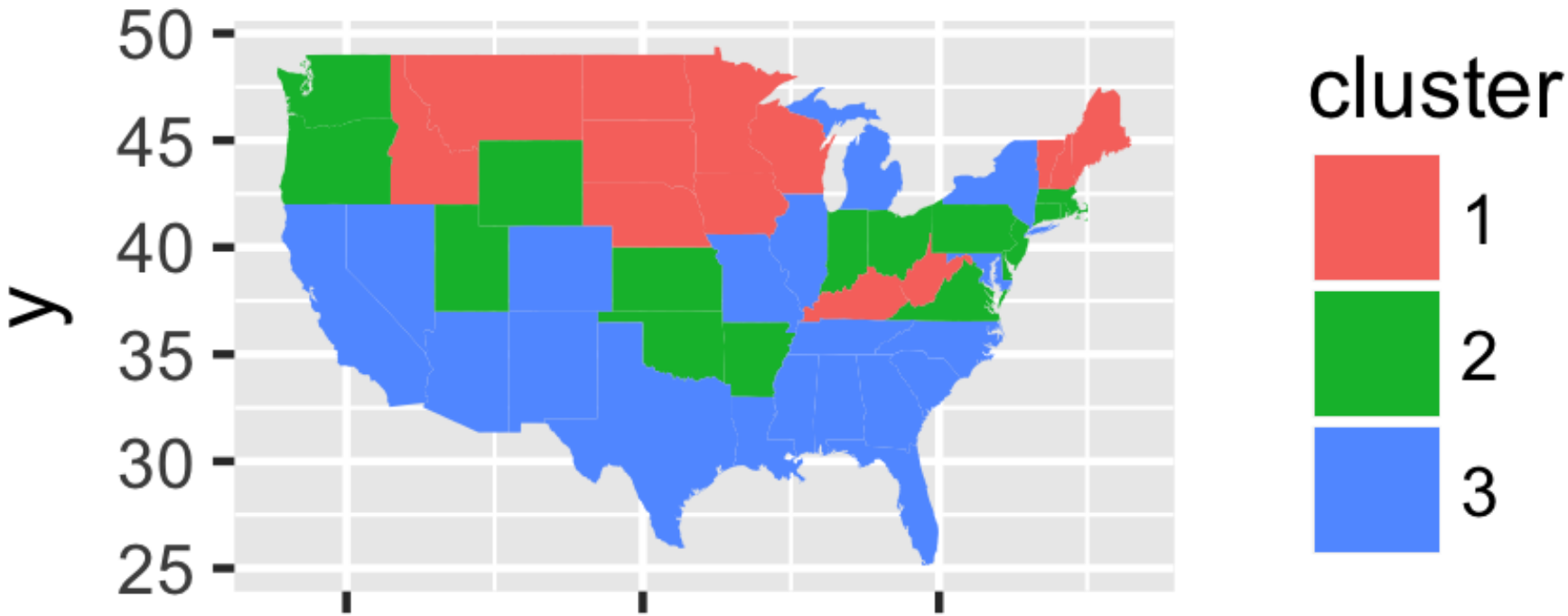
k-means

```
res_km <- eclust(USdf, "kmeans", k = 3)
```



Chloropleth

```
library(maps)
UScat <-data.frame(state = tolower(names(res_km$cluster)), clu
states_map <-map_data("state")
ggplot(UScat, aes(map_id = state)) +
    geom_map(aes(fill = cluster), map = states_map) + expand_l
```



Dendogram

```
res.hc <- eclust(USdf, "hclust")
fviz_dend(res.hc, rect = TRUE)
```