Université Paris Cité

Data analysis and Exploration using R

Dimensionality reduction

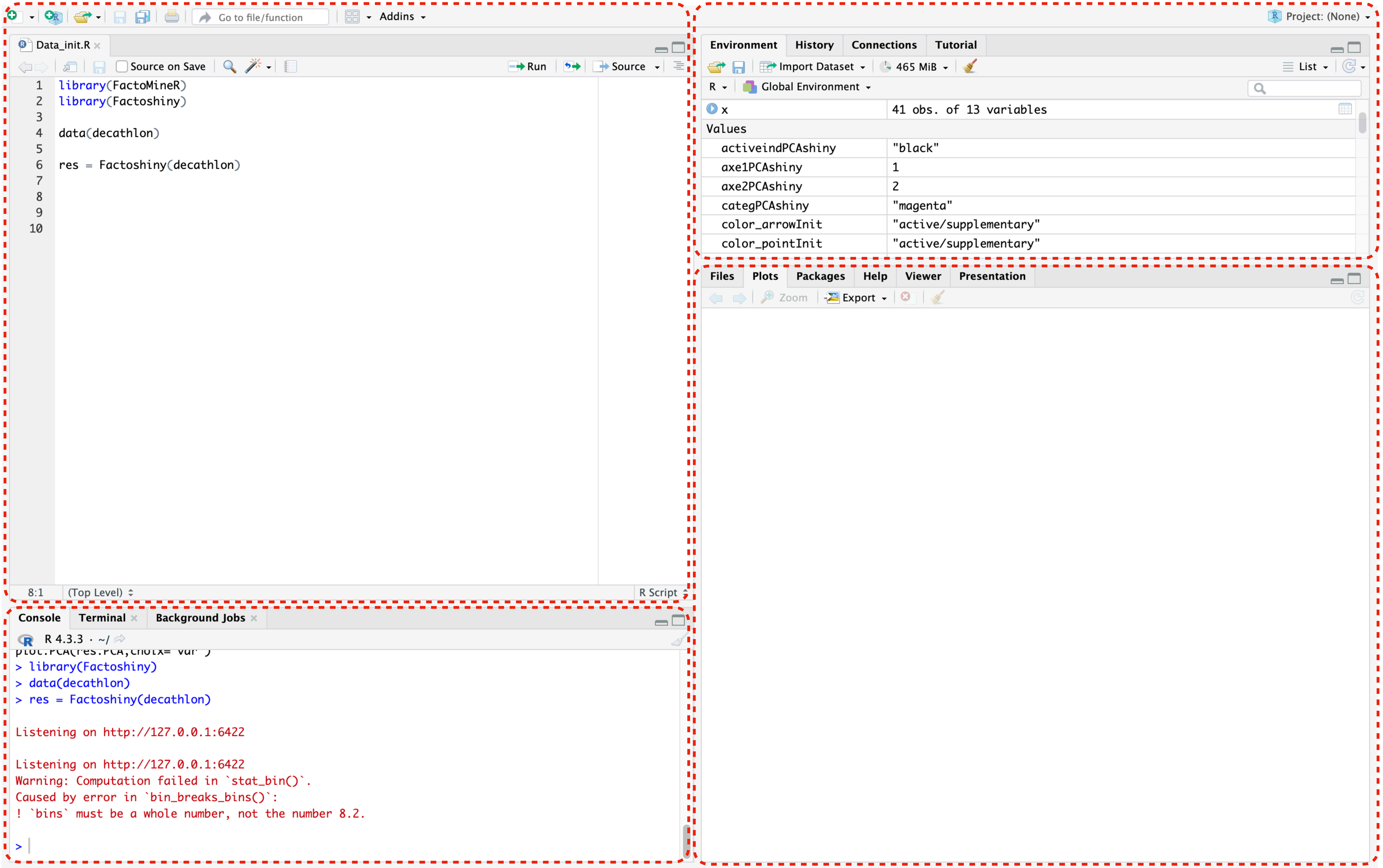
Amine FERDJAOUI

Master I Machine Learning pour la Science des données - Année universitaire 2025/2026

Introduction to R

Code Editor

R console
Terminal



Environment
Variables
History

Plots
Help
Files

R: Introduction

- Install and load library

```
> install.packages("ggplot2") # Install new library
> library(ggplot2) # Load library
```

- Visualise documentation for a function or library

```
> ?mean # or help(mean)
> help("PCA", package = "FactoMineR")
> example(mean)
```

- Load preloaded datasets

```
> data(cars)
> library(help = “datasets”)
> view(cars)
```

R: Data Frames

A data frame is a table of data in R:

- each row = one individual (or observation),
- each column = one variable (or attribute),
- columns can be of different types (numeric, text, factor, etc.).

```
> class(cars)
> is.data.frame(cars)
```

```
> df = data.frame(
    Nom = c("Alice", "Bob", "Clara"),
    Age = c(23, 25, 22),
    Sexe = c("F", "M", “F")
  )
```

R: Manipulate Data Frames

```
> head(cars)
> summary(cars)
> head(mtcars)
> names(mtcars) # Show column names
> mtcars$hp <- NULL # Delete a column
> mtcars <- mtcars[-2, ]  # Delete a row
> mtcars$new_var <- 1:nrow(mtcars) # Add a column
```

R: Read csv

- Example of csv file

```
Name; City; Sallary; Year
Alpha; Paris; 22000; 2023
Beta; Lyon; 69500; 2023
Gamma; Marseille; 33400; 2023
Delta; Paris; 12000; 2024
```

- Read csv file

```r
data <- read.csv("ventes.csv", sep = ";", dec = ".", header = TRUE, row.names = 1)
```

Delimiter (eg. , or \t or ;)

Decimal separator (. or ,)

Whether the first like is columns names

Use first column as index

TP1: R basics

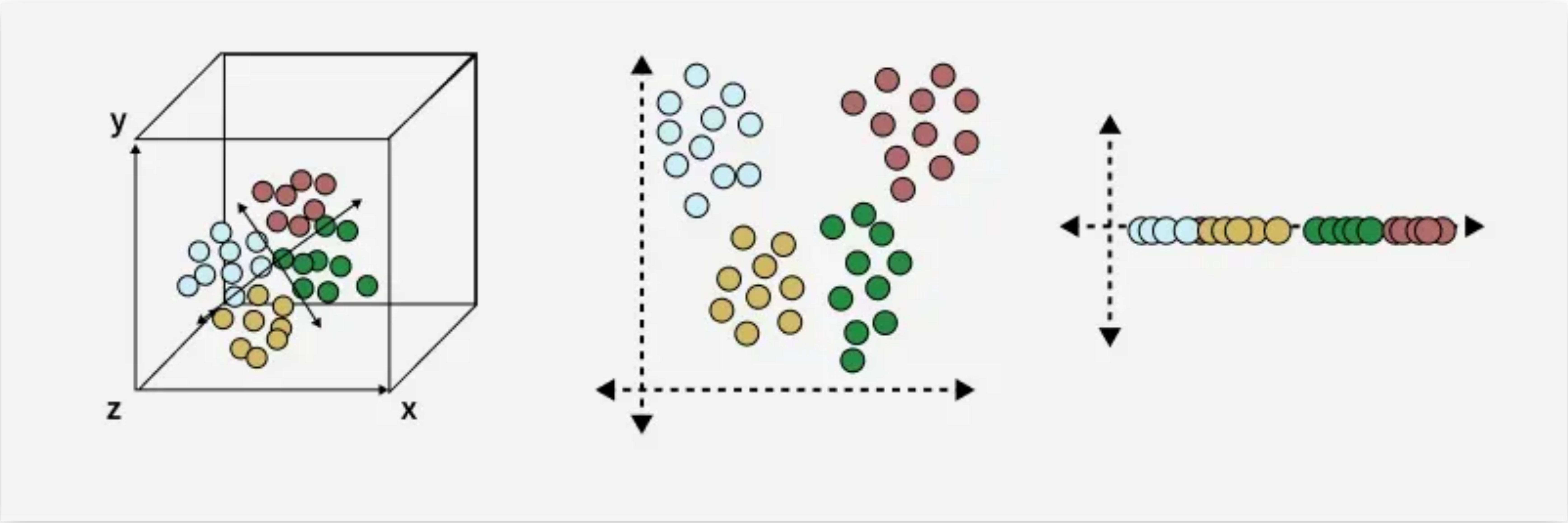
A envoyer à : mlds.git@gmail.com

1. Charger un dataset intégré
    1. Chargez le jeu de données **mtcars**.
    2. Visualiser les données sous forme de table.
    3. Affichez les 5 premières lignes et le résumé des variables.
2. Modifier le dataset
    1. Supprimez la colonne drat.
    2. Ajoutez une colonne prix avec des valeurs aléatoires entre 10000 et 40000.
    3. Supprimez la première ligne du tableau.
3. Ajouter une nouvelle ligne
    1. Créez une ligne avec vos propres valeurs et ajoutez-la à la fin.
4. Sauvegarder et réimporter
    1. Sauvegardez votre table CSV en utilisant write.csv (utiliser ?write.csv pour afficher l'aide.)
    2. Réimportez-la avec read.csv() et vérifiez les données.

Dimensionality Reduction

What is dimensionality reduction

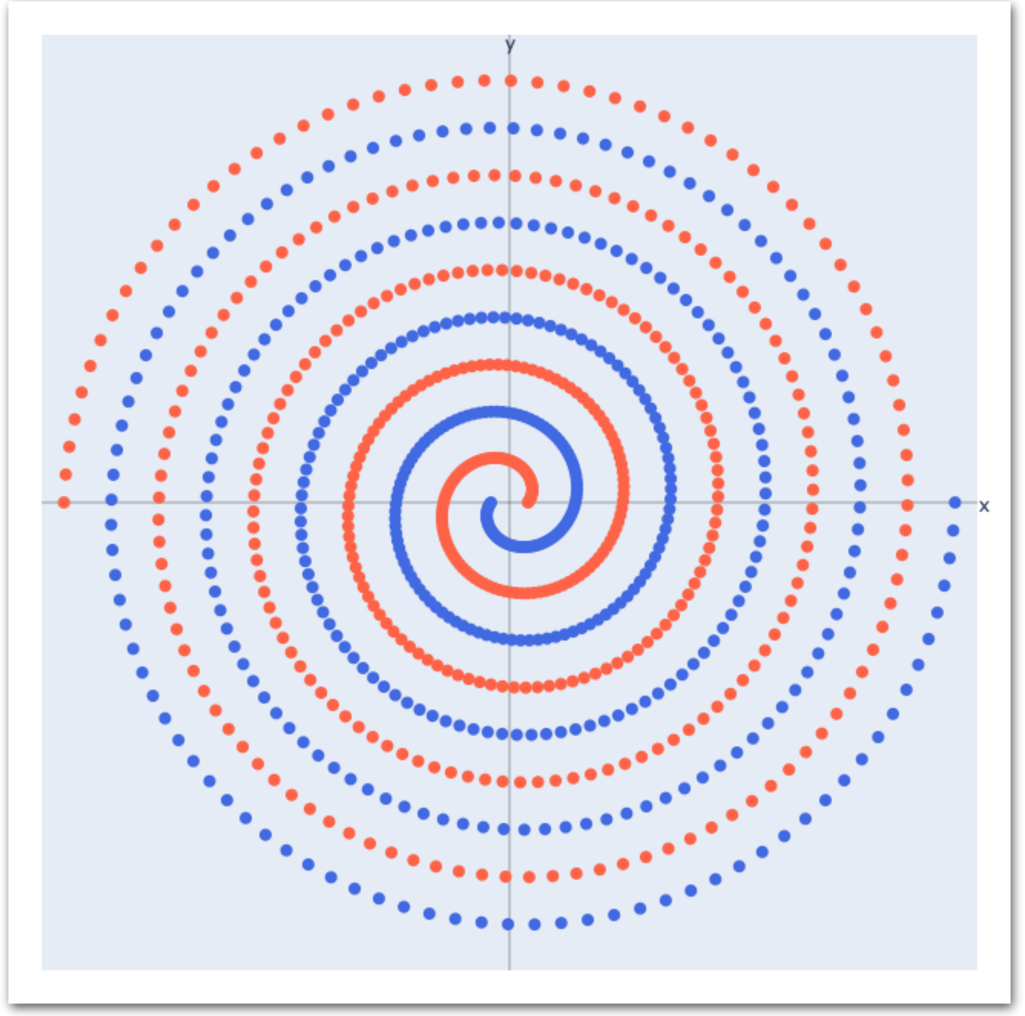
- Dimensionality reduction projects high-dimensional data into a lower-dimensional space while keeping as much useful information as possible.
- It can be used to:
    - Simplify visualization (e.g., 2D/3D plots of complex datasets).
    - Reduce noise and improve machine learning performance.



PCA: Principal Component Analysis

Principal Component Analysis (PCA) is a **linear** and **fast** method based on linear algebra.

- Finds axes (principal components) that maximize the variance of the data.
- Each component is a linear combination of the original variables.
- Useful for reducing dimensionality while keeping most of the variance.
- Limitation: cannot capture non-linear structures in the data (relation between data is non linear)



Dimensionality Reduction using FACTOMINER

PCA using FactoMineR

- FactoMineR is an R package (developed by the team of François Husson) dedicated to multivariate exploratory data analysis. The main goal is to simplify complex multivariate analyses and make them accessible and interpretable.
- It provides functions for the most common multivariate methods, such as:
  - Principal Component Analysis (**PCA**)
  - Correspondence Analysis (**CA**)
  - Multiple Correspondence Analysis (**MCA**)
  - Hierarchical Clustering (**HCPC**)
  - and several extensions (**MFA**, **MFAmix**, etc.).
- FactoMineR automatically produces:
  - Tables of eigenvalues, contributions, and squared cosines ($\cos^2$)
  - Graphical outputs (individuals, variables, biplots)
  - Interpretation aids (which variables/individuals influence each axis)

PCA using FactoMineR

- Load Library

```
> install.packages(c("FactoMineR", "factoextra"))
> library(FactoMineR)
> library(factoextra)
```

- Run PCA on USArrests dataset

```
> res.pca <- PCA(USArrests)
```

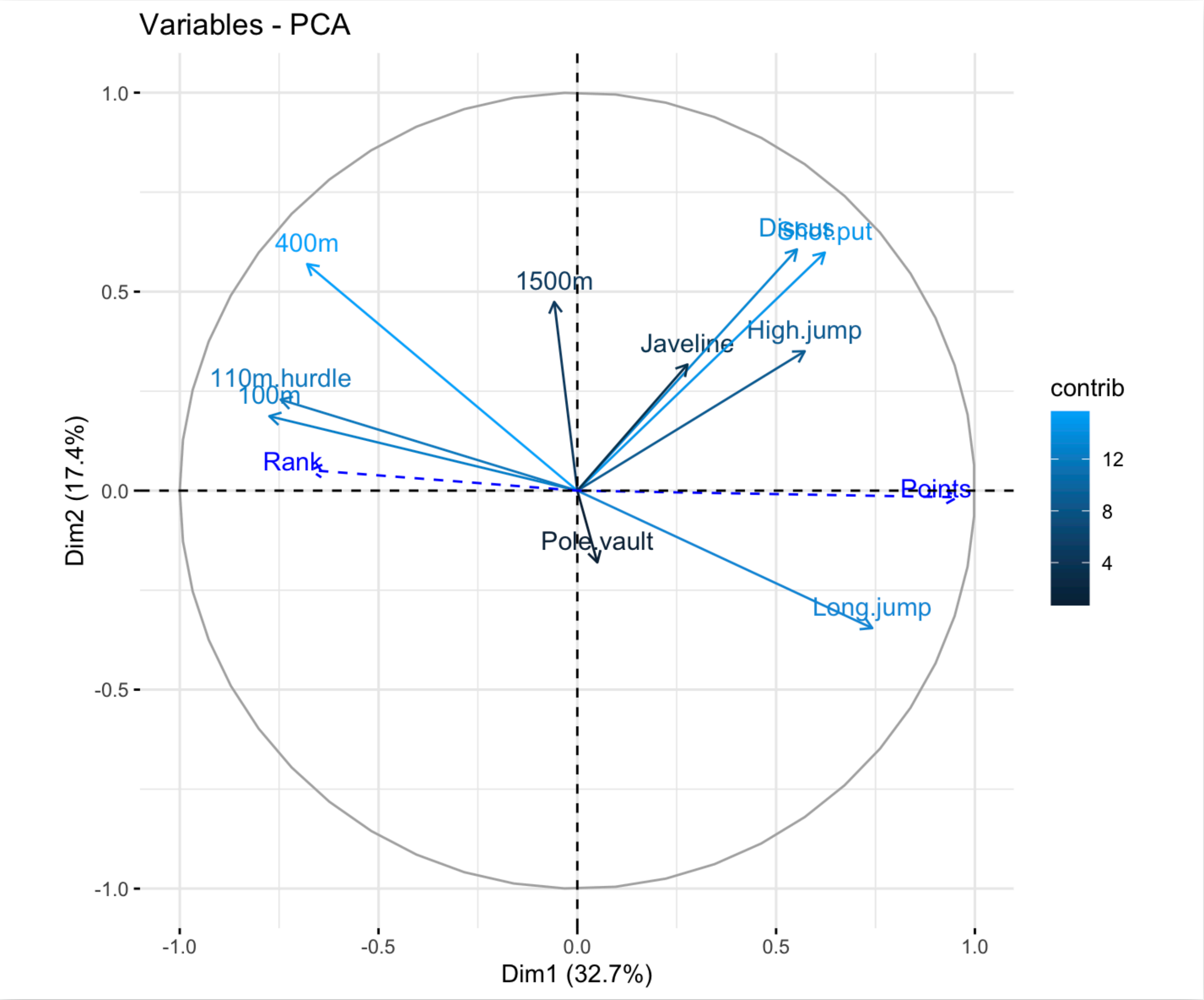
- Basic plots

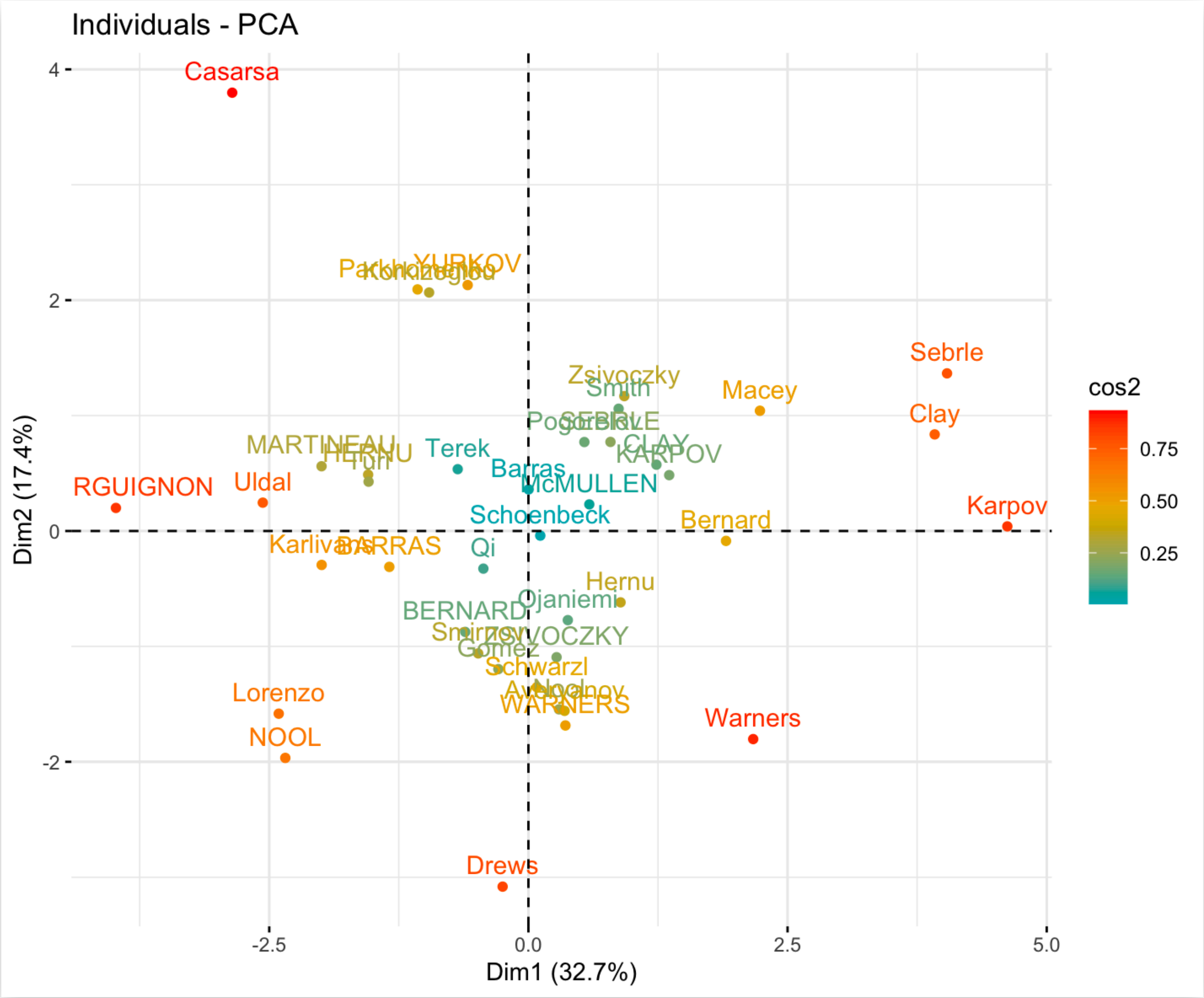
```
> plot(res.pca, choix = “var”) # Plot of variables
> plot(res.pca, choix = “ind") # Plot of individuals
```

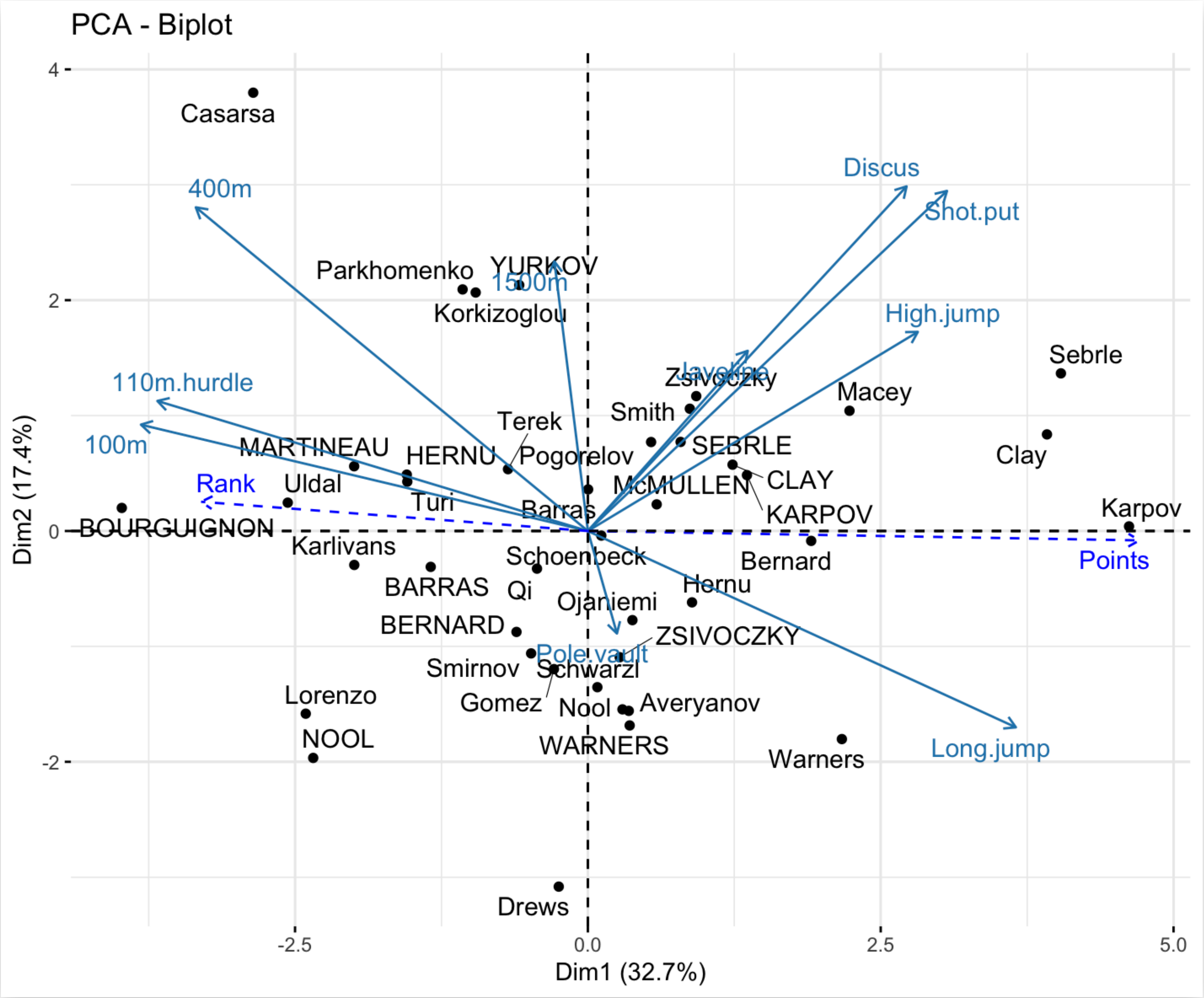
PCA using FactoMineR and factoextra

- Use factoextra for better visualization

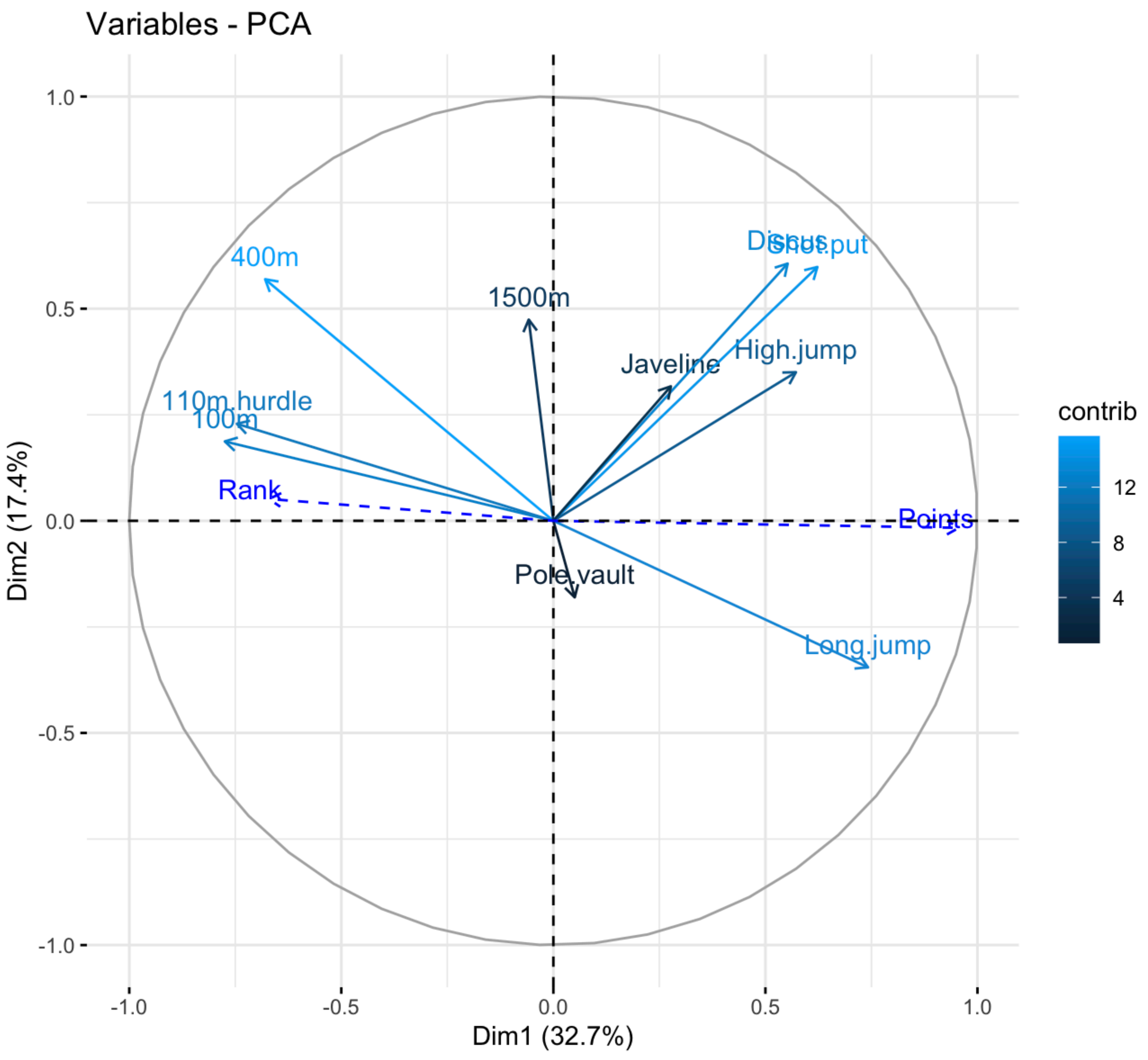
```r
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))  # Scree plot
fviz_pca_var(res.pca, col.var = "contrib")            # Variables
fviz_pca_ind(res.pca, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
fviz_pca_biplot(res.pca, repel = TRUE)                # Biplot
```







PCA using FactoMine: Variables plot interpretation



- Each arrow represents a quantitative variable.
- The **direction** and **length** of an arrow show how much that variable contributes to the **axes**. Longer arrows → better represented on the plane (higher $\cos^2$).
- The angle between arrows indicates **correlation** between variables:
  - Small angle (close arrows) → strong positive correlation.
  - Opposite directions (180°) → strong negative correlation.
  - Perpendicular (90°) → very weak correlation.
- Variables close to the same axis are the ones that define that component the most.