Université Paris Cité

Data analysis and Exploration using R

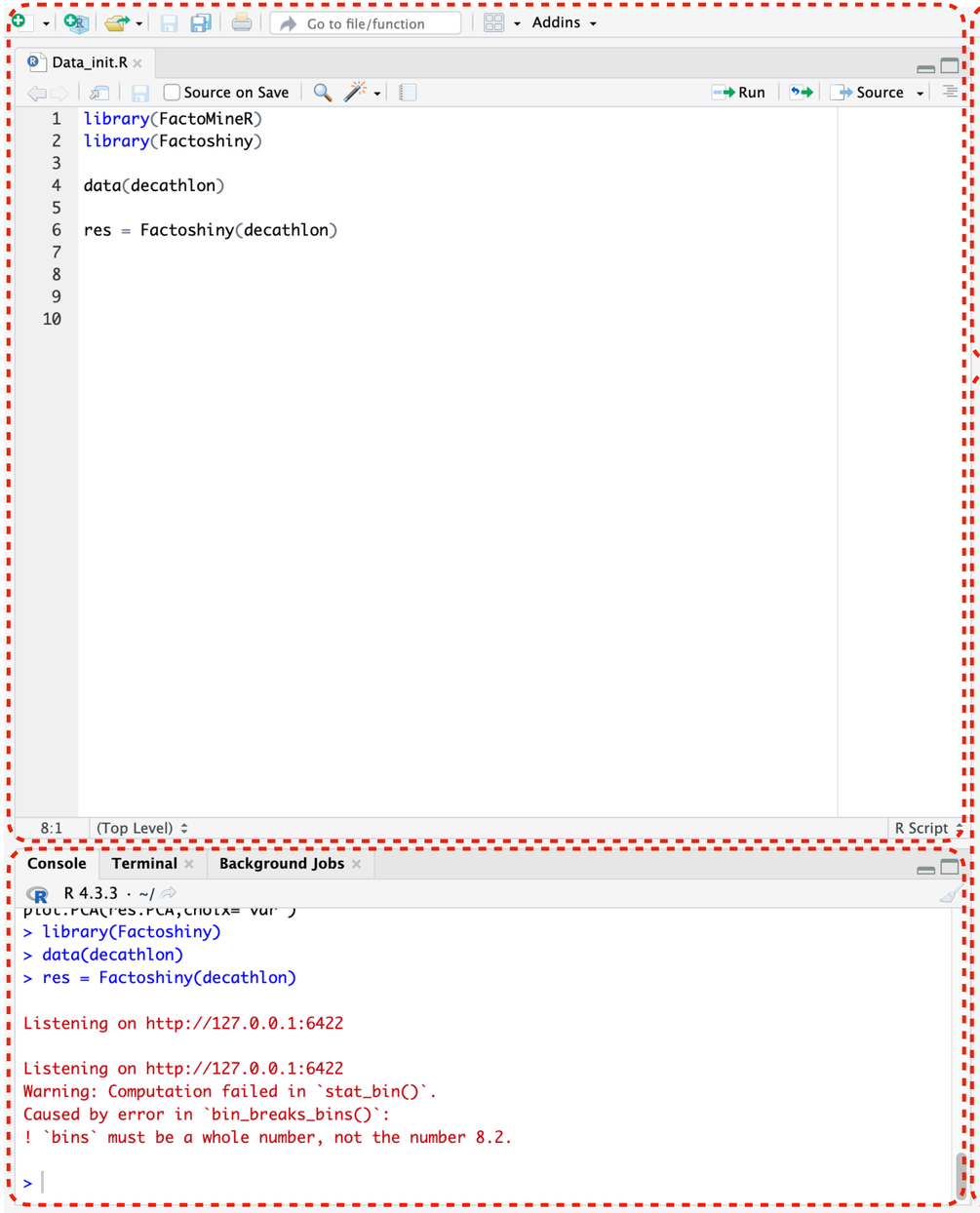
Dimensionality reduction

Amine FERDJAOUI

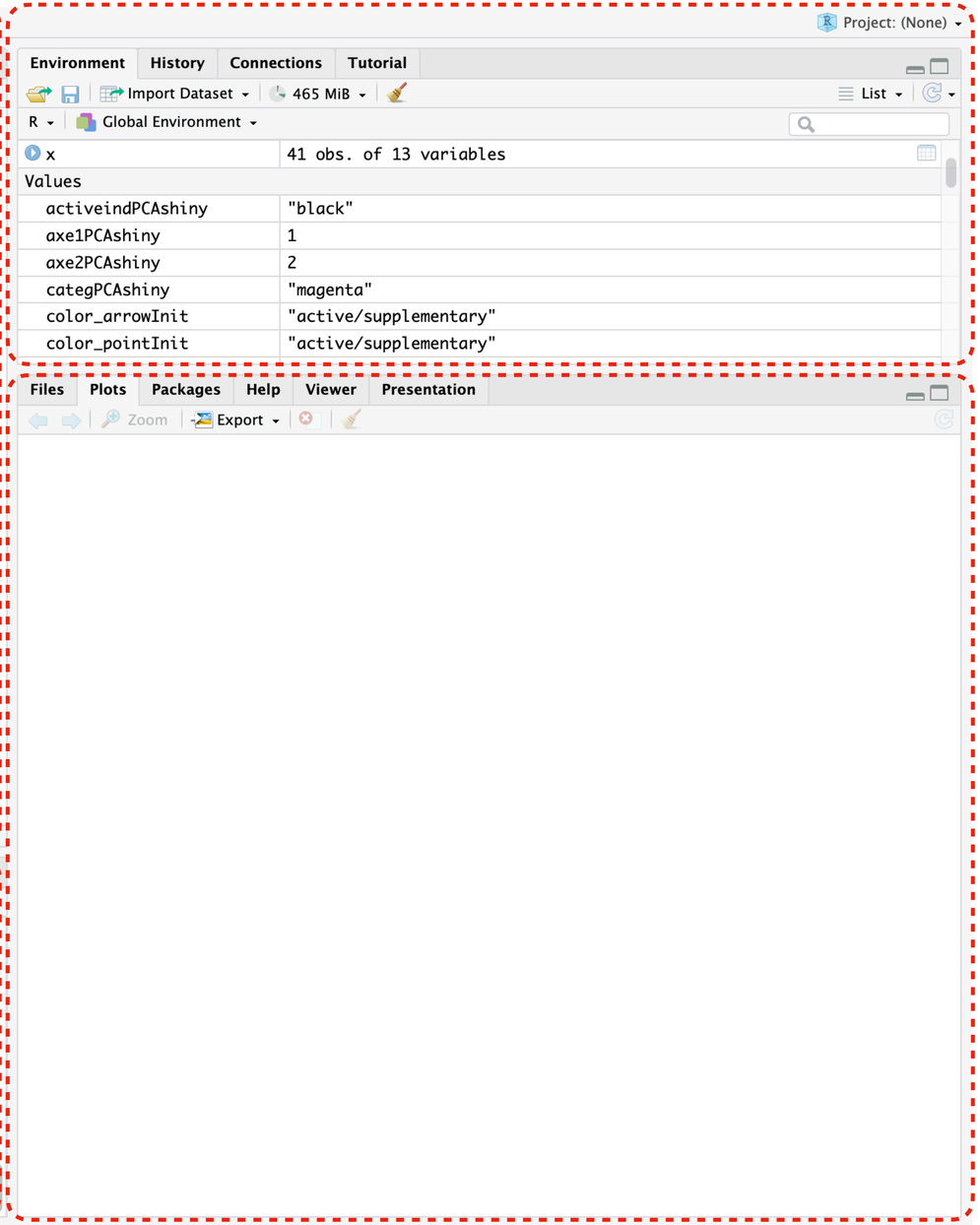
Master I Machine Learning pour la Science des données - Année universitaire 2025/2026

Introduction to 

Code Editor



R console
Terminal



Environment
Variables
History

Plots
Help
Files

R: Introduction

https://github.com/AmineFrj/PCA

- 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

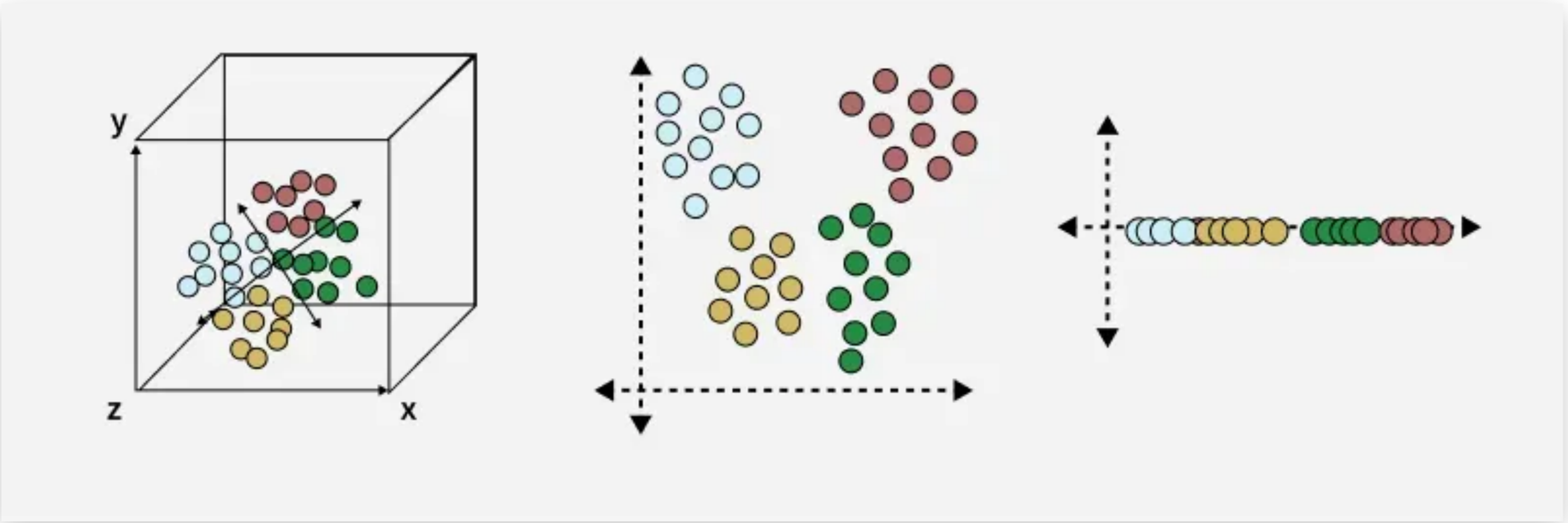
A envoyer à : mlds.git@gmail.com

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

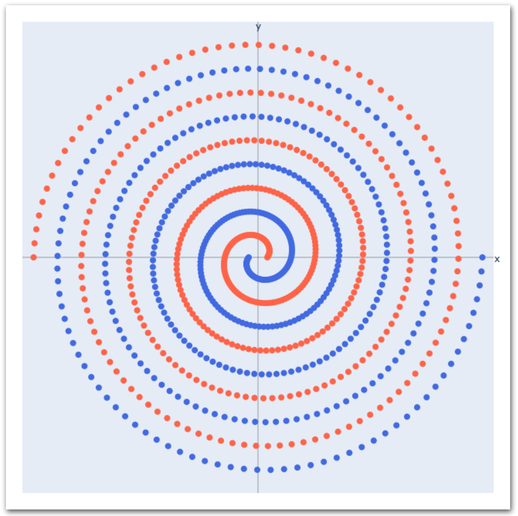
- Dimensionality reduction projects high-dimensional data into a lower-dimensional space while keeping as much useful information as possible.
- It can be used to:
  - Simplify visualization (e.g., 2D/3D plots of complex datasets).
  - Reduce noise and improve machine learning performance.



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)
```

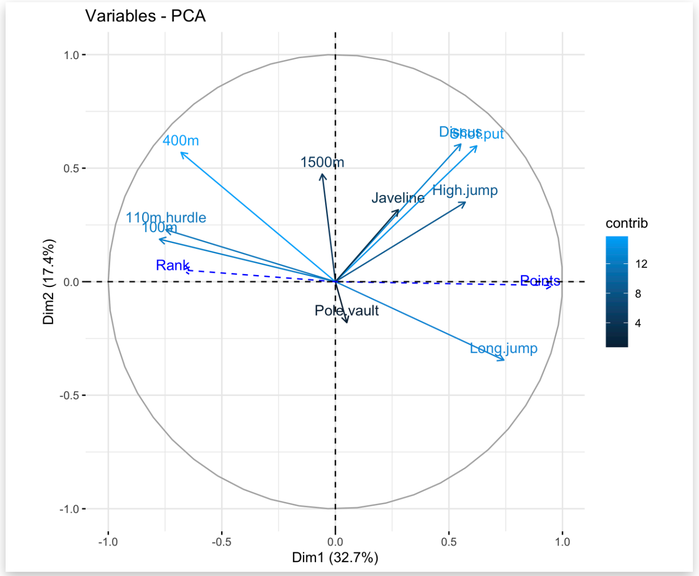
- Basic plots

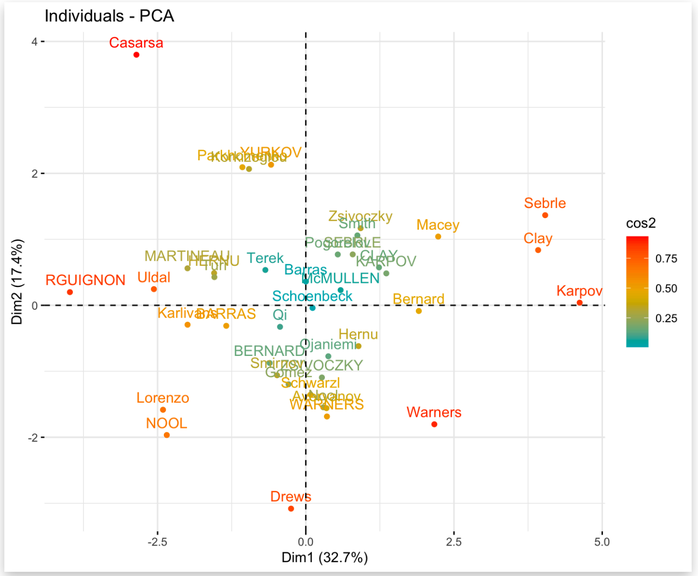
```
> plot(res.pca, choix = ‘‘var’’) # Plot of variables
> plot(res.pca, choix = ‘‘ind") # Plot of individuals
```

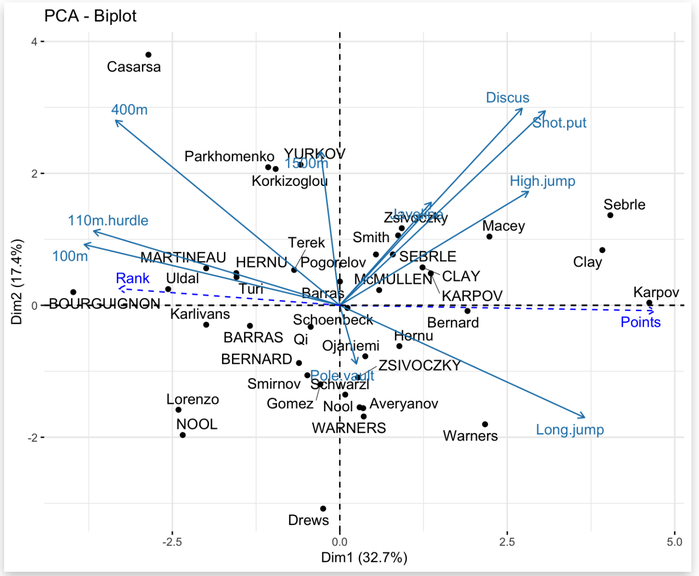
PCA using FactoMineR and factoextra

- Use factoextra for better visualization

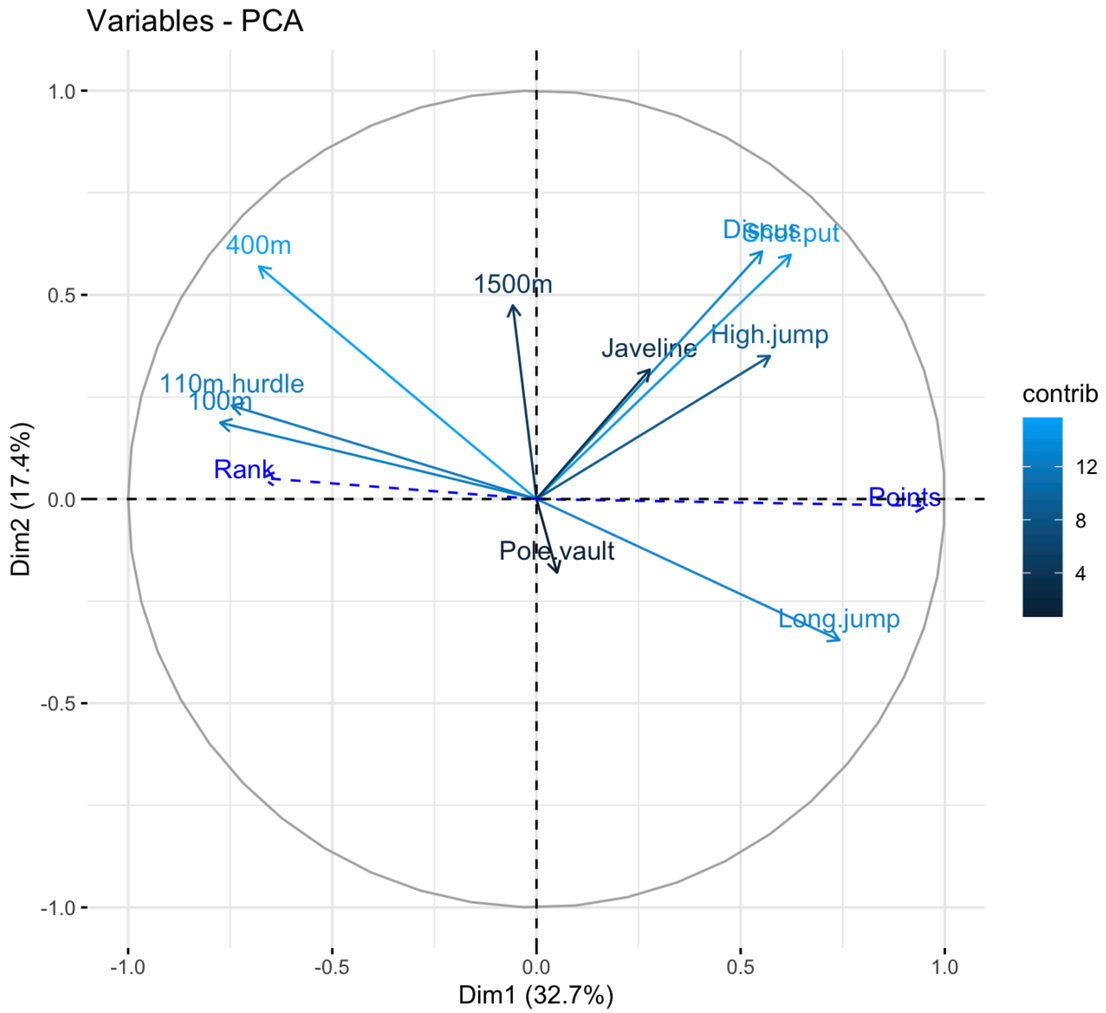
```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))  # Scree plot
fviz_pca_var(res.pca, col.var = "contrib")              # Variables
fviz_pca_ind(res.pca, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
fviz_pca_biplot(res.pca, repel = TRUE)                  # Biplot
```







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 $\cos^2$).
- Variables close to the same axis are the ones that define that component the most.
- 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.

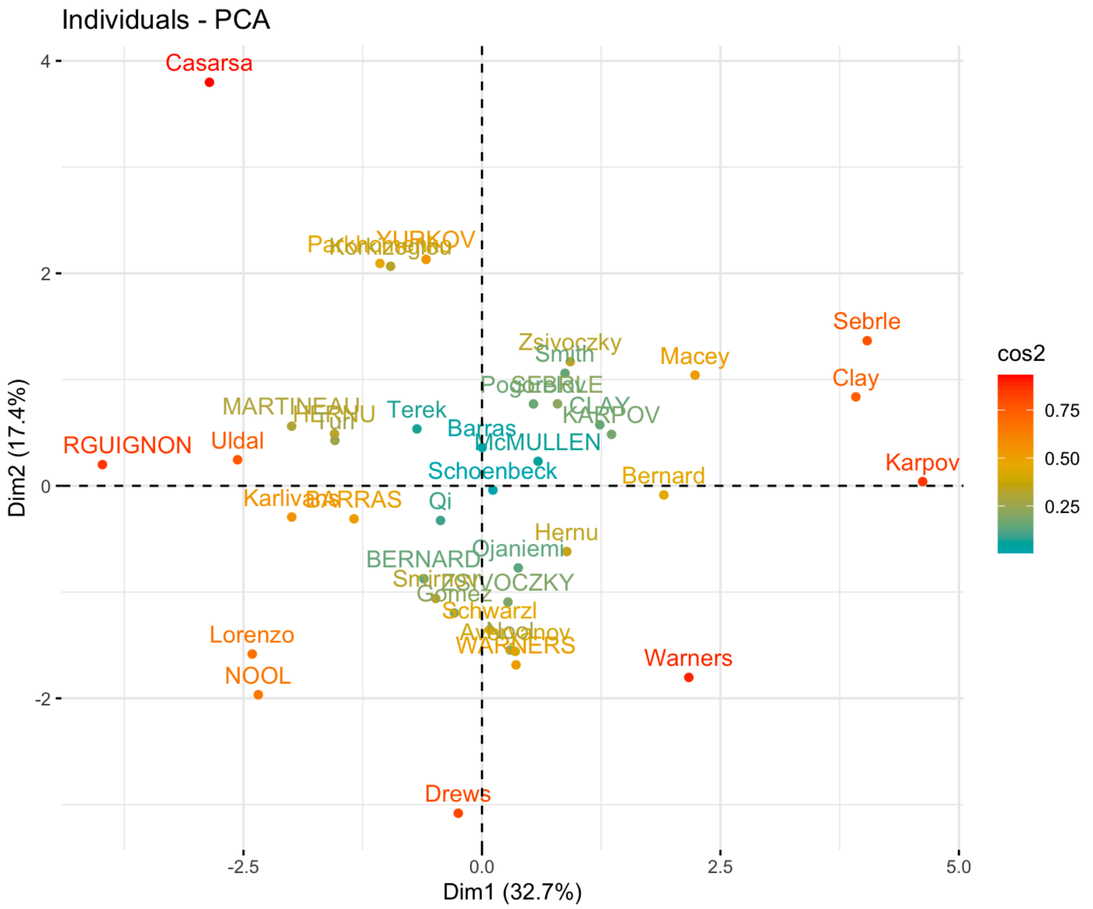
TP2: PCA Variables

A envoyer à : mlds.git@gmail.com

Objectif. Réaliser une ACP sur les données decathlon. Les données portent sur des athlètes ayant participé à un décathlon (10 épreuves d'athlétisme). Chaque ligne correspond à un athlète, chaque colonne à une épreuve : 100m, longueur, poids, hauteur, 400m, 110m haies, disque, perche, javelot et 1500m. L'objectif est de comprendre comment ces épreuves sont reliées entre elles, quelles dimensions principales structurent la performance globale.

1. Préparation et exploration
   - Sélectionner les variables quantitatives correspondant aux dix épreuves.
   - Examiner la structure et le résumé statistique du jeu de données.
   - Expliquer brièvement le type de variables et les unités mesurées.
2. Réaliser une ACP
   - Effectuer une ACP sur les dix épreuves uniquement et considérer le reste en variables supplémentaires.
   - Afficher le cercle des corrélations des variables avec FactoMineR avec factoextra.
   - Identifier les variables fortement corrélées, ceux qui sont bien représentées, et ceux qui contribuent le plus aux deux premiers axes.
   - Interpréter les deux premiers axes : que semble représenter le premier axe ? Que semble représenter le deuxième axe ?
   - Repérer les grands groupes de variables et les types de performances qu'ils décrivent.

PCA using FactoMineR: Variables plot



- Each point represents an individual (here: an athlete).
- The **position** of each individual is determined by their scores (coordinates) on the principal components (Dim1, Dim2, etc.).
- Individuals that are **close** to each other on the plot have **similar** performance profiles across all variables.
- The axes (Dim1, Dim2) represent **underlying dimensions** (combinations of variables that summarize performance).
- The color scale ($\cos^2$) shows the quality of representation:
  - High $\cos^2$ → the individual is well represented on this plane.
  - Low $\cos^2$ → less well represented; the individual's pattern **may appear on another dimension**.
- Individuals far from the origin are more extreme or distinctive (very high or very low scores).

TP3: PCA Variables

A envoyer à : mlds.git@gmail.com

Après avoir étudié les **relations entre variables**, on cherche maintenant à comprendre **comment les athlètes se répartissent** dans l'espace défini par les deux premières dimensions de l'ACP.

1. Générer le plot des individus (plan 1–2).
2. Observer la répartition générale des athlètes :
   - Le nuage est-il homogène ou structuré ?
   - Observe-t-on des groupes distincts ou des valeurs extrêmes ?
3. Identifier les individus les plus éloignés du centre
   - Que signifie leur position dans le plan ?
   - Quelle information donne la distance à l'origine ?
4. Repérer les groupes d'athlètes proches les uns des autres :
   - Que peut-on en déduire sur leurs profils de performance ?
   - Quels athlètes semblent partager des caractéristiques communes ?
5. En observant la disposition générale :
   - Comment se distribuent les athlètes le long du premier axe (Dim 1) ?
   - Et le long du second axe (Dim 2) ?

Dimensionality Reduction using Factoshiny

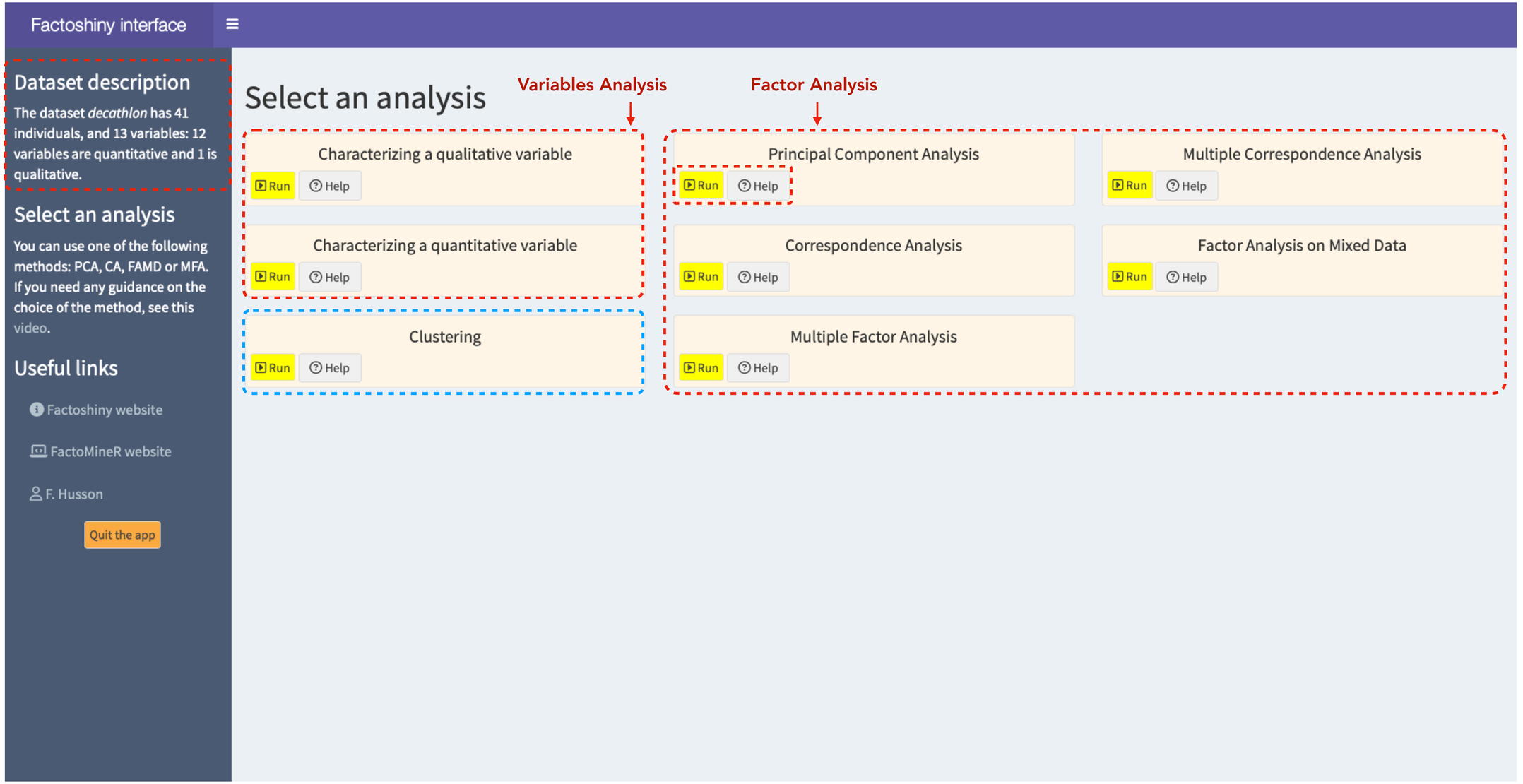
Factoshiny

- Factoshiny is a graphical and interactive interface built on top of the FactoMineR package.
- It allows users to perform multivariate analyses (PCA, CA, MCA, HCPC, MFA, etc.) without writing R code.
- Designed for teaching, quick exploration, and reporting.
- Ideal for presenting analyses to non-technical audiences.
- **How It Works.** Launch Factoshiny with one line of R code:

```
> res = Factoshiny(dataset)
```

- The app opens a web interface (via Shiny) in your browser.

PCA using R


Factoshiny interface
Dataset description
The dataset decathlon has 41 individuals, and 13 variables: 12 variables are quantitative and 1 is qualitative.
Select an analysis
You can use one of the following methods: PCA, CA, FAMD or MFA. If you need any guidance on the choice of the method, see this video.
Useful links
Factoshiny website
FactoMineR website
F. Husson
Quit the app
Select an analysis
Variables Analysis
Factor Analysis
Characterizing a qualitative variable
Run
Help
Characterizing a quantitative variable
Run
Help
Clustering
Run
Help
Principal Component Analysis
Run
Help
Multiple Correspondence Analysis
Run
Help
Correspondence Analysis
Run
Help
Factor Analysis on Mixed Data
Run
Help
Multiple Factor Analysis
Run
Help

PCA using R

Parameters

Results

PCA parameters

Graphical options

Perform clustering after leaving PCA app?

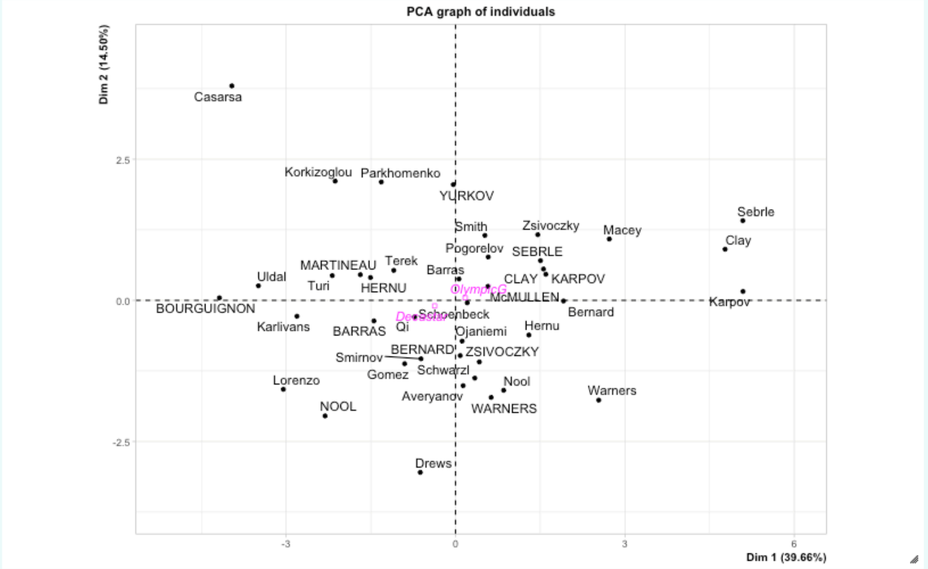
Automatic report

Get the PCA code

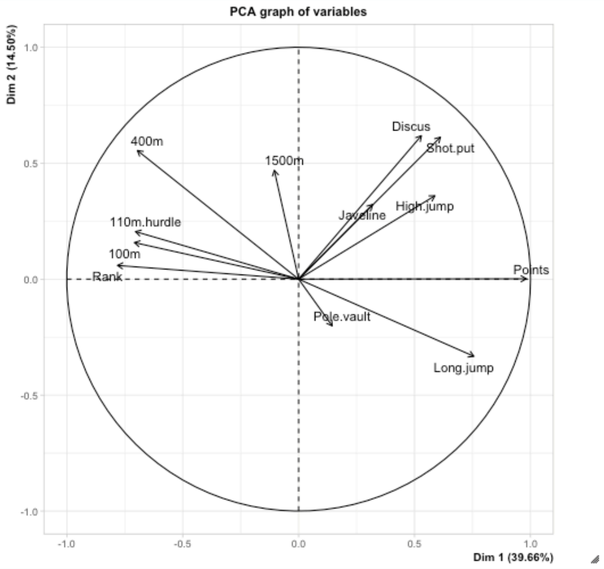
Quit the app

PCA on the dataset decathlon

Graphs | Values | Automatic description of axes | Summary of dataset | Data

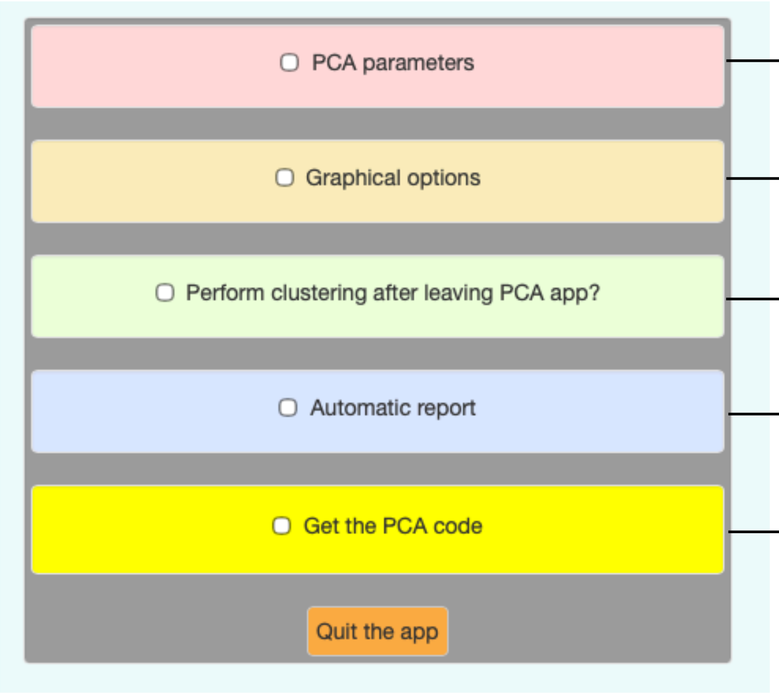


Download as jpg | png | pdf



Download as jpg | png | pdf

PCA using R



Parameters choice: supplementary variables and individuals, and handle missing values.

Plots parameters: eg. choice of principle components, title, label size

Perform clustering (after quitting application)

Generate a report to interpret PCA results

Retrieve personalized code to reproduce results in R studio

TP4: PCA Variables

A envoyer à : mlds.git@gmail.com

Les données proviennent de la World Health Organization (WHO) et regroupent, pour différents pays, des indicateurs moyens de santé, de conditions de vie et de développement socio-économique.
L'objectif de cette étude est de comprendre quels facteurs influencent le plus l'espérance de vie et comment les pays se regroupent selon leurs profils de santé.

1. Importez le jeu de données Life Expectancy – Averaged Dataset dans R, et spécifiez le pays comme indice de ligne.
2. Lancez Factoshiny et effectuez une analyse en composantes principales (ACP) sur les variables quantitatives uniquement.
3. Analyse des variables.
    a. Changez la palette de couleur selon la contribution et garder uniquement les 15 première variables.
    b. Identifiez les variables positivement corrélées et négativement corrélées.
    c. Repérez les variables qui contribuent le plus à la formation du premier axe.
    d. Que semble représenter l'axe 1 ?
    e. Que semble représenter l'axe 2 ?

TP4: PCA Variables

A envoyer à : mlds.git@gmail.com

4. Analyse des individus.
    a. Quels pays sont proches les uns des autres ? Que peut-on dire de leur profil de santé ?
    b. Changez la coloration des individus selon la contribution et garder uniquement les 100 premiers.
    c. Qu'observez-vous sur la relation entre variables et groupes de pays ?
    d. Les pays à forte espérance de vie se situent-ils dans la même direction que certaines variables (PIB, vaccination, etc.) ?
5. En colorant les points selon la région (continent) uniquement :
    a. Quelle régions regroupent les pays à plus forte espérance de vie ?
    b. Quelles régions sont associées à des niveaux plus faibles de développement ou de santé ?
    c. En affichant à nouveau les noms des pays, les pays d'une même région sont-ils regroupés ou dispersés ? Quel continent semble le plus hétérogène en termes de profil de santé ?
6. **Conclusion.** Rédigez un court texte pour résumer vos observations.