Introduction to R, Rstudio, RAnalyticalFlow

Cours M1 IA
Promotion 16
« Systèmes Intelligents & Multimédia »

- Open source, development-flexible, extensible
- Large number of statistical and numerical methods
- High quality visualization and graphical tools
- Extended by a very large collection of rapidly developing packages

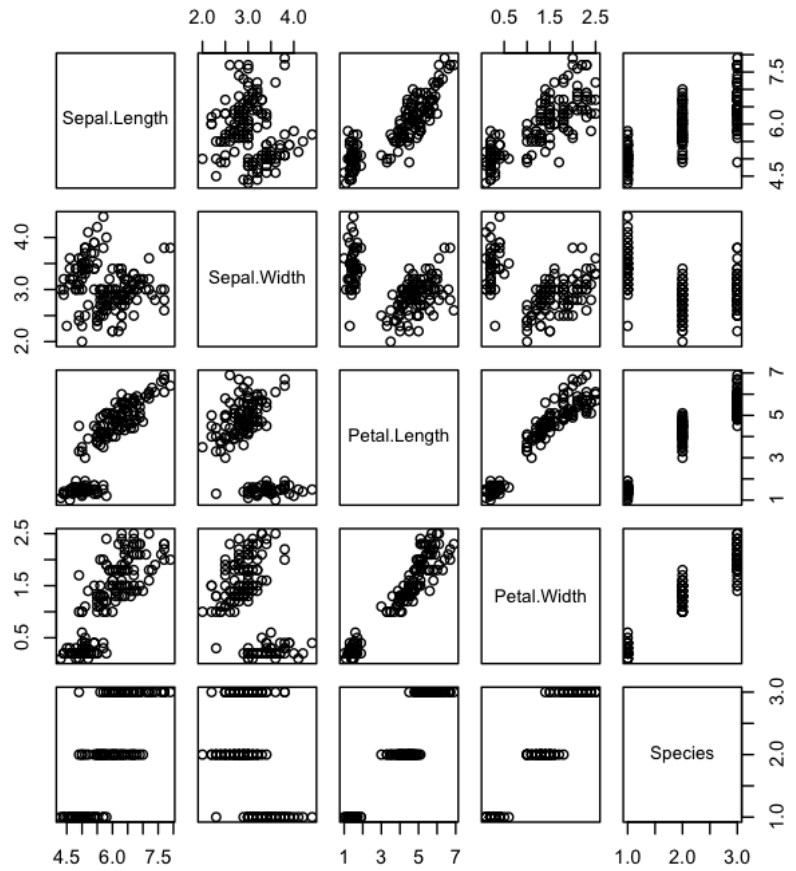
Console R

/Volumes/DataHD/Users/jdz

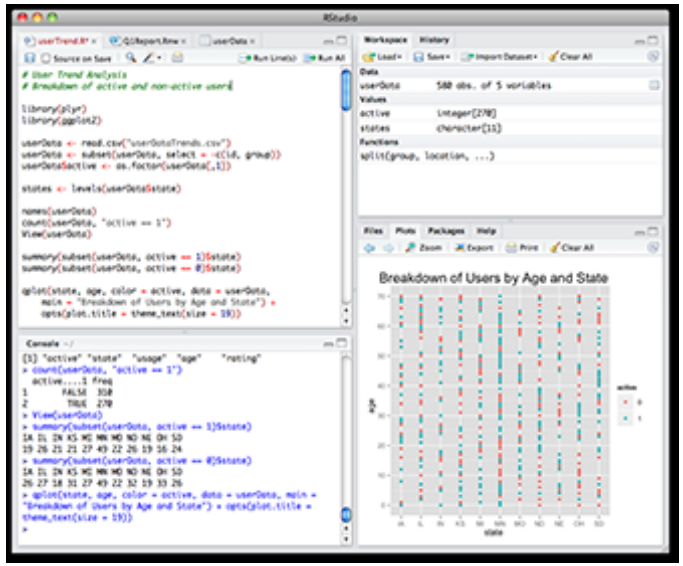
Recherche dans l'aide

```
>
>
>
>
>
>
>
> data(iris)
>
> plot(iris)
>
>
>
```

Quartz 2 [*]

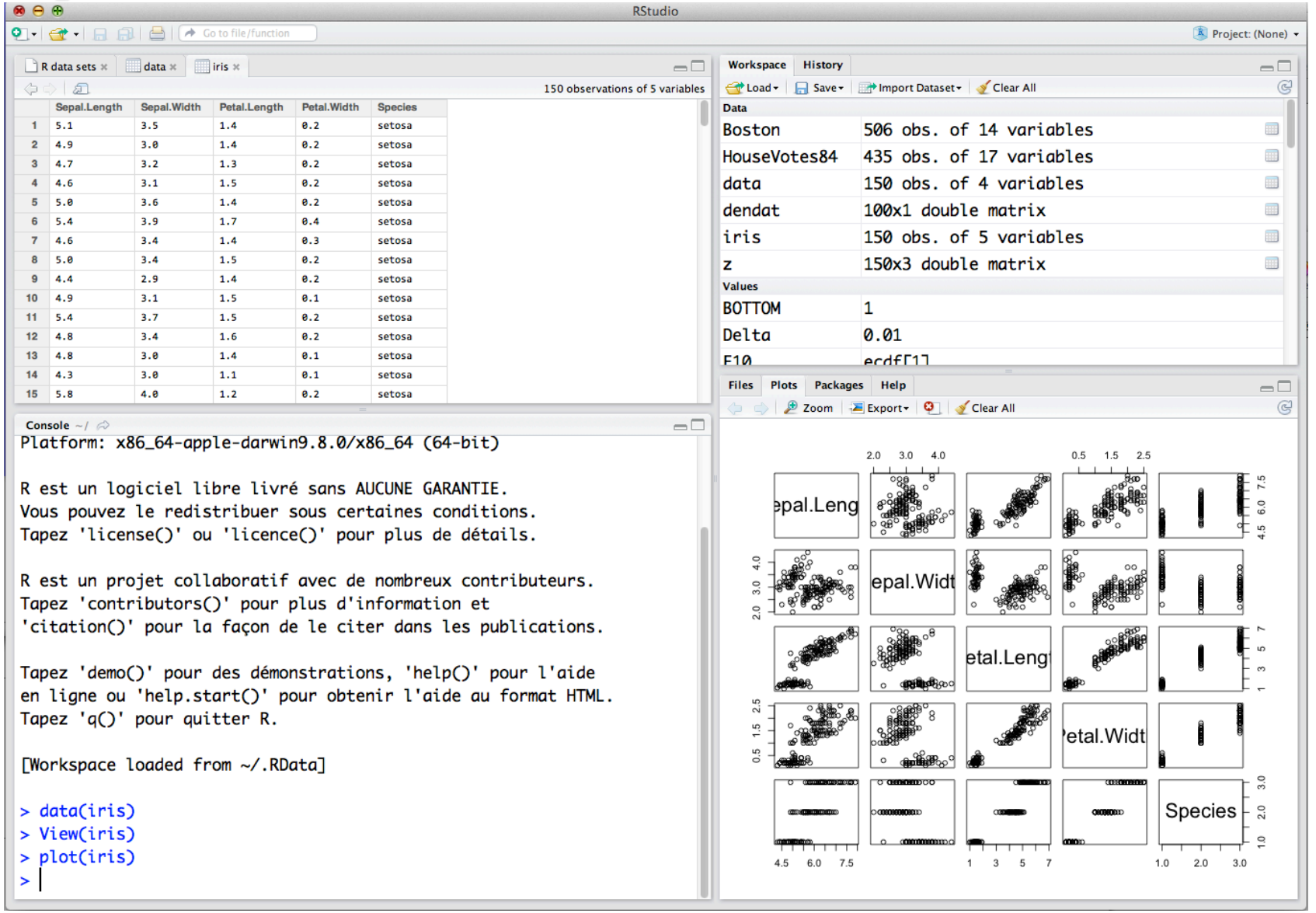


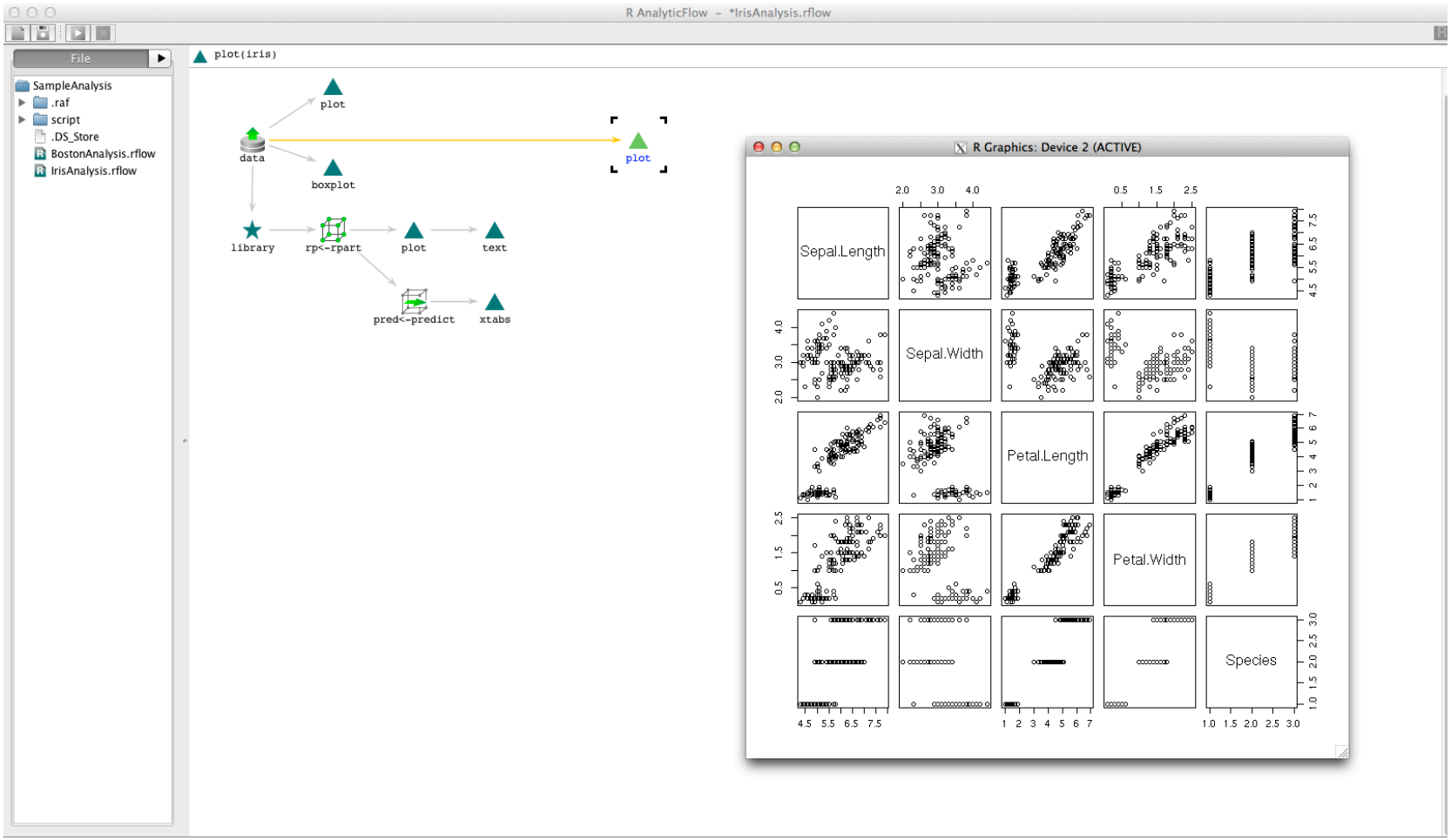
RStudio



- 4 windows
  - Editor,
  - Console,
  - History,
  - Files/plots
- Code completion
- Easy access to help (F1)
- One step Sweave (latex/R) pdf generation
- Searchable history
- Keyboard Shortcuts
  - http://www.rstudio.org/docs/using/keyboard_shortc

RStudio


RStudio
Go to file/function
Project: (None)
R data sets
data
iris
150 observations of 5 variables
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
7 4.6 3.4 1.4 0.3 setosa
8 5.0 3.4 1.5 0.2 setosa
9 4.4 2.9 1.4 0.2 setosa
10 4.9 3.1 1.5 0.1 setosa
11 5.4 3.7 1.5 0.2 setosa
12 4.8 3.4 1.6 0.2 setosa
13 4.8 3.0 1.4 0.1 setosa
14 4.3 3.0 1.1 0.1 setosa
15 5.8 4.0 1.2 0.2 setosa
Console ~/
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)
R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.
R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.
Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.
[Workspace loaded from ~/.RData]
> data(iris)
> View(iris)
> plot(iris)
>
Workspace History
Load Save Import Dataset Clear All
Data
Boston 506 obs. of 14 variables
HouseVotes84 435 obs. of 17 variables
data 150 obs. of 4 variables
dendat 100x1 double matrix
iris 150 obs. of 5 variables
z 150x3 double matrix
Values
BOTTOM 1
Delta 0.01
Files Plots Packages Help
Zoom Export Clear All
Species


R AnalyticFlow - *IrisAnalysis.rflow
File
SampleAnalysis
.raf
script
.DS_Store
BostonAnalysis.rflow
IrisAnalysis.rflow
plot(iris)
plot
data
boxplot
plot
library
rpc-rpart
plot
text
pred<-predict
xtabs
R Graphics: Device 2 (ACTIVE)
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species

Descriptive Statistics in R

Mean	`> mean(x);` `> mean(x,trim=a)`
Median	`> median(x)`
Mode	`> sort(table(x))`
Standard deviation	`> sd(x)`
Variance	`> var(x)`
the median absolute deviation	`> mad(c(x))`
interquartile range	`> IQR(x)`
Range	`> range(x)`

Scatterplot Matrix

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	6.5	2.8	4.6	1.5	versicolor
5	5.7	2.8	4.5	1.3	versicolor
6	6.3	3.3	4.7	1.6	versicolor
7	6.7	2.5	5.8	1.8	virginica
8	....	.....	....	....	.....

- Iris dataset
- 150 flowers
- 5 variables

CC BY Goingslo, flickr

Les données des indiens Pima

- The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix.
- 9 Variables (8 continuous, 1 categorical)
  1) pregnant: Number of times pregnant
  2) Glucose : Plasma glucose concentration at 2 hours in an oral glucose tolerance test
  3) Diastolic : Diastolic blood pressure (mm Hg)
  4) Triceps : Triceps skin fold thickness (mm)
  5) Insulin : 2-Hour serum insulin (mu U/ml)
  6) Bmi : Body mass index (weight in kg/(height in metres squared))
  7) Diabetes : Diabetes pedigree function
  8) Age : Age (years)
  9) Test : diabetes (coded 0 if negative, 1 if positive)

Intermede en R : quantiles

```
# Chargement de la librairie
library(faraway)
# Chargement des données PIMA
data(pima)
?pima
str(pima)
summary(pima)
pima$diastolic
# Calcul la somme des femmes qui ont 0
# pour la variable diastolic
sum(pima$diastolic == 0)

# On remplace les valeurs nuls par NA
pima$diastolic[ pima$diastolic == 0 ] <-
 NA
pima$glucose[ pima$glucose == 0 ] <- NA
pima$triceps [ pima$triceps == 0 ] <- NA
pima$insulin [ pima$insulin == 0 ] <- NA
pima$bmi [ pima$bmi == 0 ] <- NA
```

Intermede en R : ScatterPlot

```
# Modification des facteurs
pima$test <- factor(pima$test)
levels(pima$test) <- c("negativ", "positiv")
summary(pima$test)

#some boxplot
boxplot(pima$triceps, main="Skin at triceps in mm")

#parallel boxplots
boxplot(diabetes ~ test , pima, main="Parallele
 Boxplots")

### scatterplot
plot(diastolic ~ bmi,pima)

### scatterplotmatrix
plot(pima, pch="+",main="Scatterplotmatrix")
```

Intermede en R : PCA

```
# The ground data are set to be the pima data
MyData <- pima

# Performs a principal components analysis on
 GroundData
MyData.pca <- prcomp(MyData[,-9])
# Display the coefficients of the two first
 components
signif(MyData.pca$rotation[,1],digits = 2)
signif(MyData.pca$rotation[,2],digits = 2)

# The Data2D is built from the two first components
# and the column result of the test
class <- MyData[,9]
Data2D <- data.frame(cbind(MyData.pca$x[,1:2],
 class))
names(Data2D) <- c("PC1","PC2","test")
```