METODI INFORMATICI PER IL DATA SCIENCE – INTRODUZIONE

Vedremo

- come essere uno Scienziato dei dati
- i Tipi di Dati
- i cinque passi della Scienza dei dati

Obiettivi della Lezione

La Scienza dei dati è una disciplina che ha sperimentato una rapida crescita negli ultimi decenni

Gli Stati Uniti hanno recentemente nominato il primo “esperto di scienza dei dati”, Data Scientist, **Dhanurjay “DJ” Patil** su sollecitazione delle società che si occupano di nuove tecnologie

Per spiegare il Data Science utilizzeremo R per tutti gli esempi di codice. Per questo dovrete avere in dotazione un computer (Linux, Mac o Windows) dotato di R 4.1 e Rstudio

Preliminari: R

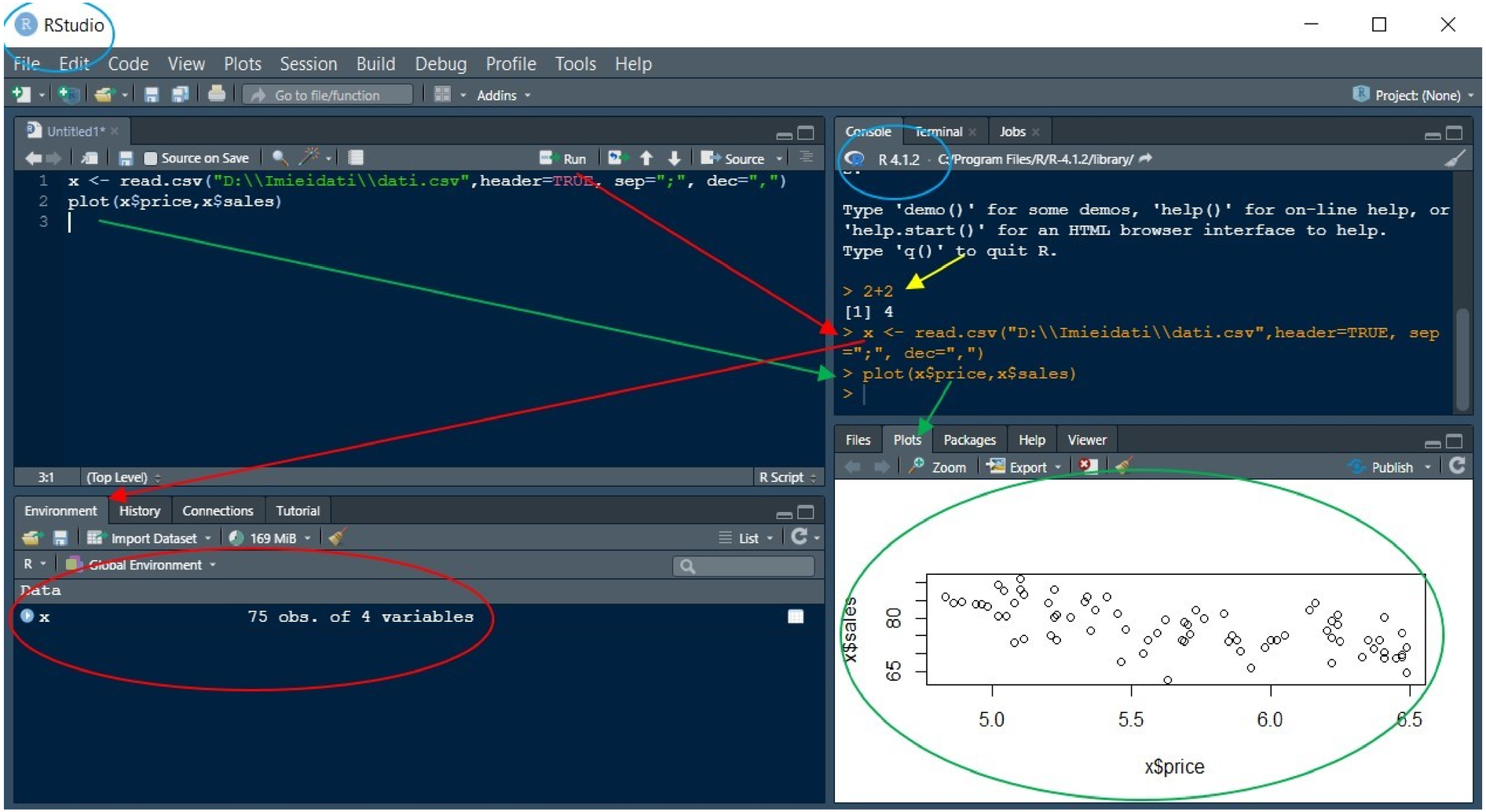
- R è un linguaggio di programmazione per il calcolo statistico e la grafica supportato dalla **R Foundation for Statistical Computing** (https://www.r-project.org/foundation/)
- Fondatore **Ross Ihaka** e **Robert Gentlemen** (1995) ed è attualmente sviluppato da **R Development Core Team**
- Gran parte del sistema stesso è scritto nel dialetto R di S. Per attività ad alta intensità di calcolo, **codice C, C ++ e Fortran collegato e chiamato in fase di esecuzione**
- S è un linguaggio di programmazione, sviluppato da **John Chambers** (Bell laboratori). **Bell Labs ha sviluppato anche Unix** e **C**
- Dal 1997, sviluppo internazionale

Preliminari: R (segue)

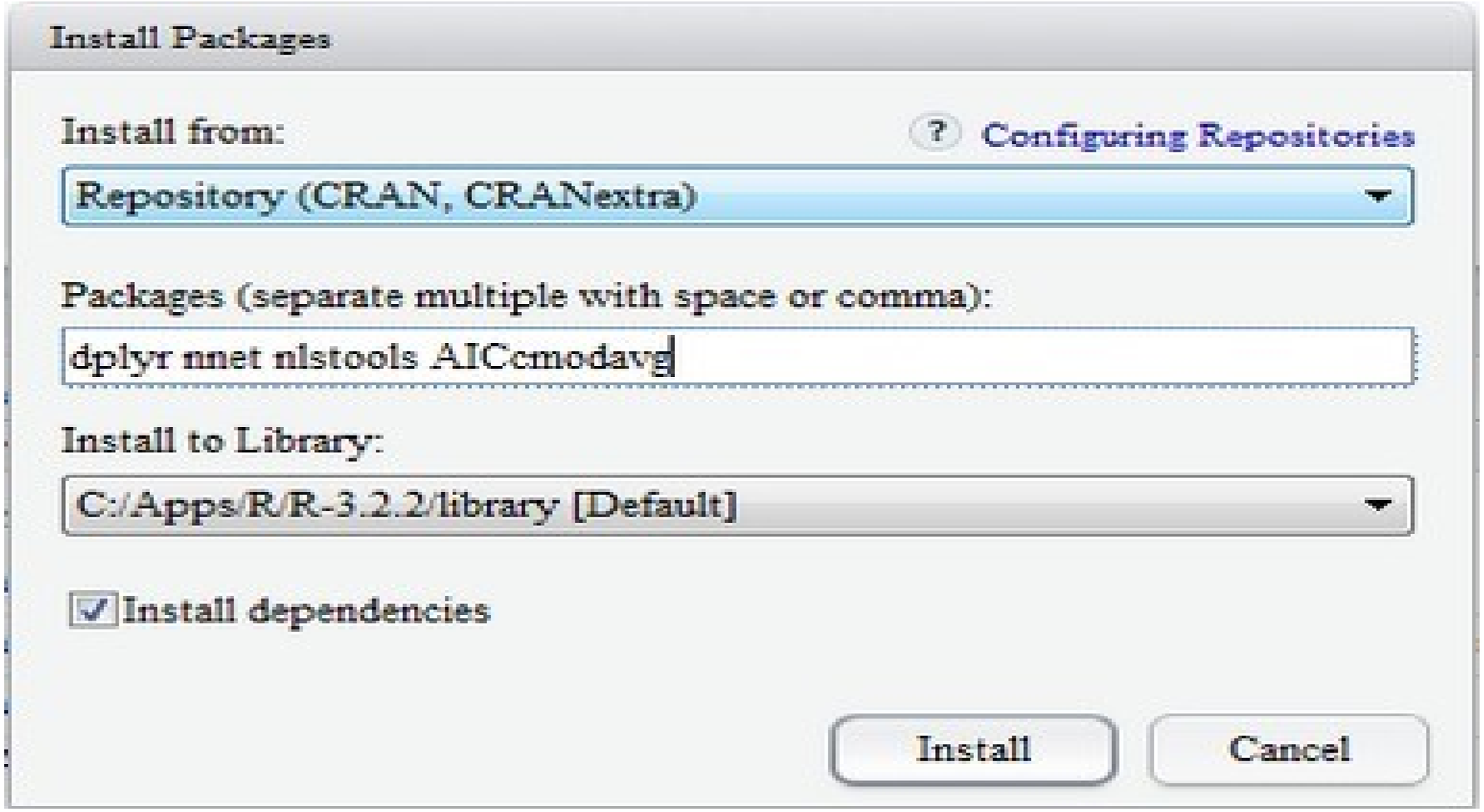
- R è disponibile come Software Libero secondo i termini del Software Libero Licenza pubblica generale GNU della Fondazione sotto forma di codice sorgente
- Compila e funziona su un'ampia varietà di piattaforme UNIX e simili sistemi (inclusi FreeBSD e Linux), Windows e MacOS
- Il codice sorgente è disponibile e può essere modificato (open source)
- http://www.r-project.org
- http://cran.r-project.org
- http://www.rstudio.com/
- molti corsi on line o libri (http://adv-r.had.co.nz/)

RStudio e R

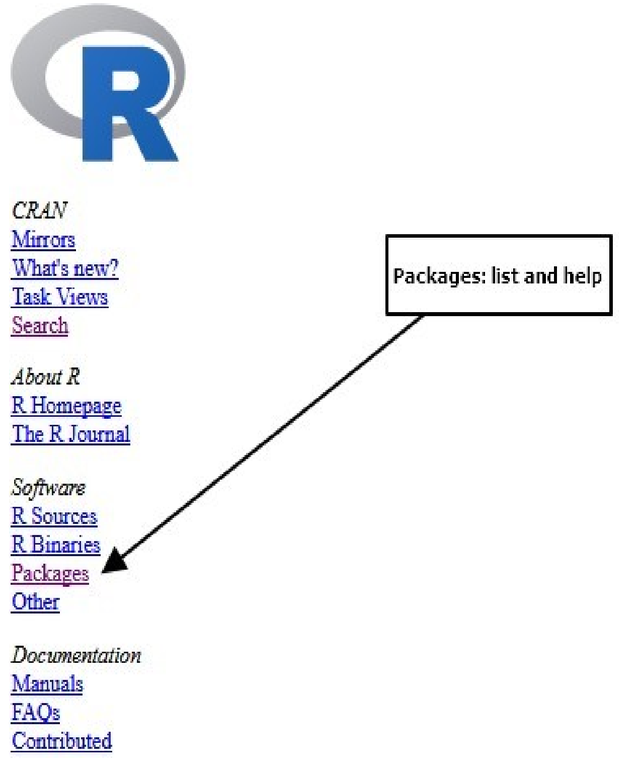
RStudio è un Integrated Development Environment (IDE) per R. Area divisa in 4 panel (Editor, Console, Environment e Plots)



R base e pacchetti (Packages)



https://cran.r-project.org/



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be co
do not know what this means, you probably do not want to do it!

- The latest release (2019-04-26, Planting of a Tree) R-3.6.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corre
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently
email.

Nella console digitare library(**nomepacchetto**)

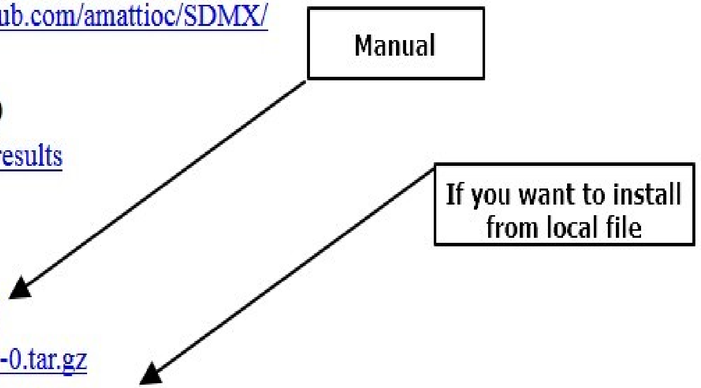
Packages manuale

RJSDMX: R Interface to SDMX Web Services

Provides functions to retrieve data and metadata from providers that disseminate data by means of SDMX web services. SDMX (Statistical Data and Metadata eXchange) is a standard that has been developed with the aim of simplifying the exchange of statistical information. More about the SDMX standard and the SDMX Web Services can be found at: <http://sdmx.org>.

Version:	2.1-0
Depends:	R (≥ 3.0.0), rJava (≥ 0.8-8), zoo (≥ 1.6-4)
Published:	2018-08-22
Author:	Attilio Mattiocco, Diana Nicoletti, Gianpaolo Lopez, Banca d'Italia
Maintainer:	Attilio Mattiocco <attilio.mattiocco at bancaditalia.it>
BugReports:	https://github.com/amattioc/SDMX/issues
License:	EUPL
Copyright:	Banca d'Italia
URL:	https://github.com/amattioc/SDMX/
NeedsCompilation:	no
SystemRequirements:	Java (>= 7)
CRAN checks:	RJSDMX results

Downloads:



Reference manual:	RJSDMX.pdf
Package source:	RJSDMX_2.1-0.tar.gz
Windows binaries:	r-devel: RJSDMX_2.1-0.zip, r-release: RJSDMX_2.1-0.zip, r-oldrel: RJSDMX_2.1-0.zip
OS X binaries:	r-release: RJSDMX_2.1-0.tgz, r-oldrel: RJSDMX_2.1-0.tgz
Old sources:	RJSDMX archive

Era dell’informazione

- **Nel Diciannovesimo secolo,** con l’era industriale, il genere umano inizia ad esplorare la produzione industriale attraverso l’uso di gigantesche invenzioni meccaniche
- **Nel Ventesimo secolo,** ormai abili nella creazione di grandi macchine, ci si pone l’obiettivo di renderle sempre più piccole e veloci. L’era industriale viene rimpiazzata dall’**era dell’informazione**. Le macchine sono utilizzate per raccogliere e conservare informazioni

Era dei dati

- L’era dell’informazione, nella continua ricerca di generare dati, ha fatto letteralmente esplodere la produzione di dati elettronici. Secondo le stime, nel 2011 abbiamo creato circa 1.800 miliardi di GB di dati
- Abbiamo un’enorme quantità di dati e ne creiamo sempre di più. Abbiamo costruito macchine sempre più piccole in grado di raccogliere dati “24/7” e ora il nostro compito consiste nel capirne il senso
- Siamo **nell’era dei dati**

Tipi di dati

- **Dati strutturati (organizzati):** dati ordinati in una struttura a righe e colonne, dove ogni riga rappresenta un'unica osservazione e le colonne rappresentano le caratteristiche di tale osservazione
- **Dati non strutturati (non organizzati):** dati in formato libero, normalmente testo, audio grezzo o segnali che devono essere analizzati meglio per poter essere organizzati

La **Scienza dei dati** è l'arte e la scienza che consiste nel trarre conoscenza dai dati

Dati quantitativi e qualitativi

Nella maggior parte dei casi, quando si parla di dati quantitativi, si parla di un dataset strutturato con una rigida struttura a righe e colonne (perché i dati non strutturati difficilmente hanno delle caratteristiche così ben separate). Un motivo in più per cui il passo di pre-elaborazione è così importante

- **Dati quantitativi:** dati che possono essere descritti tramite numeri. Su di essi è possibile eseguire semplici operazioni matematiche, compresa la somma
- **Dati qualitativi:** dati che non possono essere descritti tramite numeri e semplici operazioni matematiche. Questi dati, in genere, vengono descritti usando delle categorie e un linguaggio “naturale”

Dati strutturati: Aspetti statistici

- **Cross-section:** le osservazioni disponibili sono relative a individui diversi
- **Serie storiche:** le osservazioni riferite a una (**serie storiche univariate**) o più (**serie storiche multivariate**) grandezze sono protratte nel **tempo**
- **Panel**

Dati Cross-section e in Serie storica

Dati Cross-section

Plausibile considerare un insieme di *N* dati come realizzazioni di *N* variabili casuali indipendenti e identicamente distribuite (i.i.d.)

Serie storiche

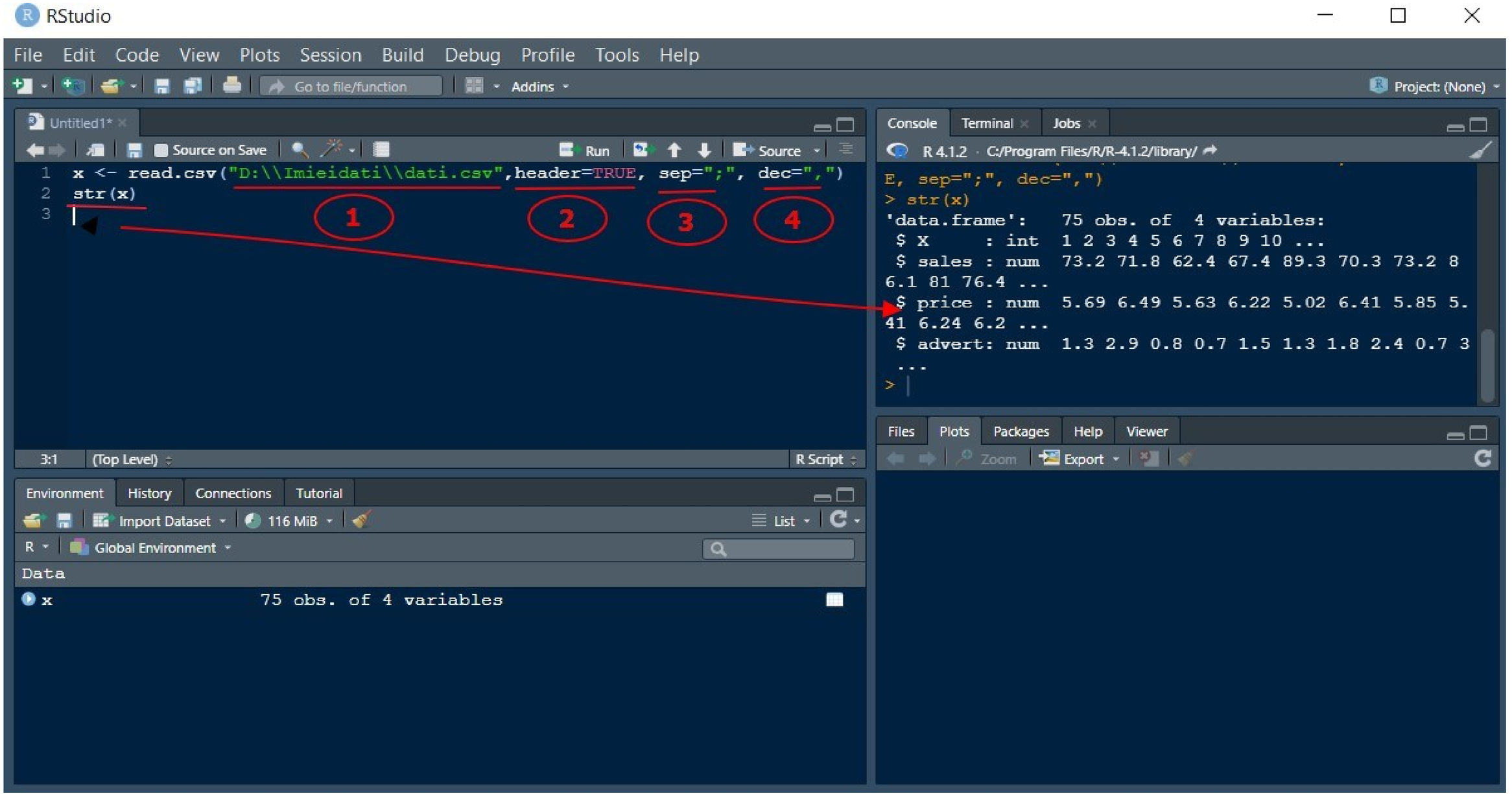
Fondamentale il **tempo**, che ha una **direzione**

- la Serie storica ha una **memoria** (**persistenza**)
- la caratteristica che distingue le Serie storiche dai dati Cross-section è che nelle Serie storiche l'**ordine delle osservazioni** è rilevante

Passi per la Scienza dei dati

1) Porre una domanda interessante
2) Ottenere i dati
3) Esplorare i dati
4) Creare un modello per i dati
5) Comunicare e presentare i risultati

Ottenere i dati: File locale



Ottenere i dati: Database

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Untitled1*

Source on Save | Run | Source

```
1  library(RJSDMX)
2  library("TSsdmx")
3  eurostat <- TSconnect("sdmx", dbname="EUROSTAT")
4  getProviders()
5  dati <- TSget('sts_inpr_m.M.PROD.B-D.NSA|SCA.I15.IT|FR',
6                eurostat,start=2000)
7  ts.plot(dati,col=c("red","black","blue","green"))
8
```

5:1 (Top Level) R Script

Console Terminal Jobs

R 4.1.2 · C:/Program Files/R/R-4.1.2/library/

```
> getProviders()
 [1] "ABS"                "ECB"
 [3] "EUROSTAT"           "ILO"
 [5] "ILO_Legacy"         "IMF2"
 [7] "IMF_SDMX_CENTRAL"   "INEGI"
 [9] "INSEE"              "ISTAT"
[11] "ISTAT_CENSUS_AGR"   "ISTAT_CENSUS_IND"
[13] "ISTAT_CENSUS_POP"   "NBB"
[15] "OECD"               "OECD_RESTR"
[17] "StatsEE"            "UIS"
[19] "UNDATA"             "WB"
[21] "WITS"
>
```

Environment History Connections Tutorial

Import Dataset | 279 MiB | List

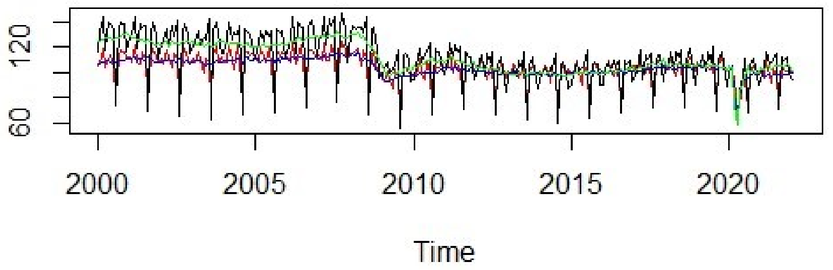
R | Global Environment

Data

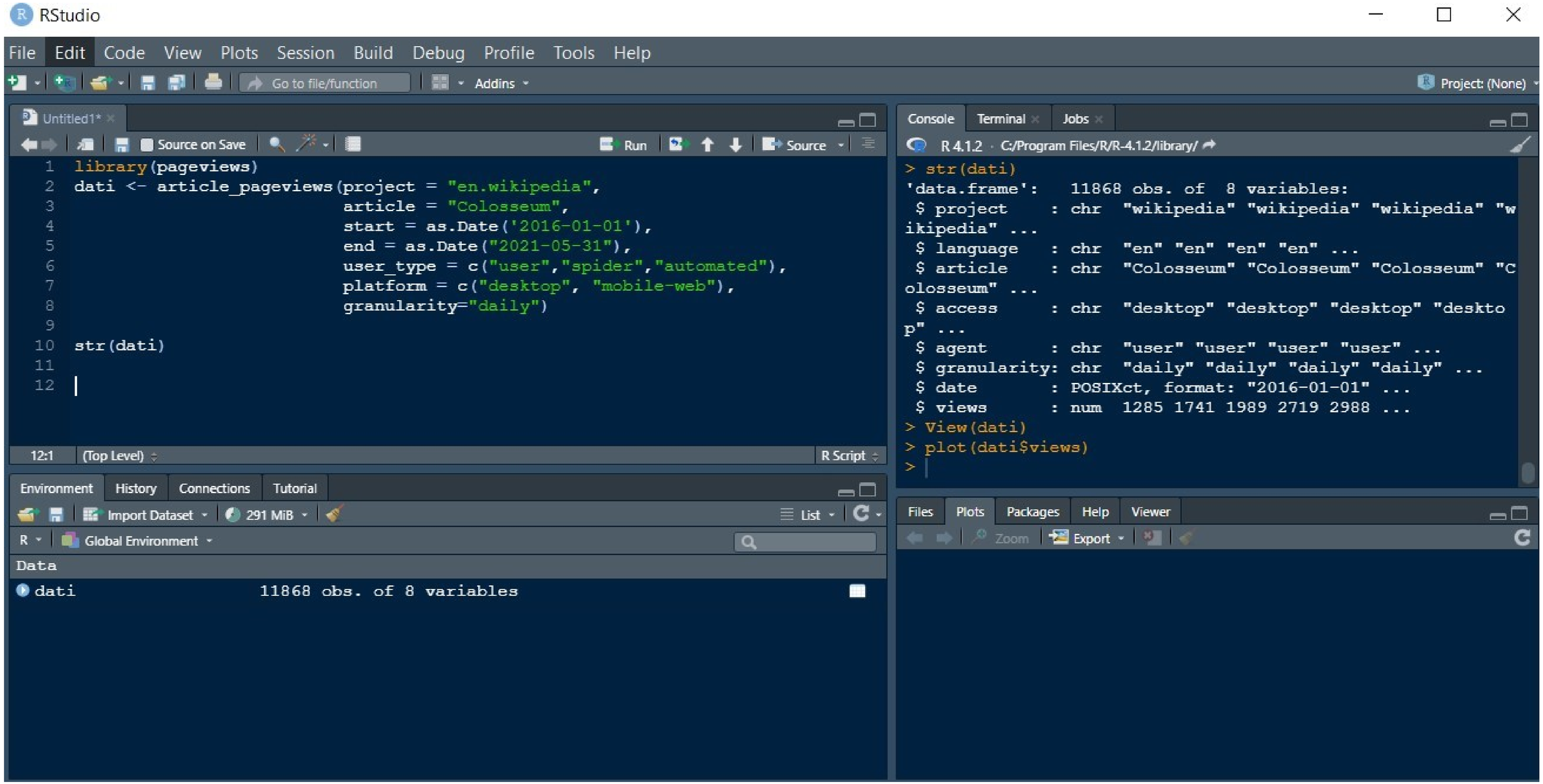
dati	Time-Series [1:266, 1:4] from 2000 to 2022: 106 108 120…
eurostat	Formal class TSsdmxConnection

Files Plots Packages Help Viewer

Zoom | Export | Publish



Ottenere i dati: Dati non strutturati



Esplorare i dati

```
> mydata <- mtcars
>str(mydata)                    # struttura
'data.frame': 32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160  160  108  258  360 ...
 $ hp  : num  110  110  93  110  175  105  245  62  95  123 ...
 $ drat: num  3.9  3.9  3.85  3.08  3.15  2.76  3.21  3.69  3.92  3
 $ wt  : num  2.62  2.88  2.32  3.21  3.44 ...
 $ qsec: num  16.5  17  18.6  19.4  17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Prime statistiche sui dati

```
> class(mydata)                    # "data.frame"
> names(mydata)                    # show list components
> dim(mydata)                      # dimensions of object, if any
> min(mydata); max(mydata)# minimo e massimo
> range(mydata)                 # range
> mean(mydata);median(mydata) #media e mediana
> sd(mydata); # standard deviation
```

Riepilogo sui dati

```
> summary(mydata[,1:3])
      mpg              cyl             disp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8  
 Median :19.20   Median :6.000   Median :196.3  
 Mean   :20.09   Mean   :6.188   Mean   :230.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0  
```

Una prima analisi dei dati

```
> # Tabella di contingenza
> mytable <- table(mtcars$cyl,mtcars$am)
> # cyl will be rows, am will be columns
> mytable
    0  1
4   3  8
6   4  3
8  12  2
```

Verso un Modello dei dati

Covarianza: misura della direzione di una relazione lineare tra due variabili

La relazione tra due variabili può essere studiata attraverso:

- **Analisi di correlazione:**
  - esiste una associazione tra le variabili?
  - c'è una relazione tra consumo e reddito?
- **Analisi di regressione:**
  - come varia il valore di una variabile in conseguenza della variazione di un'altra variabile?
  - analizza la forma della relazione tra variabili

Anscombe data

Anscombe ha utilizzato questo set di dati per dimostrare quando le statistiche riassuntive sono inadeguate per descrivere l'associazione

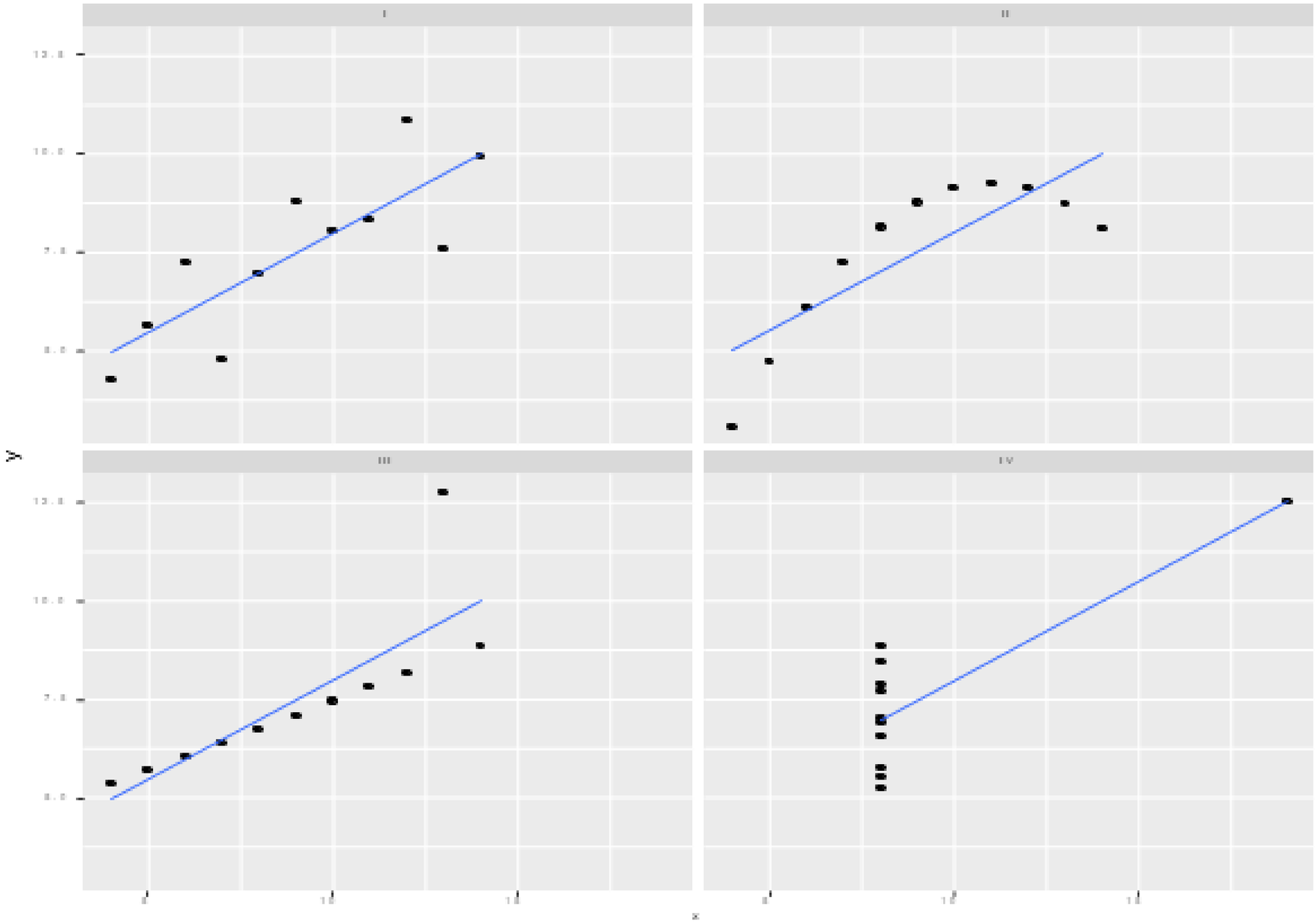
	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.0	9.1	7.5	6.6
2	8	8	8	8	7.0	8.1	6.8	5.8
3	13	13	13	8	7.6	8.7	12.7	7.7
4	9	9	9	8	8.8	8.8	7.1	8.8
5	11	11	11	8	8.3	9.3	7.8	8.5
6	14	14	14	8	10.0	8.1	8.8	7.0
7	6	6	6	8	7.2	6.1	6.1	5.2
8	4	4	4	19	4.3	3.1	5.4	12.5
9	12	12	12	8	10.8	9.1	8.2	5.6
10	7	7	7	8	4.8	7.3	6.4	7.9
11	5	5	5	8	5.7	4.7	5.7	6.9

>

Anscombe data

set	`mean(x)`	`sd(x)`	`mean(y)`	`sd(y)`	`cor(x, y`
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<db
I	9	3.32	7.50	2.03	0.8
II	9	3.32	7.50	2.03	0.8
III	9	3.32	7.5	2.03	0.8
IV	9	3.32	7.50	2.03	0.8

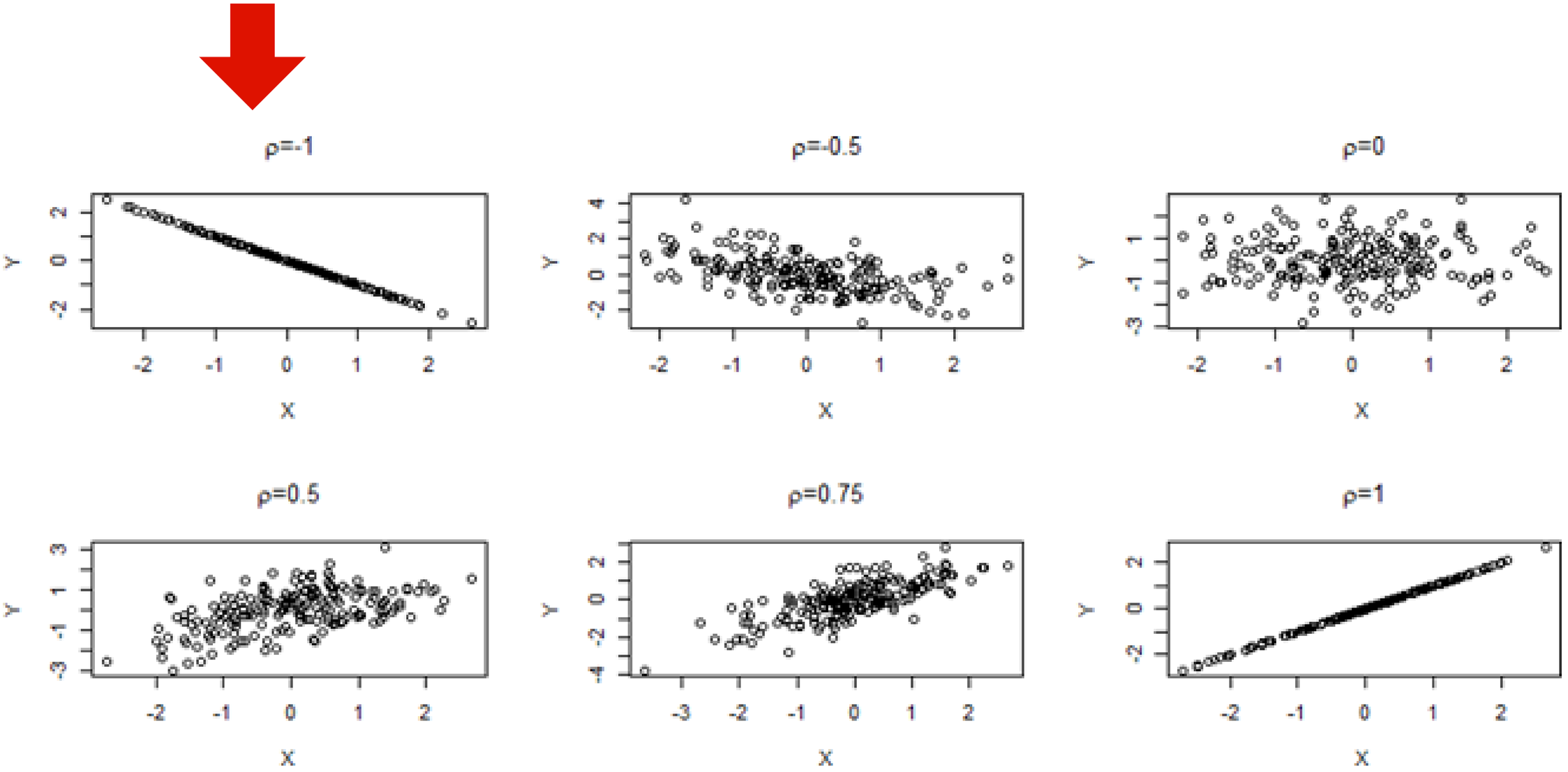
Anscombe data



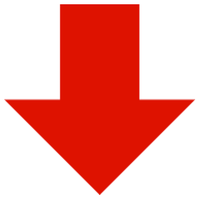
Scatterplot

- Esiste una relazione che può essere descritta da una linea retta (il che significa che c’èuna relazione lineare)?
- Esiste una relazione, che non sia lineare?
- Se il grafico a dispersione delle variabili assomiglia a una nuvola, non vi è alcuna relazione tra le variabili senza procedere ad ulteriori analisi

Correlazione



Riepilogo e conclusioni finali



- la Terminologia di base
- i Tipi di dati e come manipolarli
- i 5 Passi della Scienza dei dati