Lecture Series: Mastering Data Visualization using R

Session-1: Importance of data visualization and data structure in R

Jaynal Abedin
PhD Student, National University of Ireland Galway

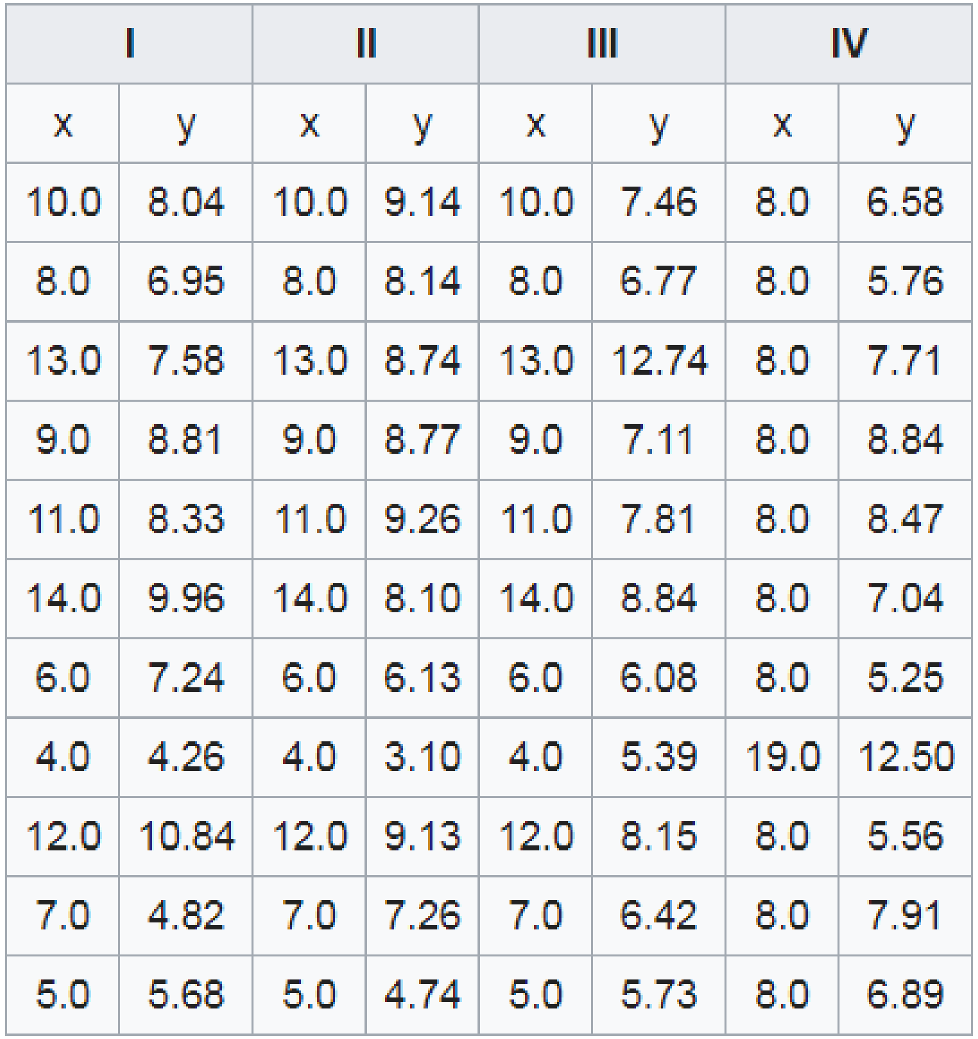
Lecture Series: Mastering Data Visualization using R

Session-1: Importance of data visualization and data structure in R

At the end of this session you will be able

- to understand, the importance of data visualization
- to understand data structure used in R
- to know most popular libraries in R for data visualization

Why we visualize data?



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Can you say anything about this **FOUR** separate dataset?

Why we visualize data?

- Graph is universal language that convey message directly to the mind
- By graphs we can compare, contrasts and show differences
- A graph can show systematic structure

> There is magic in graphs. The profile of a curve reveals in a flash a whole situation -the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces……. Graphs are dynamic, dramatic. They may epitomize an epoch, each dot a fact, each slope an event, each curve a history. Wherever there are data to record, inferences to draw, or facts to tell, graphs furnish the unrivalled means whose power we are just beginning to realize and to apply

Henry D. Hubbard, National Bureau of Standards, Washinton, D.C

Graphic Presentation, Willard C. Brinton, 1939

MAGIC IN GRAPHS

THERE is a magic in graphs. The profile of a curve reveals in a flash a whole situation —the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces.

Graphs carry the message home. A universal language, graphs convey information directly to the mind. Without complexity there is imaged to the eye a magnitude to be remembered. Words have wings, but graphs interpret. Graphs are pure quantity, stripped of verbal sham, reduced to dimension, vivid, unescapable.

Graphs are all inclusive. No fact is too slight or too great to plot to a scale suited to the eye. Graphs may record the path of an ion or the orbit of the sun, the rise of a civilization, or the acceleration of a bullet, the climate of a century or the varying pressure of a heart beat, the growth of a business, or the nerve reactions of a child.

The graphic art depicts magnitudes to the eye. It does more. It compels the seeing of relations. We may portray by simple graphic methods whole masses of intricate routine, the organization of an enterprise, or the plan of a campaign. Graphs serve as storm signals for the manager, statesman, engineer; as potent narratives for the actuary, statist, naturalist; and as forceful engines of research for science, technology and industry. They display results. They disclose new facts and laws. They reveal discoveries as the bud unfolds the flower.

The graphic language is modern. We are learning its alphabet. That it will develop a lexicon and a literature marvelous for its vividness and the variety of application is inevitable.

Graphs are dynamic, dramatic. They may epitomize an epoch, each dot a fact, each slope an event, each curve a history. Wherever there are data to record, inferences to draw, or facts to tell, graphs furnish the unrivalled means whose power we are just beginning to realize and to apply.

HENRY D. HUBBARD
National Bureau of Standards
Washington, D. C.

GRAPHIC PRESENTATION

By

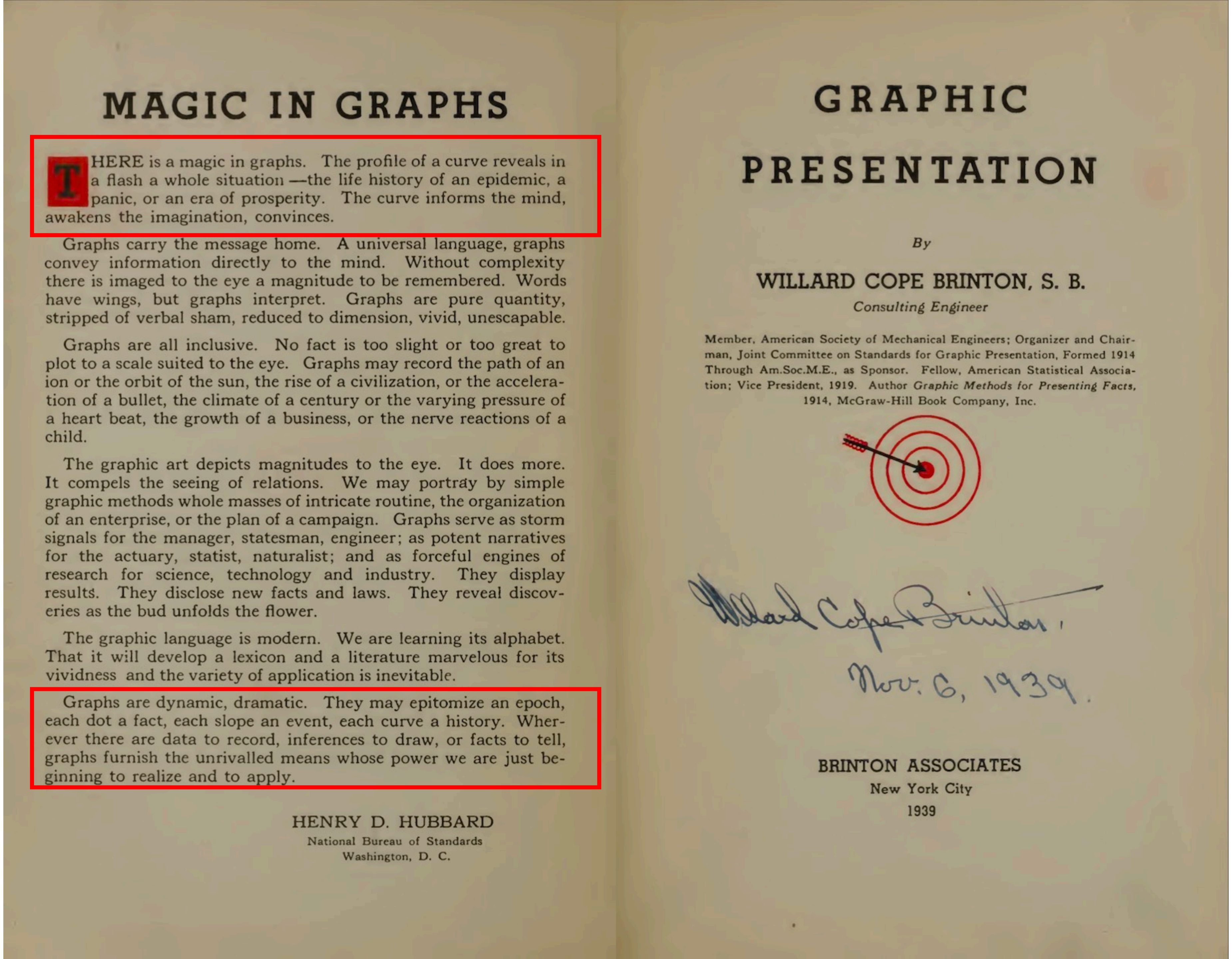
WILLARD COPE BRINTON, S. B.
Consulting Engineer

Member, American Society of Mechanical Engineers; Organizer and Chairman, Joint Committee on Standards for Graphic Presentation, Formed 1914 Through Am.Soc.M.E., as Sponsor. Fellow, American Statistical Association; Vice President, 1919. Author *Graphic Methods for Presenting Facts,* 1914, McGraw-Hill Book Company, Inc.

Willard Cope Brinton.

Nov. 6, 1939.

BRINTON ASSOCIATES
New York City
1939



MAGIC IN GRAPHS

THERE is a magic in graphs. The profile of a curve reveals in a flash a whole situation —the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces.

Graphs carry the message home. A universal language, graphs convey information directly to the mind. Without complexity there is imaged to the eye a magnitude to be remembered. Words have wings, but graphs interpret. Graphs are pure quantity, stripped of verbal sham, reduced to dimension, vivid, unescapable.

Graphs are all inclusive. No fact is too slight or too great to plot to a scale suited to the eye. Graphs may record the path of an ion or the orbit of the sun, the rise of a civilization, or the acceleration of a bullet, the climate of a century or the varying pressure of a heart beat, the growth of a business, or the nerve reactions of a child.

The graphic art depicts magnitudes to the eye. It does more. It compels the seeing of relations. We may portray by simple graphic methods whole masses of intricate routine, the organization of an enterprise, or the plan of a campaign. Graphs serve as storm signals for the manager, statesman, engineer; as potent narratives for the actuary, statist, naturalist; and as forceful engines of research for science, technology and industry. They display results. They disclose new facts and laws. They reveal discoveries as the bud unfolds the flower.

The graphic language is modern. We are learning its alphabet. That it will develop a lexicon and a literature marvelous for its vividness and the variety of application is inevitable.

Graphs are dynamic, dramatic. They may epitomize an epoch, each dot a fact, each slope an event, each curve a history. Wherever there are data to record, inferences to draw, or facts to tell, graphs furnish the unrivalled means whose power we are just beginning to realize and to apply.

HENRY D. HUBBARD
National Bureau of Standards
Washington, D. C.

GRAPHIC PRESENTATION

By

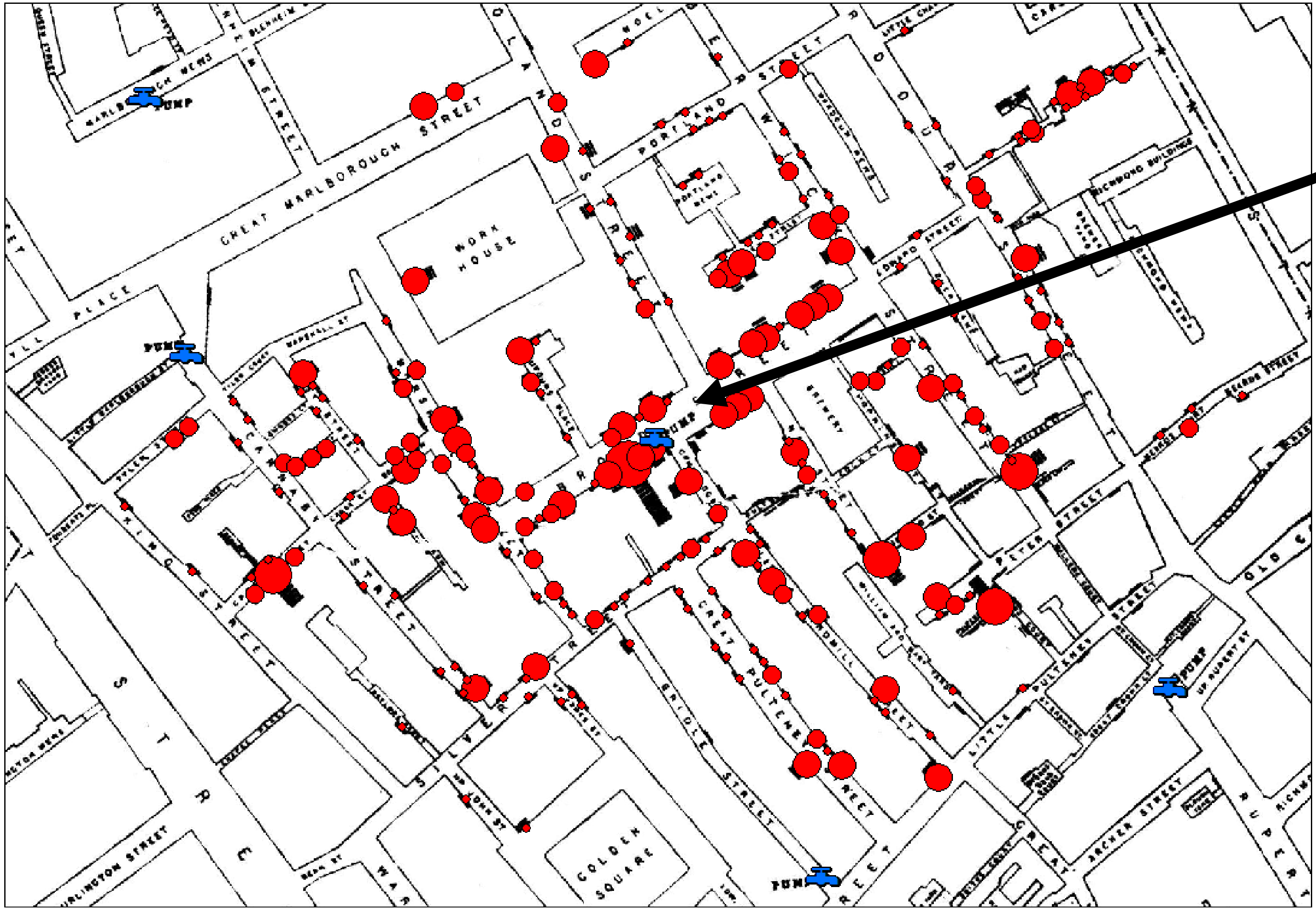
WILLARD COPE BRINTON, S. B.
Consulting Engineer

Member, American Society of Mechanical Engineers; Organizer and Chairman, Joint Committee on Standards for Graphic Presentation, Formed 1914 Through Am.Soc.M.E., as Sponsor. Fellow, American Statistical Association; Vice President, 1919. Author *Graphic Methods for Presenting Facts*, 1914, McGraw-Hill Book Company, Inc.

Willard Cope Brinton.
Nov. 6, 1939.

BRINTON ASSOCIATES
New York City
1939

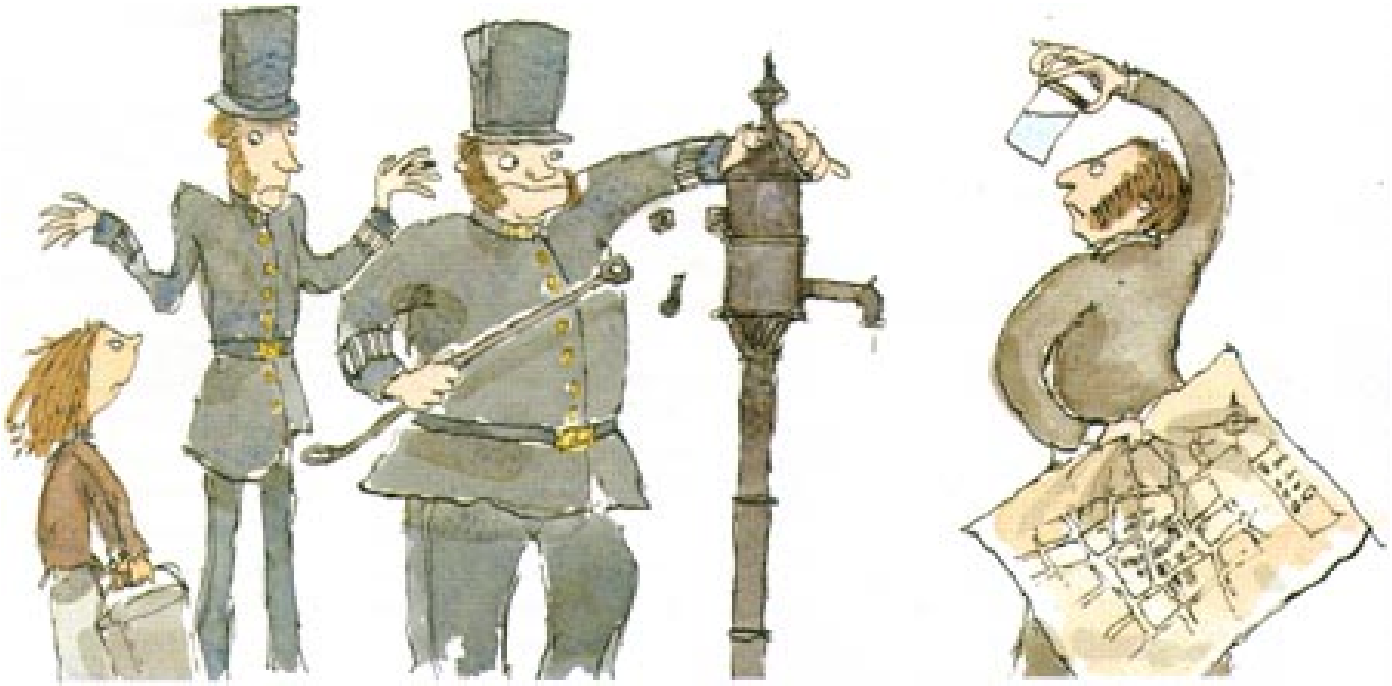
In **1854**, British doctor **John Snow** had a very difficult time to convince doctors and scientists that ***<u>cholera spread through contaminated drinking water</u>***



Most of the fatal cases found around Board Street and it was the pump nearby was the source of water supply

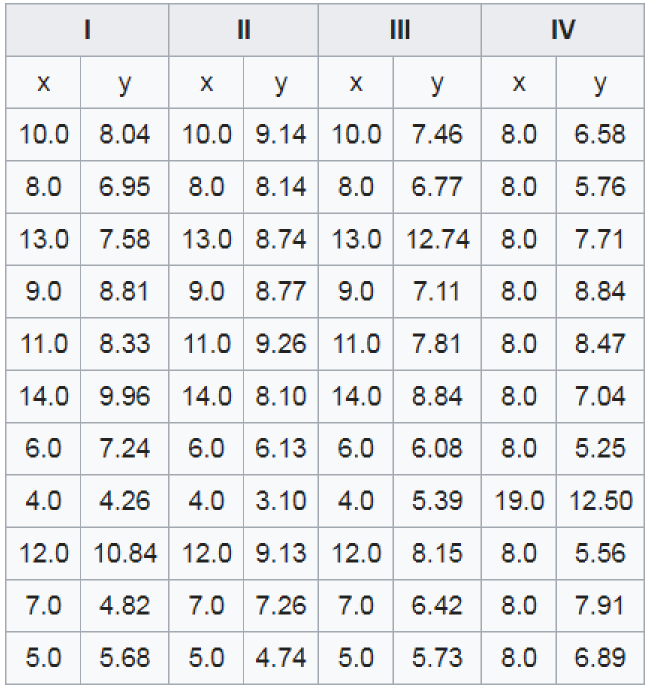
A popular **bubbly drink** of the time was called “sherbet”, which was a spoonful of powder that fizzed **when mixed with water**.

In the **Broad Street area of Soho**, that **water usually came from the Broad Street pump and was, Snow believed, the source for many cases**.

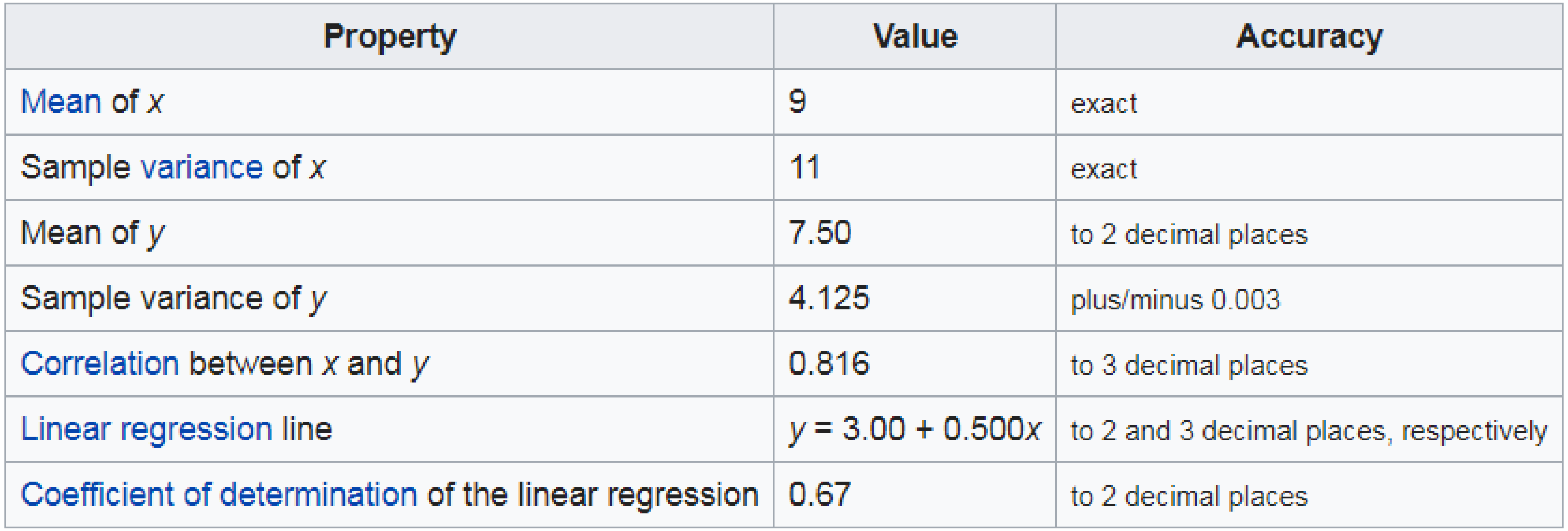


On 7 September 1854, Snow took his research to the town officials and convinced them to take the handle off the pump, making it impossible to draw water.

……….. And the cholera outbreak stopped!!!!

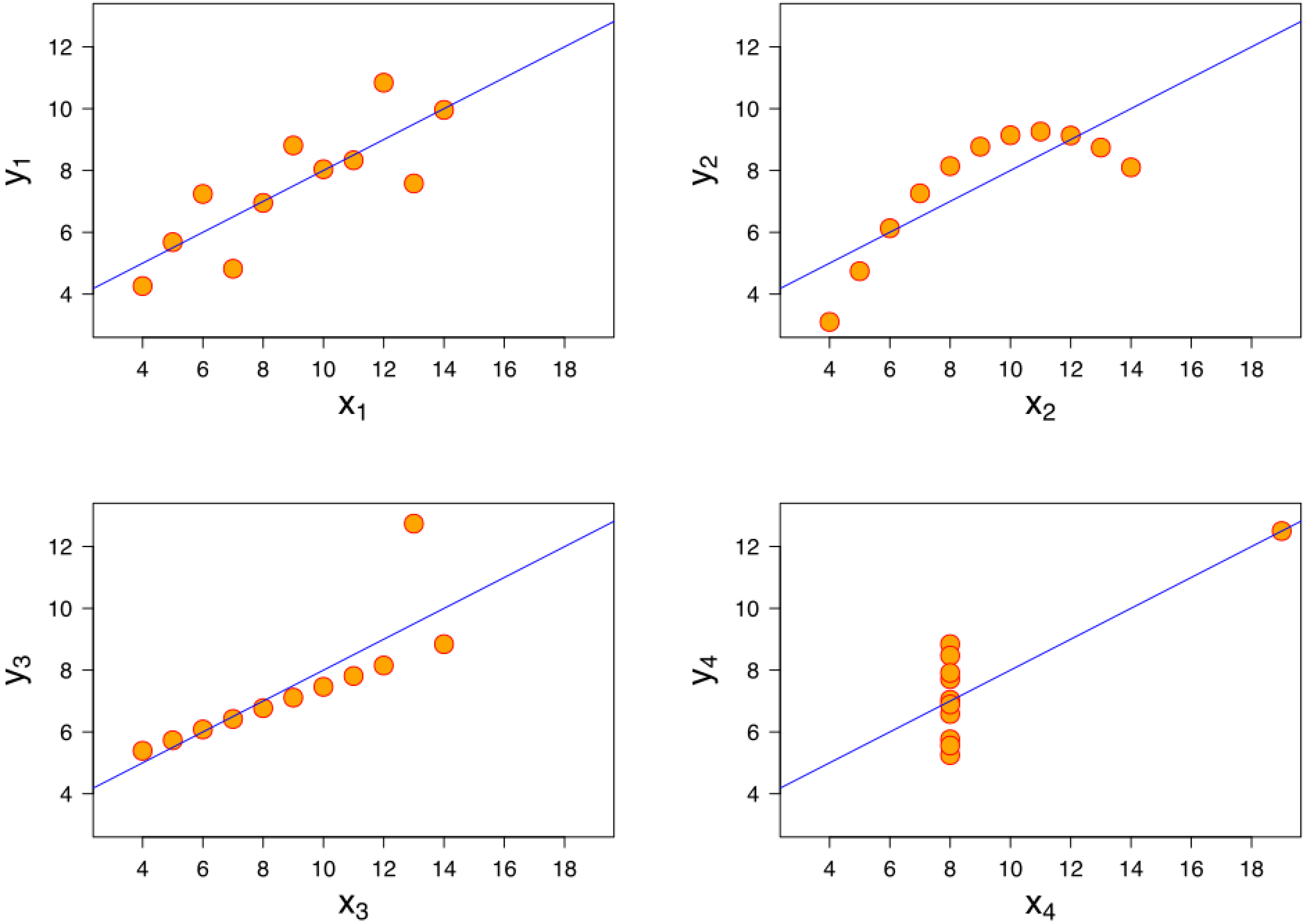


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	plus/minus 0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

Is the numerical summary enough to find the pattern hidden inside the data?????

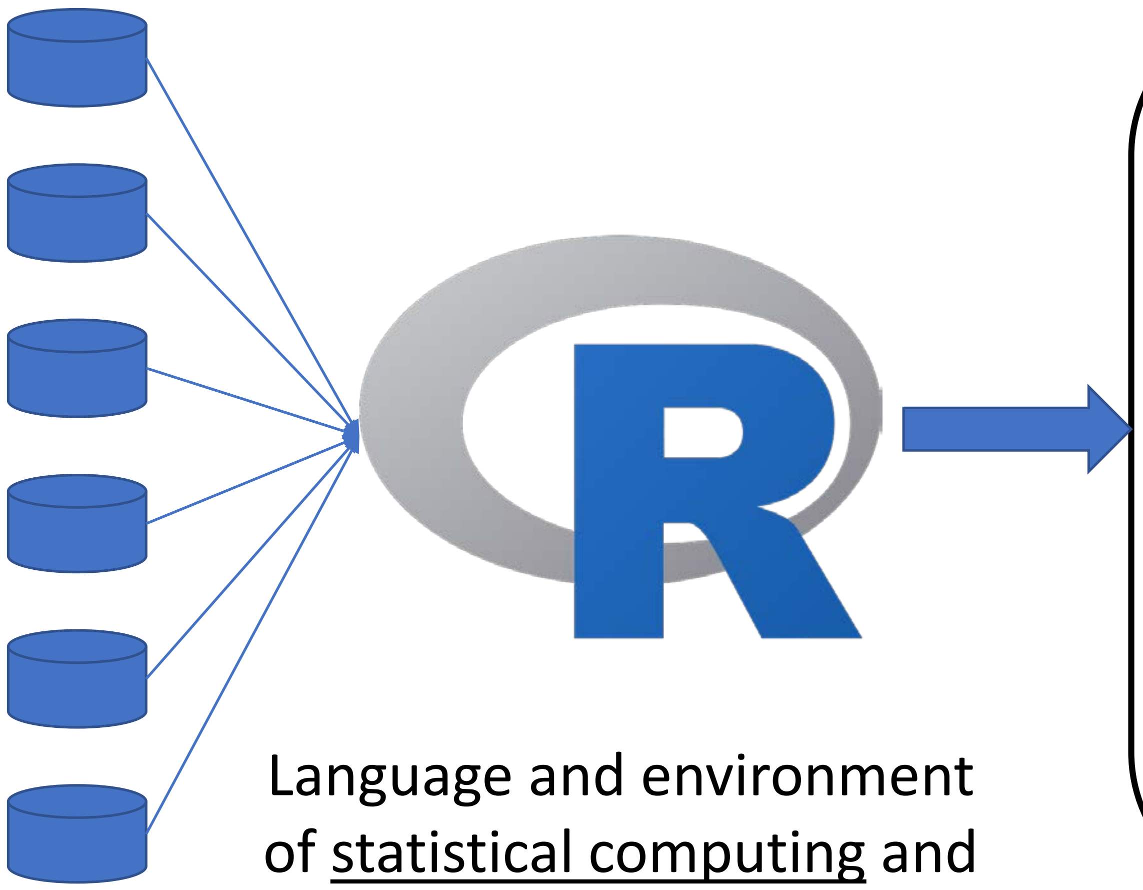


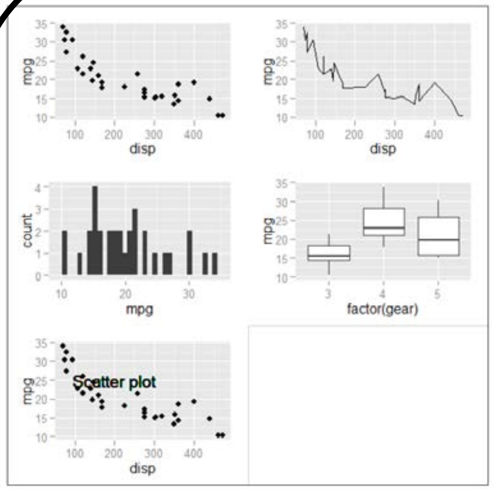
Only the numerical summary is not enough, need to visualize the data

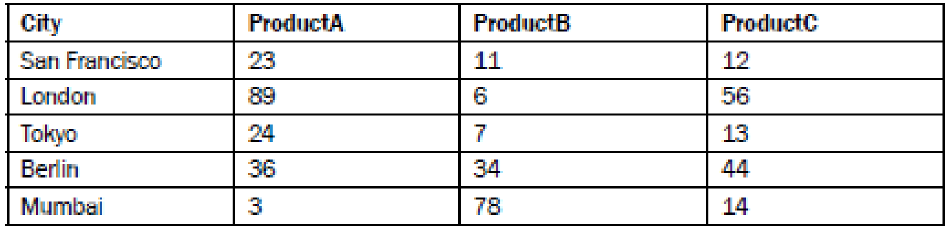
What is R?

Data

Graphs, Report and analytical dashboard







City	ProductA	ProductB	ProductC
San Francisco	23	11	12
London	89	6	56
Tokyo	24	7	13
Berlin	36	34	44
Mumbai	3	78	14

Language and environment of statistical computing and graphics

Data types in R

Data types are usually linked with measurement scale, such as nominal, ordinal, interval and ratio

- Character
- Complex
- Numeric (Integer, Double)
- Logical

Data types in R

Creating various types of data and test

```
cVec <- c("Cricket", "Football", "Basketball", "Rugby")

cVec <- c(game1="Cricket", game2="Football", game3="Basketball", game4="Rugby")

nVec <- c(1:10)
Lvec <- c(TRUE, FALSE, FALSE, TRUE)

nVec2 <- runif( n=5, min=10, max=20)
cVec2 <- sample(x=letters, size= 5, replace = F)
Lvec2 <- nVec2>=13

mixVec <- c(1, 3, "Cricket", "Football", "Basketball", "Rugby")
```

is.x() is a generic function to test the R object’s type, such as, is.character()

Data structure in R

The way of organizing data in R is usually termed as data structure

- Vector (it contains only one type of data, either numeric or character or logical or complex
- Matrix
- Array
- Data frame
- List

Creating a single variable: vector

- The default function `c()`, is the most convenient way to create a single variable (vector) in R
- This function takes the element of the vector as input separated by a comma as : `c(item1, item2, item3)`
- To assign a variable name (also called object in R), we use assignment operator "<-", as: `a <- c(3,5)`, this tells R that the value 3 and 5 is assigned to an object called "a"

Matrix in R

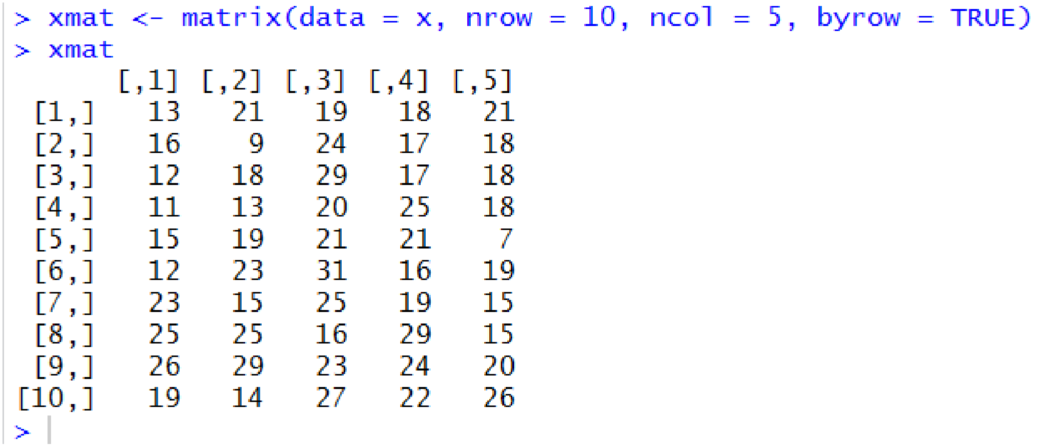
- A matrix in R is a two-dimensional representation of a dataset
- Each column **must** be same length
- Each column **must contain same types of data** either all numeric, or all character, or all logical or all complex number
- You cannot store different types of data in different columns

```
x <- c(13, 21, 19, 18, 21, 16, 21, 24, 17, 18,
       12, 18, 29, 17, 18, 11, 13, 20, 25, 18,
       15, 19, 21, 21, 7, 12, 23, 31, 16, 19,
       23, 15, 25, 19, 15, 25, 25, 16, 29, 15,
       26, 29, 23, 24, 20, 19, 14, 27, 22, 26)
```

```
xmat <- matrix(data = x, nrow = 10, ncol = 5, byrow = TRUE)
```

The output will look like the following image in next slide

Matrix in R



```
> xmat <- matrix(data = x, nrow = 10, ncol = 5, byrow = TRUE)
> xmat
      [,1] [,2] [,3] [,4] [,5]
 [1,]   13   21   19   18   21
 [2,]   16    9   24   17   18
 [3,]   12   18   29   17   18
 [4,]   11   13   20   25   18
 [5,]   15   19   21   21    7
 [6,]   12   23   31   16   19
 [7,]   23   15   25   19   15
 [8,]   25   25   16   29   15
 [9,]   26   29   23   24   20
[10,]   19   14   27   22   26
> 
```

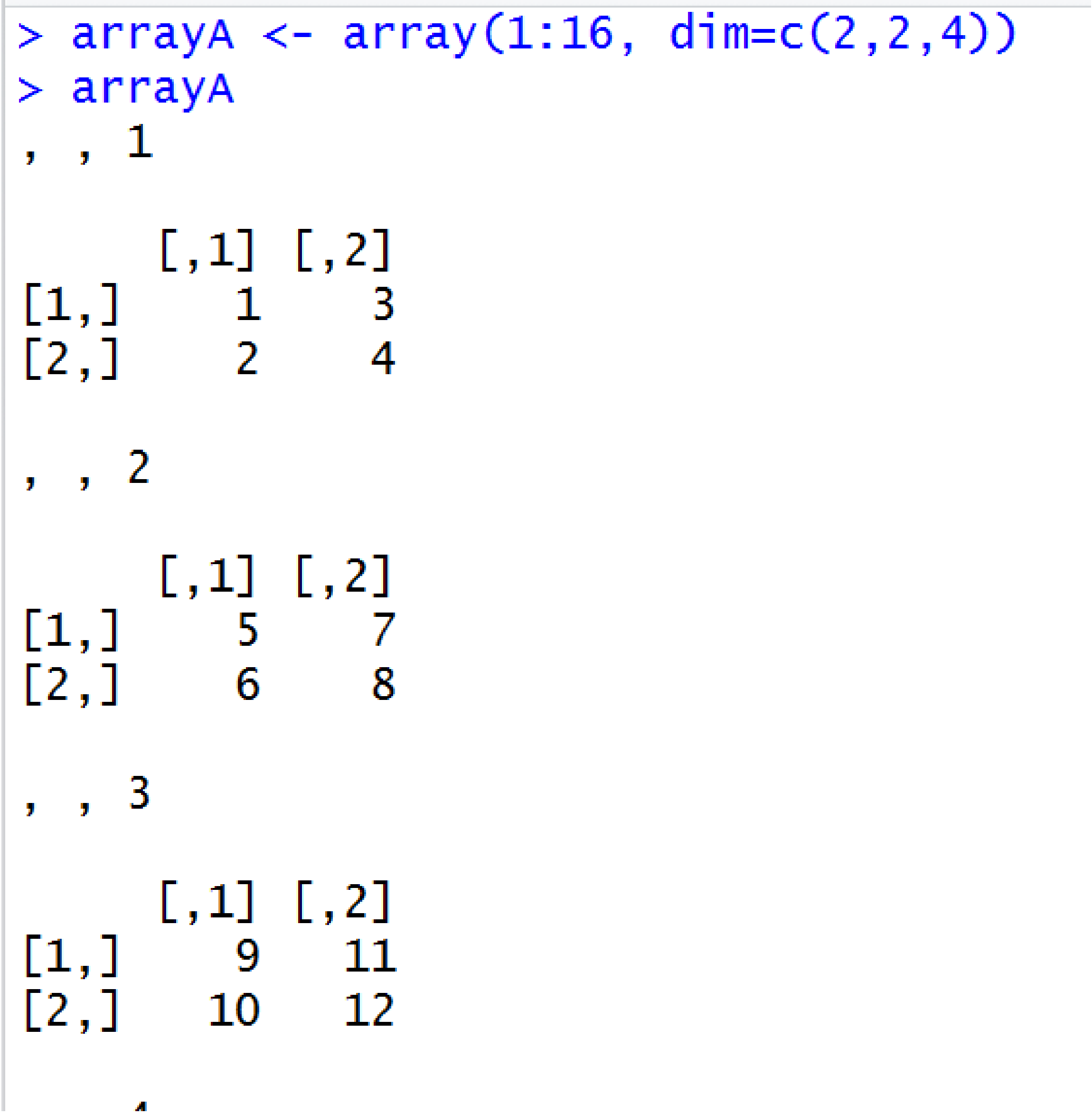
We had a vector “x” and now it is converted into a matrix, the question is, how the elements will be organized? Where to place which value?

Data frame in R

- A data frame is also a two-dimensional representation of data
- Each column **must** be same length
- The columns of a data frame could be a mix of numeric, character, logical and complex number

Array in R

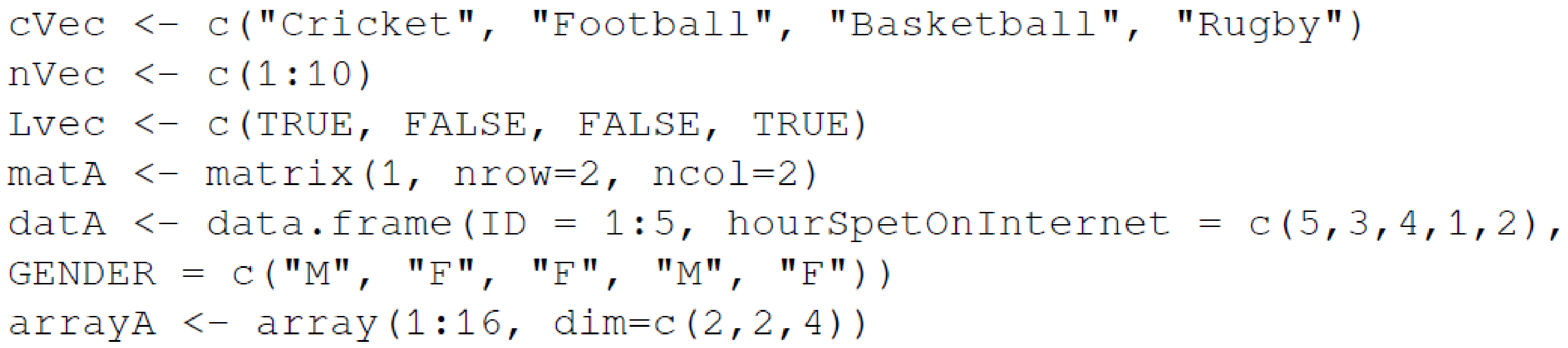
- The way to organize data with more than two dimension
- Each component **must** same types of data
- The number of dimension could be any



```
> arrayA <- array(1:16, dim=c(2,2,4))
> arrayA
, , 1

     [,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2

     [,1] [,2]
[1,]    5    7
[2,]    6    8

, , 3

     [,1] [,2]
[1,]    9   11
[2,]   10   12
```

List in R

- List is the natural generalization of a data frame
- It can contain heterogeneous data, e.g., mix of numeric and character with different length
- It can contain a vector, matrix, data frame and a list itself



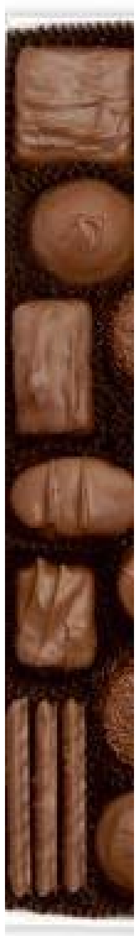
```
cVec <- c("Cricket", "Football", "Basketball", "Rugby")
nVec <- c(1:10)
Lvec <- c(TRUE, FALSE, FALSE, TRUE)
matA <- matrix(1, nrow=2, ncol=2)
datA <- data.frame(ID = 1:5, hourSpetOnInternet = c(5,3,4,1,2),
GENDER = c("M", "F", "F", "M", "F"))
arrayA <- array(1:16, dim=c(2,2,4))
```

listA <- list(cVec, nVec, Lvec, matA, datA, arrayA)

listB <- list(vector1 = cVec, vector2 = nVec, vector3 = Lvec, matrix1 = matA, data1 = datA, array1 = arrayA)

How can we access elements of an array?

Data structure in R (Intuitive analogy)



Matrix

All milk chocolates with different ***colour*** and ***shapes*** and possibly from different brand

Data Frame

Along with ***milk chocolate*** it could contain ***dark chocolates***, ***nuts*** and other things from variety of brands

List

It contains everything

Vector

Milk chocolate
with different shapes

R Libraries and Repositories

- **Library:** A customized collection of pre-defined functions to perform particular tasks
- **Repository:** Collection of available libraries:
  - CRAN
  - CRAN (extras)
  - BioC software
  - BioC annotation
  - BioC experiment
  - BioC extra
  - R-Forge
  - rforge.net
  - Omegahat

Installing R Libraries

- **Net Installation**
  - Open R console
  - Connect the PC with internet
  - Type:
    `install.packages(“pkgname”,lib=“location”,dependencies=TRUE)`

Note: Location is the storage folder where the installation file will be stored. By default it will go into “My documents”

```
install.packages("ggplot2", dependencies = TRUE, lib = "c:/rPackages")
```

R libraries for data visualization

The following are the most popular library for data visualization in R

- `lattice` (pre-ggplot2 era popular library for data visualization)
- `ggplot2` (Implementation of grammar of graphics)
- `ggiraph` (to make the graph interactive)
- `dygraphs` (Creating HTML/Java Script graph of time series data)
- `googleVis` (Creating interactive visualization by connecting R with Google API)
- `ggmap, leaflet` (to create spatial visualization)
- `plotly` (creating interactive graph for web application)

We will primarily use ggplot2 for the examples and also show the other options when necessary

Materials to study further

- [Book]: Graphic Presentation, Willard C. Brinton, 1939
- [Book]: The visual display of quantitative information, Edward R. Tufte, 2001
- [Book]: ggplot2: Elegant Graphics for Data Analysis, Hadley Wickham, 2009

The images and excerpt of text about Dr. John Snow has been used from the following website

http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html

Lecture Series: Mastering Data Visualization using R

Session-2: Introduction to **`ggplot2`**

At the end of this session you will be able

- to understand, the syntax structure of ggplot2 library
- to map and define aesthetic of data
- to create basic graphs using `ggplot2`

Date: **14 Dec 2017**

Time: 10:00 pm Bangladesh

Duration: 30 min

f LIVE