Introduction to Multivariate Analysis

Woojoo Lee

Instructor: Dr.Woojoo Lee

- Office: 5N 435B
- Phone: (032)860-7649
- E-mail is the best way to contact me.

Prerequisites

- Basic Statistics, Basic Calculus, Basic linear algebra

Statistical software

- We will use R (http://www.r-project.org) for data analysis.

R programming

- R is a free statistical package.
- There are huge amounts of resources for R.

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

From

http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all&_r=0

Statistical Software (SAS & R)

Features	Stata	SPSS	SAS	R
Learning curve	Steep/gradual	Gradual/flat	Pretty steep	Pretty steep
User interface	Programming/point-and-click	Mostly point-and-click	Programming	Programming
Data manipulation	Very strong	Moderate	Very strong	Very strong
Data analysis	Powerful	Powerful	Powerful/versatile	Powerful/versatile
Graphics	Very good	Very good	Good	Excellent
Cost	Affordable (perpetual licenses, renew only when upgrade)	Expensive (but not need to renew until upgrade, long term licenses)	Expensive (yearly renewal)	Open source

* http://www.r-project.org/index.html

R in the enterprise

- R is often used as an analysis platform fc ɜny.

History of R

- Brief history
  - **1993**: Research project in Auckland, NZ (Ross Ihake, Robert Gentleman)
  - **1995**: R Released as open-source software
  - **1997**: R core group formed
  - **2000**: R 1.0.0 released (February 29)
  - **2003**: R Foundation founded
  - **2004**: First international user conference in Vienna
  - **2015:** R Consortium founded

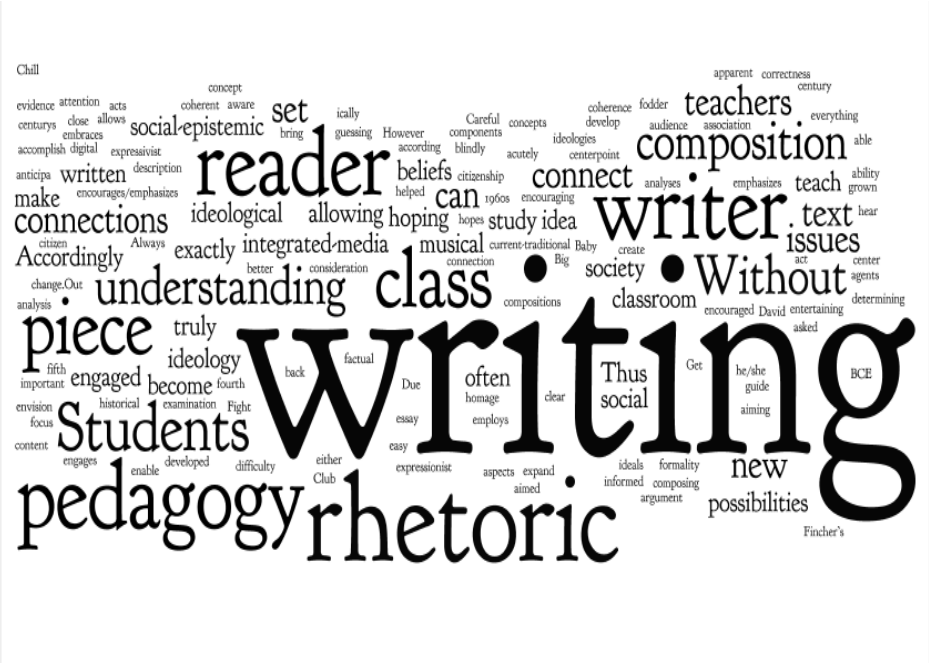
History of R

- Brief history

Using R

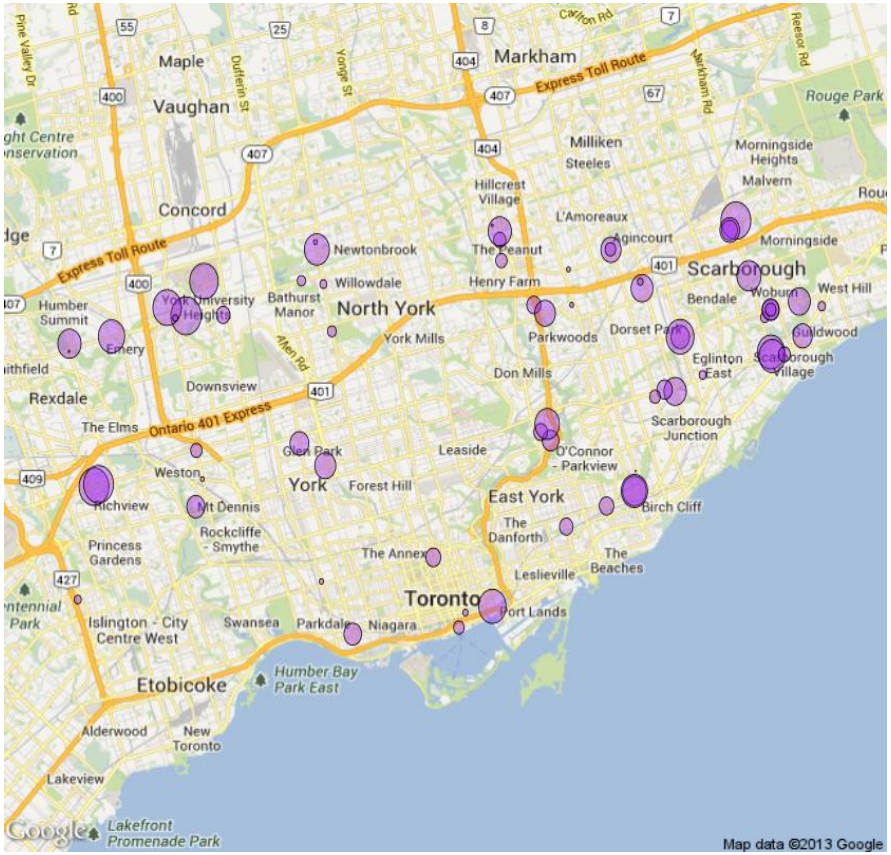
- wordcloud package

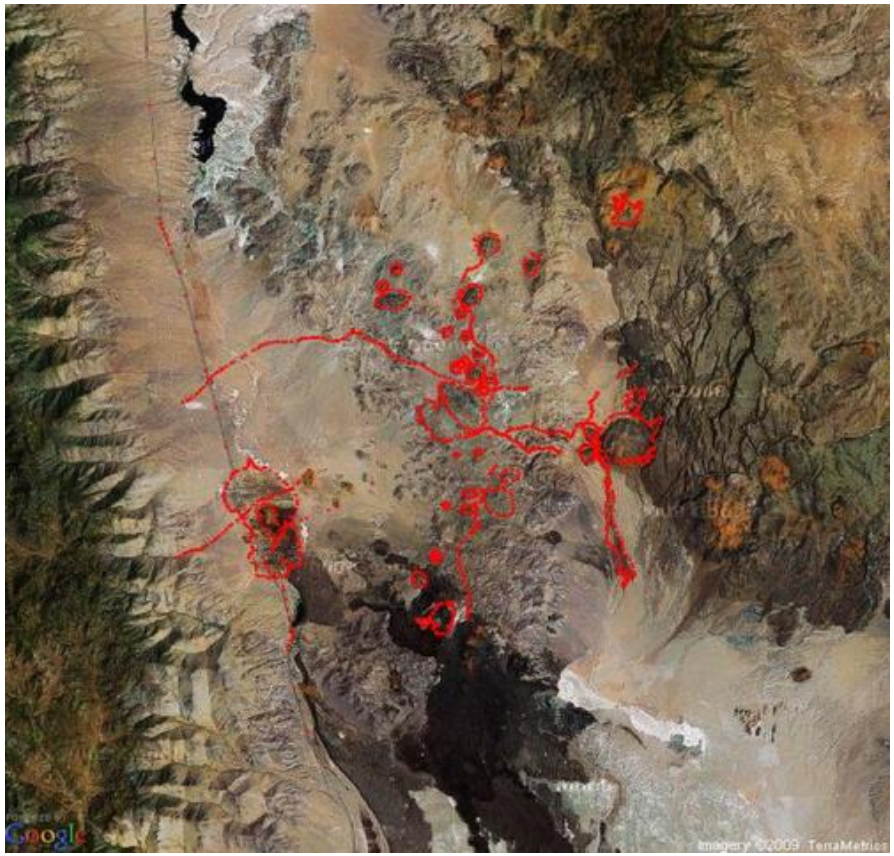
```
require(tm)
require(wordcloud)
data(crude)
                    ⋮
wordcloud(d$word,d$freq)
```



Using R

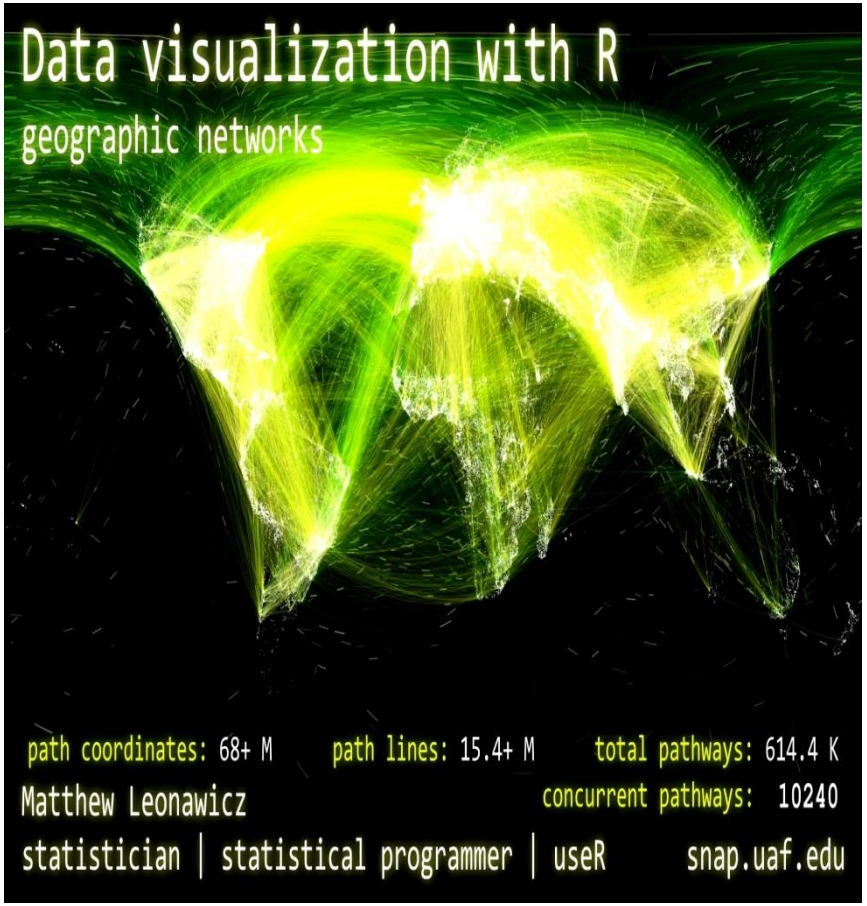
- RgoogleMaps package

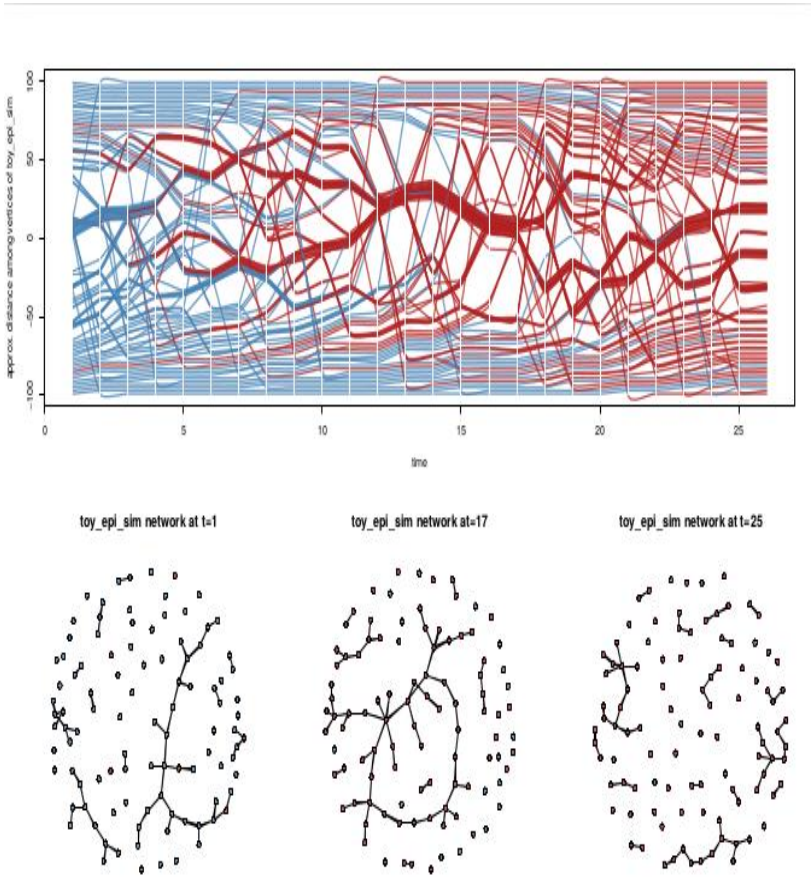




Using R

- Animation package





Overall grade will be evaluated based on

- Midterm exam : 30 %
- Final exam : 50%
- Attendance : 5 %
- Assignment : 10 %
- Discussion : 5 %

Late Homework Policy:

- No late homework will be accepted.

Main text

- Authors: R.A. Johnson and D.W. Wichern
- Title: Applied Multivariate Statistical Analysis
- Publisher: Pearson (2007)

Other recommended text (We sometimes use this as a supplementary material.)

- Authors: W. Hardle and L. Simar
- Title: Applied Multivariate Statistical Analysis
- You may download the pdf file from

http://www.leg.ufpr.br/lib/exe/fetch.php/wiki:internas:biblioteca:applied_multivariate_statistics.pdf

1주	강의주제	Some aspects of multivariate data
	강의내용	We discuss basic characteristics of multivariate data.
	시험 및 과제	
2주	강의주제	Introduction to multivariate data analysis
	강의내용	We discuss the basic purposes of multivariate data analysis.
	시험 및 과제	
3주	강의주제	Basic matrix algebra
	강의내용	We discuss basic linear algebra used in multivariate data analysis.
	시험 및 과제	
4주	강의주제	Multivariate Gaussian distribution
	강의내용	We discuss the definition of multivariate Gaussian distribution and its properties.
	시험 및 과제	
5주	강의주제	Inference about a mean vector
	강의내용	We investigate sampling distributions from multivariate Gaussian distribution.
	시험 및 과제	
6주	강의주제	Inference about a mean vector and hypothesis testing
	강의내용	We discuss Hotelling's T^2 and confidence interval.
	시험 및 과제	
7주	강의주제	Principal component analysis (PCA)
	강의내용	We introduce PCA and discuss its meaning. We will examine some examples and do computational practice.
	시험 및 과제	
8주	강의주제	Midterm exam
	강의내용	Review and evaluation

9주	강의주제	Factor analysis
	강의내용	We introduce factor analysis and its basic theory.
	시험 및 과제	
10주	강의주제	Factor analysis in practice
	강의내용	We examine examples and do computational practice.
	시험 및 과제	
11주	강의주제	Discrimination analysis
	강의내용	We introduce discrimination analysis and its basic theory.
	시험 및 과제	
12주	강의주제	Discrimination analysis in practice
	강의내용	We examine examples and do computational practice.
	시험 및 과제	
13주	강의주제	Clustering analysis
	강의내용	We introduce clustering analysis and its basic theory.
	시험 및 과제	
14주	강의주제	Clustering analysis in practice
	강의내용	We examine examples and do computational practice.
	시험 및 과제	
15주	강의주제	Further topics and discussion
	강의내용	Some advanced multivariate analysis techniques are introduced.
	시험 및 과제	
16주	강의주제	Final exam
	강의내용	Review and evaluation
	시험 및 과제	

Course schedule: we will deal with

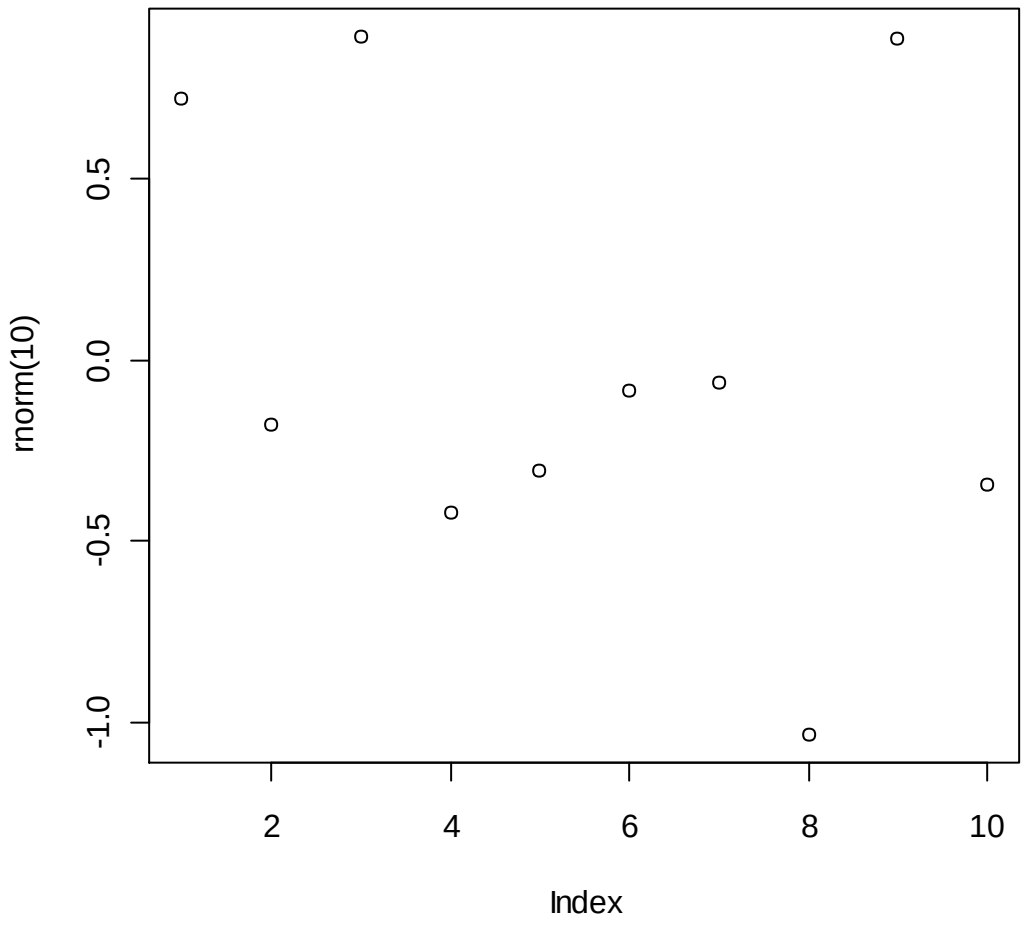
- Some multivariate distributions
- Basic hypothesis testing for mean vectors
- Principal Component Analysis
- Factor Analysis
- Discrimination Analysis
- Clustering Analysis

R examples

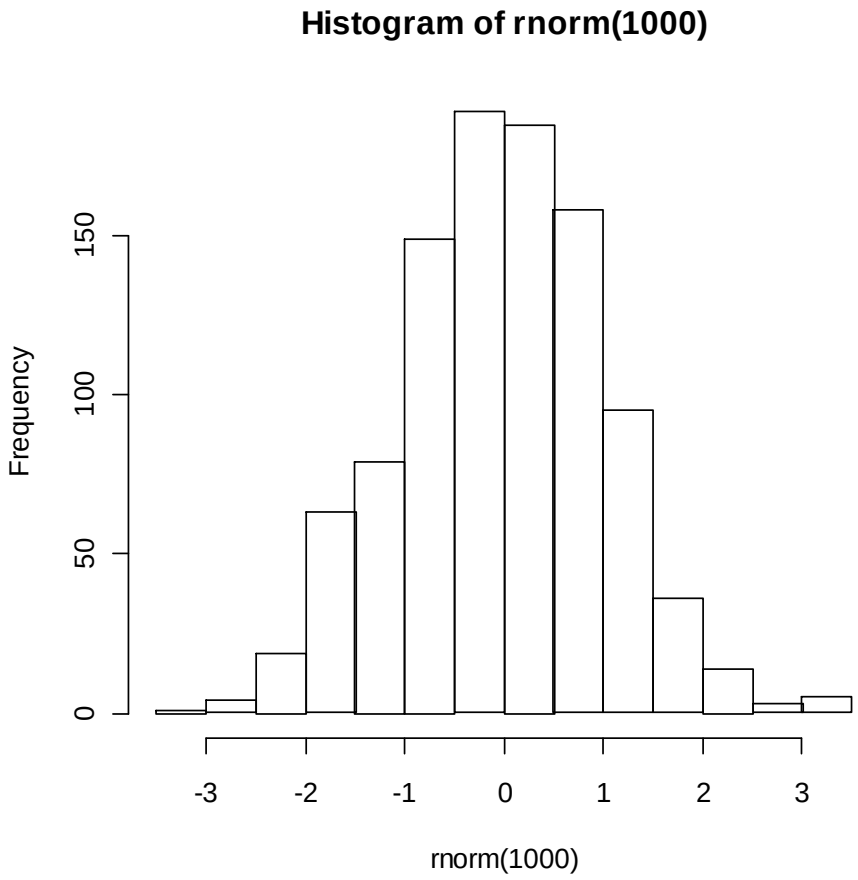
- 3+3
- exp(-1.2)
- log(2)
- log(5,base=10)
- rnorm(10)
- runif(10)

Remark) Functions (exp, log etc.) are indicated by the presence of parentheses.

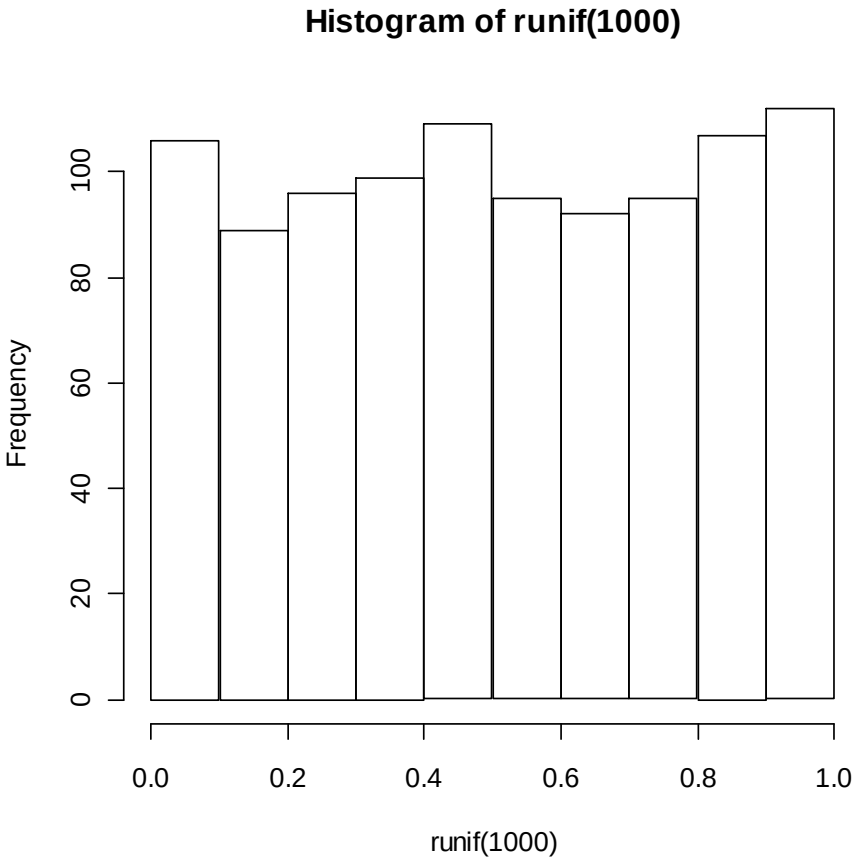
plot(rnorm(10))



hist(rnorm(1000))



hist(runif(1000))



We can assign values to variables by using "$\leftarrow$" or "=".

- x$\leftarrow$2
- y$\leftarrow$2+4
- z$\leftarrow$x+y
- print(x)

Note that variable names are case-sensitive in R.

Basically, R is based on vectorized arithmetic.
Common arithmetic operations (and functions) work elementwise on vectors.

- $x <- c(1, 2, 3, 4, 5)$
- x+5
- 2*x
- x*x
- log(x)

- $x <- c(1, 2, 3, 4, 5)$
- $y <- c(5, 4, 3, 2, 1)$
- x+y
- x*y
- crossprod(x,y)

There are many useful functions in R.

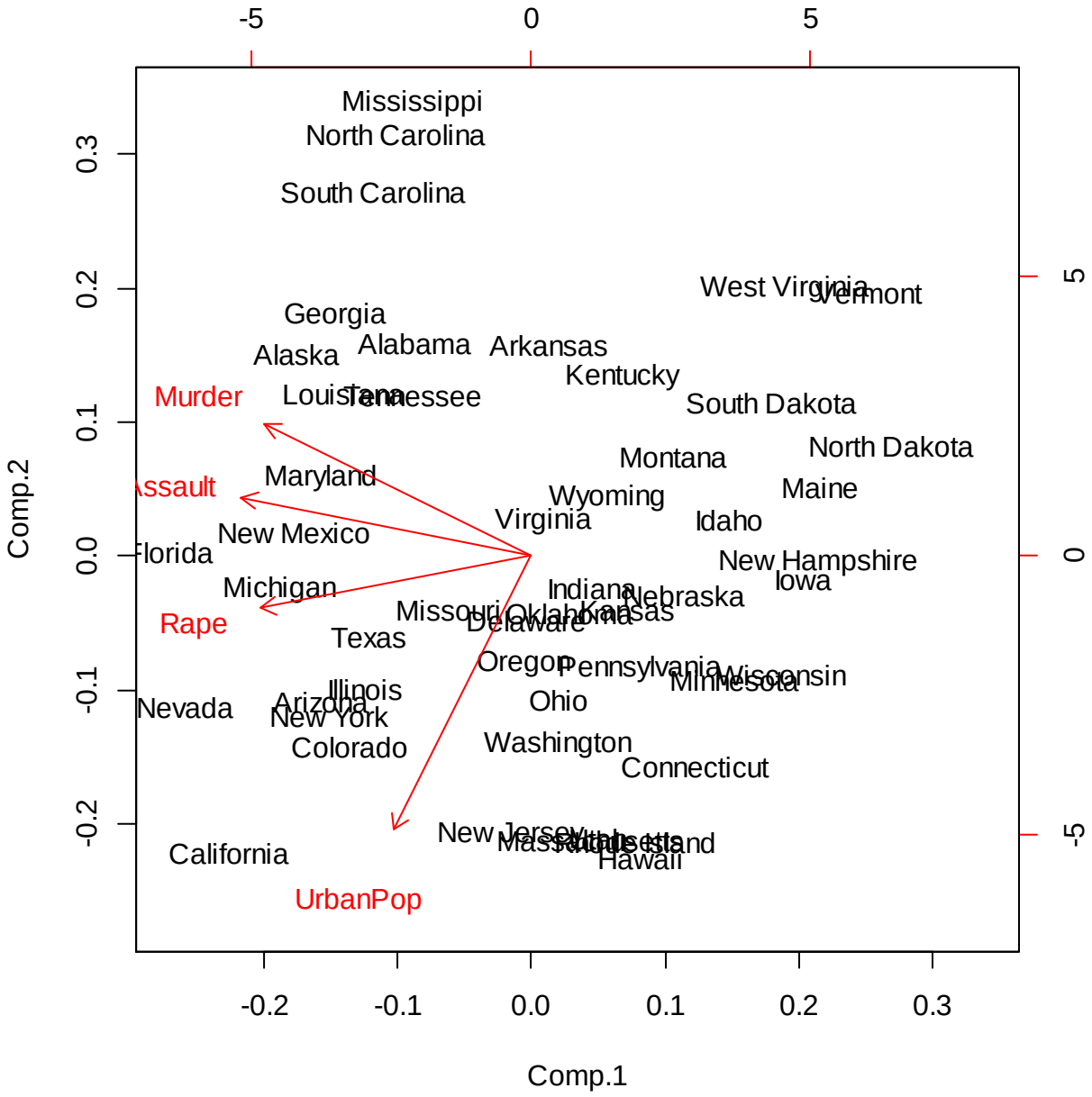
- $x <- \mathrm{c}(1, 2, 3, 4, 5)$
- sum(x)
- mean(x)
- var(x)
- sd(x)
- median(x)
- summary(x)

Especially, there are many graphical tools.

R has too many functions to remember them all.
When the name of topic is "A",

- ?A
- help(A) or help("A")

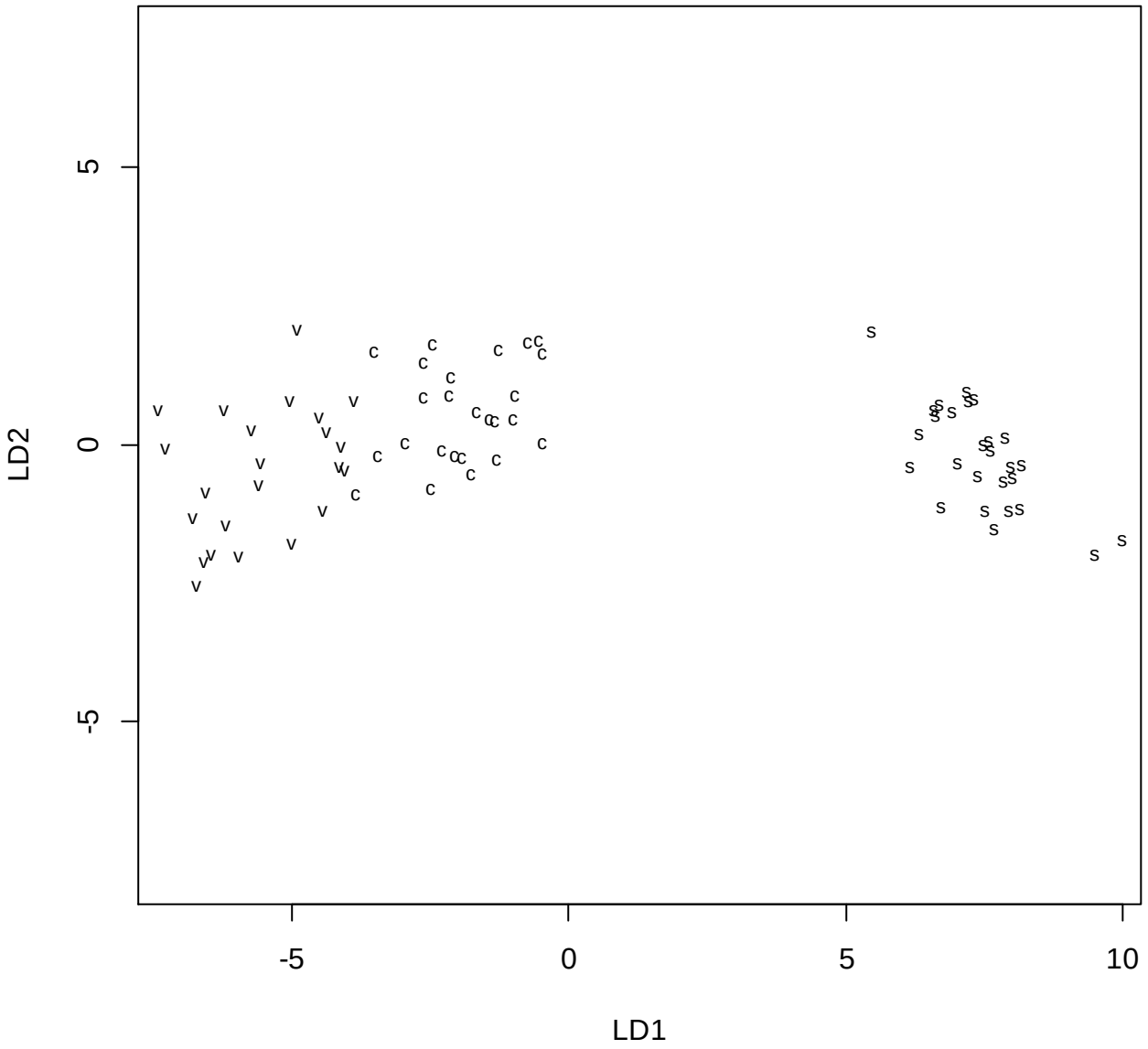
?princomp


-5
0
5
Mississippi
North Carolina
South Carolina
West Virginia
Vermont
Georgia
Alabama
Arkansas
Alaska
Kentucky
Murder
Louisiana
Tennessee
South Dakota
North Dakota
Montana
Maryland
Assault
Wyoming
Maine
Virginia
Idaho
New Mexico
Florida
New Hampshire
Iowa
Michigan
Indiana
Nebraska
Missouri
Oklahoma
Kansas
Rape
Delaware
Texas
Oregon
Pennsylvania
Wisconsin
Minnesota
Illinois
Nevada
Arizona
Ohio
New York
Colorado
Washington
Connecticut
New Jersey
California
Hawaii
UrbanPop
0.3
0.2
0.1
0.0
-0.1
-0.2
Comp.2
5
0
-5
-0.2
-0.1
0.0
0.1
0.2
0.3
Comp.1

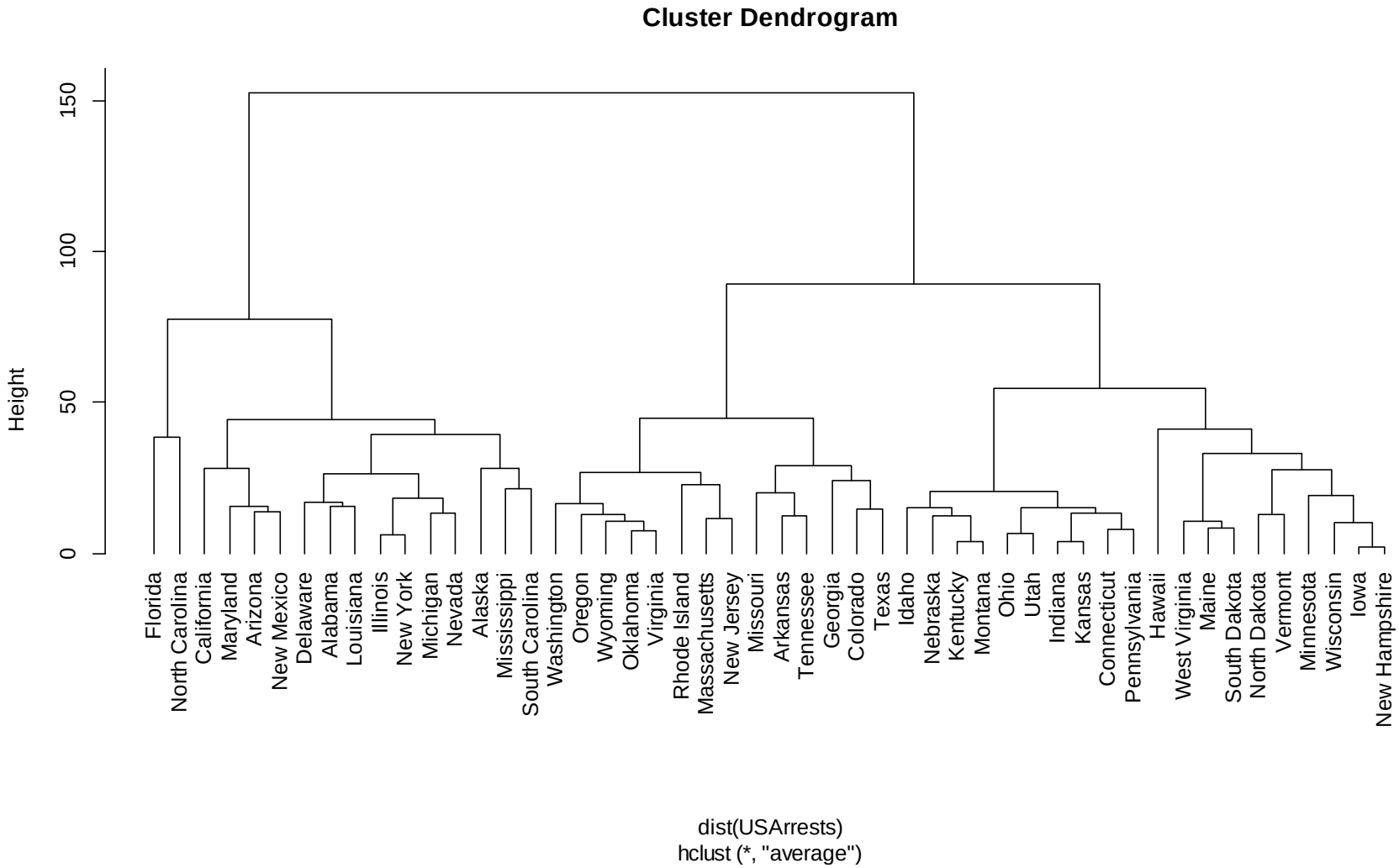
?factanal

```
> factanal(m1, factors = 3) # varimax is the default

Call:
factanal(x = m1, factors = 3)

Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005

Loadings:
   Factor1 Factor2 Factor3
v1 0.944   0.182   0.267
v2 0.905   0.235   0.159
v3 0.236   0.210   0.946
v4 0.180   0.242   0.828
v5 0.242   0.881   0.286
v6 0.193   0.959   0.196

                Factor1 Factor2 Factor3
SS loadings       1.893   1.886   1.797
Proportion Var    0.316   0.314   0.300
Cumulative Var    0.316   0.630   0.929

The degrees of freedom for the model is 0 and the fit was 0.4755
```

```
library(MASS)
?lda
```



?hclust


Cluster Dendrogram
Height
0
50
100
150
Florida
North Carolina
California
Maryland
Arizona
New Mexico
Delaware
Alabama
Louisiana
Illinois
New York
Michigan
Nevada
Alaska
Mississippi
South Carolina
Washington
Oregon
Wyoming
Oklahoma
Virginia
Rhode Island
Massachusetts
New Jersey
Missouri
Arkansas
Tennessee
Georgia
Colorado
Texas
Idaho
Nebraska
Kentucky
Montana
Ohio
Utah
Indiana
Kansas
Connecticut
Pennsylvania
Hawaii
West Virginia
Maine
South Dakota
North Dakota
Vermont
Minnesota
Wisconsin
Iowa
New Hampshire
dist(USArrests)
hclust (*, "average")