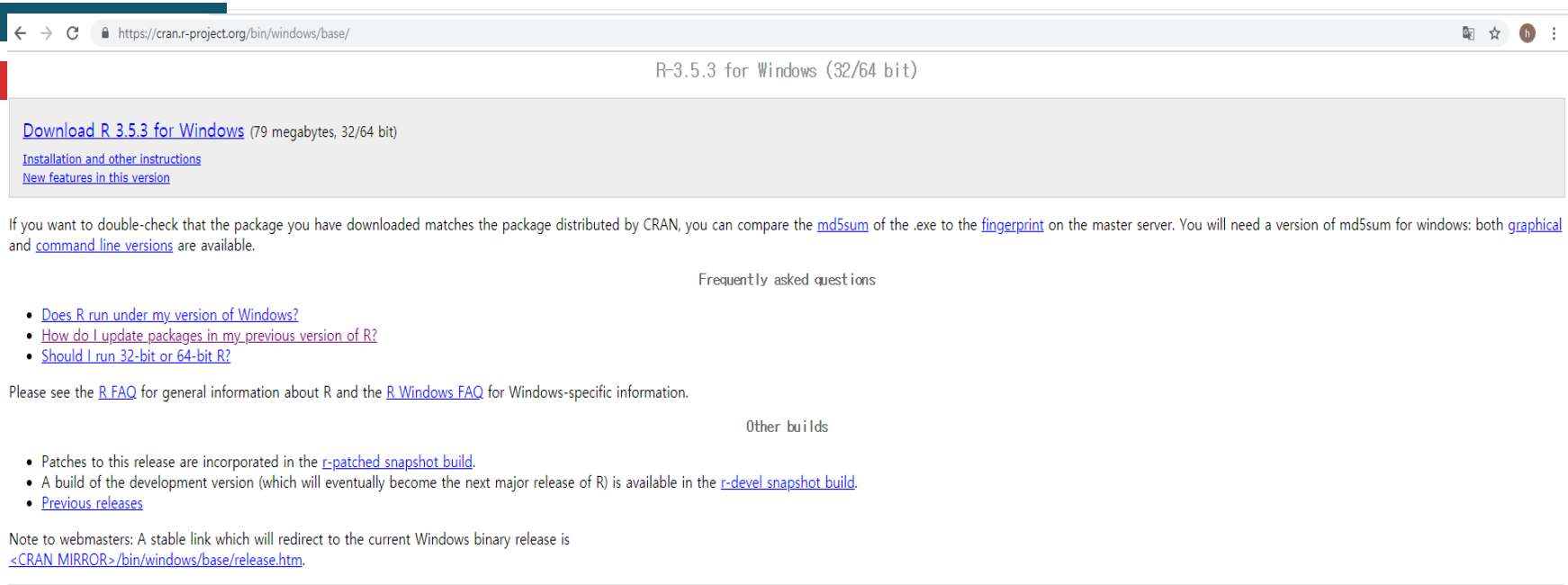
서강대학교 정보통신대학원
통계기반 데이터 분석 3주

2019년 3월 정화민 교수

R 설치



R 프로그래밍 특성 (무료 – 오픈 소스)

- In-Memory Computing
  빠른 처리 속도(H/W 메모리 크기에 영향 받음)
- Object-oriented programming
데이터, 함수, 차트 등 모든 것이 object로 관리
- Package
최신의 알고리즘 및 방법론이 Package로 제공됨
- Visualization
분석에 통찰을 부여할 수 있는 그래픽에 대한 지원

R Studio 설치

https://www.rstudio.com/products/rstudio/download/

RStudio

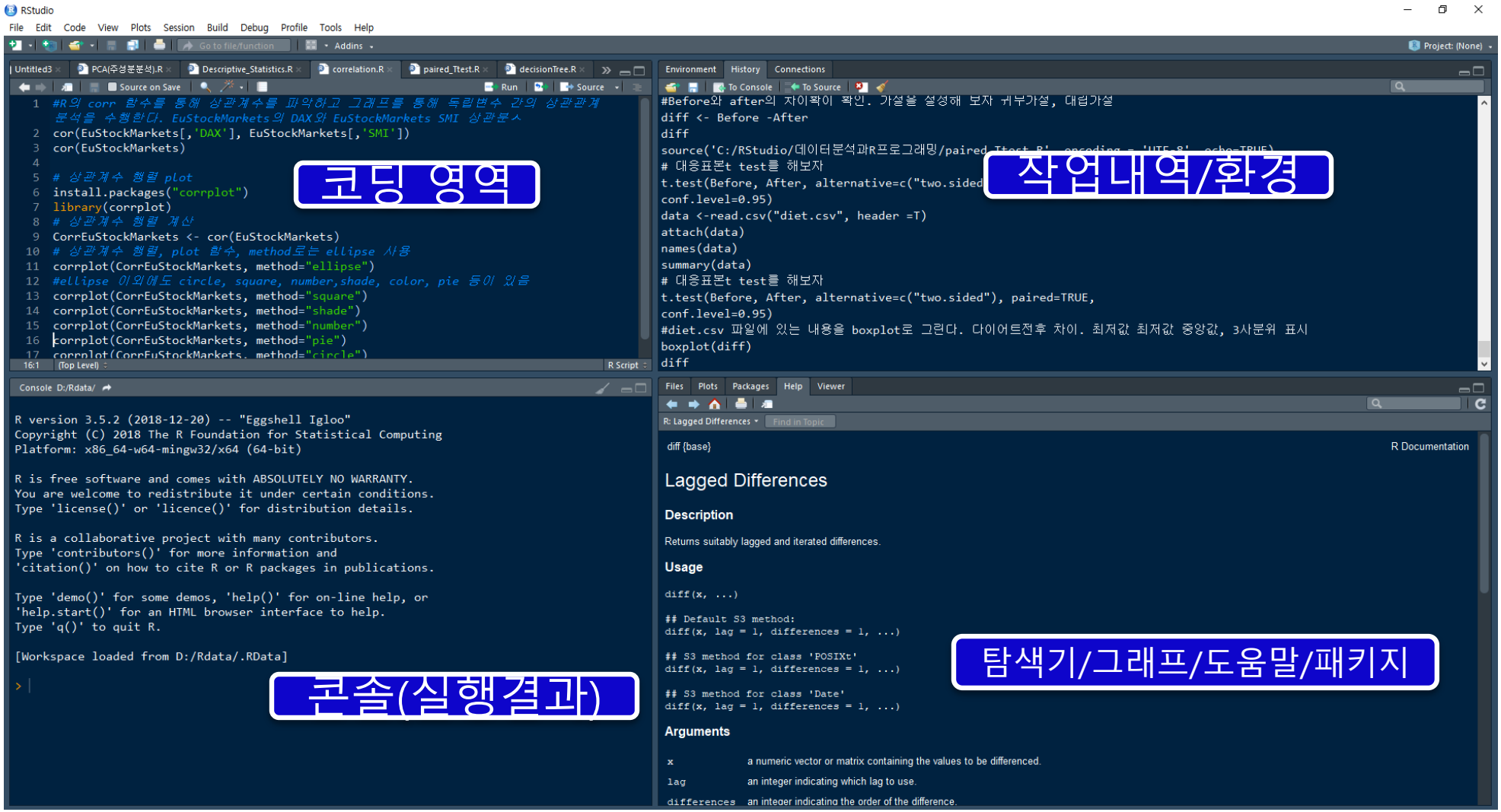
Products Resources Pricing About Us Blogs

Choose Your Version of RStudio

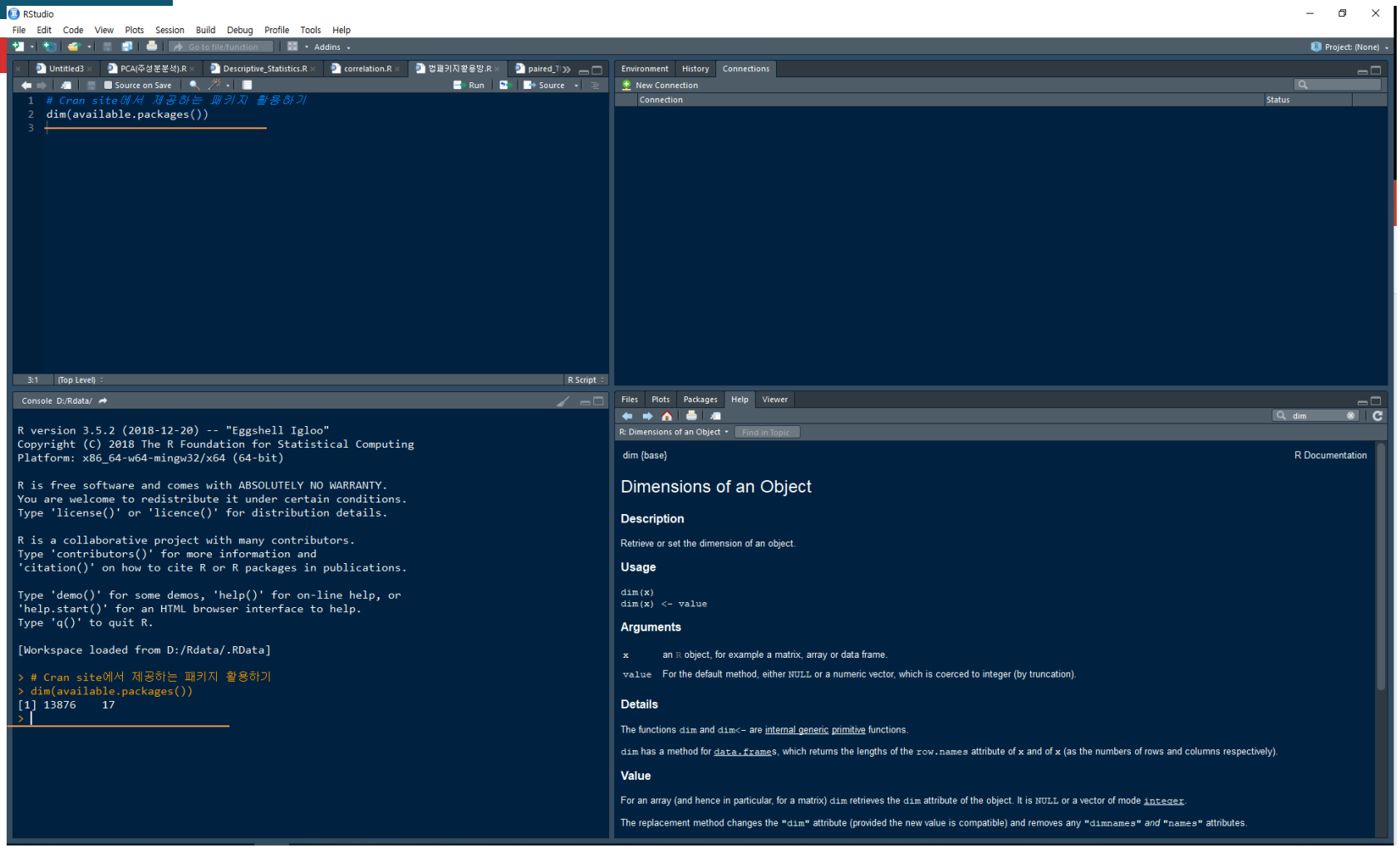
RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. Learn More about RStudio features.

	RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server Pro + RStudio Connect Commercial License
	FREE	$995 per year	FREE	$9,995 per year	$29,995 per year
	DOWNLOAD	BUY	DOWNLOAD	DOWNLOAD	TALK
	Learn More	Learn More	Learn More	Learn More	Learn More
Integrated Tools for R	●	●	●	●	●
Priority Support		●		●	●
Access via Web Browser			●	●	●
Enterprise Security				●	●
Project Sharing				●	●
Manage Multiple R Sessions & Versions				●	●
Admin Dashboard				●	●
Load Balancing				●	●

R Studio 화면

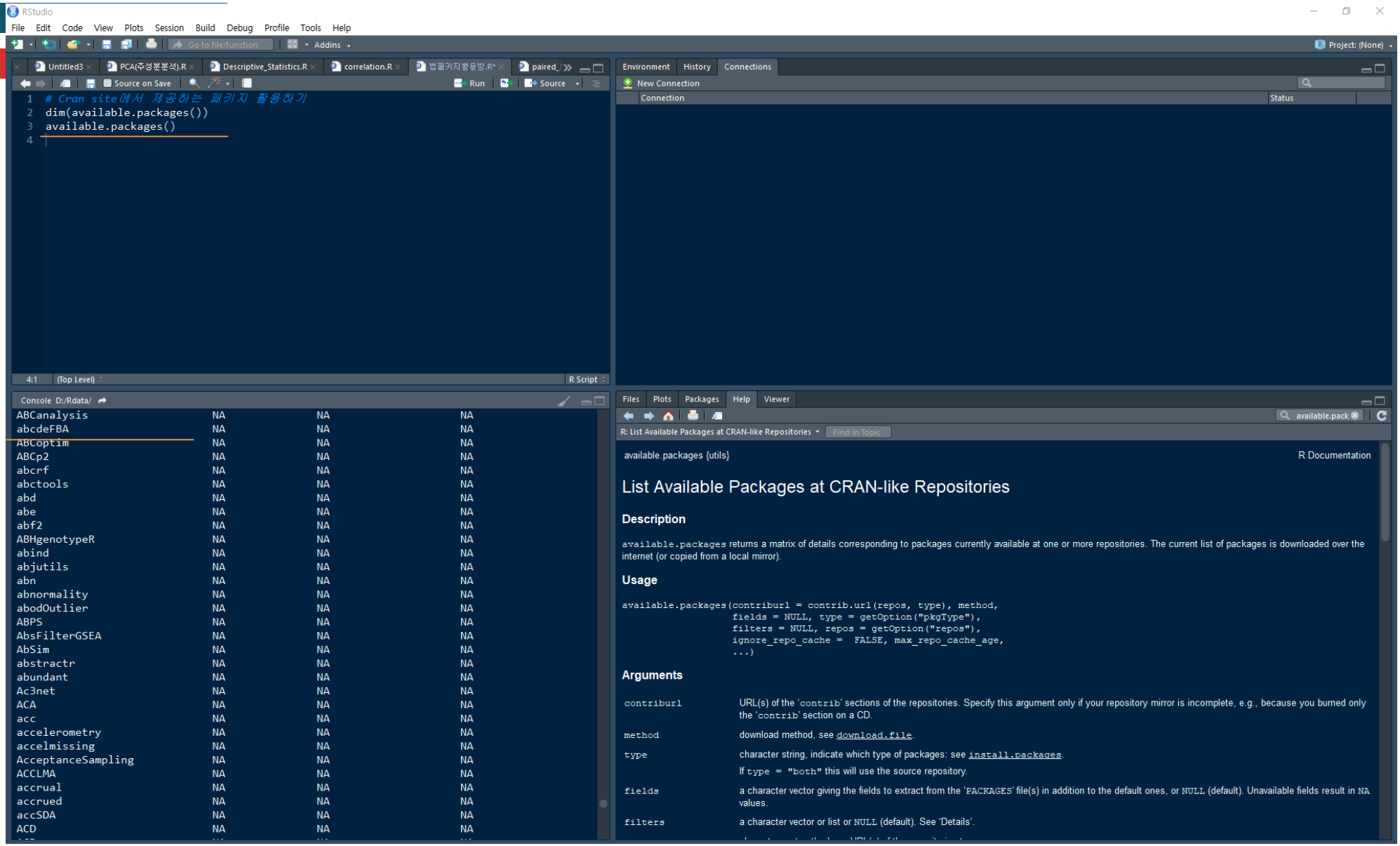

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
코딩 영역
작업내역/환경
콘솔(실행결과)
탐색기/그래프/도움말/패키지
#R의 corr 함수를 통해 상관계수를 파악하고 그래프를 통해 독립변수 간의 상관관계 분석을 수행한다. EuStockMarkets의 DAX와 EuStockMarkets SMI 상관분석
cor(EuStockMarkets[,'DAX'], EuStockMarkets[,'SMI'])
cor(EuStockMarkets)
# 상관계수 행렬 plot
install.packages("corrplot")
library(corrplot)
# 상관계수 행렬 계산
CorrEuStockMarkets <- cor(EuStockMarkets)
# 상관계수 행렬, plot 함수, method로는 ellipse 사용
corrplot(CorrEuStockMarkets, method="ellipse")
#ellipse 이외에도 circle, square, number,shade, color, pie 등이 있음
corrplot(CorrEuStockMarkets, method="square")
corrplot(CorrEuStockMarkets, method="shade")
corrplot(CorrEuStockMarkets, method="number")
corrplot(CorrEuStockMarkets, method="pie")
corrplot(CorrEuStockMarkets, method="circle")
#Before와 after의 차이확이 확인. 가설을 설정해 보자 귀무가설, 대립가설
diff <- Before -After
diff
source('C:/RStudio/데이터분석과R프로그래밍/paired_Ttest.R', encoding = 'UTF-8', echo=TRUE)
# 대응표본t test를 해보자
t.test(Before, After, alternative=c("two.sided
conf.level=0.95)
data <-read.csv("diet.csv", header =T)
attach(data)
names(data)
summary(data)
# 대응표본t test를 해보자
t.test(Before, After, alternative=c("two.sided"), paired=TRUE,
conf.level=0.95)
#diet.csv 파일에 있는 내용을 boxplot로 그린다. 다이어트전후 차이. 최저값 최저값 중앙값, 3사분위 표시
boxplot(diff)
diff
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from D:/Rdata/.RData]
Lagged Differences
Description
Returns suitably lagged and iterated differences.
Usage
diff(x, ...)
Default S3 method:
diff(x, lag = 1, differences = 1, ...)
S3 method for class 'POSIXt'
diff(x, lag = 1, differences = 1, ...)
S3 method for class 'Date'
diff(x, lag = 1, differences = 1, ...)
Arguments
x a numeric vector or matrix containing the values to be differenced.
lag an integer indicating which lag to use.
differences an integer indicating the order of the difference.

Package 사용



Cran site에서 제공하는 패키지 활용하기
dim(available.packages()) 결과값 13876개 Package

available.packages()



Package 사용예

(1) 패키지(Package)
패키지 개념 : 데이터+기능(함수)+알고리즘 (압축),
R 활용 및 응용은 패키지 활용 능력
(2) 패키지 설치
형식) install.packages("패키지명"), install.packages("ggplot2")
(3) 패키지 설치 확인

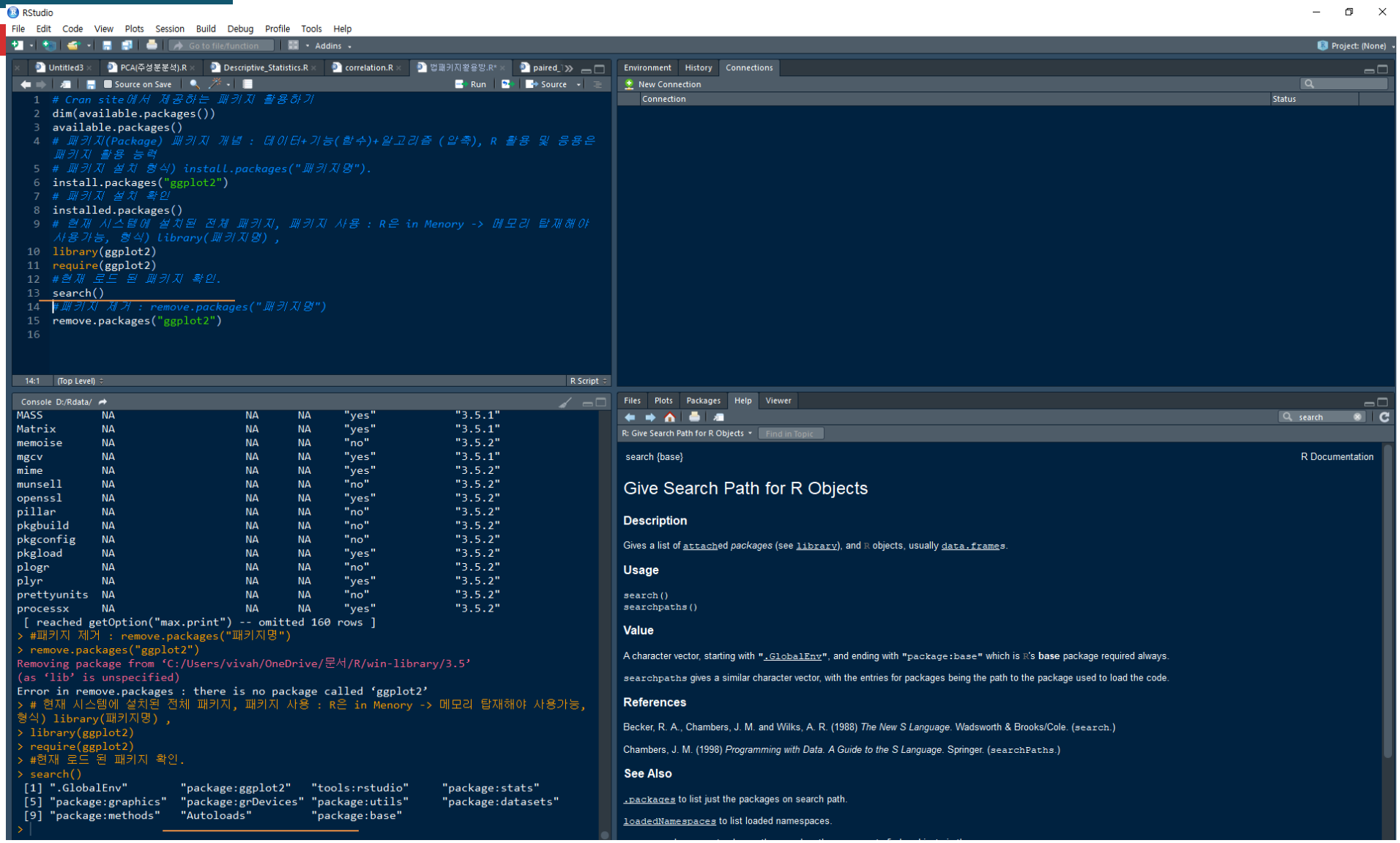
installed.packages() # 현재 시스템에 설치된 전체 패키지
(4) 패키지 사용 : R은 in Menory -> 메모리 탑재해야 사용가능
형식) library(패키지명) , library(ggplot2), require(ggplot2)

(5) 현재 로드 된 패키지 확인. search()

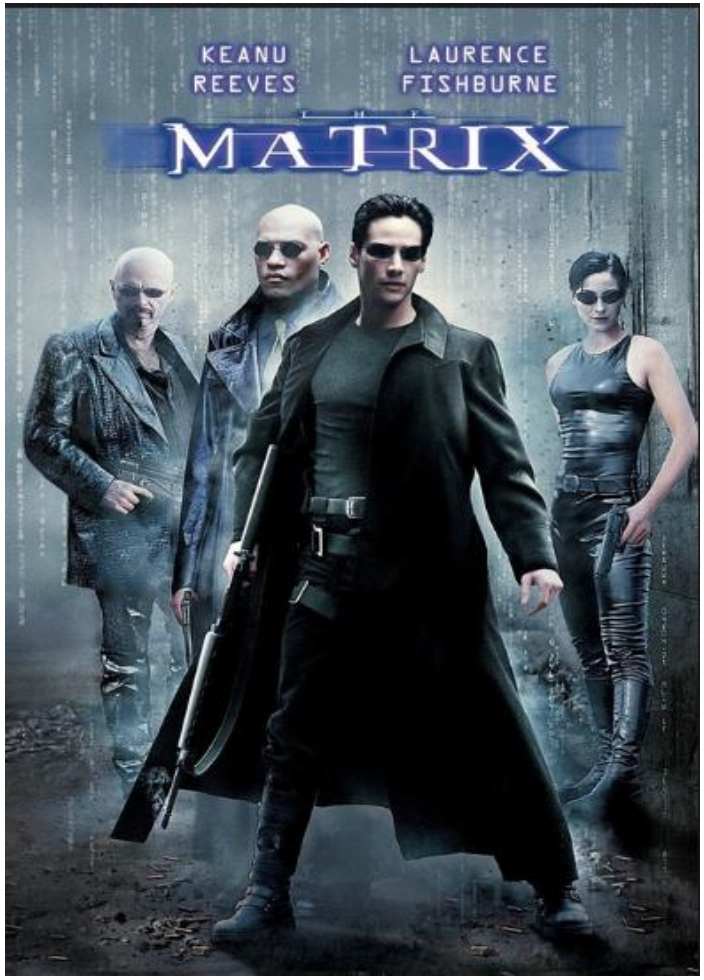
(6) 패키지 제거 : remove.packages("패키지명")

remove.packages("ggplot2")

Package 실행


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled3
PCA(주성분분석).R
Descriptive_Statistics.R
correlation.R
법패키지꿀용방.R*
paired_
Source on Save
Run
Source
1 # Cran site에서 제공하는 패키지 활용하기
2 dim(available.packages())
3 available.packages()
4 # 패키지(Package) 패키지 개념 : 데이터+기능(함수)+알고리즘 (압축), R 활용 및 응용은 패키지 활용 능력
5 # 패키지 설치 형식) install.packages("패키지명").
6 install.packages("ggplot2")
7 # 패키지 설치 확인
8 installed.packages()
9 # 현재 시스템에 설치된 전체 패키지, 패키지 사용 : R은 in Menory -> 메모리 탑재해야 사용가능, 형식) library(패키지명) ,
10 library(ggplot2)
11 require(ggplot2)
12 #현재 로드 된 패키지 확인.
13 search()
14 #패키지 제거 : remove.packages("패키지명")
15 remove.packages("ggplot2")
16
14:1 (Top Level) R Script
Environment History Connections
New Connection
Connection Status
Console D:/Rdata/
MASS NA NA NA "yes" "3.5.1"
Matrix NA NA NA "yes" "3.5.1"
memoise NA NA NA "no" "3.5.2"
mgcv NA NA NA "yes" "3.5.1"
mime NA NA NA "yes" "3.5.2"
munsell NA NA NA "no" "3.5.2"
openssl NA NA NA "yes" "3.5.2"
pillar NA NA NA "no" "3.5.2"
pkgbuild NA NA NA "no" "3.5.2"
pkgconfig NA NA NA "no" "3.5.2"
pkgload NA NA NA "yes" "3.5.2"
plogr NA NA NA "no" "3.5.2"
plyr NA NA NA "yes" "3.5.2"
prettyunits NA NA NA "no" "3.5.2"
processx NA NA NA "yes" "3.5.2"
[reached getOption("max.print") -- omitted 160 rows]
> #패키지 제거 : remove.packages("패키지명")
> remove.packages("ggplot2")
Removing package from 'C:/Users/vivah/OneDrive/문서/R/win-library/3.5'
(as 'lib' is unspecified)
Error in remove.packages : there is no package called 'ggplot2'
> # 현재 시스템에 설치된 전체 패키지, 패키지 사용 : R은 in Menory -> 메모리 탑재해야 사용가능, 형식) library(패키지명) ,
> library(ggplot2)
> require(ggplot2)
> #현재 로드 된 패키지 확인.
> search()
[1] ".GlobalEnv" "package:ggplot2" "tools:rstudio" "package:stats"
[5] "package:graphics" "package:grDevices" "package:utils" "package:datasets"
[9] "package:methods" "Autoloads" "package:base"
Files Plots Packages Help Viewer
R: Give Search Path for R Objects
search {base} R Documentation
Give Search Path for R Objects
Description
Gives a list of attached packages (see library), and R objects, usually data.frames.
Usage
search()
searchpaths()
Value
A character vector, starting with ".GlobalEnv", and ending with "package:base" which is R's base package required always.
searchpaths gives a similar character vector, with the entries for packages being the path to the package used to load the code.
References
Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole. (search.)
Chambers, J. M. (1998) Programming with Data. A Guide to the S Language. Springer. (searchPaths.)
See Also
.packages to list just the packages on search path.
loadedNamespaces to list loaded namespaces.

Package : 영화 매트릭스



https://www.youtube.com/watch?v=2svLxh03bYw

변수(Variable)와 데이터 유형(Data type)

1) 변수(Variable)

- 값을 저장하는 메모리 이름
- R은 모두 변수가 객체(데이터+함수+차트 등)

2) 변수명 작성 규칙

- 첫번째 단어 : 영문자 시작
- 두번째 단어 : 숫자, 언더바(_), 점(name.id) 사용 가능
- 대소문자 구분
- 의미를 파악할 수 있는 이름으로 지정

3) 데이터 유형(Data Type)

유형(Type)	값(value)	예
숫자형(Numeric)	정수, 실수	1, 2, 3, 22, 23
문자형(Character)	문자, 문자열	"대한민국"
논리형(Logical)	참, 거짓	TRUE 또는 T FALSE 또는 F
기타	결측치, 정의되지 않음	NA, Null

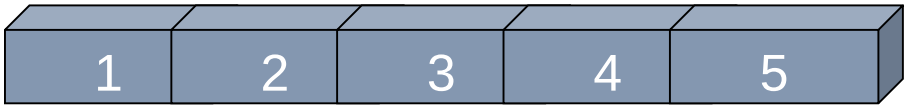
자료구조

R에서 제공하는 주요 자료구조(객체 타입)

1. Vector : 동일 데이터 타입을 갖는 1차원 배열
2. Matrix : 동일 데이터 타입을 갖는 2차원 배열
3. Array : 동일 데이터 타입을 갖는 3차원 배열
4. Data Frame : 열 단위로 서로 다른 데이터 타입을 갖는 배열
5. List : 서로 다른 데이터 구조(Vector, Data Frame, Array, List) 중첩

1. Vector

1) Vector 자료구조



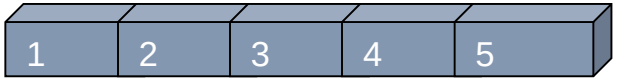
Vector

- ✓ R의 기본 데이터 구조
- ✓ 1차원 배열 형태
  - 접근 : [index] : 1부터 시작
- ✓ 동일한 타입의 데이터만 저장 가능
- ✓ X <-c(1, 10, 13) 숫자형 벡터
- ✓ Y <-c("대한민국", "자동차") 문자형 벡터
- ✓ Z <-c(TRUE, FALSE) 논리형 벡터
- ✓여기서 C는 concentration (연결하다)

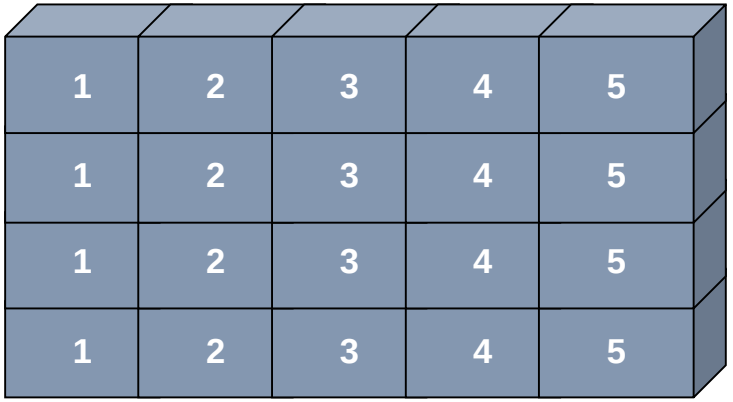
2. Matrix

2) Matrix 자료구조

- ✓ 동일 데이터 타입을 갖는 2차원 배열
- ✓ 행렬(matrix) 객체



Vector



Matrix

```
> #행렬
> mx=matrix(c(1,2,3,4,5,6,7,8), ncol=2)
> mx
     [,1] [,2]
[1,]    1    5
[2,]    2    6
[3,]    3    7
[4,]    4    8
```

3. Array

3) Array : 동일 데이터 타입을 갖는 다차원 배열

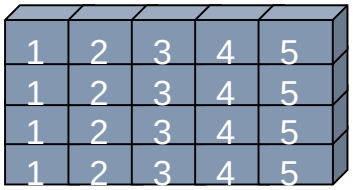
- ✓ 3차원 배열 객체 생성
- ✓ R에서 활용도 낮음

d <- c(1:12) # 12개 벡터 객체 생성

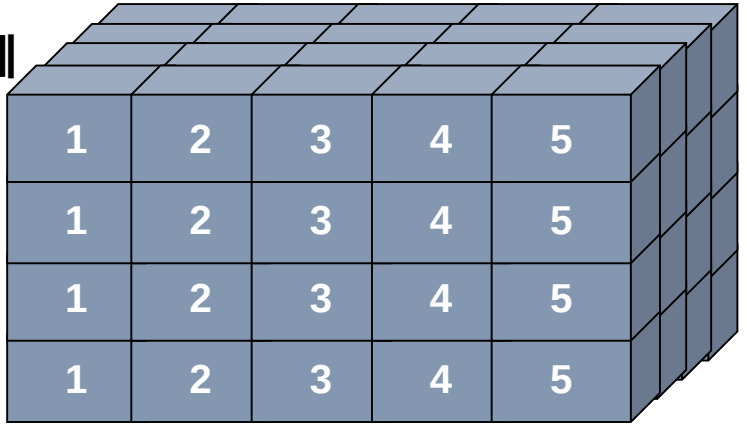
arr <- array(d, c(3,2,2)) # 3행2열 구조 2개

arr #1~6(1면), 7~12(2면) -> 3차원 배열 객체

Vector



Matrix

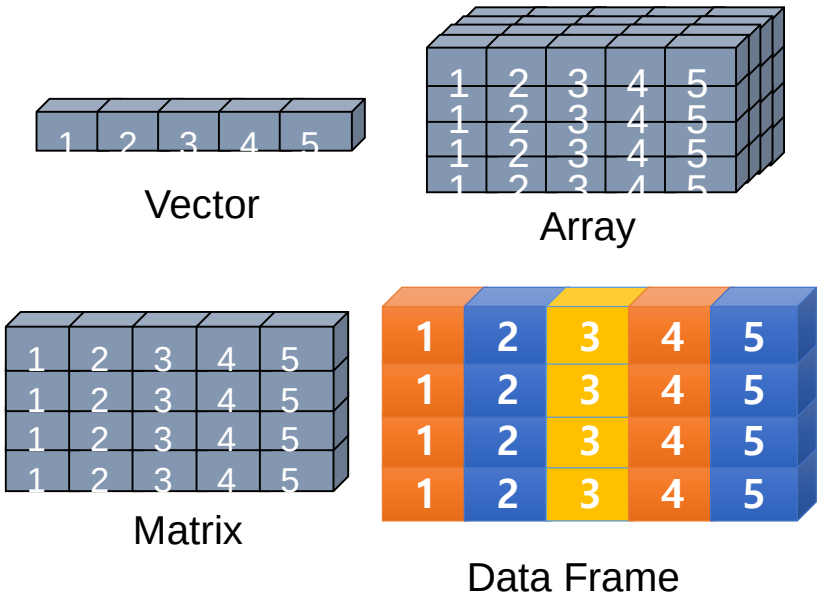


Array

4. Data Frame

5) Data Frame

- ✓ **리스트 보다 활용범위 넓다.**
- ✓ **DBMS의 테이블 구조와 유사**
  (서로 다른 데이터 타입을 갖는 컬럼)
- ✓ **가장 많이 사용하는 객체 타입**
- ✓ **list와 Vector 혼합형**
  - 컬럼 구성 : list, list 구성 : vector
- ✓ **data frame 생성방법**
  - Vector, Matrix, txt/excel/csv 파일



5. Data Frame

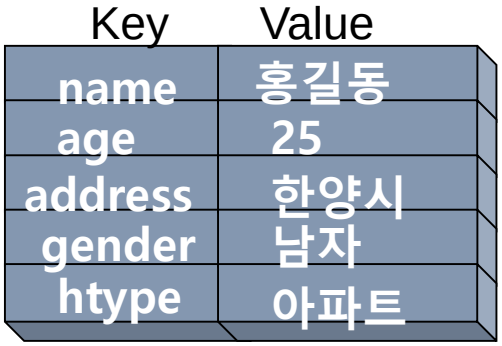
- **Data Frame 특징**
  1. 형식) data.frame(컬럼1, 컬럼2.. 컬럼n)
  2. 컬럼 단위로 서로 다른 자료형 가능
  3. 모든 컬럼은 크기가 동일해야 함

```
> #데이터 프레임
> income = c(10,100,200,300)
> car=c("Samsung","KIA","Hyundai","BMW")
> marriage=c(FALSE, FALSE, TRUE, FALSE)
> mydat=data.frame(income,car,marriage)
> mydat
  income     car marriage
1     10 Samsung    FALSE
2    100     KIA    FALSE
3    200 Hyundai     TRUE
4    300     BMW    FALSE
```

5. List

4) List : 서로 다른 데이터 구조

- ✓ Vector, Data Frame, Array, List의 중첩 구조
- ✓ c(구조체), python(딕션너리)
- ✓ 함수 내에서 여러 값을 하나의 키로 묶어서 반환할 경우 유용함




```
member <- list(name="홍길동",
               age = 25,address="한양",
               gender="남자", htype="아파트")
```