Recipes 데이터 분석

강사 : 문성민

What is R?

Data Analysis Tool : R

R 설치

http://www.r-project.org/

1) CRAN

About R
What is R?
Contributors
Screenshots
What's new?

Download, Packages
CRAN

2) 국가선택

Korea
http://cran.nexr.com/
http://healthstat.snu.ac.kr/CRAN/
http://cran.biodisk.org/

3) 운영체제 선택

Download and Install R

Precompiled binary distributions of these versions of R:

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

4) 모드 선택

Subdirectories:

base

contrib

Rtools

Please do not submit to Windows binaries.

5) 다운로드

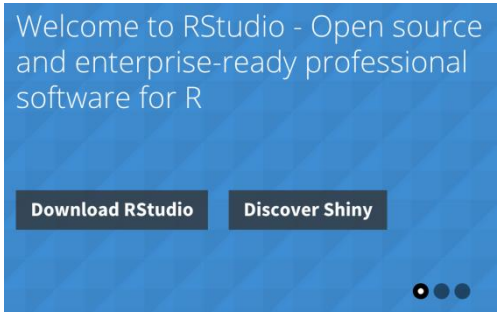
Download R 3.1.2 for Windows (54 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

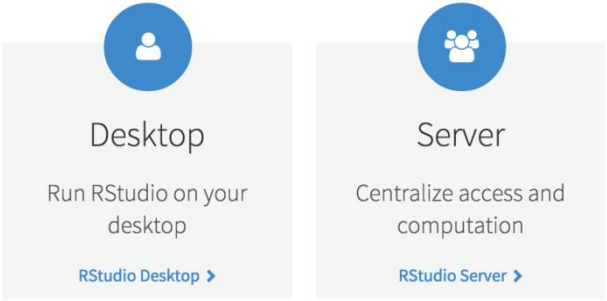
R studio 설치

http://www.rstudio.com/

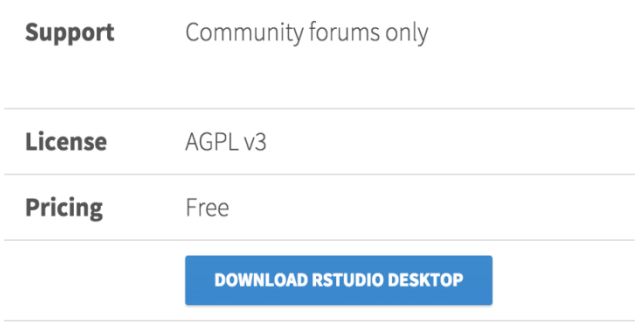
1) Download Main



2) 설치 위치 선택



3) 상품 선택



4) 다운로드

Installers for ALL Platforms

Installers	Size	Date	MD5
RStudio 0.98.1102 - Windows XP/Vista/7/8	47.4 MB	2015-02-07	553b53f8b467ba31f21c672686662152
RStudio 0.98.1102 - Mac OS X 10.6+ (64-bit)	43.7 MB	2015-02-07	045e903ad09e9c8dbf65cf08ff16023d
RStudio 0.98.1102 - Debian 6+/Ubuntu 10.04+ (32-bit)	49.5 MB	2015-02-07	90ba83bf5a791ca3bcc12e1faf37d5ae
RStudio 0.98.1102 - Debian 6+/Ubuntu 10.04+ (64-bit)	51.4 MB	2015-02-07	f4d479f62352c5a709d330f67ef310dc
RStudio 0.98.1102 - Fedora 13+/RedHat 7+/openSUSE 11.4+ (32-bit)	49.9 MB	2015-02-07	91b64c1bbedfde387b523aa0cc0036df
RStudio 0.98.1102 - Fedora 13+/RedHat 7+/openSUSE 11.4+ (64-bit)	51.5 MB	2015-02-07	dac3eb2127d82fa0ef35e8c4773c1f6a

R 이란?

- 개발(Development)
  - 뉴질랜드 오클랜드 대학 로스 이하카, 로버트 젠틀맨이 최초 개발
  - R-Core Team 1997
- 환경(environment)
  - 대화식 프로그램 수행
  - 대용량 데이터 관리 및 처리
  - 행렬연산
  - 그래픽환경
- 확장성 및 범용성
  - Linux, Mac, Windows 운영체제에서 사용 가능
  - Java, C, Fortran 프로그래밍 언어에 인터페이스 제공
  - DBMS 데이터 접근 용이
  - Embedded R in Excel
- Free software and Open source
  - GPL(General Public License) 개념으로 CRAN(Comprehensive R Archive Network)에서 배포

Descriptive Statistic

Descriptive Statistic

- 기술통계학(Descriptive statistics)

  - 관심의 대상이 되는 자료에 대해 그림 및 수치를 사용하여 정리하고 요약하는 방법
  - 추론통계를 위한 사전단계로 수집된 자료의 분석에 초점을 둔다.
  - 수치를 활용한 방법 : 비율, 지수, 평균, 분산 등
  - 그림을 활용한 방법 : 막대그래프, 히스토그램, 상자그림 등

- 수식(Formuler)

$$\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n} \qquad S=\sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}} \qquad S^2=\frac{\sum(X_i-\bar{X})^2}{n-1}$$

$\bar{X}$=평균, S(σ)=표준편차, $S^2(\sigma^2)$=분산(자료의 흩어진 정도에 대한 척도)

Descriptive Statistic

- 기술통계학(Descriptive statistics)을 사용하여 데이터 분석하기

- 데이터 설명하기

Student Admissions at UC Berkeley

Description

Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

Usage

```
UCBAdmissions
```

Format

A 3-dimensional array resulting from cross-tabulating 4526 observations on 3 variables. The variables and their levels are as follows:

No	Name	Levels
1	Admit	Admitted, Rejected
2	Gender	Male, Female
3	Dept	A, B, C, D, E, F

- 데이터 확인하기(head,str)

```
> head(DF)
     Admit Gender Dept Freq
1 Admitted   Male    A  512
2 Rejected   Male    A  313
3 Admitted Female    A   89
4 Rejected Female    A   19
5 Admitted   Male    B  353
6 Rejected   Male    B  207
> str(DF)
'data.frame':   24 obs. of  4 variables:
 $ Admit : Factor w/ 2 levels "Admitted","Rejected": 1 2 1 2 1 2 1 2 1 2 ...
 $ Gender: Factor w/ 2 levels "Male","Female": 1 1 2 2 1 1 2 2 1 1 ...
 $ Dept  : Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 2 2 2 2 3 3 ...
 $ Freq  : num  512 313 89 19 353 207 17 8 120 205 ...
```

- Data.frame형태로 가져오기

```
> DF=as.data.frame(UCBAdmissior
> DF
      Admit Gender Dept Freq
1  Admitted   Male    A  512
2  Rejected   Male    A  313
3  Admitted Female    A   89
4  Rejected Female    A   19
5  Admitted   Male    B  353
6  Rejected   Male    B  207
7  Admitted Female    B   17
8  Rejected Female    B    8
9  Admitted   Male    C  120
10 Rejected   Male    C  205
11 Admitted Female    C  202
12 Rejected Female    C  391
13 Admitted   Male    D  138
14 Rejected   Male    D  279
15 Admitted Female    D  131
16 Rejected Female    D  244
17 Admitted   Male    E   53
18 Rejected   Male    E  138
19 Admitted Female    E   94
20 Rejected Female    E  299
21 Admitted   Male    F   22
22 Rejected   Male    F  351
23 Admitted Female    F   24
24 Rejected Female    F  317
```

Descriptive Statistic

- 기술통계학(Descriptive statistics)을 사용하여 데이터 분석하기

- 데이터 적용하기

```
> attach(DF)
The following object is masked from Exam_1 (pos = 3):

    Gender

The following object is masked from Exam_1 (pos = 4):

    Gender
```

- 최대값 구하기

```
> max(Freq)
[1] 512
```

- 평균값 구하기

```
> mean(Freq)
[1] 188.5833
```

- 분산 구하기

```
> var(Freq)
[1] 19617.82
```

- 최소값 구하기

```
> min(Freq)
[1] 8
```

- 중앙값 구하기

```
> median(Freq)
[1] 170
```

- 표준편차 구하기

```
> sd(Freq)
[1] 140.0636
```

- 데이터 요약하기

```
> summary(DF)
     Admit       Gender   Dept       Freq
 Admitted:12   Male  :12   A:4   Min.   :  8.0
 Rejected:12   Female:12   B:4   1st Qu.: 80.0
                           C:4   Median :170.0
                           D:4   Mean   :188.6
                           E:4   3rd Qu.:302.5
                           F:4   Max.   :512.0
```

- 데이터 적용 해지하기

```
> detach(DF)
```

Descriptive Statistic

- 기술통계학(Descriptive statistics)을 사용하여 데이터 분석하기

- 변수생성

```
> a=c(1,2,3,1,2,3,4,1,2,3,4,1,2,1,3,4,1,1)
> b=c(5,6,7,5,6,5,6,7,8,5,6,8,5,5,6,7,5,5)
```

- 도수 분포표 만들기

```
> table(a)
a
1 2 3 4
7 4 4 3
> table(b)
b
5 6 7 8
8 5 3 2
```

- 도수 분할표 만들기

```
> table(a,b)
   b
a   5 6 7 8
  1 5 0 1 1
  2 1 2 0 1
  3 2 1 1 0
  4 0 2 1 0
```

- UCBAdmissions데이터를 활용하여 도수 분할표 만들기

```
> xtabs(Freq ~ Gender + Admit, DF)
        Admit
Gender   Admitted Rejected
  Male       1198     1493
  Female      557     1278
```

Descriptive Statistic

- Pressure데이터를 활용한 선 그래프
- 데이터 확인

pressure {datasets} R Documentation

Vapor Pressure of Mercury as a Function of Temperature

Description

Data on the relation between temperature in degrees Celsius and vapor pressure of mercury in millimeters (of mercury).

Usage

```
pressure
```

Format

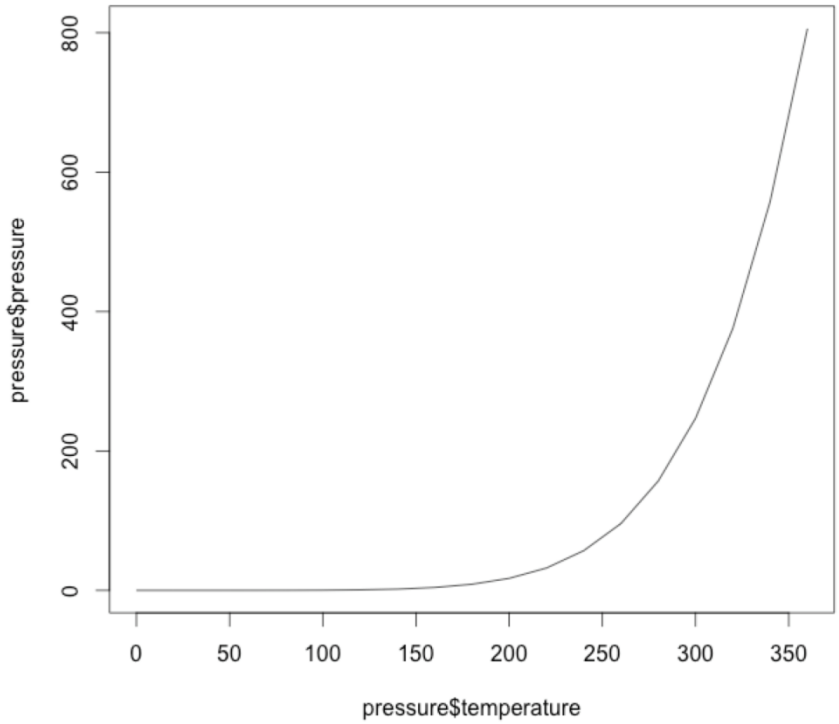
A data frame with 19 observations on 2 variables.

[, 1] temperature numeric temperature (deg C)
[, 2] pressure numeric pressure (mm)

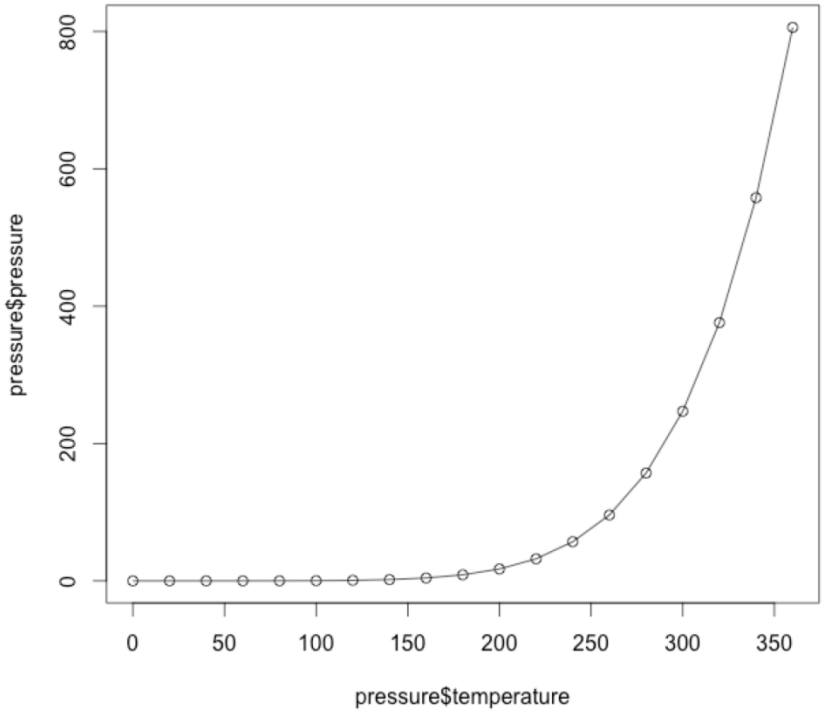
pressure데이터 설명 : 1 미리리터의 수은의 증기압과 섭씨온도 사이의 관계
변수설명
temperature = 섭씨온도
pressure = 1 미리리터의 수은의 증기압력

Descriptive Statistic

- Pressure데이터를 활용한 선 그래프




```
> plot(pressure$temperature,pressure$pressure,type="l")
```




```
> points(pressure$temperature,pressure$pressure)
```

Descriptive Statistic

- mtcars데이터를 활용한 산점도, 히스토그램
  - 데이터 확인

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 variables.

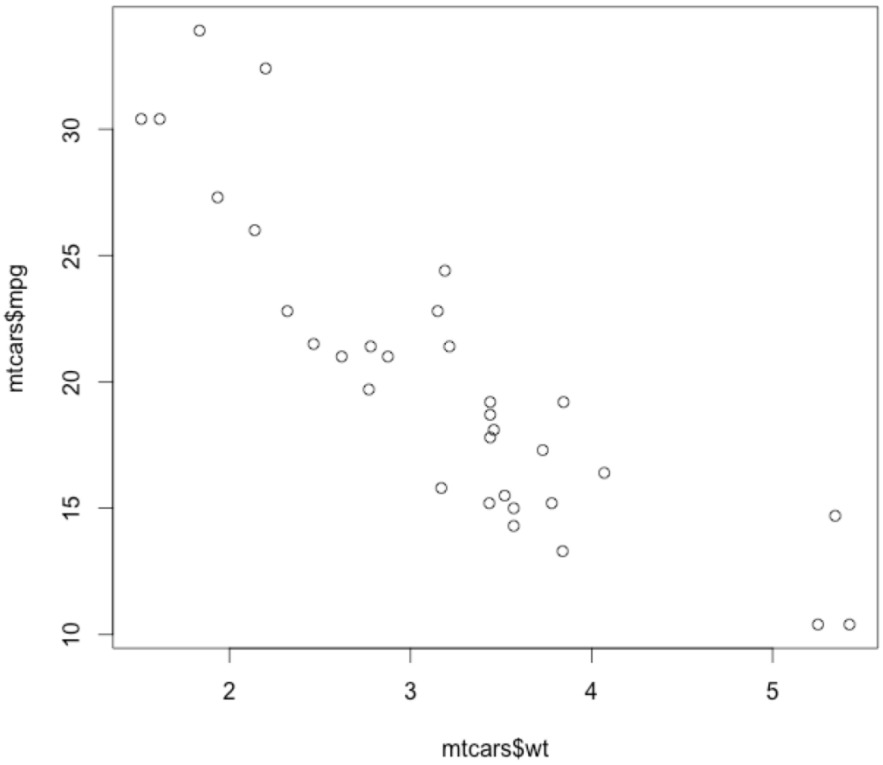
[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (lb/1000)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

mtcars데이터 설명
이 데이터는 1974 년 모터 트렌드 미국 잡지에서 추출하였으며 1973년-1974년도 모델의 32종의 자동차들의 연비등 자동차의 10가지 중요요소를 보여준다.

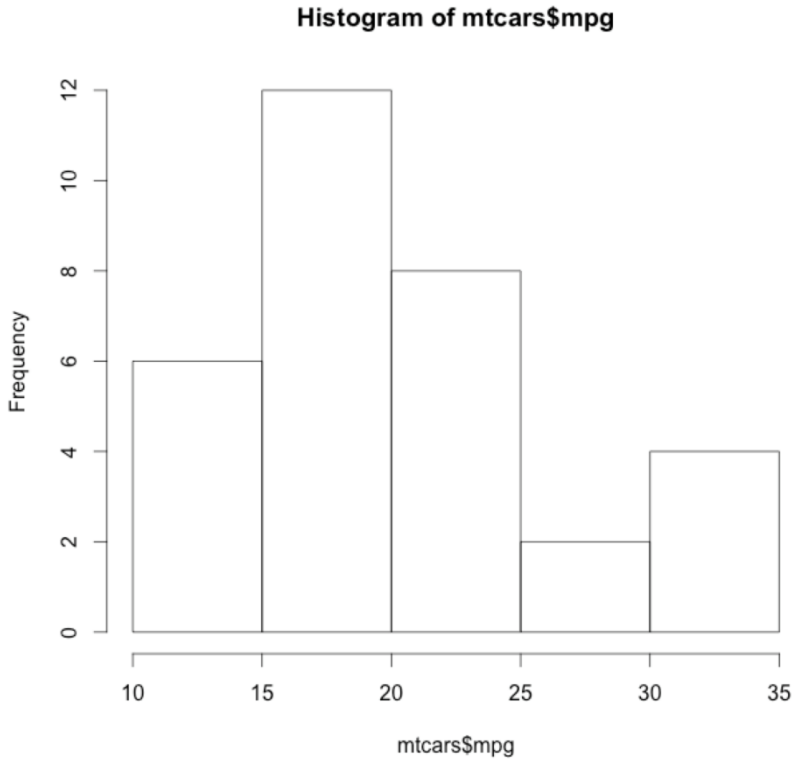
변수설명
mpg = 마일 / (US) 갤런
cyl = 실린더의 수
disp = 변위 (cu.in.)
hp = 총 마력
drat = 리어 액슬 비율
wt = 무게 (파운드 / 1000)
qsec = 1/4 마일 시간
vs = V / S
am = 변속기 (0 = 자동, 1 = 수동)
gear = 기어의 수
carb = 기화기의 수

Descriptive Statistic

- mtcars데이터를 활용한 산점도, 히스토그램




```
>
> plot(mtcars$wt,mtcars$mpg)
>
```




```
>
> hist(mtcars$mpg)
>
```

Descriptive Statistic

- BOD데이터를 활용한 막대 그래프
- 데이터 확인

BOD {datasets} R Documentation

Biochemical Oxygen Demand

Description

The BOD data frame has 6 rows and 2 columns giving the biochemical oxygen demand versus time in an evaluation of water quality.

Usage

```
BOD
```

Format

This data frame contains the following columns:

`Time`

A numeric vector giving the time of the measurement (days).

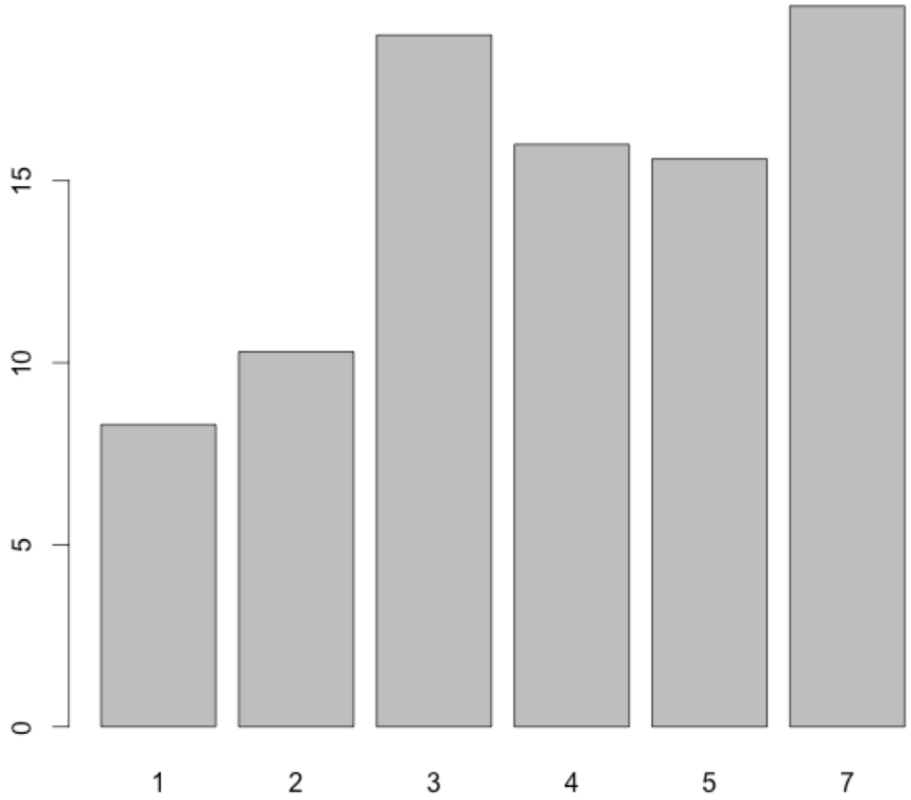
`demand`

A numeric vector giving the biochemical oxygen demand (mg/l).

BOD데이터 설명 : 수질평가 시간과 생화학적 산소요구량의 관계를 보여주는 자료
변수설명
time = 수질 평가 측정 시간 , A numeric vector giving the time of the measurement (days).
demand = 산소요구량 , A numeric vector giving the biochemical oxygen demand (mg/l).

Descriptive Statistic

- BOD데이터를 활용한 막대 그래프




```
>
> barplot(BOD$demand,names.arg=BOD$Time)
>
```

Descriptive Statistic

- ToothGrowth데이터를 활용한 상자그림

- 데이터 확인

ToothGrowth {datasets} R Documentation

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Description

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

Usage

```
ToothGrowth
```

Format

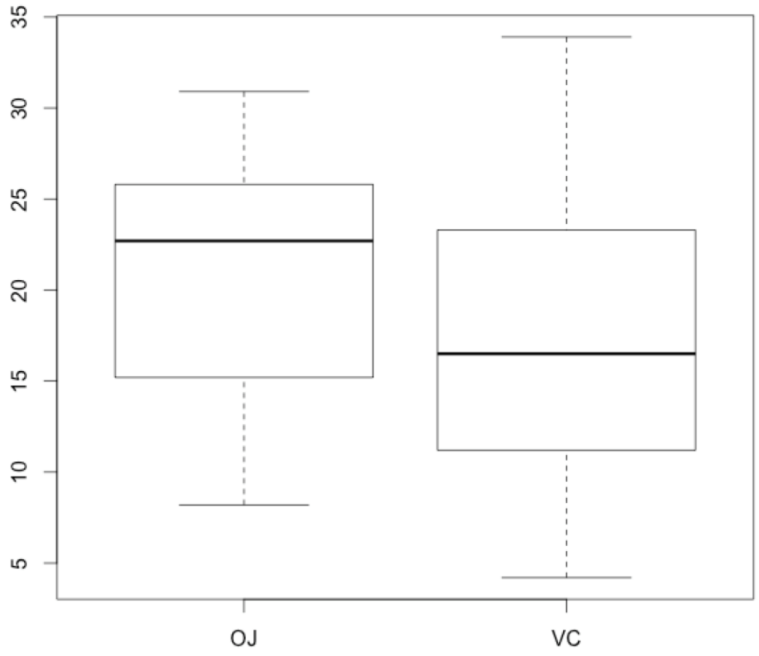
A data frame with 60 observations on 3 variables.

[,1] len numeric Tooth length
[,2] supp factor Supplement type (VC or OJ).
[,3] dose numeric Dose in milligrams.

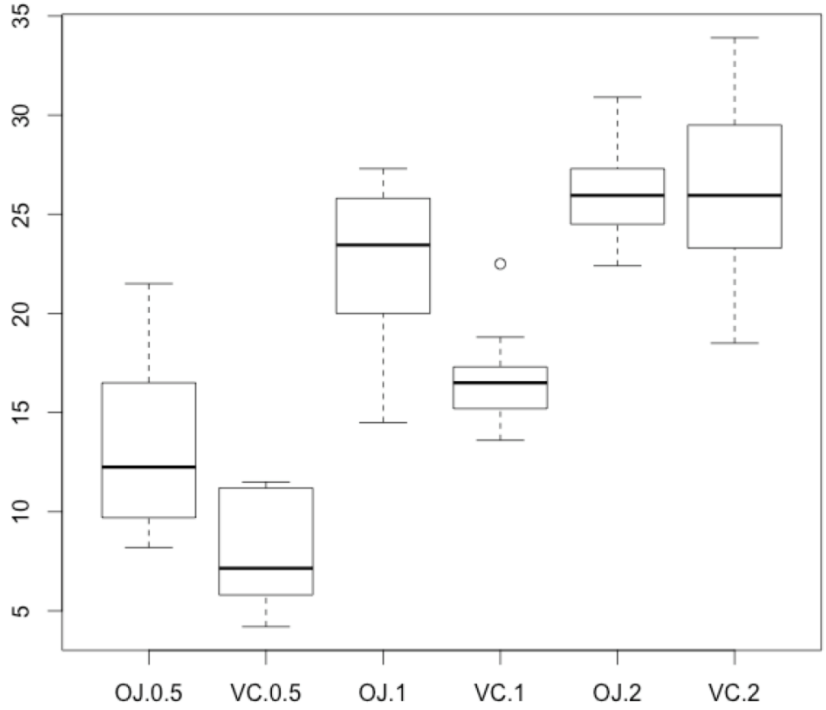
ToothGrowth데이터 설명 :데이터는 각 기니피그 아세포치아의 길이에 대해 두가지 전송방법(오렌지쥬스, 아스코르브산)과 비타민C의 세레벨(0.5mg,1mg,2mg)을 혼합하여 대입하였을때의 반응을 비교한 데이터이다.

변수설명
len 기니피그치아의 길이(Tooth length)
supp 두가지 전송방법(Supplement type (VC or OJ))
dose 비타민 C의 레벨(Dose in milligrams.)

- ToothGrowth데이터를 활용한 상자그림




```
>
> boxplot(len ~ supp, data = ToothGrowth)
>
```




```
>
> boxplot(len ~ supp + dose, data = ToothGrowth)
>
```