

# Big Data



About..

컴퓨터소프트웨어공학과  
김 원 일

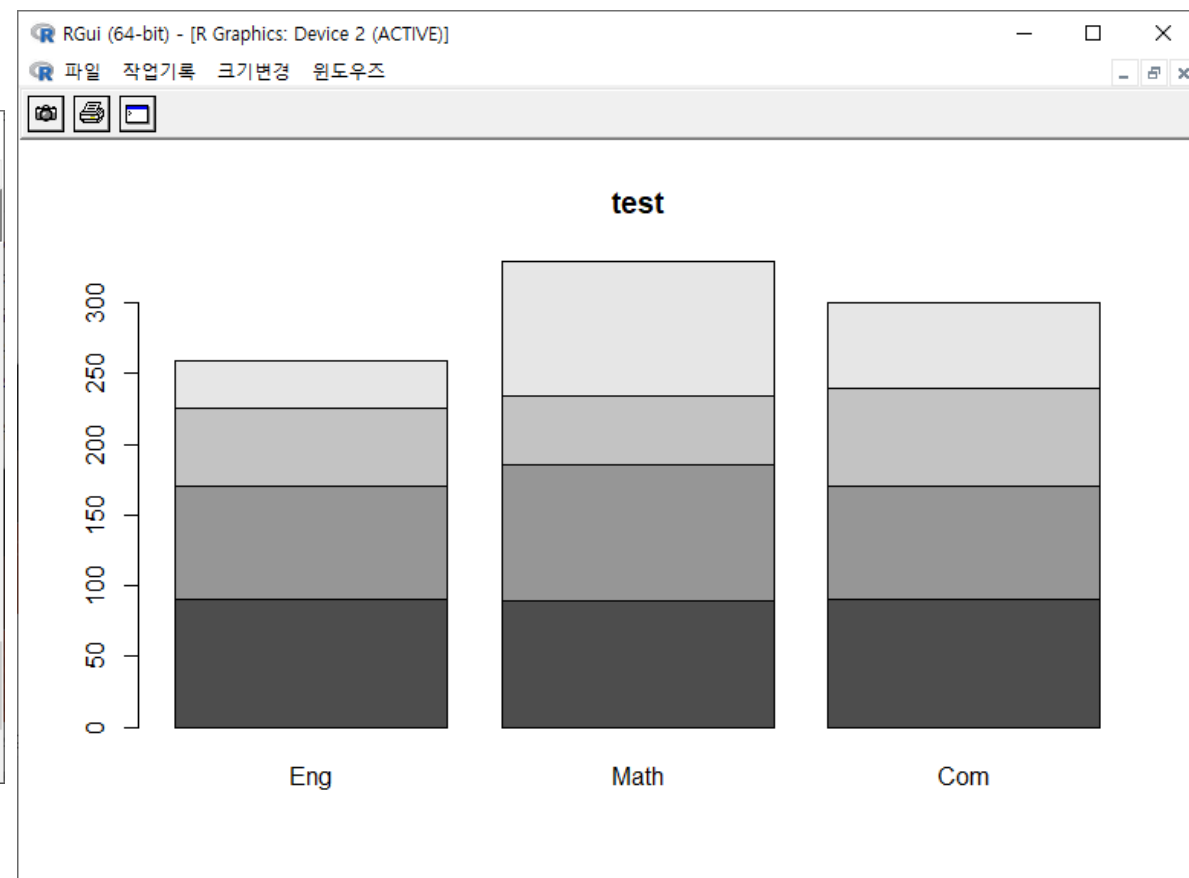


## 간단한 그래프 - 1

- matrix를 통한 그래프 작성
  - 이전 예제의 score matrix를 이용하여 그래프 그리기
  - barplot( )으로 간단한 막대 그래프 그리기 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> class( score )
[1] "matrix" "array"
> score
      Eng Math Com
kim    90   89  90
lee    80   96  80
park   55   49  70
choi   34   95  60
> barplot( score, main = "test" )
> |
```





## 간단한 그래프 - 2

### • 그래프가 보이지 않는 경우

- barplot( ) 옵션 변경 후, 실행 시 그래프가 나타나지 않는 경우
- 상단 메뉴의 "윈도우즈"에 생성된 그래프를 선택해야 변경된 그래프 확인이 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

[1] "matrix" "array"
> score
      Eng Math Com
kim    90    89  90
lee    80    96  80
park   55    49  70
choi   34    95  60
> barplot( score, main = "test" )
> barplot( score, main = "test", col = "blue" )
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

[1] "matrix" "
> score
      Eng Math
kim    90    89  90
lee    80    96  80
park   55    49  70
choi   34    95  60
> barplot( score, main = "test" )
> barplot( score, main = "test", col = "blue" )
> |
```

윈도우즈 메뉴:

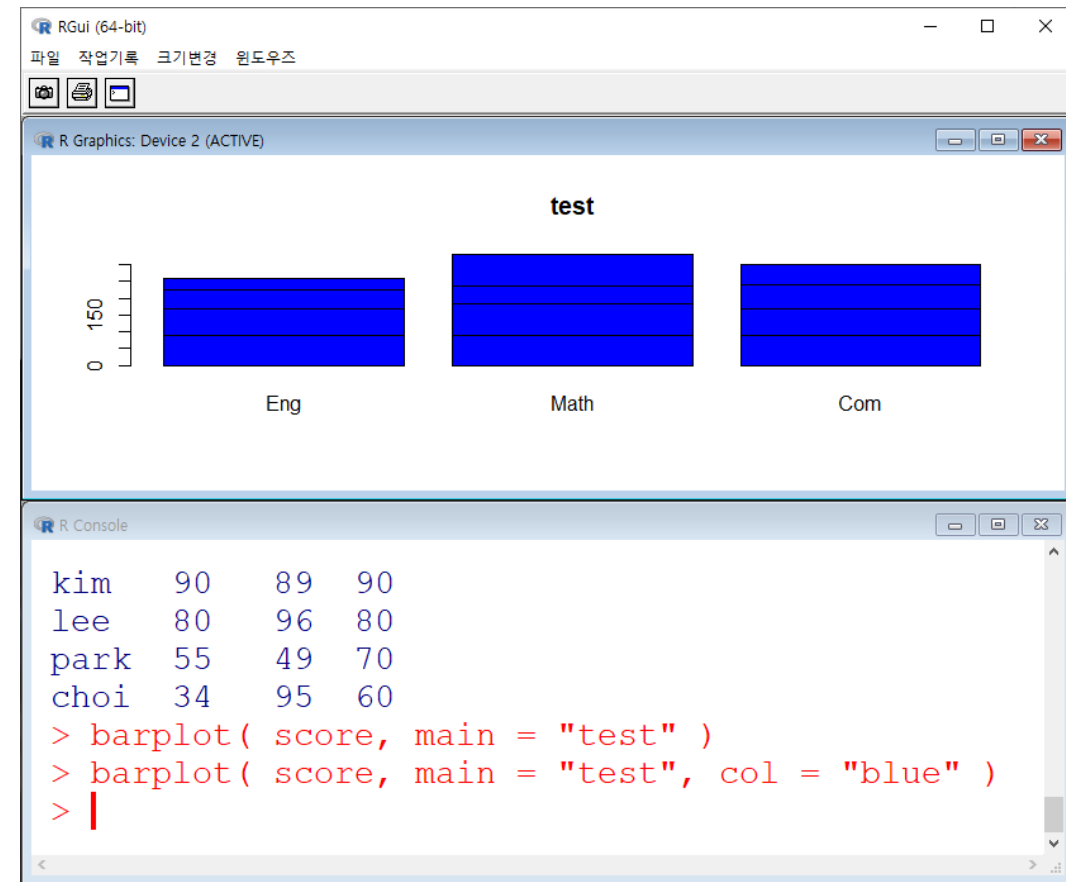
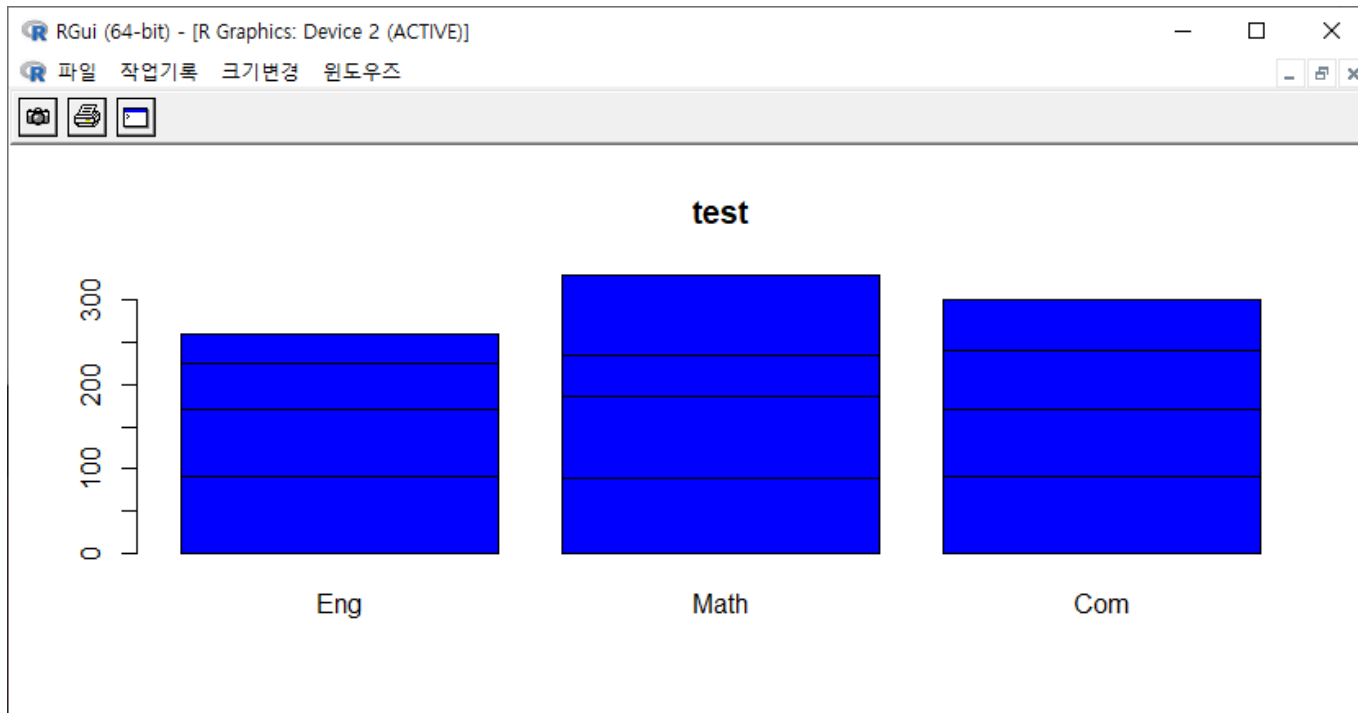
- 계단식
- 창을 가로로 정렬
- 창을 세로로 정렬
- 아이콘 배열
- ☒ 1 R Console
- ☐ 2 R Graphics: Device 2 (ACTIVE)



## 간단한 그래프 - 3

### • 변경된 그래프를 확인 가능

- 상단 메뉴의 선택이 없으면 지속적으로 표시되지 않아 문제가 있다고 판단할 수 있음
- 이미 그려진 그래프에 업데이트만 수행하므로 확인이 필요함
- 창 배치를 통해 지속적으로 확인하면서 진행해야 함





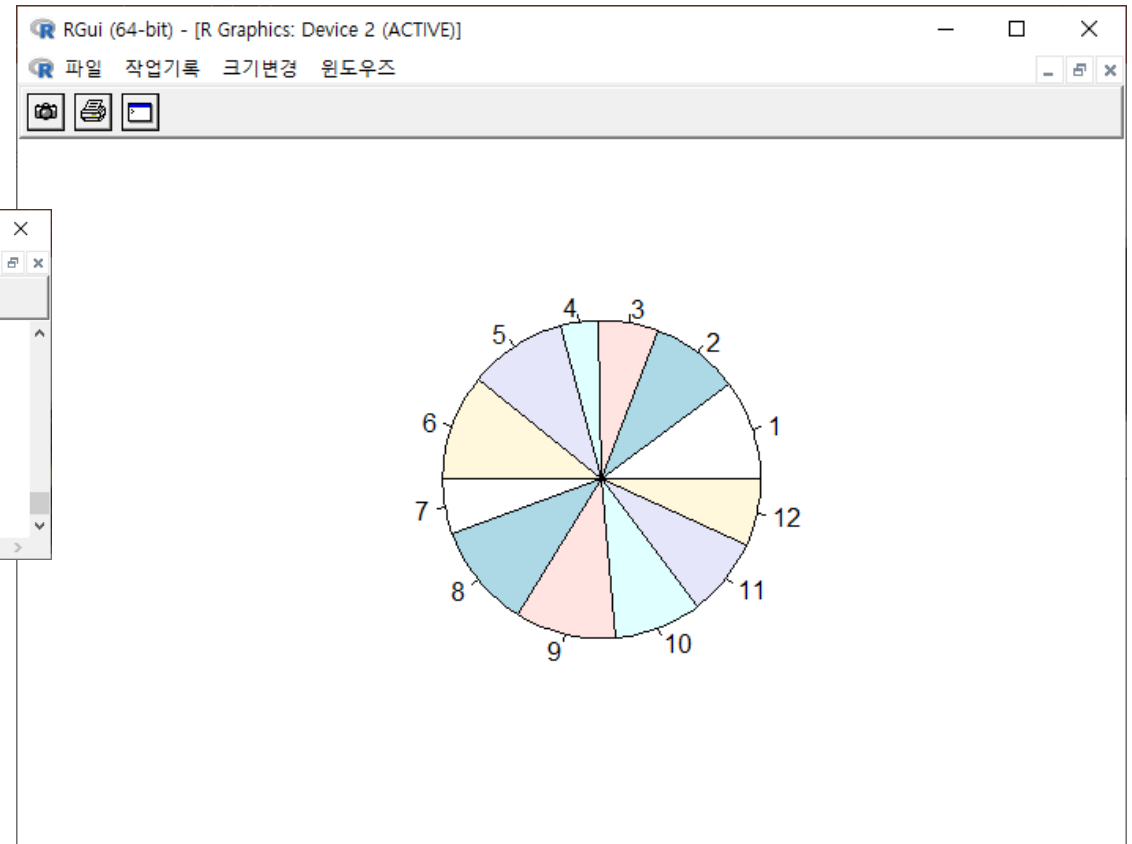
## 간단한 그래프 - 4

### • 파이 그래프

- 원형으로 데이터의 분포를 확인할 수 있도록 지원
- `pie( )` 함수를 통해 간단하게 그래프 작성 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> barplot( score, main = "test" )
> barplot( score, main = "test", col = "blue" )
> pie( score )
> |
```



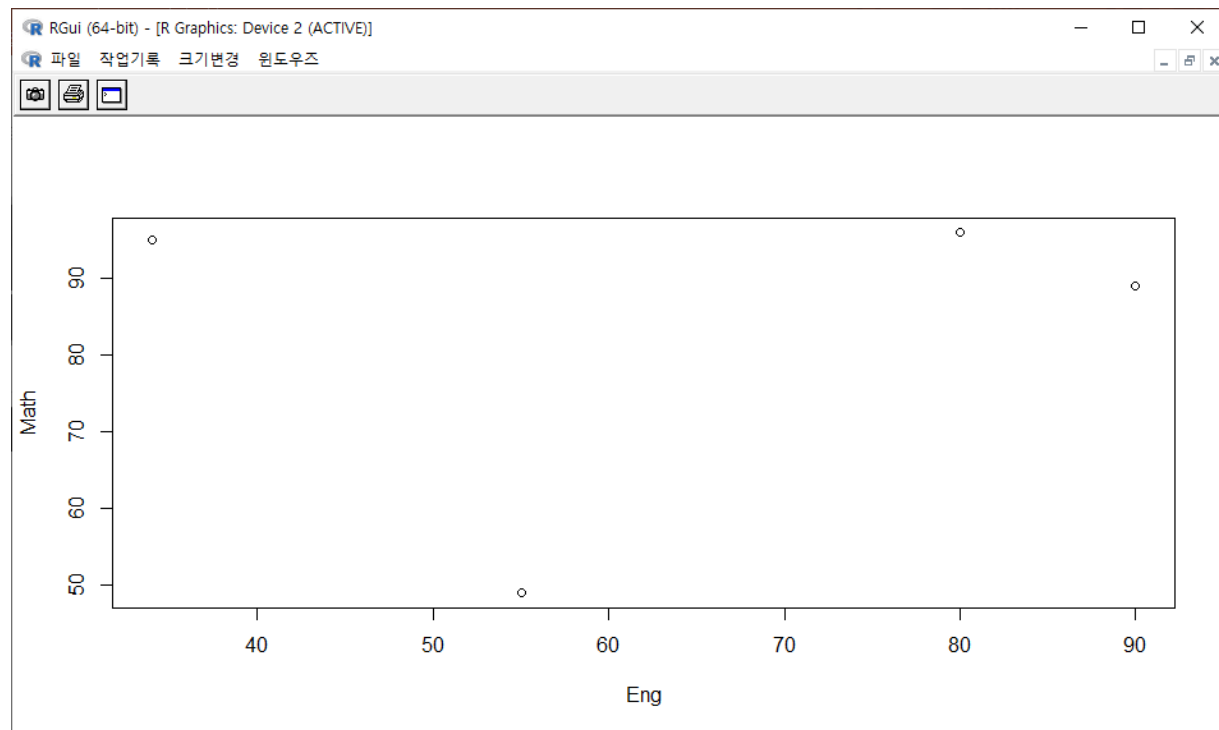


## • 단순 그래프

- 전달된 matrix를 이용하여 간단하게 위치를 점으로 표시
- plot( ) 함수로 간단하게 표시 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> barplot( score, main = "test" )
> barplot( score, main = "test", col = "blue" )
> pie( score )
> plot( score )
> |
```





## • 데이터 분석 절차 - 1

### - 문제 정의

- 데이터 분석은 해결하려는 문제를 명확히 정의하는 것에서 시작
- 문제 해결을 위해 어떤 데이터를 수집하고 분석할 지 계획할 수 있음

### - 데이터 수집

- 문제가 명확히 정의되면 문제를 해결하기 위해 필요한 데이터가 무엇인지 파악
- 필요한 데이터들을 수집하는 과정
- 문제 해결에 필요한 데이터는 다양한 형태로 존재 ex) 데이터베이스, 엑셀 파일, 문서 등

### - 데이터 전처리

- 수집 데이터는 바로 분석에 사용할 수 없을 때가 많음 ex) 단위, 결측값, 이상값
- 수집 데이터를 분석 가능한 형태로 정돈하는 데이터 정제 또는 전처리(data preprocessing) 과정 필요

문제 정의

데이터 수집

데이터 전처리

데이터 탐색

데이터 분석

결과보고



## • 데이터 분석 절차 - 2

### - 데이터 탐색

- 분석을 위해 정돈된 데이터 자체를 이해하고 파악하는 가벼운 데이터 분석 과정
- 비교적 간단하고 쉬운 통계 기법을 적용하여 전반적인 데이터의 내용을 파악
- 탐색적 데이터 분석(EDA, Exploratory Data Analysis)이라고도 함

### - 데이터 분석

- 데이터 탐색에서 파악한 정보를 바탕으로 보다 심화된 분석 수행

### - 결과 보고

- 데이터 분석 및 해석을 보고서 형태로 작성
- 최초 정의했던 문제를 해결하는 데 도움이 되는 내용으로 요약
- 데이터 시각화 기술이 중요하게 활용됨

문제 정의

데이터 수집

데이터 전처리

데이터 탐색

데이터 분석

결과보고





# 단일 데이터 분석 예제

## • 데이터 분석

- 데이터 분석은 수집 데이터의 종류에 따라 다른 형태의 도구가 활용
- 단일 변수 데이터 분석
  - vector와 같은 형태로 수집된 데이터를 이용하여 분석 진행
- 다중 변수 데이터 분석
  - matrix나 data frame 형태로 수집된 데이터를 이용하여 분석 진행

```
RGU (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> install.packages( 'carData' )
'C:/Users/YUHAN/Documents/R/win-library/4.1'의 위치에 패키지(들)을 설치합니다.
(왜냐하면 'lib'가 지정되지 않았기 때문입니다)
--- 현재 세션에서 사용할 CRAN 미러를 선택해 주세요 ---
URL 'https://cloud.r-project.org/bin/windows/contrib/4.1/carData_3.0-4.zip'을 시도함$
Content type 'application/zip' length 1822339 bytes (1.7 MB)
downloaded 1.7 MB

package 'carData' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\Public\Documents\ESTsoft\CreatorTemp\Rtmpe4db5P\downloaded_packages
> library( catData )
library(catData)에서 다음과 같은 에러가 발생했습니다: 'catData'이라고 불리는 패키지$
> library( carData )
> |
```



# 단일 데이터 분석 예제



## • 데이터 분석 예제 - 1

- 타이타닉 호 승객 데이터를 이용한 데이터 분석 예제
- carData 패키지에 포함된 데이터로 실제 승객 데이터를 제공
  - 이름, 생존여부, 성별, 나이, 선실 등급으로 구성된 data frame

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]

> library( carData )
> TitanicSurvival
```

	survived	sex	age	passengerClass
Allen, Miss. Elisabeth Walton	yes	female	29.0000	1st
Allison, Master. Hudson Trevor	yes	male	0.9167	1st
Allison, Miss. Helen Loraine	no	female	2.0000	1st
Allison, Mr. Hudson Joshua Crei	no	male	30.0000	1st
Allison, Mrs. Hudson J C (Bessi	no	female	25.0000	1st
Anderson, Mr. Harry	yes	male	48.0000	1st
Andrews, Miss. Kornelia Theodos	yes	female	63.0000	1st
Andrews, Mr. Thomas Jr	no	male	39.0000	1st
Appleton, Mrs. Edward Dale (Cha	yes	female	53.0000	1st





## • 데이터 분석 예제 - 3

- 데이터 탐색으로 처리에 필요한 정보들을 획득
- 문자열로 구성된 levels 정보를 이용하여 선실 등급 획득
- `sum( matrix )`를 통해 선실 등급의 합계 획득

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

[1259] 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd
[1276] 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd
[1293] 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd 3rd
Levels: 1st 2nd 3rd
> classTable <- table( classInfo )
> classTable
classInfo
1st 2nd 3rd
323 277 709
> sum( classTable )
[1] 1309
> |
```



## 단일 데이터 분석 예제



### • 데이터 분석 예제 - 4

- 탑승객의 선실 등급에 대한 비율을 계산
- 숫자 형식의 표현은 시각적으로 이해하는데 어려움

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]

> classTable
classInfo
1st 2nd 3rd
323 277 709
> sum( classTable )
[1] 1309
> barplot( classTable, main="class", xlab="level", ylab="count" )
> classTable/sum( classTable )
classInfo
      1st      2nd      3rd
0.2467532 0.2116119 0.5416348
> |
```



## 단일 데이터 분석 예제



### • 데이터 분석 예제 - 5

- 데이터 셋을 이용한 데이터를 분석
- 시각적으로 표현하여 분석 결과를 보기 편하게 작성
- 색상을 적용하여 보다 시각적인 효과를 줄 수 있음

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

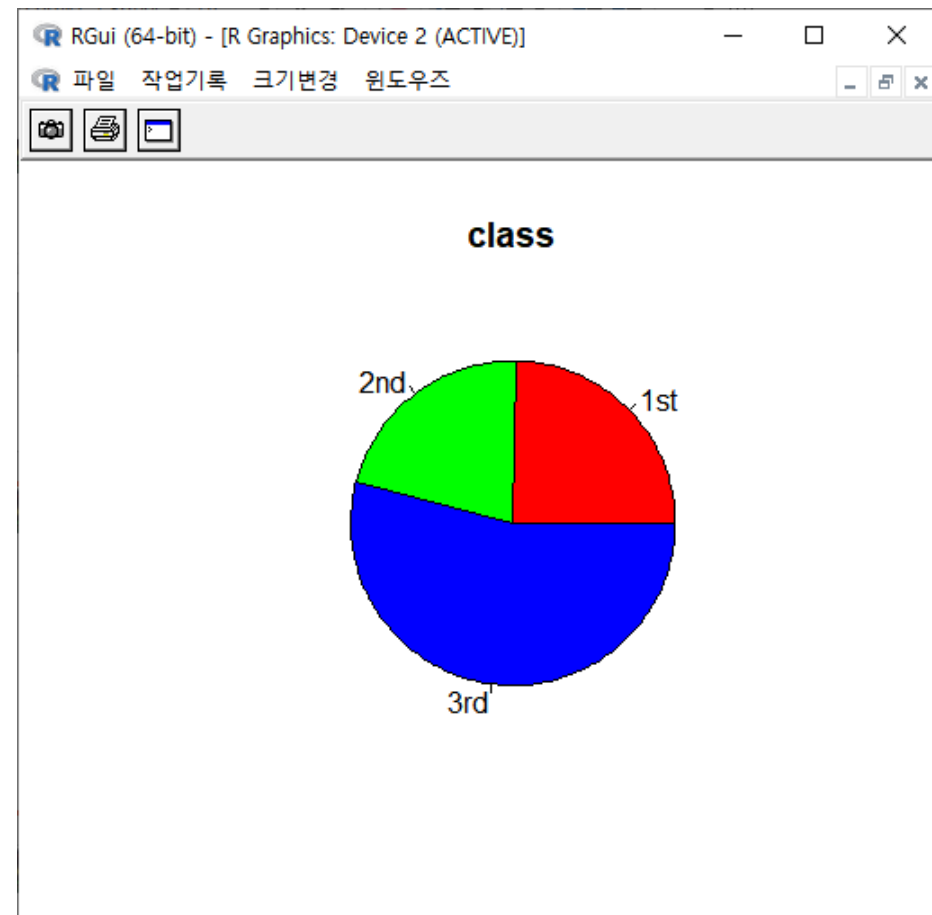
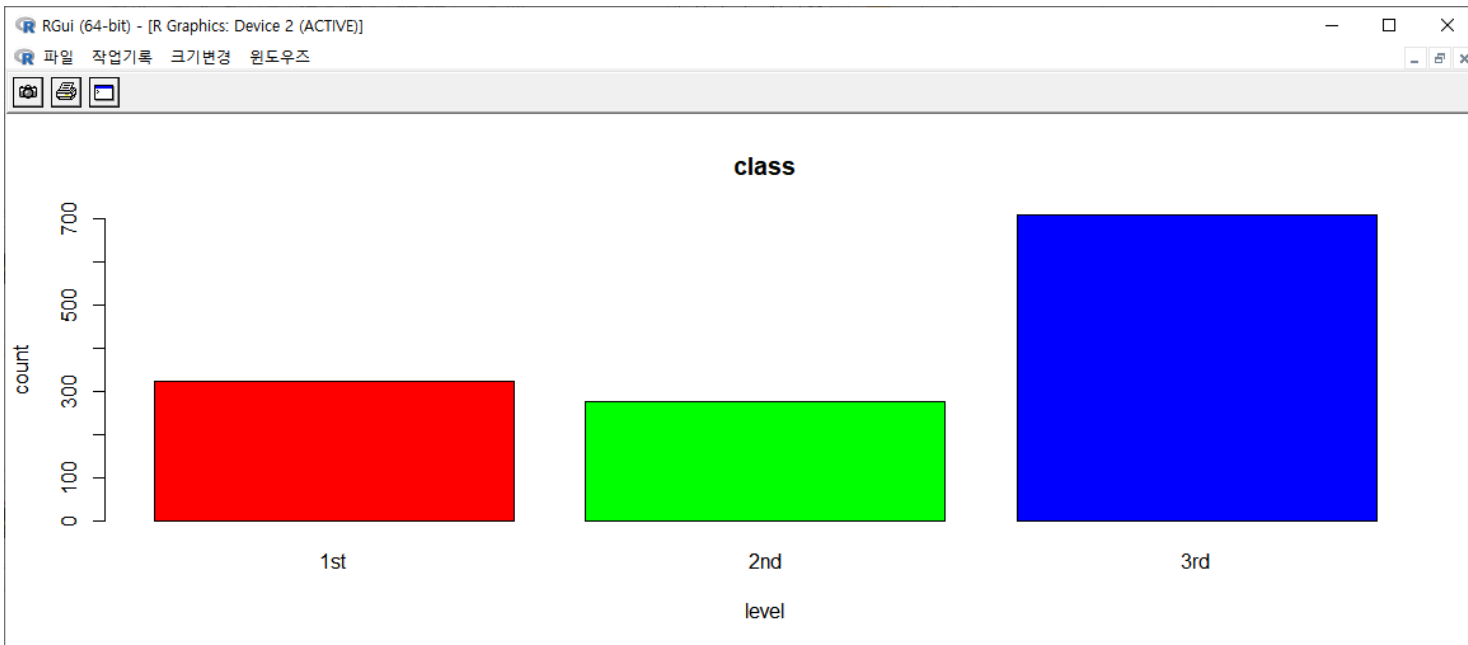
> barplot( classTable, main="class", xlab="level", ylab="count",
+ col = c( 'red', 'green', 'blue' ) )
> pie( classTable, main="class", col=c( 'red', 'green', 'blue' ) )
> |
```



# 단일 데이터 분석 예제

## • 데이터 분석 예제 - 6

- 막대 그래프와 파이 그래프로 시각적인 표현
- 사람이 인지하기가 보다 편함
- 결과
  - 3등급 선실의 인원이 대다수
  - 1등급 보다 2등급 선실 인원이 더 적었음 ...





## • 데이터 분석 예제 - 7

- 탑승자들의 나이를 vector로 생성하여 획득

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons: File Explorer, Run, Save, Print, Copy, Paste, Undo, Redo, Stop, Run]

> ageData <- TitanicSurvival$age
> ageData
 [1] 29.0000  0.9167  2.0000 30.0000 25.0000 48.0000 63.0000 39.0000
 [9] 53.0000 71.0000 47.0000 18.0000 24.0000 26.0000 80.0000      NA
[17] 24.0000 50.0000 32.0000 36.0000 37.0000 47.0000 26.0000 42.0000
[25] 29.0000 25.0000 25.0000 19.0000 35.0000 28.0000 45.0000 40.0000
[33] 30.0000 58.0000 42.0000 45.0000 22.0000      NA 41.0000 48.0000
[41]      NA 44.0000 59.0000 60.0000 41.0000 45.0000      NA 42.0000
[49] 53.0000 36.0000 58.0000 33.0000 28.0000 17.0000 11.0000 14.0000
[57] 36.0000 36.0000 49.0000      NA 36.0000 76.0000 46.0000 47.0000
[65] 27.0000 33.0000 36.0000 30.0000 45.0000      NA      NA 27.0000
```





## 단일 데이터 분석 예제



### • 데이터 분석 예제 - 8

- 획득된 데이터에 대한 요약 정리 함수로 정보 확인
- 숫자로 구성된 데이터로는 전체적인 확인이 어려움

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

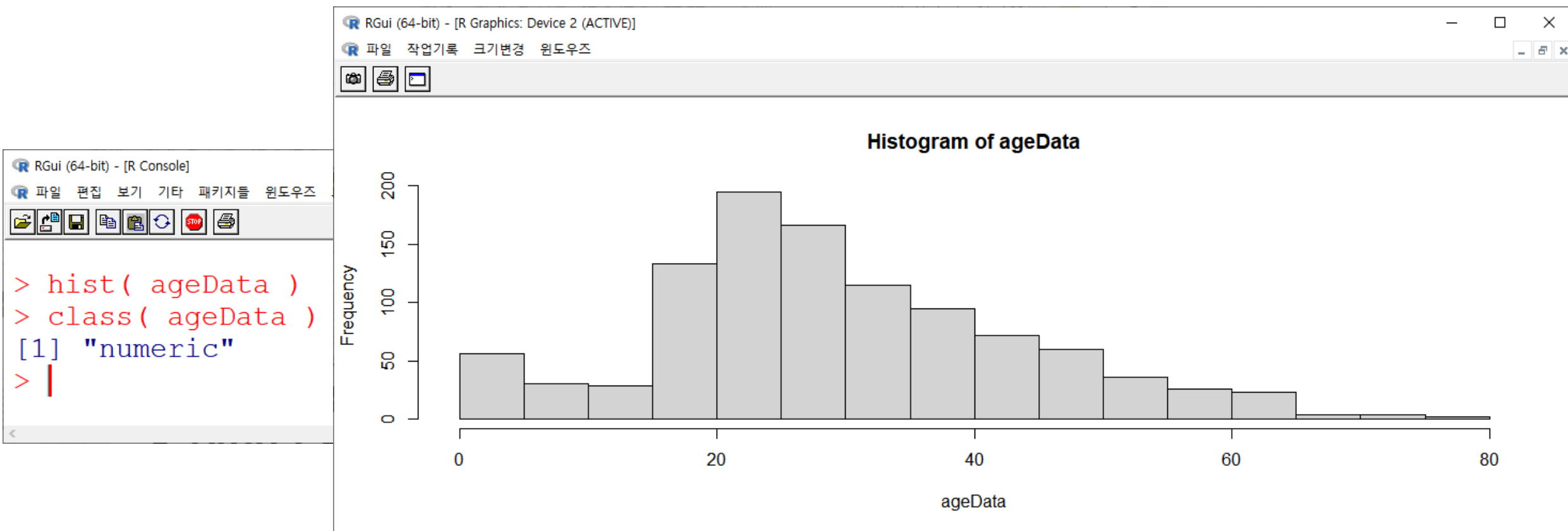
[1257]  7.0000  9.0000 29.0000 36.0000 18.0000 63.0000      NA 11.5000
[1265] 40.5000 10.0000 36.0000 30.0000      NA 33.0000 28.0000 28.0000
[1273] 47.0000 18.0000 31.0000 16.0000 31.0000 22.0000 20.0000 14.0000
[1281] 22.0000 22.0000      NA      NA      NA 32.5000 38.0000 51.0000
[1289] 18.0000 21.0000 47.0000      NA      NA      NA 28.5000 21.0000
[1297] 27.0000      NA 36.0000 27.0000 15.0000 45.5000      NA      NA
[1305] 14.5000      NA 26.5000 27.0000 29.0000

>
> summary( ageData )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
0|.1667 21.0000 28.0000 29.8811 39.0000 80.0000   263
```



## • 데이터 분석 예제 - 8

- 히스토그램을 이용한 전체 탑승객의 나이 분포에 대한 분석 결과

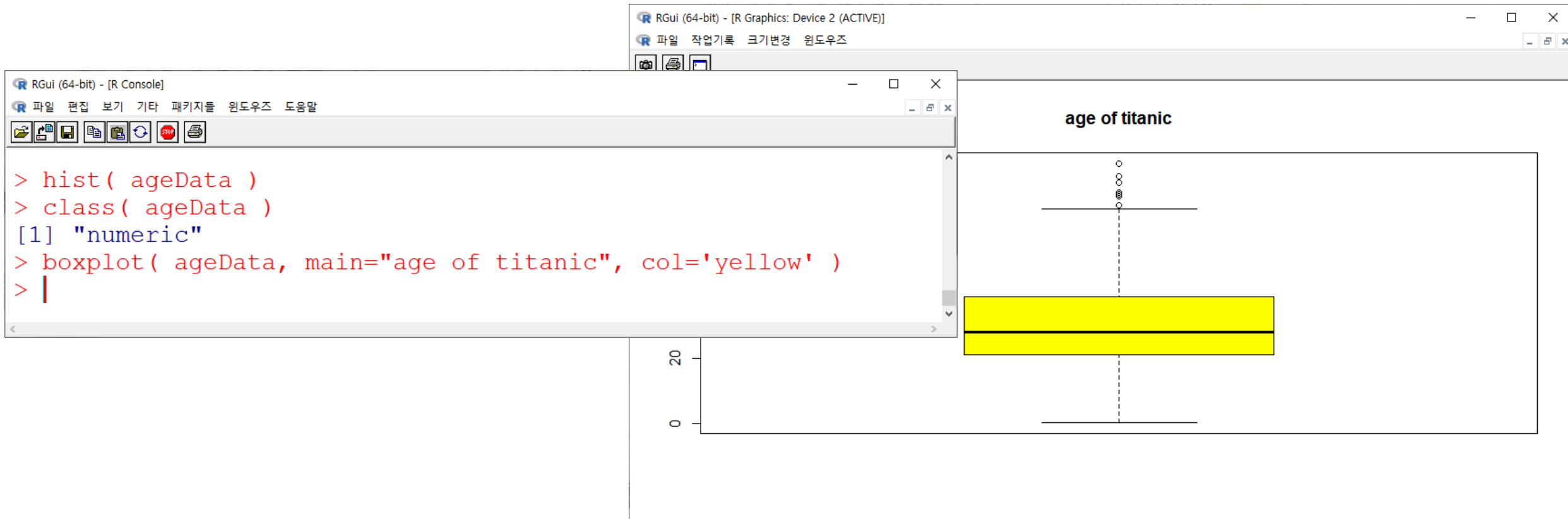




## 단일 데이터 분석 예제

### • 데이터 분석 예제 - 9

- boxplot( )으로 탑승자의 나이 대별 분석
- 노란색 상자가 절대 다수를 차지하는 나이 대를 나타냄
- 거의 중간에 박스가 위치함으로 전체적으로 고른 형태의 나이 분포를 보임





# 다중 데이터 분석 예제



## • 데이터 분석 예제 - 1

- 다중 데이터 분석 시 데이터 간의 관계에 집중해야 함
- 온도에 따른 기압 정보를 통해 데이터 분석을 수행
- R에 내장되어 있는 기본 데이터 셋

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]
> pressure
  temperature pressure
1           0  0.0002
2          20  0.0012
3          40  0.0060
4          60  0.0300
5          80  0.0900
6         100  0.2700
7         120  0.7500
8         140  1.8500
9         160  4.2000
10        180  8.8000
```



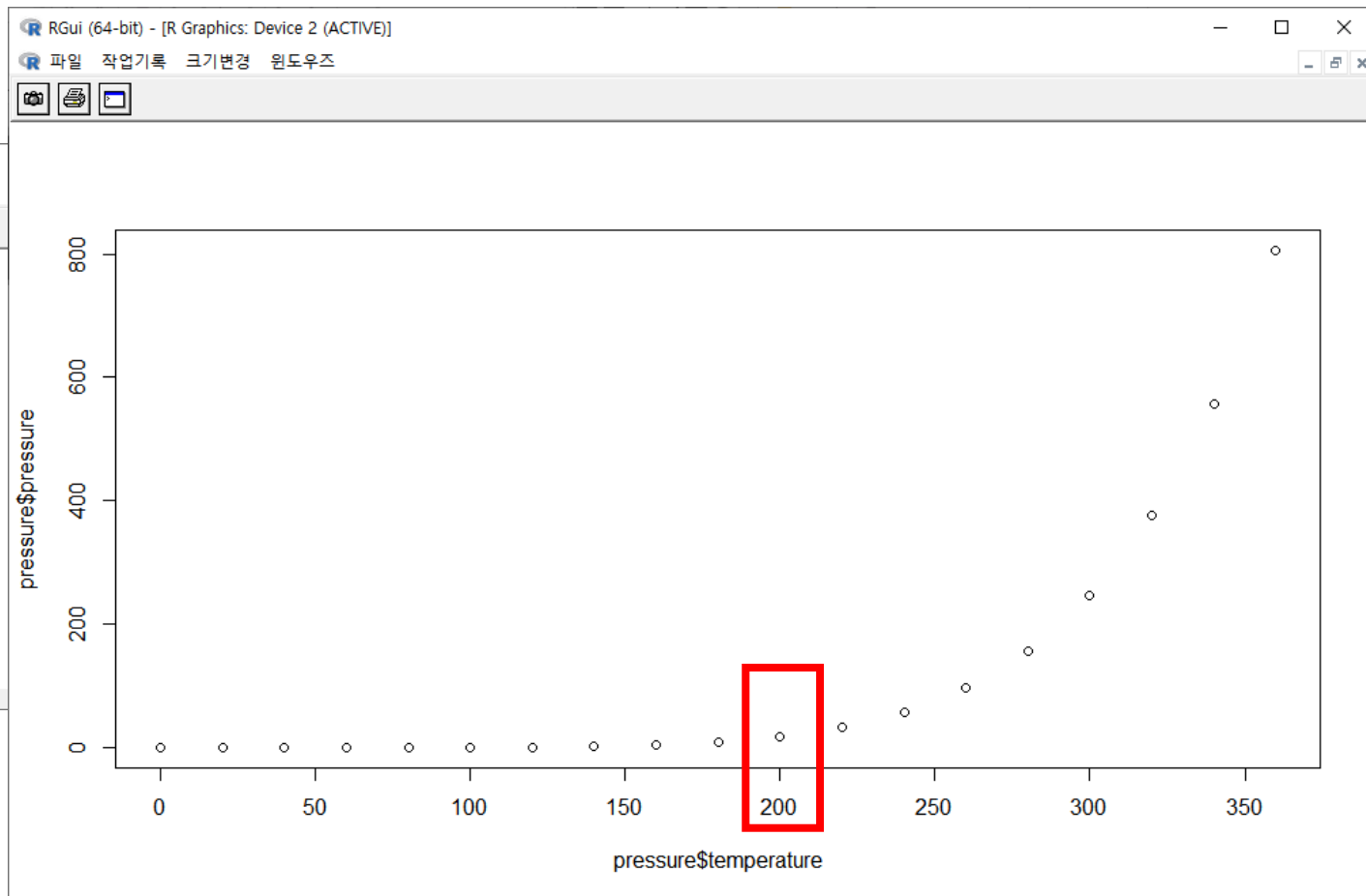
# 다중 데이터 분석 예제

## • 데이터 분석 예제 - 2

- 온도와 기압 간의 관계를 확인. 200도부터 기압이 크게 증가하는 것을 확인
- 여러 정보를 이용한 분석이 가능함

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

12      220  32.1000
13      240  57.0000
14      260  96.0000
15      280 157.0000
16      300 247.0000
17      320 376.0000
18      340 558.0000
19      360 806.0000
> plot( pressure$temperature, pressure$pressure )
> |
```





# 다중 데이터 분석 예제

## • 데이터 분석 예제 - 3

- cor( matrix, data frame )으로 데이터 간의 상관 관계 확인
- 포함된 데이터 가운데 연산 가능한 데이터로 정보 제공
- iris, pressure 데이터 셋으로도 확인 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> is <- subset( iris[ , 1:4 ] )
> cor( is )
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

```
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> cor( pressure )
```

	temperature	pressure
temperature	1.0000000	0.7577923
pressure	0.7577923	1.0000000

```
> |
```



## 다중 데이터 분석 예제



### • 데이터 분석 예제 - 4

- state.x77 데이터 셋은 미국 50주의 1977년의 통계 데이터
- R에 기본 내장되어 있는 테스트용 데이터 셋
- 인구, 수입, 범죄, 기대수명, 살인, 고등학교 졸업률 등을 포함

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> tmp <- data.frame( state.x77 )
> head( tmp )
      Population  Income Illiteracy Life.Exp Murder HS.Grad Frost  Area
Alabama      3615   3624         2.1   69.05   15.1    41.3    20  50708
Alaska        365   6315         1.5   69.31   11.3    66.7   152  566432
Arizona      2212   4530         1.8   70.55    7.8    58.1    15  113417
Arkansas      2110   3378         1.9   70.66   10.1    39.9    65   51945
California    21198  5114         1.1   71.71   10.3    62.6    20  156361
Colorado      2541   4884         0.7   72.06    6.8    63.9   166  103766
> plot( tmp )
> |
```



## • 데이터 분석 예제 - 5

- cor( matrix )로 데이터 간의 상관 관계를 확인
- 데이터들 간의 계산 가능한 정보들을 기반으로 상대적인 정보 확인 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말
[Icons]

> cor( tmp )
```

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
Population	1.00000000	0.2082276	0.10762237	-0.06805195	0.3436428	-0.09848975	-0.3321525	0.02254384
Income	0.20822756	1.0000000	-0.43707519	0.34025534	-0.2300776	0.61993232	0.2262822	0.36331544
Illiteracy	0.10762237	-0.4370752	1.00000000	-0.58847793	0.7029752	-0.65718861	-0.6719470	0.07726113
Life.Exp	-0.06805195	0.3402553	-0.58847793	1.00000000	-0.7808458	0.58221620	0.2620680	-0.10733194
Murder	0.34364275	-0.2300776	0.70297520	-0.78084575	1.0000000	-0.48797102	-0.5388834	0.22839021
HS.Grad	-0.09848975	0.6199323	-0.65718861	0.58221620	-0.4879710	1.00000000	0.3667797	0.33354187
Frost	-0.33215245	0.2262822	-0.67194697	0.26206801	-0.5388834	0.36677970	1.0000000	0.05922910
Area	0.02254384	0.3633154	0.07726113	-0.10733194	0.2283902	0.33354187	0.0592291	1.00000000

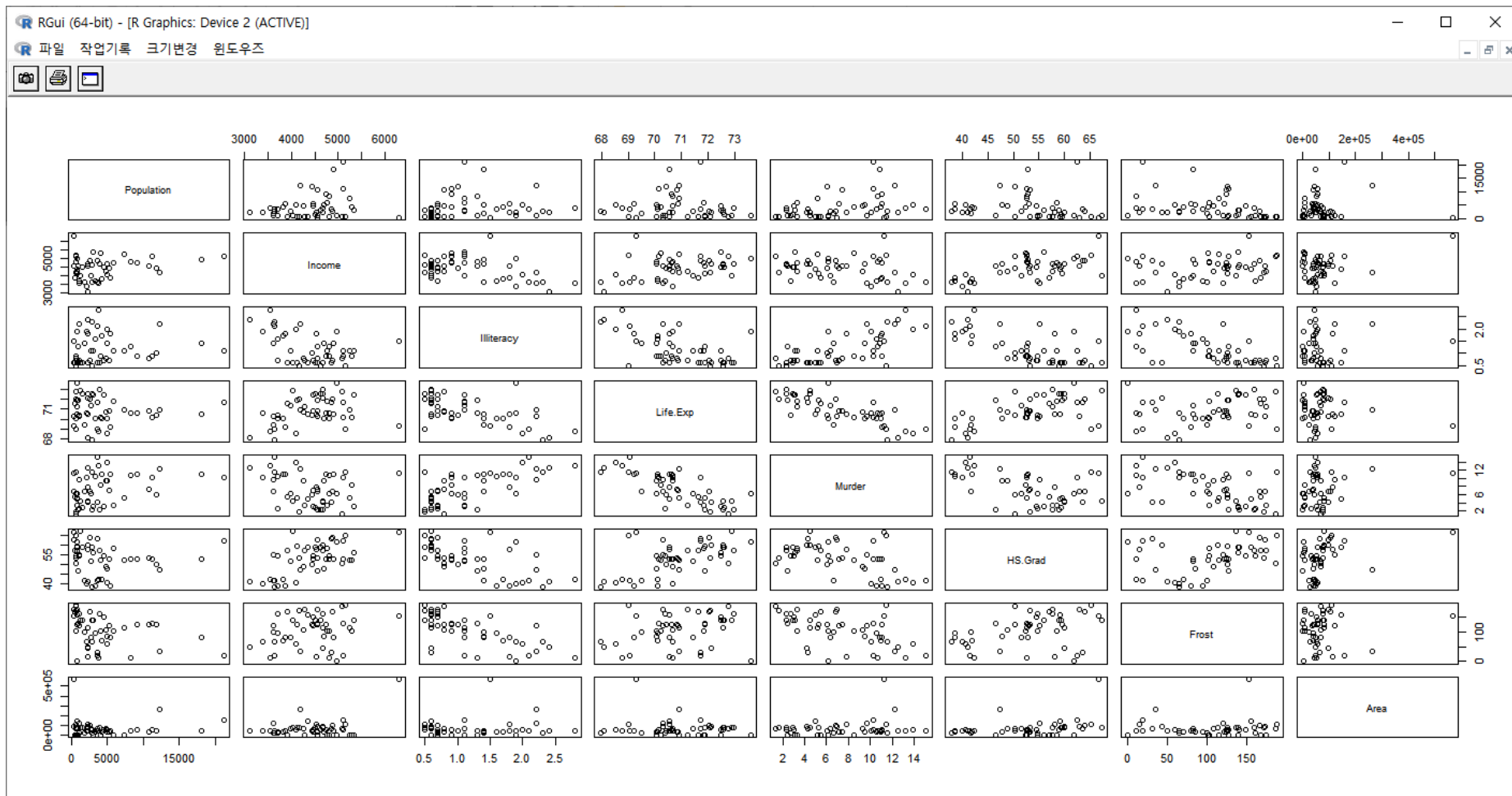
```
> |
```

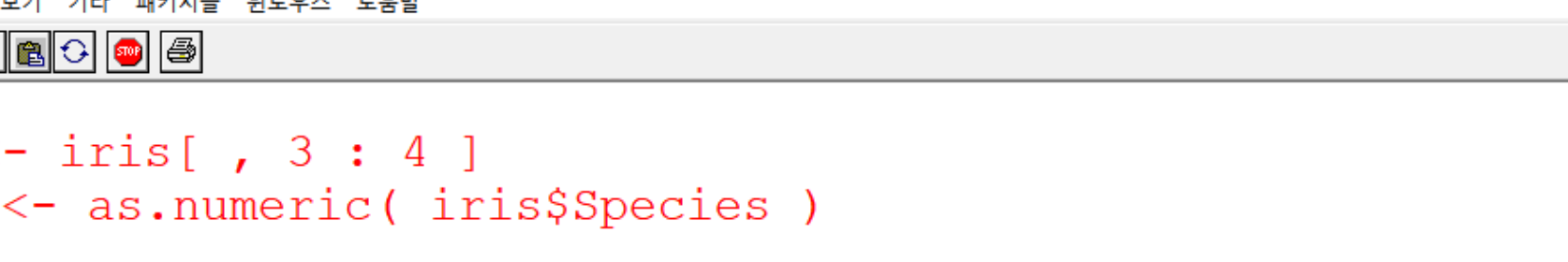




# 다중 데이터 분석 예제

- 데이터 분석 예제 - 4
  - 그래프로도 확인이 가능





The screenshot shows the RGui (64-bit) - [R Console] window. The title bar includes standard window controls (minimize, maximize, close) and a menu bar with options: 파일 (File), 편집 (Edit), 보기 (View), 기타 (Other), 패키지들 (Packages), 윈도우즈 (Windows), and 도움말 (Help). Below the menu bar is a toolbar with icons for file operations (open, save, print, etc.) and a stop button. The console area displays the following R code and its output:

```
> is <- iris[ , 3 : 4 ]
> ptr <- as.numeric( iris$Species )
> ptr
```

[1]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
[32]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
[63]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[94]	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
[125]	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

The prompt character > is followed by a vertical bar |.

