

Lecture 01: R 기초 - R로 데이터 요약하기

개요

- R과 Rstudio 설치
- R의 기본적인 문법 이해
- R을 이용한 자료 요약
- R 패키지 ggplot2을 이용하여 그림 그리기

R 소개

프로그래밍 언어 R

- R은 프로그래밍 언어
 - 프로그래밍 언어: 컴퓨터에게 일을 시키기 위한 수단
 - 컴퓨터: 기계어 사용, 사용자(인간): 사람의 언어(주로 영어) 사용
 - 컴퓨터에게 일을 시키기 위해서 만든 의사소통 도구
- R: interpreted language
 - compiled language
 - 컴퓨터에게 할 일을 미리 적어 놓고, 컴퓨터 언어로 미리 번역
 - 인간이 적어 놓은 책을 기계어로 번역하는 과정이 반드시 필요
 - 일단 번역된 작업은 아주 빠르게 실행 가능
 - interpreted language
 - 컴퓨터에게 할 일을 필요할 때마다 컴퓨터 언어로 동시 통역
 - 인간의 명령을 한꺼번에 번역하는 것이 아니라 필요할 때마다 해석
 - trial and error 방식의 대화형 작업이 가능
 - why interpreted language?
 - 분석 작업은 수 많은 시도와 수정 작업이 필요
 - 한 번에 완벽한 분석 과정을 만드는 것은 불가능
 - 컴퓨터와 대화하면서 분석 진행

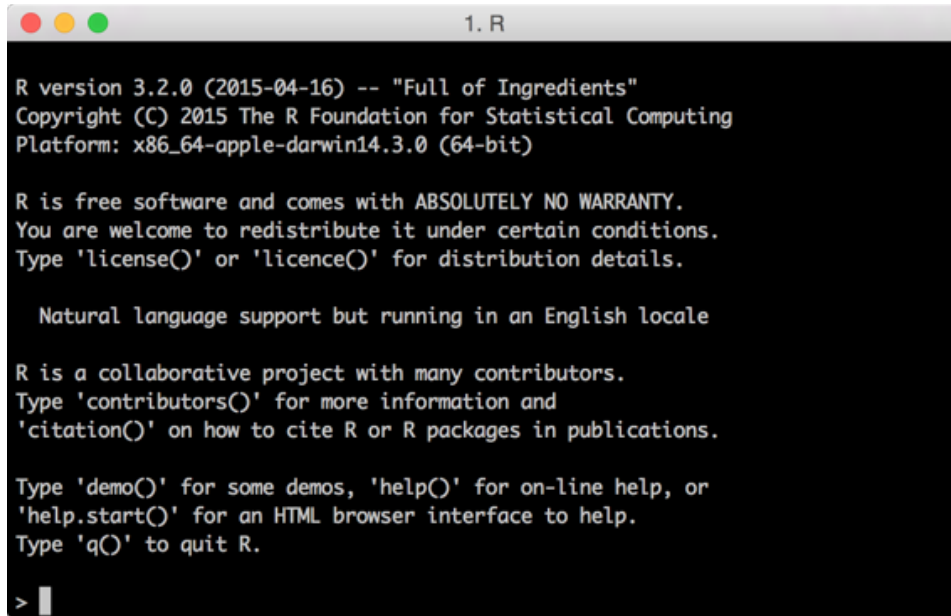
통계 분석 툴 R

- 프로그래밍 언어를 이용한 통계 분석 도구
- 오픈소스 프로젝트: 공짜, 사용자 책임
- 전세계 모든 통계학자나 분석가가 사용
- 새로운 방법론을 개발하면 R package로 만들어서 무료로 배포
 - http://cran.rstudio.com/web/packages/available_packages_by_name.html
- 거의 모든 분석 방법론은 R을 이용해서 편하게 적용 가능

Rstudio 소개

- <http://www.rstudio.com/>
- 오픈소스 IDE(integrated development environment), 공짜
- Desktop 버전과 Server 버전이 있음
- 현재 가장 대중적으로 쓰이는 IDE for R

R 실행 화면



```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin14.3.0 (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.

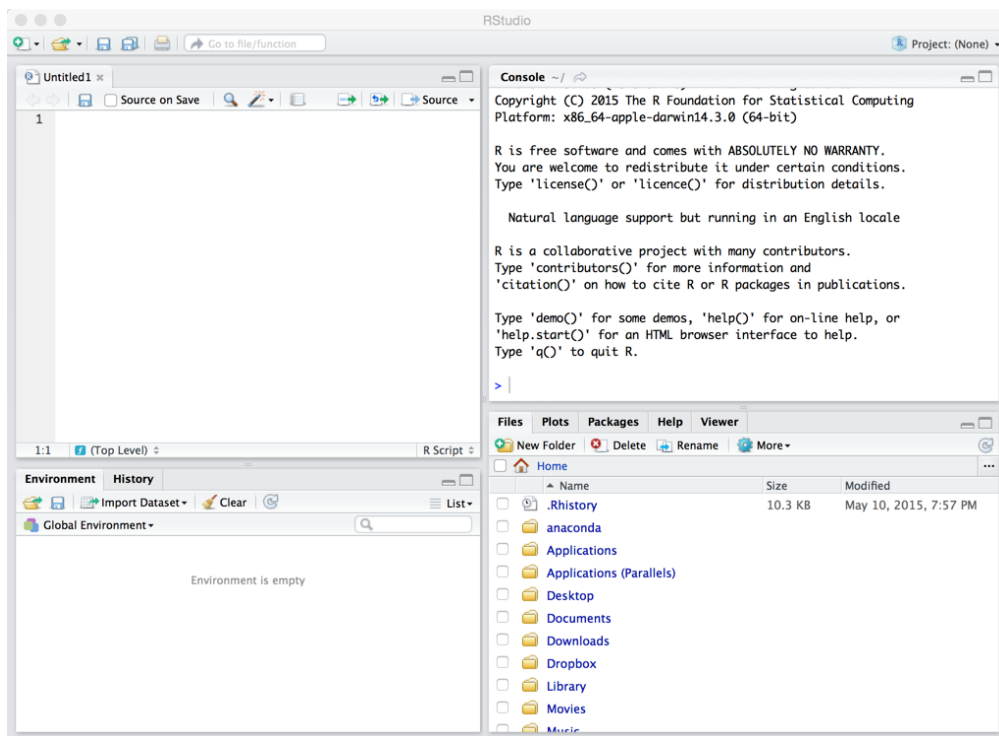
Natural language support but running in an English locale

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.

> |
```

Rstudio 실행 화면



rstudio 설치

- 반드시 R을 먼저 설치한 후에 설치
- 윈도우에서 **관리자 계정**으로 설치
- 윈도우에서 RStudio 설치 시 **경로명에 한글이나 특수문자가 포함되지 않아야** 합니다.
 - 윈도우 계정 이름이 한글인 경우, temp 폴더 경로에 한글이 포함되어 문제가 생기는 경우가 있습니다.
 - '제어판 / 사용자 계정 / 환경 변수 변경'에서 '사용자 변수' tmp temp 값을 'c:\Temp'로 바꾸면 해결됩니다.

R 기초

project 만들기

getwd() - 현재의 working directory를 확인
'File - New Project...' or Project 메뉴
모든 작업은 프로젝트 단위로 구분, 서로 영향을 주지 않도록 한다.

preference 변경

'Tools - Global Options...'
원하는 모습으로 세팅하기

console 환경

command line
대화형 작업 환경

script 작성

컴퓨터에게 시킬 일들을 절차적으로 정리한 문서
대화형 작업으로 할 수 없는 과업 수행
대부분의 분석 과정은 script로 작성하여 수행

keyboard shortcut

마우스 사용 최소화
재현 가능한 분석(reproducible analysis)

type of object

"To understand computations in R, two slogans are helpful:

- *Everything that exists is an object.*
- *Everything that happens is a function call."*

— John Chambers

object type을 따지는 이유는 type에 따라서 쓸 수 있는 function이 다르기 때문.

```
1 + 2  
'a' + 'b'
```

atomic vector

가장 기본이 되는 object type

데이터 분석에서 **vector**는 연산의 기본 단위.

```
1:10
c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10) # concatenate
is.vector(1:10)
length(1:10)
```

중요!: 하나의 벡터는 오직 하나의 데이터 타입(type of data)만 저장할 수 있다.

types of data

6가지: double, integer, character, logical, complex, raw

doubles

일반적인 숫자, 'numeric' 이라고도 부름.

```
typeof(pi)
```

integer

정수.

```
typeof(1L)
typeof(1)
typeof(1:10)
```

character

문자

```
typeof(c('one', 'two'))
```

중요!: object의 이름과 character data를 혼동하면 안된다.

logical

논리

```
typeof(1 > 2)
typeof(c(TRUE, FALSE))
c(T, F)
```

Attributes

vector에 속성을 부여하여 class를 바꾼다.

```
v.10 <- 1:10
attributes(v.10)
dim(v.10) <- c(2, 5)
attributes(v.10)
class(v.10)
```

factor

범주형 자료(categorical data): 성별, 학점 등

```

factor class
v.10 <- 1:10
f.10 <- factor(v.10)
class(f.10)
attributes(f.10)
typeof(f.10)

```

중요!: 마치 **character data type**처럼 보이지만, 실제로는 **integer**.

indexing

```

v.10 <- 1:10
v.10[1]
v.10[2:4]
dim(v.10) <- c(2, 10)
v.10[1, 2]
v.10[1, 3:4]
v.10[-1, c(2, 5)]
v.10[1, ]
v.10[, 2]

```

R 데이터 읽기

txt, csv 파일 읽기

<http://goo.gl/eBU8B1>

가장 기본이 되는 데이터 저장 방식

```
read.table(file)
```

중요한 argument

```

header
stringsAsFactors
na.strings

```

데이터 살펴보기

```

head(raw.diamonds)
View(raw.diamonds)
str(raw.diamonds)

```

R packages

다른 사람들(통계학자, 프로그래머 등)이 만들어놓은 툴

```

install.packages('ggplot2')
library('ggplot2')

```

R ggplot2

'the grammar of graphics'에 근거해 만든 그림 그리기 package(Hadley Wicham)
기본 문법이 쉽고, 캠퍼스에 레이어(layer)를 더하는 방식으로 그리기.

<http://docs.ggplot2.org/current/>

EDA(exploratory data analysis)

- John Tukey, 탐색적 자료 분석
 - <https://books.google.com/books?id=UT9dAAAAIAAJ>

"The processes of criminal justice are clearly divided between the search for the evidence—the responsibility of the police and other investigative forces—and the evaluation of the evidence's strength—a matter for juries and judges. ... Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone—as the first step."

- 증거를 찾는 활동 vs. 증거의 강한 정도를 밝히는 활동
 - EDA: 증거를 찾는 활동, 단서 찾기
 - 추론통계: 증거의 유의성을 밝히는 활동

Central Tendency

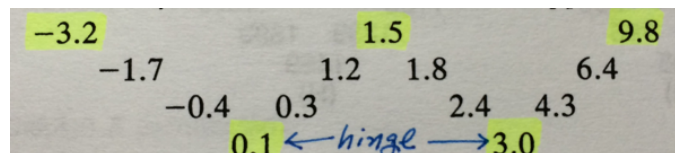
- 자료의 중심을 나타내면 대표값
- 평균, 중위수, 최빈값 등
- http://en.wikipedia.org/wiki/Central_tendency

Dispersion

- 자료의 퍼짐을 나타내는 대표값
- 분산, IQR,
- http://en.wikipedia.org/wiki/Statistical_dispersion

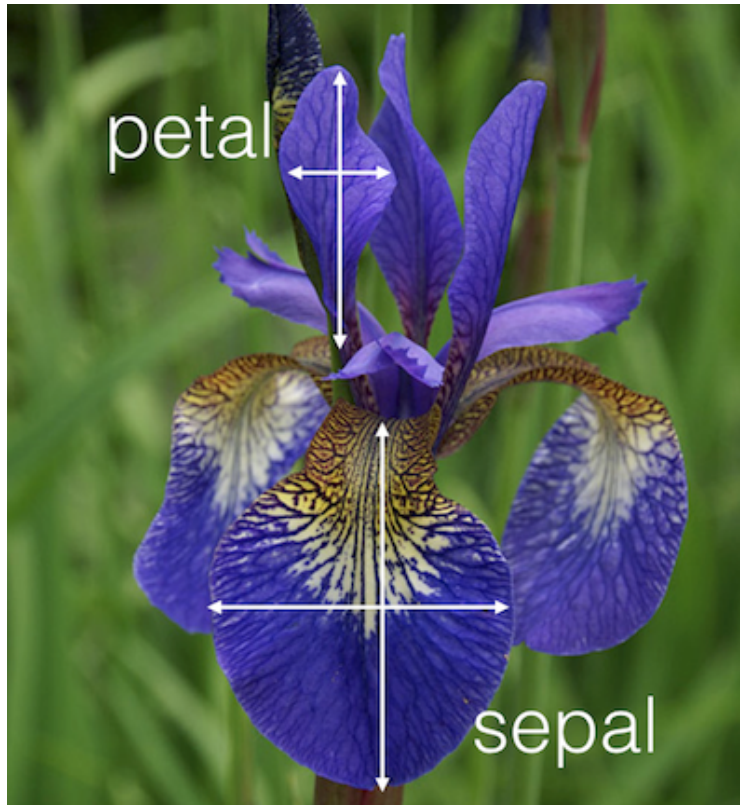
5 number summary

- 자료를 5개의 숫자로만 요약
 - minimum: 최소값
 - first quartile: 첫 번째 사분위수
 - median (middle value): 중위수 = 두 번째 사분위수
 - third quartile: 세 번째 사분위수
 - maximum: 최대값



iris data

http://en.wikipedia.org/wiki/Iris_flower_data_set



출처: http://sebastianraschka.com/Articles/2014_python_lda.html

```
> summary(iris)
  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width      Species
Min.   :4.300    Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
1st Qu.:5.100    1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800    Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843    Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400    3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900    Max.   :4.400   Max.   :6.900   Max.   :2.500
```

참고: http://en.wikipedia.org/wiki/Five-number_summary

상자 그림(box plot)

- 5-number summary의 시각화 버전
- 4등분 경계를 수직선 위에 표현.
- IQR(inter-quartile range = third quartile - first quartile): 박스의 길이

- inner fence: $1.5 * IQR \rightarrow$ outlier
- outer fence: $3.0 * IQR \rightarrow$ extreme

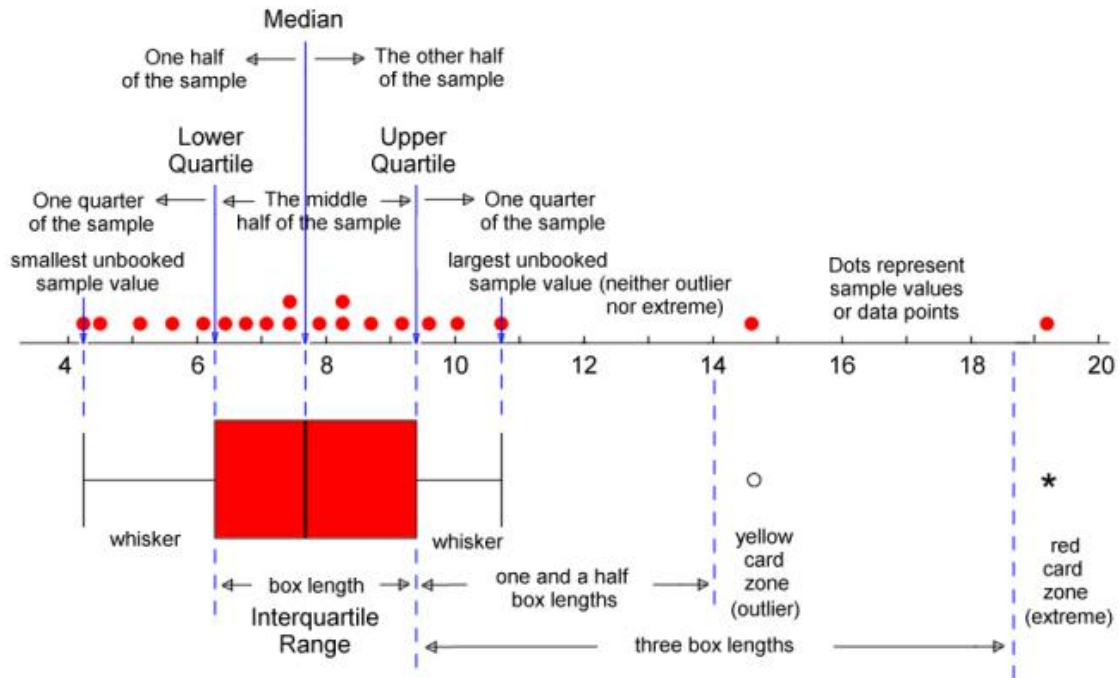
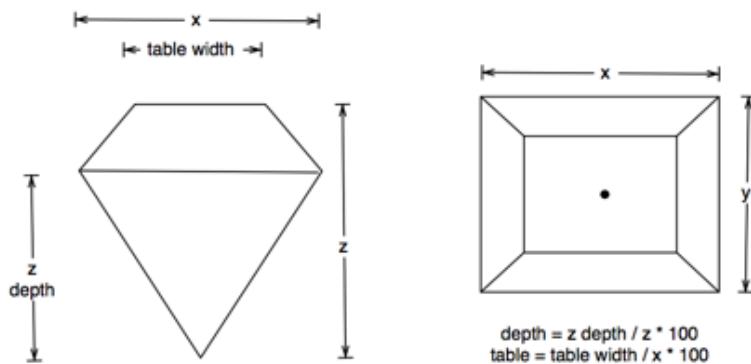


그림 출처: <http://web.pdx.edu/~stipakb/download/PA551/boxplot.html>

data: diamonds



Details

- price. price in US dollars (\\$326–\\$18,823)
- carat. weight of the diamond (0.2–5.01)
- cut. quality of the cut (Fair, Good, Very Good, Premium, Ideal)

- colour. diamond colour, from J (worst) to D (best)
- clarity. a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
- x. length in mm (0–10.74)
- y. width in mm (0–58.9)
- z. depth in mm (0–31.8)
- depth. total depth percentage = $z / \text{mean}(x, y) = 2 * z / (x + y)$ (43–79)
- table. width of top of diamond relative to widest point (43–95)

<http://www.diamondse.info/>

```
library(ggplot2)
data(diamonds)
```

가격에 영향을 주는 요소는 무엇일까?

```
ggplot(diamonds, aes(x=carat, y=price)) + geom_point()
```