Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

457.212 Statistics for Civil & Environmental Engineers
In-Class Material: Class 01
Introduction to Course & "R" Software Environment

1. Instructor: Prof. Junho Song (http://systemreliability.wordpress.com)

2. Course introduction: Check syllabus and calendar (available at http://etl.snu.ac.kr/)

3. What is "R"?

(a) Past, present and future of R (https://simplystatistics.org/2018/07/12/use-r-keynote-2018/)

"R is a **free** software environment for **statistical computing** and **graphics**"
- https://www.r-project.org/

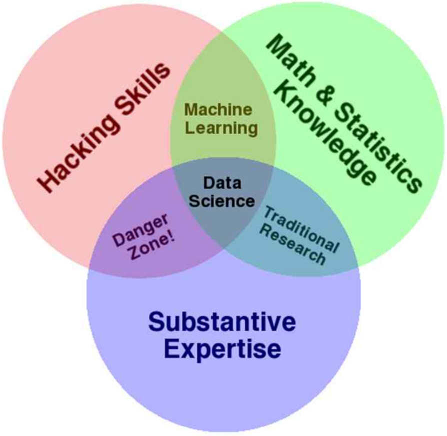
"The tidyverse (R package collection) is an opinionated collection of R packages designed for **data science**"
- https://www.tidyverse.org/

(b) Why named R?

R was developed by **R**oss Ihaka and **R**obert Gentleman (initial version and first stable beta versions released in 1995 and 2000) using the **S** programming language.

(c) But… why in this course, Prof. Song?

- "Practice does not make perfect. Only perfect practice makes perfect." – Vince Lombardi: What is a perfect practice to learn p_________ and s_______?
- Introduction to civil and environmental engineering via r____ d____
- They say it's a golden age of data science. But who is a d____ s______? This course provides math & statistics knowledge in the context of civil and environmental engineering (i.e. domain expertise). What about introduction to "hacking skills"?



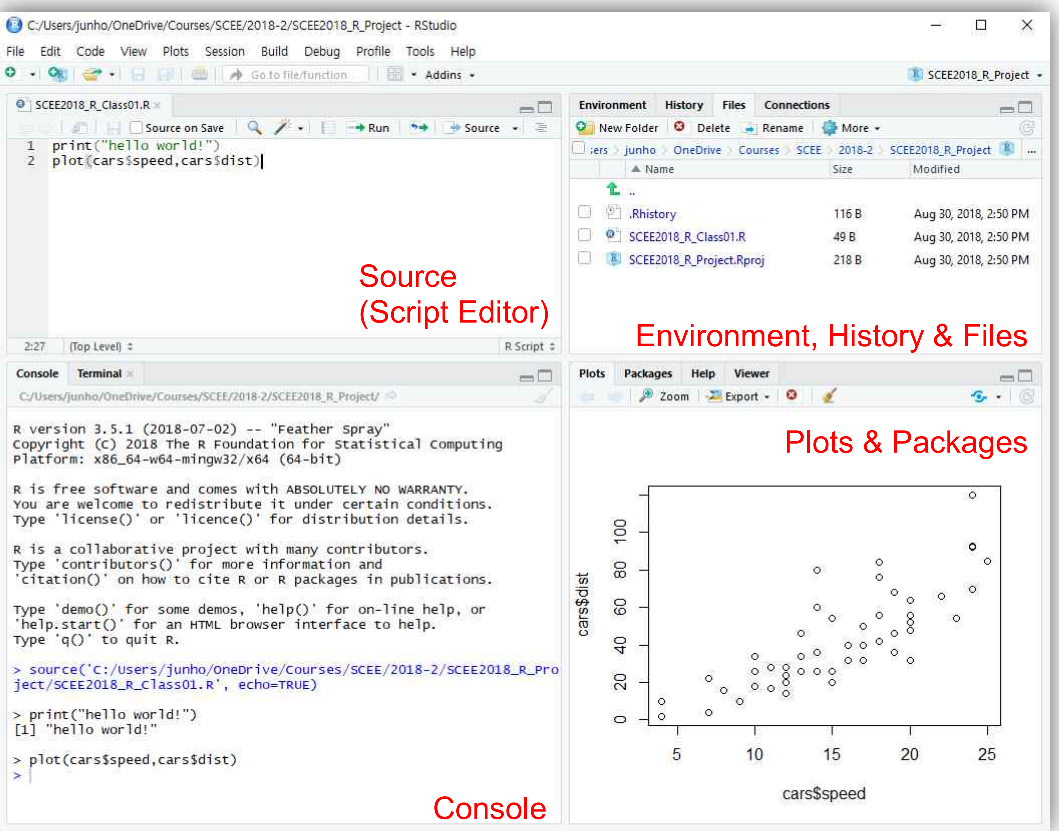
The Data Science Venn Diagram (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram)

4. Recommended software environment for R: R Studio

(a) First, download and install R “base” on your computer: https://www.r-project.org/

(b) Download and install the most popular integrated development environment (IDE) of R, called RStudio from https://www.rstudio.com/products/rstudio/

(c) The interface of RStudio:


C:/Users/junho/OneDrive/Courses/SCEE/2018-2/SCEE2018_R_Project - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
SCEE2018_R_Project
SCEE2018_R_Class01.R
Source on Save
Run
Source
2:27
(Top Level)
R Script
Source
(Script Editor)
Environment History Files Connections
New Folder Delete Rename More
:ers › junho › OneDrive › Courses › SCEE › 2018-2 › SCEE2018_R_Project
Environment, History & Files
Console Terminal
C:/Users/junho/OneDrive/Courses/SCEE/2018-2/SCEE2018_R_Project/
Console
Plots Packages Help Viewer
Zoom
Export
Plots & Packages
cars$dist
0
20
40
60
80
100
5
10
15
20
25
cars$speed

```
print("hello world!")
plot(cars$speed,cars$dist)
```

Name	Size	Modified
..		
.Rhistory	116 B	Aug 30, 2018, 2:50 PM
SCEE2018_R_Class01.R	49 B	Aug 30, 2018, 2:50 PM
SCEE2018_R_Project.Rproj	218 B	Aug 30, 2018, 2:50 PM

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
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.

> source('C:/Users/junho/OneDrive/Courses/SCEE/2018-2/SCEE2018_R_Pro
ject/SCEE2018_R_Class01.R', echo=TRUE)

> print("hello world!")
[1] "hello world!"

> plot(cars$speed,cars$dist)
> 
```

(d) Tips & recommendations

- Create a Project by “File→New Project” menu
- Choose “New Directory” option to create a working folder for your project
- Create an R Script file by “File→New File→R Script” menu
- You can create your own interface layout at “View→Panes→Pane Layout” Menu
- “Ctrl+Enter”: Run the commands in the current line or selected area in the Editor
- “Ctrl+Shift+Enter” or “Ctrl+Shift+S”: Run the entire script with/without echoes
- Set your working folder as the default one by “Set as Working Directory”

(e) References for R and RStudio

- The R Project (https://www.r-project.org/): Check “Manuals” and “Books” sections
- RStudio Website: Check “Resources” section

5. Basics of the R language

(a) "R may be seen as a powerful calculator" (Härdle et al. 2017)

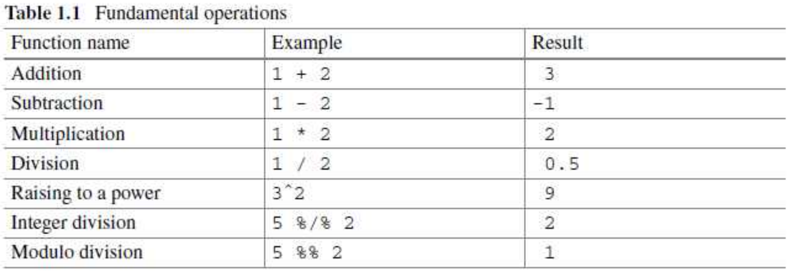


Table 1.1 Fundamental operations

Function name	Example	Result
Addition	`1 + 2`	`3`
Subtraction	`1 - 2`	`-1`
Multiplication	`1 * 2`	`2`
Division	`1 / 2`	`0.5`
Raising to a power	`3^2`	`9`
Integer division	`5 %/% 2`	`2`
Modulo division	`5 %% 2`	`1`

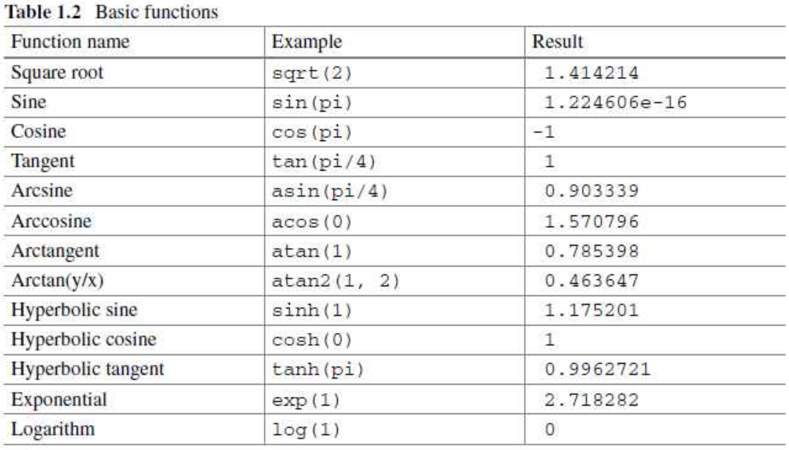


Table 1.2 Basic functions

Function name	Example	Result
Square root	`sqrt(2)`	`1.414214`
Sine	`sin(pi)`	`1.224606e-16`
Cosine	`cos(pi)`	`-1`
Tangent	`tan(pi/4)`	`1`
Arcsine	`asin(pi/4)`	`0.903339`
Arccosine	`acos(0)`	`1.570796`
Arctangent	`atan(1)`	`0.785398`
Arctan(y/x)	`atan2(1, 2)`	`0.463647`
Hyperbolic sine	`sinh(1)`	`1.175201`
Hyperbolic cosine	`cosh(0)`	`1`
Hyperbolic tangent	`tanh(pi)`	`0.9962721`
Exponential	`exp(1)`	`2.718282`
Logarithm	`log(1)`	`0`

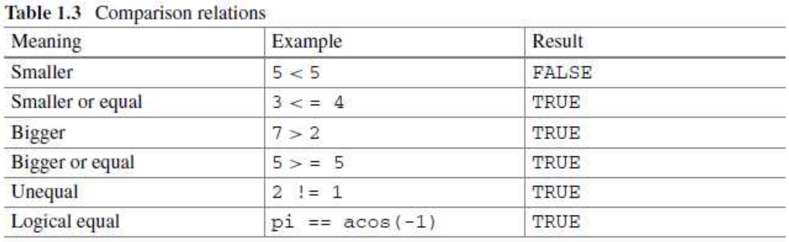


Table 1.3 Comparison relations

Meaning	Example	Result
Smaller	`5 < 5`	`FALSE`
Smaller or equal	`3 <= 4`	`TRUE`
Bigger	`7 > 2`	`TRUE`
Bigger or equal	`5 >= 5`	`TRUE`
Unequal	`2 != 1`	`TRUE`
Logical equal	`pi == acos(-1)`	`TRUE`

Can you guess the result of the followings? (Hint: Euler's identity)

```
i = complex(real=0, imaginary=1)
exp(i*pi) == cos(pi)+i*sin(pi)
exp(i*pi) != cos(pi)+i*sin(pi)
```

(b) Assigning variables (cf. deleting: rm(*variable*) or rm(list=ls()))

```
a = pi; a
b <- 2.0; b
c1 <- a; c1
b -> c2; c2
sin(c1*c2)
```

6. Array

(a) There are three types/categories of data: *numerical*, *character* and *logical*

```
x = 1.25 # numerical
y = "Hello" # character
z = x > 3 # logical
```

(b) Data structures in R: Arrays, data frames and lists

	Array	Data Frame	List
Data type	Same	Flexible	Flexible
Length	Same (for matrix)	Same	Flexible

(c) **Vector**: one-dimensional array

Creating a vector using c() function – “concatenate”

```
v1 = c(1.3, 2*pi, sqrt(2))
v1[3]
```

Basic vector operations

```
v2 = v1*2
v3 = v1+3
v4 = v2*v3 # Element-wise multiplication
v5 = v3*v3^(-1) # Element-wise inversion and product
v1 <= 1.3 # Element-wise logical operation (>,<,>=,<=,==,!=)
```

Creating a vector using colon, array() or seq() functions

```
c(1:10)
c(1,3:10)

array(1:3,8) # Repeat 1:3, but cannot exceed 8 elements

seq(1,10) # From 1 to 10 (default increment = 1)
seq(1,10, by=2) # Increment = 2
seq(1,10, length.out=3) # Increment determined to have 3 elements
```

Selecting, excluding and locating elements

```
v1[c(1,3)] # 1st and 3rd element only
v1[v1>1.3] # elements greater than 1.3 only

v1[-2] # Excluding 2nd element

which(v1>1.3) # Locating elements satisfying the equation
```

Note: ( ) is used to provide arguments for functions while [ ] is used to access elements or part of arrays or data frames.

(d) **Matrix**: two-dimensional array

Creating a matrix using matrix() or transforming a vector to matrix by dim()

```
matrix(0,2,5) # matrix of zeros, 2x5
matrix(1:12, nrow=3) # Put 1 to 12 into 3x? matrix (column-wise)
matrix(1:12, ncol=3, byrow=TRUE) # row-wise

m = 1:6
dim(m) = c(2,3) # Change the dimension to 2x3 matrix
```

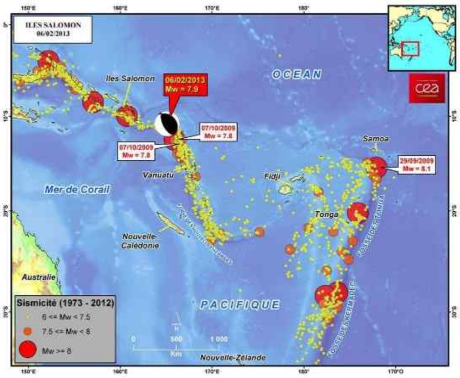
Locating element, row, column or submatrix

```
y = matrix(c(1,3:17),4,4) # 4x4 matrix with 1,3,…,17

y[2,3] # Element at 2nd row and 3rd column
y[2,] # 2nd row
y[,2] # 2nd column
y[3] # 3rd element in the column-wise sequence
y[1:2, 3:4] # submatrix: 1st and 2nd rows + 3rd and 4th columns
```

(e) Example (numeric) matrix available in R base package: *quakes*

- Details: https://rdrr.io/r/datasets/quakes.html
- Numeric matrix array with ______ rows and _______ columns



(Figure credit: http://www-dase.cea.fr/actu/dossiers_scientifiques/2013-02-06/index_en.html)

```
?quakes # '?' to get help regarding data, packages and functions
dim(quakes) # check the dimension of the matrix
summary(quakes) # summarize the data set
pairs(quakes, main = "Fiji Earthquakes, N = 1000", cex.main = 1.2,
pch = "o") # Create a scatterplot matrices

# Exercise 01: access 1st and 2nd columns and assign variables
"Latitude" and "Longitude" respectively


# Exercise 02: create a scatter plot for pairs of Longitude (x-axis)
and Latitude (y-axis)
```

7. **Data Frame**

(a) Looks like a matrix, but its columns can have different data types

(b) Example of constructing a data frame using data.frame() function

```
cities = c("Berlin", "New York", "Paris", "Tokyo")
```

```
area = c(892, 1214, 105, 2188)
population = c(3.4, 8.1, 2.1, 12.9)
continent = factor(c("Europe", "North America", "Europe", "Asia"))
myframe = data.frame(cities, area, population, continent)

# Assign names to rows
rownames(myframe) = c("Berlin", "New York", "Paris", "Tokyo")
# Assign names to columns (default: original vector names)
colnames(myframe) = c("City", "Area", "Pop.", "Continent")
```

Note: factor() function is needed when used as indicators

(c) Augmenting the data frame with a new vector

```
f = factor(c("Inland", "Coastal", "Inland", "Coastal"))
myframe = data.frame(myframe, f)
colnames(myframe)[5] = "Prox.Sea"
```

Alternatively,

```
myframe = cbind(myframe, "Prox.Sea" = factor(c("Inland", "Coastal",
"Inland", "Coastal")))
```

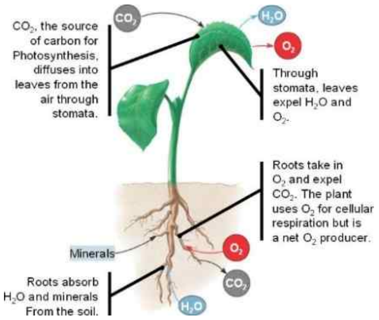
(d) Addressing one particular column in a data frame

```
myframe$Area
myframe[,2]
myframe[,"Area"]

myframe["Area"] # gives a subdata frame
```

(e) Example data frame available in R base package: *CO2*

- Details: https://rdrr.io/rforge/MEMSS/man/CO2.html
- Data frame with _____ rows and _____ columns


CO_2, the source of carbon for Photosynthesis, diffuses into leaves from the air through stomata.
CO_2
H_2O
O_2
Through stomata, leaves expel H_2O and O_2.
Roots take in O_2 and expel CO_2. The plant uses O_2 for cellular respiration but is a net O_2 producer.
Minerals
O_2
CO_2
Roots absorb H_2O and minerals From the soil.
H_2O

(Figure credit: http://bio1903.nicerweb.com/Locked/media/ch37/plant_nutrient.html)

```
?CO2
dim(CO2)
summary(CO2)

# Exercise 01: access 4th and 5th columns of the data frame and
store their first seven elements in "conc_Qn1" and "uptake_Qn1",
respectively

# Exercise 02: Show the scatter plot for conc_Qn1 and uptake_Qn1
```