Data Science

SIT22009: Introduction to Data Science

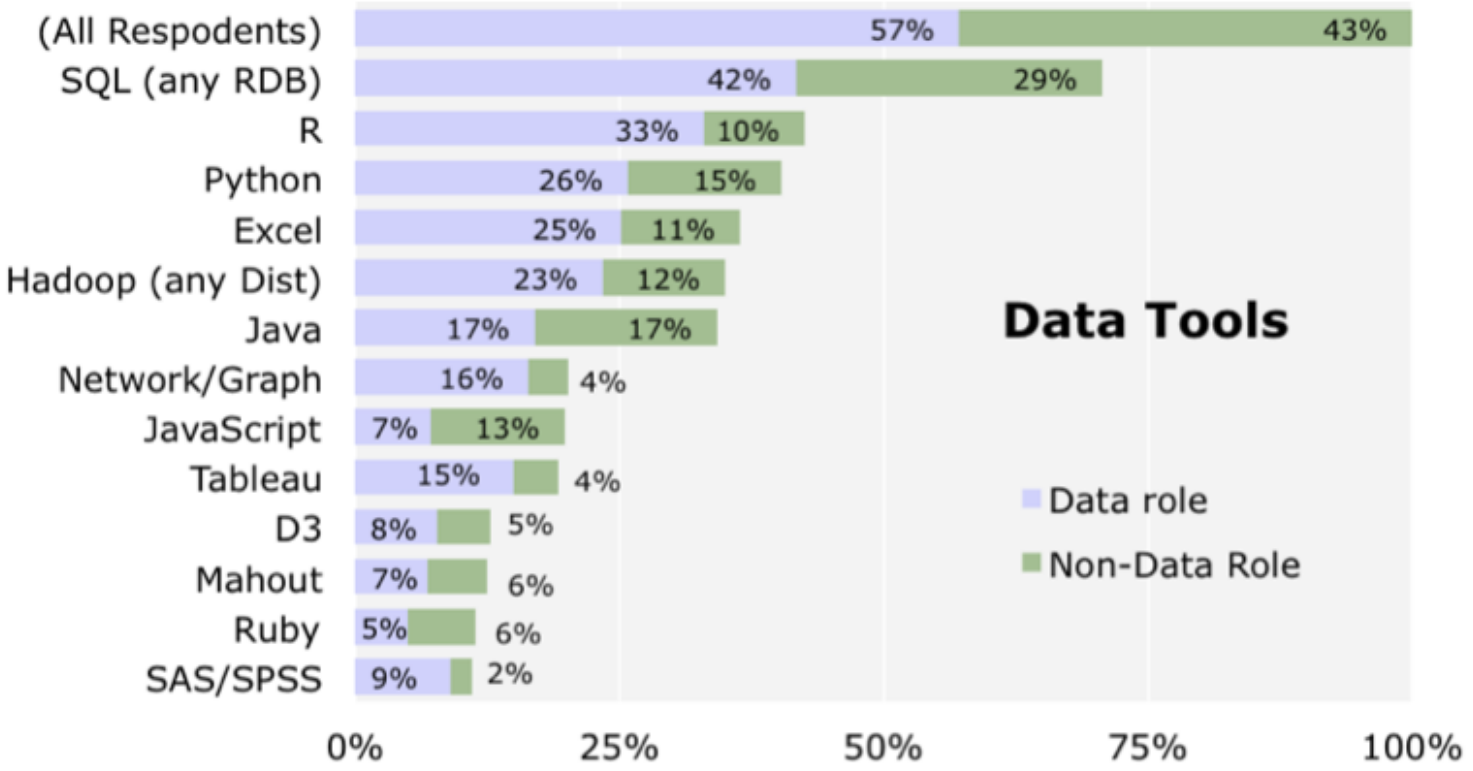
Presented by Hyebong Choi

Contents

- Course Description
- Toy examples

- 3~4 weeks: Getting familiar with **R**
- 1 weeks: Basic concept of **Data Science**
  - Understanding typical process of Big data project and the role of data scientist
- 3 weeks: Data Preparation
  - Data Loading, Data Verification, Data Cleaning
- Midterm test
- 2 weeks: Basic concept of modeling and evaluation
- 5 weeks: A few modeling method with single variable and multivariable
- Final test

R



- Most common tools for data scientist other than DBMS
- Cover wide range of data scientist – Data engineer, Statistician, Data main expert, … rather than just computer engineer
- Providing thousands of ready-to-use powerful packages for data scientist
- Well documented

http://blog.revolutionanalytics.com/2014/01/in-data-scientist-survey-r-is-the-most-used-tool-other-than-databases.html

Installation

- R
  - Download from https://www.r-project.org/
- R studio
  - Integrated Developing Environment **IDE** for R
  - Download from https://www.rstudio.com/
- R studio server
  - R working environment available from remote server
  - Accounts are already setup for all enrolled students

R Installation

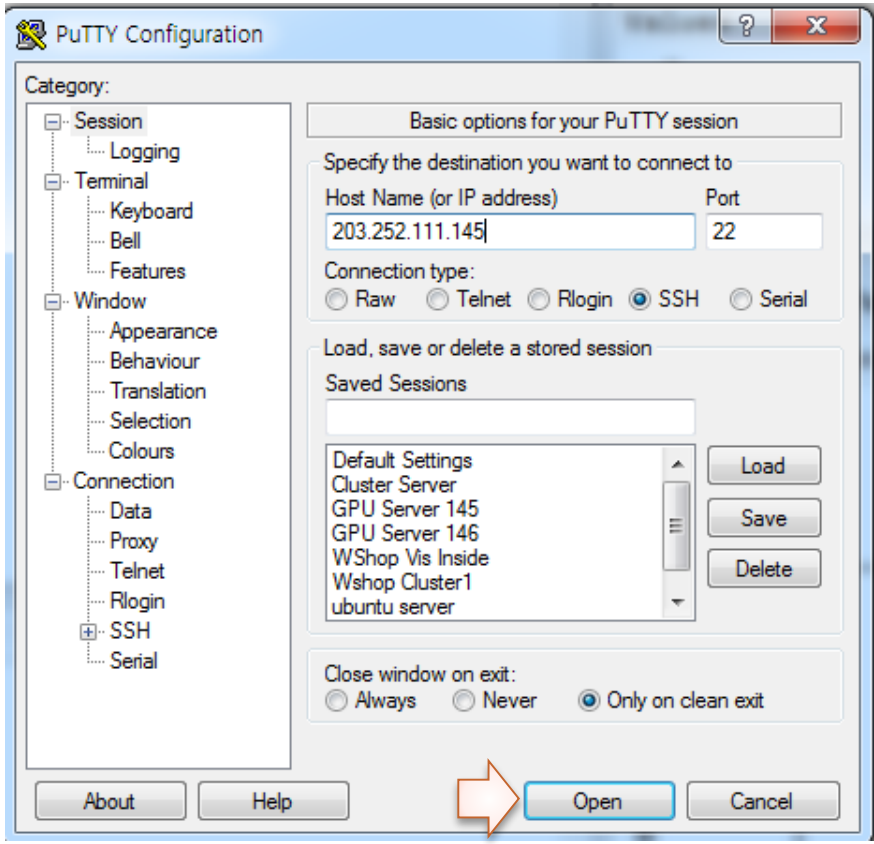
- on Windows
  - https://youtu.be/MFfRQuQKGYg
- on Mac
  - https://youtu.be/Ywj6yNfc5nM
- on Linux
  - (Ubuntu) https://youtu.be/GsuA5ugYqyw

R Studio Server

- R Studio Server is available for this class
  - http://220.149.111.120
  - ID : st[student_ID] e.g. st21012345
  - password : [student_ID]
- pros
  - no installation needed
  - no dedicated machine needed
- cons
  - no access from outside campus
  - cannot guarantee to support all students

R Studio Server

- If you want to change your password…
  - download “putty” or other ssh client
  - access 220.149.111.120: (port 22)
  - access with your ID/passwd
  - “passwd” is the command to change password



Toy examples

```
> print("Hello, Welcome to Data Science")
[1] "Hello, Welcome to Data Science"
```

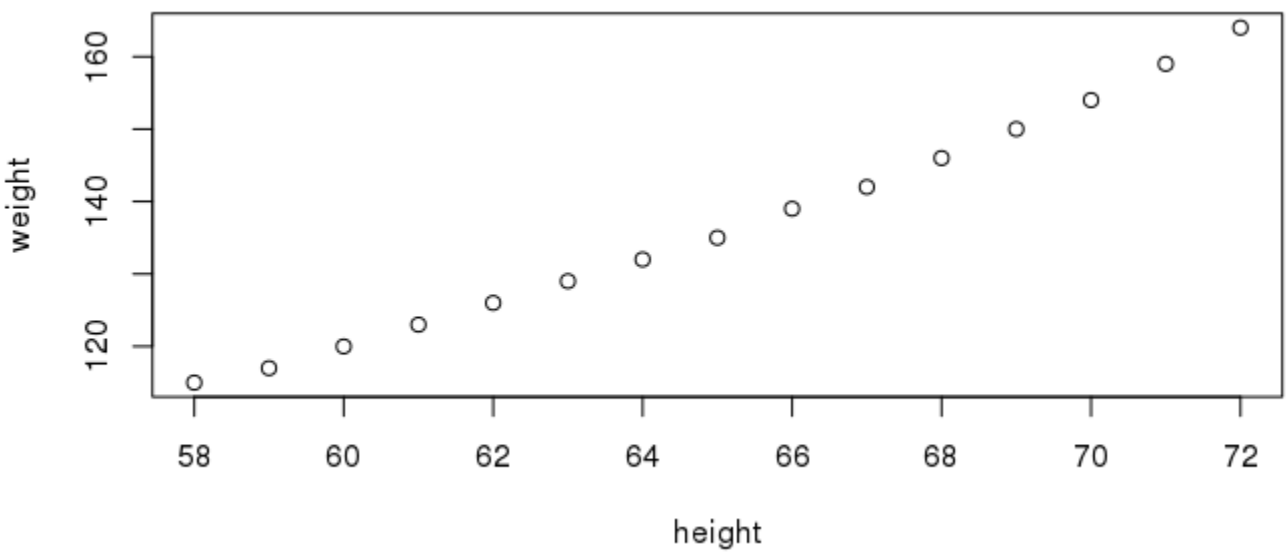
Toy examples

```
> seq(1:100)
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24
 [25]  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
 [49]  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96
 [97]  97  98  99 100
```

Toy examples

```
library(help="datasets")
data(women)
plot(women)
summary(women)
```

	height	weight
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142
11	68	146




```
     height          weight
 Min.   :58.0   Min.   :115.0
 1st Qu.:61.5   1st Qu.:124.5
 Median :65.0   Median :135.0
 Mean   :65.0   Mean   :136.7
 3rd Qu.:68.5   3rd Qu.:148.0
 Max.   :72.0   Max.   :164.0
```

References

- Practical Data Science with R, by Nina Zumel and John Mount
- R을 이용한 데이터 분석 실무, 서민구, 길벗