

---

# Review of the course “R for Data Science” Part 01(Talk 01~04)

By Haoran Nie @ HUST Life ST

Partially translated by Rui Zhu @ HUST Life ST

双语版

This work is licensed under CC BY-NC-SA 4.0

## 目录

Review of the course “R for Data Science” Part 01(Talk 01~ 04)	I
Multi-omics data analysis and visualisation, #1	1
Install R . . . . .	1
Rstudio . . . . .	1
R language basics, part 1	2
基础数据类型 . . . . .	2
数字 . . . . .	2
逻辑符号 . . . . .	2
字符串 . . . . .	2
简单数据类型 . . . . .	3
数据类型之间的转换 . . . . .	3
一些特殊值 . . . . .	4
Vectors and Matrix . . . . .	4
矩阵由函数 <code>matrix()</code> 定义, 比如: . . . . .	5
加减乘除逻辑运算老一套 . . . . .	5
通过 Console window 管理变量 . . . . .	5
vector 算术 <b>vectorisation</b> : R 最重要的一个概念 . . . . .	6
matrix 算术 . . . . .	6
更多 matrix 相关函数 . . . . .	6
The hierarchy of R's vector types . . . . .	8

<b>R language basics, part 2</b>	<b>9</b>
data.frame . . . . .	9
What is a data.frame? . . . . .	9
Usage of head() and tail() . . . . .	9
Structure of data.frame & tibble . . . . .	9
Make a new data.frame . . . . .	10
How to add row(s)/col(s) to an existing data.frame . . . . .	10
tibble . . . . .	11
Make new tibble . . . . .	11
tibble 元素替换 . . . . .	14
Manipulate the tibble . . . . .	14
tibble to data.frame . . . . .	14
Differences between tibble and data.frame . . . . .	14
Tibble 按顺序计算列 . . . . .	14
data.frame causes trouble when fetching subset operations . . . . .	15
tibble allows controlled data type conversion . . . . .	15
Recycling . . . . .	15
data.frame will do partial matching, while tibble will <b>NEVER</b> do it. . . . .	16
Advanced tips for using data.frame and tibble . . . . .	16
attach() and detach() . . . . .	17
with() . . . . .	17
within() . . . . .	18
File IO . . . . .	19
Read from files . . . . .	19
Write to files . . . . .	20
<b>R language basics, part 3: factor</b>	<b>23</b>
IO and working enviroment management . . . . .	23
Start a new RStudio session by creating a new project . . . . .	24
Working Space . . . . .	25
Variables in working space in RStudio . . . . .	25
Save and restore work space . . . . .	26

---

Save selected variables . . . . .	26
Close and (re)open a project . . . . .	26
Open a project . . . . .	27
Factors . . . . .	27
Play around with <code>levels()</code> . . . . .	28
Use factor to clean data . . . . .	30
factor 在做图中的应用（真正精髓） . . . . .	32
Using <code>factor</code> to vchange values . . . . .	33
Delete useless <code>levels</code> . . . . .	34
Advance usage . . . . .	37
一些勘误 . . . . .	38
vector 和 factor 有什么区别? . . . . .	38
data.frame 与 tibble 的区别 . . . . .	38

# Review of the course “R for Data Science” Part 02(Talk 05~08)

By Haoran Nie @ HUST Life ST

Partically translated by Rui Zhu @ HUST Life ST

双语版

This work is licensed under CC BY-NC-SA 4.0

## 目录

<b>Review of the course “R for Data Science” Part 02(Talk 05~ 08)</b>	<b>I</b>
<b>R for bioinformatics, data wrangler, part 1</b>	<b>1</b>
Pipe in R . . . . .	1
What is pipe in R? . . . . .	1
egs: . . . . .	2
Data Wrangler - dplyr . . . . .	4
What is dplyr? . . . . .	4
e.g. . . . .	5
查看 mouse.tibble 的内容 . . . . .	5
分析任务 . . . . .	6
用 dplyr 实现 . . . . .	6
检查运行结果 . . . . .	6
<b>R for bioinformatics, data wrangler, part 2</b>	<b>8</b>
tidyr . . . . .	8
Data Wrangler - tidyr . . . . .	8
宽数据的特点 . . . . .	8
优点: . . . . .	8
缺点: . . . . .	8

The usage of <code>tidyr</code> . . . . .	8
If you meet NA in the 1st example, you can do like this: . . . . .	9
More functions in <code>tidyr</code> : (See @ <a href="https://r4ds.hadley.nz/data-tidy.html">https://r4ds.hadley.nz/data-tidy.html</a> ) . . . . .	9
<code>tidyr::separate()</code> . . . . .	9
<code>tidyr::unite()</code> . . . . .	10
<b>R for bioinformatics, Strings and regular expression</b>	<b>11</b>
<b><code>stringr</code></b> . . . . .	11
Also notice other famous packages used to manipulating string: . . . . .	11
Usage of <code>writeLines()</code> (from official R Documentation) . . . . .	11
Difference between double quote(“ ”) and single quote(‘ ’) . . . . .	12
Some of the functions in the <code>stringi</code> package are similar in function to those that come with the system. . . . .	12
Some of the functions in the <code>stringr</code> package are similar in function to those that come with the system. . . . .	14
string length . . . . .	14
string combine . . . . .	15
string comparison . . . . .	15
(In the slide) Difference between <code>toupper()</code> , <code>tolower()</code> and <code>stri_reverse()</code>	16
Tricks . . . . .	17
Regex - Regular Expression . . . . .	17
tasks of regular expression . . . . .	20
useful tools . . . . .	21
<code>str_extract</code> vs. <code>str_match</code> . . . . .	21
<code>str_extract_all</code> 和 <code>str_match_all</code> . . . . .	21
<b>R for bioinformatics, data iteration &amp; parallel computing</b>	<b>23</b>
TOC . . . . .	23
Iteration Basics . . . . .	23
for loop , getting data ready . . . . .	23
apply functions . . . . .	24
Something about <code>tapply()</code> : . . . . .	25
Differences between <code>apply</code> in base R and the package <code>dplyr</code> : . . . . .	26

More on iteration: <b>purrr</b> package . . . . .	27
About <b>purrr</b> (from official website <a href="https://purrr.tidyverse.org">https://purrr.tidyverse.org</a> ) . . . . .	27
Detailed Usage . . . . .	28
Examples . . . . .	29
<b>map</b> 的高阶应用 . . . . .	31
<b>split</b> 与 <b>group_by</b> 的区别 . . . . .	32
(in the slide) Function <b>reduce()</b> and <b>accumulate()</b> . . . . .	32
Parallel Computing . . . . .	34
并行计算介绍 . . . . .	34
Related Packages . . . . .	34
Step-by-step Guidance . . . . .	34
(in the slide) Function <b>foreach()</b> . . . . .	36
Simple usage . . . . .	36
嵌套 (nested) <b>foreach</b> . . . . .	37

# Review of the course “R for Data Science” Part 03(Talk 09~12)

By Haoran Nie @ HUST Life ST

Partially translated by Rui Zhu @ HUST Life ST

双语版

This work is licensed under CC BY-NC-SA 4.0

## 目录

Review of the course “R for Data Science” Part 03(Talk 09~ 12)	I
R for bioinformatics, data visualisation	1
TOC . . . . .	1
Basic plot functions using R . . . . .	1
Dot plot 散点图 . . . . .	1
High-level and low-level . . . . .	3
图形相关参数（系统函数） . . . . .	4
调整 par() 参数前请备份 . . . . .	4
常用图形参数及调整: margin . . . . .	5
常用图形参数及调整: 多 panel . . . . .	5
重要概念: 图形设备 . . . . .	5
图形设备: cont. . . . .	5
常用图形设备: pdf() . . . . .	6
请尽量使用 pdf 作为文件输出格式 . . . . .	6
ggplot2 . . . . .	6
Some basic parameters of ggplot2 . . . . .	6
fill 与 colour 有什么区别??? . . . . .	8
Coordinate System 坐标系 . . . . .	9
faceting . . . . .	10
layered grammer (图层语法) 的成分 . . . . .	10

如何在一张图中画多个 panel? . . . . .	10
<code>cowplot::plot_grid</code> parameters . . . . .	11
用 <code>draw_plot</code> 调整 graph 的相对大小 . . . . .	12
use <code>gridExtra::grid.arrange</code> to arrange multiple graphs . . . . .	12
Different layouts . . . . .	13
在图中加入公式和统计信息 . . . . .	13
In talk09 . . . . .	14
equation 的其它写法 (更复杂难懂) . . . . .	14
希腊字符 . . . . .	15
<b>R for bioinformatics, data summarisation and statistics</b>	<b>17</b>
TOC . . . . .	17
Vector Summarization . . . . .	17
Describe Normal Distribution . . . . .	17
Functions to generate random normal distrubions . . . . .	18
Other regular distributions . . . . .	18
uniform distribution 的各种函数 . . . . .	18
other distributions, cont. . . . .	19
量化描述数据 . . . . .	19
量化描述函数 . . . . .	19
<code>quantile</code> and <code>summary</code> . . . . .	19
<code>table</code> 函数 . . . . .	20
<code>count</code> in <code>dplyr</code> . . . . .	20
<code>ntile</code> 函数的参数 . . . . .	20
<code>cut</code> 函数 . . . . .	20
Statistics . . . . .	20
Parametric tests . . . . .	20
how to detect outlier ?? . . . . .	20
Non-parametric Comparison . . . . .	26



<b>Linear and nonlinear regression</b>	<b>28</b>
TOC . . . . .	28
Linear Regression . . . . .	28
Fitting a Linear Regression Model: . . . . .	28
glm vs. lm . . . . .	29
glm 还可用于其它类型数据的分析 . . . . .	29
glm 的 Poisson regression (family=poisson) . . . . .	29
Other Useful Functions for Linear Regression Analysis: . . . . .	30
Nonlinear Regression . . . . .	32
Fitting a Nonlinear Regression Model: . . . . .	32
Other Useful Functions for Nonlinear Regression Analysis: . . . . .	32
Modeling and Prediction . . . . .	34
Modeling and Prediction Steps: . . . . .	35
<b>K-fold &amp; X times cross-validation</b> . . . . .	<b>36</b>
K-fold Cross-Validation: . . . . .	36
X times Cross-Validation: . . . . .	36
Implementation in R: . . . . .	37
External Validation . . . . .	38
Steps for External Validation: . . . . .	38
Implementation in R: . . . . .	38
<b>Machine learning basics</b>	<b>40</b>
TOC . . . . .	40
机器学习可分为以下几类 . . . . .	40
1. 回归算法 . . . . .	40
2. 基于实例的算法 . . . . .	41
3. 决策树学习 . . . . .	41
4. 贝叶斯方法 . . . . .	41
5. 基于核的算法 . . . . .	42
6. 聚类算法 . . . . .	42
7. 降低维度算法 . . . . .	42
8. 关联规则学习 . . . . .	42

9. 集成算法 . . . . .	43
10. 人工神经网络 . . . . .	43
section 3: 随机森林 . . . . .	43
随机森林 – Random forest . . . . .	43
决策树 - decision tree . . . . .	43
Steps to Implement Random Forest in R: . . . . .	44
Example - Random Forest for Regression: . . . . .	45
Feature Selection . . . . .	45
Feature Selection Techniques in R: . . . . .	45