

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

By Haoran Nie @ HUST Life ST

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

Contents

| | |
|---|---|
| Review of the course “R for Data Science” Part 01(Talk 01~ 04) | I |
| Multi-omics data analysis and visualisation, #1 | 1 |
| R language basics, part 1 | 1 |
| Fundamental Data Type | 1 |
| Simple Data Types | 1 |
| Conversion between data types | 1 |
| Some special values in matrices | 2 |
| Vectors and Arrays | 2 |
| Vector manipulation | 2 |
| The hierarchy of R's vector types | 3 |
| R language basics, part 2 | 4 |
| data.frame | 4 |
| What is a data.frame? | 4 |
| Usage of head() and tail() | 4 |
| Components of data.frame and common functions | 4 |
| Structure of data.frame & tibble | 5 |
| Make a new data.frame | 5 |
| How to add row(s)/col(s) to an existing data.frame | 5 |
| tibble | 6 |
| Make new tibble | 6 |
| Manipulate the tibble | 7 |
| tibble to data.frame | 7 |
| Differences between tibble and data.frame | 7 |
| Tibble evaluates columns sequentially | 7 |
| data.frame causes trouble when fetching subset operations | 7 |
| tibble allows controlled data type conversion | 7 |
| Recycling | 7 |
| data.frame will do partial matching, while tibble will NEVER do it. | 8 |
| Advanced tips for using data.frame and tibble | 8 |
| attach() and detach() | 9 |

| | |
|---|-----------|
| with() | 9 |
| within() | 9 |
| File IO | 11 |
| Read from files | 11 |
| Write to files | 12 |
| R language basics, part 3: factor | 13 |
| IO and working enviroment management | 13 |
| Start a new RStudio session by creating a new project | 14 |
| Working Space | 15 |
| Variables in working space in RStudio | 15 |
| Save and restore work space | 15 |
| Save selected variables | 15 |
| Close and (re)open a project | 16 |
| Open a project | 16 |
| Factors | 17 |
| Play around with levels() | 17 |
| Use factor to clean data | 17 |
| Usage of factors in drawing plots | 18 |
| Using factor to vchange values | 20 |
| Delete useless levels | 21 |
| Advance usage | 23 |

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

By Haoran Nie @ HUST Life ST

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

To reduce the size, all the codes listed will **NOT** include the output as picture.

Contents

| | |
|---|----------|
| Review of the course “R for Data Science” Part 02(Talk 05~ 08) | I |
| R for bioinformatics, data wrangler, part 1 | 1 |
| TOC | 1 |
| pipe | 1 |
| dplyr | 1 |
| tidyr, part 1 | 1 |
| Pipe in R | 1 |
| What is pipe in R? | 1 |
| Other kinds of pipe | 1 |
| Data Wrangler - dplyr | 2 |
| What is dplyr? | 2 |
| R for bioinformatics, data wrangler, part 2 | 2 |
| TOC | 3 |
| tidyr | 3 |
| Data Wrangler - tidyr | 3 |
| The usage of tidyr | 3 |
| What’s the difference between wide and long data? | 3 |
| If you meet NA in the 1st example, you can do like this: | 4 |
| More functions in tidyr: (See @ https://r4ds.hadley.nz/data-tidy.html) | 4 |
| R for bioinformatics, Strings and regular expression | 5 |
| TOC | 5 |
| stringr | 5 |
| Basics | 5 |
| Also notice other famous packages used to manipulating string: | 5 |
| Usage of <code>writeLines()</code> (from official R Documentation) | 6 |
| Difference between double quote(“”) and single quote(‘’) | 7 |

| | |
|---|-----------|
| Some of the functions in the stringr package are similar in function to those that come with the system. | 7 |
| Some of the functions in the stringi package are similar in function to those that come with the system. | 8 |
| (In the slide) Difference between toupper() , tolower() and stri_reverse() | 9 |
| Tricks | 10 |
| Regex - Regular Expression | 10 |
| R for bioinformatics, data iteration & parallel computing | 11 |
| TOC | 11 |
| Iteration Basics | 12 |
| for loop , getting data ready | 12 |
| apply functions | 12 |
| Something about tapply() : | 13 |
| Differences between apply in base R and the package dplyr : | 14 |
| More on iteration: purrr package | 15 |
| About purrr (from official website https://purrr.tidyverse.org) | 15 |
| Detailed Usage | 15 |
| Examples | 16 |
| (in the slide) Function reduce() and accumulate() | 18 |
| Parallel Computing | 19 |
| Related Packages | 19 |
| Step-by-step Guidance | 19 |
| (in the slide) Function foreach() | 20 |
| Nested foreach | 21 |

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

By Haoran Nie @ HUST Life ST

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

Contents

| | |
|---|----------|
| 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 | 1 |
| High-level and low-level | 3 |
| Graphics-related parameters (system functions) | 4 |
| ggplot2 | 4 |
| Some basic parameters of ggplot2 | 4 |
| Coordinate System | 6 |
| faceting | 6 |
| Different layouts | 7 |
| Formulas | 7 |
| R for bioinformatics, data summarisation and statistics | 8 |
| TOC | 8 |
| Vector Summarization | 8 |
| Describe Normal Distribution | 8 |
| Functions to generate random normal distrubions | 9 |
| Other regular distributions | 9 |
| Quantitative descriptive data | 9 |
| Quantitative descriptive function | 9 |
| Statistics | 10 |
| Parametric tests | 10 |
| Non-parametric Comparison | 14 |

| | |
|---|-----------|
| Linear and nonlinear regression | 15 |
| TOC | 15 |
| Linear Regression | 15 |
| Fitting a Linear Regression Model: | 16 |
| Other Useful Functions for Linear Regression Analysis: | 16 |
| Nonlinear Regression | 18 |
| Fitting a Nonlinear Regression Model: | 18 |
| Other Useful Functions for Nonlinear Regression Analysis: | 18 |
| Modeling and Prediction | 20 |
| Modeling and Prediction Steps: | 20 |
| K-fold & X times cross-validation | 21 |
| K-fold Cross-Validation: | 21 |
| X times Cross-Validation: | 21 |
| Implementation in R: | 22 |
| External Validation | 23 |
| Steps for External Validation: | 23 |
| Importance of External Validation: | 23 |
| Implementation in R: | 23 |
| Machine learning basics | 24 |
| TOC | 24 |
| Machine Learning Algorithms Generalization | 24 |
| Random Forest in Machine Learning using R | 25 |
| Steps to Implement Random Forest in R: | 25 |
| Example - Random Forest for Regression: | 26 |
| Feature Selection | 26 |
| Feature Selection Techniques in R: | 26 |
| Implementation Considerations: | 27 |