BITTIGER

DS 501 Data scientist express bootcamp

Week 1 [Ella]

版权声明

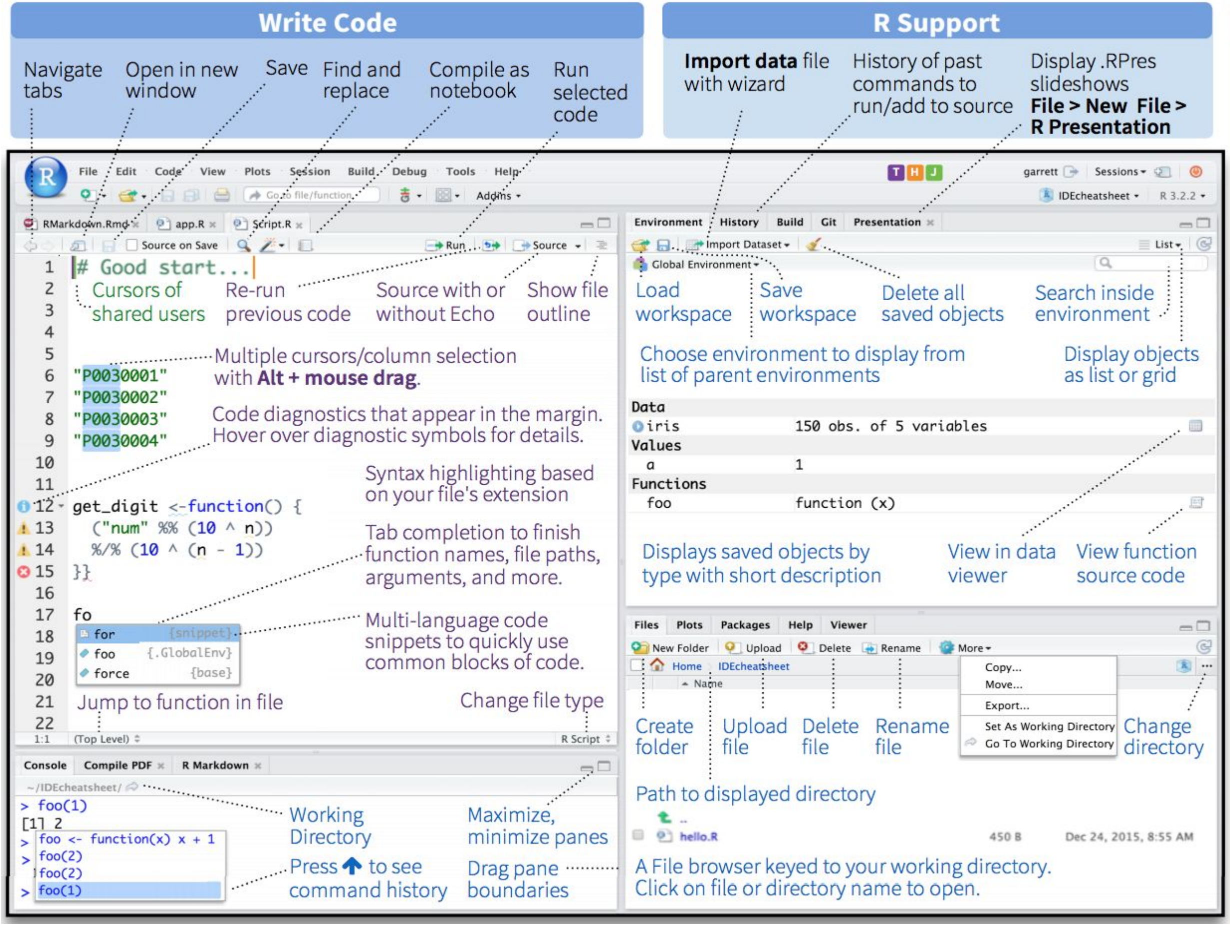
Copyright Policy

Summary

- Get familiar with R studio
- Common data structures in R
- Data exploration
  - Numeric variables
  - Categorical variables
  - Numeric variable with numerical response
  - Categorical variable with numerical response
  - Numeric variable with categorical response
  - Categorical variable with categorical response

R Studio

R material list



Write Code

Navigate tabs

Open in new window

Save

Find and replace

Compile as notebook

Run selected code

R Support

Import data file with wizard

History of past commands to run/add to source

Display .RPres slideshows **File > New File > R Presentation**

```
# Good start...
"P0030001"
"P0030002"
"P0030003"
"P0030004"
get_digit <-function() {
  ("num" %% (10 ^ n))
  %/% (10 ^ (n - 1))
}}
fo
```

Cursors of shared users

Re-run previous code

Source with or without Echo

Show file outline

Multiple cursors/column selection with **Alt + mouse drag**.

Code diagnostics that appear in the margin. Hover over diagnostic symbols for details.

Syntax highlighting based on your file's extension

Tab completion to finish function names, file paths, arguments, and more.

Multi-language code snippets to quickly use common blocks of code.

Jump to function in file

Change file type

```
~/IDEcheatsheet/
> foo(1)
[1] 2
> foo <- function(x) x + 1
> foo(2)
> foo(2)
> foo(1)
```

Working Directory

Maximize, minimize panes

Press ↑ to see command history

Drag pane boundaries

Load workspace

Save workspace

Delete all saved objects

Search inside environment

Choose environment to display from list of parent environments

Display objects as list or grid

```
Data
iris        150 obs. of 5 variables
Values
a           1
Functions
foo         function (x)
```

Displays saved objects by type with short description

View in data viewer

View function source code

Create folder

Upload file

Delete file

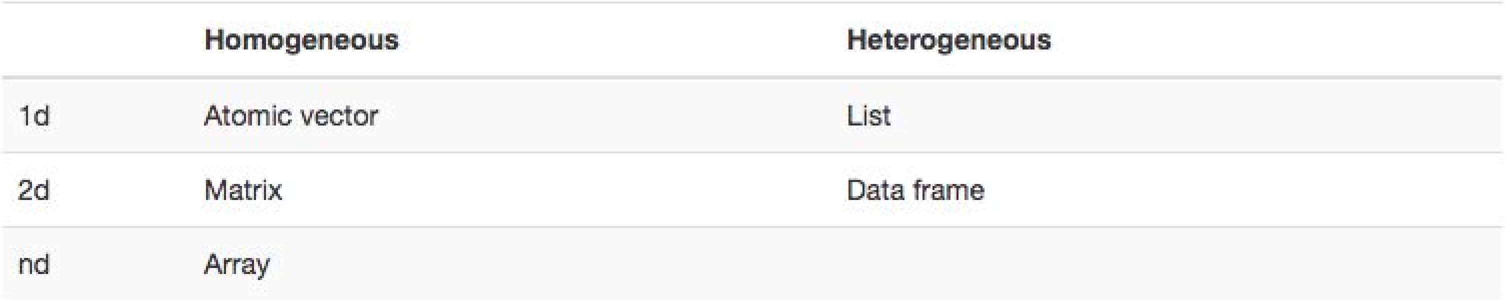
Rename file

Change directory

Path to displayed directory

A File browser keyed to your working directory. Click on file or directory name to open.

Data structure



	Homogeneous	Heterogeneous
1d	Atomic vector	List
2d	Matrix	Data frame
nd	Array	

- Dimensions
- homogeneous: all columns must be of the same type
- heterogeneous: columns can be of different types

Load data in R

- Download data
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- Prepare workspace
  - Getwd(), setwd()
- Load in data
  - .txt file: read.table()
  - **.csv file: read.csv()**
  - Excel file: readWorksheetFromFile from **library**
  - Json file, XML file, HTML table,
  - Other stats software files
  - Relational database and non-relational database

Getting to know the data

- Metadata
  - summary(), str(), dim(), head(), colnames()/rownames(), length(), unique()
- Categorical variable
  - table(), barplot(), pie()
- Continuous variable or ordinal categorical
  - **by**(), apply()...
  - mean(), median(), sd(), quantiles(), density(), boxplot()
  - plot(), lines() to visualize results

How to understand data?

- Data exploration
    - Numeric variables
        - mean(), sd(), quantile(), boxplot(), density()...
    - Categorical variables
        - Sort(), table(), barplot()...
    - Numeric variable with numerical response
        - Cor, library(corrplot)...
    - Categorical variable with numerical response
        - boxplot(), by(), apply(), library(lattice)...
    - Numeric variable with categorical response
        - boxplot(), library(tabplot)...
    - Categorical variable with categorical response

Summary

- Get familiar with R studio
- Common data structures in R
- Data exploration