



BITTIGER

DS 501 Data scientist express bootcamp

Week 1 Practice



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

The screenshot shows the RStudio interface with the following components and annotations:

- Source Editor:** Contains R code with annotations for "Cursors of shared users", "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", and "Change file type".
- Environment Pane:** Shows the "Global Environment" with a list of objects. Annotations include "Load workspace", "Save workspace", "Delete all saved objects", "Search inside environment", "Choose environment to display from list of parent environments", and "Display objects as list or grid".
- Data Viewer:** Displays the "iris" dataset with 150 observations and 5 variables. Annotations include "Displays saved objects by type with short description", "View in data viewer", and "View function source code".
- Files Pane:** Shows the file browser for the working directory. Annotations include "Create folder", "Upload file", "Delete file", "Rename file", "Path to displayed directory", "A File browser keyed to your working directory. Click on file or directory name to open.", and "Change directory".
- Console:** Shows the command history. Annotations include "Working Directory", "Maximize, minimize panes", "Press ↑ to see command history", and "Drag pane boundaries".



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



HW

- Examine relationship between each feature and response (int_rate)
 - Pick 5 categorical and 5 numeric features, which you think are the most predictive with reasoning
- How to generate potential useful features from existing data?
- What other questions you can explore using this data set?