



# BITTIGER

DS 501 Data scientist express bootcamp

*Week 1 [Ella]*

## 版权声明

所有太阁官方网站以及在第三方平台课程中所产生的课程内容, 如文本, 图形, 徽标, 按钮图标, 图像, 音频剪辑, 视频剪辑, 直播流, 数字下载, 数据编辑和软件均属于太阁所有并受版权法保护。

对于任何尝试散播或转售BitTiger的所属资料的行为, 太阁将采取适当的法律行动。



有关详情, 请参阅

<https://www.bittiger.io/termsfuse> <https://www.bittiger.io/termservice>

## Copyright Policy

All content included on the Site or third-party platforms as part of the class, such as text, graphics, logos, button icons, images, audio clips, video clips, live streams, digital downloads, data compilations, and software, is the property of BitTiger or its content suppliers and protected by copyright laws.

Any attempt to redistribute or resell BitTiger content will result in the appropriate legal action being taken.



We thank you in advance for respecting our copyrighted content.

For more info:

see <https://www.bittiger.io/termsfuse>

and <https://www.bittiger.io/termservice>



## 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**

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 Panel:** 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 Panel:** 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





# 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