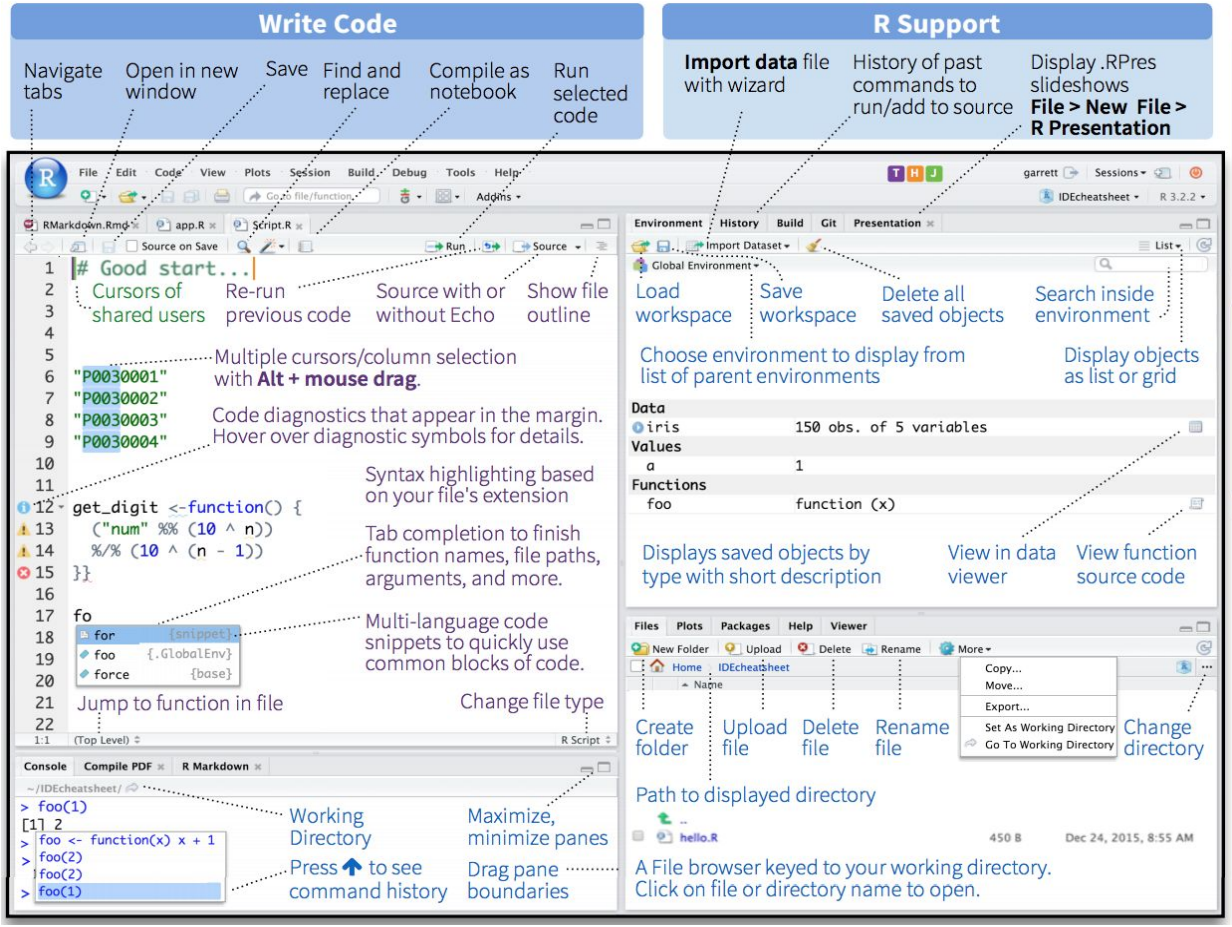
BITTIGER

DS 501 Data scientist express bootcamp

Week 1 Practice

R Studio



R material list

Data structure

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

- Dimensions
- homogeneous: all contents must be of the same type
- heterogeneous: the contents 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
  - Pick 5 categorial and 5 continuous features, which you think are the most predictive with reasoning
- How to generate potential useful features from existing features?
  - For example, total square in a house