Data Analysis and Visualization in R

REU Data Carpentry Workshop

June 27th-28th

Elizabeth McDaniel, University of Wisconsin-Madison

After this lesson, you will be able to:

- Use RStudio for writing R scripts
- Implement the basic building blocks of an R script
- Load your data into R
- Manipulate data within R
- Visualize data within R

Getting Started

- Understand the difference between an Rscript and RStudio
- Identify the different panels of an Rstudio window
- Use Rstudio to find help on R functions
- Describe how to troubleshoot problems with the wider R community in different settings

What is R?

- “R” refers to both the programming language itself, and the software platform to interpret scripts written in it

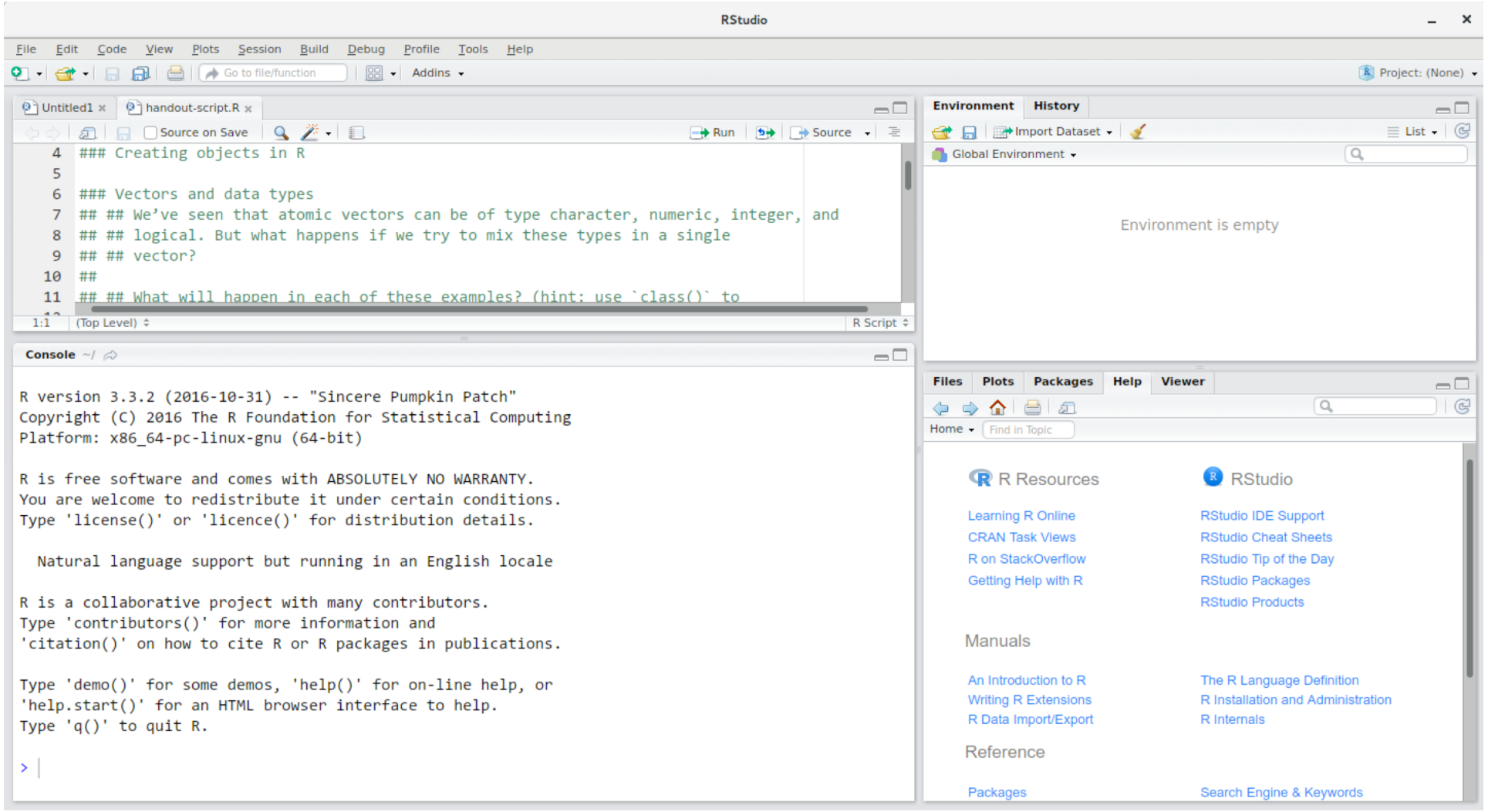
Why learn R?

- Less pointing and clicking involved
- Writing scripts makes analyses more reproducible than clicking a series of buttons
- Interdisciplinary (field-specific extensions or packages) for genomics, statistical simulations, etc.
- Takes data of all shapes and sizes
- Produce high-quality graphics without $$$
- Large, welcoming, open source community always developing new packages

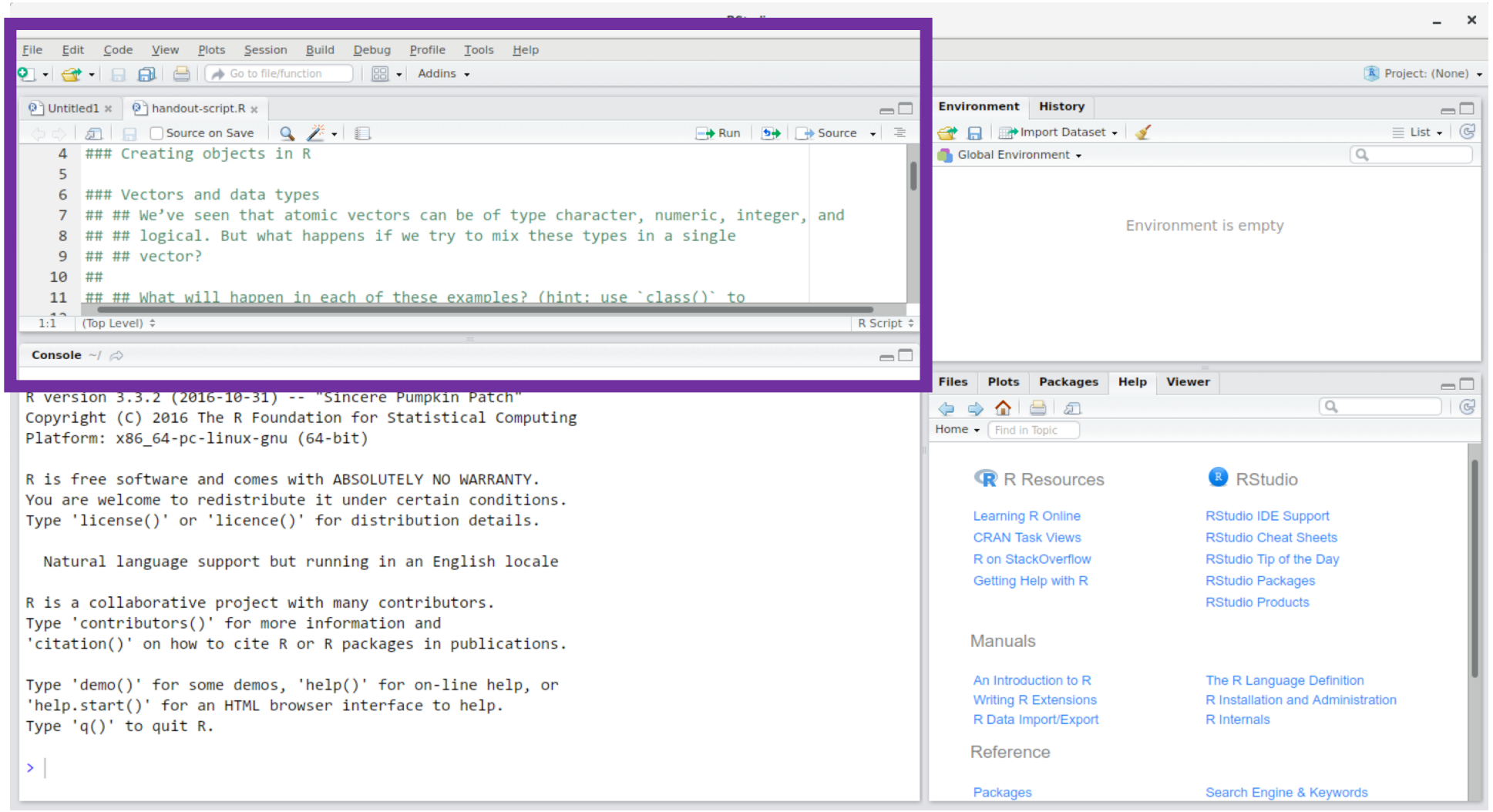
What is RStudio?

- Rstudio is an **I**ntegrated **D**evelopment **E**nvironment (IDE) for working with R
- Can write code, navigate files, view variables you created, and inspect plots within RStudio
- Lots of other IDEs that can suit a wide variety of programming languages

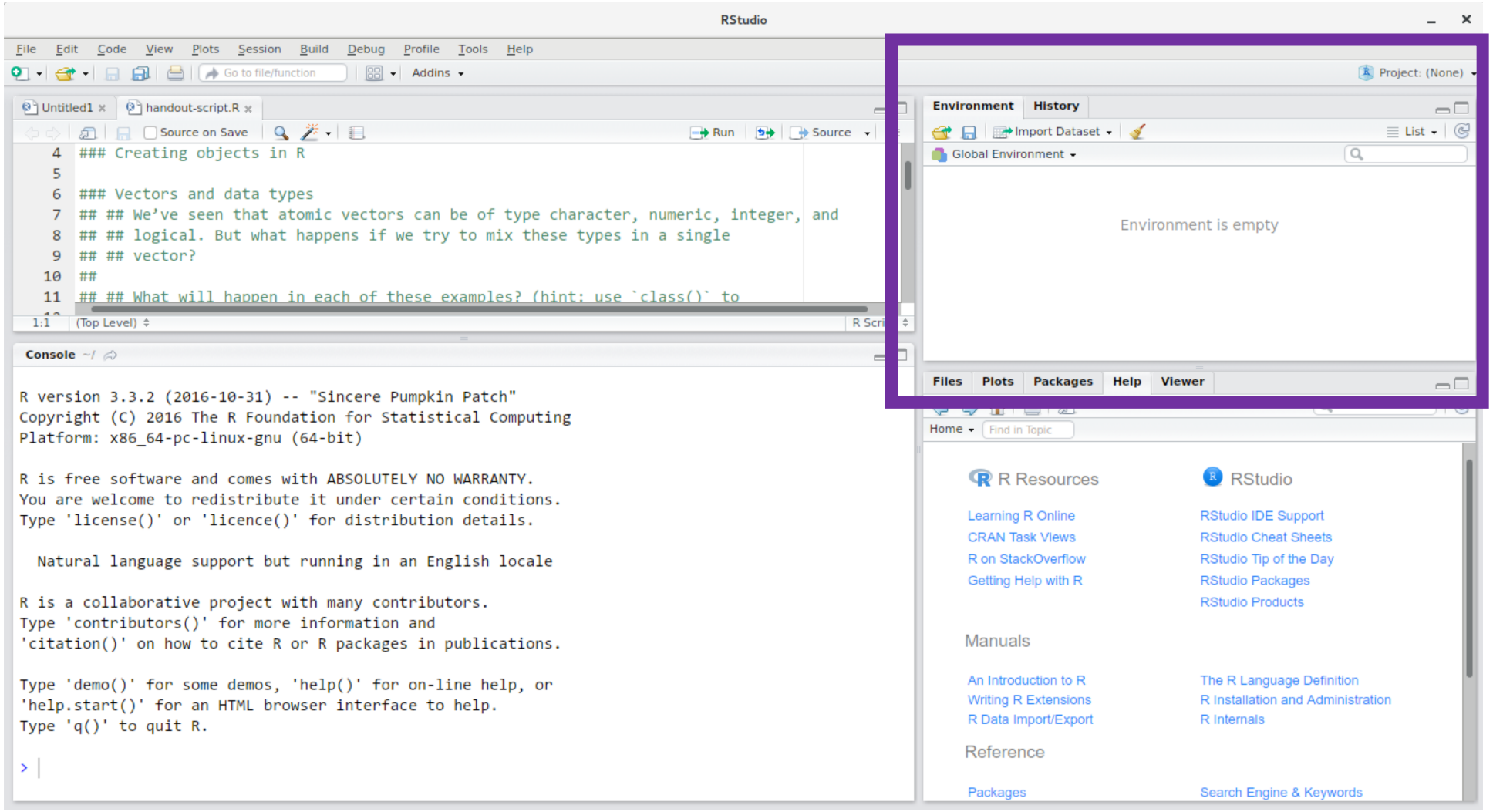
The RStudio Interface



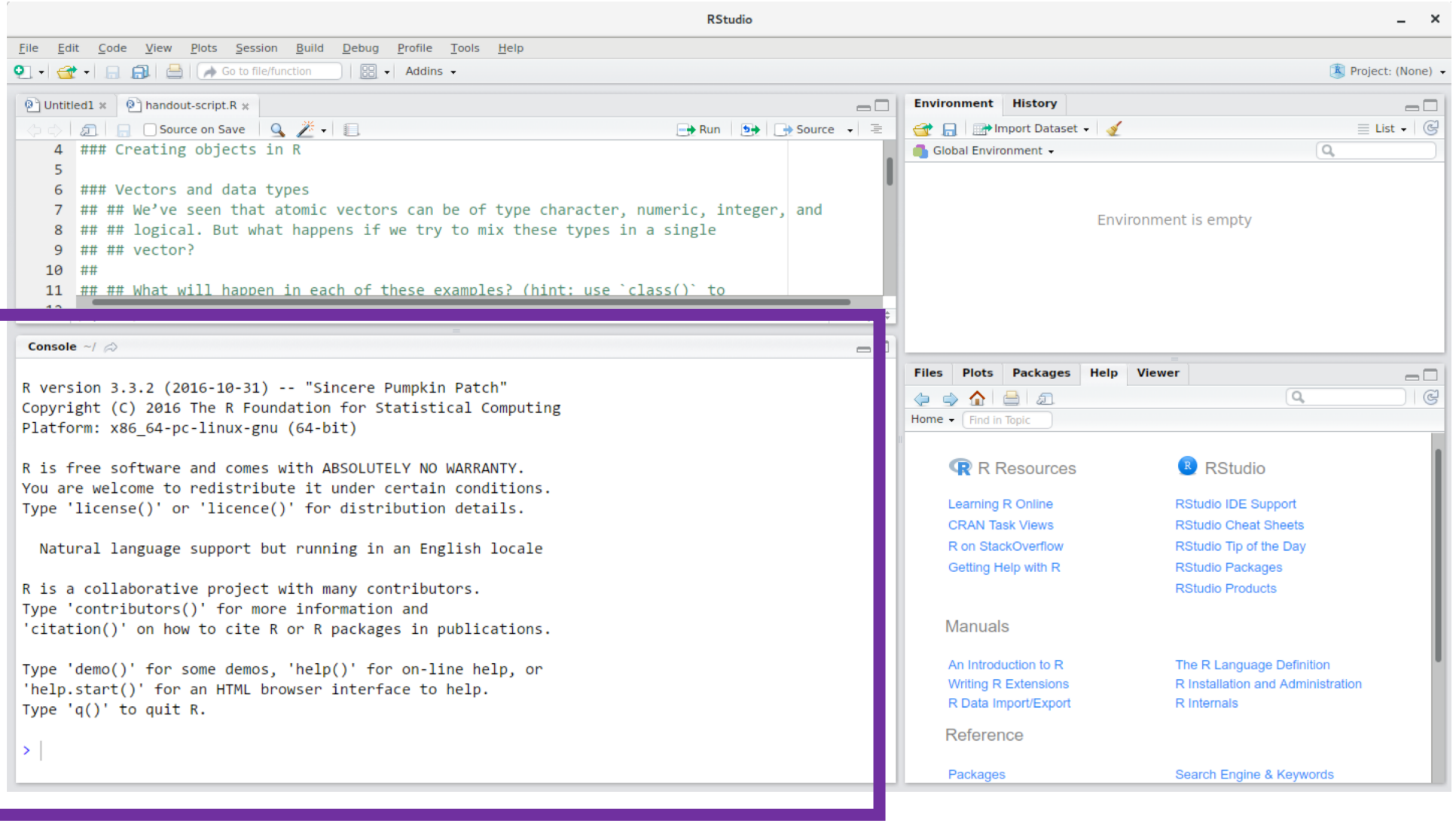
The RStudio Interface



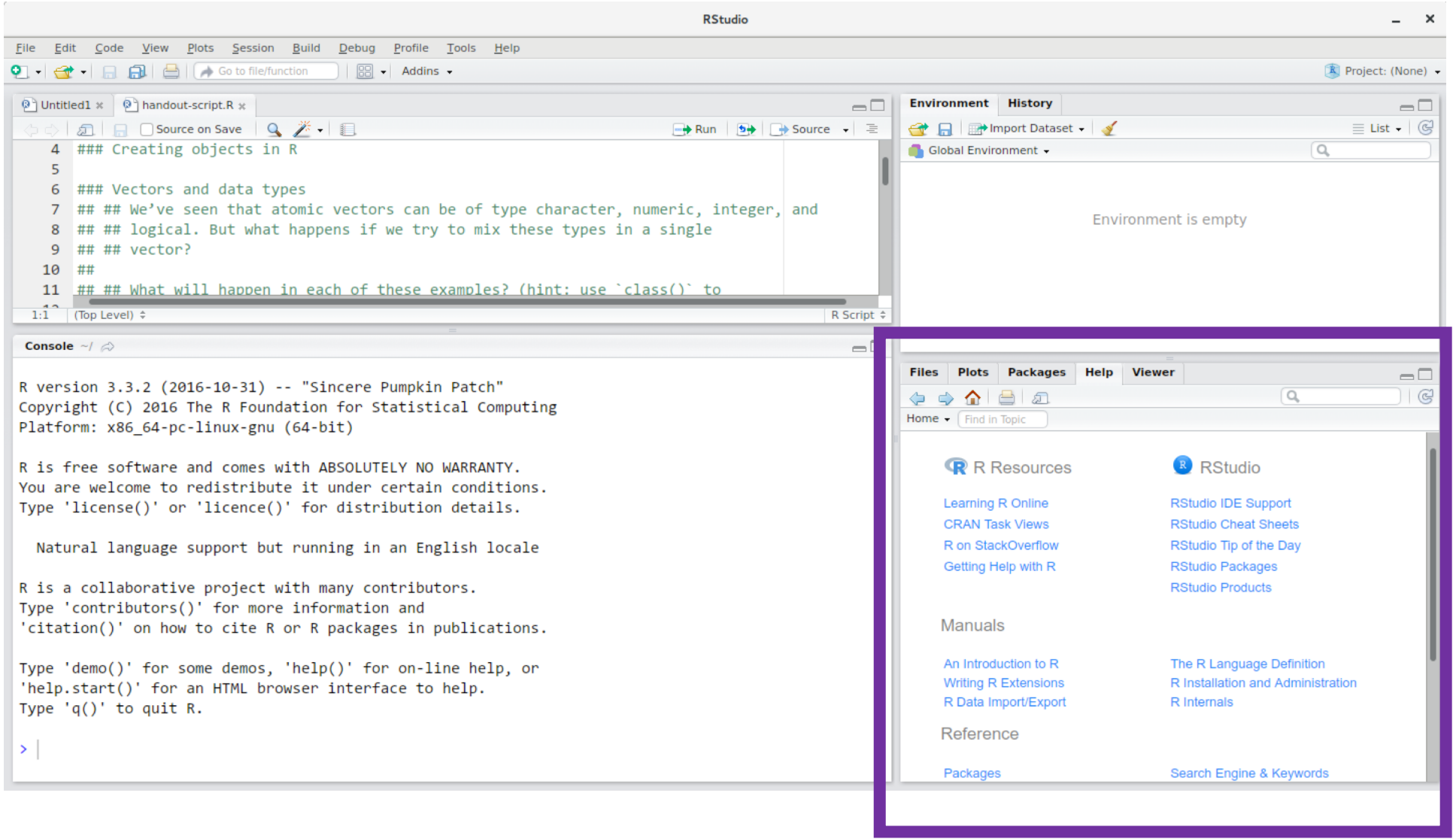
The RStudio Interface



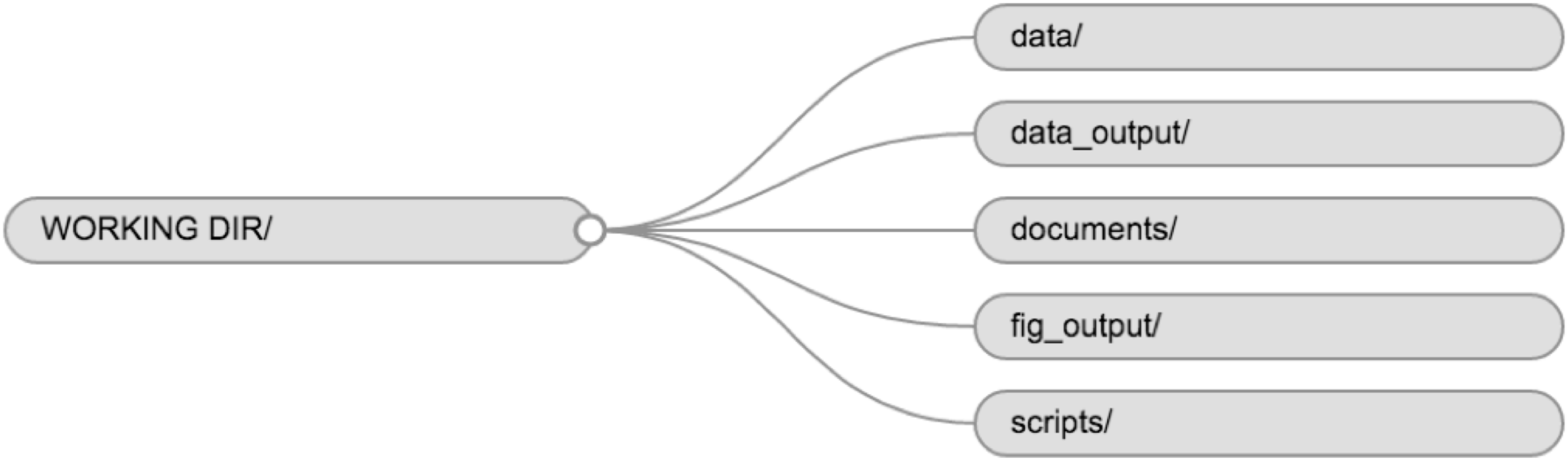
The RStudio Interface



The RStudio Interface



Your Working Directory



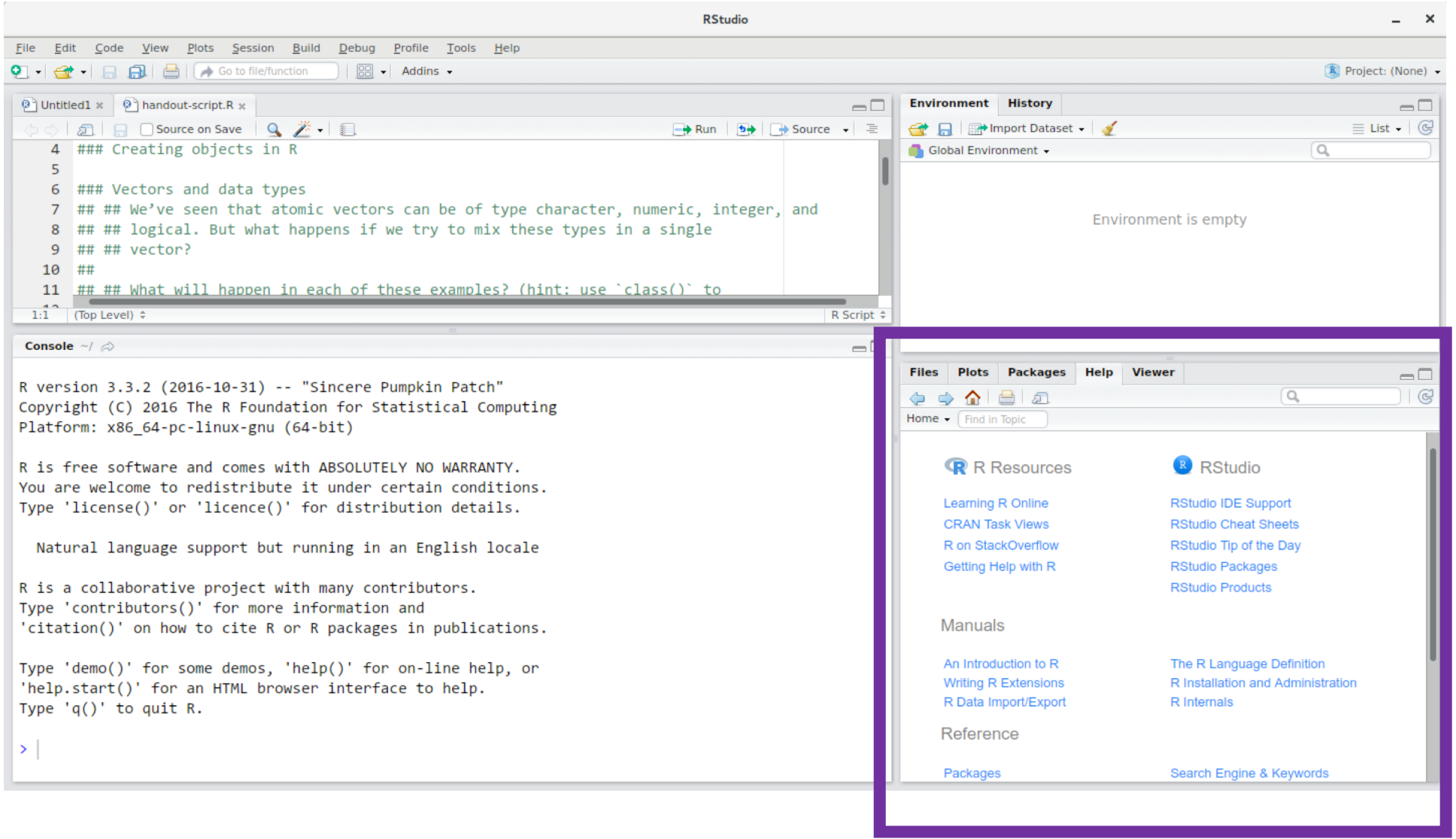
Working in RStudio

1. Start RStudio
2. File > New Project > New Directory
3. Enter a name for the new folder, which will be your **working directory**
4. Click Create Project

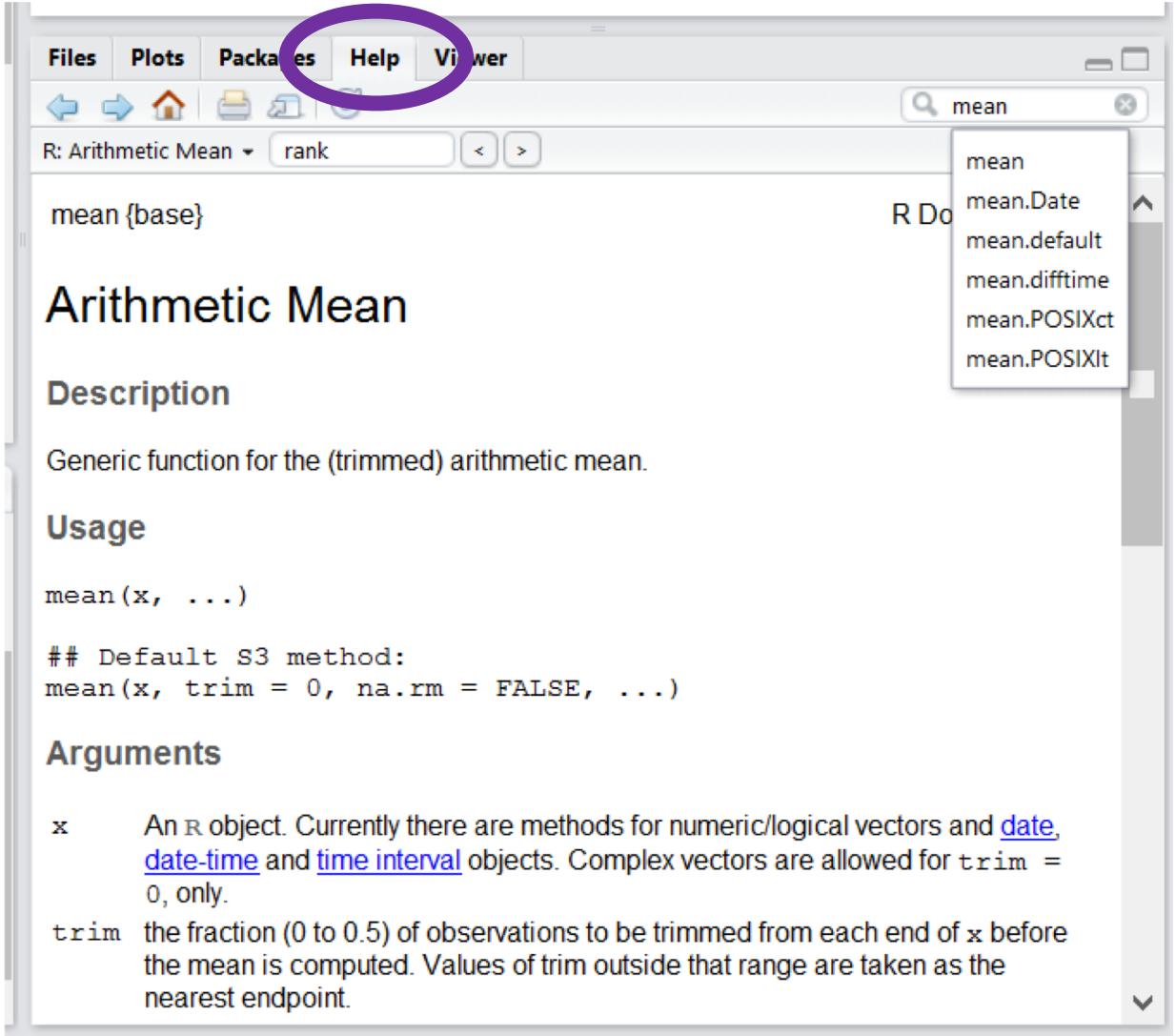
How do I find help?

Built-In RStudio help interface

The RStudio Interface

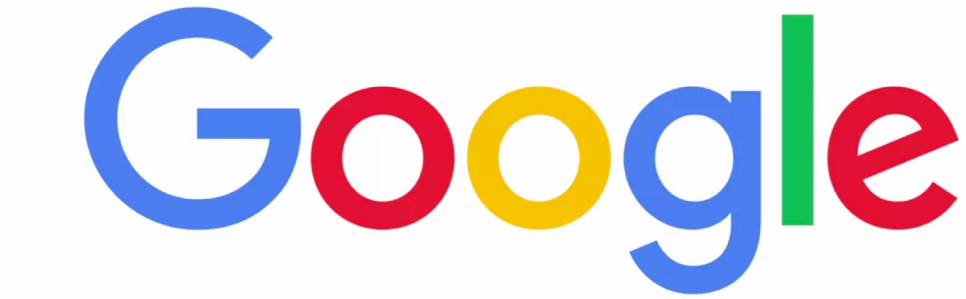


The RStudio Interface



How do I find help?

Google is a great place to start!



What information do I include when I ask for help?

- Make your code human-readable
- Include the output of your session info, which includes information about your platform, versions of R and packages you are using

✓ Getting Started

- ✓Understand the difference between an Rscript and RStudio
- ✓Identify the different panels of an Rstudio window
- ✓Use Rstudio to find help on R functions
- ✓Describe how to troubleshoot problems with the wider R community

Introduction to R

- Terminology:
  - object
  - assign
  - call
  - function
  - arguments
  - options

Introduction to R

- Assign values and names to objects in R
- Use comments to give future you/collaborators helpful information
- Solve arithmetic operations in R
- Call functions and use arguments to change default parameters
- Inspect vectors and manipulate their content
- Subset and extract values from vectors
- Analyze vectors with missing data

Objects

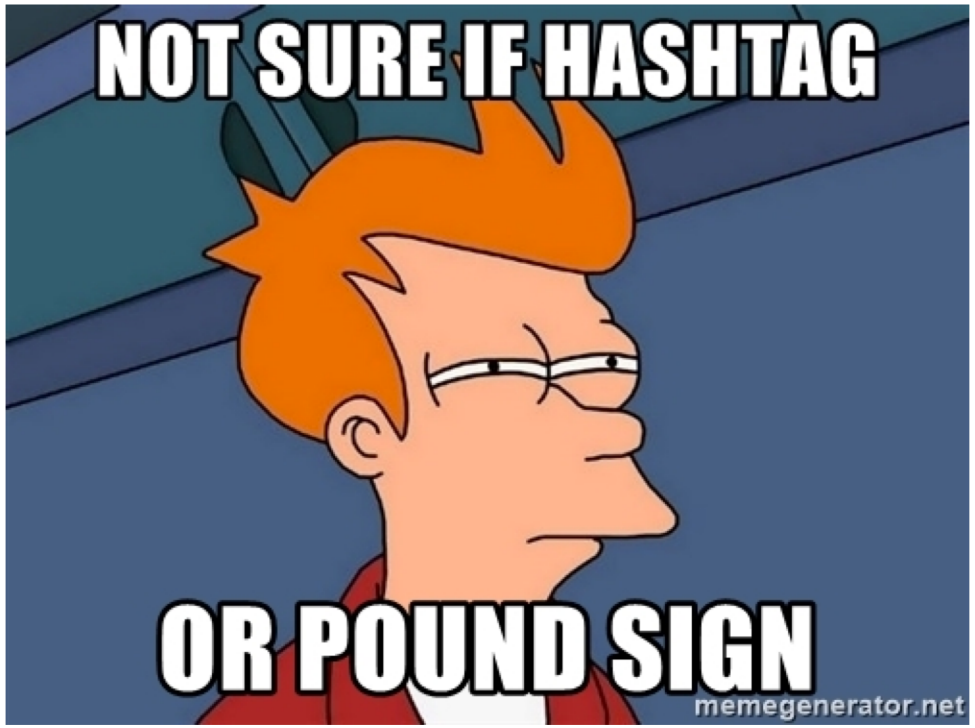
- Assign a value:
  - Give a name to the object
  - Use the <- assignment operator
- Can be any name (with no spaces!), but cannot start with a number. Names are case sensitive. Cannot use a fundamental function (if, else, for) as an object name

Comments

“Your closest collaborator is you six months ago,
but you don’t reply to emails.
- Mark Holder”
- Karl Broman

Comments

- Leave yourself useful notes about why you wrote lines of code a certain way (you won't remember…trust me)
- The comment character in R (and most programming languages) is the # symbol



Vector

- Most common data type in R
- Composed of a series of values, either numbers or characters
- Assign a series of values to a vector using the c() function, which **c**ombines the arguments to form a vector

Subsetting Vectors

- Extract one or several specific values from a vector
- Provide an indices with brackets

Conditional Subsetting

- TRUE or FALSE logical factors
- Combining multiple tests
- & = both conditions are satisfied
- | = one OR the other condition is satisfied
- %in% test if any elements of a search vector are found

Missing Data

- R was designed to analyze datasets, which may sometimes have missing data
- Missing data in vectors is represented as NA

Introduction to R

- Terminology:
  - object
  - assign
  - call
  - function
  - arguments
  - options

✓ Introduction to R

- ✓Assign values and names to objects in R
- ✓Use comments to give future you/collaborators helpful information
- ✓Solve arithmetic operations in R
- ✓Call functions and use arguments to change default parameters
- ✓Inspect vectors and manipulate their content
- ✓Subset and extract values from vectors
- ✓Analyze vectors with missing data

Loading Data into R

- Terminology:
  - Data frame
  - Factor
  - String

Loading Data into R

- Load external data from a .csv file into a data frame
- Describe what a data frame is
- Summarize the contents of a data frame
- Use indexing to subset specific parts of a data frame
- Describe a factor
- Reorder and rename factors

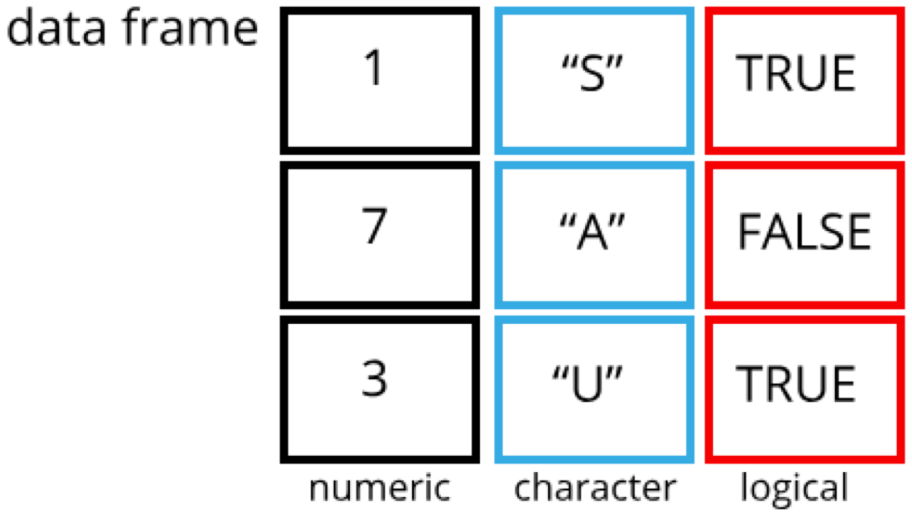
Survey Data

- Studying species repartition and weight of animals caught in plots in our study area

Column	Description
record_id	Unique id for the observation
month	month of observation
day	day of observation
year	year of observation
plot_id	ID of a particular plot
species_id	2-letter code
sex	sex of animal (“M”, “F”)
hindfoot_length	length of the hindfoot in mm
weight	weight of the animal in grams
genus	genus of animal
species	species of animal
taxon	e.g. Rodent, Reptile, Bird, Rabbit
plot_type	type of plot

Data Frames

- Data structures for tabular data, further used for statistics and plotting
- Representation of data in the format of a table, columns are vectors with the same length, and contains the same type of data



Indexing and Subsetting Data Frames

- Survey data frame contains rows and columns = 2 dimensions
- Extract information from the data frame, need to specify “coordinates”
- Row numbers come first, then the column number

Factors

- Factors represent categorical data
- Several of the columns of our surveys data frame contain integers
- However, columns such as genus, species, sex, plot_type… are categorical data
- Factors are stored as integers with labels, and can either be ordered or unordered
- Contain pre-defined set of values = levels

Loading Data into R

- Terminology:
  - Data frame
  - Factor
  - String

✓ Loading Data into R

- ✓Load external data from a .csv file into a data frame
- ✓Describe what a data frame is
- ✓Summarize the contents of a data frame
- ✓Use indexing to subset specific parts of a data frame
- ✓Describe a factor
- ✓Reorder and rename factors

Manipulating and Analyzing Data

- Terminology:
  - pipe operator
  - split-apply-combine concept

Manipulating and Analyzing Data

- Describe the purpose of the dplyr and tidyr packages
- Select columns and filter rows
- Use the pipe %>% operator
- Understand and use the split-apply-combine concept for data analysis
- Apply summary statistics and combine results
- Describe the concept of wide and long table formats, and reshape data to and from these formats
- Export data to a .csv file

Packages in R

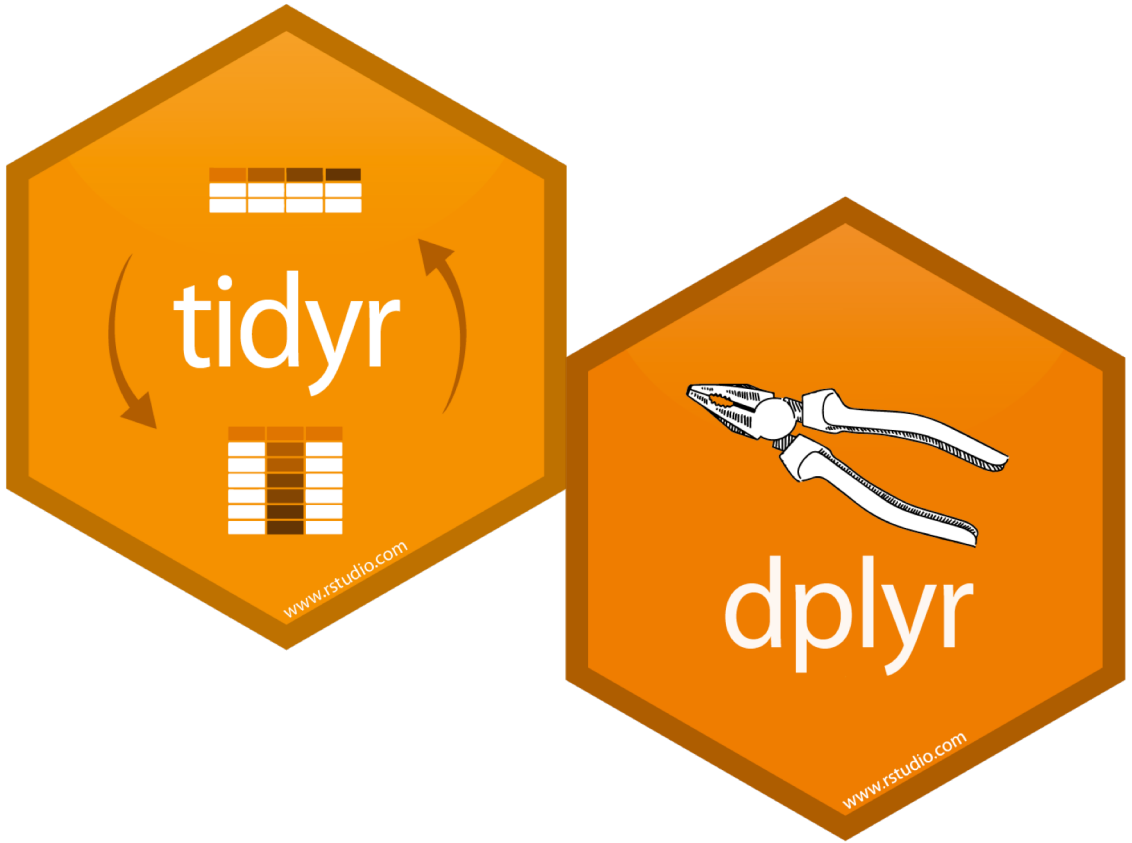
- Packages are extensions of R with additional functions that do not come in base R

tidyverse

- The tidyverse is an umbrella of a suite of packages: tidyr, dplyr, ggplot2, and others.
- Addresses 3 common issues when doing data analysis with functions in base R:
  1. Results from a base R function depend on type of data
  2. R expression used in a non standard way, can be confusing for new learners
  3. Hidden arguments, contains default operations that new learners don't have to be aware of

dplyr and tidyr Packages

- dplyr is a package for manipulating tabular data easily
- tidyr enables conversion between different data formats for plotting and analysis purposes



Tibbles

- When a dataset is read into R using read_csv from the tidyverse package readr, the class of the data is a tbl_df or a “tibble”
- This data structure is very similar to a data frame, with the differences are:
  1. Displays data type of each column under its name, but only prints the first few rows of data and as many columns as will fit on the screen
  2. Columns of class character are never converted into factors

Select Columns and Filter Rows

- Selecting columns of a data frame
- Filter rows by specific criteria

Multiple Operations

- What if you want to select columns and filter rows at the same time?
  - You could do intermediate steps
  - Nested functions

Pipes

- Pass information from one program to another, useful for doing many operations on the same dataset
- Pipes in R are represented by %>%, which the shortcut is Cmd+shift+M or Ctrl+Shift+M

Mutate

- Create new columns based on values in existing columns

Split-apply-combine Approach

- Many data analysis tasks can be tackled with the split-apply-combine approach
- Split the data into groups
- Apply some analysis to each group
- Combine the results
- Exemplified through the dplyr group_by() function

Packages

- ffff