BEWiSE Introduction to R

Stacey Borrego

1/1/2022

## Helpful Links

* Data Carpentry: [Data Analysis and Visualization in R for Ecologists](https://datacarpentry.org/R-ecology-lesson/index.html)
* Tidyverse: [R packages for Data Science](tidyverse.org)
* [R for Data Science](https://r4ds.had.co.nz/) by Hadley Wickham and Garret Grolemund
* [Advanced R](http://adv-r.had.co.nz/) by Hadley Wickham
* [R packages](https://r-pkgs.org/) by Hadley Wickham
* [ggplot2: Elegant graphics for data analysis](https://ggplot2-book.org/) by Hadley Wickham

### Simple math

In the **CONSOLE** section of Rstudio, type in the following expressions.  
Hit ENTER after each expression.

3 + 5  
12 / 7  
3 \* 4  
8 ^ 2  
log(8)  
  
3 + 5 / 2  
(3 + 5) / 2

### Creating objects in R

#### Working with variables/objects

Assign values to a variable/object

weight\_kg <- 55

* Assignment operator (<-)
* Cannot start with numbers
* Case sensitive
* Avoid function names
* Avoid dots (.)
* Be consistent in the styling of your code
  + [R Style Guide](https://google.github.io/styleguide/Rguide.html)
  + Styles can include “lower\_snake”, “UPPER\_SNAKE”, “lowerCamelCase”, “UpperCamelCase”, etc
  + We will use “lower\_snake” for readibility during this workshop

Print the value of an object

weight\_kg

Simple math with objects

2.2 \* weight\_kg

Reassign object’s value

weight\_kg  
  
weight\_kg <- 65  
weight\_kg  
  
2.2 \* weight\_kg

Assign a new object a value containing another object

weight\_lb <- 2.2 \* weight\_kg  
weight\_lb

#### Saving code in a script

* New script
  + File > New File > R Script (Ctrl + Shift + N)
* Save the file
  + File > Save (Ctrl + S)

#### Running code in a script

* Select Run button at the top of the script window
* Select code or move cursor to the line to be run and type Ctrl + ENTER

2 + 2

#### Add comments in a script

The comment character in R is #. Anything to the right of a # in a script will be ignored by R. Comment out a section by selecting and pressing Ctrl + Shift + C

# This is a comment.  
# The sum of 2 + 2  
2 + 2

#### Functions and arguments

* Functions are scripts that automate more commands
* Functions are available in base R, by importing packages, or by making them yourself
* Usually require one or more inputs called **arguments**
* Functions typically return a **value**

# This is a function call  
sqrt(10)  
weight\_kg <- sqrt(10)  
weight\_kg

* Information about functions can be found in the Help window by typing ? and the function name. Example: ?round
* You can also use the function args() to see the arguments of a function.

round(3.14159)  
  
?round  
args(round)  
  
round(3.14159, digits = 2)

### Vectors and data types

* A vector is composed by a series of values, which can be either numbers or characters.
  + Quotes must surround characters.
  + In a vector, all of the elements are the same type of data.
* We can assign a series of values to a vector using the c() function.

# Numerical vector  
weight\_g <- c(50, 60, 65, 82)  
weight\_g  
  
# Character vector  
animals <- c("mouse", "rat", "dog")  
animals

#### Explore the content of vector

There are many functions that allow you to inspect the content of a vector.

* length() tells you how many elements are in a particular vector
* class() indicates what kind of object you are working with
* str() provides an overview of the structure of an object and its elements.

# Length  
length(weight\_g)  
length(animals)  
  
# Class  
class(weight\_g)  
class(animals)  
  
# Structure  
str(weight\_g)  
str(animals)

#### Add elements to a vector

# Add to the end of the vector  
weight\_g <- c(weight\_g, 90)  
  
# Add to the beginning of the vector  
weight\_g <- c(30, weight\_g)   
  
# Inspect the modified vector  
weight\_g

#### Data Types

An atomic vector is the simplest R data type and is a linear vector of a single type.

There are four common types of atomic vectors: character, numeric or double, logical, and integer. There are two rare types that will not be discussed: complex and raw.

* character
* numeric or double
* logical for TRUE and FALSE (the boolean data type)
* integer for integer numbers (e.g., 2L, the L indicates to R that it’s an integer)
* complex to represent complex numbers with real and imaginary parts (e.g., 1 + 4i)
* raw for bitstreams

typeof(weight\_g)  
typeof(animals)

#### Data Structures

Vectors are one of the many data structures that R uses.

* lists (list)
* matrices (matrix)
* data frames (data.frame)
* factors (factor)
* arrays (array)

#### Coercion

All elements of an atomic vector must be the same type, so when you attempt to combine different types they will be coerced to the most flexible type. Types from least to most flexible are: logical, integer, double, and character.

num\_char <- c(1, 2, 3, "a")  
typeof(num\_char)  
str(num\_char)  
  
num\_logical <- c(1, 2, 3, TRUE, FALSE)  
typeof(num\_logical)  
str(num\_logical)  
  
char\_logical <- c("a", "b", "c", TRUE)  
typeof(char\_logical)  
str(char\_logical)  
  
tricky <- c(1, 2, 3, "4")  
typeof(tricky)  
str(tricky)

### Conditional subsetting

To extract one or several values from a vector, provide one or several indices in square brackets.

animals <- c("mouse", "rat", "dog", "cat")  
  
animals[2]  
  
animals[c(3, 1)]  
  
animals[1:3]  
  
more\_animals <- animals[c(1, 4, 1, 3, 1, 2)]  
more\_animals

Subset by using a logical vector. TRUE will select the element with the same index, while FALSE will not.

weight\_g <- c(21, 34, 39, 54, 55)  
  
weight\_g[c(TRUE, FALSE, FALSE, TRUE, TRUE)]  
  
weight\_g > 50  
  
weight\_g[weight\_g > 50]

Combine multiple tests using & (both conditions are true, **AND**) or | (at least one of the conditions is true, **OR**).

Helpful operators

* > greater than
* < less than
* <= less than or equal to
* == equal to, test for numerical equality between the left and right hand sides

weight\_g  
  
weight\_g[weight\_g > 30 & weight\_g < 50]  
  
weight\_g[weight\_g <= 30 | weight\_g == 55]  
  
weight\_g[weight\_g >= 30 & weight\_g == 21]

Search for strings in a vector

animals <- c("mouse", "rat", "dog", "cat", "cat")  
animals  
  
animals[animals == "cat" | animals == "rat"]  
  
animals %in% c("rat", "cat", "dog", "duck", "goat", "bird", "fish")  
  
animals[animals %in% c("rat", "cat", "dog", "duck", "goat", "bird", "fish")]

### Missing data

* Missing data are represented in vectors as NA.
* You can add the argument na.rm = TRUE to calculate the result as if the missing values were removed (rm stands for **r**e**m**ove)
* is.na() identifies missing elements
* na.omit() returns the object with incomplete cases removed
* complete.cases() returns a logical vector indicating which cases are complete

heights <- c(2, 4, 4, NA, 6)  
  
mean(heights)  
max(heights)  
  
mean(heights, na.rm = TRUE)  
max(heights, na.rm = TRUE)

heights  
  
heights[is.na(heights)]  
heights[!is.na(heights)]  
  
na.omit(heights)  
  
heights[complete.cases(heights)]

### Data Frames

When we load data into R, it may be stored as an object of class data frame.

A data frame is the representation of data in the format of a table where the columns are vectors that all have the same length. Because columns are vectors, each column must contain a single type of data (e.g., characters, integers, factors).

For example, here is an example data frame comprising a numeric, character, and logical vector.

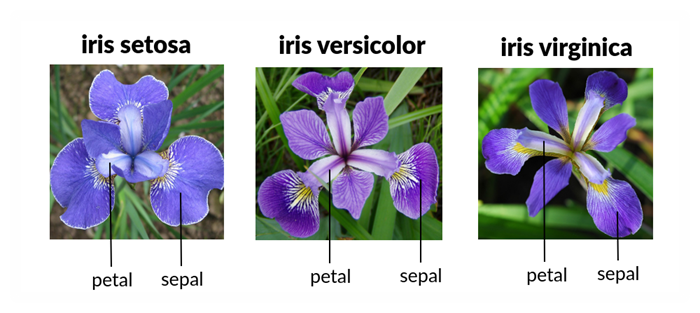
## numeric character logical  
## 1 1 S TRUE  
## 2 7 A FALSE  
## 3 3 M TRUE

### Starting with Data

This is what the data will look like:

|  |  |
| --- | --- |
| Column | Description |
| Sepal.Length | Sepal length in centimeters |
| Sepal.Width | Sepal width in centimeters |
| Petal.Length | Petal length in centimeters |
| Petal.Width | Petal width in centimeters |
| Species | Name of species name |

# To see data sets available  
data()  
  
# Load the data set of choice, use the data set name as the argument to `data()`  
data(iris)



iris

Open the dataset in RStudio’s Data Viewer

# Check the data type of the data just loaded  
class(iris)  
  
# View the whole data set  
view(iris)

Inspect the data

# See a few rows of the data  
head(iris)  
tail(iris)  
  
# See a specific number of rows of the data  
head(iris, n=10)  
tail(iris, n=10)

* Size:
  + dim(iris) - returns a vector with the number of rows in the first element, and the number of columns as the second element (the dimensions of the object)
  + nrow(iris) - returns the number of rows
  + ncol(iris) - returns the number of columns
* Content:
  + head(iris) - shows the first 6 rows
  + tail(iris) - shows the last 6 rows
* Names:
  + names(iris) - returns the column names (synonym of colnames() for data.frame objects)
  + rownames(iris) - returns the row names
* Summary:
  + str(iris) - structure of the object and information about the class, length and content of each column
  + summary(iris) - summary statistics for each column

### Indexing and subsetting data frames

A data frame has rows and columns in 2 dimensions. To extract specific data, specify the “coordinates” in [ ] indicating row numbers first followed by column numbers.

# dim() gives the output of the rows and columns of the object  
dim(iris)  
  
# Extract specific values by specifying row and column indices  
# in the format:   
# data\_frame[row\_index, column\_index]  
  
# Extract the first row and first column from surveys  
iris[1, 1]  
  
# Extract the first row and sixth column  
iris[1, 6]   
  
# To select all columns, leave the column index blank  
# Extract the first row and all columns   
iris[1, ]  
  
# Extract all rows and the first column  
iris[, 1]  
  
# An even shorter way to select first column across all rows:  
iris[1] # No comma!   
  
# Select multiple rows or columns with vectors  
# Extract the first three rows of the 5th and 6th column  
iris[c(1, 2, 3), c(5, 6)]   
  
# We can use the : operator to create those vectors   
iris[1:3, 5:6]   
  
# This is equivalent to head(surveys)  
iris[1:6, ]  
  
# Subsetting with single square brackets ("[]") always returns a data frame.  
# If you want a vector, use double square brackets ("[[]]")  
  
# For instance, to get a column as a vector:  
iris[[1]]  
iris[["Species"]]  
  
# To get the first value in our data frame:  
iris[[1, 1]]  
iris[1, "Species"]  
  
# Data frames can be subset by calling their column names directly  
# Use the $ operator with column names to return a vector  
iris$Species

# Exceptions: subsetting the whole data frame, except the first column  
iris[, -1]   
  
# Exceptions: subsetting all columns of the data frame but only the first six rows  
iris[-(7:nrow(iris)), ]

### Plotting

Helpful links:

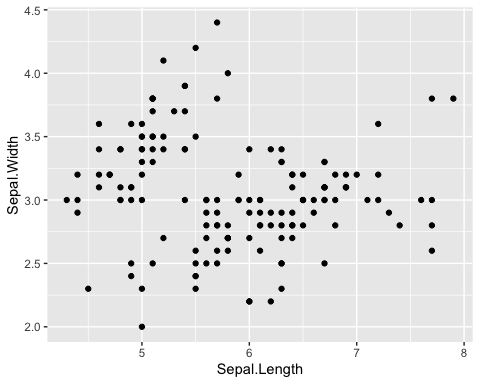
* [ggplot2: Elegant Graphics for Data Analysis](https://ggplot2-book.org/getting-started.html#basic-use)

Plotting can be done in many different ways in R. Base R has several basic plotting functions which can be valuable for a quick peek at your data. However, a common and customizable option is to use the package ggplot2.

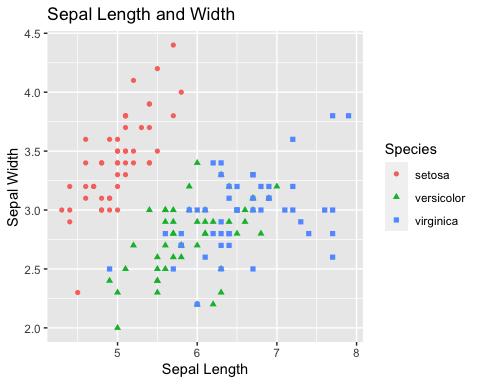
Every ggplot2 plot has three key components:

* **Data** (data)
* A set of **aesthetic mappings** between variables in the data and visual properties (aes(x, y))
* At least one **layer** which describes how to render each observation. Layers are usually created with a geom function. It is important to note that layers are added with a +.

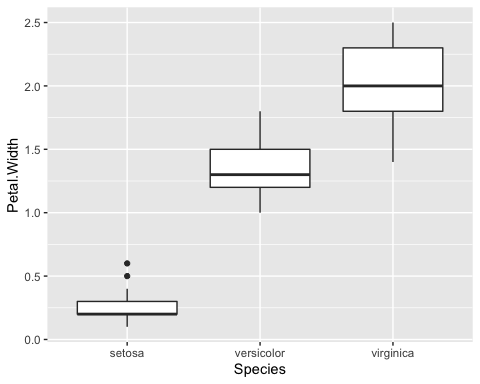
library(ggplot2)  
  
# Scatter Plot  
plot\_1 <- ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width))   
  
# Basic scatter plot adding a `geom\_point()` layer  
plot\_1 +   
 geom\_point()



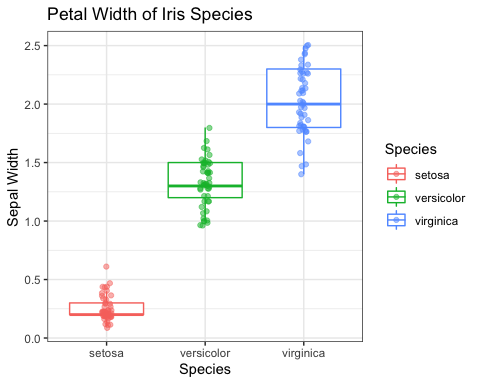
# Colorful scatter plot adding aesthetic mappings to the `geom\_point()` layer and labels  
plot\_1 +   
 geom\_point(aes(color = Species, shape = Species)) +  
 labs(  
 x = "Sepal Length",  
 y = "Sepal Width",  
 title = "Sepal Length and Width")



# Boxplot  
plot\_2 <- ggplot(data = iris, aes(x = Species , y = Petal.Width))   
  
# Basic boxplot adding a `geom\_boxplot()` layer  
plot\_2 +  
 geom\_boxplot()



# Colorful boxplot adding more layers and aesthetic mappings to the data  
plot\_2 +   
 geom\_boxplot(aes(color = Species),   
 outlier.colour = NA,   
 position = "dodge") +  
 geom\_jitter(aes(color = Species,   
 x = Species),   
 position = position\_jitter(width = .05),   
 alpha = 0.5) +  
 scale\_x\_discrete() +  
 labs(  
 x = "Species",   
 y = "Sepal Width",  
 title = "Petal Width of Iris Species") +  
 theme\_bw()



## Extra Information

### Factors

This can be the cause of many issues when plotting data. Always check if the data is factored and determined whether is should or should not be to get the correct plot.

R has a special class for working with categorical data, called factors. Once created, factors can only contain a pre-defined set of values, known as levels. Factors are stored as integers associated with labels and they can be ordered or unordered. While factors look (and often behave) like character vectors, they are actually treated as integer vectors by R. So you need to be very careful when treating them as strings.

When working with data in a data frame, the columns that contain text are not automatically coerced (= converted) into the factor data type, but once we have loaded the data we can do the conversion using the factor() function

head(iris)  
iris$Species <- factor(iris$Species)  
head(iris)  
  
summary(iris$Species)

By default, R always sorts levels in alphabetical order. In the example below, R will assign 1 to the level “experimental” and 2 to the level “wild type” (because “e” comes before “w”, even though the first element in this vector is “wild type”).

sample <- factor(c("wild type", "experimental", "wild type", "experimental"))  
  
# Different ways to see the levels of factored data  
sample  
levels(sample)  
nlevels(sample)

Reorder factors

sample <- factor(sample, levels = c("wild type", "experimental"))  
  
sample  
levels(sample)  
nlevels(sample)

#### Converting Factors

Sometimes it is necessary to convert data from one type to another

sample  
  
as.character(sample)  
sample\_char <- as.character(sample)  
sample\_char  
  
as.numeric(sample)

year\_fct <- factor(c(1990, 1983, 1977, 1998, 1990))  
levels(year\_fct)  
  
as.numeric(year\_fct)   
as.character(year\_fct)  
  
as.numeric(as.character(year\_fct))  
as.numeric(levels(year\_fct))[year\_fct]

Notice that in the levels() approach, three important steps occur:

* Obtain all the factor levels using levels(year\_fct)
* Convert these levels to numeric values using as.numeric(levels(year\_fct))
* Access these numeric values using the underlying integers of the vector year\_fct inside the square brackets

#### Renaming Factors

# Inspect the contents of the column   
summary(iris$Species)  
  
# Plot the column of the data   
plot(iris$Species)

## Data Carpentry: Manipulating, analyzing and exporting data with tidyverse

### Data manipulation using dplyr and tidyr

[dplyr](https://dplyr.tidyverse.org/) is a package for helping with tabular data manipulation. It pairs nicely with [tidyr](https://tidyr.tidyverse.org/) which enables you to swiftly convert between different data formats for plotting and analysis.

The [tidyverse](https://www.tidyverse.org/) package is an “umbrella-package” that installs tidyr, dplyr, and several other useful packages for data analysis, such as ggplot2, tibble, etc.

dplyr functions

* select() subset columns
* filter() subset rows on conditions
* mutate() create new columns by using information from other columns
* group\_by() and summarize() create summary statistics on grouped data
* arrange() sort results
* count() count discrete values

### Selecting columns and filtering rows

* select() subset columns
  + The first argument is the data frame (surveys), and the subsequent arguments are the columns to keep.
* filter() subset rows on conditions

library(tidyverse)  
select(iris,   
 Species, Petal.Length, Petal.Width)

Select all columns except certain ones by using a “-” in front of the variable to exclude it

select(iris,   
 -Sepal.Length, -Sepal.Width)

Choose rows based on specific criterion with filter()

filter(iris, Species == "setosa")

### Pipes

What if you want to select and filter at the same time? There are three ways to do this: use intermediate steps, nested functions, or pipes.

Pipes let you take the output of one function and send it directly to the next, which is useful when you need to do many things to the same dataset. Pipes in R look like %>% and are made available via the magrittr package, installed automatically with dplyr. If you use RStudio, you can type the pipe with Ctrl + Shift + M if you have a PC or Cmd + Shift + M if you have a Mac.

filter(iris,   
 Petal.Length < 5)  
select(iris,   
 Species, Petal.Length, Petal.Width)  
  
iris %>%  
 filter(Petal.Length < 5) %>%  
 select(Species, Petal.Length, Petal.Width)

Create a new object with this smaller version of the data by assigning it a new name

iris\_petals <- iris %>%  
 filter(Petal.Length < 5) %>%  
 select(Species, Petal.Length, Petal.Width)  
  
iris\_petals

#### Mutate

Create new columns based on the values in existing columns using mutate()

iris %>%  
 mutate(petal\_ratio = Petal.Width / Petal.Length)  
  
  
iris %>%  
 mutate(petal\_ratio = Petal.Width / Petal.Length,  
 petal\_percent = petal\_ratio \* 100)

The first few rows of the output contain NAs. To remove those we can insert a filter() in the chain of functions combined with is.na() to identify all NAs and the ! operator to negate the result

iris %>%  
 filter(!is.na(Petal.Length)) %>%  
 mutate(petal\_ratio = Petal.Width / Petal.Length) %>%  
 tail()

#### Split-apply-combine data analysis and the summarize() function

Many data analysis tasks can be approached using the split-apply-combine paradigm.

* Split the data into groups
* Apply some analysis to each group
* Combine the results

Key functions of dplyr for this workflow are group\_by() and summarize(). group\_by() is often used together with summarize(), which collapses each group into a single-row summary of that group.

iris %>%  
 group\_by(Species) %>%  
 summarize(petal\_width\_mean = mean(Petal.Width, na.rm = TRUE))

Once the data are grouped, you can also summarize multiple variables at the same time. Add a column indicating the minimum and maximum width for each species.

iris %>%  
 filter(!is.na(Petal.Width)) %>%  
 group\_by(Species) %>%  
 summarize(petal\_width\_mean = mean(Petal.Width),  
 petal\_width\_min = min(Petal.Width),  
 petal\_width\_max = max(Petal.Width))

### Exporting plots

ggsave() allows you easily change the dimension and resolution of your plot by adjusting the appropriate arguments (width, height and dpi)

my\_plot <- ggplot(data = iris,  
 aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +  
 geom\_point() +  
 facet\_wrap(vars(Species)) +  
 labs(title = "Petal width by species",  
 x = "Sepal Length",  
 y = "Sepal Width") +  
 theme\_bw() +  
 theme(axis.text.x = element\_text(colour = "grey20", size = 12, angle = 90,  
 hjust = 0.5, vjust = 0.5),  
 axis.text.y = element\_text(colour = "grey20", size = 12),  
 text = element\_text(size = 16))  
  
ggsave("path/name\_of\_file.png", my\_plot, width = 15, height = 10)