Worksheet 1: R Basics

Amanda Regan

2022-12-14

This is the first in a series of worksheets for History 8510 at Clemson University. The goal of these worksheets is simple: practice, practice, practice. The worksheet introduces concepts and techniques and includes prompts for you to practice in this interactive document. When you are finished, you should change the author name (above), knit your document to a pdf, and upload it to canvas.

What is R?

To start let's define what exactly R is. R is a language and environment for statistical computing and graphics. R provides a variety of statistical and graphical techniques and its very extensible which makes it an ideal language for historians.

Foundational Concepts

Values

There are several kinds of variables in R. Numeric, logical, and strings.

R takes inputs and returns an output. So for example, if our input is a number the output will be a number. Simply typing a number, below or in the console, will return a number.

```
5
```

```
## [1] 5
```

Same thing happens if we add a number with a variable.

```
5.5483
```

```
## [1] 5.5483
```

Numbers can be used to do arithmetic.

```
5 + 5
```

```
## [1] 10
```

(1) You try, multiple two numbers.

(2) Can you multiply two numbers and then divide by the result?

The next type of value is a string. Strings are lines of text. Sometimes these are referred to as character vectors. To create a string, you add text between two quotation marks. For example:

```
"Go Tigers"
```

```
## [1] "Go Tigers"
```

(3) Try to create your own string.

You can't add strings using `+` like you can numbers. But there is a function called `paste()` which concatenates character vectors into one. This function is *very* useful and one you'll use a lot in a variety of circumstances. Here's what that looks like:

```
paste("Hello", "Clemson Graduate Students")
```

```
## [1] "Hello Clemson Graduate Students"
```

(4) Try it, add two strings together like the above example.

(5) Can you explain what happened in 2-3 sentences?

test answer

The last type are logical values like `TRUE` and `FALSE`. (Note that these are all caps.)

```
TRUE
```

```
## [1] TRUE
```

```
FALSE
```

```
## [1] FALSE
```

These logical values are really useful for testing a statement and comparing values to each other. For example:

```
3 < 4
```

```
## [1] TRUE
```

3 is indeed less than 4 so the return value is `TRUE`. Here are a few more examples:

```
5 == 10
```

```
## [1] FALSE
```

```
3 < 4
```

```
## [1] TRUE
```

```
3 == 4 | 3
```

```
## [1] TRUE
```

(6) Explain, what does the code on each line above do?

(7) Create your own comparison.

Values are great but they are made so much more powerful when combined with **Variables** which are a crucial building block of any programming language. Simply put, a variable stores a value. For example if I want x to equal 5 I can do that like this:

```
x <- 5
```

`<-` is known as an assignment operator. Technically, you could use `=`: here too but it is considered bad practice and can cause complicated issues when you write more advanced code. So its important to stick with `<-` whenever you are coding in R.

I can also add a string or character vector to a variable.

```
x <- "Go Tigers!"
```

Variable names can be almost anything.

```
MyFavoriteNumber <- 25
```

Variable or object names must start with a letter, and can only contain letters, numbers, _, and .. You want your object names to be descriptive, so you'll need a convention for multiple words. People do it different ways. I tend to use periods but there are several options:

```
i_use_snake_case
otherPeopleUseCamelCase
some.people.use.periods
And_aFew.People_RENOUNCEconvention
```

Whatever you prefer, be consistent and your future self will thank you when your code gets more complex.

(8) You try, create a variable and assign it a number:

(9) Can you assign a string to a variable?

(10) Once we've assigned a variable we can use that variable to run calculations just like we did with raw numbers.

```
x <- 25
x * 5
```

```
## [1] 125
```

R can also handle more complex equations.

```
(x + x)/10
```

```
## [1] 5
```

```
(x + x * x) - 100
```

```
## [1] 550
```

And we could store the output of a calculation in a new variable:

```
My.Calculation <- (x + x)*10
```

(11) You try. Assign a number to x and a number to y. Add those two numbers together.

(12) Can you take `x` and `y` and multiply the result by 5?

(13) Try creating two variables with names other than `x` and `y`. Descriptive names tend to be more useful. Can you multiply the contents of your variables?

(14) Try creating two variables that store strings. Can you concatenate those two variables together?

If we have a lot of code and rely on just variables, we're going to have a lot of variables. That's where vectors come into play. Vectors allow you to store multiple values. All variables in R are actually already vectors. That's why when R prints an output, there is a `[1]` before it. That means there is one item in that vector.

```
myvalue <- "George Washington"
myvalue
```

```
## [1] "George Washington"
```

In this instance "George Washington" is the only item in the variable myvalue. But we could add more. To do that we use the `c()` function which combines values into a vector.

```
myvalue <- c("George Washington", "Franklin Roosevelt", "John Adams")
myvalue
```

```
## [1] "George Washington"  "Franklin Roosevelt" "John Adams"
```

You'll notice that the output still only shows `[1]` but that doesn't mean there is only one item in the list. It simply means George Washington is the first. If we use `length()` we can determine the number of items in this vector list.

```
length(myvalue)
```

```
## [1] 3
```

We could get the value of the 2nd or 3rd item in that list like this:

```
myvalue[2]
```

```
## [1] "Franklin Roosevelt"
```

We could also create a vector of numbers:

```
my.numbers <- c(2, 4, 6, 8)
```

And we could then do calculations based on these values:

```
my.numbers * 2
```

```
## [1]  4  8 12 16
```

(15) Explain in a few sentences, what happened in the code above?

Lets try something slightly different.

```
my.numbers[3] * 2
```

```
## [1] 12
```

(16) Explain in a few sentences, what happened in the code above?

(17) You try, create a list of five items and store it in a descriptive variable.

R also has **built in functions**. We've already used a couple of these: `paste()` and `c()`. But there are others, like `sqrt()` which does what you think it does, finds the square root of a number.

```
sqrt(1000)
```

```
## [1] 31.62278
```

Many functions have options that can be added to them. For example, the `round()` function allows you to include an option specifying how many digits to round to.

You can run it without that option and it'll use the default:

```
round(15.492827349)
```

```
## [1] 15
```

Or we can tell it to round it to 2 decimal places.

```
round(15.492827349, digits = 2)
```

```
## [1] 15.49
```

How would you know what options are available for each function in R? Every function and package in R comes with **documentation** or a **manual** that is built into R studio and can be pulled by by typing a question mark in front of the function in your console. These packages will commonly give you examples of how to use the function and syntax for doing so. They are incredibly useful.

We can ask R to pull up the documentation like this:

```
?round()
```

(18) Now you try, find the documentation for the function `signif()`.

In real life, you typically you wouldn't want to store this code in your script file. You probably don't need to pull up the documentation for the function every time you run that piece of code. But for the purposes of this worksheet we're adding it to our `.Rmd` document.

(19) Use the console to find the documentation for `floor()`? Try it and then tell me, what does that function do? <

Data Frames & Packages

R is the language of choice for most data scientists and that is because of its powerful suite of data analysis tools. Some are built into R, like the `floor()` which we looked at above. But others come from packages. Most programming languages have some sort of package system although every language calls it a slightly different thing. Ruby has gems, python has eggs, and php has libraries. In R these are called packages or libraries and they are hosted by the Comprehensive R Archive Network or more commonly, CRAN. CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. You didn't know it but you already used CRAN when you looked up the documentation for `floor()` above. If you go to the CRAN webpage and look at the list of available R packages you'll see just how many there are!

So there are many packages but there are also some that are indispensable and that you'll use over and over for this class. We'll get into some of those in the next two worksheets but for now lets look at one basic package for data.

Lets start with the `tibble()` package which allows you to create and work with **data frames**. What is a data frame? Think of it as a spreadsheet. Each column can contain data - including numbers, strings, and logical values.

Lets load the `tibble()` package. If you have this package installed this line of code will work. If not, you'll get an error that says something like: `## Error in library("tibble"): there is no package called 'tibble'`. If that's the case (it probably is), then no worries - we can install it. To install a package run `install.packages("tibble")` in your console and R will download and install the package.

```
library(tibble)
```

While we're at it, lets also install the data package that I've created for our class. This package contains a variety of historical datasets that you can use to complete your assignments this semester. However, this package is not on CRAN. That's okay though, we can still install it from github. To do that we'll need to use the `devtools` package. We'll install devtools and then use it to install our class's package which is hosted on GitHub.

First, install the `devtools` package.

Our class's package is called `DigitalMethodsData` and its hosted on GitHub. Use the directions on the repository page and what you learned about packages above to install this package.

After installing the package, we should load it:

```
library(DigitalMethodsData)
```

What kind of datasets are available in this package? We can use the help documentation to find out.

Run `help(package="DigitalMethodsData")` in your console to pull up a list of datasets included in this package.

For demonstration purposes we'll use the `gayguides` data here. Pull up the help documentation for this package. What is the scope of this dataset?

<

To use a dataset included in this package we first need to load it. We can do so like this:

```
data(gayguides)
```

Notice that you now have a loaded dataset in your environment pane. It shows us that there are 60,698 observations (rows) of data and 14 variables.

Let's now look at our data. You can use the `head()` function to return the first part of an object or dataset.

```
head(gayguides)
```

```
##   X   ID        title                   description streetaddress
## 1 1 3213       'B.A.'                (woods & ponds)
## 2 2 2265 'B.A.' Beach                      2 mi. E.        Rte. 2
## 3 3 3269 'B.A.' Beach             nr. Salt Air Beach
## 4 4 3388 'B.A.' Beach nr. Evergreen Floating Bridge
## 5 5 3508 'B.A.' Beach          (2 mi. E. on Rte. 2)
## 6 6 5116 'B.A.' Beach nr. Evergreen Floating Bridge
##             type amenityfeatures           city state Year notes      lat
## 1 Cruising Areas     Cruisy Area    Lake Placid    NY 1982       44.27949
## 2 Cruising Areas     Cruisy Area           Troy    NY 1981       42.72841
## 3 Cruising Areas     Cruisy Area Salt Lake City    UT 1981       40.74781
## 4 Cruising Areas     Cruisy Area        Seattle    WA 1981       47.60621
## 5 Cruising Areas     Cruisy Area           Troy    NY 1982       42.72841
## 6 Cruising Areas     Cruisy Area        Seattle    WA 1983       47.60621
##          lon
## 1  -73.97987
## 2  -73.69178
## 3 -112.18727
## 4 -122.33207
## 5  -73.69178
## 6 -122.33207
##                                                                      status
## 1 Location could not be verified. General city or location coordinates used.
## 2 Location could not be verified. General city or location coordinates used.
## 3 Location could not be verified. General city or location coordinates used.
## 4 Location could not be verified. General city or location coordinates used.
## 5 Location could not be verified. General city or location coordinates used.
## 6 Location could not be verified. General city or location coordinates used.
```

That gives us the first 6 rows of data. Its really useful if you just want to peak into the dataset but don't want to print out all 60k+ rows.

(20) The default is to print six rows of data. Can you modify the above code to print out the first 10 rows?

(21) Can you find the last 6 rows of data?

(22) Reflect on the previous two prompts. How did you figure these out? <

The `str()` function is another very useful function for understanding a dataset. It compactly displays the internal structure of an R object.

```
str(gayguides)
```

```
## 'data.frame':    60698 obs. of  14 variables:
##  $ X             : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID            : chr  "3213" "2265" "3269" "3388" ...
##  $ title         : chr  "'B.A.'" "'B.A.' Beach" "'B.A.' Beach" "'B.A.' Beach" ...
```

```
##  $ description    : chr  "(woods & ponds)" "2 mi. E." "nr. Salt Air Beach" "nr. Evergreen Floating B
##  $ streetaddress  : chr  "" "Rte. 2" "" "" ...
##  $ type           : chr  "Cruising Areas" "Cruising Areas" "Cruising Areas" "Cruising Areas" ...
##  $ amenityfeatures: chr  "Cruisy Area" "Cruisy Area" "Cruisy Area" "Cruisy Area" ...
##  $ city           : chr  "Lake Placid" "Troy" "Salt Lake City" "Seattle" ...
##  $ state          : chr  "NY" "NY" "UT" "WA" ...
##  $ Year           : int  1982 1981 1981 1981 1982 1983 1983 1983 1983 1984 ...
##  $ notes          : chr  "" "" "" "" ...
##  $ lat            : num  44.3 42.7 40.7 47.6 42.7 ...
##  $ lon            : num  -74 -73.7 -112.2 -122.3 -73.7 ...
##  $ status         : chr  "Location could not be verified. General city or location coordinates used.
```

There is a ton of useful info here. First, we see that this is a data.frame and that it has 60698 obs. of 14 variables. Second, we see the names of all the variables (columns) in the dataset. Lastly, it shows us what type of data is contained in each variable. For example, the variable `state` is a chr or character vector while `Year` (note the capitalization) is numeric.

The `$` operator allows us to get a segment of the data.

```
head(gayguides$title)
```

```
## [1] "'B.A.'"       "'B.A.' Beach" "'B.A.' Beach" "'B.A.' Beach" "'B.A.' Beach"
## [6] "'B.A.' Beach"
```

While that's useful, this particular dataset happens to be sorted by title. So the first rows are all labeled some version of B.A. beach. That's not very useful. What if we want to see the 500th item in this list? Well, we can pull that up like this:

```
gayguides$title[500]
```

```
## [1] "'Rest Stop'"
```

(23) Break this down. Whats going on here? What do each of the elements of this code mean? (gayguides, $, title, and [500]) <

(24) The Year variable contains the year (numeric) of each entry. Can you find the earliest year in this dataset? (Using code, don't cheat and use the documentation!)

(25) How about the latest year?

(26) Add another dataset from `DigitalMethodsData`. How many observations are included in this dataset? Run through some of the examples from above to learn about the dataset.

(27) What did you learn about the dataset? What can you tell me about it? <

CONGRATS! You've completed your first worksheet for Digital Methods II. That wasn't so bad, right?