Intermediate R Handout

Data Carpentry contributors & Montana State University R Workshops Team

2/22/2024

Learning Objectives

- Use relational operators
- Join sequences of relational operations together with logical operators
- Create conditional statements (`ifelse()`)
- Create `for` loops
- Create functions

Orientation of/for the workshop

- This workshop assumes some basic familiarity with working in `R` such as what you might obtain in the "Introduction to R" workshop or in a statistics course that uses `R` heavily, such as STAT 217 or STAT 411/511. If you have not interacted with `R` previously, some of the assumptions of your background for this workshop might be a barrier. We would recommend getting what you can from this workshop and you can always revisit the materials at a later date after filling in some of those basic `R` skills. We all often revisit materials and discover new and deeper aspects of the content that we were not in a position to appreciate in a first exposure.

- In order to focus this workshop on coding, we developed this interactive website for you to play in a set of "sandboxes" and try your hand at implementing the methods we are discussing. When each code chunk is ready to run (all can be edited, many have the code prepared for you), you can click on "Run Code". This will run `R` in the background on a server. For the "Challenges", you can get your answer graded although many have multiple "correct" answers, so don't be surprised if our "correct" answer differs from yours. The "Solution" is also provided in some cases so you can see a solution - but you will learn more by trying the challenge first before seeing the answer. Each sandbox functions independently, which means that you can pick up working at any place in the documents and re-set your work without impacting other work (this is VERY different from how `R` usually works!). Hopefully this allows you to focus on the code and what it does… The "Start over" button can be used on any individual sandbox or you can use the one on the left tile will re-set all the code chunks to the original status.

- These workshops are taught by Greta Linse and Esther Birch and co-organized by the MSU Library, Statistical Consulting and Research Services (SCRS), and the Department of Mathematical Sciences. More details on us and other workshops are available at the end of the session or via https://www.montana.edu/datascience/training/#workshop-recordings.

- First, we will start with the building blocks of conditional statements, the relational operator. Next, we will join sequences of relational operations together with logical operators. We will then use these relational operators inside conditional statements (`ifelse()`).

- Finally, we will dive into two methods to avoid copying and pasting your code numerous times to accomplish the same task. The `for` loop will be introduced for procedures done multiple times in *one* location in your code, and functions will be introduced for procedures done *throughout* your code.

Let's get started!

Refresher

This workshop covers content that will require that we remember how to extract elements from vectors and dataframes. Let's work through the following warm-ups to refresh how we can use these operations.

Extracting Elements from a Vector

To extract an element from a vector we use the bracket (`[]`) notation. Recall, a vector has only one dimension, so inside the brackets goes *one* number. To extract elements of a vector, we can give their corresponding index, starting from the number one. (`R` uses a one-based numbering system.)

```
x <- c(5.4, 6.2, 7.1, 4.8, 7.5)

x[1]  ## Extracts the first element of x
```

```
## [1] 5.4
```

```
x[1:3]  ## Extracts a slice of x, from the first index to the third index
```

```
## [1] 5.4 6.2 7.1
```

```
x[-2]  ## Extracts everything in x BUT the second index
```

```
## [1] 5.4 7.1 4.8 7.5
```

Challenge 1

Why does `R` produce an error for the following code?

```
x <- c(5.4, 6.2, 7.1, 4.8, 7.5)
```

```
x[-1:3]
```

Extracting Elements from a Dataframe

A dataframe is a list of vectors, where each vector is permitted to have a different data type. Because dataframes have two dimensions (columns and rows), we are able to extract two types of elements.

To extract a column from a dataframe we can use the familiar `$` operator:

```
example_df <- data.frame(A = c(1, 5, 9, 13), B = c(2, 6, 10, 14), C = c(3, 7, 11,
    15), D = c(4, 8, 12, 16))

example_df$A  ## Extracts the A column from the dataframe
```

```
## [1]  1  5  9 13
```

This is useful if we only wish to extract a *single* vector from our dataframe. If we want multiple vectors, then using matrix notation is more simple.

To extract a column from a dataframe using brackets (`[row, column]`):

```
example_df <- data.frame(A = c(1, 5, 9, 13), B = c(2, 6, 10, 14), C = c(3, 7, 11,
    15), D = c(4, 8, 12, 16))
```

```
example_df[1, 1]  ## Extracts the first row, first column entry
```

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

```
example_df[, 1]  ## Extracts EVERY row in the first column
```

```
## [1]  1  5  9 13
```

```
example_df[1, ]  ## Extracts EVERY column in the first row
```

```
##   A B C D
## 1 1 2 3 4
```

Challenge 2 (Part 1)

I want to extract the 3rd and 4th columns. What is wrong with my code?

```
## Change the code below to extract the 3rd and 4th columns

# example_df[, 3, 4]
```

```
##    C  D
## 1  3  4
## 2  7  8
## 3 11 12
## 4 15 16
```

Challenge 2 (Part 2)

How would you change it to extract these columns? What if I wanted to extract the 2nd and 4th columns?

```
## Write code to extract the 2nd and 4th columns
```

```
example_df[, c(2, 4)]
```

```
##    B  D
## 1  2  4
## 2  6  8
## 3 10 12
## 4 14 16
```

Relational Operators

Relational operators tell how one object relates to another, where the output is a logical value. There are a variety of relational operators that you have used before in mathematics but take on a slightly different feel in data science. We will walk through a few examples of different types of relational operators and the rest will be left as exercises.

```
w <- 10.2
x <- 1.3
y <- 2.8
z <- 17.5
dna1 <- "attattaggaccaca"
dna2 <- "attattaggaacaca"
```

```
w <- 10.2
x <- 1.3
y <- 2.8
```

```
z <- 17.5
dna1 <- "attattaggaccaca"
dna2 <- "attattaggaacaca"
```

Equality & Inequality

This type of operator tells whether an object is equivalent to another object (`==`), or its negation (`!=`).

```
dna1 == dna2
```

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

```
dna1 != dna2
```

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

Greater & Less

These statements should be familiar, with a bit of a notation twist. To write a strict greater than or less than statement you would use `>` or `<` in your statement. To add an equality constraint, you would add a `=` sign after the inequality statement (`>=` or `<=`).

```
w > 10
```

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

```
x > y
```

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

Inclusion

This type of statement checks if a character, number, or factor are included in a vector. The `%in%` operator tells you whether a value is included in an object. These values can be linked together into a vector, and the output will be a vector of logical values.

```
colors <- c("green", "pink", "red")

"blue" %in% colors
```

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

```
numbers <- c(1, 2, 3, 4, 5)

5 %in% numbers
```

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

```
some_letters <- c("a", "b", "c", "d", "e")

c("a", "b") %in% some_letters
```

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

Challenge 3

Write `R` code to see if:

- 2 * x + 0.2 is equal to y
- "hello" is greater than or equal to "goodbye"

- TRUE is greater than FALSE
- dna1 is longer than 5 bases (use nchar() to figure out how long a string is)

```
# Challenge 3 R code goes here
```

```
2 * x + 0.2 == y
```

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

```
"hello" > "goodbye"
```

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

```
TRUE > FALSE
```

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

```
nchar(dna1)
```

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

Comparing Decimal Valued Numbers

Why does the output from `2 * x + 0.2 == y` not seem right? What do you think might be going on?

General `R` advice is that the `==` operator should only be used for comparisons of integer and Boolean data types. To store decimal valued data types (doubles) `R`, like other programming systems, uses a format (binary floating-point) that doesn't accurately represent a number like 1.3. When the code is interpreted by `R`, the "1.3" is rounded to the nearest number in floating-point format, which results in a small rounding error even before the calculation happens.

This is why, general `R` advice is to use the `all.equal(x, y)` function to check for equality of two doubles. This function, applied to the two arguments, "is a utility to compare `R` objects `x` and `y` testing *near equality*." This means that the function allows for a little bit of wiggle room in the rounding errors that are associated with doubles.

```
# Verify the result using all.equal(..., ...)
```

Comparing Characters

How did `R` know how to compare the words `hello` and `goodbye`?

```
"hello" > "goodbye"
```

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

```
# What does this output suggest?
Sys.getlocale(category = "LC_COLLATE")
```

```
## [1] "English_United States.utf8"
```

Challenge 4

What is going on in the code below? Why is `R` giving an error?

```
some_letters <- c("a", "b", "c", "d", "e")
```

```
some_letters != c("a", "c")
```

```
## Warning in some_letters != c("a", "c"): longer object length is not a multiple
## of shorter object length
## [1] FALSE  TRUE  TRUE  TRUE  TRUE
```

Why does R give TRUE as the third element of this vector, when `some_letters[3] != "c"` is obviously false?

Recycling

When you use `!=` or `==`, R tries to compare each element of the left argument with the corresponding element of its right argument. Then what happens when you compare vectors of different lengths?

When one vector is shorter than the other, it gets *recycled*:

In this case, R **repeats** `c("a", "c")` as many times as necessary to match the length of the `some_letters` vector. So, we get `c("a", "c", "a", "c", "a")`. Since the recycled `"a"` doesn't match the third element of `some_letters`, the value of `!=` is in fact `TRUE`.

Note: We got lucky here! R output an error because the length of our vectors was not a constant multiple (e.g. 2 times as long). If they were, R would have carried out the **same** recycling procedure, but **would not** have output an error!

This is why the inclusion (`%in%`) operator is so great! It does carry out the procedure we wish for R to do, without any silly recycling! To exclude values you place a `!` in front of the **entire** statement, not directly in front of the `%in%`.

```
!c("a", "b") %in% some_letters
```

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

Logical Values

You can also use logical values (`TRUE`, `FALSE`) to extract elements of a vector.

```
some_letters[c(TRUE, TRUE, FALSE, FALSE, FALSE)]
```

```
## [1] "a" "b"
```

```
## Extracts the first two elements of the some_letters vector
```

This should give you an intuition as to how we might be able to use the results of relational statements to extract the elements of the vector that satisfy the relation.

Which

The `which` statement returns the *indices* of a vector, where a relational statement is satisfied (evaluates to `TRUE`).

```
x_2 <- c(3, 5, 7, 9, 11, 13, 15)

which(x_2 > 8)
```

```
## [1] 4 5 6 7
```

```
which(x_2 == 7)
```

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

```
## Mini-challenge: How would you use the indices from these 'which' statements
## to extract the elements of x_2 that meet the criteria?
```

Matrices

The above relational and which statements can be applied to a matrix, or to a subset of the matrix (e.g., a vector). The relational and `which` statements are applied element wise (a step-by-step progression through the matrix/vector entries).

```r
# The following are dating data, of one person's messages received per day for
# one week

messages <- data.frame(okcupid = c(16, 9, 13, 5, 2, 17, 14), match = c(17, 7, 5,
    16, 8, 13, 14))
# This makes the data from OkCupid the first column and the data from Match the
# second column

row.names(messages) <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday",
    "Saturday", "Sunday")
```

Challenge 5

Use the `messages` matrix to return a matrix of *logical* values which answer the following question:

1. For what days were the number of messages at either site greater than 12?

```r
# The following are dating data, of one person's messages received per day for
# one week

messages <- data.frame(okcupid = c(16, 9, 13, 5, 2, 17, 14), match = c(17, 7, 5,
    16, 8, 13, 14))
# This makes the data from OkCupid the first column and the data from Match the
# second column

row.names(messages) <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday",
    "Saturday", "Sunday")
```

```r
# Create a logical matrix to provide information that would allow you to answer
# the question: For what days were the number of messages at either site
# greater than 12?
```

```r
messages > 12
```

```
##           okcupid match
## Monday       TRUE  TRUE
## Tuesday     FALSE FALSE
## Wednesday    TRUE FALSE
## Thursday    FALSE  TRUE
## Friday      FALSE FALSE
## Saturday     TRUE  TRUE
## Sunday       TRUE  TRUE
```

Challenge 6

Use the messages matrix to return the *rows* of `messages` which answer the following questions (Use rows of the matrix to answer):

1. when were the messages at OkCupid equal to 13?
2. when were the messages at OkCupid greater than Match?

```
# Use rows of the matrix to answer:

# when were the messages at OkCupid equal to 13?


# when were the messages at OkCupid greater than Match?
```

```
messages[which(messages[, 1] == 13), ]
```

```
##           okcupid match
## Wednesday      13     5
```

```
messages[which(messages[, 1] > messages[, 2]), ]
```

```
##           okcupid match
## Tuesday         9     7
## Wednesday      13     5
## Saturday       17    13
```

Subset

The `subset` command takes in an object (vector, matrix, data frame) and returns the subset of that object where the entries meet the relational conditions specified (evaluates to `TRUE`).

```
x_3 <- c(3, 5, 7, 9, 11, 13, 15)

subset(x_3, x_3 > 6)
```

```
## [1]  7  9 11 13 15
```

Challenge 7

Using the okcupid "data" from above, answer the following question:

1. Change the which() statement code to a subset() statement, extracting the days that the number of messages at OkCupid greater than the messages at Match.

```
# Change the which() statement code to a subset() statement, extracting the
# days that the number of messages at OkCupid greater than the messages at
# Match
```

```
subset(messages, messages[, 1] > messages[, 2])
```

```
##           okcupid match
## Tuesday         9     7
## Wednesday      13     5
## Saturday       17    13
```

Logicals

These statements allow for us to change or combine the results of the relational statements we discussed before, using **and**, **or**, or **not**.

"and" statements (`&`)

These statements evaluate to `TRUE` only if ***every*** relational statement evaluates to `TRUE`.

- `(3 < 5) & (9 > 7)` would evaluate to `TRUE` because ***both*** relational statements are `TRUE`

- `(3 > 5) & (9 > 7)`, this would evaluate to `FALSE` as only one of the relational statements is `FALSE` (the first one)

"or" statements (`|`)

These statements evaluate to `TRUE` if ***at least one*** relational statement evaluates to `TRUE`.

- `(3 > 5) | (9 > 7)` would evaluate to `TRUE` because ***one*** of the relational statements is `TRUE` (the second one)

- `(3 > 5) | (9 < 7)` would evaluate to `FALSE` as ***both*** relational statements evaluate to `FALSE`

"not" statements (`!`)

These statements convert (negate) the statement it proceeds, changing `TRUE` to `FALSE` and `FALSE` to `TRUE`.

- `is.numeric(5)` would evaluate to `TRUE` because 5 is a number

- `!is.numeric(5)` would evaluate to `FALSE` as it negates the statement it proceeds (`!TRUE = FALSE`)

  NOTE:

  The `&&` and `||` logical statements ***do not*** evaluate the same as their single counterparts. Instead, these logical operators evaluate to `TRUE` or `FALSE` based ***only*** on the first element of the statement, vector, or matrix.

Below are examples of using logicals to evaluate vectors. For example, in the first line of code three statements are checked, `TRUE & TRUE`, `TRUE & FALSE`, and `FALSE & FALSE`. The first statement is "TRUE" since both elements were "TRUE". The second and third statements are "FALSE" since at least one element was "FALSE".

```r
c(TRUE, TRUE, FALSE) & c(TRUE, FALSE, FALSE)
```

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

```r
c(TRUE, TRUE, FALSE) | c(TRUE, FALSE, FALSE)
```

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

```r
c(TRUE, TRUE, FALSE) && c(TRUE, FALSE, FALSE)
```

```
## Warning in c(TRUE, TRUE, FALSE) && c(TRUE, FALSE, FALSE): 'length(x) = 3 > 1'
## in coercion to 'logical(1)'

## Warning in c(TRUE, TRUE, FALSE) && c(TRUE, FALSE, FALSE): 'length(x) = 3 > 1'
## in coercion to 'logical(1)'
```

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

Challenge 8

Using the okcupid vector from above, answer the following questions:
1. Is the last day of the week under 5 messages or above 10 messages?
(HINT: `last <- tail(messages$okcupid, n=1)` could be helpful)
2. Is the last day of the week between 15 and 20 messages, excluding 15 but including 20?

```
# Is the last day of the week under 5 messages or above 10 messages?  (hint:
# last <- tail(messages$okcupid, n=1) could be helpful)

# Is the last day of the week between 15 and 20 messages, excluding 15 but
# including 20?

# Make sure you test your code with some other values
```

```
last <- tail(messages$okcupid, 1)

last < 5 | last > 10
```

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

```
last > 15 & last <= 20  #Or last > 15 & (last < 20 | last == 20)
```

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

The `subset` command (from before) can also accept more than one relational condition if joined by logicals.

```
bad_days <- subset(messages, okcupid < 6 | match < 6)

bad_days
```

```
##           okcupid match
## Wednesday      13     5
## Thursday        5    16
## Friday          2     8
```

```
good_days <- subset(messages, okcupid > 10 & match > 10)

good_days
```

```
##          okcupid match
## Monday        16    17
## Saturday      17    13
## Sunday        14    14
```

Conditional Statements

Conditional statements utilize relational and logical statements to change the results of your `R` code. You may have encountered an `if else` statement before (or not), but let's breakdown exactly what `R` is doing when it evaluates them.

If Statements

First, let's start with an `if` statement, the often overlooked building block of the `if else` statement. The `if` statement is structured as follows:

```
if(condition){
              statement
}
```

- the condition inside the parentheses (a relational statement) is what the computer executes to check its logical value,
  - if the condition evaluates to `TRUE` then the statement inside the curly braces (`{}`) is output, and
  - if the condition is `FALSE` nothing is output.

NOTE: In `R` the `if` statement, as described above, will only accept a **single** value (not a vector or matrix).

```r
y <- -3  ## Try changing this number, for example y <- 5

if (y < 0) {
    "y is a negative number"
}
```

```
## [1] "y is a negative number"
```

Challenge 9

Using the `last` number from Challenge 8, write an `if` statement that prints "You're popular!" if the number of messages from okcupid exceeds 10.

```r
# hint: use the last day of the week for okcupid that you made above
last <- tail(messages$okcupid, 1)
```

```r
last <- tail(messages$okcupid, 1)

if (last > 10) {
    "You're popular!"
}
```

```
## [1] "You're popular!"
```

If/Else Statements

Since whenever an `if` statement evaluates to `FALSE` nothing is output, you might see why an `else` statement could be beneficial! An `else` statement allows for a different statement to be output whenever the `if` condition evaluates to `FALSE`. The `if else` statement is structured as follows:

```
if(condition){
              statement1
}
else{
              statement2
}
```

- again, the `if` condition is executed first,
  - if it evaluates to `TRUE` then the first statement (`statement1`) is output,
  - if the condition is `FALSE` the computer moves on to the `else` statement, and the second statement (`statement2`) is output.

NOTE: In `R` the `if else` statement, as described above, will only accept a **single** value (not a vector or matrix).

```r
y <- -3  ## Try changing this number, for example y <- 5

if (y < 0) {
    "y is a negative number"
} else {
    "y is either positive or zero"
}
```

```
## [1] "y is a negative number"
```

NOTE: `R` accepts both `if else` statements structured as outlined above, but also `if else` statements using the built-in `ifelse()` function. This function accepts **both singular and vector** inputs and is structured as follows:

```
ifelse(condition, statement1, statement2)
```

where the first argument is the conditional (relational) statement, the second argument is the statement that is evaluated when the condition is `TRUE` (`statement1`), and the third statement is the statement that is evaluated when the condition is `FALSE` (`statement2`).

```r
y <- -3  ## Try changing this number, for example y <- 5

ifelse(y < 0, "y is a negative number", "y is either positive or zero")
```

```
## [1] "y is a negative number"
```

Challenge 10

Using the "if" statement from Challenge 9, add the following else statement: When the "if"-condition on messages is not met, R prints out "Send more messages!".

```r
# Challenge: Rewrite the if, else function from above using R's built in
# ifelse() function.
last <- tail(messages$okcupid, 1)
```

```r
last <- tail(messages$okcupid, 1)

ifelse(last > 10, "You're popular!", "Send more messages!")
```

```
## [1] "You're popular!"
```

Else/If Statements

On occasion, you may want to add a third (or fourth, or fifth, ...) condition to your `if else` statement, which is where the `else if` statement comes in. The `else if` statement is added to the `if else` as follows:

```
if(condition1){
              statement1
}
else if(condition2){
              statement2
}
else{
              statement3
}
```

- The `if` condition (`condition1`) is executed first,
  - if it evaluates to `TRUE` then the first statement (`statement1`) is output,
  - if the condition is `FALSE` the computer moves on to the `else if` condition,
- Now the second condition (`condition2`) is executed,
  - if it evaluates to `TRUE` then the second statement (`statement2`) is output,
  - if the condition is `FALSE` the computer moves on to the `else` statement, and
- the third statement (`statement3`) is output.

```r
y <- -3  ## Try changing this number, for example y <- 5 and y <- 0

if (y < 0) {
    "y is a negative number"
} else if (y == 0) {
    "y is zero"
} else {
    "y is positive"
}
```

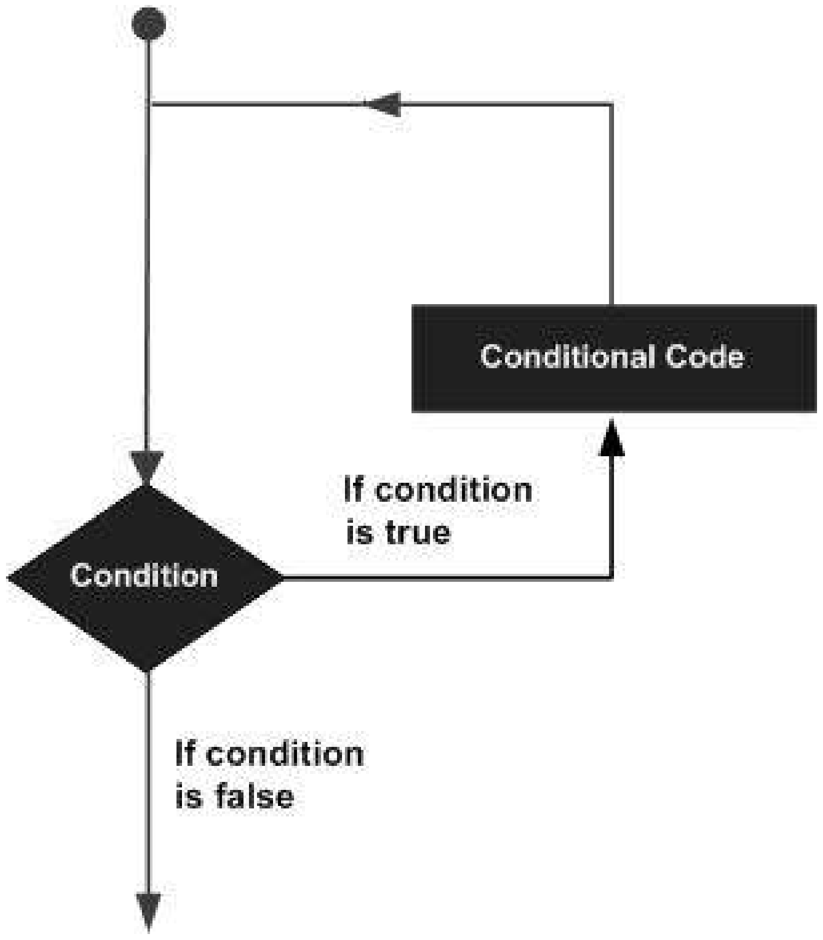
```
## [1] "y is a negative number"
```

> NOTE: Conditional statements should be written so that the components are mutually exclusive (an observation can only belong to one piece).
>
> The modulo (`%%`) returns the remainder of the division of the number to the left by the number on the right, for example 5 modulo 3 or 5 %% 3 is 2.

```r
x <- 6

if (x%%2 == 0) {
    "x is divisible by 2"
} else if (x%%3 == 0) {
    "x is divisible by 3"
} else {
    "x is not divisible by 2 or 3"
}
```

```
## [1] "x is divisible by 2"
```

Loops



Loops are a popular way to iterate or replicate the same set of instructions many times. It is no more than creating an automated process by organizing a sequence of steps in your code that need to be repeated. In general, the advice of many `R` users would be to learn about loops, but once you have a clear understanding of them to instead use vectorization, when your current iteration doesn't depend on the previous iteration.

Loops will give you a detailed view of the process that is happening and how it relates to the data you are manipulating. Once you have this understanding, you should put your effort into learning about vectorized alternatives as they pay off in efficiency. These loop alternatives (the `apply` and `purrr` families) will be covered in a later workshop.

Typically in data science, we separate loops into two types. The loops that execute a process a specified number of times, where the "index" or "counter" is incremented after each cycle are part of the `for` loop family. Other loops that only repeat themselves until a conditional statement is evaluated to be `FALSE` are part of the `while` loop family.

For Loops

In the `for` loop figure above:

- The starting point (black dot) represents the initialization of the object(s) being used in the `for` loop (i.e., the variables). `R` requires you to initialize the objects you will use in your loop **before** you use them, it does not automatically create the objects for you!
- The diamond shapes represent the repeat condition the computer is required to evaluate. The computer evaluates the conditional statement (`i in sequence`) as either `TRUE` or `FALSE`. In other terms, you are

testing if the current value of `i` (the index/counter) is within the specified range of values (`sequence`), where this range is either defined in the initialization or directly within the condition statement (like `1:100`).

- The rectangle represents the set of instructions to execute for every iteration. This could be a simple statement, a block of instructions, or another loop (nested loops).

The computer marches through this process until the repeat decision (condition) evaluates to `FALSE` (`i` is not in `sequence`).

Notation The `for` loop repeat statement is placed between parentheses after the `for` statement. Directly after the repeat statement, in curly braces, are the instructions that are to be repeated. You will find the formatting that you like best for things such as loops and functions, but here is our preferred syntax:

```
num <- 100
for(i in 1:num){
              statement
}
```

Alternatively, you could use:

```
sequence <- 1:100

for(i in sequence){
              statement
}
```

```
for (i in 1:10) {
    print(i)
}
```

Example:

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
```

```
i  ## should be 10 (the last number of the sequence 1:10), unless the loop had issues!
```

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

Now, let's add the corresponding entries of two vectors together!

```
x <- seq(1, 10, by = 1)
y <- seq(10, 28, by = 2)

## create an empty vector to store added quantities
z <- rep(NA, length(x))


for (i in 1:10) {
```

```r
    # write a loop to store the entries of x added to the entries of y in z

}

z  ## examine the resulting vector
```

```r
x <- seq(1, 10, by = 1)
y <- seq(10, 28, by = 2)
```

Challenge 11

How could the above calculation be done outside of a loop?

```r
# code to execute adding of the x and y vectors from the example above, NOT in
# a loop
```

```r
z <- x + y
z
```

```
##  [1] 11 14 17 20 23 26 29 32 35 38
```

NOTE: Like we used in the above `for` loop, it is often useful, to add a `print()` statement inside the loop. This allows for you to verify that the process is executing correctly and can save you some major headaches!

Challenge 12

Recursive `for-loop`!

Read in the BlackfootFish dataset and modify the code to write a `for-loop` to find the indices needed to sample every 7^{th} row from the dataset, starting with the 1^{st} row, until you've sampled 1200 rows. This would allow us to split the data set into "testing" and "training" samples, possibly using the `testing` data for assessing a model we develop using the "training" data.

```r
BlackfootFish <- read.csv("https://raw.githubusercontent.com/saramannheimer/data-science-r-workshops/ma

n_testing <- 1200

samps <- rep(NA, n_testing)
## Initializing the samps vector, for storing indicies

samps[1] <- 1
## Setting first sample index to 1 (first row)

# Code snippet:
# Create the code for the process will be executed at each step
# e.g., How will you get the next sample after 1?

for(i in 2:#ending index here){
  samps[i] <- # process you execute at every index
}

testing <- BlackfootFish[samps, ]
```

```
training <- BlackfootFish[-samps, ]
```

```
BlackfootFish <- read.csv("https://raw.githubusercontent.com/saramannheimer/data-science-r-workshops/ma
    header = TRUE)

n_testing <- 1200

samps <- rep(NA, n_testing)
## Initializing the samps vector, for storing indices

samps[1] <- 1
## Setting first sample index to 1 (first row)

# Code snippet: Create the code for the process will be executed at each step
# e.g., How will you get the next sample after 1?

for (i in 2:n_testing) {
    samps[i] <- samps[i - 1] + 7
}

testing <- BlackfootFish[samps, ]
training <- BlackfootFish[-samps, ]
```

Before we move on, is there a better way to split a data set into training and testing versions than taking every 7th observation?

Nesting For Loops The section title suggest that `for` loops can also be nested, you're probably wondering how this would look.

In nested `for` loops, you have two indices keeping track of where you are in the loop. The outer loop will have an index, say `i`, and the inner loop will have an index, say `j`. Let's look at the nested `for` loop below and see if we can figure out how the loop proceeds through the indices.

```
for (i in 1:5) {
    for (j in c("a", "b", "c", "d", "e")) {
        print(paste(i, j))
    }
}
```

```
## [1] "1 a"
## [1] "1 b"
## [1] "1 c"
## [1] "1 d"
## [1] "1 e"
## [1] "2 a"
## [1] "2 b"
## [1] "2 c"
## [1] "2 d"
## [1] "2 e"
## [1] "3 a"
## [1] "3 b"
## [1] "3 c"
## [1] "3 d"
## [1] "3 e"
## [1] "4 a"
## [1] "4 b"
## [1] "4 c"
## [1] "4 d"
## [1] "4 e"
## [1] "5 a"
## [1] "5 b"
## [1] "5 c"
## [1] "5 d"
## [1] "5 e"
```

> QUESTION:
>
> *When would you possibly use this in your code?*

Well, suppose you wish to manipulate a matrix by setting its elements to specific values, based on their row and column position. You will need to use nested `for` loops in order to assign each of the matrix's entries a value.

- The index `i` runs over the rows of the matrix
- The index `j` runs over the columns of the matrix

> *But when should you use a loop? Couldn't you just repeat the set of instructions the number of times you want to do it?*

A rule of thumb is that if you need to perform the same action in one place of your code three or more times, then you are better served by using a loop. A loop makes your code more compact, readable, maintainable, and saves you some typing! If you have the same process ***multiple*** places in your code, that's where functions come in handy!

Functions

What is a function? In rough terms, a function is a set of instructions that you would like repeated (similar to a loop), but are more efficient to be self-contained in a sub-program and called upon when needed. A function is a piece of code that carries out a specific task, accepting arguments (inputs), and returning an output.

Functions gather a sequence of operations into a whole, preserving it for ongoing use. Functions provide:

- a name we can remember and invoke it by,
- relief from the need to remember individual operations,
- a defined set of inputs and expected outputs, and
- rich connections to the larger programming environment.

As the basic building block of most programming languages, user-defined functions constitute "programming" as much as any single abstraction can. If you have written a function, you are a computer programmer!

At this point, you may be familiar with a smattering of `R` functions in a few different packages (e.g., `mean`, `table`, `lm`, `glm`, `str`, `tapply`, etc.). Some of these functions take multiple arguments and return a single output, but the best way to learn about the inner workings of functions is to write your own!

User Defined Functions

A function allows for us to repeat several operations with a single command. Take for instance a function which converts heights from feet and inches to centimeters, named `feet_inch_to_cm()`.

```r
feet_inch_to_cm <- function(feet, inches) {
    inches <- feet * 12 + inches
    cm <- inches * 2.54

    return(cm)
}
```

```r
feet_inch_to_cm <- function(feet, inches) {
    inches <- feet * 12 + inches
    cm <- inches * 2.54

    return(cm)
}
```

Challenge 13

Use the `feet_inch_to_cm` function to convert your height to cm!

```r
## Your code goes here. NOTE, the solution will use a height of 5' 7'
```

```r
feet_inch_to_cm(5, 7)
```

```
## [1] 170.18
```

We define `feet_inch_to_cm()` by assigning it to the output of function. The list of argument names are contained within parentheses. Next, the body of the function (the statements that are executed when it runs) is contained within the curly braces (`{}`). The statements in the body are indented by two spaces. This makes the code easier to read but does not affect how the code operates.

It is useful to think of creating functions like writing a cookbook.

1. Define the "ingredients" that your function needs. In this case, we only need two ingredients to use our function: "feet" and "inches".
2. State what we to do with the ingredients. In this case, we are taking our ingredient and applying a set of mathematical operators to it.
3. Declare what the final product (output) of your function will be. This typically lives in a `return()` statement, but we will see other methods to use.

When we call the function, the values we pass to it as arguments are assigned to those variables so that we can use them inside the function. Inside the function, we use a return statement to send a result back to whoever asked for it.

Function Scoping

When writing and using functions, you should be careful to consider where each of the variables you are using are defined. Variables that are defined *within* a function belong to a local environment, and are not accessible from outside of the function. Take this example:

```r
x <- 10
## globally defined variable

f <- function() {
    ## locally defined x and y
    x <- 1
    y <- 2
    c(x, y)
}

f()
```

```
## [1] 1 2
```

```r
x
```

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

We notice that the globally defined value of `x` is not called inside the function, as the function defines `x` within itself, so it does not search for a value of `x`! Now look at another example, where `x` is ***not*** defined within the function.

```r
g <- function() {
    ## locally defined y
    y <- 2
    c(x, y)
}

x <- 15

g()
```

```
## [1] 15  2
```

```r
x <- 20

g()
```

```
## [1] 20  2
```

Summary

1. When you call a function, a new environment is created for the function to do its work in.

2. The new environment is populated with the argument values passed in to the function.

3. Objects are looked for *first* in the function environment.

4. If objects are not found in the function environment, they are then looked for in the global environment.

- This scoping issue emphasizes the need for using good global variable and data frame names that are informative and explicit, not reused, and also not function or typical argument names.

Why Write a Function?

Similar to loops, you will often find that you have performed the same task multiple times. A good rule of thumb is that if you've copied and pasted the same code twice (so you have three copies), you should write a function instead!

Example of Scaling a Variable without a Function If we wanted to rescale every quantitative variable in a dataset so that the variables have values between 0 and 1. We could use the following formula:

$$y_{scaled} = \frac{y_i - min\{y_1, y_2, ..., y_n\}}{max\{y_1, y_2, ..., y_n\} - min\{y_1, y_2, ..., y_n\}}$$

The following R code would carry out this rescaling procedure for the `length` and `weight` columns of the `BlackfootFish` data:

```r
BlackfootFish$length_scaled <- (BlackfootFish$length - min(BlackfootFish$length,
    na.rm = TRUE))/(max(BlackfootFish$length, na.rm = TRUE) - min(BlackfootFish$length,
    na.rm = TRUE))

BlackfootFish$weight_scaled <- (BlackfootFish$weight - min(BlackfootFish$weight,
    na.rm = TRUE))/(max(BlackfootFish$weight, na.rm = TRUE) - min(BlackfootFish$length,
    na.rm = TRUE))
```

We could continue to copy and paste for other columns of the data, but you can probably get the idea. This process of duplicating an action multiple times makes it difficult to understand the intent of the process. Additionally, it makes it very difficult to spot the mistakes.

> QUESTION:
>
> *Did you spot the mistake in the weight conversion?*

Often you will find yourself in the position of needing to find a function that performs a specific task, but you do not know of a function or a library that would help you. You could spend time Googling for a solution, but in the amount of time it takes you to find something, it is possible that you could have already written your own function!

Example of Scaling a Variable *with* a Function Let's transform the repeated process above into a rescaling function.

The following snippet of code rescales the `length` column to be between 0 and 1:

```r
head(BlackfootFish$length - min(BlackfootFish$length, na.rm = TRUE))/(max(BlackfootFish$length,
    na.rm = TRUE) - min(BlackfootFish$length, na.rm = TRUE))
```

```
## [1] 0.2226804 0.1979381 0.1525773 0.2103093 0.3288660 0.1309278
```

Our goal is to turn this snippet of code into a general rescale function that we can apply to any numeric vector.

- The first step is to examine the process to determine how many inputs there are.
- The second step is to change the code snippet to instead refer to these inputs using temporary variables.

For this process there is one input, the numeric vector to be rescaled (currently `BlackfootFish$length`). We instead want to refer to this input using a temporary variable. It is common (in R) to refer to a generic vector of data as `x`.

Challenge 14

1. Modify the code snippet above to instead refer to a temporary variable `x`, make sure your code does not depend on a specific dataset.
2. Save the rescaled vector in a variable called `x_rescaled`

```r
# code modification with temporary variable 'x'
x <- BlackfootFish$length
```

```r
x <- BlackfootFish$length
x_rescaled <- (x - min(x, na.rm = TRUE))/(max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
```

But now take a closer look at our process. Is there any duplication? An obvious duplicated statement is `min(x, na.rm = TRUE)`. It would make more sense to calculate the minimum of the data once, store the result, and then refer to it when needed. We also notice that we use the maximum value of `x`, so we could instead calculate the range of `x` and refer to the first (minimum) and second (maximum) elements when they are needed.

Note: The `range` function in `R` takes a single numeric input and returns the minimum and maximum (so this is not the "range" statistic but the information to calculate it):

```r
range(seq(1, 5))
```

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

What should we call this variable containing the range of `x`? Some important advice on naming variables in `R`:

- Make sure that the name you choose for your variable ***is not*** a reserved word in `R` (`range`, `mean`, `sd`, `hist`, etc.).
    - A way to avoid this is to use the help directory to see if that name already exists (`?functionName`).
- It is possible to overwrite the name of an existing variable, but it is ***not*** recommended!
- Variable names cannot begin with a number.
- Keep in mind that capitalization matters in `R`!

We will call this intermediate variable `x_range` (for range of the x variable).

Challenge 15

Create another intermediate variable called `x_range` that contains the range of the intermediate variable `x`, using the `range()` function. Make sure that you specify the `na.rm` option to ignore any NAs in the input vector.

```r
x <- BlackfootFish$length
# Create x_range = the range of x


# Rewrite the code snippet from Exercise 14 to now refer to x_range instead of
# min(x) and max(x)
```

```r
x <- BlackfootFish$length

x_range <- range(x, na.rm = TRUE)
x_rescaled <- (x - x_range[1])/(x_range[2] - x_range[1])
```

How does this process end up with us writing our own function? What do you need in order to write a function?

- The task the function will solve (what you've been copying and pasting) **and**
- The inputs of the function (the intermediate variables)

We now have all these pieces ready to put together. It's time to write the function!

In `R` the function you define will have the following construction:

```
myFunction <- function(argument1, argument2,...){
                body
}
```

Challenge 16

Using the template above, write a function named `rescale` that rescales a vector to be between 0 and 1. The function should take a single argument, `x`.

```
# Use the function template to create a function named rescale that performs
# the process outlined above, using the intermediate variables you previously
# defined

myfunction <- function(arg1, ...) {
    # body
}
```

```
rescale <- function(x) {
    x_range <- range(x, na.rm = TRUE)
    x_rescaled <- (x - x_range[1])/(x_range[2] - x_range[1])
    return(x_rescaled)
}
```

Once you have written your function, the next step is to test it out. The simplest place to start is to set value(s) for the function's argument(s) (e.g., `x <- c(1, 2, 3, 4, 5)`). You can then run the code *inside* the function, line by line, to make sure that each line successfully executes for this variable. This is what we call unit testing in data science. This is a similar process to testing out a `for` loop, by setting the value of the index (`i <- 1`) and running the inside of the loop.

Challenge 17

Unit test your function on a simple vector, with the same name as your function's input. What do you expect the values of the test to be after you input it into your function?

```
rescale <- function(x) {
    x_range <- range(x, na.rm = TRUE)
    x_rescaled <- (x - x_range[1])/(x_range[2] - x_range[1])
    return(x_rescaled)
}
```

```
# Copy the body code from the function to unit test the code in your function
# on a simple vector named x <- c(1, 2, 3, 4, 5)
```

```
x <- c(1, 2, 3, 4, 5)

x_range <- range(x, na.rm = TRUE)
x_rescaled <- (x - x_range[1])/(x_range[2] - x_range[1])

x_rescaled
```

```
## [1] 0.00 0.25 0.50 0.75 1.00
```

If your unit test was successful, now you can move on to testing your function on the test data and then with your real data.

Challenge 18

Test your function on the `length` column of the BlackfootFish dataset, saving the results as `BlackfootFish$length_scaled`. Inspect the rescaled values and run `summary()` on the new variable. Do they look correct?

```
# Test your function out on the BlackfootFish data's length variable
```

```
BlackfootFish$length_scaled <- rescale(BlackfootFish$length)
summary(BlackfootFish$length_scaled)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1763  0.2412  0.2542  0.3216  1.0000
```

Process of Creating a Function

You should **never** start writing a function with the function template. Instead, you start with a problem that you need to solve and work through the following steps:

1. Define the problem you need to solve.
2. Get a working snippet of code that solves the simple problem. (*Where you start when you're copying and pasting*)
3. Rewrite the code snippet to use temporary variables.
4. Rewrite the code for clarity (remove redundancy or multiple calculations of the same value).
5. Turn everything into a function! The code snippet forms the body of the function, the temporary variables are the arguments, and you choose the name of your function.
6. Test your function! Start with a simple example and then go big!

Challenge 19 (Part 1)

Putting it All Together! Start by creating a data subset with just the species, length, and weight variables from the BlackfootFish dataset. Call the data subset `BlackfootFish_subset`.

```
BlackfootFish_subset <- data.frame(species = BlackfootFish$species, length = BlackfootFish$length,
    weight = BlackfootFish$weight)
```

Challenge 19 (Part 2)

Now, given the matrix of species, length, and weights above, use the tools from the workshop (`for` loop, function, conditional, and/or relational statements) to:

- compute Fulton's condition factor of each fish ($condition = \frac{100000*weight}{length^3}$), making any fish with weight or lengths of 0s to be NAs on the condition factor since those are obvious data entry errors.

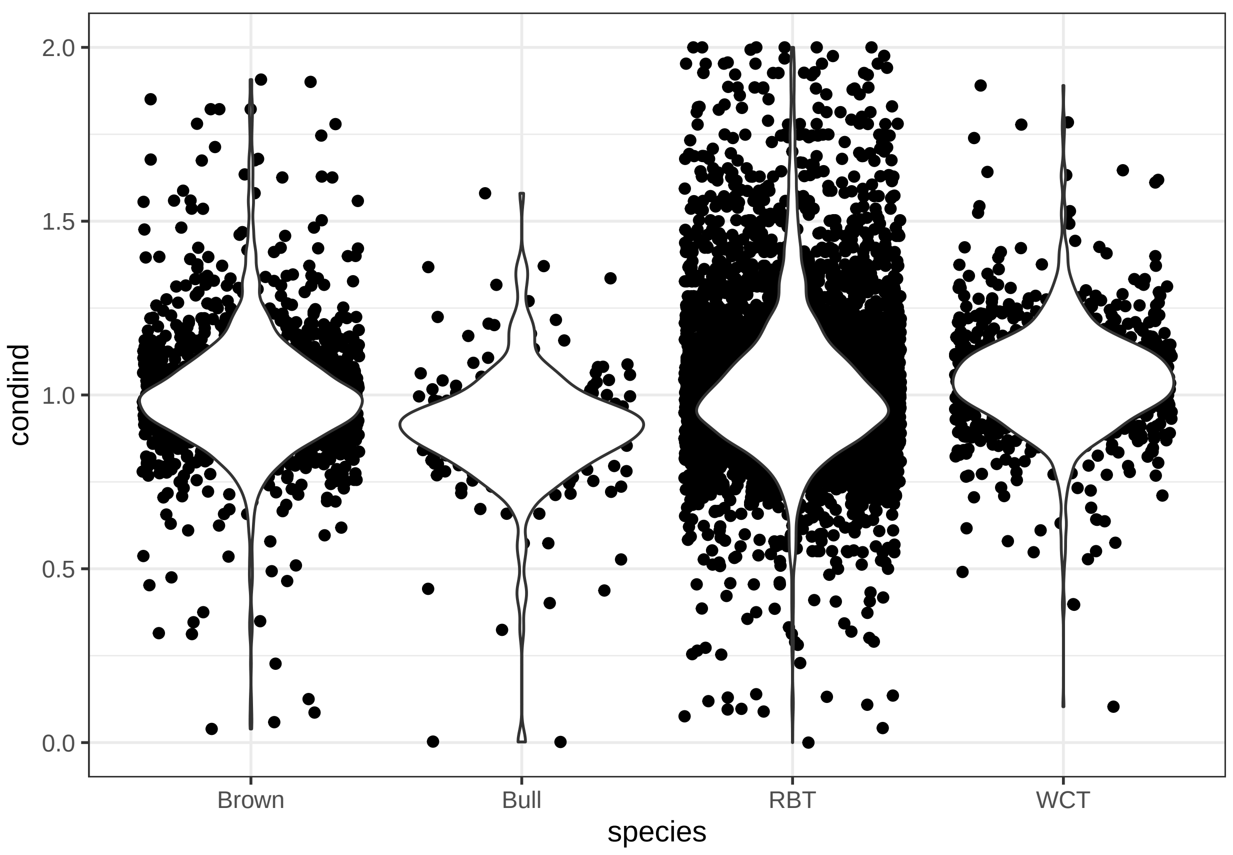
- then remove the fish from the dataset whose condition index is NA or more than 2 (values larger than 2 are not typical and might suggest further data entry errors).
- If you attended our data visualization workshop, try to make a plot of condition factors based on the fish species with jittered points and violins.

```r
condition_index <- function(x, y) {
    ifelse(x == 0 | y == 0, c_in <- NA, c_in <- 1e+05 * (x/(y^3)))

    return(c_in)
}

BlackfootFish_subset <- data.frame(species = BlackfootFish$species, length = BlackfootFish$length,
    weight = BlackfootFish$weight)
```

```r
condition_index <- function(x, y) {
    ifelse(x == 0 | y == 0, c_in <- NA, c_in <- 1e+05 * (x/(y^3)))

    return(c_in)
}


BlackfootFish_subset$condind <- condition_index(BlackfootFish_subset$weight, BlackfootFish_subset$lengt

summary(BlackfootFish_subset$condind)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##     0.000     0.910     1.000     9.936     1.112 19411.084       864
```

```r
D1 <- subset(BlackfootFish_subset, !is.na(condind) & condind <= 2)

library(ggplot2)

D1 %>%
    ggplot(aes(x = species, y = condind)) + geom_jitter() + geom_violin() + theme_bw()
```



References

R for Data Science (Grolemund & Wickham, 2017)

- Functions: https://r4ds.had.co.nz/functions.html
- Iteration: https://r4ds.had.co.nz/iteration.html

Advanced R (Wickham, 2015)

- Subsetting: https://adv-r.hadley.nz/subsetting.html
- Control Flow: https://adv-r.hadley.nz/control-flow.html
- Functions: https://adv-r.hadley.nz/functions.html
- Conditionals: https://adv-r.hadley.nz/conditions.html
- Environments: https://adv-r.hadley.nz/environments.html

Suggestions for your own work

The goal of this workshop was to teach you to write code in `R` to learn how to use relational operators, create conditional sequences, and write loops and functions. The first workshop in our series contains more information on how to get started working in `R` using RStudio (see http://www.montana.edu/datascience/training/). The second workshop in our series contains information on how to write code to visualize data using `ggplot2`. The codechunks in this interactive document mimic the codechunks you can use on your own projects in RMarkdown but you will need to download and install both `R` and RStudio on your own computer.

Montana State University R Workshops Team

These materials were adapted from materials generated by the Data Carpentries (https://datacarpentry.org/) and were originally developed at MSU by Dr. Allison Theobold. The workshop series is co-organized by the Montana State University Library, Department of Mathematical Sciences, and Statistical Consulting and Research Services (SCRS, https://www.montana.edu/statisticalconsulting/). SCRS is supported by Montana INBRE (National Institutes of Health, Institute of General Medical Sciences Grant Number P20GM103474). The workshops for 2021-2022 are supported by Faculty Excellence Grants from MSU's Center for Faculty Excellence.

Research related to the development of these workshops appeared in:

- Allison S. Theobold, Stacey A. Hancock & Sara Mannheimer (2021) Designing Data Science Workshops for Data-Intensive Environmental Science Research, *Journal of Statistics and Data Science Education*, 29:sup1, S83-S94, DOI: 10.1080/10691898.2020.1854636

The workshops for 2023-2024 involve modifications of materials and are being taught by:

Greta Linse

- Greta Linse is the Interim Director of Statistical Consulting and Research Services (https://www.montana.edu/statisticalconsulting/) and the Project Manager for the Human Ecology Learning and Problem Solving (HELPS) Lab (https://helpslab.montana.edu). Greta has been teaching, documenting and working with statistical software including R and RStudio for over 10 years.

Sara Mannheimer

- Sara Mannheimer is an Associate Professor and Data Librarian at Montana State University, where she helps shape practices and theories for curation, publication, and preservation of data. Her research examines the social, ethical, and technical issues of a data-driven world. She is the project lead for the MSU Dataset Search, and she is working on a book about data curation to support responsible qualitative data reuse and big social research.

The materials have also been modified and improved by:

- Dr. Mark Greenwood
- Harley Clifton
- Eliot Liucci
- Dr. Allison Theobold