Introduction to R

September, 2024

Table of Contents

# 1 Course Logistics and Expectations

## 1.1 Course Description

In this course, you will learn the fundamentals of R, a statistical programming language, and use it to wrangle data for analysis and visualization. The programming skills you will learn are transferable to learn more about R independently and other high-level languages such as Python. At the end of the class, you will be reproducing analysis from a scientific publication!

## 1.2 Learning Objectives

After taking this course, you will be able to:

* **Analyze** Tidy datasets in the R programming language via data wrangling, summary statistics, and visualization.
* **Describe** how the R programming environment interpret complex expressions made out of functions, operations, and data structures, in a step-by-step way.
* **Apply** problem solving strategies to debug broken code.

## 1.3 Course Website

All course information will be available here:

<https://hutchdatascience.org/Intro_to_R>

Course discussions will be done in the class slack Workspace. Invites will be sent before class.

Lab Assignments will be done in the class Posit.cloud workspace. Students should register at <https://posit.cloud> before the lab. Link to join the workspace will be sent out before the first lab.

## 1.4 DaSL Courses are a Psychologically Safe Space

We want everyone to feel ok with asking questions. That’s why we adhere to the [Participation Guidelines](https://hutchdatascience.org/communitystudios/guidelines/) for each course.

Be respectful of each other and how we learn differently. It is never ok to disparage people for their questions.

## 1.5 Office Hours

Office Hours will be held via Teams on Fridays. Feel free to drop into the office hours and work and ask questions as needed.

## 1.6 Clinical Network Issues

We know that learners on the Clinical network are having issues accessing material, including websites. We are working to figure out good workarounds for it.

If you are connected via VPN, we recommend that you disconnect it while working.

## 1.7 Slack

If you haven’t yet joined the FH Data Slack, you can join here: <https://hutchdatascience.org/joinslack/>

Look for the #dasl-s4-intro-to-r channel - that’s where we’ll have conversations and field questions.

## 1.8 Instructor

[Ted Laderas, PhD](https://laderast.github.io)  
[tladera2@fredhutch.org](mailto:tladera2@fredhutch.org)  
**Preferred Method of Contact**: Email/Slack  
**Expected Response Time**: 24hrs

I’ve been teaching R for over 10 years, and have been an active user and data scientist for over 20. I write a lot, including on Data Science, Mental Health, and Bioinformatics.

I’m always excited to see my learners surpass me, and if you are curious enough, I guarantee you will.

## 1.9 Words of Encouragement

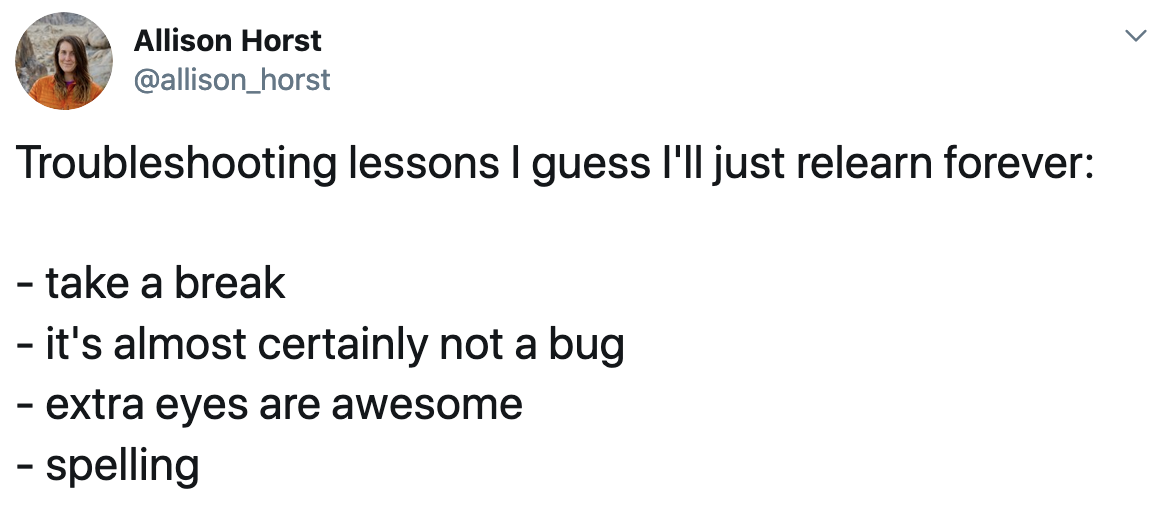
This was adopted from Andrew Heiss. Thanks!

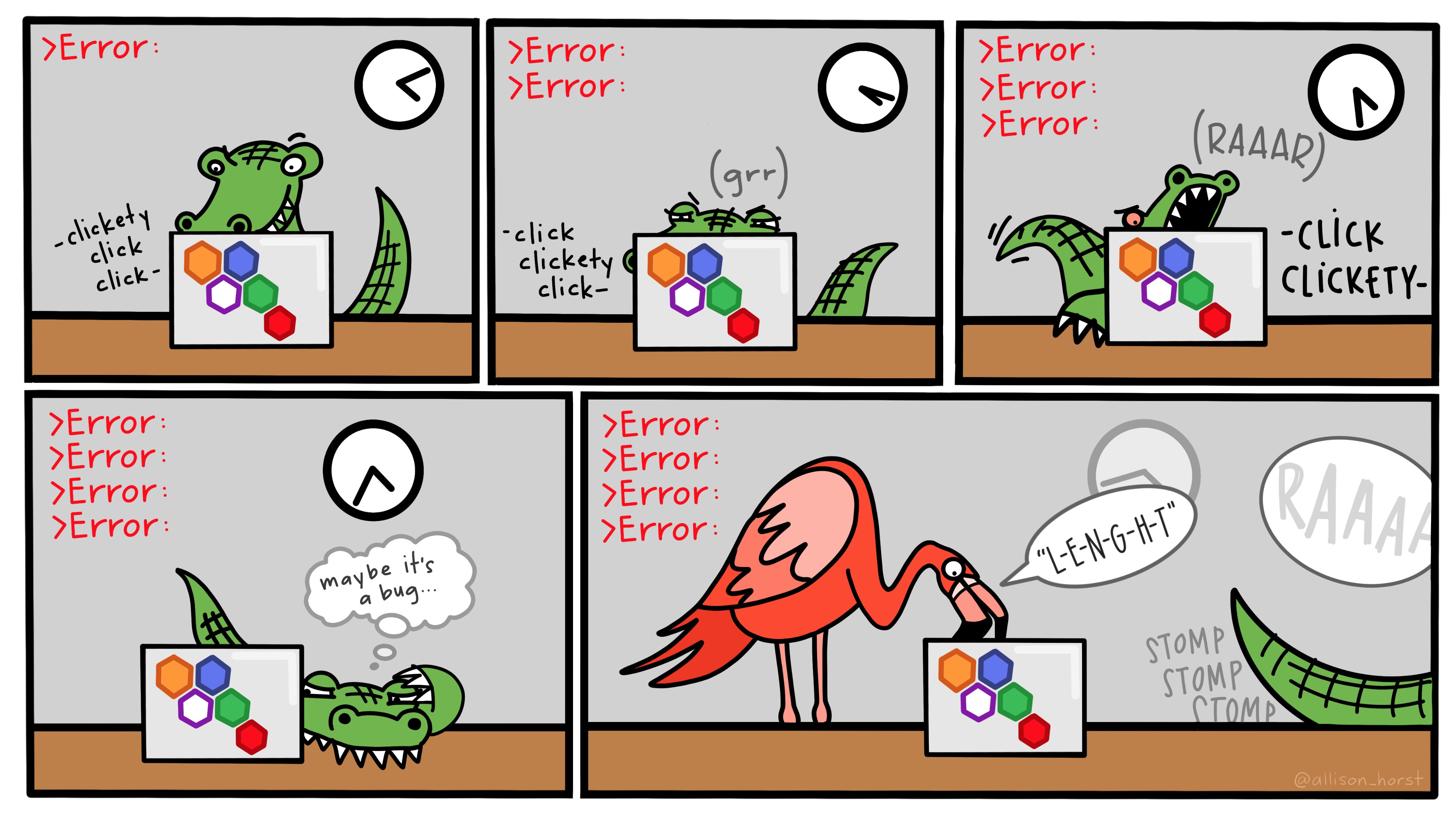
I *promise* you can succeed in this class.

Learning R can be difficult at first—it’s like learning a new language, just like Spanish, French, or Chinese. Hadley Wickham—the chief data scientist at RStudio and the author of some amazing R packages you’ll be using like ggplot2—[made this wise observation](https://r-posts.com/advice-to-young-and-old-programmers-a-conversation-with-hadley-wickham/):

It’s easy when you start out programming to get really frustrated and think, “Oh it’s me, I’m really stupid,” or, “I’m not made out to program.” But, that is absolutely not the case. Everyone gets frustrated. I still get frustrated occasionally when writing R code. It’s just a natural part of programming. So, it happens to everyone and gets less and less over time. Don’t blame yourself. Just take a break, do something fun, and then come back and try again later.

Even experienced programmers find themselves bashing their heads against seemingly intractable errors. If you’re finding yourself taking way too long hitting your head against a wall and not understanding, take a break, talk to classmates, e-mail me, etc.



[](https://twitter.com/allison_horst/status/1213275783675822080)

## 1.10 LeaRning is Social

* Be curious, not afraid.
* Know that if you have a question, other people will have it.
* Asking questions is our way of taking care of others

The students who have a bad time in my workshops and courses are the ones who don’t work with each other to learn. We are a learning community, and we should help each other to learn.

Find a buddy to work with - and check in with them during class and out of class

If you understand something and someone is struggling with it, try and help them. If you are struggling, take a breath, and try to pinpoint what you are struggling with.

Our goal is to be better programmers each day, not to be the perfect programmer. There’s no such thing as a perfect programmer. I’ve been learning new things almost every day.

## 1.11 Course Times

I know that everyone is busy, and we’ll do our best to accomodate everyone’s schedule.

Classes will be recorded, but please do not use this as an excuse to miss class. Again, those who are curious and ask questions will learn quite a bit.

## 1.12 Class Schedule

There are two sections of Intro to R.

* A hybrid (in-person and online) session on Wednesdays (12-1:30 PM PST)
* A completely remote session on Thursdays (2-3:30 PM PST)

When you are enrolled, we will send you teams invites for your section. Please note that we are at capacity for in-person. So if you have enrolled as online, please stay online.

The Hybrid Sections will be held in the Data Science Lab Lounge - Arnold M1- and online. Please note that I will in town and teaching in person on the starred (\*) dates below.

Dates when I am not on campus, you are free to attend in the DaSL lounge, but I will be teaching Remotely.

If you are remote, feel free to jump between either sessions.

| Week | Subject | Hybrid Section Dates | Remote Session Dates |
| --- | --- | --- | --- |
| 1\* | Introduction to R/RStudio | September 25 | September 26 |
| 2 | Data Structures | October 2 | October 3 |
| 3\* | Data Visualization | October 9 | October 10 |
| 4 (optional) | Community Session | October 16 | October 16 |
| 5\* | Data Wrangling 1 | October 23 | October 24 |
| 6 | Data Wrangling 2 | October 30 | October 31 |
| 7\* (optional) | Community Session | November 6 | November 6 |
| 8 | Wrap-up/Discuss Code-a-thon | November 13 | November 14 |

Note that the Community Sessions are Shared between the two sections.

More details about the Code-a-thon to come.

## 1.13 Community Sessions

Two times this quarter we will have learning community sessions, to talk about applications of what we’re learning. These sessions are optional, but will help you solidify your learning during the course.

These dates are:

October 16 at 12-1:30 PM November 6 at 12-1:30 PM

These dates will be sent to you when you register for the course.

## 1.14 Patient / Clinical Data is a No on Posit Cloud

The Posit Cloud workspace is for your learning. Please do not put any patient or clinical information on there.

## 1.15 Offerings

This course is taught on a regular basis at [Fred Hutch Cancer Center](https://www.fredhutch.org/) through the [Data Science Lab](https://hutchdatascience.org/). Announcements of course offering can be found [here](https://hutchdatascience.org/training/). If you wish to follow the course content asynchronously, you may access the course content on this website and [exercises and solutions on Posit Cloud](https://posit.cloud/content/8245357). The Posit Cloud compute space can be copied to your own workspace for personal use, and you can get started via this [introduction](https://hutchdatascience.org/Intro_to_R/intro-to-computing.html#posit-cloud-setup). Or, you can access the [exercises and solutions on GitHub](https://github.com/fhdsl/Intro_to_R_Exercises).

# 2 Intro to Computing

Welcome to Introduction to R! Each week, we cover a chapter, which consists of a lesson and exercise. In our first week together, we will look at big conceptual themes in programming, see how code is run, and learn some basic grammar structures of programming.

## 2.1 Goals of the course

In the next 6 weeks, we will explore:

* Fundamental concepts in high-level programming languages (R, Python, Julia, WDL, etc.) that is transferable: *How do programs run, and how do we solve problems using functions and data structures?*
* Beginning of data science fundamentals: *How do you translate your scientific question to a data wrangling problem and answer it?*
* 
* Data science workflow. Image source: [R for Data Science](https://r4ds.hadley.nz/whole-game).
* Find a nice balance between the two throughout the course: we will try to reproduce a figure from a scientific publication using new data.

## 2.2 What is a computer program?

* A sequence of instructions to manipulate data for the computer to execute.
* A series of translations: English <-> Programming Code for Interpreter <-> Machine Code for Central Processing Unit (CPU)

We will focus on English <-> Programming Code for R Interpreter in this class.

More importantly: **How we organize ideas <-> Instructing a computer to do something**.

## 2.3 A programming language has following elements:

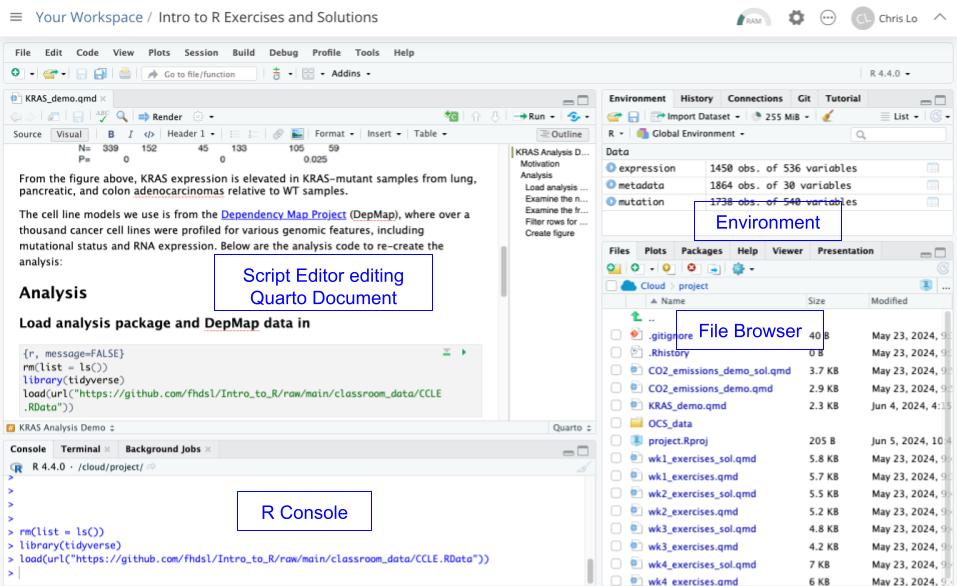
* Grammar structure to construct expressions
* Combining expressions to create more complex expressions
* Encapsulate complex expressions via **functions** to create modular and reusable tasks
* Encapsulate complex data via **data structures** to allow efficient manipulation of data

## 2.4 Posit Cloud Setup

Posit Cloud (the website version of RStudio) is an Integrated Development Environment (IDE). Think about it as Microsoft Word to a plain text editor. It provides extra bells and whistles to using R that is easier for the user.

Let’s open up the KRAS analysis in Posit Cloud. If you are taking this course while it is in session, the project name is probably named “KRAS Demo” in your Posit Cloud workspace. If you are taking this course on your own time, open up [“Intro to R Exercises and Solutions” project](https://posit.cloud/content/8245357).

Once you have opened the project, open the file “KRAS\_demo.qmd” from the File Browser, and you should see something like this:



Today, we will pay close attention to:

* R Console (Interpreter): You give it one line of R code, and the console executes that single line of code; you give it a single piece of instruction, and it executes it for you.
* Script Editor: where many lines of R code are typed and saved as a text document. To run the script, the Console will execute every single line of code in the document. The document you have opened in the script editor is a Quarto Document. A Quarto Document has chunks of plain text *and* R code, which helps us understand better the code we are writing.
* Environment: Often, your code will store information in the Environment, so that information can be reused. For instance, we often load in data and store it in the Environment, and use it throughout rest of your R code.

The first thing we will do is see the different ways we can run R code. You can do the following:

1. Type something into the R Console and type enter, such as 2+2. The R Console will run it and give you an output.
2. Scroll down the Quarto Document, and when you see a chunk of R Code, click the green arrow button. It will copy the R code chunk to the R Console and run all of it. You will likely see variables created in the Environment as you load in and manipulate data.
3. Run every single R code chunk in the Quarto Document by pressing the Run button at the top left corner of the Script Editor. It will generate an output document with all the code run.

Remember that the *order* that you run your code matters in programming. Your final product would be the result of Option 3, in which you run every R code chunk from start to finish. However, sometimes it is nice to try out smaller parts of your code via Options 1 or 2. But you will be at risk of running your code out of order!

Quarto is great for data science work, because:

* It encourages reproducible data analysis, when you run your analysis from start to finish.
* It encourages excellent documentation, as you can have code, output from code, and prose combined together.
* It is flexible to other programming languages, such as Python.

More options and guides can be found in [Introduction to Quarto](https://quarto.org/docs/get-started/hello/rstudio.html).

### 2.4.1

Now, we will get to the basics of programming grammar.

## 2.5 Grammar Structure 1: Evaluation of Expressions

* **Expressions** are be built out of **operations** or **functions**.
* Operations and functions combine **data types** to return another data type.
* We can combine multiple expressions together to form more complex expressions: an expression can have other expressions nested inside it.

For instance, consider the following expressions entered to the R Console:

18 + 21

## [1] 39

max(18, 21)

## [1] 21

max(18 + 21, 65)

## [1] 65

18 + (21 + 65)

## [1] 104

nchar("ATCG")

## [1] 4

Here, our input **data types** to the operation are **numeric** in lines 1-4 and our input data type to the function is **character** in line 5.

Operations are just functions in hiding. We could have written:

sum(18, 21)

## [1] 39

sum(18, sum(21, 65))

## [1] 104

Remember the function machine from algebra class? We will use this schema to think about expressions.

![Function machine from algebra class.](data:text/html; charset=iso-8859-1;base64,PCFET0NUWVBFIEhUTUwgUFVCTElDICItLy9JRVRGLy9EVEQgSFRNTCAyLjAvL0VOIj4KPGh0bWw+PGhlYWQ+Cjx0aXRsZT40MDQgTm90IEZvdW5kPC90aXRsZT4KPC9oZWFkPjxib2R5Pgo8aDE+Tm90IEZvdW5kPC9oMT4KPHA+VGhlIHJlcXVlc3RlZCBVUkwgd2FzIG5vdCBmb3VuZCBvbiB0aGlzIHNlcnZlci48L3A+CjwvYm9keT48L2h0bWw+Cg==)

Function machine from algebra class.

If an expression is made out of multiple, nested operations, what is the proper way of the R Console interpreting it? Being able to read nested operations and nested functions as a programmer is very important.

3 \* 4 + 2

## [1] 14

3 \* (4 + 2)

## [1] 18

Lastly, a note on the use of functions: a programmer should not need to know how the function is implemented in order to use it - this emphasizes [abstraction and modular thinking](#X4289ef2db5109e8791918a3c7b0cce6f2f89fd2), a foundation in any programming language.

### 2.5.1 Data types

Here are some data types that we will be using in this course:

* **Numeric**: 18, 21, 65, 1.25
* **Character**: “ATCG”, “Whatever”, “948-293-0000”
* **Logical**: TRUE, FALSE

## 2.6 Grammar Structure 2: Storing data types in the environment

To build up a computer program, we need to store our returned data type from our expression somewhere for downstream use. We can assign a variable to it as follows:

x = 18 + 21

If you enter this in the Console, you will see that in the Environment, the variable x has a value of 39.

### 2.6.1 Execution rule for variable assignment

Evaluate the expression to the right of =.

Bind variable to the left of = to the resulting value.

The variable is stored in the environment.

<- is okay too!

The environment is where all the variables are stored, and can be used for an expression anytime once it is defined. Only one unique variable name can be defined.

The variable is stored in the working memory of your computer, Random Access Memory (RAM). This is temporary memory storage on the computer that can be accessed quickly. Typically a personal computer has 8, 16, 32 Gigabytes of RAM. When we work with large datasets, if you assign a variable to a data type larger than the available RAM, it will not work. More on this later.

Look, now x can be reused downstream:

x - 2

## [1] 37

y = x \* 2

## 2.7 Grammar Structure 3: Evaluation of Functions

A function has a **function name**, **arguments**, and **returns** a data type.

### 2.7.1 Execution rule for functions:

Evaluate the function by its arguments, and if the arguments are functions or contains operations, evaluate those functions or operations first.

The output of functions is called the **returned value**.

sqrt(nchar("hello"))

## [1] 2.236068

(nchar("hello") + 4) \* 2

## [1] 18

## 2.8 Tips on writing your first code

Computer = powerful + stupid

Even the smallest spelling and formatting changes will cause unexpected output and errors!

* Write incrementally, test often
* Check your assumptions, especially using new functions, operations, and new data types.
* Live environments are great for testing, but not great for reproducibility.
* Ask for help!

## 2.9 Exercises

You can find [exercises and solutions on Posit Cloud](https://posit.cloud/content/8245357), or on [GitHub](https://github.com/fhdsl/Intro_to_R_Exercises).

# 3 Working with data structures

In our second lesson, we start to look at two **data structures**, **vectors** and **dataframes**, that can handle a large amount of data.

## 3.1 Vectors

In the first exercise, you started to explore **data structures**, which store information about data types. You played around with **vectors**, which is a ordered collection of a data type. Each *element* of a vector contains a data type, and there is no limit on how big a vector can be, as long the memory use of it is within the computer’s memory (RAM).

We can now store a vast amount of information in a vector, and assign it to a single variable. We can now use operations and functions on a vector, modifying many elements within the vector at once! This fits with the feature of “encapsulate complex data via data structures to allow efficient manipulation of data” described in the first lesson!

We often create vectors using the combine function, c() :

staff = c("chris", "shasta", "jeff")  
chrNum = c(2, 3, 1)

If we try to create a vector with mixed data types, R will try to make them be the same data type, or give an error:

staff = c("chris", "shasta", 123)  
staff

## [1] "chris" "shasta" "123"

Our numeric got converted to character so that the entire vector is all characters.

### 3.1.1 Using operations on vectors

Recall from the first class:

* Expressions are be built out of **operations** or **functions**.
* Operations and functions combine **data types** to return another data type.

Now that we are working with data structures, the same principle applies:

* Operations and functions combine **data structures** to return another data structure (or data type!).

What happens if we use some familiar operations we used for numerics on a numerical vector? If we multiply a numerical vector by a numeric, what do we get?

chrNum = chrNum \* 3  
chrNum

## [1] 6 9 3

All of chrNum’s elements tripled! Our multiplication operation, when used on a *numeric vector with a numeric*, has a *new* meaning: it multiplied all the elements by 3. Multiplication is an operation that can be used for multiple data types or data structures: we call this property **operator overloading**. Here’s another example: *numeric vector multiplied by another numeric vector*:

chrNum \* c(2, 2, 0)

## [1] 12 18 0

but there are also limits: a numeric vector added to a character vector creates an error:

#chrNum + staff

When we work with operations and functions, we must be mindful what inputs the operation or function takes in, and what outputs it gives, no matter how “intuitive” the operation or function name is.

### 3.1.2 Subsetting vectors explicitly

In the exercise this past week, you looked at a new operation to subset elements of a vector using brackets.

Inside the bracket is either a single numeric value or an a **numerical indexing vector** containing numerical values. They dictate which elements of the vector to return.

staff[2]

## [1] "shasta"

staff[c(1, 2)]

## [1] "chris" "shasta"

small\_staff = staff[c(1, 2)]

In the last line, we created a new vector small\_staff that is a subset of the staff given the indexing vector c(1, 2). We have three vectors referenced in one line of code. This is tricky and we need to always refer to our rules step-by-step: evaluate the expression right of the =, which contains a vector bracket. Follow the rule of the vector bracket. Then store the returning value to the variable left of =.

Alternatively, instead of using numerical indexing vectors, we can use a **logical indexing vector**. The logical indexing vector must be the *same length* as the vector to be subsetted, with TRUE indicating an element to keep, and FALSE indicating an element to drop. The following block of code gives the same value as before:

staff[c(TRUE, FALSE, FALSE)]

## [1] "chris"

staff[c(TRUE, TRUE, FALSE)]

## [1] "chris" "shasta"

small\_staff = staff[c(TRUE, TRUE, FALSE)]

### 3.1.3 Subsetting vectors implicitly

Here are two applications of subsetting on vectors that need distinction to write the correct code:

1. **Explicit subsetting**: Suppose someone approaches you a length 10 vector of people’s ages, and say that they want to subset to the 1st, 3rd, and 9th elements.
2. **Implicit subsetting**: Suppose someone approaches you a length 10 vector of people’s ages, and say that they want to subset to elements >50 age.

Consider the following vector.

age = c(89, 70, 64, 90, 66, 71, 55, 60, 30, 16)

We could subset age **explicitly** two ways. Suppose we want to subset the 1st and 5th, and 9th elements. One can do it with numerical indexing vectors:

age[c(1, 5, 9)]

## [1] 89 66 30

or by **logical indexing vectors**:

age[c(TRUE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, TRUE, FALSE)]

## [1] 89 66 30

and you can do it in one step as we have done so, or two steps by storing the indexing vector as a variable. *Either ways is fine.*

num\_idx = c(1, 5, 9)  
age[num\_idx]

## [1] 89 66 30

logical\_idx = c(TRUE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, TRUE, FALSE)  
age[logical\_idx]

## [1] 89 66 30

For implicit subsetting, we don’t know which elements to select off the top of our head! (We could count, but this method does not scale up.)

Rather, we can figure out which elements to select by using a **comparison operator**, which returns a logical indexing vector.

age > 50

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

The comparison operator > compared the numeric value of age to see which elements of age is greater than 50, and then returned a logical vector that has TRUE if age is greater than 50 at that element and FALSE otherwise.

Then,

indexing\_vector = age > 50  
age[indexing\_vector]

## [1] 89 70 64 90 66 71 55 60

#or  
age[age > 50]

## [1] 89 70 64 90 66 71 55 60

To summarize:

Subset a vector **implicitly**, in 3 steps:

1. Come up with a criteria for subsetting: “I want to subset to values greater than 50”.
2. We can use a **comparison operator** to create a **logical indexing vector** that fits this criteria.

age > 50

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

1. Use this logical indexing vector to subset.

age[age > 50]

## [1] 89 70 64 90 66 71 55 60

#or  
idx = age > 50  
age[idx]

## [1] 89 70 64 90 66 71 55 60

And you are done.

### 3.1.4 Comparison Operators

We have the following comparison operators in R:

< less than

<= less or equal than

== equal to

!= not equal to

> greater than

>= greater than or equal to

You can also put these comparison operators together to form more complex statements, which you will explore in this week’s exercise.

Another example:

age\_90 = age[age == 90]  
age\_90

## [1] 90

age\_not\_90 = age[age != 90]  
age\_not\_90

## [1] 89 70 64 66 71 55 60 30 16

For most of our subsetting tasks on vectors (and dataframes below), we will be encouraging implicit subsetting. The power of implicit subsetting is that you don’t need to know what your vector contains to do something with it! This technique is related to *abstraction* in programming mentioned in the first lesson: by using expressions to find the specific value you are interested instead of *hard-coding* the value explicitly, it generalizes your code to handle a wider variety of situations.

## 3.2 Dataframes

Before we dive into dataframes, check that the tidyverse package is properly installed by loading it in your R Console:

library(tidyverse)

Here is the data structure you have been waiting for: the **dataframe**. A dataframe is a spreadsheet such that each column must have the same data type. Think of a bunch of vectors organized as columns, and you get a dataframe.

For the most part, we load in dataframes from a file path (although they are sometimes created by combining several vectors of the same length, but we won’t be covering that here):

load(url("https://github.com/fhdsl/S1\_Intro\_to\_R/raw/main/classroom\_data/CCLE.RData"))

### 3.2.1 Using functions and operations on dataframes

We can run some useful functions on dataframes to get some useful properties, similar to how we used length() for vectors:

nrow(metadata)

## [1] 1864

ncol(metadata)

## [1] 30

dim(metadata)

## [1] 1864 30

colnames(metadata)

## [1] "ModelID" "PatientID" "CellLineName"   
## [4] "StrippedCellLineName" "Age" "SourceType"   
## [7] "SangerModelID" "RRID" "DepmapModelType"   
## [10] "AgeCategory" "GrowthPattern" "LegacyMolecularSubtype"  
## [13] "PrimaryOrMetastasis" "SampleCollectionSite" "Sex"   
## [16] "SourceDetail" "LegacySubSubtype" "CatalogNumber"   
## [19] "CCLEName" "COSMICID" "PublicComments"   
## [22] "WTSIMasterCellID" "EngineeredModel" "TreatmentStatus"   
## [25] "OnboardedMedia" "PlateCoating" "OncotreeCode"   
## [28] "OncotreeSubtype" "OncotreePrimaryDisease" "OncotreeLineage"

The last function, colnames() returns a character vector of the column names of the dataframe. This is an important property of dataframes that we will make use of to subset on it.

We introduce an operation for dataframes: the dataframe$column\_name operation selects for a column by its column name and returns the column as a vector. For instance:

metadata$OncotreeLineage[1:5]

## [1] "Ovary/Fallopian Tube" "Myeloid" "Bowel"   
## [4] "Myeloid" "Myeloid"

metadata$Age[1:5]

## [1] 60 36 72 30 30

We treat the resulting value as a vector, so we can perform implicit subsetting:

metadata$OncotreeLineage[metadata$OncotreeLineage == "Myeloid"]

## [1] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [8] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [15] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [22] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [29] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [36] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [43] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [50] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [57] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [64] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"  
## [71] "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid" "Myeloid"

The bracket operation [ ] on a dataframe can also be used for subsetting. dataframe[row\_idx, col\_idx] subsets the dataframe by a row indexing vector row\_idx, and a column indexing vector col\_idx.

metadata[1:5, c(1, 3)]

## ModelID CellLineName  
## 1 ACH-000001 NIH:OVCAR-3  
## 2 ACH-000002 HL-60  
## 3 ACH-000003 CACO2  
## 4 ACH-000004 HEL  
## 5 ACH-000005 HEL 92.1.7

We can refer to the column names directly:

metadata[1:5, c("ModelID", "CellLineName")]

## ModelID CellLineName  
## 1 ACH-000001 NIH:OVCAR-3  
## 2 ACH-000002 HL-60  
## 3 ACH-000003 CACO2  
## 4 ACH-000004 HEL  
## 5 ACH-000005 HEL 92.1.7

We can leave the column index or row index empty to just subset columns or rows.

metadata[1:5, ]

## ModelID PatientID CellLineName StrippedCellLineName Age SourceType  
## 1 ACH-000001 PT-gj46wT NIH:OVCAR-3 NIHOVCAR3 60 Commercial  
## 2 ACH-000002 PT-5qa3uk HL-60 HL60 36 Commercial  
## 3 ACH-000003 PT-puKIyc CACO2 CACO2 72 Commercial  
## 4 ACH-000004 PT-q4K2cp HEL HEL 30 Commercial  
## 5 ACH-000005 PT-q4K2cp HEL 92.1.7 HEL9217 30 Commercial  
## SangerModelID RRID DepmapModelType AgeCategory GrowthPattern  
## 1 SIDM00105 CVCL\_0465 HGSOC Adult Adherent  
## 2 SIDM00829 CVCL\_0002 AML Adult Suspension  
## 3 SIDM00891 CVCL\_0025 COAD Adult Adherent  
## 4 SIDM00594 CVCL\_0001 AML Adult Suspension  
## 5 SIDM00593 CVCL\_2481 AML Adult Mixed  
## LegacyMolecularSubtype PrimaryOrMetastasis SampleCollectionSite  
## 1 Metastatic ascites  
## 2 Primary haematopoietic\_and\_lymphoid\_tissue  
## 3 Primary Colon  
## 4 Primary haematopoietic\_and\_lymphoid\_tissue  
## 5 bone\_marrow  
## Sex SourceDetail LegacySubSubtype CatalogNumber  
## 1 Female ATCC high\_grade\_serous HTB-71  
## 2 Female ATCC M3 CCL-240  
## 3 Male ATCC HTB-37  
## 4 Male DSMZ M6 ACC 11  
## 5 Male ATCC M6 HEL9217  
## CCLEName COSMICID PublicComments  
## 1 NIHOVCAR3\_OVARY 905933   
## 2 HL60\_HAEMATOPOIETIC\_AND\_LYMPHOID\_TISSUE 905938   
## 3 CACO2\_LARGE\_INTESTINE NA   
## 4 HEL\_HAEMATOPOIETIC\_AND\_LYMPHOID\_TISSUE 907053   
## 5 HEL9217\_HAEMATOPOIETIC\_AND\_LYMPHOID\_TISSUE NA   
## WTSIMasterCellID EngineeredModel TreatmentStatus OnboardedMedia PlateCoating  
## 1 2201 MF-001-041 None  
## 2 55 MF-005-001 None  
## 3 NA Unknown MF-015-009 None  
## 4 783 Post-treatment MF-001-001 None  
## 5 NA MF-001-001 None  
## OncotreeCode OncotreeSubtype OncotreePrimaryDisease  
## 1 HGSOC High-Grade Serous Ovarian Cancer Ovarian Epithelial Tumor  
## 2 AML Acute Myeloid Leukemia Acute Myeloid Leukemia  
## 3 COAD Colon Adenocarcinoma Colorectal Adenocarcinoma  
## 4 AML Acute Myeloid Leukemia Acute Myeloid Leukemia  
## 5 AML Acute Myeloid Leukemia Acute Myeloid Leukemia  
## OncotreeLineage  
## 1 Ovary/Fallopian Tube  
## 2 Myeloid  
## 3 Bowel  
## 4 Myeloid  
## 5 Myeloid

head(metadata[, c("ModelID", "CellLineName")])

## ModelID CellLineName  
## 1 ACH-000001 NIH:OVCAR-3  
## 2 ACH-000002 HL-60  
## 3 ACH-000003 CACO2  
## 4 ACH-000004 HEL  
## 5 ACH-000005 HEL 92.1.7  
## 6 ACH-000006 MONO-MAC-6

The bracket operation on a dataframe can be difficult to interpret because multiple expression for the row and column indicies is a lot of information for one line of code. You will see easier-to-read functions for dataframe subsetting in the next lesson.

Lastly, try running View(metadata) in RStudio Console…whew, a nice way to examine your dataframe like a spreadsheet program!

## 3.3 Exercises

You can find [exercises and solutions on Posit Cloud](https://posit.cloud/content/8245357), or on [GitHub](https://github.com/fhdsl/Intro_to_R_Exercises).

# 4 Data Visualization

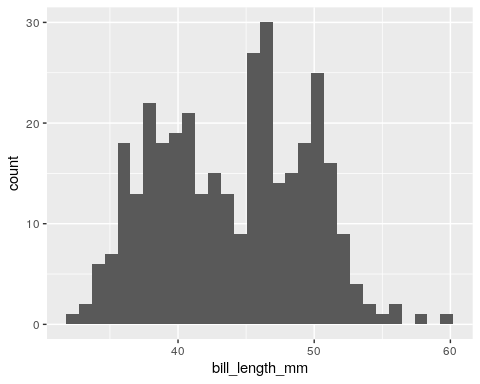
Now that we have learned basic data structures in R, we can now learn about how to do visualize our data. There are several different data visualization tools in R, and we focus on one of the most popular, “Grammar of Graphics”, or known as “ggplot”.

The syntax for ggplot will look a bit different than the code we have been writing, with syntax such as:

ggplot(penguins) + aes(x = bill\_length\_mm) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



The output of all of these functions, such as from ggplot() or aes() are not data types or data structures that we are familiar with…rather, they are graphical information.

You should be worried less about how this syntax is similar to what we have learned in the course so far, but to view it as a new grammar (of graphics!) that you can “layer” on to create more sophisticated plots.

To get started, we will consider these most simple and common plots:

**Univariate**

* Numeric: histogram
* Character: bar plots

**Bivariate**

* Numeric vs. Numeric: Scatterplot, line plot
* Numeric vs. Character: Box plot

Why do we focus on these common plots? Our eyes are better at distinguishing certain visual features more than others. All of these plots are focused on their position to depict data, which gives us the most effective visual scale.



Image Source: <https://www.oreilly.com/library/view/visualization-analysis-and/9781466508910/K14708_C005.xhtml>

## 4.1 Grammar of Graphics

The syntax of the grammar of graphics breaks down into 4 sections.

Data

Mapping to data

Geometry

Additional settings

You add these 4 sections together to form a plot.

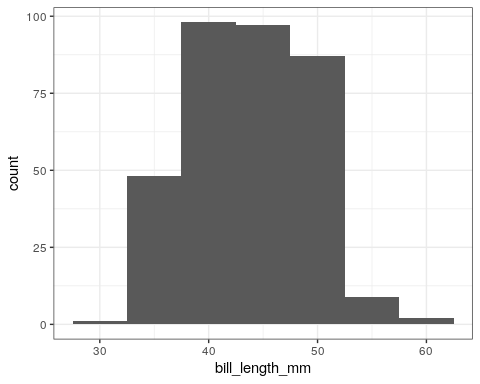
### 4.1.1 Histogram

ggplot(penguins) + aes(x = bill\_length\_mm) + geom\_histogram() + theme\_bw()



With options:

ggplot(penguins) + aes(x = bill\_length\_mm) + geom\_histogram(binwidth = 5) + theme\_bw()



### 4.1.2 Bar plots

ggplot(penguins) + aes(x = species) + geom\_bar()



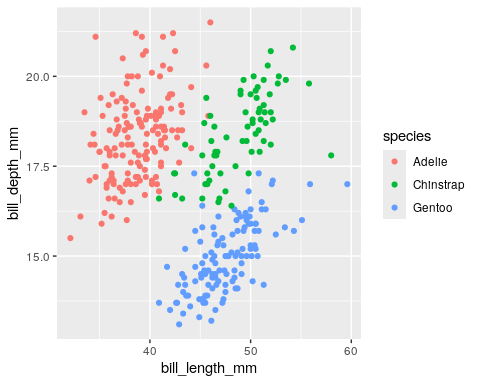
### 4.1.3 Scatterplot

ggplot(penguins) + aes(x = bill\_length\_mm, y = bill\_depth\_mm) + geom\_point()



### 4.1.4 Multivaraite Scatterplot

ggplot(penguins) + aes(x = bill\_length\_mm, y = bill\_depth\_mm, color = species) + geom\_point()



### 4.1.5 Multivaraite Scatterplot

ggplot(penguins) + aes(x = bill\_length\_mm, y = bill\_depth\_mm) + geom\_point() + facet\_wrap(~species)



### 4.1.6 Line plot?

ggplot(penguins) + aes(x = bill\_length\_mm, y = bill\_depth\_mm) + geom\_line()



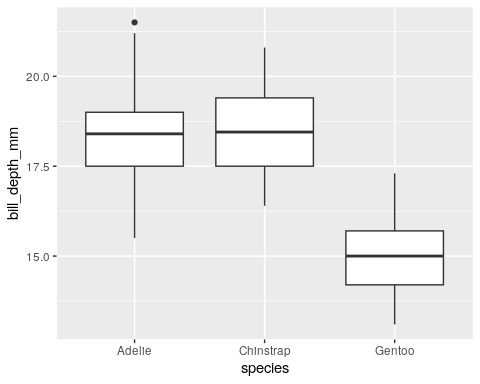
### 4.1.7 Grouped Line plot?

ggplot(penguins) + aes(x = bill\_length\_mm, y = bill\_depth\_mm, group = species) + geom\_line()



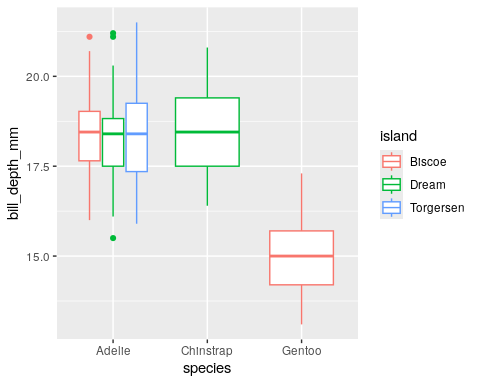
### 4.1.8 Boxplot

ggplot(penguins) + aes(x = species, y = bill\_depth\_mm) + geom\_boxplot()



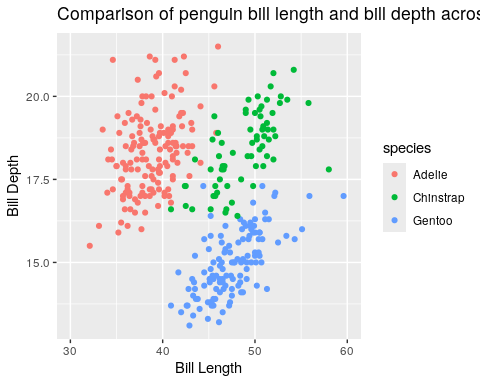
### 4.1.9 Grouped Boxplot

ggplot(penguins) + aes(x = species, y = bill\_depth\_mm, color = island) + geom\_boxplot()



### 4.1.10 Some additional options

ggplot(data = penguins) + aes(x = bill\_length\_mm, y = bill\_depth\_mm, color = species) + geom\_point() + labs(x = “Bill Length”, y = “Bill Depth”, title = “Comparison of penguin bill length and bill depth across species”) + scale\_x\_continuous(limits = c(30, 60))



## 4.2 Summary of options

data

geom\_point: x, y, color, shape

geom\_line: x, y, group, color

geom\_histogram: x, y, fill

geom\_bar: x, fill

geom\_boxplot: x, y, fill, color

facet\_wrap

labs

scale\_x\_continuous

scale\_y\_continuous

scale\_x\_discrete

scale\_y\_discrete

Consider the esquisse package to help generate your ggplot code via drag and drop.

An excellent ggplot “cookbook” can be found [here](https://r-graphics.org/).

## 4.3 Exercises

You can find [exercises and solutions on Posit Cloud](https://posit.cloud/content/8245357), or on [GitHub](https://github.com/fhdsl/Intro_to_R_Exercises).

# 5 Data Wrangling with Tidy Data, Part 1

From our first two lessons, we are now equipped with enough fundamental programming skills to apply it to various steps in the data science workflow, which is a natural cycle that occurs in data analysis.



Data science workflow. Image source: [R for Data Science.](https://r4ds.hadley.nz/whole-game)

For the rest of the course, we focus on *Transform* and *Visualize* with the assumption that our data is in a nice, “Tidy format”. First, we need to understand what it means for a data to be “Tidy”.

## 5.1 Tidy Data

Here, we describe a standard of organizing data. It is important to have standards, as it facilitates a consistent way of thinking about data organization and building tools (functions) that make use of that standard. The principles of **tidy data**, developed by Hadley Wickham:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

If you want to be technical about what variables and observations are, Hadley Wickham describes:

A **variable** contains all values that measure the same underlying attribute (like height, temperature, duration) across units. An **observation** contains all values measured on the same unit (like a person, or a day, or a race) across attributes.



A tidy dataframe. Image source: [R for Data Science](https://r4ds.hadley.nz/data-tidy).

## 5.2 Examples and counter-examples of Tidy Data:

Consider the following three datasets, which all contain the exact same information:

table1

## # A tibble: 6 × 4  
## country year cases population  
## <chr> <dbl> <dbl> <dbl>  
## 1 Afghanistan 1999 745 19987071  
## 2 Afghanistan 2000 2666 20595360  
## 3 Brazil 1999 37737 172006362  
## 4 Brazil 2000 80488 174504898  
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

This table1 satisfies the the definition of Tidy Data. The observation is a country’s year, and the variables are attributes of each country’s year.

head(table2)

## # A tibble: 6 × 4  
## country year type count  
## <chr> <dbl> <chr> <dbl>  
## 1 Afghanistan 1999 cases 745  
## 2 Afghanistan 1999 population 19987071  
## 3 Afghanistan 2000 cases 2666  
## 4 Afghanistan 2000 population 20595360  
## 5 Brazil 1999 cases 37737  
## 6 Brazil 1999 population 172006362

Something is strange able table2. The observation is still a country’s year, but “type” and “count” are not clear attributes of each country’s year.

table3

## # A tibble: 6 × 3  
## country year rate   
## <chr> <dbl> <chr>   
## 1 Afghanistan 1999 745/19987071   
## 2 Afghanistan 2000 2666/20595360   
## 3 Brazil 1999 37737/172006362   
## 4 Brazil 2000 80488/174504898   
## 5 China 1999 212258/1272915272  
## 6 China 2000 213766/1280428583

In table3, we have multiple values for each cell under the “rate” column.

## 5.3 Our working Tidy Data: DepMap Project

The [Dependency Map project](https://depmap.org/portal/) is a multi-omics profiling of cancer cell lines combined with functional assays such as CRISPR and drug sensitivity to help identify cancer vulnerabilities and drug targets. Here are some of the data that we have public access. We have been looking at the metadata since last session.

* Metadata
* Somatic mutations
* Gene expression
* Drug sensitivity
* CRISPR knockout
* and more…

Let’s see how these datasets fit the definition of Tidy data:

| Dataframe | The observation is | Some variables are | Some values are |
| --- | --- | --- | --- |
| metadata | Cell line | ModelID, Age, OncotreeLineage | “ACH-000001”, 60, “Myeloid” |
| expression | Cell line | KRAS\_Exp | 2.4, .3 |
| mutation | Cell line | KRAS\_Mut | TRUE, FALSE |

## 5.4 Transform: “What do you want to do with this dataframe”?

Remember that a major theme of the course is about: **How we organize ideas <-> Instructing a computer to do something.**

Until now, we haven’t focused too much on how we organize our scientific ideas to interact with what we can do with code. Let’s pivot to write our code driven by our scientific curiosity. After we are sure that we are working with Tidy data, we can ponder how we want to transform our data that satisfies our scientific question. We will look at several ways we can transform tidy data, starting with subsetting columns and rows.

Here’s a starting prompt:

In the metadata dataframe, which rows would you filter for and columns would you select that relate to a scientific question?

We should use the implicit subsetting mindset here: ie. “I want to filter for rows such that the Subtype is breast cancer and look at the Age and Sex.” and *not* “I want to filter for rows 20-50 and select columns 2 and 8”.

*Notice that when we filter for rows in an implicit way, we often formulate our criteria about the columns.*

(This is because we are guaranteed to have column names in dataframes, but not usually row names. Some dataframes have row names, but because the data types are not guaranteed to have the same data type across rows, it makes describing by row properties difficult.)

Let’s convert our implicit subsetting criteria into code!

metadata\_filtered = filter(metadata, OncotreeLineage == "Breast")  
breast\_metadata = select(metadata\_filtered, ModelID, Age, Sex)  
  
head(breast\_metadata)

## ModelID Age Sex  
## 1 ACH-000017 43 Female  
## 2 ACH-000019 69 Female  
## 3 ACH-000028 69 Female  
## 4 ACH-000044 47 Female  
## 5 ACH-000097 63 Female  
## 6 ACH-000111 41 Female

Here, filter() and select() are functions from the tidyverse package, which we have to install and load in via library(tidyverse) before using these functions.

### 5.4.1 Filter rows

Let’s carefully a look what how the R Console is interpreting the filter() function:

* We evaluate the expression right of =.
* The first argument of filter() is a dataframe, which we give metadata.
* The second argument is strange: the expression we give it looks like a logical indexing vector built from a comparison operator, but the variable OncotreeLineage does not exist in our environment! Rather, OncotreeLineage is a column from metadata, and we are referring to it as a **data variable** in the context of the dataframe metadata. So, we make a comparison operation on the column OncotreeLineage from metadata and its resulting logical indexing vector is the input to the second argument.
  + How do we know when a variable being used is a variable from the environment, or a data variable from a dataframe? It’s not clear cut, but here’s a rule of thumb: most functions from the tidyverse package allows you to use data variables to refer to columns of a dataframe. We refer to documentation when we are not sure.
  + This encourages more *readable* code at the expense of consistency of referring to variables in the environment. The authors of this package [describes this trade-off](https://dplyr.tidyverse.org/articles/programming.html#data--and-env-variables).
* Putting it together, filter() takes in a dataframe, and an logical indexing vector described by data variables as arguments, and returns a data frame with rows that match condition described by the logical indexing vector.
* Store this in metadata\_filtered variable.

### 5.4.2 Select columns

Let’s carefully a look what how the R Console is interpreting the select() function:

* We evaluate the expression right of =.
* The first argument of filter() is a dataframe, which we give metadata.
* The second and third arguments are **data variables** referring the columns of metadata.
  + For certain functions like filter(), there is no limit on the number of arguments you provide. You can keep adding data variables to select for more column names.
* Putting it together, select() takes in a dataframe, and as many data variables you like to select columns, and returns a dataframe with the columns you described by data variables.
* Store this in breast\_metadata variable.

## 5.5 Summary Statistics

Now that your dataframe has be transformed based on your scientific question, you can start doing some analysis on it! A common data science task is to examine summary statistics of a dataset, which summarizes the observations of a variable in a numeric summary.

If the columns of interest are numeric, then you can try functions such as mean(), median(), mode(), or summary() to get summary statistics of the column. If the columns of interest is character or logical, then you can try the table() function.

All of these functions take in a vector as input and not a dataframe, so you have to access the column as a vector via the $ operation.

mean(breast\_metadata$Age, na.rm = TRUE)

## [1] 50.96104

table(breast\_metadata$Sex)

##   
## Female Unknown   
## 91 1

## 5.6 Pipes

Often, in data analysis, we want to transform our dataframe in multiple steps via different functions. This leads to nested function calls, like this:

breast\_metadata = select(filter(metadata, OncotreeLineage == "Breast"), ModelID, Age, Sex)

This is a bit hard to read. A computer doesn’t care how difficult it is to read this line of code, but there is a lot of instructions going on in one line of code. This multi-step function composition will lead to an unreadable pattern such as:

result = function3(function2(function1(dataframe, df\_col4, df\_col2), arg2), df\_col5, arg1)

To untangle this, you have to look into the middle of this code, and slowly step out of it.

To make this more readable, programmers came up with an alternative syntax for function composition via the **pipe** metaphor. The ideas is that we push data through a chain of connected pipes, in which the output of a pipe becomes the input of the subsequent pipe.

Instead of a syntax like result2 = function3(function2(function1(dataframe))),

we linearize it with the %>% symbol: result2 = dataframe %>% function1 %>% function2 %>% function3.

In the previous example,

result = dataframe %>% function1(df\_col4, df\_col2) %>%  
 function2(arg2) %>%  
 function3(df\_col5, arg1)

This looks much easier to read. Notice that we have broken up one expression in to three lines of code for readability. If a line of code is incomplete (the first line of code is piping to somewhere unfinished), the R will treat the next line of code as part of the current line of code.

Try to rewrite the select() and filter() function composition example above using the pipe metaphor and syntax.

## 5.7 Exercises

You can find [exercises and solutions on Posit Cloud](https://posit.cloud/content/8245357), or on [GitHub](https://github.com/fhdsl/Intro_to_R_Exercises).

# 6 Data Wrangling with Tidy Data, Part 2

Today, we will continue learning about common functions from the Tidyverse that is useful for Tidy data manipulations.

## 6.1 Modifying and creating new columns in dataframes

The mutate() function takes in the following arguments: the first argument is the dataframe of interest, and the second argument is a *new or existing data variable* that is defined in terms of *other data variables*.

We create a new column olderAge that is 10 years older than the original Age column.

metadata$Age[1:10]

## [1] 60 36 72 30 30 64 63 56 72 53

metadata2 = mutate(metadata, olderAge = Age + 10)  
metadata2$olderAge[1:10]

## [1] 70 46 82 40 40 74 73 66 82 63

Here, we used an operation on a column of metadata. Here’s another example with a function:

expression$KRAS\_Exp[1:10]

## [1] 4.634012 4.638653 4.032101 5.503031 3.713696 3.972693 3.235727 4.135042  
## [9] 9.017365 3.940167

expression2 = mutate(expression, log\_KRAS\_Exp = log(KRAS\_Exp))  
expression2$log\_KRAS\_Exp[1:10]

## [1] 1.533423 1.534424 1.394288 1.705299 1.312028 1.379444 1.174254 1.419498  
## [9] 2.199152 1.371223

### 6.1.1 Alternative: Creating and modifying columns via $

Instead of mutate() function, we can also create a new or modify a column via the $ symbol:

expression2 = expression  
expression2$log\_KRAS\_Exp = log(expression2$KRAS\_Exp)

## 6.2 Merging two dataframes together

Suppose we have the following dataframes:

expression

| ModelID | PIK3CA\_Exp | log\_PIK3CA\_Exp |
| --- | --- | --- |
| “ACH-001113” | 5.138733 | 1.636806 |
| “ACH-001289” | 3.184280 | 1.158226 |
| “ACH-001339” | 3.165108 | 1.152187 |

metadata

| ModelID | OncotreeLineage | Age |
| --- | --- | --- |
| “ACH-001113” | “Lung” | 69 |
| “ACH-001289” | “CNS/Brain” | NA |
| “ACH-001339” | “Skin” | 14 |

Suppose that I want to compare the relationship between OncotreeLineage and PIK3CA\_Exp, but they are columns in different dataframes. We want a new dataframe that looks like this:

| ModelID | PIK3CA\_Exp | log\_PIK3CA\_Exp | OncotreeLineage | Age |
| --- | --- | --- | --- | --- |
| “ACH-001113” | 5.138733 | 1.636806 | “Lung” | 69 |
| “ACH-001289” | 3.184280 | 1.158226 | “CNS/Brain” | NA |
| “ACH-001339” | 3.165108 | 1.152187 | “Skin” | 14 |

We see that in both dataframes, the rows (observations) represent cell lines with a common column ModelID, so let’s merge these two dataframes together, using full\_join():

merged = full\_join(metadata, expression, by = "ModelID")

The number of rows and columns of metadata:

dim(metadata)

## [1] 1864 30

The number of rows and columns of expression:

dim(expression)

## [1] 1450 536

The number of rows and columns of merged:

dim(merged)

## [1] 1864 565

We see that the number of *columns* in merged combines the number of columns in metadata and expression, while the number of *rows* in merged is the larger of the number of rows in metadata and expression : full\_join() keeps all observations common to both dataframes based on the common column defined via the by argument.

Therefore, we expect to see NA values in merged, as there are some cell lines that are not in expression dataframe.

There are variations of this function depending on your application:



Given xxx\_join(x, y, by = "common\_col"),

* full\_join() keeps all observations.
* left\_join() keeps all observations in x.
* right\_join() keeps all observations in y.
* inner\_join() keeps observations common to both x and y.

## 6.3 Grouping and summarizing dataframes

In a dataset, there may be multiple levels of observations, and which level of observation we examine depends on our scientific question. For instance, in metadata, the observation is cell lines. However, perhaps we want to understand properties of metadata in which the observation is the cancer type, OncotreeLineage. Suppose we want the mean age of each cancer type, and the number of cell lines that we have for each cancer type.

This is a scenario in which the *desired rows are described by a column*, OncotreeLineage, and the columns, such as mean age, need to be *summarized from other columns.*

As an example, this dataframe is transformed from:

| ModelID | OncotreeLineage | Age |
| --- | --- | --- |
| “ACH-001113” | “Lung” | 69 |
| “ACH-001289” | “Lung” | 23 |
| “ACH-001339” | “Skin” | 14 |
| “ACH-002342” | “Brain” | 23 |
| “ACH-004854” | “Brain” | 56 |
| “ACH-002921” | “Brain” | 67 |

into:

| OncotreeLineage | MeanAge | Count |
| --- | --- | --- |
| “Lung” | 46 | 2 |
| “Skin” | 14 | 1 |
| “Brain” | 48.67 | 3 |

We use the functions group\_by() and summarise() :

metadata\_by\_type = metadata %>%   
 group\_by(OncotreeLineage) %>%   
 summarise(MeanAge = mean(Age, rm.na=TRUE), Count = n())

Or, without pipes:

metadata\_by\_type\_temp = group\_by(metadata, OncotreeLineage)  
metadata\_by\_type = summarise(metadata\_by\_type\_temp, MeanAge = mean(Age, rm.na=TRUE), Count = n())

The group\_by() function returns the identical input dataframe but remembers which variable(s) have been marked as grouped:

head(group\_by(metadata, OncotreeLineage))

## # A tibble: 6 × 30  
## # Groups: OncotreeLineage [3]  
## ModelID PatientID CellLineName StrippedCellLineName Age SourceType  
## <chr> <chr> <chr> <chr> <dbl> <chr>   
## 1 ACH-000001 PT-gj46wT NIH:OVCAR-3 NIHOVCAR3 60 Commercial  
## 2 ACH-000002 PT-5qa3uk HL-60 HL60 36 Commercial  
## 3 ACH-000003 PT-puKIyc CACO2 CACO2 72 Commercial  
## 4 ACH-000004 PT-q4K2cp HEL HEL 30 Commercial  
## 5 ACH-000005 PT-q4K2cp HEL 92.1.7 HEL9217 30 Commercial  
## 6 ACH-000006 PT-ej13Dz MONO-MAC-6 MONOMAC6 64 Commercial  
## # ℹ 24 more variables: SangerModelID <chr>, RRID <chr>, DepmapModelType <chr>,  
## # AgeCategory <chr>, GrowthPattern <chr>, LegacyMolecularSubtype <chr>,  
## # PrimaryOrMetastasis <chr>, SampleCollectionSite <chr>, Sex <chr>,  
## # SourceDetail <chr>, LegacySubSubtype <chr>, CatalogNumber <chr>,  
## # CCLEName <chr>, COSMICID <dbl>, PublicComments <chr>,  
## # WTSIMasterCellID <dbl>, EngineeredModel <chr>, TreatmentStatus <chr>,  
## # OnboardedMedia <chr>, PlateCoating <chr>, OncotreeCode <chr>, …

The summarise() returns one row for each combination of grouping variables, and one column for each of the summary statistics that you have specified.

Functions you can use for summarise() must take in a vector and return a simple data type, such as any of our summary statistics functions: mean(), median(), min(), max(), etc.

The exception is n(), which returns the number of entries for each grouping variable’s value.

You can combine group\_by() with other functions. See this [guide](https://dplyr.tidyverse.org/articles/grouping.html).

## 6.4 Appendix: How functions are built

As you become more independent R programmers, you will spend time learning about new functions on your own. We have gone over the basic anatomy of a function call back in the first lesson, but now let’s go a bit deeper to understand how a function is built and how to call them.

Recall that a function has a **function name**, **input arguments**, and a **return value**.

*Function definition consists of assigning a* ***function name*** *with a “function” statement that has a comma-separated list of named* ***function arguments****, and a* ***return expression****. The function name is stored as a variable in the global environment.*

In order to use the function, one defines or import it, then one calls it.

Example:

addFunction = function(num1, num2) {  
 result = num1 + num2   
 return(result)  
}  
result = addFunction(3, 4)

With function definitions, not all code runs from top to bottom. The first four lines defines the function, but the function is never run. It is called on line 5, and the lines within the function are executed.

When the function is called in line 5, the variables for the arguments are reassigned to function arguments to be used within the function and helps with the modular form.

To see why we need the variables of the arguments to be reassigned, consider the following function that is *not* modular:

x = 3  
y = 4  
addFunction = function(num1, num2) {  
 result = x + y   
 return(result)  
}  
result = addFunction(10, -10)

Some syntax equivalents on calling the function:

addFunction(3, 4)  
addFunction(num1 = 3, num2 = 4)  
addFunction(num2 = 4, num1 = 3)

but this *could* be different:

addFunction(4, 3)

With a deeper knowledge of how functions are built, when you encounter a foreign function, you can look up its help page to understand how to use it. For example, let’s look at mean():

?mean  
  
Arithmetic Mean  
  
Description:  
  
 Generic function for the (trimmed) arithmetic mean.  
  
Usage:  
  
 mean(x, ...)  
   
 ## Default S3 method:  
 mean(x, trim = 0, na.rm = FALSE, ...)  
   
Arguments:  
  
 x: An R object. Currently there are methods for numeric/logical  
 vectors and date, date-time and time interval objects.  
 Complex vectors are allowed for ‘trim = 0’, only.  
  
 trim: the fraction (0 to 0.5) of observations to be trimmed from  
 each end of ‘x’ before the mean is computed. Values of trim  
 outside that range are taken as the nearest endpoint.  
  
 na.rm: a logical evaluating to ‘TRUE’ or ‘FALSE’ indicating whether  
 ‘NA’ values should be stripped before the computation  
 proceeds.  
  
 ...: further arguments passed to or from other methods.

Notice that the arguments trim = 0, na.rm = FALSE have default values. This means that these arguments are *optional* - you should provide it only if you want to. With this understanding, you can use mean() in a new way:

numbers = c(1, 2, NA, 4)  
mean(x = numbers, na.rm = TRUE)

## [1] 2.333333

## 6.5 Exercises

You can find [exercises and solutions on Posit Cloud](https://posit.cloud/content/8245357), or on [GitHub](https://github.com/fhdsl/Intro_to_R_Exercises).

# 7 Cheatsheet

Here is a summary of expressions we learned in class.

Recall that we focused on English <-> Programming Code for R Interpreter in this class. Many of the functions we learned require the “Tidyverse” library to run.

## 7.1 Basic Data Types

| English | R Language |
| --- | --- |
| Numeric | 2 + 3 |
| Character | "hello", "123" |
| Logical | TRUE, FALSE |

## 7.2 Vectors

| English | R Language |
| --- | --- |
| Create a vector with some elements | vec = c(1, -4, -9, 12)  names = c("chris", "hannah", "chris", NA) |
| Compute length of a vector | length(vector) |
| Access the second element of names | names[2] |

## 7.3 Conditional Operations

Often to create a logical indexing vector for subsetting

| English | R Language |
| --- | --- |
| vec is greater than 0 | vec > 0 |
| vec is between 0 and 10 | vec >= 0 & vec <= 10 |
| vec is between 0 and 10, exclusively | vec > 0 & vec < 10 |
| vec is greater than 4 or less than -4 | vec > 4 | vec < -4 |
| names is “chris” | names == "chris" |
| names is not “chris” | names != "chris" |
| The non-missing values of names | !is.na(names) |

## 7.4 Subsetting vectors

| English | R Language |
| --- | --- |
| Subset vec to the first 3 elements | vec[c(1, 2, 3)]  or  vec[1:3]  or  vec[c(TRUE, TRUE, TRUE, FALSE)] |
| Subset vec to be greater than 0 | vec[vec > 0] |
| Subset names to have “chris” | vec[vec == "chris"] |

## 7.5 Dataframes

| English | R Language |
| --- | --- |
| Load a dataframe from CSV file “data.csv” | dataframe = read\_csv("data.csv") |
| Load a dataframe from Excel file “data.xlsx” | dataframe = read\_excel("data.xlsx") |
| Compute the dimension of dataframe | dim(dataframe) |
| Access a column “subtype” of dataframe as a vector | dataframe$subtype |
| Subset dataframe to columns “subtype”, “diversity”, “outcome” | select(dataframe, subtype, diversity, outcome) |
| Subset dataframe to rows such that the outcome is greater than zero, and the subtype is “lung”. | filter(dataframe, outcome > 0 & subtype == "lung") |
| Create a new column “log\_outcome” so that it is the log transform of “outcome” column | dataframe$log\_outcome = log(dataframe$outcome)  or  dataframe = mutate(dataframe, log\_outcome = log(outcome) |

## 7.6 Summary Statistics of a Dataframe’s column

| English | R Language |
| --- | --- |
| Mean of dataframe’s “outcome” column | mean(dataframe$outcome) |
| Mean of dataframe’s “outcome” column, removing NA values | mean(dataframe$outcome, na.rm = TRUE) |
| Max of dataframe’s “outcome” column | max(dataframe$outcome) |
| Min of dataframe’s “outcome” column | min(dataframe$outcome) |
| Count of dataframe’s “subtype” column | table(dataframe$subtype) |

## 7.7 Dataframe transformations

| English | R Language |
| --- | --- |
| Merge dataframe df1 and df2 by common column “id”, using all common entities. | full\_join(df1, df2, "id") |
| Group dataframe by “subtype” column, and summarise the mean “outcome” value for each “subtype” value, and get the total elements for each “subtype” value. | dataframe\_grouped = group\_by(dataframe, subtype)  dataframe\_summary = summarise(dataframe\_grouped, mean\_outcome = mean(outcome), n\_sample = n()) |

# About the Authors

These credits are based on our [course contributors table guidelines](https://www.ottrproject.org/more_features.html#giving-credits-to-contributors).

| Credits | Names |
| --- | --- |
| **Pedagogy** |  |
| Lead Content Instructor(s) | Chris Lo |
| Lecturer | Chris Lo |
| Content Author(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | If any other authors besides lead instructor |
| Content Contributor(s) (include section name/link in parentheses) - make new line if more than one section involved | Wrote less than a chapter |
| Content Editor(s)/Reviewer(s) | Checked your content |
| Content Director(s) | Helped guide the content direction |
| Content Consultants (include chapter name/link in parentheses or word “General”) - make new line if more than one chapter involved | Gave high level advice on content |
| Acknowledgments | Gave small assistance to content but not to the level of consulting |
| **Production** |  |
| Content Publisher(s) | Helped with publishing platform |
| Content Publishing Reviewer(s) | Reviewed overall content and aesthetics on publishing platform |
| **Technical** |  |
| Course Publishing Engineer(s) | Helped with the code for the technical aspects related to the specific course generation |
| Template Publishing Engineers | [Candace Savonen](https://www.cansavvy.com/), [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/) |
| Publishing Maintenance Engineer | [Candace Savonen](https://www.cansavvy.com/) |
| Technical Publishing Stylists | [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/), [Candace Savonen](https://www.cansavvy.com/) |
| Package Developers ([ottrpal](https://github.com/jhudsl/ottrpal)) [Candace Savonen](https://www.cansavvy.com/), [John Muschelli](https://johnmuschelli.com/), [Carrie Wright](https://carriewright11.github.io/) |  |
| **Art and Design** |  |
| Illustrator(s) | Created graphics for the course |
| Figure Artist(s) | Created figures/plots for course |
| Videographer(s) | Filmed videos |
| Videography Editor(s) | Edited film |
| Audiographer(s) | Recorded audio |
| Audiography Editor(s) | Edited audio recordings |
| **Funding** |  |
| Funder(s) | Institution/individual who funded course including grant number |
| Funding Staff | Staff members who help with funding |

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value  
## version R version 4.3.2 (2023-10-31)  
## os Ubuntu 22.04.4 LTS  
## system x86\_64, linux-gnu  
## ui X11  
## language (EN)  
## collate en\_US.UTF-8  
## ctype en\_US.UTF-8  
## tz Etc/UTC  
## date 2024-09-19  
## pandoc 3.1.1 @ /usr/local/bin/ (via rmarkdown)  
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date (UTC) lib source  
## bookdown 0.39.1 2024-06-11 [1] Github (rstudio/bookdown@f244cf1)  
## cachem 1.0.8 2023-05-01 [1] RSPM (R 4.3.0)  
## cli 3.6.2 2023-12-11 [1] RSPM (R 4.3.0)  
## devtools 2.4.5 2022-10-11 [1] RSPM (R 4.3.0)  
## digest 0.6.34 2024-01-11 [1] RSPM (R 4.3.0)  
## ellipsis 0.3.2 2021-04-29 [1] RSPM (R 4.3.0)  
## evaluate 0.23 2023-11-01 [1] RSPM (R 4.3.0)  
## fastmap 1.1.1 2023-02-24 [1] RSPM (R 4.3.0)  
## fs 1.6.3 2023-07-20 [1] RSPM (R 4.3.0)  
## glue 1.7.0 2024-01-09 [1] RSPM (R 4.3.0)  
## htmltools 0.5.7 2023-11-03 [1] RSPM (R 4.3.0)  
## htmlwidgets 1.6.4 2023-12-06 [1] RSPM (R 4.3.0)  
## httpuv 1.6.14 2024-01-26 [1] RSPM (R 4.3.0)  
## knitr 1.47.3 2024-06-11 [1] Github (yihui/knitr@e1edd34)  
## later 1.3.2 2023-12-06 [1] RSPM (R 4.3.0)  
## lifecycle 1.0.4 2023-11-07 [1] RSPM (R 4.3.0)  
## magrittr 2.0.3 2022-03-30 [1] RSPM (R 4.3.0)  
## memoise 2.0.1 2021-11-26 [1] RSPM (R 4.3.0)  
## mime 0.12 2021-09-28 [1] RSPM (R 4.3.0)  
## miniUI 0.1.1.1 2018-05-18 [1] RSPM (R 4.3.0)  
## pkgbuild 1.4.3 2023-12-10 [1] RSPM (R 4.3.0)  
## pkgload 1.3.4 2024-01-16 [1] RSPM (R 4.3.0)  
## profvis 0.3.8 2023-05-02 [1] RSPM (R 4.3.0)  
## promises 1.2.1 2023-08-10 [1] RSPM (R 4.3.0)  
## purrr 1.0.2 2023-08-10 [1] RSPM (R 4.3.0)  
## R6 2.5.1 2021-08-19 [1] RSPM (R 4.3.0)  
## Rcpp 1.0.12 2024-01-09 [1] RSPM (R 4.3.0)  
## remotes 2.4.2.1 2023-07-18 [1] RSPM (R 4.3.0)  
## rlang 1.1.4 2024-06-04 [1] CRAN (R 4.3.2)  
## rmarkdown 2.27.1 2024-06-11 [1] Github (rstudio/rmarkdown@e1c93a9)  
## sessioninfo 1.2.2 2021-12-06 [1] RSPM (R 4.3.0)  
## shiny 1.8.0 2023-11-17 [1] RSPM (R 4.3.0)  
## stringi 1.8.3 2023-12-11 [1] RSPM (R 4.3.0)  
## stringr 1.5.1 2023-11-14 [1] RSPM (R 4.3.0)  
## urlchecker 1.0.1 2021-11-30 [1] RSPM (R 4.3.0)  
## usethis 2.2.3 2024-02-19 [1] RSPM (R 4.3.0)  
## vctrs 0.6.5 2023-12-01 [1] RSPM (R 4.3.0)  
## xfun 0.44.4 2024-06-11 [1] Github (yihui/xfun@9da62cc)  
## xtable 1.8-4 2019-04-21 [1] RSPM (R 4.3.0)  
## yaml 2.3.8 2023-12-11 [1] RSPM (R 4.3.0)  
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library  
##   
## ──────────────────────────────────────────────────────────────────────────────

# 8 References