R for Environmental Chemists

David Hall, Steven Kutarna, Kristen Yeh, Hui Peng and Jessica D'eon

Last built on: 2021-07-14

Contents

Preface

Howdy,

This website is more-or-less the living results of a collaborative project between us. We're not trying to be an exhaustive resource for all environmental chemist. Rather, we're focusing on developing broadly applicable data science course content (tutorials and recipes) based in `R` for undergraduate environmental chemistry courses and research. Note that none of this has been reviewed yet and is not implemented in any capacity in any curriculum.

This book will ultimately be broken up into four sections:

- **Section 1: Getting Setup in R** is a general guide for the complete novice that will help you install, setup, and run R code. It features a useful tutorial exercise to make sure you have a working script before starting courses.
- **Section 2: Data Analysis in R** introduces data analysis workflows and showcases *how* you can use R and the `tidyverse` to tackle the vast majority of the data science/analysis problems you'll encounter in undergraduate environmental chemistry courses.
- **Section 3** explores the theory behind the most common data analysis practices in environmental chemistry. These include linear regression analysis, plotting, etc.
- **Section 4** consist of chapters specific to individual laboratory experiments. They rely upon knowledge from the previous three sections to introduce concepts unique to individual labs.

We recommend everyone reads over the first two sections in sequential order to understand the general R workflow employed in this, and all other, environmental chemistry courses.

Authors

If you have any questions/comments/suggestions/concerns please email:

- Dave at davidross.hall@mail.utoronto.ca
- Steven at steven.kutarna@mail.utoronto.ca
- Kristen at kristen.yeh@mail.utoronto.ca
- Dr. D'eon at jessica.deon@utoronto.ca

Section 1: Getting Setup in R

Chapter 1

Installing R

You've probably heard of coding and the R language, but figure out how to get started can be a hurdle; at least it was for us. This chapter will cover installing the software you'll need for coding in R.

1.1 Prerequisite software

Before we get started, you'll need to download the following open source and free software:

- **R**
- **RStudio**
- `tidyverse` suite of packages

Read on for instructions on downloading all three.

1.1.1 R

R is the programming language we'll code in. R is hosted on the Comprehensive R Archive Network (CRAN) and is one of the most popular programming languages for statisticians and scientist alike. You can download the latest build for your operating system here.

A quick aside, but coding is simply writing instructions for the computer to execute. To do this, we need a language that both we, humans, and the computer can understand. For our needs we'll use R, and like any language R has it's own syntax, rules, and quirks. We'll revisit this in subsequent sections.

1.1.2 Downloading RStudio IDE

RStudio is the *integrated development environment* of choice when working with R. It's where you'll actually be typing your code and interacting with R. Again, R is a language, and you need somewhere to write it down to make use of it. Writing in English can be done with a pencil and notepad or a word processor filed with useful tools to help you write. This is what RStudio is for R. You can download the latest version of RStudio here.

Once you have R and RStudio installed go ahead and open up RStudio. Once you open RStudio, you'll be greeted with an interface divided into numerous panes We've highlighted the major ones in the image below:

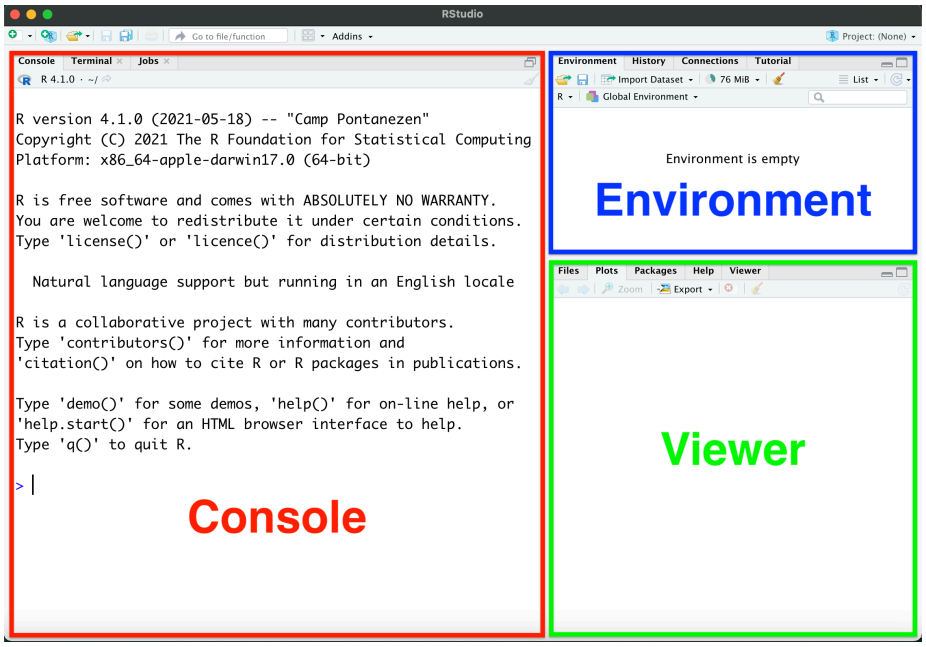


Figure 1.1: The RStudio interface with annottated regions

Each pane serves a specific role:

- **The console** allows you to directly type and run your code. It also provides messages, warnings, and errors from any code you run.
- **The environment** window lists all variables, data, and functions you've created since the start of your coding session.

- **The viewer** shows your outputs, help documents, etc. which each has their own tab.

1.1.3 Installing packages

Packages are previously written snippets of code that extend the capabilities of base R. Typically packages are created to address specific issues or workflows in different types of analysis. This book will make frequent use of a family of packages called the `tidyverse`. These packages all share a common thought process and integrate naturally with one another.

You can download the entire suite of `tidyverse` packages by simply copy and pasting the following code into the console and pressing 'enter.'

```
install.packages("tidyverse")
```

You'll see a flurry of lines printed to the console indicating the status of the installation. Once installed you won't be able to use these functions until you load it with `library()`. Enter the code below into the console to load the `tidyverse` package.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1.2 Running R Code

As we've already seen, you can run bits of R code directly from the console. Throughout the book, code you can copy and run will look like this:

```
2 + 2
```

```
## [1] 4
```

Noticed that both the code (the first part) and what the code outputs (the second part) are shown. Throughout this book code outputs will be proceeded

by ##. You can run code directly from the console. It's handy for short and sweet snippets of code, something that can be typed in a single line. Examples of this is the `install.packages()` function, or to use R as a calculator:

```
2 * 3
```

```
## [1] 6
```

```
pi * (10/2)
```

```
## [1] 15.70796
```

However, working like this isn't very useful Imagine printing a book one sentence at a time, you couldn't really go back and edit earlier work because it's already printed. That's why we write out code in *scripts*. *Scripts* are similar to recipes, in that they're a series of instructions that R evaluates from the top of the script to the bottom. More importantly, writing your code out in a script makes it *more readable* to humans (presumably this includes you). Don't undervalue the usefulness of legible code. Your code will evaluate in seconds or minutes whereas it may take you hours to understand what it does.

Let's open up a new script in RStudio by going to *File->New File->R Script*, or by clicking on the highlighted button in the image below.

This should open up a new window in the RStudio interface, as shown in the following image.

You can copy and paste the code above into the script, save it, edit it, etc. and ultimately run specific lines of code by highlighting them and pressing Ctrl+Enter (Cmd+Enter on Mac), or by clicking the "Run" button in the top right corner of the Scripts window. Whenever you copy code blocks from this website (or other online sources). If you're reading this book online, you can easily copy an entire block of code using the `copy` button in the top right corner of the code block.

We'll dive into the basics of coding in R in the next chapter.

1.3 Customizing RStudio

As many of us spend an absurd amount of time staring at bright screens, some of you may be interested in setting your RStudio to Dark Mode.

You can customize the appearance of your RStudio interface by clicking *Tools->Global Options*, or *RStudio->Preferences* on Mac, then clicking "Appearance" on the left. Select your preferred Editor Theme from the list.

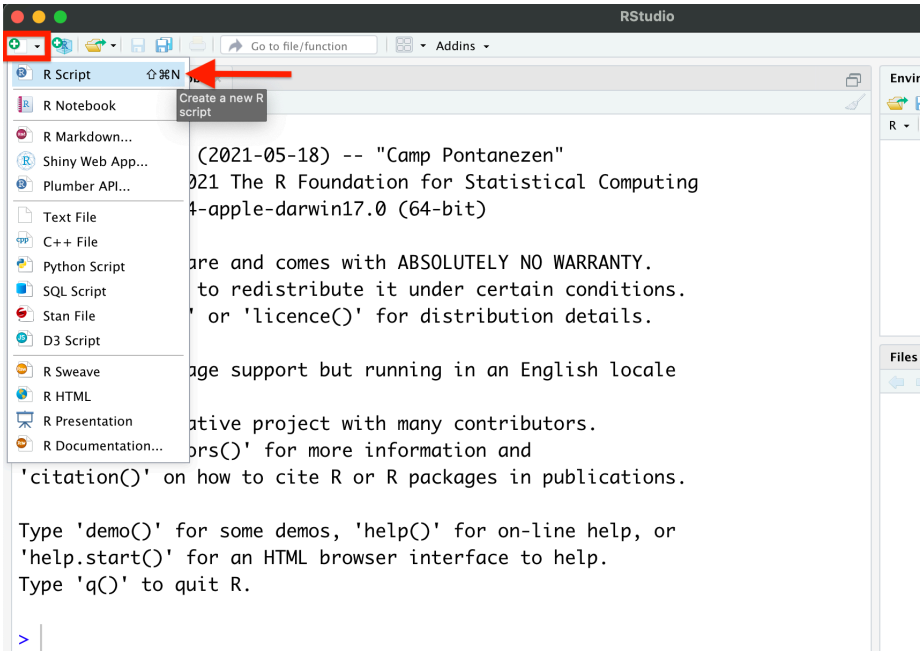


Figure 1.2: Figure 2.5: Opening a new script in RStudio.

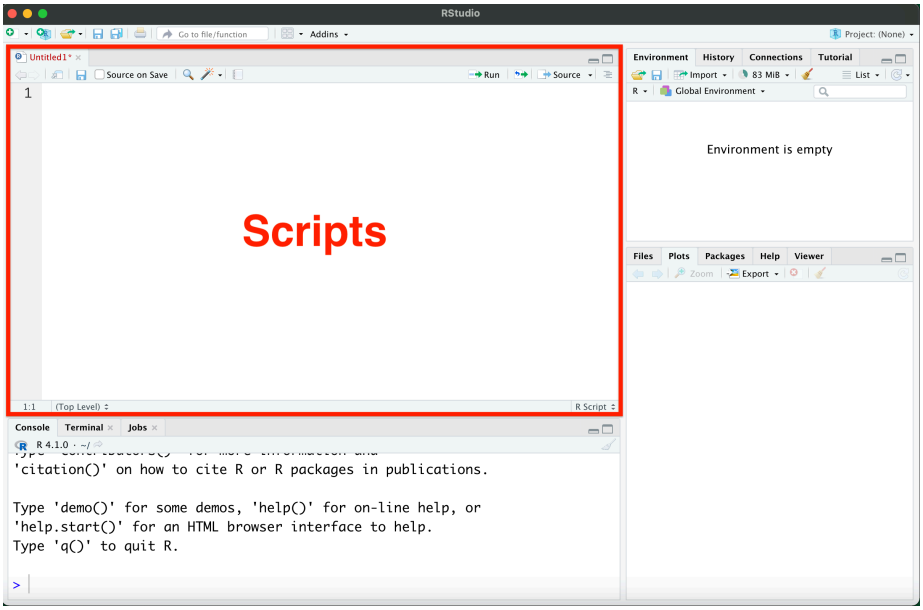


Figure 1.3: Figure 2.6: Scripts window in RStudio.

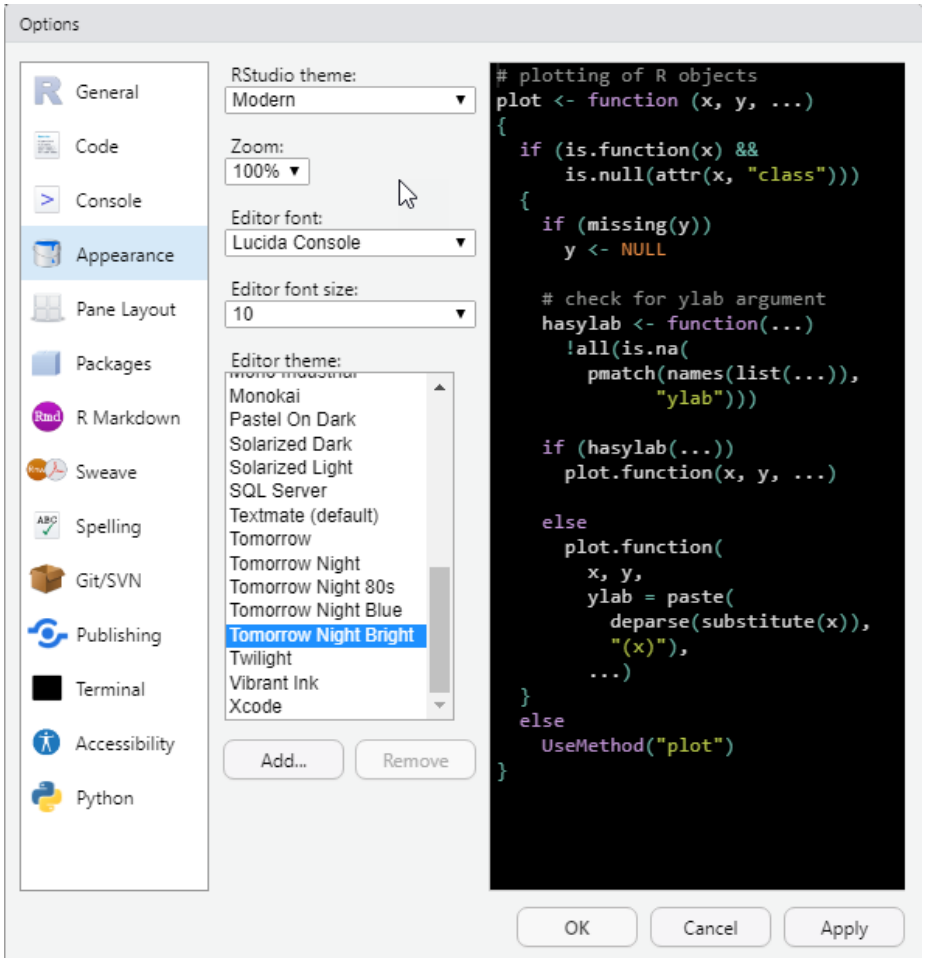


Figure 1.4: Figure 2.4: RStudio Appearance customization window.

1.4 Where to get help

While it's often tempting to contact your TA or Professor at the first sign of trouble, it's often better to try and resolve your issues on your own, especially if they're related to technical issues in R. Given the popularity of R, if you've run into an issue, someone else has too and they complained about it and someone else has solved it! An often unappreciated aspect of coding/data science is knowing how to get help, how to search for it, and how to translate someone's solutions to your unique situation.

Places to get help include:

- Google, Stack Overflow, etc. When in doubt Google it.
- Using built-in documentation (`?help`)
- reference book such as the invaluable *R for Data Science*, which inspired this entire project.
- All else fails, holler at your TA/profs.

Chapter 2

R coding basics

Now that you know how to navigate and run code in RStudio, we'll take a look at the basics of R. As we're chemist first, and not computer programmers, we'll try and avoid as much of the nitty-gritty underneath the hood aspects of R. However, a risk of this approach is being unable to understand errors and warnings preventing your code from running. As such, we'll introduce the most important and pertinent aspects of the R language to meet your environmental chemistry needs.

2.1 Variables

We've already talked about how R can be used like a calculator:

```
(1000 * pi) / 2
```

```
## [1] 1570.796
```

```
(2 * 3) + (5 * 4)
```

```
## [1] 26
```

But managing these inputs and outputs is simplified with **variables**. Variables in R, like those you've encountered in math class, can only have one value, and you can reference or pass that value along by referring the variable name. And, unlike the variables in math classes, you can change that value whenever you want. Another way to think about it is to treat variables in R like a box which you use to store a value. Then you can simply open the box somewhere else without having to worry about the hassle of what's inside.

You can assign the outputs variables using `<-`, as shown below.

```
x <- 12
x
```

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

In addition to reading code top to bottom, you often *read it from right to left*. `x <- 12` would be read as "take the value 12 and store it into the variable x." The second line of code, `x`, simply returns the value stored inside `x`. Note that when a variable is typed on it own, R will print out it's contents. You can now use this variable in snippets of code:

```
x
```

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

```
x <- x * 6.022e23
x
```

```
## [1] 7.2264e+24
```

Remember, we're evaluating from right to left, so the code above is taking the number `6.022e23` and multiplying it by the value of `x`, which is 12 and storing that value back into `x`. Note that variable names are case sensitive, so if your variable is named `x` and you type `X` into the console, R will not be able to print the contents of `x`. Variable names can consist of letters, numbers, dots (`.`) and/or underlines (`_`). Here are some rules and guidelines for naming variables in R:

- **Rules** dictated by R
  - names must begin with a letter or with the dot character. `var` and `.var` are acceptable.
  - Variable names *cannot* start with a number or the `.` character cannot be preceded by number. `var1` is acceptable, `1var` and `.1var` are not.
  - Variable names *cannot* contain a space. `var 1` is interpreted as two separate values, `var` and `1`.
  - Certain words are reserved for R, and cannot be used as variable names. These include, but are not limited to, `if`, `else`, `while`, `function`, `for`, `in`, `next`, `break`, `TRUE`, `FALSE`, `NULL`, `Inf`, `NA`, and `NaN`

Good names for variables are short, sweet, and easy to type while also being somewhat descriptive. For example, let's say you have an air pollution data set. A good name to assign the data set to would be `airPol` or `air_pol`, as these names tell us what is contained in the data set and are easy to type. A bad name for the data set would be `airPollution_NOx_O3_June20_1968`. While this name is much more descriptive than the previous names, it will take you a long time to type, and will become a bit of a nuisance when you have to type it 10+ times to refer to the data set in a single script. Please refer to the *Style Guide* found in *Advanced R* by H. Wickham for more information.

Lastly, R evaluates code from top-to-bottom of your script. So if you reference a variable it must have already been created at an earlier point in your script. For example:

```
y + 1
```

```
## Error in eval(expr, envir, enclos): object 'y' not found
```

```
y <- 12
```

The code above returns the `object 'y' not found` error because we're adding `+ 1` to `y` which hasn't been created yet, it's created on the next line. These errors also pop up when you edit your code without clearing your workplace. All variables created in a session are stored in the working environment so you can call them, even if you change your code. This means you can accidentally reference a variable that isn't reproduced in the latest iteration of your code. Consequently, a good practice is to frequently clear your work-space suing the 'broom' button in the *environment* pane. This will help you to ensure the code you're writing will is organized in the correct order; see Saving R scripts for why this is important.

2.2 Data Types

Data types refer to how data is stored and handled by and in R. You can really get into the weeds on this, but we'll focus on the most common types so you can get started on your work. Firstly, here are the data types you'll likely be working with:

- **character**: `"a"`, `"howdy"`, `"1"`, is used to represent string values in R. Basically it's text that you'd read. Note the quotation marks and the fact that `"1"`, despite being a number, is stored as a character.
- **numeric** (real or decimal): `2`, `3.14`, `6.022e23`.
- **integer**: `2L`, note the 'L' tells R this is an integer.

- **logical**: either `TRUE` or `FALSE`

There are some helpful functions to test the data type of a value in R as a frequent source of error and frustration are values stored in the wrong data type.

```
x <- "6"
x / 2
```

```
## Error in x/2: non-numeric argument to binary operator
```

You'll see the error above frequently, and it's simply saying you're trying to do math on something you can't do math on. You might think if `x` is 6, why can't I divide it by 2? Let's see what type of data `x` is:

```
is.numeric(x)   # test if numeric
```

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

```
is.logical(x)   # test if logical
```

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

```
is.integer(x)   # test if integer
```

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

```
is.character(x) # test if character
```

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

So the value of `x` is a character, in other words R treats it as a word, and we can't do math on that (think, how would you divide a word by a number?). So let's convert the data type of `x` to numeric to proceed.

```
x
```

```
## [1] "6"
```

```
x <- as.numeric(x)
is.numeric(x)
```

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

```
x
```

```
## [1] 6
```

```
x / 2
```

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

So we've converted our character string `"6"` to the numerical value `6`. Keep in mind there are other conversion functions which are described elsewhere, but you can't always convert types. In the above example we could convert a character to numeric because it was ultimately a number, but we couldn't do the same if the value of `x` was `"six"`.

2.3 Data Structures

Data structures refers to how R stores data. Again, it's easy to get lost in the weeds here so we'll focus on the most common and useful data types for your work which will overwhelmingly be data frames. Data frames consist of data stored in rows and columns. If you've ever worked with a spreadsheet (i.e. *Excel*), it's essentially that with the caveat that *all data stored in a column must be of the same type*. Again, different columns can have different data types, but *within* a column all the data needs to be the same type. R will convert your data otherwise to make it all the same. A common error is a single character in a column of numerical values leading to the entire column to be interpreted as character values; similar to what we discussed above. Errors like this most often stem from mistakes in recording and importing your data so be careful!

Let's import some data to see a data frame:

```
airPol <- read_csv("data/2018-07-01_60430_Toronto_ON.csv")

# data table so you can browse the data online
DT::datatable(airPol)
```

Show 10 entries Search:

	naps	city	p	latitude	longitude	date.time	pollutant	concentration
1	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T00:00:00Z	O3	46
2	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T00:00:00Z	NO2	8
3	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T00:00:00Z	SO2	0
4	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T01:00:00Z	O3	33
5	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T01:00:00Z	NO2	14
6	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T01:00:00Z	SO2	0
7	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T02:00:00Z	O3	33
8	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T02:00:00Z	NO2	11
9	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T02:00:00Z	SO2	0
10	60430	Toronto	ON	43.709444	-79.5435	2018-07-01T03:00:00Z	O3	34

Showing 1 to 10 of 507 entries Previous 1 2 3 4 5 … 51 Next

The above is air quality data measured in downtown Toronto around July 2018. Some of the variables are:

- `naps`, `city`, `p`, `latitude`, `longitude` to tell you where the data was measured.
- `data.time` for when the measurements were taking.
- `pollutant` for the chemical measured
- `concentration` for the measured concentration in parts-per-million (ppm).

This data is stored **tidy**, which is to say each column is a variable and each row is an observation. So reading the first row, we know that the Toronto 60430 station on 2018-07-01 at midnight measured ambient O_3 concentrations of 46 ppm. The concept of tidy data is important and is integral to working in R. It's discussed further in Tidying your data.

2.3.1 Other data structures

R has several other data structures. They aren't as frequently used, but it's worth being aware of their existence. Other structures include:

- **Vectors** contain multiple elements *of the same type*; either numeric, character (text), logical, or integer. Vectors are created using `c()`, which is short for combine. A data frame is just multiple vectors arranged into columns. Some examples of vectors are shown below.

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

```
## [1] 1 2 3 4 5
```

```
char <- c("blue", "green", "red")
char
```

```
## [1] "blue"  "green" "red"
```

```
log <- c(T, T, T, F, F, F)
log
```

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

- **Lists** are similar to vectors in that they are one dimensional data structures which contain multiple elements. However, lists can contain multiple elements of different types, while vectors only contain a single type of data. You can create lists using `list()`. Some examples of lists are shown below. You can use `str()` to reveal the different components of a list, in a more detailed format than if you were to simply type the assigned name of the list.

```
hi <- list("Hello", c(5,10,15,20), c(T, T, F))
str(hi)
```

```
## List of 3
##  $ : chr "Hello"
##  $ : num [1:4] 5 10 15 20
##  $ : logi [1:3] TRUE TRUE FALSE
```

```
hi
```

```
## [[1]]
## [1] "Hello"
##
## [[2]]
## [1]  5 10 15 20
##
## [[3]]
## [1]  TRUE  TRUE FALSE
```

There are many freely available resources online which dive more in depth into different data structures in R. If you are interested in learning more about different structures, you can check out the *Data structure* chapter of *Advanced R* by Hadley Wickham.

2.4 R packages and functions

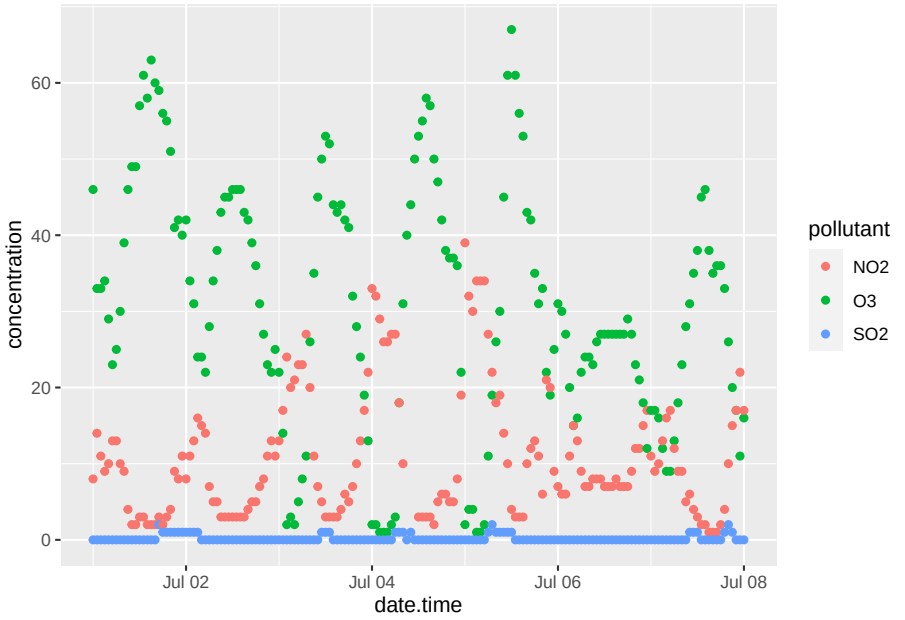
Functions are bits of code written to execute a specific task. We've already used several functions such as `library()` to import packages, and `read_csv()` to read the air quality data as seen in Data Structures above. Functions offer a convenient means to reduce the amount of typing while making code more reliable and readable. Some of these functions are built into R, such as `library()`, but often people write new functions to improve upon base R to help it meet the needs of it's users, such as the `read_csv()`. A collection of functions for a similar tasks is stored in a package, such as the `tidyverse` suite of packages which contains functions for plotting (`ggplot2`), reading data `read_csv()` and more.

Let's take a look at one of the functions you'll be using the most: `ggplot` from the `ggplot2` package which is included in the `tidyverse`.

2.4.1 ggplot2

`ggplot` allows you to create a variety of visualizations to explore and communicate your data and results. Like every function, `ggplot` has required *arguments*, i.e. data and instructions you pass to the function. The required arguments for this function are the *data* to be plotted and the aesthetic *mappings* for how the plot should look. Using our loaded air quality data from above:

```
ggplot(data = airPol,
       aes(x = date.time,
           y = concentration,
           colour = pollutant)) +
  geom_point()
```

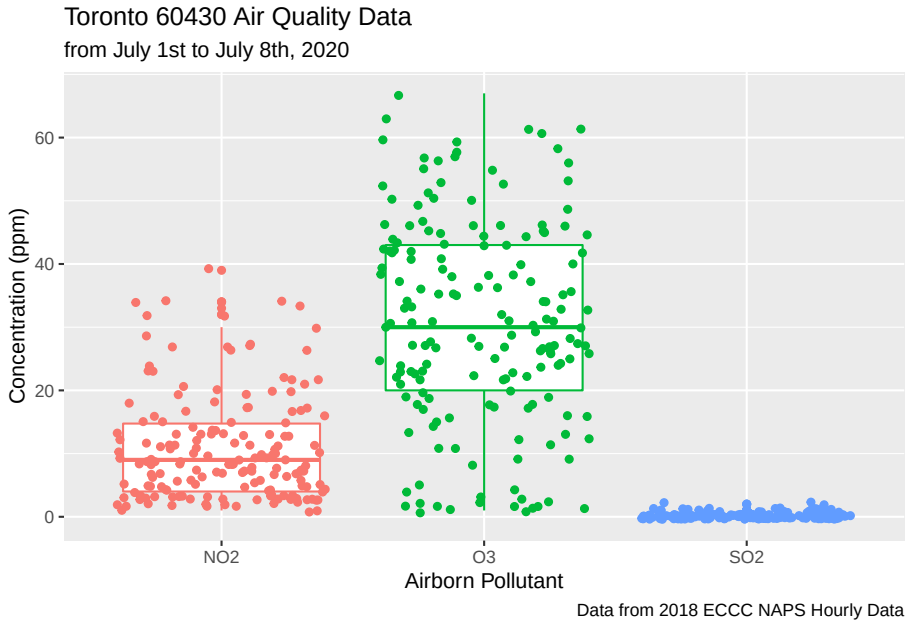


Breaking this down:

- We're calling `ggplot()` in the first line and passing the arguments inside the parentheses
- We're specifying that we want to plot the loaded air pollution data with `data = airPol`
- We insert a `,` to separate each argument
- We specified the *aesthetics* (`aes()`) arguments:
  - `x = data.time` means our x-axis will be the data in the `data.time` column.
  - `y = concentration` means our y-axis will be the data in the `concentration` column.
  - `colour = pollutant` means we colour each point based on `pollutant` column.
- We add a `+` on the second last line of code as this so we can *add* components to `ggplot()`
- And finally we add `geom_point()` to specify what type of plot we want; in this case it's a scatter plot.
  - `geoms` are layers that combine data, aesthetic mappings, and other data to create a plot.

`ggplot()` allows us to quickly create numerous plots of our data to aid our analysis. We can pass more than geoms to ggplot to improve our graphics. We can even stack geoms!

```
ggplot(data = airPol,
       aes(x = pollutant,
           y = concentration,
           colour = pollutant)) +
  geom_boxplot() +
  geom_jitter() +
  labs(title = "Toronto 60430 Air Quality Data",
       subtitle = "from July 1st to July 8th, 2020",
       x = "Airborn Pollutant",
       y = "Concentration (ppm)",
       caption = "Data from 2018 ECCC NAPS Hourly Data") +
  theme(legend.position = "none")
```



This plot looks more complicated then the previous one, but it's the same data plotted slightly differently and with a few bells and whistles:

- We specified that the `pollutant` column would be the x-axis, i.e. the three pollutants.
- We kept the y-axis and colour the same.
- `geom_boxplot()` creates a box-plot summarizing the spread of our data.
- `geom_jitter()` is overlaid so we see all the individual points in our data set; this is useful to make sure stuff isn't found in clusters.
- Annotated the plot using `labs()` including title, subtitle, x- and y-axis, and a caption. Useful for publications.

- Made some final aesthetic changes using `theme()`
  - specifically we removed the legend using `legend.position = "none"`.

This covers the basics of `ggplot()` but there's scores more you can do with this functions, and it's extended even further with packages. All of this is discussed in more detail in the Visualizations chapter.

2.4.2 Function Documentation

An oft unappreciated aspect of packages is that they not only contain functions we can use, but documentation. Documentation provides a description of the function (what it does), what arguments it takes, details, and working examples.Often the easiest way to learn how to use a function is to take a working example and change it bit by bit to see how it works etc. To see documentation check the "help" tab in the "outputs" window or type a question mark in front of a functions name:

```
# Takes you to the help document for the ggplot function
?ggplot
```

You can also write you're own functions. Please see Programming with R for additional details.

Now that you're familiar with navigating RStudio and some basic coding building blocks, let's move over to Chapter 3, where we'll review a normal workflow in R.

Chapter 3

Working with R

Now that you have some tools in your arsenal, let's discuss how to use them. After all knowing how to hammer a nail doesn't make you a carpenter. A coherent workflow is often overlooked by student's when they learn R, but a little bit of time setting up your work at the beginning will save you plenty of heartaches down the line. And just like there's a common workflow in any chemistry lab (pre-lab, collect reagents, conduct experiment, etc.) there's a common workflow when working with R. This is by no means the only way to work, but it's tried and true and will serve you well as you tackle your coursework. Let's take a step back from the scripts and consider *where* we're working on your computer.

3.1 Paths and directories

Before you get started with running your code, it is good to know where your analysis is actually occurring, or where your **working directory** is. The working directory is the folder where R looks for files that you have asked it to import, and the folder where R stores files that you have asked it to save.

RStudio displays the current working directory at the top of the console, as shown below, but can also be printed to the console using the command `getwd()`.

By default, R usually sets the working directory to the home directory on your computer. The `~` symbol denotes the home directory, and can be used as a shortcut when writing a path that references the home directory.

You can change the working directory using `setwd()` and an absolute file path. Absolute paths are references to files which point to the same file, regardless of what your working directory is set to. In Windows, absolute paths begin with `"C:"`, while they begin with with a slash in Mac and Linux (i.e.,

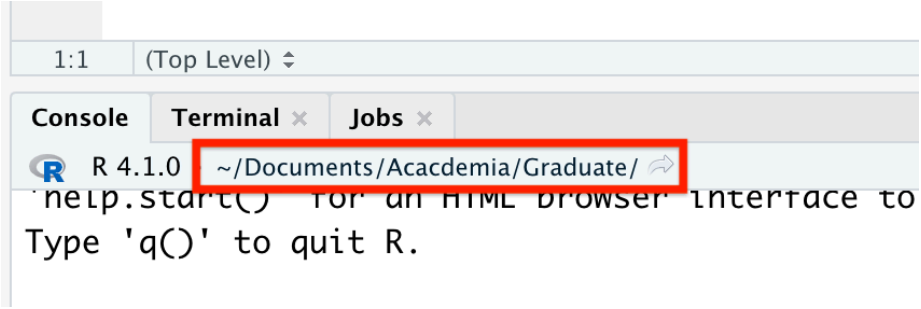


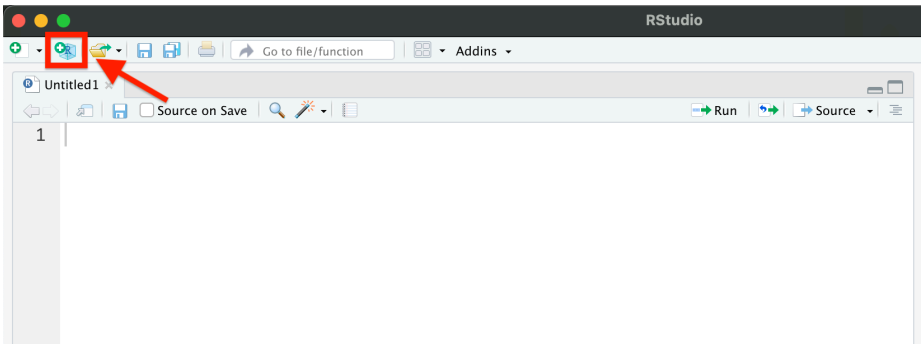
Figure 3.1: Working directory path displayed in the RStudio console

`"/Users/Vinny/Documents"`). It is important to note that absolute paths and `setwd()` should **never** be used in your scripts because they hinder sharing of code – no one else will have the same file configuration as you do. If you share your script with your TA or Prof, they will not be able to access the files you are referencing in an absolute path. Thus, they will not be able to run the code as-is in your script.

In order to overcome the use of absolute paths and `setwd()`, we strongly recommend that you conduct all work in RStudio within an **R project**. When you create an R project, R sets the working directory to a file folder of your choice. Any files that your code needs to run (i.e., data sets, images, etc.) are placed within this folder. You can then use relative paths to refer to data files in the project folder, which is much more conducive to sharing code with colleagues, TAs, and Profs.

3.2 Creating an RStudio project

Let's go ahead and create a new **R Project**. Go to *File->New Project*, or click the button highlighted in the image below. Click *New Directory*, then *New Project*.



You may want your project directory to be a sub-folder of an existing directory on your computer which already contains your data sets. If this is the case, click *Existing Directory* instead of *New Directory* at the previous step, and then select the folder of your choice.

Next, you'll be asked to choose a sub-directory name and location. Enter your selected name and choose an appropriate location for the folder on your computer. Click *Create Project*, and you should now see your chosen file path displayed in the bottom-right window:

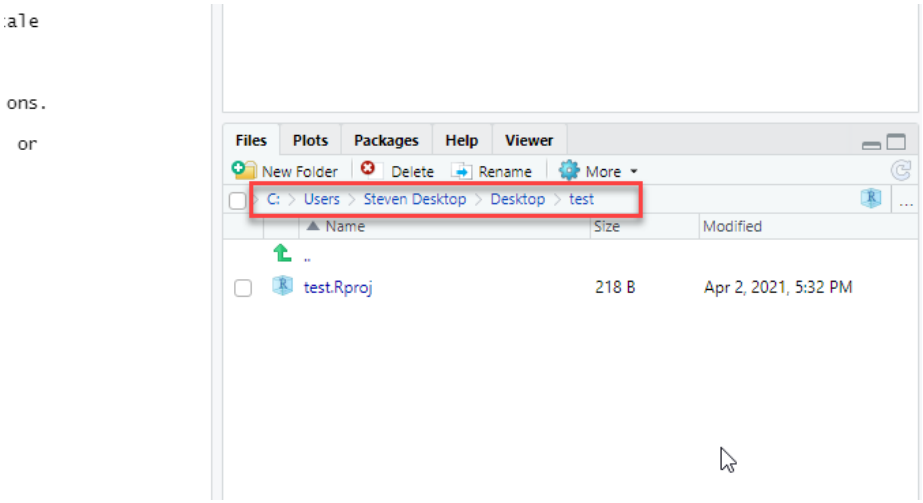


Figure 3.2: RStudio Project Folder

As mentioned previously, you can also view the file path to your project directory using `getwd()`. The output should match the file path shown in the image above.

When working on assignments for coursework, it is good practice to create a new R project for each assignment you work on. You should store the data, images, and any other files required for that assignment within the folder for the designated R project. You can create sub-folders for data and images, however, you may want to avoid making too many nested sub-folders, as this will make your paths long and tiresome to type.

3.2.1 The value of R projects

To demonstrate the benefit of working in a project directory rather than using absolute paths, let's review a quick example.

Let's say you have a data set called `absorbance.csv`, which is stored on your computer in the lengthy file path `/Users/Your_Name/Documents/School/Undergrad/Second_Year/CHM210/Assignm` You want to import the contents of this data set into R using `read_csv`.

Improper referencing: When working *outside* of an R project, you would need to reference the full, absolute file path in your scripts in order for R to recognize the file you are looking for. If you wanted to import the file, you would need to type something like this:

```
abs <- read_csv("/Users/Your_Name/Documents/School/Undergrad/Second_Year/CHM210/Assignm
```

While this does the job, it is extremely tedious to type the entire file path without typos, and this also hinders sharing your work with colleagues. Other students/your instructors will not have `absorbance.csv` stored in the same, lengthy file path which you have referenced above. Thus, if they try to run your script, this line of code will throw an error, as there is no file named `absorbance.csv` on their computer at the given file path.

Proper referencing: When working *inside* of an R project, you would set your project directory to the folder `/Assignment1`, using the pop-up windows that appear after clicking *File->New Project->Existing Directory*. By default, whenever you open the R project, the working directory will automatically be set to `/Assignment1`, the folder containing the data set of interest. If you wanted to import the file now, you would write the following command:

```
abs <- read_csv("absorbance.csv")
```

This is much simpler to type, and is much more compatible with sharing your work. Even if you don't share your entire project directory with your colleagues, they should still be able to run this line of code in the script, as long as they have the `absorbance.csv` data set in their current working directory folder.

Not only does working within an R project make your scripts much easier to share with colleagues, TAs, and Profs, but it also makes it easier for you to resume working on your code after you have closed the RStudio application. Think of your scripts as tabs in a web browser. Sometimes a project may require you to have several scripts open at once.

If you are working *outside* of an R project and have multiple scripts open, all of the scripts will close automatically when you quit RStudio. The next time you open RStudio you'll have to manually locate and open up each of the scripts you were working on previously, which can be tedious if they're not stored in convenient locations.

If you are working *inside* of an R project and have multiple scripts open, R will leave the scripts open within the project even after you have quit RStudio. The next time you open RStudio and your project, the script tabs will remain open, allowing you to easily pick up where you left off.

You can try this out for yourself. Open up a new script in your current project in RStudio using *File->New File->R Script*. (If you don't have a project open

currently, go to *File->New File*, click *New Directory*, then *New Project*.) Type in whatever you want. If you can't think of anything, here's an example:

```
# R projects are life savers
# wow
# blessed
```

Save the script by going to *File->Save*, or by clicking the button highlighted in the image below. Keep the script open, but close RStudio.

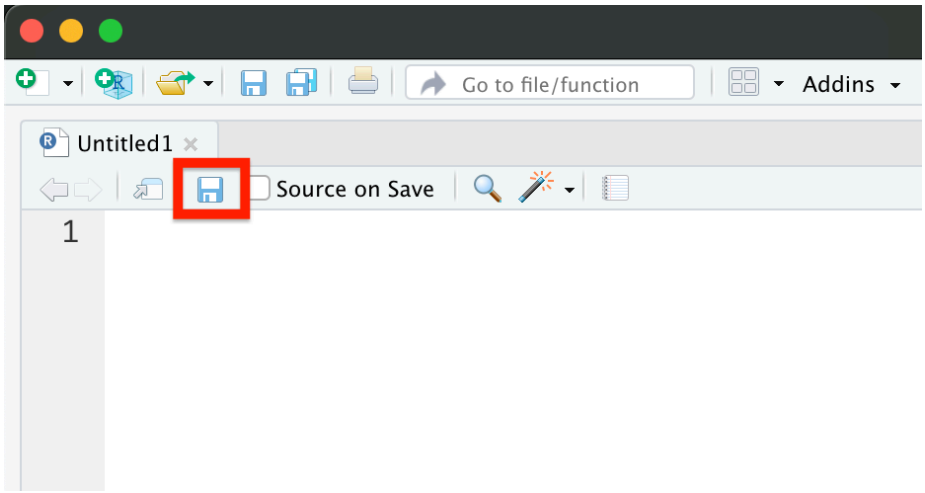


Figure 3.3: Save button in RStudio.

When you re-open RStudio, the script will still be there.

Leave the script open. Let's close the R project now. You can close the current project by going to *File->Close Project*, or by clicking the downwards arrow in the top right corner of RStudio, highlighted in the image below. Choose *Close Project* from the drop down menu.

Now close RStudio. When you open RStudio again, no script will be open! The same is true when you work outside of an R project.

3.3 Workspace and what's real

We've already mentioned the *environment* pane that displays objects present in your R session. While they are useful to work with, they're not *real*. That is to say, if you closed your R session, those objects would be lost. And while RStudio allows you to save a working environment (and it's associated objects), it's best to embrace that *only your scripts are real*. You can't readily share your

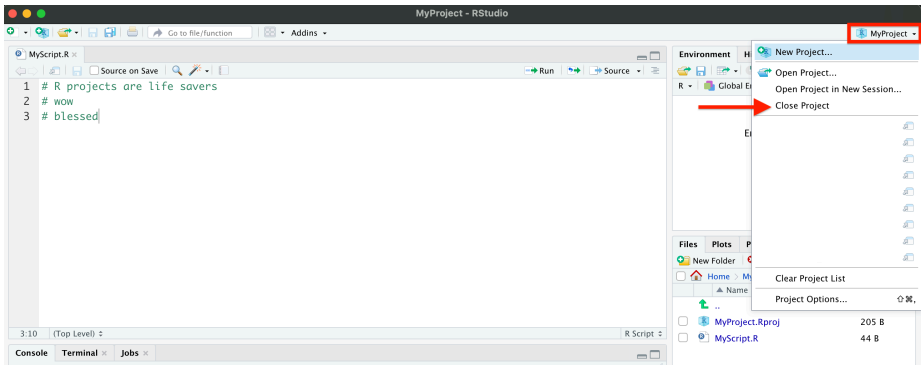


Figure 3.4: How to close an R Project.

working environment, and even so it's bad practice as you may be reference a previous iteration of an object giving you erroneous results. Think back to the chemistry lab where although you may jot notes down on loose leaf, only what's written in your lab book is considered real... we'll that's how it's supposed to work anyways.

The idea is everything you need can be generated from the original data and the instructions in your script Anyone should be able to take your data and your code and get the same results you got. This is paramount for reproducibility of your work and your results.

3.4 Saving R scripts

You can save an R script to a `.r` file by going to *File->Save* or by clicking the 'save' button in the top left of your script. Code saved to a `.r` file is considered *real*. Variables, plots, or data sets that only exist in your work-space (shown in the Environment window) are not. Whenever you close RStudio, any objects in R that are not considered *real* will be lost in that R session. Furthermore when you need to share your code (for school or publication) you'll need to share your data and your script, but never your work-space. This is to increase predictability and helps people (and you) to make sure your work is reproducible, an under appreciate hallmark of science.

3.4.1 What should I save?

At this point in the chapter, two things should be clear:

1. R scripts saved to `.R` files are real.

2. Objects in your work-space/environment are not real, and will not be available to you after you close and re-open RStudio unless you re-run the code used to generate the work-space.

So what is important to save in R, and how often should you save these files?

It is paramount that you save the scripts you code in, and that you save them regularly. Even if you've made small notation changes to the code, it is always a good idea to save your changes to the script before closing RStudio, as there is a good chance you will not remember the minor differences upon returning. You want to make sure that even if you lose an object in your environment, your script still contains the code you used to generate that object. You also want to make sure that you generate the object before you call it in part of another command, so that when you run your scripts from top-to-bottom, the variables are generated in the work-space before they are referenced by later commands.

3.4.2 Saving objects

In some cases, your code may be used to generate large data structures which require quite a bit of input to create. It can be quite tedious to re-run the code used to generate these large data sets every time you open RStudio, and you might find yourself wanting to save the data structure to a *real* file that you can simply import the next time you open the application. Most often this will be an intermediate step of your data analysis in the form of a data frame. To save a data frame as a `.csv` file you use `write.csv()`.

```
# dummy data frame to save
df <- data.frame(x = c(1,2,3),
                 y = c("yes", "no", "maybe"))

write.csv(x = df,
          file = "testData.csv")
```

Breaking it down:

- we created a dummy data frame `df`
- we called `write.csv()` and
  - `x = df` specifies we want to save the data.frame `df`
  - `file = "data/testData.csv"` specifies *where* we want the file to save (in the *data* sub-directory, more later), and *what* our file will be called (*testData.csv*). It's important to specify the file extension so R knows how to save it.

3.5 Script formatting

You should now be familiar with how to open the Scripts window, as well as some of the advantages of typing your code into this window rather than into the console directly. Before you write your first script, let's review some basic script formatting.

Before you enter any code into your script, it is good practice to fill the first few lines with text comments which indicate the script's title, author, and creation or last edit date. You can create a comment in a script by typing `#` before your text. An example is given below.

```
#Title: Ozone time series script
#Author: Georgia Green
#Date: January 8, 2072
```

Below your script header, you should include any packages that need to be loaded for the script to run. Including the necessary packages at the top of the script allows you, and anyone you share your code with, to easily see what packages they need to install. This also means that if you decide to run an entire script at once, the necessary packages will always be loaded before any subsequent code that requires those packages to work.

The first few lines of your scripts should look something like the following.

```
#Title: Ozone time series script
#Author: Georgia Green
#Date: January 8, 2072

#import packages
library(tidyverse) # for readr & ggplot2
```

The rest of your script should be dedicated to executable code. It is good practice to include text comments throughout the script, in between different chunks of code, to remind yourself what the different sections of code are for (i.e., `#import packages` in the above example). This also makes it easy for anyone you share your code with to understand what you're trying to do with different sections within the script.

You can also use headers and sub-headers in your scripts using `#`, `##`, and `###` before your text and `---` after as shown below:

```
# Section ----
## Subsection ----
### Sub-subsection ----
```

Headings and subheadings are picked up by RStudio and displayed in the Document Outline box. You can open the Document Outline box by clicking the button highlighted in the image below. Use of these headings allows easy navigation of long scripts, as you can navigate between sections using the Document Outline box.

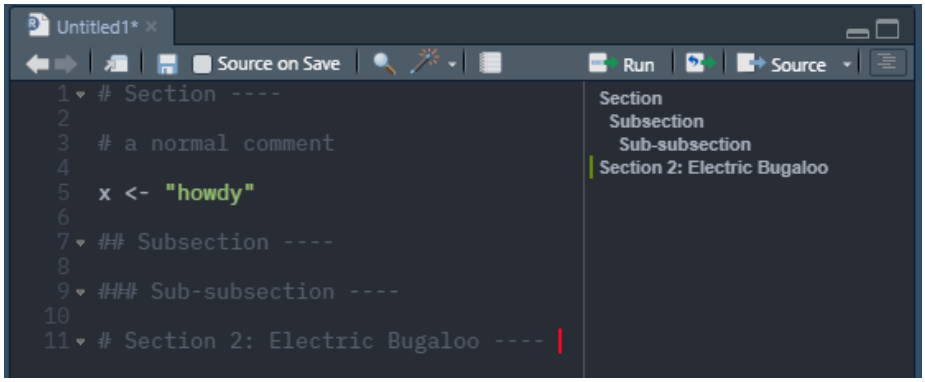


Figure 3.5: Example script headings, document outlines, and comments. Note the "—" which specifies a comment is a header.

3.6 Viewing data and code simultaneously

Before we get into more about coding and workflows, you may want to know how to view your scripts and data side-by-side. You can open a script, plot, or data set in a new window by clicking and dragging the tab in RStudio (may not be compatible with Mac), or by clicking the button highlighted in the image below.

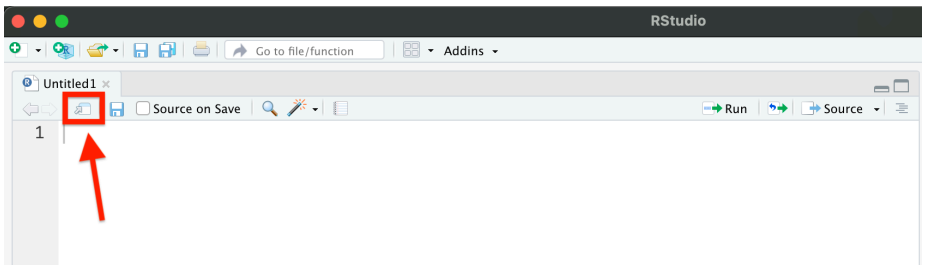


Figure 3.6: How to open an R script/plot/data set in a new window.

Now that you're familiar with navigating RStudio and some basic coding building blocks, let's move over to Chapter 3, where we'll review a normal workflow in R.

3.7 Troubleshooting error messages

In the previous section, you were introduced to your first error message in R, and we briefly discussed how to resolve the issue.

As you become more familiar with R and start using more complex functions, you will become better acquainted with error messages in R, and how to deal with them accordingly.

We'll go through a few examples of error messages in the following sections, as well as how to read the errors, and how to fix your code to resolve the issues.

3.7.1 Script diagnostics

When writing code in the Script window, RStudio will highlight any syntax errors in your code with a red squiggly line and an 'x' in the side bar, as shown below. You can hover over the 'x' to see what is causing the error.

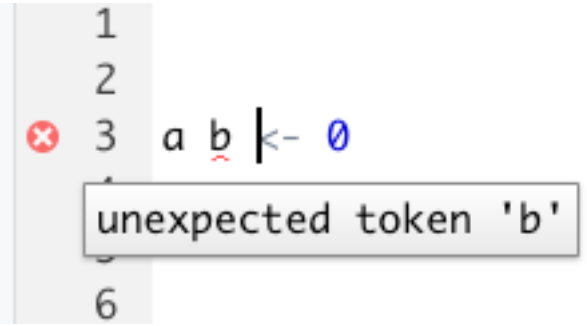


Figure 3.7: Figure 3.8: RStudio highlights syntax errors in the Scripts window.

In the above message, R is telling you that it is not sure what to do with `b`. As mentioned previously, variable assignment is done in the format `name <- assignment`. However, in the above example, the variable assignment statement is written as `name name <- assignment`. Since variable names cannot contain spaces, R reads `a b` as two separate input variable names, not as a single string. If you wanted to assign a value of 0 to both `a` and `b`, you would need to write the statement once per variable, as shown below.

```
a <- 0
b <- 0
```

Let's look at another example. Some functions require you to write code with nested parentheses. A good example would be the `aes()` argument that is called inside of `ggplot()`, as shown below.

```
#plot ozone concentration vs. time
ggplot(data = airPol, aes(x = Time, y = Concentration))
```

(For more detail about importing and using `ggplot2`, please re-visit Chapter 2, section 2.3.4, or see Chapter 11.)

If you were to forget one of the parentheses in the previous line of code, RStudio would highlight it similar to below:

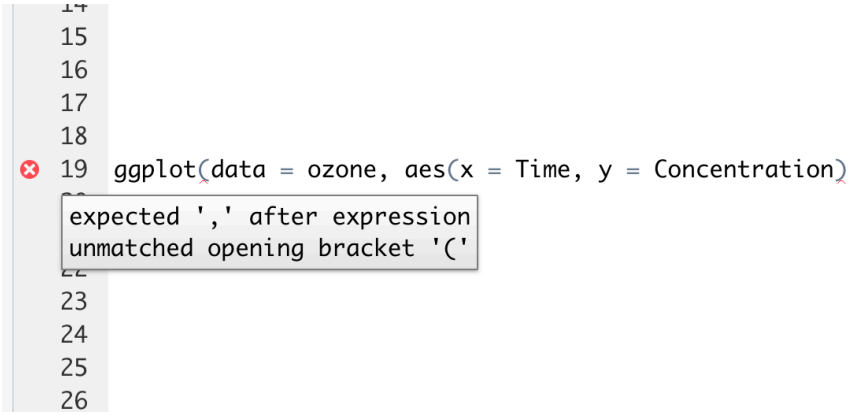


Figure 3.8: Figure 3.9: RStudio highlights unmatched parentheses in the script window.

Here R is telling you that you have an unmatched opening bracket. To resolve the error, simply add a closing bracket to match.

The `expected ',' after expression` is a common error that you will see accompanying unmatched opening brackets. Sometimes you might get this error in the console after running code that is missing a bracket somewhere. It is good practice to check your parentheses a few times before running your code to make sure that all the commands are closed, and that R doesn't keep waiting for you to continue inputting code after you've click *Run*. If you notice that the `>` in your R console has turned into a `+`, this is likely because you've just run a command that is missing a closing bracket, and thus, R is not aware that your code is finished. Simply input a closing bracket into the console, and the `>` should return.

3.7.2 Reading error codes

While the script window is very useful for pointing out syntax errors in your code, there are many other errors that can arise in RStudio which the script window is not able to capture. These are generally errors that arise from trying to execute your code, rather than from mistakes in your syntax.

The following is a prime example of such an error.

```
q <- 8 + "hi"
```

```
## Error in 8 + "hi": non-numeric argument to binary operator
```

Here we are trying to add a numeric value (8) to a character string ("hi"), then set the sum of the two to variable `q`. R has given us an error in return, because there is no logical way for R to add a numeric value to non-numeric text.

The error indicates that we have passed a `non-numeric argument to binary operator`, meaning we have used a non-numeric data type for an expression which is exclusively reserved for numeric data. If you try to add, divide or multiply two character strings using arithmetic operations in the console, you will get the same error.

```
"hey" * "hi"
```

```
## Error in "hey" * "hi": non-numeric argument to binary operator
```

It is important to be aware of these error codes as many functions require specific data types as their inputs. You can always look at the required data type by looking at the documentation for the function (generally, this can be viewed by typing `?function` into the console, where `function` is the name of the function). If the function requires numeric data, inputting character strings or logical values will throw the errors shown above. If the function requires logical values, inputting numeric data or character strings will throw the errors shown above.

In order to avoid these errors, make sure that you are using the right type of data in your functions. You can always check your data type using `class()`. Some examples are shown below.

```
class("hi")
```

```
## [1] "character"
```

```
class(10)
```

```
## [1] "numeric"
```

```
class(1L)
```

```
## [1] "integer"
```

```
class(TRUE)
```

```
## [1] "logical"
```

Now that you're familiar with working in RStudio, saving your projects, scripts, and data, let's move over to Chapter 4, where we'll discuss the advantages and disadvantages of using `.Rmd` documents instead of `.R` scripts.

Chapter 4

Using R Markdown

In a nutshell, R Markdown allows you to analyse your data with R and write your report in the same place (this entire book was written with rmarkdown). This has loads of benefits including a *reproducible workflow*, and streamlined thinking. No more flipping back and forth between coding and writing to figure out what's going on.

Let's run some simple code as an example:

```
# Look at me go mom
x <- 2+2
x
```

```
## [1] 4
```

What we've done here is write a snippet of R code, ran it, and printed the results (as they would appear in the console). While the above code isn't anything special, we can extend this concept so that our R markdown document contains any data, figures or plots we generate throughout our analysis in R.

Pretty neat, eh? You might not think so, but let's imagine a scenario you'll encounter soon enough. You're about to submit your assignment, you've spent hours analyzing your data and beautifying your plots. Everything is good to go until you notice at the last minute you were supposed to *subtract* value `x` and not value `y` in your analysis. If you did all your work in *Excel* (tsk tsk), you'll need to find the correct worksheet, apply the changes, reformat your plots, and import them into word (assuming everything is going well, which is never does with looming deadlines). Now if you did all your work in R markdown, you go to your one `.rmd` document, briefly apply the changes and re-compile your document.

4.1 Deeper look into rmarkdown

What we've done here is write a snippet of R code, ran it, and printed the results (as they would appear in the console). While the above code isn't anything special, we can extend this concept so that our R markdown document contains any data, figures or plots we generate throughout our analysis in R. For example:

```
library(tidyverse)
library(knitr)

airPol <- read_csv("data/2018-01-01_60430_Toronto_ON.csv")

ggplot(data = airPol,
       aes(x = date.time,
           y = concentration,
           colour = pollutant)) +
  geom_line() +
  theme_classic()
```

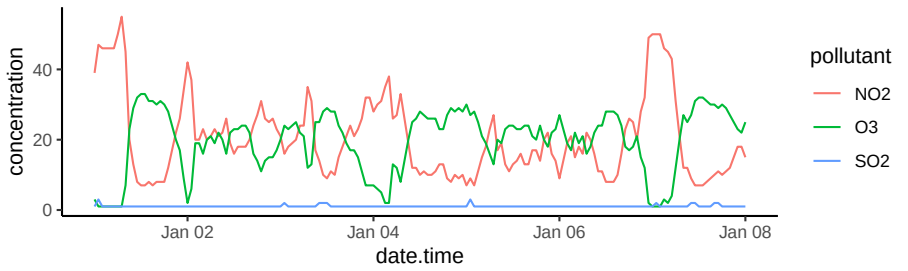


Figure 4.1: Time series of 2018 ambient atmospheric O_3, NO_2, and SO_2 concentrations (ppb) in downtown Toronto

```
sumAirPol <- airPol %>%
  drop_na() %>%
  group_by(city, naps, pollutant) %>%
  summarize(mean = mean(concentration),
            sd = sd(concentration),
            min = min(concentration),
            max = max(concentration))

knitr::kable(sumAirPol, digits = 1)
```

city	naps	pollutant	mean	sd	min	max
Toronto	60430	NO2	20.5	11.5	7	55
Toronto	60430	O3	19.7	8.7	1	33
Toronto	60430	SO2	1.1	0.3	1	3

Pretty neat, eh? You might not think so, but let's imagine a scenario you'll encounter soon enough. You're about to submit your assignment, you've spent hours analyzing your data and beautifying your plots. Everything is good to go until you notice at the last minute you were supposed to *subtract* value `x` and not value `y` in your analysis. If you did all your work in *Excel* (tsk tsk), you'll need to find the correct worksheet, apply the changes, reformat your plots, and import them into word (assuming everything is going well, which is never does with looming deadlines). Now if you did all your work in R markdown, you go to your one `.rmd` document, briefly apply the changes and re-compile your document.

4.2 Getting started with rmarkdown

As you've already guessed, R markdown documents use R and are most easily written and assembled in the R Studio IDE. If you have not done so, revisit Chapter 1:Installing R. Once setup with R and R Studio, you'll need to install the `rmarkdown` and `tinytex` packages. In the console, simply run the following code:

```
# These are large packages so it'll take a couple of minutes to install

install.packages("rmarkdown") # downloaded from CRAN

install.packages("tinytex")
tinytex::install_tinytex()  # install TinyTeX
```

The `rmarkdown` package is what we'll use to generate our documents, and the `tinytex` package enables compiling documents as PDFs. There's a lot more going on behind the scenes, but you shouldn't need to worry about it.

Now that everything is setup, you can create your first R Markdown document by opening up R Studio, selecting `FILE -> NEW FILE -> Rmarkdown`. A dialog box will appear asking for some basic input parameters for your R markdown document. Add your title and select PDF as your default output format (you can always change these later if you want). A new file should appear that's already populated with some basic script illustrating the key components of an R markdown document.

4.2.1 Understanding rmarkdown

Your first reaction when you opened your newly created R markdown document is probably that it doesn't look anything at all like something you'd show your prof. You're right, what you're seeing is the plain text code which needs to be compiled (called *knit* in R Studio) to create the final document. When you create a R markdown document like this in R Studio a bunch of example code is already written. You can compile this document (see below) to see what it looks like, but let's break down the primary components. At the top of the document you'll see something that looks like this:

```
---
title: "Temporal Analysis of Foot Impacts While Birling Down the White Water"
author: "Jean Guy Rubberboots"
date: "24/06/2021"
output: pdf_document
---
```

This section is known as the *preamble* and it's where you specify most of the document parameters. In the example we can see that the document title is "Temporal Analysis of Foot Impacts While Birling Down the White Water", it's written by Jean Guy Rubberboots, on the 24th of June, and the default output is a PDF document. You can modify the preamble to suit your needs. For example, if you wanted to change the title you would write `title: "Your Title Here"` in the preamble. Note that none of this is `R` code, rather it's `YAML`, the syntax for the document's metadata. Apart from what's shown you shouldn't need to worry about this much, just remember that indentation in YAML matters.

Reading further down the default R markdown code, you'll see different blocks of text. In R markdown anything you write will be interpreted as body text (i.e. stuff like this that you want folks to read) in the knitted document. **To actually run R code** you'll need to see the next section.

4.2.2 Runnign code in rmarkdown

There's two ways to write `R` code in markdown:

- **Setup a code chunk**. Code chunks start with three back-ticks like this: ```` ```{r} ````, where `r` indicates you're using the R language You end a code chunk using three more backticks like this ```` ``` ````.
  - Specify code chunks options in the curly braces. i.e. ```` ```{r, fig.height = 2} ```` sets figure height to 2 inches. See the *Code Chunk Options* section below for more details.

- **Inline code expression**, which starts with `` `r `` and ends with `` ` `` in the body text.
  - Earlier we calculated `x <- 2 + 2`, we can use inline expressions to recall that value (ex. We found that x is 4)

A screenshot of how this document, the one you're reading, appeared in R Studio is shown in the image below.

To actually run your R code you have two options. The first is to run the individual chunks using the *Run current chunk* button (See figure 2). This is a great way to tinker with your code before you compile your document. The second option is to compile your entire document using the *Knit document* button (see Figure 2). Knitting will sequentially run all of your code chunks, generate all the text, knit the two together and output a PDF. You'll basically save this for the end.

Note all the code chunks in a single markdown document work together like a normal R script. That is if you assign a value to a variable in the first chunk, you can call this variable in the second chunk; the same applies for libraries. **Also note that every time you compile a markdown document, it's done in a "fresh" R session.** If you're calling a variable that exist in your working environment, but isn't explicitly created in the markdown document you'll get an error.

4.2.3 Generating final report

To create a PDF to hand in you'll need to compile, or knit, your entire markdown document as mentioned above. To knit (or compile) your R markdown script, simply click the *knit* button in R Studio (yellow box, Figure 2). You can specify what output you would like and R Studio will (hopefully) compile your script.

If you want to test how your code chunks will run, R Studio shows a little green 'play button' on the top right of every code chunk. this is the 'run current chunk' button, and clicking it will run your code chunk and output whatever it would in the final R markdown document. This is a great way to tweak figures and codes as it avoids the need to compile the entire document to check if you managed to change the lines from 'black' to 'blue' in your plot.

4.3 So now what do I do with R Markdown?

You do science and you write it down!

In all seriousness though, this document was only meant to introduce you to R markdown, and to make the case that you should use it for your coursework. A

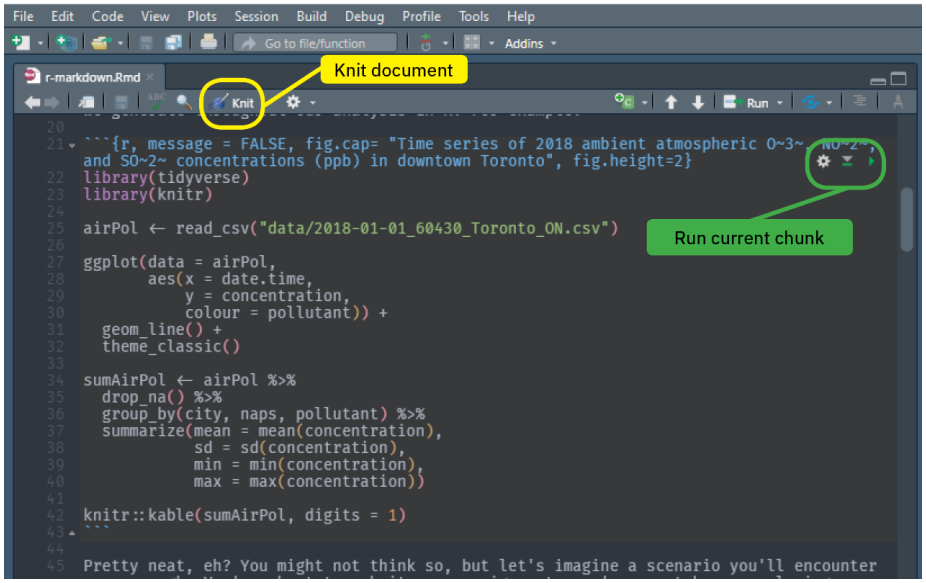


Figure 4.2: How this document, the one you're reading, appeared in RStudio; to see the final results scroll up to Figure 1. Note the "knit" and "run current chunk" buttons.

couple of the most useful elements are talked about below, and there is a wealth of helpful resources for formatting your documents. Just remember to keep it simple, there's no need to reinvent the wheel. The default R markdown outputs are plenty fine with us.

4.3.1 R Markdown resources and further reading

There's a plethora of helpful online resources to help hone your R markdown skills. We'll list a couple below (the titles are links to the corresponding document):

- Chapter 2 of the *R Markdown: The Definitive Guide* by Xie, Allair & Grolemund (2020). This is the simplest, most comprehensive, guide to learning R markdown and it's available freely online.
- *The R markdown cheat sheet*, a great resource with the most common R markdown operations; keep on hand for quick referencing.
- *Bookdown: Authoring Books and Technical Documents with R Markdown* (2020) by Yihui Xie. Explains the `bookdown` package which greatly expands the capabilities of R markdown. For example, the table of contents of this document is created with `bookdown`.

4.3.2 R code chunk options

You can specify a number of options for an individual R code chunk. You include these at the top of the code chunk. For example the following code tells markdown you're running code written in `R`, that when you compile your document this code chunk should be evaluated, and that the resulting figure should have the caption "Some Caption." A list of code chunk options is shown below:

````
```{r, eval = FALSE, fig.cap = "Some caption"}

# some code to generate a plot worth captioning.

```
````

option	default	effect
eval	TRUE	whether to evaluate the code and include the results
echo	TRUE	whether to display the code along with its results
warning	TRUE	whether to display warnings
error	FALSE	whether to display errors
message	TRUE	whether to display messages
tidy	FALSE	whether to reformat code in a tidy way when displaying it
fig.width	7	width in inches for plots created in chunk
fig.height	7	height in inches for plots created in chunk
fig.cap	NA	include figure caption, must be in quotation makrs ("")

4.3.3 Inserting images

Images not produced by R code can easily be inserted into your document. The markdown code isn't R code, so between paragraphs of bodytext insert the following code. Note that compiling to PDF, the LaTeX call will place your image in the "optimal" location, so you might find your image isn't exactly where you though it would be. A quick google search can help you out if this is problem.

```
![Caption for the picture.](path/to/image.png){width=50%, height=50%}
```

Note that in the above the use of image atributes, the `{width=50%, height=50%}` at the end. This is how you'll adjust the size of your image. Other dimensions you can use include `px`, `cm`, `mm`, `in`, `inch`, and `%`.

4.3.4 Generating Tables

There's multiple methods to create tables in R markdown. Assuming you want to display results calculated through R code, you can use the `kable()` function. Please consult Chapter 10 of the *R Markdown Cookbook* for additional support.

Alternatively, if you want to create simple tables manually use the following code in the main body, outside of an `R` code chunk. You can increase the number of rows/columns and the location of the horizontal lines. To generate more complex tables, see the `kable()` function and the `kableExtra` package.

```
Header 1 | Header 2| Header 3
---------|---------|---------|
Row 1    | Data    | Some other Data
Row 2    | Data    | Some other Data
---------|---------|---------|
```

Header 1	Header 2	Header 3
Row 1	Data	Some other Data
Row 2	Data	Some other Data

4.3.5 Spellcheck in R Markdown

While writing an R markdown document in R studio, go to the `Edit` tab at the top of the window and select `Check Spelling`. You can also use the `F7` key as a shortcut. The spell checker will literally go through every word it thinks you've misspelled in your document. You can add words to it so your spell checker's utility grows as you use it. **Note** that the spell check with also check your R code; be wary of changing words in your code chunks because you may get an error down the line.

4.3.6 R markdown syntax

- **Inline formatting**; *which* is $_{\text{used}}$ to $^{\text{format}}$ `your text`.

1. Numbered lists

- Normal lists
  - Lists
- **Block-level elements**, i.e. you're section headers

Example R markdown syntax used for formatting shown above:

```
- **Inline formatting**; *which* is ~used~ to ^format^ `your text`.
1. Numbered lists
- Normal lists
  - Lists

- **Block-level elements**, i.e. your section headers

# Headers
## Headers
### Headers
```

Chapter 5

R Tutorial Exercise

With the information presented in Chapters 1 to 4 you have the skills to start your data analysis. We've create a brief tutorial that covers the major elements introduced. At the end of this tutorial you'll have visualized a small subset of real Environment and Climate Change Canada (ECCC) National Airborne Pollution Surveillance Program (NAPS) data. More importantly, you'll have a properly setup project with working code and rmarkdown documents that you can recycle and re-purpose for your upcoming course work. After all, a beautiful aspect of coding is recycling it in future work to save you hassle.

In brief, the tutorial tasks are:

1. Copy the template project from the GitHub repository here; there are instructions on downloading on the repo's `README`.
2. Install the following packages if you haven't already, you can copy the code below and run it in the console:

- `tidyverse`
- `rmarkdown`
- `tinytex`: needed to generate PDF files, more info here

```
install.packages("tidyverse")
install.packages("rmarkdown")
install.packages("tinytex")

tinytex::install_tinytex() # This will take ~5 mins, so grab a coffee
```

3. Follow the instructions written in the template PDF to modify the `rmarkdown` file to analyze a dataset of your choice.

5.1 Expected outcome

There's a lot of upfront work with this tutorial, but if you've completed it successfully and generated your own markdown file analyzing your select dataset you'll be well on your way to tackling the upcoming course labs/work as you'll have:

- Created a working *RStudio* project, which you can copy and reuse for your future projects.
- Working *rmarkdown* document, similar to the reports you'd hand in during class.
- Working R code showcasing basics of `ggplot`.

Section 2: Data Analysis in R

Chapter 6

Intro to Data Analysis

This section will teach you **how** to use R to meet your data analysis needs using a common workflow. Whether it takes 10 minutes or 10 hrs, *you'll use this workflow for every data analysis project*. By explicitly understanding the workflow steps, and how to execute them in R, you'll be more than capable of expanding the limited tools learned from this book to any number of data analysis projects you'll soon encounter.

The explicit workflow we'll be teaching was originally described by Wickham and Grolemund, and consists of six key steps:

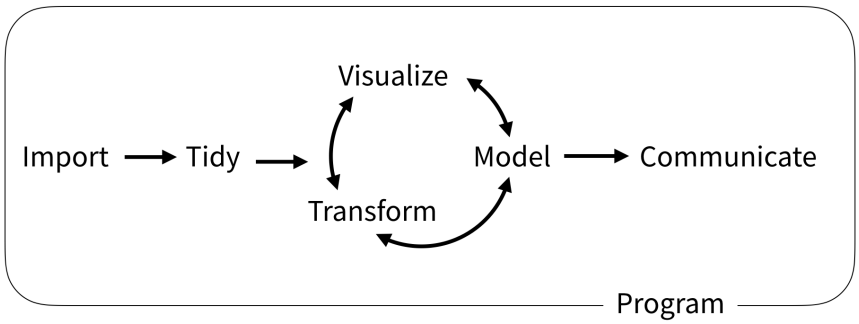


Figure 6.1: Data science workflow describes by Wickham and Grolemund; image from *R for Data Science*, Wickham and Grolemund (2021)

- **Import** is the first step and consist of getting your data into R. Seems obvious, but doing it correctly will save you time and headaches down the line.

- **Tidy** refers to organizing your data in a *tidy* manner where each variable is a column, and each observation a row. This is often the least intuitive part about working with R, especially if you've only used Excel, but it's critical. If you don't tidy your data, you'll be fighting it every step of the way.
- **Transform** is anything you do to your data including any mathematical operations or narrowing in on a set of observations. It's often the first stage of the cycle as you'll need to transform your data in some manner to obtain a desired plot.
- **Visualize** is any of the plots/graphics you'll generate with R. Take advantage of R and plot often, it's the easiest way to spot an errors.
- **Model** is an extension of mathematical operations to help understand your data. The *linear regressions* needed for a calibration curve are an example of a model.
- **Communicate** is the final step and is where you share the *knowledge* you've squeezed out of the information in the original data.

The *Transform*, *Visualize*, and *Model* cycle exists because these steps often feed into one another. For example, you'll often transform your data, make a quick model, then visualize it to see how it performs. Other times, you'll visualize your data to see what type of model can explain it, and if any transformations are necessary. This is the beauty of R (and coding in general). Once you've setup everything, these steps are fairly simple to execute allowing you to quickly explore your data from a number of different angles. The next section will explore the theory (the **why**) behind these steps, and introduce some tools you can use to better explore your data.

6.1 Further Reading

In case it hasn't been apparent enough, this entire endeavour was inspired by the *R for Data Science* reference book by Hadley Wickham and Garrett Grolemund. Every step described above is explored in more detail in their book, which can be read freely online at https://r4ds.had.co.nz/. We strongly encourage you to read through the book to supplement your R data analysis skills.

Chapter 7

Importing data into R

Unlike *Excel*, you can't copy and paste your data into R (or RStudio). Instead you need to *import* your data into R so you can work with it. This chapter will discuss how your data is stored, and how to import it into R (with some accompanying nuances).

7.1 How data is stored

While there are a myriad of ways data is stored, notably raw instrument often record results in a proprietary vendor format, the data you're likely to encounter in an undergraduate lab will be in the form of a `.csv` or *comma-separated values* file. As the name implies, values are separated by commas (go ahead and open any `.csv` file in any text editor to observe this). Essentially you can think of each line as a row and commas as separating values into columns, which is exactly how R and *Excel* handle `.csv` files.

7.2 read_csv

Importing a `.csv` file into R simply requires the `read.csv` or the `read_csv` function from tidyverse. The first variable is the most important as it's the file path. Recall that R, unless specified, uses relative referencing. So in the example below we're importing the `ATR_plastics.csv` from the `data` sub-folder in our project by specifying `"data/ATR_plastics.csv"` and assigning it to the variable `atr_plastics`. Note the inclusion of the file extension.

```
atr_plastics <- read_csv("data/ATR_plastics.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   wavenumber = col_double(),
##   EPDM = col_double(),
##   Polystyrene = col_double(),
##   Polyethylene = col_double(),
##   `Sample: Shopping bag` = col_double()
## )
```

A benefit of using `read_csv` is that it prints out the column specifications with each column's name (how you'll reference it in code) and the column value type. Columns can have different data types, but a data type must be consistent within any given column. Having the columns specifications is a good way to ensure R is correctly reading your data. The most common data types are:

- **int** for integer values (*-1,1, 2, 10, etc.*)
- **dbl** for doubles or real numbers (*-1.20, 0.0, 1.200, 1e7, etc.*)
- **chr** for character vectors or strings (*"A," "chemical," "Howdy ma'am," etc.*)
  - note numbers can be encoded as strings, so while you might read "1" as a number, R treats it as a character, limiting how you can use this value.
- **lgl** for logical values, either `TRUE` or `FALSE`

We can also quickly inspect either through the *Environment* pane in *RStudio* or quickly with the `head()` function. Note the column specifications under the column name.

```
head(atr_plastics)
```

```
## # A tibble: 6 x 5
##   wavenumber  EPDM Polystyrene Polyethylene `Sample: Shopping bag`
##        <dbl> <dbl>       <dbl>        <dbl>                  <dbl>
## 1       550. 0.212      0.0746     0.000873                 0.0236
## 2       551. 0.212      0.0746     0.000834                 0.0238
## 3       551. 0.213      0.0745     0.000819                 0.0239
## 4       552. 0.213      0.0745     0.000825                 0.0239
## 5       552. 0.214      0.0745     0.000868                 0.0240
## 6       553. 0.214      0.0746     0.000949                 0.0240
```

Note how the first line of the `ATR_plastics.csv` has been interpreted as columns names (or *headers*) by R. This is common practice, and gives you a

handle by which you can manipulate your data. If you did not intend for R to interpret the first row as headers you can suppress this with the additional argument `col_names = FALSE`.

```
head(read_csv("data/atr_plastics.csv", col_names = FALSE))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_character(),
##   X2 = col_character(),
##   X3 = col_character(),
##   X4 = col_character(),
##   X5 = col_character()
## )
```

```
## # A tibble: 6 x 5
##   X1         X2        X3          X4           X5
##   <chr>      <chr>     <chr>       <chr>        <chr>
## 1 wavenumber EPDM      Polystyrene Polyethylene Sample: Shopping bag
## 2 550.0952   0.2119556 0.07463058  0.000873196  0.02364882
## 3 550.5773   0.2124079 0.07455246  0.000834192  0.02382648
## 4 551.0594   0.2128818 0.07450471  0.000819447  0.02387163
## 5 551.5415   0.2133267 0.07449704  0.000825491  0.02391921
## 6 552.0236   0.2137241 0.07452058  0.000868397  0.02396947
```

Note in the example below that since the headers are now considered data, the entire column is interpreted as character values. This will happen if a single non-numeric character is introduced in the column, so beware of typos when recording data! If we wanted to skip rows (i.e. to avoid blank rows at the top of our `.csv`), we can use the `skip = n` to skip n rows:

```
head(read_csv("data/atr_plastics.csv", col_names = FALSE, skip = 1))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double()
## )
```

```
## # A tibble: 6 x 5
##       X1    X2     X3       X4     X5
##    <dbl> <dbl>  <dbl>    <dbl>  <dbl>
## 1  550. 0.212 0.0746 0.000873 0.0236
## 2  551. 0.212 0.0746 0.000834 0.0238
## 3  551. 0.213 0.0745 0.000819 0.0239
## 4  552. 0.213 0.0745 0.000825 0.0239
## 5  552. 0.214 0.0745 0.000868 0.0240
## 6  553. 0.214 0.0746 0.000949 0.0240
```

7.2.1 Tibbles vs. data frames

Quick eyes will notice the first line outputted above is `# A tibble: 6 x 5`. `tibbles` are a variation of `data.frames` introduced in section one, but built specifically for the `tidyverse` family of packages. While `data.frames` and `tibbles` are often interchangeable, it's important to be aware of the difference in case you do run into a rare conflict. In these situations you can readily transform a `tibble` into a `data.frame` by coercion with the `as.data.frame()` function, and vice-versa with the `as_tibble()` function.

```
class(as.data.frame(atr_plastics))
```

```
## [1] "data.frame"
```

7.3 Importing other data types

There are other functions to import different types of tabular data which all function like `read_csv`, such as `read_tsv` for tab-separate value files (`.tsv`) and `read_excel` and `read_xlsx` from the `readxl` package to import *Excel* files. Note most *Excel* files have probably been formatted for legibility (i.e. merged columns), which can lead to errors when importing into R. If you plan on importing *Excel* files, it's probably best to open them in *Excel* to remove any formatting, and then save as `.csv` for smoother importing into R.

7.4 Saving data

As you progress with your analysis you may want to save intermediate or final datasets. This is readily accomplished using the `write.csv` (base R) or `write_csv` (tidyverse) functions. Similar rules apply to how we used `read_csv`, but now the second argument specifies the save location and file name, the first

argument is which `tibble`/`data.frame` we're saving. Note that R *will not* create a folder this way, so if you're saving to a sub-folder you'll have to make sure it exists or create it yourself.

```
write_csv(atr_plastics, "data/ATRSaveExample.csv")
```

A benefit of `write_csv` is that it will always save in `UTF-8` encoding and ISO8601 time format. This standardization makes it easier to share your `.csv` files with collaborators/yourself.

7.5 Further Reading

See Chapters 10 and 11 of *R for Data Science* for some more details on `tibbles` and `read_csv`.

Chapter 8

Tidying your data

You might not have explicitly thought about how you store your data, whether working in *Excel* or elsewhere. Data is data after all. But having your data organized in a systematic manner that is conducive to your goal is paramount for working not only with R, but all of your experimental data. This chapter will introduce the concept of *tidy* data, and some of the tools of the *dplyr* package to get there. Lastly we'll offer some tips for how you should record *your* data in the lab. A bit of foresight and consistency can eliminate hours of tedious work down the line.

8.1 What is tidy data?

Tidy data has "…each variable in a column, and each observation in a row…" (Wickham 2014) This may seem obvious to you, but let's consider how data is often recorded in lab, as exemplified in Figure 8.1A. Here the instrument response of two chemicals (*A* and *B*) for two samples (*blank* and *unknown*) are recorded. Note how the samples are on each row and the chemical are columns. However, someone else may record the same data differently as shown in Figure 8.1B, with the samples occupying distinct columns, and the chemical rows. Either layout may work well, but analyzing both would require re-tooling your approach. This is where the concept of *tidy* data comes into play. By reclassifying our data into *observations* and *variables* we can restructure out data into a common format: the *tidy* format (Figure 8.1C).

In the *tidy* or *long* format, we reclassified out data into three variables (*Sample*, *Chemical*, and *Reading*). This makes the observations clearer as now we know we measured two chemicals (*A* and *B*) in two samples (*blank* and *unknown*) and we've explicitly declared the *Reading* variable for our measured instrument response, which was only implied in the original layouts. Moreover, we can

A.

Sample	Chemical A	Chemical B
blank	0	0
unknown	1	2

B.

Chemical	blank	unknown
A	0	1
B	0	2

C.

Sample	Chemical	Reading
blank	A	0
blank	B	0
unknown	A	1
unknown	B	2

Figure 8.1: (A and B) The same data can be recorded in multiple formats. (C) The same data in the tidy format. Note how the tidy data typically has more rows, hence why it's sometimes refered to as 'long' data.

read across a row to get the gist of one data point (i.e. "Our blank has a reading of 0 for Chemical A"). Again we haven't changed any information, we've simply reorganized our data to be clearer, consistent, and compatible with the `tidyverse` suit of tools.

This might seem pedantic now, but as you progress you'll want to reuse code you've previously written. This is greatly facilitated by making every data set as consistently structured as possible, and the *tidy* format is an ideal starting place.

8.2 Tools to tidy your data

Now one of the more laborious parts of data science is tidying your data. If you can follow the tips in the Tips for recording data section, but the truth is you often won't have control. To this end, the `tidyverse` offers several tools, notable `dplyr` (pronounces 'd-pliers'), to help you get there.

Let's revisit our spectroscopy data from the previous chapter:

```
atr_plastics <- read_csv("data/ATR_plastics.csv")

# This just outputs a table you can explore within your browser
DT::datatable(atr_plastics)
```

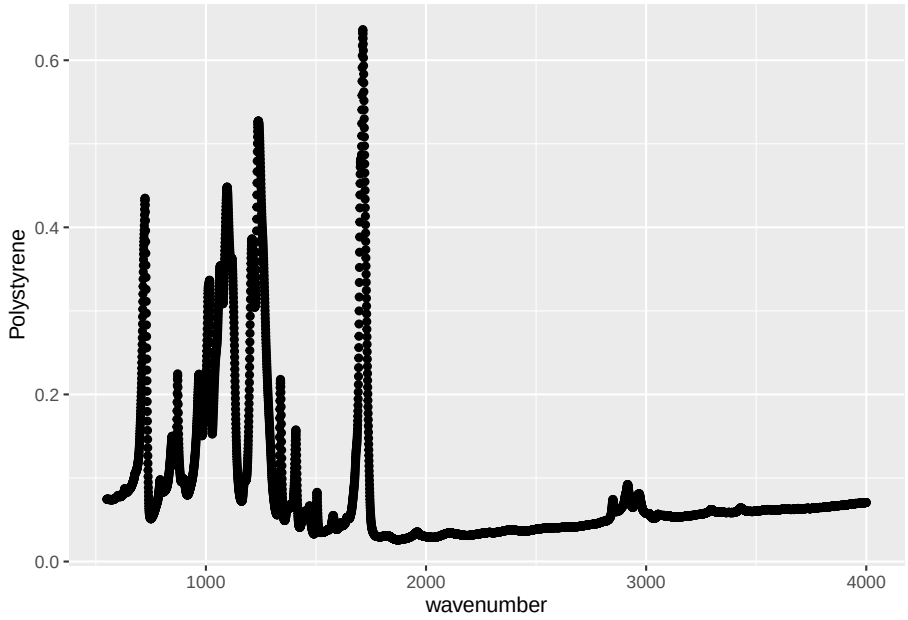
Show 10 entries Search:

	wavenumber	EPDM	Polystyrene	Polyethylene	Sample: Shopping bag
1	550.0952	0.2119556	0.07463058	0.000873196	0.02364882
2	550.5773	0.2124079	0.07455246	0.000834192	0.02382648
3	551.0594	0.2128818	0.07450471	0.000819447	0.02387163
4	551.5415	0.2133267	0.07449704	0.000825491	0.02391921
5	552.0236	0.2137241	0.07452058	0.000868397	0.02396947
6	552.5057	0.2140997	0.07455528	0.000949017	0.02402323
7	552.9879	0.2145173	0.07458498	0.001008869	0.02408946
8	553.47	0.2150495	0.07458818	0.0010252	0.02413738
9	553.9521	0.2156501	0.07456694	0.001053734	0.02419313
10	554.4342	0.2163207	0.07452466	0.001066902	0.02424033

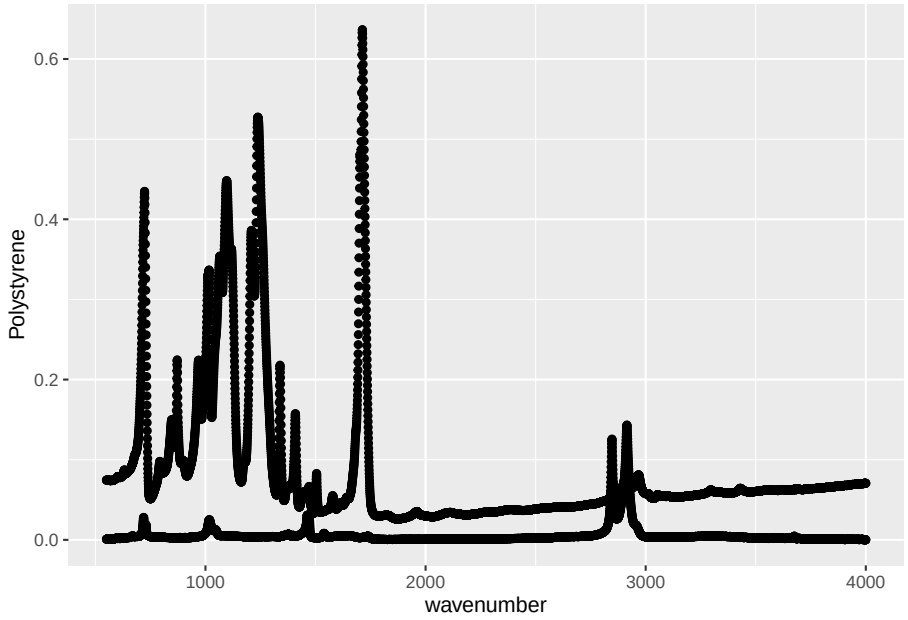
Showing 1 to 10 of 7,157 entries Previous 1 2 3 4 5 … 716 Next

As we can see this our ATR spectroscopy results of several plastics, as recorded for a *CHM 317* lab, is structured similarly to the example in Figure 8.1A. The ATR absorbance spectra of the four plastics are recorded in separate columns. Again, this format makes intuitive sense when recording in the lab, and for working in Excel, but isn't the friendliest with R. In the example below we can only specify one `y` value for `ggplot` to plot. In our example it's the absorbance spectrum of `Polystyrene`. However, if wanted to plot the other spectra for comparison, we'd need to repeat our `geom_point` call.

```
# Plotting Polystyrene absorbance spectra
ggplot(data = atr_plastics,
       aes( x = wavenumber,
            y = Polystyrene)) +
  geom_point()
```




```
# Plotting Polystyrene and Polyethylene absorbance spectra
ggplot(data = atr_plastics,
       aes( x = wavenumber,
            y = Polystyrene)) +
  geom_point() +
  geom_point(data = atr_plastics,
             aes(x = wavenumber,
                 y = Polyethylene))
```



8.2.1 Making data 'longer'

While code above works, it's not particularly handy and undermines much of the utility of `ggplot` because the data *isn't* tidy. Fortunately the `pivot_longer` function can easily restructure our data into the *long* format to work with `ggplot`. Let's demonstrate that:

```
atr_long <- atr_plastics %>%
  pivot_longer(cols = -wavenumber,
               names_to = "sample",
               values_to = "absorbance")

head(atr_long)
```

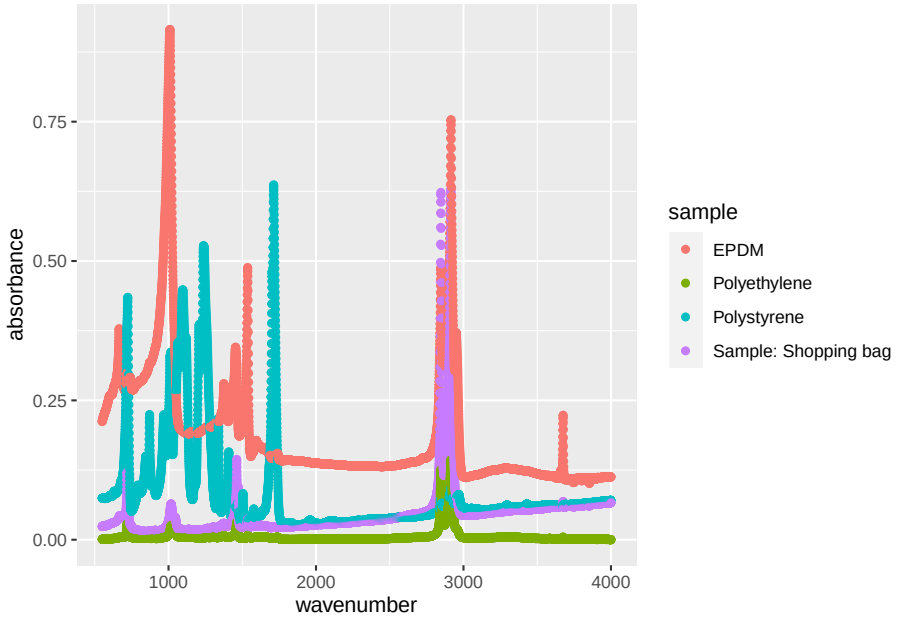
```
## # A tibble: 6 x 3
##   wavenumber sample                absorbance
##        <dbl> <chr>                      <dbl>
## 1       550. EPDM                     0.212
## 2       550. Polystyrene              0.0746
## 3       550. Polyethylene             0.000873
## 4       550. Sample: Shopping bag     0.0236
## 5       551. EPDM                     0.212
## 6       551. Polystyrene              0.0746
```

Let's break down the code we've executed via the `pivot_longer` function:

1. `cols = -wavenumber` specifies that we're selecting every other column *but* wave number.
    - we could have just as easily specified each column individually using `cols = c("EPDM",...)` but it's easier to use `-` to specify what we *don't* want to select.
2. `names_to = "sample"` specifies that the column header (i.e. names) be converted into an observation under the `sample` column.
3. `values_to = "absorbance"` specifies that the absorbance values under each of the selected headers be placed into the `aborsbance` column.

Now that we've reclassified out data into the 'longer,' we can exploit the explicitly introduced *sample* variable to easily plot all of our spectra:

```
ggplot(data = atr_long,
       aes(x = wavenumber,
           y = absorbance,
           colour = sample)
       ) +
  geom_point()
```



We'll talk more about `ggplot` in the *Visualizations* chapter, but for now you can understand how our code could scale to accommodate any number of different

samples, whereas the previous attempt would require an explicit call to each column.

`pivot_longer` has many other features that you can take advantage of. We highly recommend reading the examples listed on the pivot_longer page to get a better sense of the possibilities. For example it's common to record multiple observations in a single column header, i.e. `Chemical_A_0_mM`. We can exploit common naming conventions like this to easily split up these observations as shown below.

```
head(example)
```

```
##   wavelength_nm Chemical_A_0_mM Chemical_A_1_mM Chemical_B_0_mM Chemical_B_1_mM
## 1           488               0               1               2              NA
## 2           572               0               5               7              20
```

```
example_long <- example %>%
 pivot_longer(
   cols = starts_with("Chemical"),
   names_prefix = "Chemical_",
   names_to = c("Chemical", "Concentration", "Conc_Units"),
   names_sep = "_",
   values_to = "Absorbance",
   values_drop_na = TRUE
 )

head(example_long)
```

```
## # A tibble: 6 x 5
##   wavelength_nm Chemical Concentration Conc_Units Absorbance
##           <dbl> <chr>    <chr>         <chr>            <dbl>
## 1           488 A        0             mM                   0
## 2           488 A        1             mM                   1
## 3           488 B        0             mM                   2
## 4           572 A        0             mM                   0
## 5           572 A        1             mM                   5
## 6           572 B        0             mM                   7
```

8.2.2 Making data 'wider'

Sometimes packages or circumstances will require you reformat your data into a matrix or 'wide' format (notable the `matrixStats` and `matrixTests` packages). You can accomplish this using the `pivot_wider` function, which operates inverse to the `pivot_longer` function described above. For example the input

`names_from` is used to specify which variables are to be converted to headers. You can read up on the `pivot_wider` function here

8.2.3 Seperating columns

Sometimes your data has already been recorded in a tidy-ish fashion, but there may be multiple observations recorded under one apparent variable, something like `1 mM` for concentration. As it stands we cannot easily access the numerical value in the concentration recording because R will encode this as a string due to the `mM`. We can **separate** data like this using the `seperate` function, which operates similarly to how `pivot_longer` breaks up headers.

```
# Example with multiple encoded observations
sep_example
```

```
##          sample reading
## 1  Toronto_O3_1      10
## 2  Toronto_O3_2      22
## 3 Toronto_NO2_1      30
```

The example above is something you'll come across in the lab, most often with the sample names you'll pass along to your TA. You've crammed as much information as possible into that name so you and them know exactly what's being analyzed. In this example, the sample name contains the location (`Toronto`), the chemical measured (`O3` or `NO2`) and the replicate number (i.e. `1`). Using the `seperate` function we can split up these three observations so we can properly group our data later on in our analysis.

```
# Separating observations

sep_example %>%
  separate(
    col = sample,
    into = c("location", "chemical", "replicateNum"),
    sep = "_",
    remove = TRUE,
    convert = TRUE)
```

```
##   location chemical replicateNum reading
## 1  Toronto       O3            1      10
## 2  Toronto       O3            2      22
## 3  Toronto      NO2            1      30
```

Again, let's break down what we did with the `separate` function:

1. `col = sample` specifies we're selecting the `sample` column
2. `into = c(...)` specifies what columns we're separating our name into.
3. `sep = "_"1` specifies that each element is separated by an underscore (_); you can use `sep = " "` if they were separated by spaces.
4. `remove = TRUE` removes the original sample column, no need for duplication; setting this to `FALSE` would keep the original column.
5. `convert = TRUE` converts the new columns to the appropriate data format. In the original column ,the replicate number is a character value because it's part of a string, `convert` ensures that it'll be converted to a numerical value.

Again it's paramount to **be consistent when recording data**.

8.2.4 Uniting/combining columns

The opposite of the `separate` function is the `unite` function. You'll use it far less often, but you should be aware of it as it may come in handy. You can use it for combining strings together, or prettying up tables for publication/presentations. You can read more about the `unite` function here

8.2.5 Renaming columns/headers

Sometimes a name is lengthy, or cumbersome to work with in R. While something like `This_is_a_valid_header` is valid and compatible with R and tidyverse functions, you may want to change it to make it easier to work with (i.e. less typing). Simply use the `rename` function:

```
colnames(badHeader)
```

```
## [1] "UVVis_Wave_Length_nM" "Absorbance"
```

```
colnames(rename(badHeader, wavelength_nM = UVVis_Wave_Length_nM))
```

```
## [1] "wavelength_nM" "Absorbance"
```

8.2.6 Rounding numbers

If you want to round the numbers in your data to account for significant figures or something, you can do so using the `round` function.

```
head(example)
```

```
##   measurement absorbance     conc
## 1           A    123.123 1.100000
## 2           B    300.000 3.000022
## 3           C    175.547 1.750000
```

```
# rounding 'conc' column to 1 decimal.

example %>%
  mutate_at(vars(conc), round, 1)
```

```
##   measurement absorbance conc
## 1           A    123.123  1.1
## 2           B    300.000  3.0
## 3           C    175.547  1.8
```

8.3 Tips for recording data

In case you haven't picked up on it, tidying data in R is much easier if the data is recorded consistently. You can't always control how your data will look, but in the event that you can (i.e. your inputting the instrument readings into *Excel* on the bench top) here are some tips to make your life easier:

- *Be consistent.* If you're naming your samples make sure they all contain the same elements in the same order. The sample names `Toronto_O3_1` and `Toronto_O3_2` can easily be broken up as demonstrated in [Separating columns]; `O3_Toronto_1`, `TorontoO32`, and `Toronto_1` can't be.
- *Use as simple as possible headers.* Often you'll be pasting instrument readings into one `.csv` using *Excel* on whatever computer records the instrument readings. In these situations it's often much easier to paste things in columns. Recall the capabilities of `pivot_longer` and how you can break up names as described in Making data 'longer'. `Chemical_A_1` and `Chemical_B_2` are headers that are descriptive for your sample and can be easily pivoted into their own columns. `Chemical A 1 ( I think?!)` is a header isn't.
- *Make sure data types are consistent within a column.* This harks back to the Importing data into R chapter, but a single non-numeric character can cause R to misinterpret an entire column leading to headaches down the line.
- *Save your data in UTF-8 format.* Excel and other programs often allow you to export your data in a variety of `.csv` encodings, but this can affect how R reads when importing your data. Make sure you select `UTF-8` encoding when exporting your data.

8.4 Further reading

As always, the *R for Data Science* book goes into more detail on all of the elements discussed above. For the material covered here you may want to read Chapter 9: Tidy Data.

Chapter 9

Transform: dplyr and data manipulation

Transformation encompasses any steps you take to manipulate, reshape, refine, or transform your data. We've already touched upon some useful transformation functions in previous example code snippets, such as the `mutate` function for adding columns. This section will explore some of the most useful functionailities of the `dplyr` package, explicitly introduce the pipe operator `%>%`, and showcase how you can leverage these tools to quickly manipulate your data.

The benchmark `dplyr` functions are :

- `mutate()` to create new columns/variables from existing data
- `arrange()` to reorder rows
- `filter()` to refine observations by their values (in other words by row)
- `select()` to pick variables by name (in other words by column)
- `summarize` to collapse many values down to a single summary.

We'll go through each of these functions, but we highly recommend you read Chapter 3: Data Transformation from *R for Data Science* to get a more comprehensive breakdown of these functions. Note that the information here is based on a `tidyverse` approach, but this is only one way of doing things. Please see the Further reading section for links to other equally suitable approaches to data transformation.

Let's explore the functionality of `dplyr` using some flame absorption/emission spectroscopy (FAES) data from a *CHM317* lab. This data represents the emission signal of five sodium (Na) standards measured in triplicate:

```
# Importing using tips from Import chapter
FAES <- read_csv(file = "data/FAESdata.csv") %>% # see section on Pipe
  pivot_longer(cols = -std_Na_conc,
               names_to = "replicate",
               names_prefix = "reading_",
               values_to = "signal") %>%
  separate(col = std_Na_conc,
           into = c("type", "conc_Na", "units"),
           sep = " ",
           convert = TRUE)

DT::datatable(FAES)
```

Show 10 entries Search:

	type	conc_Na	units	replicate	signal
1	blank	0	mg/L	1	1348.5051
2	blank	0	mg/L	2	1304.0702
3	blank	0	mg/L	3	1395.7714
4	standard	0.1	mg/L	1	2947.3097
5	standard	0.1	mg/L	2	2924.3988
6	standard	0.1	mg/L	3	2927.2867
7	standard	0.2	mg/L	1	4446.4036
8	standard	0.2	mg/L	2	4453.1066
9	standard	0.2	mg/L	3	4416.439
10	standard	0.3	mg/L	1	6235.1603

Showing 1 to 10 of 15 entries Previous 1 2 Next

Note the use of `convert = TRUE` in the `separate()` call. This runs a type convert on new columns. If we didn't include this, the `conc_Na` column would be of type character because the numbers originated from a string. `convert()` ensures they're converted to numeric. **Always use `convert = TRUE`** when you separate columns.

9.1 Selecting by row or value

`filter()` allows up to subset our data based on observation (row) values.

```
filter(FAES, conc_Na == 0)
```

```
## # A tibble: 3 x 5
```

```
##   type  conc_Na units replicate signal
##   <chr>   <dbl> <chr> <chr>      <dbl>
## 1 blank       0 mg/L  1          1349.
## 2 blank       0 mg/L  2          1304.
## 3 blank       0 mg/L  3          1396.
```

Note how we need to pass logical operations to `filter()`. In the above code, we used `filter()` to get all rows where the concentration of sodium is equal to 0 (`== 0`). Note the presence of two equal signs (`==`). In R one equal sign (`=`) is used to pass an argument, two equal signs (`==`) is the logical operation "is equal" and is used to test equality (i.e. that both sides have the same value). A frequent mistake is to use `=` instead of `==` when testing for equality.

9.1.1 Logical oeprators

`filter()` can use other *relational* and *logical* operators, or combinations thereof, to improve your sub-setting. Relational operators compare values and logical operators carry out Boolean operations (TRUE or FALSE). Logical operators are used to combine multiple relational operators… let's just list what they are and how we can use them:

Operator	Type	Description
>	relational	Less than
<	relational	Greater than
<=	relational	Less than or equal to
>=	relational	Greater than or equal to
==	relational	Equal to
!=	relational	Not equal to
&	logical	AND
!	logical	NOT
|	logical	OR
is.na()	function	Checks for missing values, TRUE if NA

- Selecting all signals below a threshold value

```
filter(FAES, signal < 4450)
```

```
## # A tibble: 8 x 5
##   type     conc_Na units replicate signal
##   <chr>      <dbl> <chr> <chr>      <dbl>
## 1 blank        0   mg/L  1          1349.
## 2 blank        0   mg/L  2          1304.
## 3 blank        0   mg/L  3          1396.
## 4 standard     0.1 mg/L  1          2947.
```

```
## 5 standard     0.1 mg/L  2          2924.
## 6 standard     0.1 mg/L  3          2927.
## 7 standard     0.2 mg/L  1          4446.
## 8 standard     0.2 mg/L  3          4416.
```

- Selecting signals between values

```
filter(FAES, signal >= 4450 & signal < 8150)
```

```
## # A tibble: 6 x 5
##   type     conc_Na units replicate signal
##   <chr>      <dbl> <chr> <chr>      <dbl>
## 1 standard     0.2 mg/L  2          4453.
## 2 standard     0.3 mg/L  1          6235.
## 3 standard     0.3 mg/L  2          6207.
## 4 standard     0.3 mg/L  3          6267.
## 5 standard     0.4 mg/L  2          8141.
## 6 standard     0.4 mg/L  3          8106.
```

- Selecting all other replicates other than replicate 2

```
filter(FAES, replicate != 2)
```

```
## # A tibble: 10 x 5
##    type     conc_Na units replicate signal
##    <chr>      <dbl> <chr> <chr>      <dbl>
##  1 blank        0   mg/L  1          1349.
##  2 blank        0   mg/L  3          1396.
##  3 standard     0.1 mg/L  1          2947.
##  4 standard     0.1 mg/L  3          2927.
##  5 standard     0.2 mg/L  1          4446.
##  6 standard     0.2 mg/L  3          4416.
##  7 standard     0.3 mg/L  1          6235.
##  8 standard     0.3 mg/L  3          6267.
##  9 standard     0.4 mg/L  1          8173.
## 10 standard     0.4 mg/L  3          8106.
```

- selecting the first standard replicate OR any of the blanks.

```
filter(FAES, (type == "standard" & replicate == 1) | (type == "blank"))
```

```
## # A tibble: 7 x 5
```

```
##   type     conc_Na units replicate signal
##   <chr>      <dbl> <chr> <chr>      <dbl>
## 1 blank        0   mg/L  1          1349.
## 2 blank        0   mg/L  2          1304.
## 3 blank        0   mg/L  3          1396.
## 4 standard     0.1 mg/L  1          2947.
## 5 standard     0.2 mg/L  1          4446.
## 6 standard     0.3 mg/L  1          6235.
## 7 standard     0.4 mg/L  1          8173.
```

- removing any missing values (`NA`) using `is.na()`. Note there are no missing values in our data set so nothing will be removed, if we removed the NOT operator (`!`) we would have selected all rows *with* missing values.

```
filter(FAES, !is.na(signal))
```

```
## # A tibble: 15 x 5
##    type     conc_Na units replicate signal
##    <chr>      <dbl> <chr> <chr>      <dbl>
##  1 blank        0   mg/L  1          1349.
##  2 blank        0   mg/L  2          1304.
##  3 blank        0   mg/L  3          1396.
##  4 standard     0.1 mg/L  1          2947.
##  5 standard     0.1 mg/L  2          2924.
##  6 standard     0.1 mg/L  3          2927.
##  7 standard     0.2 mg/L  1          4446.
##  8 standard     0.2 mg/L  2          4453.
##  9 standard     0.2 mg/L  3          4416.
## 10 standard     0.3 mg/L  1          6235.
## 11 standard     0.3 mg/L  2          6207.
## 12 standard     0.3 mg/L  3          6267.
## 13 standard     0.4 mg/L  1          8173.
## 14 standard     0.4 mg/L  2          8141.
## 15 standard     0.4 mg/L  3          8106.
```

These are just some examples, but you can combine the logical operators in any way that works for you. Likewise, there are multiple combinations that will yield the same result, it's up to you do figure out which works best for you.

9.2 Arranging rows

`arrange()` simple reorders the rows based on the value you passed to it. By default it arranges the specified values into ascending order. Let's arrange our signal in increasing by increasing order:

```
arrange( FAES, signal)
```

```
## # A tibble: 15 x 5
##    type     conc_Na units replicate signal
##    <chr>      <dbl> <chr> <chr>      <dbl>
##  1 blank        0   mg/L  2          1304.
##  2 blank        0   mg/L  1          1349.
##  3 blank        0   mg/L  3          1396.
##  4 standard     0.1 mg/L  2          2924.
##  5 standard     0.1 mg/L  3          2927.
##  6 standard     0.1 mg/L  1          2947.
##  7 standard     0.2 mg/L  3          4416.
##  8 standard     0.2 mg/L  1          4446.
##  9 standard     0.2 mg/L  2          4453.
## 10 standard     0.3 mg/L  2          6207.
## 11 standard     0.3 mg/L  1          6235.
## 12 standard     0.3 mg/L  3          6267.
## 13 standard     0.4 mg/L  3          8106.
## 14 standard     0.4 mg/L  2          8141.
## 15 standard     0.4 mg/L  1          8173.
```

Since our original `FAES` data is already arranged by increasing `cong_Na` and `replicate`, let's inverse that order by arranging `conc_Na` into descending order using the `desc()` function BUT arrange the `signal` values in:

```
# Note the order of precedence
arrange(FAES, desc(conc_Na), signal)
```

```
## # A tibble: 15 x 5
##    type     conc_Na units replicate signal
##    <chr>      <dbl> <chr> <chr>      <dbl>
##  1 standard     0.4 mg/L  3          8106.
##  2 standard     0.4 mg/L  2          8141.
##  3 standard     0.4 mg/L  1          8173.
##  4 standard     0.3 mg/L  2          6207.
##  5 standard     0.3 mg/L  1          6235.
##  6 standard     0.3 mg/L  3          6267.
##  7 standard     0.2 mg/L  3          4416.
##  8 standard     0.2 mg/L  1          4446.
##  9 standard     0.2 mg/L  2          4453.
## 10 standard     0.1 mg/L  2          2924.
## 11 standard     0.1 mg/L  3          2927.
## 12 standard     0.1 mg/L  1          2947.
## 13 blank        0   mg/L  2          1304.
```

```
## 14 blank        0   mg/L  1          1349.
## 15 blank        0   mg/L  3          1396.
```

Just note with `arrange()` that `NA` values will always be placed at the bottom, whether you use `desc()` or not.

9.3 Selecting by column or variable

`select()` allows you to readily select columns by name. Note however that it will always return a tibble, even if you only select one variable/column.

```
select(FAES, signal)
```

```
## # A tibble: 15 x 1
##    signal
##     <dbl>
##  1  1349.
##  2  1304.
##  3  1396.
##  4  2947.
##  5  2924.
##  6  2927.
##  7  4446.
##  8  4453.
##  9  4416.
## 10  6235.
## 11  6207.
## 12  6267.
## 13  8173.
## 14  8141.
## 15  8106.
```

You can also select multiple columns using the same helper functions describes in Importing data into R.

```
select(FAES, conc_Na:replicate)
```

```
## # A tibble: 15 x 3
##    conc_Na units replicate
##      <dbl> <chr> <chr>
##  1     0   mg/L  1
##  2     0   mg/L  2
```

```
##  3     0   mg/L  3
##  4     0.1 mg/L  1
##  5     0.1 mg/L  2
##  6     0.1 mg/L  3
##  7     0.2 mg/L  1
##  8     0.2 mg/L  2
##  9     0.2 mg/L  3
## 10     0.3 mg/L  1
## 11     0.3 mg/L  2
## 12     0.3 mg/L  3
## 13     0.4 mg/L  1
## 14     0.4 mg/L  2
## 15     0.4 mg/L  3
```

```
# Getting columns containing the character "p"
select(FAES, contains("p"))
```

```
## # A tibble: 15 x 2
##    type     replicate
##    <chr>    <chr>
##  1 blank    1
##  2 blank    2
##  3 blank    3
##  4 standard 1
##  5 standard 2
##  6 standard 3
##  7 standard 1
##  8 standard 2
##  9 standard 3
## 10 standard 1
## 11 standard 2
## 12 standard 3
## 13 standard 1
## 14 standard 2
## 15 standard 3
```

9.4 Adding new variables

`mutate()` allows you to add new variable (read columns) to your existing data set. It'll probably be the workhorse function you'll use during your data transformation as you can readily pass other functions and mathematical operators to it to transform your data. let's suppose that our standards were diluted by a factor of 10, we can add a new column for this:

```
mutate(FAES, "dil_fct" = 10)
```

```
## # A tibble: 15 x 6
##    type     conc_Na units replicate signal dil_fct
##    <chr>      <dbl> <chr> <chr>      <dbl>   <dbl>
##  1 blank        0   mg/L  1          1349.      10
##  2 blank        0   mg/L  2          1304.      10
##  3 blank        0   mg/L  3          1396.      10
##  4 standard     0.1 mg/L  1          2947.      10
##  5 standard     0.1 mg/L  2          2924.      10
##  6 standard     0.1 mg/L  3          2927.      10
##  7 standard     0.2 mg/L  1          4446.      10
##  8 standard     0.2 mg/L  2          4453.      10
##  9 standard     0.2 mg/L  3          4416.      10
## 10 standard     0.3 mg/L  1          6235.      10
## 11 standard     0.3 mg/L  2          6207.      10
## 12 standard     0.3 mg/L  3          6267.      10
## 13 standard     0.4 mg/L  1          8173.      10
## 14 standard     0.4 mg/L  2          8141.      10
## 15 standard     0.4 mg/L  3          8106.      10
```

We can also create multiple columns in the same `mutate()` call:

```
mutate(FAES, "dil_fct" = 10, "adj_signal" = signal * dil_fct)
```

```
## # A tibble: 15 x 7
##    type     conc_Na units replicate signal dil_fct adj_signal
##    <chr>      <dbl> <chr> <chr>      <dbl>   <dbl>      <dbl>
##  1 blank        0   mg/L  1          1349.      10     13485.
##  2 blank        0   mg/L  2          1304.      10     13041.
##  3 blank        0   mg/L  3          1396.      10     13958.
##  4 standard     0.1 mg/L  1          2947.      10     29473.
##  5 standard     0.1 mg/L  2          2924.      10     29244.
##  6 standard     0.1 mg/L  3          2927.      10     29273.
##  7 standard     0.2 mg/L  1          4446.      10     44464.
##  8 standard     0.2 mg/L  2          4453.      10     44531.
##  9 standard     0.2 mg/L  3          4416.      10     44164.
## 10 standard     0.3 mg/L  1          6235.      10     62352.
## 11 standard     0.3 mg/L  2          6207.      10     62074.
## 12 standard     0.3 mg/L  3          6267.      10     62666.
## 13 standard     0.4 mg/L  1          8173.      10     81731.
## 14 standard     0.4 mg/L  2          8141.      10     81412.
## 15 standard     0.4 mg/L  3          8106.      10     81062.
```

Couple of things to note:

1. The variable we're creating needs to be in quotation marks, hence `"dil_fct"` for our dilution factor variable
2. the variables we're referencing do not need to be in quotation marks; hence `signal` because this variable already exist.
3. Note the order of precedence: `dil_fct` is created first so we can reference in the second argument, we would get an error if we swapped the order.

9.4.1 Useful mutate function

There are a myriad of functions you can make use of with the mutate function. Here are some of the mathematical operators available in R:

function.	definition
+	additon
-	subtraction
*	multiplication
/	division
^	exponent; to the power off...
log()	returns the specified base-log; see also log10() and log2()

9.5 Group and summarize data

`summarize` effectively summarized your data based on functions you've passed to it. Looking at our `FAES` data we'd probably want the mean of the triplicate signals, alongside the standard deviation. Let's see what happens when we apply the summarize function straight up:

```
summarise(FAES, "mean" = mean(signal), "stdDev" = sd(signal))
```

```
## # A tibble: 1 x 2
##    mean stdDev
##   <dbl>  <dbl>
## 1 4620.  2475.
```

This doesn't look like what we wanted. What we got was the mean and standard deviation of *all* of the signals, regardless of the concentration of the standard. Also note how we've lost the other columns/variables and are only left with the mean and stdDev. This is all because we need to **group** our observations by a variable. We can do this by using the `group_by()` function.

```
groupedFAES <- group_by(FAES, type, conc_Na)
summarise(groupedFAES, "mean" = mean(signal), "stdDev" = sd(signal))
```

```
## `summarise()` has grouped output by 'type'. You can override using the `.groups` argument.
```

```
## # A tibble: 5 x 4
## # Groups:   type [2]
##   type     conc_Na  mean stdDev
##   <chr>      <dbl> <dbl>  <dbl>
## 1 blank        0   1349.   45.9
## 2 standard     0.1 2933.   12.5
## 3 standard     0.2 4439.   19.5
## 4 standard     0.3 6236.   29.6
## 5 standard     0.4 8140.   33.4
```

Here we've created a new data set, `groupedFAES`, that we grouped by the variables `type` and `conc_Na` so we could get the mean and standard deviation of each group. Note the multiple levels of grouping. For this data set we could have omitted the `type` variable, but in larger datasets you may have multiple groupings (i.e. from different location), so you can group by multiple variables to get smaller groups.

9.5.1 Useful summarize functions

We've used the `mean()` and `sd()` functions above, but there are a host of other useful functions you can use in conjunction with summarize. See **Useful Functions** in the `summarise()` documentation (enter `?summarise`) in the console.

9.6 The pipe: chaining functions together

With the tools presented here we could do a decent job analyzing our `FAES` data. Let's say we wanted to subtract the mean of the `blank` from each `standard` signal and then get summarize those results. It would look something like this:

```
blank <- filter(FAES, type == "blank")
meanBlank <- summarize(blank, mean(signal))
meanBlank <- as.numeric(meanBlank)

paste("The mean signal from the blank triplicate is:", meanBlank)
```

```
## [1] "The mean signal from the blank triplicate is: 1349.4489"
```

```
stds_1 <- filter(FAES, type == "standard")
stds_2 <- mutate(stds_1, "cor_sig" = signal - meanBlank)
stds_3 <- group_by(stds_2, conc_Na)
stds_4 <- summarize(stds_3, "mean" = mean(cor_sig), "stdDev" = sd(cor_sig))
stds_4
```

```
## # A tibble: 4 x 3
##   conc_Na  mean stdDev
##     <dbl> <dbl>  <dbl>
## 1     0.1 1584.   12.5
## 2     0.2 3089.   19.5
## 3     0.3 4887.   29.6
## 4     0.4 6791.   33.4
```

While the code above did it's job, it's certainly wasn't easy to type and certainly not easy to read. At every step of the way we've saved our updated data outputs to a new variable (`stds_1`, `stds_2`, etc.). However, most of these intermediates aren't important, and moreover the repetitive names clutter our code. As the code above is written, we've had to pay special attending to the variable suffix to make sure we're calling the correct data set as our code has progresses. An alternative would be to reassign the outputs back to the original variable name (i.e. `stds_1 <- mutate(stds_1, ...)`), but that doesn't solve the issue of readability as there's still redundant assigning.

A solution for this is the pipe operator `%>%` (pronounced "then"), an incredibly useful tool for writing more legible and understandable code. The pipe basically changes how you read code to emphasize the functions you're working with by passing the intermediate steps to hidden processes in the background. Re-writing the code above, we'd get something like:

```
meanBlank <- FAES %>%
  filter(type =="blank") %>%
  summarise(mean(signal)) %>%
  as.numeric()

paste("The mean signal from the blank triplicate is:", meanBlank)
```

```
## [1] "The mean signal from the blank triplicate is: 1349.4489"
```

Things may look a bit different, but our underlying code hasn't changed much. What's happening is the pipe operator passes the output to the first argument of the next function. So the output of `filter...` is passed to the first argument of `sumamrise...`, and the argument we specified in `summarise` is actually the *second* argument it receives. You're probably wondering how hiding stuff makes

your code more legible, but think of `%>%` as being equivalent to "then." We can read our code as:

> "Take the `FAES` dataset, *then* filter for `type == "blank"` *then* collapse the dataset to the mean `signal` value and *then* convert to numeric value *then* pass this final output to the new variable `meanBlank`."

Not only is the pipe less typing, but the emphasis is on the functions so you can better understand what you're doing vs. where all the intermediates are going. Extending our piping to the second batch of code we get:

```
stds <- FAES %>%
  filter(type == "standard") %>%
  mutate("cor_sig" = signal - meanBlank) %>%
  group_by(conc_Na) %>%
  summarize("mean" = mean(cor_sig), "stdDev" = sd(cor_sig))

stds
```

```
## # A tibble: 4 x 3
##   conc_Na  mean stdDev
##     <dbl> <dbl>  <dbl>
## 1     0.1 1584.   12.5
## 2     0.2 3089.   19.5
## 3     0.3 4887.   29.6
## 4     0.4 6791.   33.4
```

Same thing. The underlying code hasn't changed much, but it's much more legible and we can clearly see we're subtracting the `meanBlank` value from each measured signal then summarizing the corrected signals.

9.6.1 Notes on piping

The pipe is great and especially useful with *tidyverse* packages, but it does have some limitations:

- You can't easily extract intermediate steps. So you'll need to break up your pipping chain to output any intermediate steps you can.
- The benefit of piping is legibility; this goes away as you increase the number of steps as you lose track of what's going on. Keep the piping short and thematically similar.
- Pipes are linear, if you have multiple inputs or outputs you should consider an alternative approach.

9.7 Further reading

- Chapter 5: Data Transformation of *R for Data Science* for a deeper breakdown of `dplyr` and it's functionality.
- Chapter 18: Pipes of *R for Data Science* for more information on pipes.
- Syntax equivalents: base R vs Tidyverse by Hugo Taveres for a comparison of base-R solutions to tidyverse. This entire book is largely biased towards tidyverse solutions, but there's no denying that certain base-R can be more elegant. Check out this write up to get a better idea.

Chapter 10

Programming with R

Programming is writing instructions that tell the computer what to do. Like most things, learning a little goes a long way. And like most things, it's easy to lose the forest for the trees. That's why we won't focus too much on programming (after all you're chemist not computer scientist) but we will introduce a few simple yet incredibly powerful elements of programming to help you along with your data science quest.

We'll point to several sources for further reading on functions at the end of this chapter.

10.1 Functions

Functions allow you to write general purpose code to automate common tasks. They're a great way to decrease repetition and make your code more legible and reproducible. To crease a function in R you only need `function()`:

```
funSum <- function(x,y){
  z <- x + y
  paste("The sum of", x, "+", y, "is", z, sep =" ")
}

funSum(1, 3)
```

```
## [1] "The sum of 1 + 3 is 4"
```

```
funSum("yes",3)
```

```
## Error in x + y: non-numeric argument to binary operator
```

What we've done is create a function called `funSum` which takes two numeric inputs `x` and `y`, sums the two into `z` and paste an output telling us the sum. A couple of things to note:

- We need to *explicitly* state which arguments are function will take; in this example they are `x` and `y`. Whatever we pass to `x` or `y` will be carried into the function.
- We can't sum non-numeric values, so R returns an error in the second instance
- Functions create their own environment, therefore *any variable* created inside a function only exists inside the function.
  - In the above example, `x`, `y`, and `z` only exist inside the function.
- R automatically returns whichever variable is on the last line of the body of the function; but you can explicitly ask for an output using `return()`

Let's take a look at a more practical function, something that you might actually use. In mass spectrometry, a gauge of accuracy is the *mass error*, a measure of the difference between the observed and theoretical masses, and is reported in parts-per-million (ppm). The formula for calculating mass error is:

$$Mass\ error\ (ppm) = \frac{|mass_{theoretical} - mass_{experimental}|}{mass_{theorical}} \times 10^6$$

The formula is simple enough, but you may need to calculate any number of mass errors, so it behooves us to compose a quick formula to simplify our workload:

```
ppmMS <- function(theoMZ, expMZ){

  ppm <- abs(theoMZ - expMZ)/theoMZ * 1e6
  ppm
}

# Theoretical mass = 1479.63 m/z
# experimental mass = 1480.10 m/z
ppmMS(theoMZ = 1479.63, expMZ = 1480.10)
```

```
## [1] 317.647
```

Pretty useful if you're manually checking something, but we can also use our functions into the pipe to help our data transformation:

```
# Example data
masses <- data.frame("theo" = c(1479.63, 1479.63, 1479.63),
                     "exp" = c(1478.63, 1479.63, 1480.10))

masses %>%
  mutate(massError = ppmMS(theo, exp))
```

```
##      theo     exp massError
## 1 1479.63 1478.63  675.8446
## 2 1479.63 1479.63    0.0000
## 3 1479.63 1480.10  317.6470
```

This last part is critical as *functions make your code more legible*. We can clearly read that the code above is calculating the mass error between the theoretical and experimentally observed masses. This might not be as apparent if we put in a complex mathematical formula in the middle of our pipe.

10.2 Conditional arguments

Are used to specify a path in a function depending on whether a statement is `TRUE` or `FALSE`. These are explored in greater detail via the links in the Further Reading section, but here's a quick example of a function that uses the conditional `if` statement to print out which number is largest:

```
isGreater <- function(x, y){
  if(x > y){
    return(paste(x, "is greater than", y, sep = " "))
  } else if (x < y){
    return(paste(x, "is less than", y, sep = " "))
  }
  return(paste(x, "is equal to", y, sep = " "))
}

isGreater (2, 1)
```

```
## [1] "2 is greater than 1"
```

```
isGreater (1, 2)
```

```
## [1] "1 is less than 2"
```

```
isGreater (1, 1)
```

```
## [1] "1 is equal to 1"
```

Our simple function compares two numbers, `x` and `y` and if `x > y` evaluate to `TRUE` it returns the pasted string `x is greater than y`. If `x < y` evaluates to `FALSE`, as in `y > x`, our function returns the pasted string `x is less than y`, and finally if neither `x > y` and `x < y` evaluate to TRUE, they must be equal! Therefore the final output is `x is equal to y`. This is an example of an `else if` statement. If you're simply evaluating two conditions (`TRUE` or `FALSE`) you only need the `if()` conditional, see Further Reading for more details.

10.2.1 Piping conditional statements

You can already see the potential for simply conditional statements in the pipe. However, to keep piping operations legible, `dplyr` offers the `case_when` function, which works similarly to the `else if` statements showcased above. Let's see how it works using a real world example.

In mass spectrometry undetected compounds are recorded having an intensity of 0; it's a common practice to replace 0 with $\frac{limit\ of\ detection}{2}$ for subsequent analysis However, we don't want to replace every value with $\frac{LOD}{2}$, only 0s. Let's use the `case_when()` function to create a new values with the recorded intensities

```
lod <- 4000 # previously calculated LOD
results <- data.frame("mz" = c(308.97, 380.81, 410.11, 445.34 ), # dummy data
                      "intensities" = c(0, 1000, 5000, 10000))

results %>%
  mutate(reportedIntensities = case_when(intensities < lod ~ lod/2,
                                         TRUE ~ intensities))
```

```
##       mz intensities reportedIntensities
## 1 308.97           0                2000
## 2 380.81        1000                2000
## 3 410.11        5000                5000
## 4 445.34       10000               10000
```

Firstly we're creating a new column called `reportedIntesities` using `mutate()` and using `case_when()` to conditionally fill that column. The inputs we've passed to `case_when()` are two-sided formulas. Essentially if the conditions on the left-hand side of the tilda (`~`) evaluate to `TRUE`, `case_when` will execute the

right-hand side. Thee first two-sided formula is `intensities < lod ~ lod/2` and checks if the intensities value is less than the previously calculated limit of detection. If `intensitis < lod` evaluates to `TRUE` we insert half of the LOD value for that row. If `intentisites < lod` evaluates to `FALSE`, we move onto the next two-side formula and reevaluate again. The second two-sided formula `TRUE ~ intensities` basically means for everything that's remaining (greater than LOD in our instance) just use the value from the `intensities` column.

Some ideas to consider when working with `case-when()`:

- There's no limits to the conditions you can pass to `case_when()`.
- *However* `case_when()` evaluates in order so put the more specific conditions before the more general.
- Remember that the point of `case_when()` and piping is legibility. If you're passing multiple conditions, consider writing a function using `else if` statements to keep the pipe legible.

10.3 When to use functions

A good rule when coding is **Don't Repeat yourself!**. In practice, this means don't copy and paste blocks of code to multiple parts of your script. It's more difficult to read (more lines), and if you identify an issue with one block, you'll need to hunt down all the other blocks to rectify the situation (you'll always miss something!). by using functions you'll reduce the number of lines of code, but you'll also only need to check one spot to rectify the issues.

10.4 Further Reading

These chapter has been intentional succinct. We've omitted several other aspects of programming in R such as `for` loops, and other iterative programming. To get a better sense of programming in R and to learn more, please see the following links:

- case_when(): the documentation for the `case_when()` function and several useful examples.
- Chapter 19: Functions, Chapter 20: Vectors, and Chapter 21: Iteration of *R for Data Science* by H. Wickham and G. Grolemund.

- Hands-on Programming in R by G. Grolemund for a more in-depth (but still approachable) take on programming in R.

Chapter 11

Modelling

Modelling is basically math used to describe some type of system, and they are a forte of R, a language tailor made for statistical computing… Every model has assumptions, limitations, and all around tricky bits to working. We'll discuss model fitting and break down popular models you'll encounter in specific chapters in Section 3. For now, we'll introduce the `lm()` function for generalized linear models.

Linear models are the *trend lines* you used all the way back in *CHM135*. However, if you've been exposed to these, it's most likely via *Excel*'s 'add trend line' option. While `lm()` works much the same *mathematically*, unlike *Excel*, R returns *allllll* of the model outputs. Correspondingly, it's easy to get lost between juggling R code, the endless model outputs, and keeping yourself grounded in the real science you're attempting to model.

So let's take our `lm()` function at face value and learn *how* to model in R. Again we'll touch up the details later on, but for now let's import the FAES calibration results we saw in Transform: dplyr and data manipulation. As we've already seen, our data is composed of four standards and a blank analyzed in triplicate. Since we're focusing on modelling, *we'll treat the blank as a standard in our model fitting*:

```
# Importing using tips from Import chapter

FAES <- read_csv(file = "data/FAESdata.csv") %>%
  pivot_longer(cols = -std_Na_conc,
               names_to = "replicate",
               names_prefix = "reading_",
               values_to = "signal") %>%
  separate(col = std_Na_conc,
           into = c("type", "conc_Na", "units"),
```

```
            sep = " ",
            convert = TRUE) %>%
  mutate(type = "standard")

DT::datatable(FAES)
```

Show 10 entries Search:

	type	conc_Na	units	replicate	signal
1	standard	0	mg/L	1	1348.5051
2	standard	0	mg/L	2	1304.0702
3	standard	0	mg/L	3	1395.7714
4	standard	0.1	mg/L	1	2947.3097
5	standard	0.1	mg/L	2	2924.3988
6	standard	0.1	mg/L	3	2927.2867
7	standard	0.2	mg/L	1	4446.4036
8	standard	0.2	mg/L	2	4453.1066
9	standard	0.2	mg/L	3	4416.439
10	standard	0.3	mg/L	1	6235.1603

Showing 1 to 10 of 15 entries Previous 1 2 Next

Note model is a general term, in this situation we'll be calculating a **calibration curve**. All calibration curves are models, but not all models are calibration curves.

11.1 Base R Linear Model

R's base `lm()` function for linear regression is excellent, but it's outputs have some messy quirks. It's easier to show that, so let's calculate the linear relationship between the `signal` as a function of `conc_Na`:

```
lm_fit <- lm(signal ~ conc_Na, data = FAES)
lm_fit
```

```
##
## Call:
## lm(formula = signal ~ conc_Na, data = FAES)
##
## Coefficients:
## (Intercept)      conc_Na
##        1243        16885
```

Reading the code above (recall that we're reading it from *right to left* because it's base R):

1. We're taking the FAES data
2. We're comparing `signal` (the dependent, y, variable) to `conc_Na` (the independent, x, variable) via the tilde `~`. The way to read this is: *"Signal depends on concentration"*.
3. We're comparing these two variables using the `lm()` function for generalized linear models.
4. All of the model outputs are stored in the `lm_fit` variable.

As we can see, the model outputs are pretty brief and not more more than *Excel*'s outputs. We can use `summary()` to extract more information to better understand our model:

```
summary(lm_fit)
```

```
##
## Call:
## lm(formula = signal ~ conc_Na, data = FAES)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -203.091  -86.731   -3.761  107.837  176.562
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1242.57      58.05   21.41 1.61e-11 ***
## conc_Na     16884.82     236.98   71.25  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 129.8 on 13 degrees of freedom
## Multiple R-squared:  0.9974, Adjusted R-squared:  0.9972
## F-statistic:  5077 on 1 and 13 DF,  p-value: < 2.2e-16
```

Now we have a lot more information from our model (don't worry about what everything means, it's discussed further in Section 3. For now, understand that it's a hot mess.

11.2 Cleaning up model ouputs

`summary()` provides a decent overview of our model's performance, but the outputs are difficult to work with. Let's turn to the `broom()` package to clean up our model outputs.

```
library(broom)

calCurve <- FAES %>%
  group_by(type) %>%
  nest() %>%
  mutate(fit = map(data, ~lm(signal ~ conc_Na, data = .x)),
         tidied = map(fit, tidy),
         glanced = map(fit, glance)
         )
calCurve
```

```
## # A tibble: 1 x 5
## # Groups:   type [1]
##   type     data              fit    tidied           glanced
##   <chr>    <list>            <list> <list>           <list>
## 1 standard <tibble [15 x 4]> <lm>   <tibble [2 x 5]> <tibble [1 x 12]>
```

Things look a bit more complicated than our earlier example, so let's break down our code line by line:

1. We're taking the `FAES` dataset that we created earlier.
2. `group_by(type` groups all rows by `type`, in this situation we have only one type: `standard`.
3. `nest()` collapses everything other than the `type` column into smaller dataframes. In this situation, all other information is stored as a `tibble` under the `data` column; this is the data used to calculate the linear model.
4. Withing the `mutate` function, we've created three columns: `fit`, `tidied` and `glanced`.

And it's the the `fit`, `tidied` and `glanced` that contains out cleaned up model outputs. `fit` contains the raw output from the linear regression model for `signal` as a function of `conc_Na` using the `lm()` function. The output is in the form of a list, similar to what `summary()` gave us above. Again, this is exceptionally messy, hence why we used the `tidy()`, and `glance()` function from the `broom` package . `map()` just means we're applying the function `tidy()` to the individual output list created by `lm()` and stored in the `fit` column. Note that the `tidy()` and `glanced()` outputs are tibbles. So we now have a tibble containing specific model output values (i.e. `(Intercept)`), lists (i.e. `fit`), and tibbles (`tidied`). This is known as **nested data*. We're no longer in Kansas anymore…

Anyways, let's take a look at our model results. The `glanced` tibble contains "…a concise one-row summary of the model. This typically contains values such as R^2, adjusted R^2, and residual standard error that are computed once for

the entire mode"[1] Because the data is nested, we'll need to use `unnest()` to flatten it back out into regular columns:

```
calCurve %>%
  unnest(glanced)
```

```
## # A tibble: 1 x 16
## # Groups:   type [1]
##   type   data     fit   tidied  r.squared adj.r.squared sigma statistic  p.value
##   <chr>  <list>   <lis> <list>      <dbl>         <dbl> <dbl>     <dbl>    <dbl>
## 1 stand~ <tibble~ <lm>  <tibbl~     0.997         0.997  130.     5077. 3.05e-18
## # ... with 7 more variables: df <dbl>, logLik <dbl>, AIC <dbl>, BIC <dbl>,
## #   deviance <dbl>, df.residual <int>, nobs <int>
```

What you see here is a bit more than what you'd get from *Excel*'s 'line-of-best fit' output. See the section on *Modelling* for a better breakdown of what everything means. But for now, we can see that our `r.squared` of each calibration curve is pretty good, and the `p.value` indicates each model is significant. the `adj.r.squared` is the same as `r.squared` in this situation. This is because `r.squared` will always increase if we add more exploratory variables to our model; the `adj.r.squared` accounts for the number of exploratory variables used in the model. However, in our case we only have one exploratory variable, hence they're the same.

But what about the slope and the intercept? After all, that's what we need to calculate the concentration in our unknowns. Let's take a look at `tidied` from the `tidy()` function "…which constructs a tibble that summarizes the model's statistical findings. This includes coefficients and p-values for each term in a regression…"[2]

```
# storing because we'll use it later on.

tidied <- calCurve %>%
  unnest(tidied)

tidied
```

```
## # A tibble: 2 x 9
## # Groups:   type [1]
##   type   data      fit    term   estimate std.error statistic  p.value glanced
##   <chr>  <list>    <list> <chr>     <dbl>     <dbl>     <dbl>    <dbl> <list>
## 1 stand~ <tibble ~ <lm>   (Inte~    1243.      58.0      21.4 1.61e-11 <tibble ~
## 2 stand~ <tibble ~ <lm>   conc_~   16885.     237.       71.3 3.05e-18 <tibble ~
```

[1] From the *broom* package vignette.
[2] From the *broom* package vignette.

Again, a lot more to unpack compared to *Excel*. That's because the `lm()` function in R calculates a generalized linear model. `lm()` performs a linear regression model, which we normally think of as an equation of the form $y = mx+b$. But, regression models can be expanded to account for multiple variables (hence *multiple linear regression*) of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... \beta_p x_p$$

]

where,

- y = dependent variable
- x = exploratory variable; there's no limit how many you can input
- B_0 = y-intercept (constant term)
- B_p = slope coefficient for each explanatory variable

In our situation, we only have one input variable for our model (`conc`), so the above formula collapses down to $y = \beta_0 + \beta_1 x_1$. So looking at our results above, each row corresponds to a model parameter. For each modelling parameter, we're provided an estimate of it's numerical value (`estimate`, the values we'll use to calculate concentration). The other parameters are useful to understand but not necessary at this point (again, check out the *Modelling* section).

And we can extract these values to use in subsequent calculations:

```
intercept <- as.numeric(tidied[1,5])
slope <- as.numeric(tidied[2,5])

paste("The equation of our calibration curve is: y = ", slope, "x + ", intercept, sep="
```

```
## [1] "The equation of our calibration curve is: y = 16884.8167x + 1242.56646666666"
```

11.3 Further reading

The theory and use of these models are explored in greater details in Section 3. Please read up on it for an understanding of the model outputs and how to use them in your analysis. As well, see the section on modelling in *R for Data Science*.

https://www.newyorker.com/magazine/2021/06/21/when-graphs-are-a-matter-of-life-and-death

Chapter 12

Visualizations

- theory undergirding ggplot (focus on geom_point)
  - Geoms
  - aesL x,y, ccolour, shape, size, alpha
  - arranging plots in a grid (grid.arrange) and facets…
- How to plot
  - labels
  - scales
  - annotations
  - themes
- building a plot layer by layer
- saving/exporting plots.

Visualizations have always been an important part of data science and chemsitry. Good graphics illuminate trends and pattersn you may have otherwise missed and allow us to quickly inspect thousands of values. R via the `ggplot2` package is one of, if not the premier, data visualization langauge available. This chapter will formally introduce the `ggplot2` package, explain a bit of the logic undergirding it's operation, and give you some quick examples of how it works. Section 3 will delve deeper into specific visulizations you'll use and encounter in your studies.

`ggplot2` is loaded by default with the `tidyverse` suite of pakacges. Let's revisit our spectroscopy data we encountered in Tidying your data:

```
library(tidyverse)
atr_long <- read_csv("data/ATR_plastics.csv") %>%
    pivot_longer(cols = -wavenumber,
                 names_to = "sample",
                 values_to = "absorbance")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   wavenumber = col_double(),
##   EPDM = col_double(),
##   Polystyrene = col_double(),
##   Polyethylene = col_double(),
##   `Sample: Shopping bag` = col_double()
## )
```

```
# This just outputs a table you can explore within your browser
DT::datatable(atr_long)
```

```
## Warning in instance$preRenderHook(instance): It seems your data is too big
## for client-side DataTables. You may consider server-side processing: https://
## rstudio.github.io/DT/server.html
```

Show 10 entries Search:

	wavenumber	sample	absorbance
1	550.0952	EPDM	0.2119556
2	550.0952	Polystyrene	0.07463058
3	550.0952	Polyethylene	0.000873196
4	550.0952	Sample: Shopping bag	0.02364882
5	550.5773	EPDM	0.2124079
6	550.5773	Polystyrene	0.07455246
7	550.5773	Polyethylene	0.000834192
8	550.5773	Sample: Shopping bag	0.02382648
9	551.0594	EPDM	0.2128818
10	551.0594	Polystyrene	0.07450471

Showing 1 to 10 of 28,628 entries Previous 1 2 3 4 5 … 2863 Next

12.1 Building plots ups

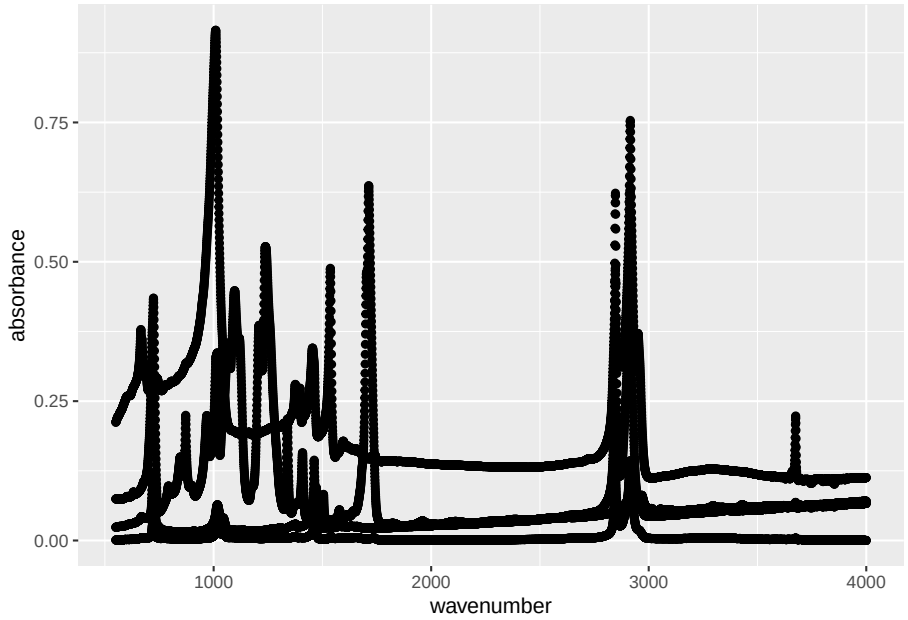
The gg in ggplot2 stands for the grammar of graphics (Wickham 2010), and it's a way to break down graphcis (plots) into small pieces that can be discussed (hence grammar). WE'll take a look at this grammar via geoms (what kind of plot), aes (aesthetic choices), etc. For now, understand that this means we need to build up graphics/plots piece-by-piece and layer-by-layer. This extends beyond code to how we code. No sense in putting lipstick on a pig. PLot often, and discard the useless ones. Take the time to pretty up your plot *after* you're satisfied with the unerlying data.

12.2 Basic plotting

ggplot2 uses geoms to specificyt what type of plot to create. Different plots are used to convey different meanings and have different strenghts and weakness. We'll explore these more in Section 3, but for now we'll focus on geom_point(), which simply plots data as poits on an [x,y] coordinate. In otherwords, a scatter plot.

Let's plot our tided atr_long data:

```
ggplot(data = atr_long,
       aes(x = wavenumber, y = absorbance)) +
  geom_point()
```



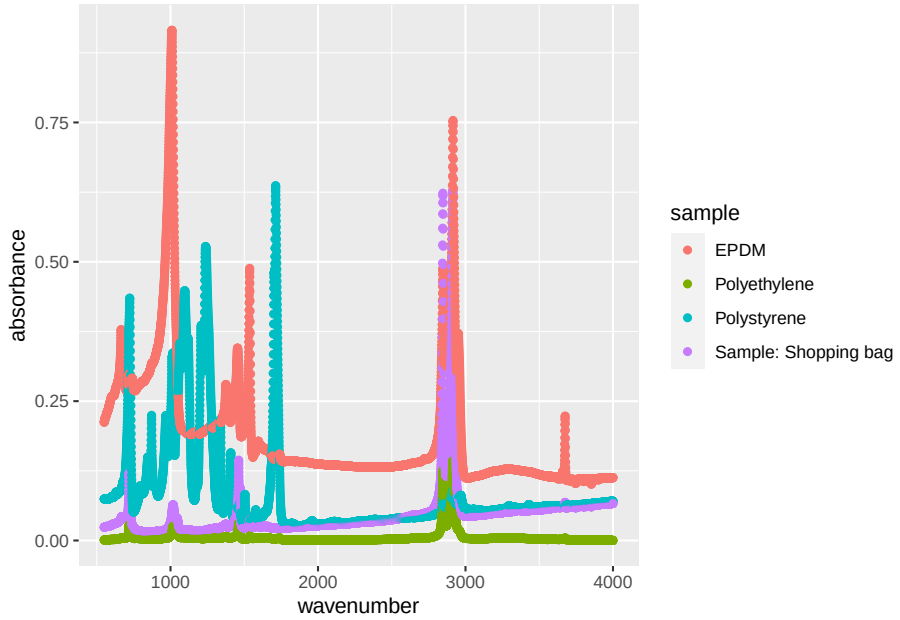
Let's ignore the plot for now and look at our code down:

1. `ggplot()` intilaizes a *ggplot object*, basically an empty plot. To this we've specified out data set (`data = atr_long`).
2. We then specified our *aesthetic mappings* via `aes()`. Here we'll pass information for how we want the plot to look. 3.To our aesthetic mappings we've specified which values from `atr_long` are supposed to be our x-axis values (`x = wavenumber`) and y-axis values (`y = absorbance`).
3. We then add the `geom_point()` layer to create a scatter plot of [x,y] points .

Now let's look at our result. What we see is a point for every recorded absorbance measurements from our ATR analysis. We can clearly see the spectra of the different plastics in our data, however they're all colours the same. This is because we've only speciies the x and y values. As far as `ggplot()` is concerned, these are the only values that mattter, but we know different.

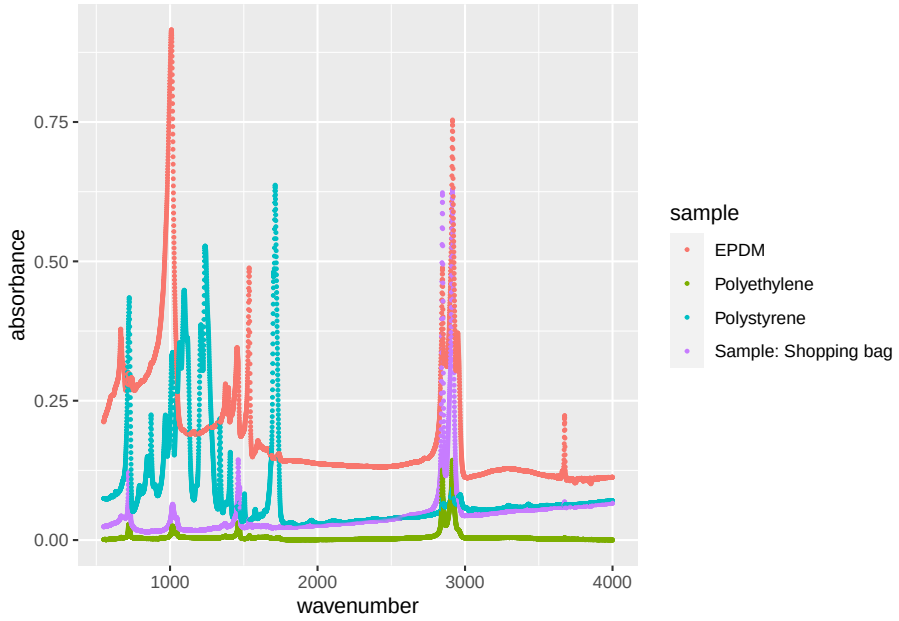
Fortuanely you can pass multiple variables to different `aes()` options to enahce our plot. For instance, we can pass the `sample` variable, which specifies which sample a spectrum originates from, to the `colour` option:

```
ggplot(data = atr_long,
       aes(x = wavenumber,
           y = absorbance,
           colour = sample)) +
  geom_point()
```



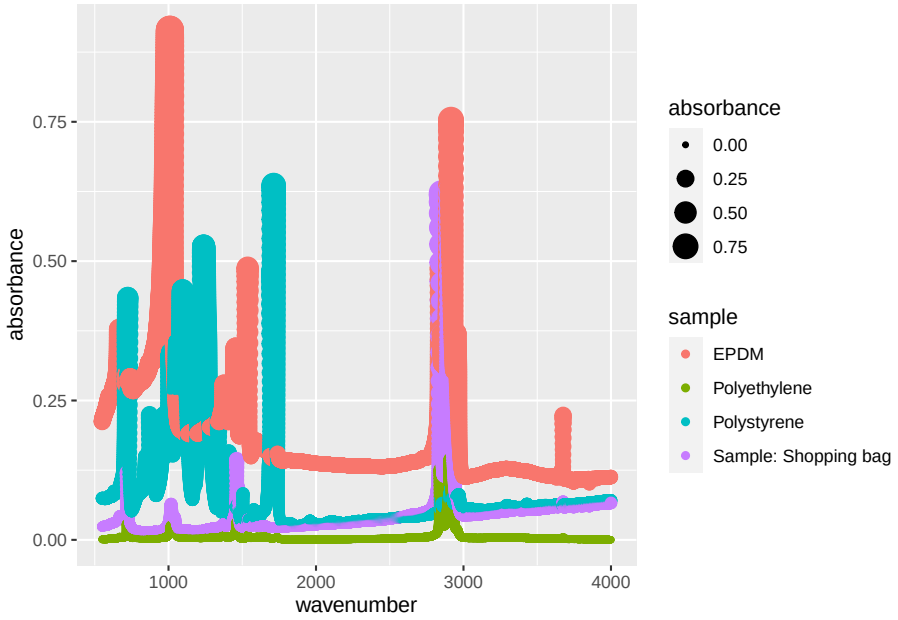
Now we have a legend which clearly specifies which points are assosciated with each sample. But now the points are too large, potentially masking certain peaks. We can adjust the size of each point as follows:

```
ggplot(data = atr_long,
       aes(x = wavenumber,
           y = absorbance,
           colour = sample)) +
  geom_point(size = 0.5)
```



We specified `size = 0.5` in the `geom_point()` call because it's a constant. We can map `size` to any continuous variable, such as the absorbance:

```
ggplot(data = atr_long,
       aes(x = wavenumber,
           y = absorbance,
           colour = sample,
           size = absorbance)) +
  geom_point()
```

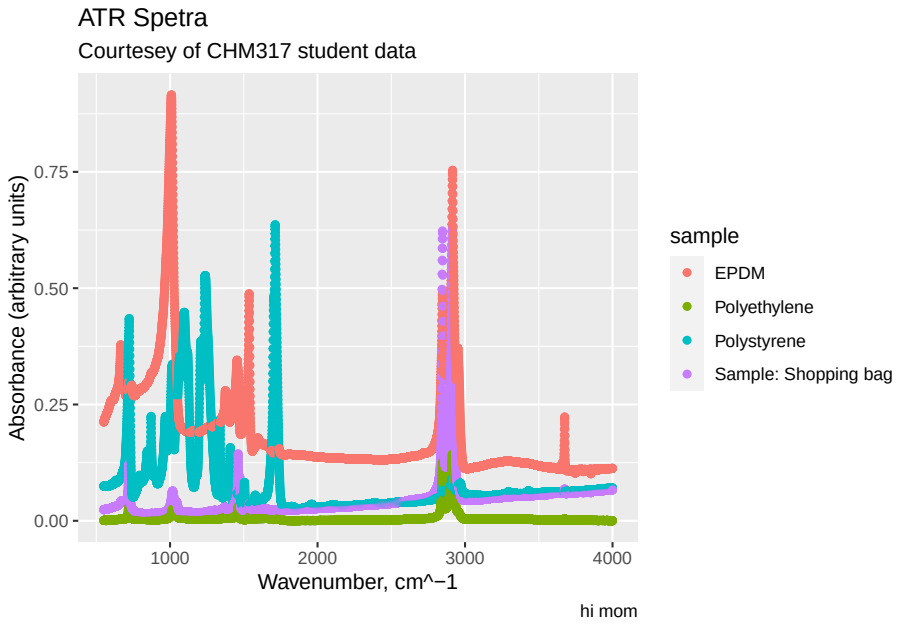


Sometimes this makes sense (i.e. a *bubble chart*) but for our example, having the size of the poitns increase as the absorbance increases doesn't provide any new information (it actually clutters our plot).

12.2.1 Changing plot labels

By default `ggplot` uses the header of the columns you passed for the `x` and `y` `aes()` options. Because headers are written for code they're often poor label titles for plots. We can specify new labels and plot titles as follows:

```
ggplot(data = atr_long,
       aes(x = wavenumber,
           y = absorbance,
           colour = sample)) +
  geom_point() +
  labs(title = "ATR Spetra",
       subtitle = "Courtesey of CHM317 student data",
       x = "Wavenumber, cm^-1",
       y = "Absorbance (arbitrary units)",
       caption = "hi mom")
```



12.3 Further reading

There's no shortage of options when playing around with `ggplot` and these will be explored in greater detail in Section 3 (including *when* you should and shouldn't do things).

Chapter 13

Communication

- R markdown
- slides
- exporting
- tips on automating Rmd generation?

Section 3: Data Analysis Toolbox

Chapter 14

Introduction

Toolbox for data analysis that will envetually cover stuff like:

- different visualizations used in CHM410/env chem w/ links to more useful websites
- statistics, but mostly just those needed for undergrad envb. chem.
- better elaboration on linear regression modelling with descriptions with all of the fit parameters.
- non-linear regression
- interactive plots (probably in the visualizations chapter above)
- how to lie w/ statistics and plots. Chapter explaining graphicacy and numeracy.

Chapter 15

Choosing Visualizations

- introduce the principle plots/visualizations used in CHM410/undergad envb chem
- scatter/jitter
- line
- box/violin plot
- histogram and density distributions
- marginal plots
- working code of example plots

Chapter 16

Linear Regression Redux

Note this needs to be cut up and reformated to elaborate more on the theory behind linear regression. - DH

Note I didn't actually take the CHM410 course, so Jess will need to review this part. As well, i expect this entire chapter will get chopped up in later drafts of the book. As well, will review modelling stuff to find an easier pay. Probably with purrr and what not. -DH

Sample prep is only half the fun when it comes to environmental chemistry. Eventually you'll want to quantify what's in your samples and to do that you'll need to construct calibration curves. This write up will use previously acquired triple-quadrupole LC-MS (hence *QQQ*) results from the 2019 CHM410 Field trip. This fieldtrip data is comprised of three datasets: `lab4_biota.csv` with the sampling information for biological samples, `liab4_sediment.csv` for sediment sampling information, and `lab4_qqq.csv` with the integrated peak areas of every analyzed sample and calibration standard. For this section you'll only need `lab4_qqq.csv` as we'll only be calculating the concentration of the samples we injected in the instrument, and not back calculating the concentration in our original samples.

Note Most of what we'll need is contained in the `tidyverse` family of packages, but you will also need the `broom` and `ggrepel` packages to make your lives easier.

16.1 Importing and tidying data

First off let's import the peak area information for all of our analytes. *Note* I cleaned up the data in Excel resulting in the `lab4_qqq.csv` file used herein. This is mostly because of issues with merged cells and multiple sheets in the

original dataset. I also took the opportunity to generated unique `sampleID` values for each sample by combining the sample name and the group letter.

```
library(tidyverse) # for dplyr, readr, stringr, and ggplot

QQQ <- read_csv("data/CHM410/lab4_qqq.csv")

head(QQQ)
```

```
## # A tibble: 6 x 29
##   group  type  sampleID  PFHxA_peakArea PFHxA_RT `13C2PFHxA_peak~ `13C2PFHxA_RT`
##   <chr>  <chr> <chr>              <dbl>    <dbl>            <dbl>           <dbl>
## 1 calCu~ std   20MC 0 p~              0     1.97           160000            3.65
## 2 calCu~ std   20MC 0.2~          28400     3.62           174000            3.62
## 3 calCu~ std   20MC 0.4~          45100     3.62           169000            3.61
## 4 calCu~ std   20MC 1 p~          90900     3.61           168000            3.61
## 5 calCu~ std   20MC 2 p~         184000     3.6            174000            3.6
## 6 calCu~ std   20MC 4 p~         384000     3.61           141000            3.6
## # ... with 22 more variables: PFHpA_peakArea <dbl>, PFHpA_RT <dbl>,
## #   PFOA_peakArea <dbl>, PFOA_RT <dbl>, 13C4PFOA_peakArea <dbl>,
## #   13C4PFOA_RT <dbl>, PFNA_peakArea <dbl>, PFNA_RT <dbl>,
## #   13C5PFNA_peakArea <dbl>, 13C5PFNA_RT <dbl>, PFDA_peakArea <dbl>,
## #   PFDA_RT <dbl>, 13C2PFDA_peakArea <dbl>, 13C2PFDA_RT <dbl>,
## #   PFHxS_peakArea <dbl>, PFHxS_RT <dbl>, 13C4PFHxS_peakArea <dbl>,
## #   13C4PFHxS_RT <dbl>, PFOS_peakArea <dbl>, PFOS_RT <dbl>,
## #   13C4PFOS_peakArea <dbl>, 13C4PFOS_RT <dbl>
```

Note how our data is in a wide format, with columns for the peak area and retention times for each targeted ion. This served the TA well when they wrote down the data from the LC-MS analysis, but let's tidy it up so it's easier to work with in R.

```
### Find cleaner way of doing this...

QQQ <- QQQ %>%
  pivot_longer(cols = -c("group", "type", "sampleID"),
               names_to = c("cmpd", "measurement"),
               names_sep = "_",
               values_to = "value") %>%
  pivot_wider(names_from = measurement,
              values_from = value)

head(QQQ)
```

```
## # A tibble: 6 x 6
```

```
##   group    type  sampleID   cmpd      peakArea    RT
##   <chr>    <chr> <chr>      <chr>        <dbl> <dbl>
## 1 calCurve std   20MC 0 ppb PFHxA            0  1.97
## 2 calCurve std   20MC 0 ppb 13C2PFHxA   160000  3.65
## 3 calCurve std   20MC 0 ppb PFHpA            0  2.26
## 4 calCurve std   20MC 0 ppb PFOA             0  2.65
## 5 calCurve std   20MC 0 ppb 13C4PFOA    244000  4
## 6 calCurve std   20MC 0 ppb PFNA             0  3.29
```

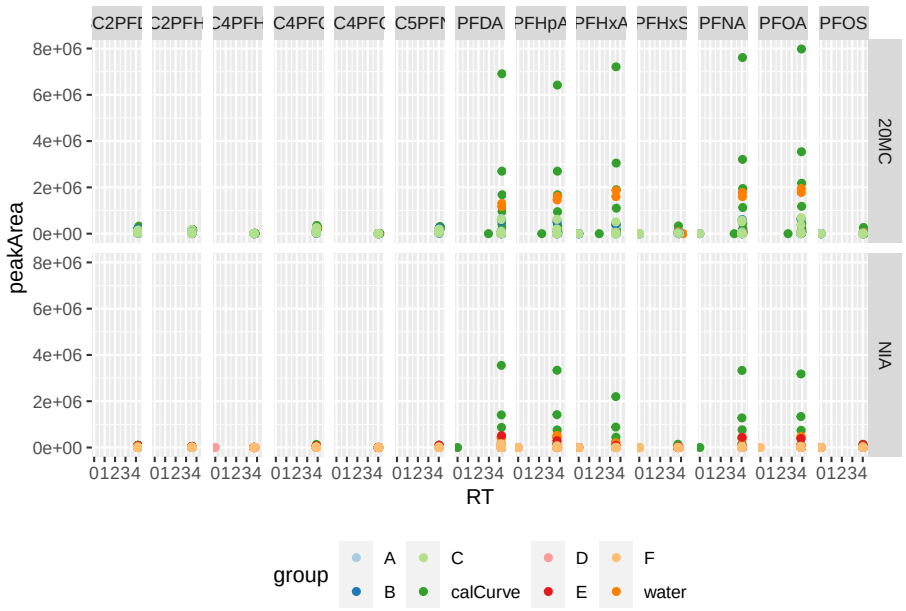
Much better, each column is a variable, and every row an observation. However, since the Lake Niamco samples were analysed at a different time then the 20 Mile Creek samples, let's quickly annotate our data to differentiate the two. Since we used (overly) descriptive sample names (stored in the `sampleID` column), we can create a new column to specify the location by searching for matching string values.

```
QQQ <- QQQ %>%
    mutate(location = case_when(
      str_detect(sampleID, regex("20MC", ignore_case=TRUE)) ~ "20MC",
      str_detect(sampleID, regex("NIA", ignore_case=TRUE)) ~ "NIA",
      TRUE ~ "NA"))
```

The code above will search through every row in the `sampleID` column. If it finds the string of characters `20MC`, which we used to denote samples from 20 Mile Creek, it will record this in a new column called `location`. Same with NIA. If neither `20MC` or `NIA` are detected, it returns `NA`. *Note*, if we had more complex `sampleID` names, we could expand our `case_when` arguments accordingly.

Let's make a quick plot to verify everything is looking alright:

```
library(RColorBrewer) # because i'm colour blind...

ggplot(QQQ, aes(x = RT, y = peakArea, colour = group)) +
  geom_point() +
  facet_grid(cols = vars(cmpd),
             rows = vars(location)) +
  theme(legend.position = "bottom") +
  scale_color_brewer(palette = "Paired")
```



Alright, busy plot, but let's see what we got. First off, this is a *small multiple*, basically a grid of small, individual, plots that share a common axis. So each small plot is our retention time (`RT`) on the x-axis vs. integrated peak area (`peakArea`) on the y-axis. Now our small multiple is organized in a grid, with the columns of the grid corresponding to the ions we analyzed, and the rows of the grid being the location grouping. So, the top-right plot shows the peak area vs. retention time of PFOS from the 20 Mile Creek samples. Lastly, the colour of a point corresponds to the group to which that value belongs. So we see multiple `calCurve` values at the same RT for a given compound, but with vastly different peak areas. This makes sense, as these are our standards from which we'll construct our calibration curve later on.

Now that we understand what we're looking at, let's inspect our data. Here are some things I noted:

- Some compounds have a `RT` and `peakArea` of 0. Now their actual concentration isn't necessarilly 0 ppb, rather the vendor software used to calculate peak areas will return `0` if a given ion wasn't detected. However, R will interpret this number literally, so we'll need to address this later on.
- Most compounds elute at approximately 4 minutes. The grouping of retention times makes sense because of the structural similarity of our targeted compounds, and the short chromtagraphy gradient.

- **However** some compounds appear to elute earlier. Often these outliers have a low `peakArea`, so they may be the result of the vendor's algorithm

integrating noise, and mislabelling it as a legitimate peak. We'll need to review this later on.
- The internal standards all appear to have similar peak areas values. This makes sense, as we've spiked in the same amount of internal standard for each sample.
- No signal is greater than a `calCurve` signal, this is good as it means all of our unknowns should fall within our calibration curve; baring matrix effects...

16.2 Normalizing QQQ Data

Now that we're all organized in terms of importing our LC-MS results, let's being the work of actually quantifying the samples we injected into the LC-MS before back calculating and quantifying our actual field samples.

First thing, we'll need to normalize our peak areas to account for matrix effects. To achieve this, we'll need to pair each analyte of interest with it's assigned internal standard. Recall however that we did not have an exact isotopic standard for each compound. The pairing, from the lab manual, is below:

Analyte	Full name	Carboxylic acid?	Sulfonic acid?	Number of Carbons	Number of perfluorinated carbons	Internal Standard to use
PFHxA	Perfluorohexanoic acid	x		6	5	13C4 PFHxA
PFHpA	Perfluoroheptanoic acid	x		7	6	13C4 PFOA
PFHxS	Perfluorohexane sulfonic acid		x	6	6	13C4 PFHxS
PFOA	Perfluorooctanoic acid	x		8	7	13C4 PFOA
PFNA	Perfluorononanoic acid	x		9	8	13C5 PFNA
PFOS	Perfluorooctane sulfonic acid		x	8	8	13C4 PFOS
PFDA	Perfluorodecanoic acid	x		10	9	13C2 PFDA

So according to the table above, both `PFHpA` and `PFOA` use `13C4PFOA` as an internal standard. Given

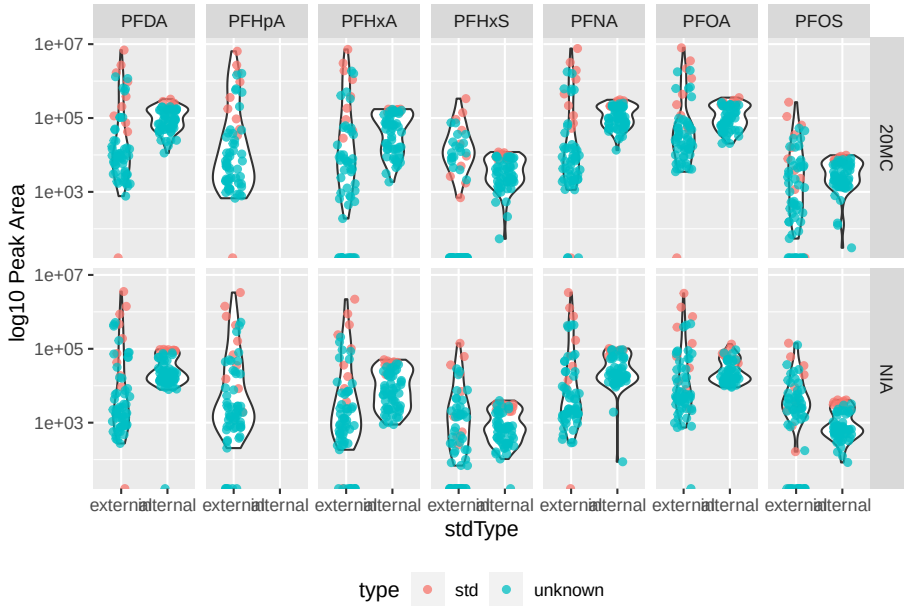
16.2.1 Assessing Internal Stds

For fun, let's gauge how much our internal standards varied between samples. After all, they're all supposed to be the same... We'll create a new column to annotate which compounds are from our internal standard (i.e. those with ^{13}C). For that, we'll recycle a bit of code from above to search for the "13C" string. Then we can compare the peak areas of the internal standards and our analytes of interest.

As for pairing, there's a couple of ways we could do this. Let's just create a new column where we remove the `13C.` string, so we get the same compounds. Some things to note about the string search:

- The regex `13C.`, will look up any string with `13C` and one additional wildcard character (noted by the `.`). This way we can account for the different numbers of 13C in the name, such as `13C4` and `13C2`.
- `PFHpA` will not have an isotopic pair. We'll need to resolve this later by subtracting the `13C4PFOA` values from it.

```
# annotating internal and external standards
QQQ <- QQQ %>%
  mutate(stdType = case_when(
      str_detect(cmpd, regex("13C", ignore_case=TRUE)) ~ "internal",
      TRUE ~ "external"))

# pairing analytes w/ internal standard

QQQ <- QQQ %>%
  mutate(pair = str_remove(cmpd, regex("13C.")))


ggplot(data = QQQ, aes(x = stdType, y = peakArea)) +
  geom_violin() +
  geom_jitter(aes(colour = type),
              position=position_jitter(0.2),
              alpha = 0.75) +
  scale_y_continuous(trans="log10") +
  facet_grid(cols = vars(pair),
             rows = vars(location)) +
  theme(legend.position = "bottom") +
  ylab("log10 Peak Area")
```



Boy howdy let's break this down. This is small multiple of violin plots that shows some neat trends. For those not in the know, a violin plot is similar to a box plot, but the width of the 'bar' is a function of the density of the data around that point. In other words, the more points at a given value, the wider the plot; we've also plotted the individual points themselves to help convey this. Violin plots help us better visualize groupings of data, and can shed some light if a grouping of data might actually be many smaller groupings.

Anyways, let's see what else this says about our data. First, we can see that there's much less variation between internal standards (the "internal" columnˮ) compared to the non-isotopically labelled analytes (the"external" column). Makes sense, all the internal standards are supposed to be the same concentration. However, even then there is still at least an order of magnitude variation in peak area for a given internal standard. Some of this is due to matrix effects, and is why we added the internal standards in the first place. However, looking at the internal standard peak areas, it appears that they cluster into two groups, albeit with some overlap. We might have missed this with a boxplot, but the violin plot helps us see it. You might imagine that the clustering is due to the differences between the external calibration curve samples, and real samples, with the latter having lowered internal standard peak areas due to matrix effects. However, looking at the colouring of the dots, we see that while the `std` internal standard peak areas are generally the highest (no matrix effect), they don't account for all of the upper cluster. So there might be another reason for this. This clustering around two points might be the result of people using two different pipettes to spike in their internal standard. The variation between pipettes could account for the clustering of the internal

standard peak areas. Speaking of two groups, again, we can see a large number of the "external" were not detected, as denoted by their peak area value of `0`. These are probably from the samples originating from the 'clean' reference site. Lastly, between internal standards, we can clearly see that sulfonic acids have a weaker instrumental response then carboxylic acid. This is most likely due to difference in ionizability between the aforementioned functional groups.

16.2.2 Normalizing peak areas

Moving onward, the entire point of spiking the same internal standard is to use those values to normalize our external measurements. For this we'll divide the external peak area by the internal peak area for a given sample. *Note this is how I was told CHM410 did it, holler it it's wrong.*

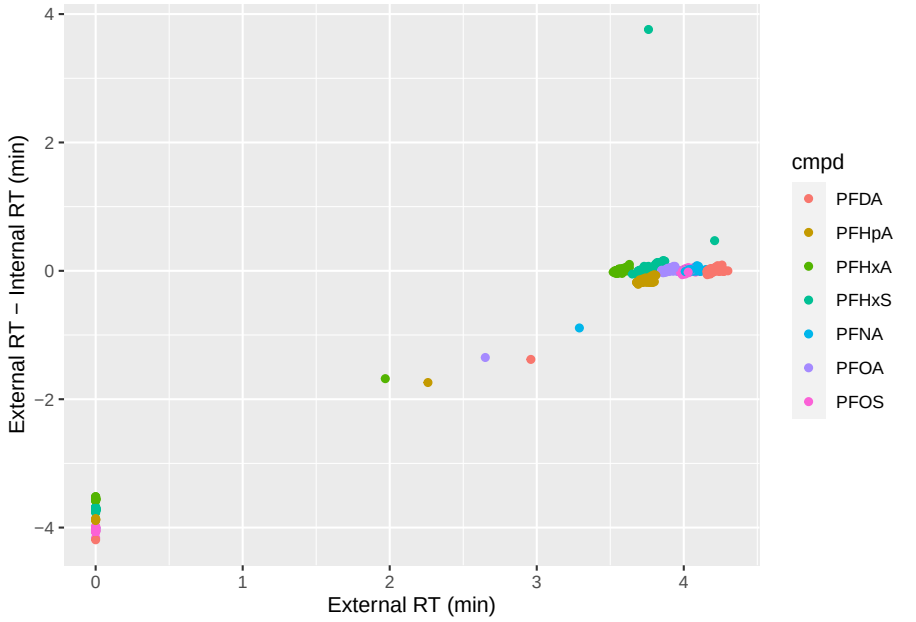
```
# Note removal of PFHpA, because it doesn't have an isotope pair
# Also note RTDif column, which is difference in retention time between internal and ex

QQQNorm <- QQQ %>%
  filter(pair != "PFHpA") %>%
  group_by(sampleID, pair) %>%
  mutate(normPeakArea = peakArea[stdType == 'external'] / peakArea[stdType == 'internal
  mutate(RTdif = RT[stdType == 'external'] - RT[stdType == 'internal'])

# Normalizing PFHpA seperatly; not removal of 13C4PFOA data at end
QQQPFHpA <- QQQ %>%
  filter(cmpd %in% c("PFHpA", "13C4PFOA")) %>%
  group_by(sampleID, ) %>%
  mutate(normPeakArea = peakArea[cmpd == 'PFHpA'] / peakArea[cmpd == '13C4PFOA']) %>%
  mutate(RTdif = RT[cmpd == 'PFHpA'] - RT[cmpd == '13C4PFOA']) %>%
  filter(cmpd %in% c("PFHpA"))

# Rejoining data and dropping internal standard values as they're no longer needed.

QQQNorm <- QQQNorm %>%
  bind_rows(QQQPFHpA) %>%
  filter(stdType == "external")
```

So the above code did double duty. First we normalized the external peak areas by the internal standard peak areas. This has a couple of consequence;

1. If a compounds wasn't detected, it has a peak area of 0. When divided by the internal peak area, the result will be zero.
2. If an internal standard wasn't detected, the external peak area will be divided by 0 resulting in `Inf`, a value of infinity, because in `R`, `1/0 = Inf`

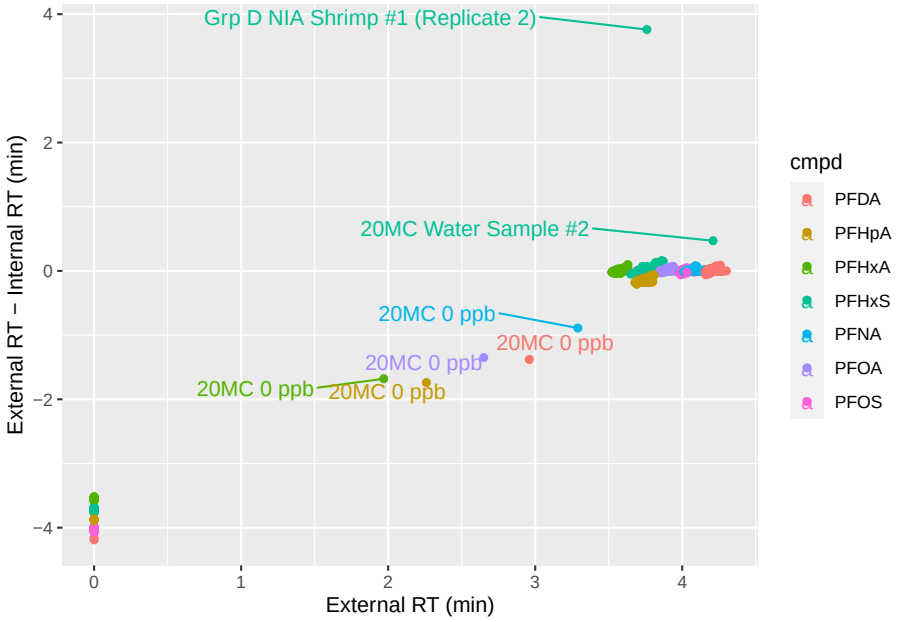
Let's see how our internal and external standards match up by comparing internal and external retention times.

```
ggplot(QQQNorm, aes(x = RT, y = RTdif, colour = cmpd)) +
  geom_point() +
  xlab("External RT (min)") +
  ylab("External RT - Internal RT (min)")
```



Alright, it appears that most of our compounds are clustering around themselves, and there's little variation among the `RTdif` axis, meaning there isn't a large difference between the retention times of the internal and external peaks. This means our peak picking algorithm chose peaks at the correct retention time. Of course, the clustering around `RT = 0` is from compounds that weren't detected, but whose internal standards were; this is fine. However, there appears to be some outliers somewhere between these two clusters. These are probably all from the same sample. Let's annotate our plot to see if this is true.

```
ggplot(QQQNorm, aes(x = RT, y = RTdif, colour = cmpd, label = sampleID)) +
  geom_point() +
    xlab("External RT (min)") +
  ylab("External RT - Internal RT (min)") +
  ggrepel::geom_text_repel(aes(label=ifelse(RTdif > 0.25 | (RTdif < -0.25 & RT > 0), as.character
```



So something strange happened during the acquisition/processing of the `20MC 0 ppb` run. Maybe it was the first one of the day, and the instrument was exceptionally noisy, leading to sloppy peak picking by the algorithm. After all, there shouldn't be any signal as the concentration of the external standard for this sample should be 0 ppb. Likewise, we don't see this effect with the NIA calibration curve standards. Something to keep in mind as we move forward.

16.3 Calculating Calibration Curves

All this and we're only here? Yup, it's important to play around with your data because you never know what you'll find. Already we've talked about issues with our internal spiking, and one of our external standard solutions. Let's move onwards to calculating our calibration curves.

First we need to get the actual concentrations from our standards to make our cal curves. You can make a data frame and match up your concentrations using `inner_join`, or you can simply extract the concentration value from the `calCurve` group `sampleIDs`, as the numerical value is located between two spaces. Just remember to convert using `as.numeric` so R knows to treat the extracted strings as numerical values and not as characters (i.e. `0.1` and not `"0.1"`).

```
QQQstds <- QQQNorm %>%
  filter(group %in% c("calCurve")) %>%
```

```
  mutate(conc = as.numeric(str_extract(sampleID, "(?<=\\s)(.*)(?=\\s)")))
```

Great, now we can simply group our standards by location and compound to compute a linear regression model for each. `R` has a plethora of built-in modelling functions, but oftentimes the output is less then intuitive. Since we'll be using the base `R` `lm` model to calculate our calibration curves, let's import the `broom` package, which is useful for cleaning up `R`'s modelling outputs (hence broom…).

```
library(broom)

calCurves <- QQQstds %>%
  group_by(location, cmpd) %>%
  nest() %>%
  mutate(fit = map(data, ~lm(normPeakArea ~ conc, data = .x)),
         tidied = map(fit, tidy),
         glanced = map(fit, glance)
         )
```

Breaking down the above code, we grouped all calibration standards by the compound and location. This way we can get a linear regression for each grouping. Now, withing the `mutate` function, we've created three columns: `fit`, `tidied` and `glanced`. The first contains the raw output from the linear regression model `lm` in the form of a list. The linear model are calculated for `normPeakArea` as a function of `conc`. This is exceptionally messy, hence why we used the `tidy`, and `glance` function from the `broom` package . `map` just means we're applying the function `tidy` to the individual output list created by `lm` and stored in the `fit` column. Note that the `tidy` and `glanced` outputs are tibbles. So we now have a tibble containing values (i.e. `location`), lists (i.e. `fit`), and tibbles (`tidied`). This is known as **nested data*. We're no longer in Kansas anymore…

Anyways, let's take a look at our model results. The `glanced` tibble contains "…a concise one-row summary of the model. This typically contains values such as R^2, adjusted R^2, and residual standard error that are computed once for the entire mode"[1]

```
calCurves %>%
  unnest(glanced)
```

```
## # A tibble: 14 x 17
## # Groups:   cmpd, location [14]
##    cmpd  location data     fit   tidied  r.squared adj.r.squared sigma statistic
##    <chr> <chr>    <list>   <lis> <list>      <dbl>         <dbl> <dbl>     <dbl>
##  1 PFHxA 20MC     <tibble~ <lm>  <tibbl~     0.996         0.995 1.05      1803.
```

[1] From the *broom* package vignette.

```
##  2 PFOA  20MC     <tibble~ <lm>  <tibbl~     0.944         0.937 1.89       134.
##  3 PFNA  20MC     <tibble~ <lm>  <tibbl~     0.998         0.998 0.556     4994.
##  4 PFDA  20MC     <tibble~ <lm>  <tibbl~     0.994         0.993 0.946     1305.
##  5 PFHxS 20MC     <tibble~ <lm>  <tibbl~     0.999         0.999 0.447     9476.
##  6 PFOS  20MC     <tibble~ <lm>  <tibbl~     0.999         0.999 0.376    11307.
##  7 PFHxA NIA      <tibble~ <lm>  <tibbl~     0.997         0.997 0.788     2742.
##  8 PFOA  NIA      <tibble~ <lm>  <tibbl~     0.946         0.939 1.94       140.
##  9 PFNA  NIA      <tibble~ <lm>  <tibbl~     0.996         0.995 0.843     1953.
## 10 PFDA  NIA      <tibble~ <lm>  <tibbl~     0.996         0.996 0.762     2085.
## 11 PFHxS NIA      <tibble~ <lm>  <tibbl~     0.999         0.999 0.719     7035.
## 12 PFOS  NIA      <tibble~ <lm>  <tibbl~     0.996         0.996 1.01      2038.
## 13 PFHpA 20MC     <tibble~ <lm>  <tibbl~     0.954         0.948 1.35       166.
## 14 PFHpA NIA      <tibble~ <lm>  <tibbl~     0.945         0.938 2.05       137.
## # ... with 8 more variables: p.value <dbl>, df <dbl>, logLik <dbl>, AIC <dbl>,
## #   BIC <dbl>, deviance <dbl>, df.residual <int>, nobs <int>
```

What you see here is a bit more than what you'd get from *Excel*'s 'line-of-best fit' output. See the section on *Modelling* for a better breakdown of what everything means. But for now, we can see that our `r.squared` of each calibration curve is pretty good, and the `p.value` indicates each model is significant. the `adj.r.squared` is the same as `r.squared` in this situation. This is because `r.squared` will always increase if we add more exploratory variables to our model; the `adj.r.squared` accounts for the number of exploratory variables used in the model. However, in our case we only have one exploratory variable, hence they're the same.

But what about the slope and the intercept? After all, that's what we need to calculate the concentration in our unknowns. Let's take a look at the `tidied` from the `tidy` function "...which constructs a tibble that summarizes the model's statistical findings. This includes coefficients and p-values for each term in a regression..."[2]

```
# storing because we'll use it later on.

tidied <- calCurves %>%
  unnest(tidied)

tidied
```

```
## # A tibble: 28 x 10
## # Groups:   cmpd, location [14]
##    cmpd  location data       fit   term    estimate std.error statistic  p.value
##    <chr> <chr>    <list>     <lis> <chr>      <dbl>     <dbl>     <dbl>    <dbl>
##  1 PFHxA 20MC     <tibble [~ <lm>  (Inter~    0.729   0.388        1.88 9.66e- 2
```

[2] From the *broom* package vignette.

```
##  2 PFHxA 20MC     <tibble [~ <lm>  conc       0.473   0.0111      42.5  1.04e-10
##  3 PFOA  20MC     <tibble [~ <lm>  (Inter~    1.24    0.694        1.78 1.12e- 1
##  4 PFOA  20MC     <tibble [~ <lm>  conc       0.231   0.0199      11.6  2.82e- 6
##  5 PFNA  20MC     <tibble [~ <lm>  (Inter~    0.289   0.205        1.41 1.95e- 1
##  6 PFNA  20MC     <tibble [~ <lm>  conc       0.415   0.00587     70.7  1.79e-12
##  7 PFDA  20MC     <tibble [~ <lm>  (Inter~    0.592   0.348        1.70 1.27e- 1
##  8 PFDA  20MC     <tibble [~ <lm>  conc       0.361   0.00999     36.1  3.78e-10
##  9 PFHxS 20MC     <tibble [~ <lm>  (Inter~    0.298   0.164        1.82 1.07e- 1
## 10 PFHxS 20MC     <tibble [~ <lm>  conc       0.459   0.00472     97.3  1.38e-13
## # ... with 18 more rows, and 1 more variable: glanced <list>
```

Again, a lot more to unpack compared to *Excel*. That's because the `lm` function in R calculates a generalized linear model. `lm` performs a linear regression model, which we normally think of as an equation of the form $y = mx + b$. But, regression models can be expanded to account for multiple variables (hence *multiple linear regression*) of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... \beta_p x_p$$

]

where,

- y = dependent variable
- x = exploratory variable; there's no limit how many you can input
- B_0 = y-intercept (constant term)
- B_p = slope coefficient for each explanatory variable

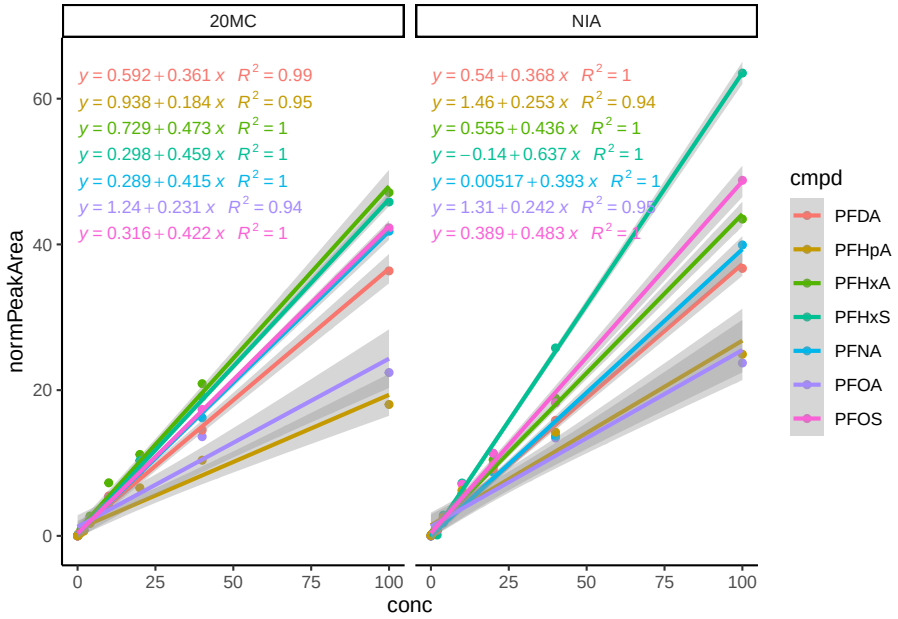
In our situation, we only have one input variable for our model (`conc`), so the above formula collapses down to $y = \beta_0 + \beta_1 x_1$. So looking at our results above, each row corresponds to a model parameter for a given compound and location. For each modelling parameter, we're provided an estimate of it's numerical value (`estimate`, the values we'll use to calculate concentration). The other parameters are useful to understand but not necessary at this point (again, check out the *Modelling* section).

16.3.1 Plotting regression curves

It's always a good idea to visualize our models fit, and it's definitely expected when it comes to calibration curves. So let's go ahead and plot ours:

```
ggplot(QQQstds, aes(x = conc, y = normPeakArea, colour = cmpd)) +
  geom_point() +
  facet_grid(cols = vars(location)) +
```

```
  geom_smooth(method='lm') +
  ggpmisc::stat_poly_eq(formula = y ~ x,
                        aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")),
                        parse = TRUE, size = 3) +
  theme_classic()
```



Note the grey area around each linear model fitting is the predicted 95% confidence interval for that model. In other words, 95% of our peak areas should fall inside those lines. Also note the difference in instrument response for different compounds. This is why true quantification requires authentic standards.

Let's extract what we need and move on:

```
terms <- tibble(cmpd = tidied$cmpd,
                location = tidied$location,
                term = tidied$term,
                estimate = tidied$estimate) %>%
  pivot_wider(names_from = "term",
              values_from = "estimate") %>%
  rename(intercept = `(Intercept)`) %>%
  rename(slope = conc)

head(terms)
```

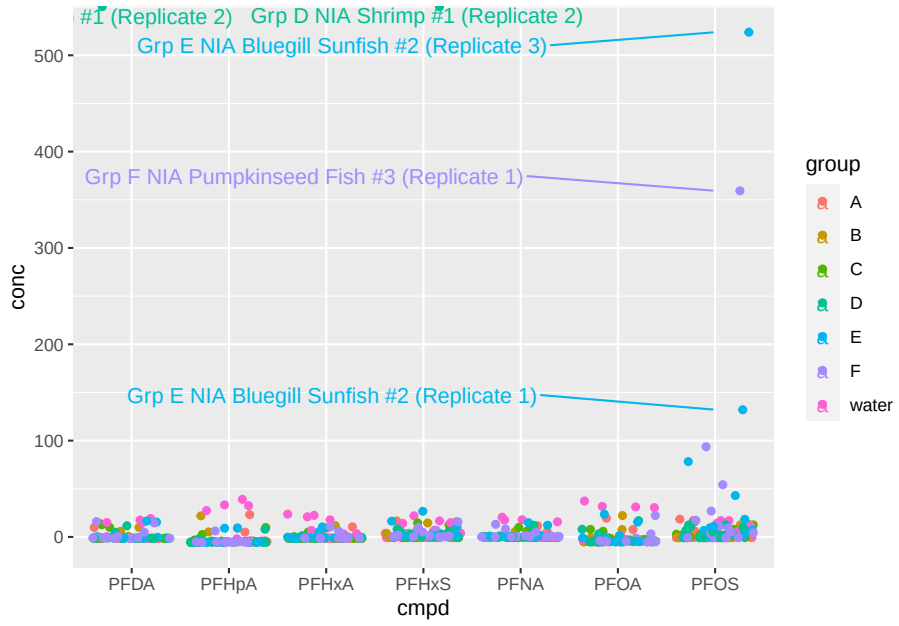
```
## # A tibble: 6 x 4
```

```
##   cmpd  location intercept slope
##   <chr> <chr>        <dbl> <dbl>
## 1 PFHxA 20MC         0.729 0.473
## 2 PFOA  20MC         1.24  0.231
## 3 PFNA  20MC         0.289 0.415
## 4 PFDA  20MC         0.592 0.361
## 5 PFHxS 20MC         0.298 0.459
## 6 PFOS  20MC         0.316 0.422
```

16.4 Quantifying sample concentrations

Let's pair up each compound with it's calibration curve terms, so we can quantify each sample:

```
unknowns <- QQQNorm %>%
  filter(type != "std") %>%
  inner_join(terms, by = c("cmpd", 'location')) %>%
  mutate(conc = (normPeakArea - intercept)/slope)

ggplot(unknowns, aes ( x = cmpd, y = conc, colour = group, label = sampleID)) +
  geom_jitter() +
  ggrepel::geom_text_repel(aes(label=ifelse(conc > 100 , as.character(sampleID),'')),hjust=0,vjus
```

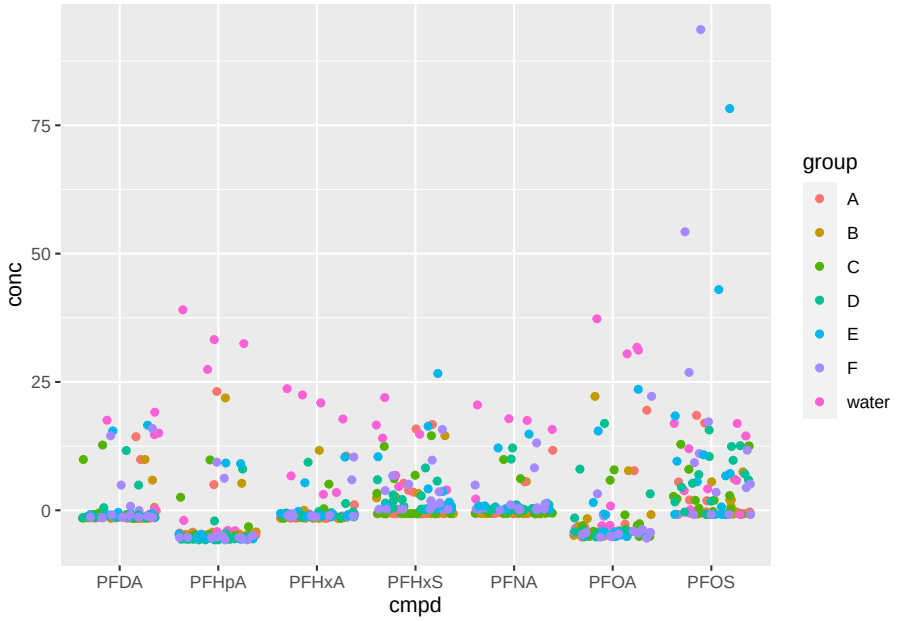


Thanks to our annotation with `ggrepel` we know from which samples our outliers came from. Note I'm using outlier here to mean anything outside our calibration curve, and not a statistical outlier, well get to that. Investigating our original data we see that:

- Neither *13C2PFDA* or *13C4PFHxS* were detected in `Grp D NIA Shrimp #1 (Replicate 2). Further inspection shows that all of the internal standards are extremely low compared to the other samples. This indicates a missloading of the internal standard, hence the out of whack normalized peak area, and subsequent concentration. - Recall in R,`$1/0 =$ `Inf, so an undetected internal standard is reported as`0‘, leading to an infinite concentration.
  - This entire sample should be removed from further analysis.
- Similar situation with `Grp E NIA Bluegill Sunfish #2 (Replicate 1)`, and although all internal standards where detected, their peak areas are close to the instrument cutoff (~1000 counts).
- Again, similar story with `Grp F NIA Pumpkinsee fish #3 (Replicate 1)`

For these samples, it may be the result of some serious matrix effects, but I doubt it. Let's just saw we'll remove them, and see what's left:

```
unknowns <- unknowns %>%
  filter(conc < 120)

ggplot(unknowns, aes ( x = cmpd, y = conc, colour = group, label = sampleID)) +
  geom_jitter()
```



So we can see many samples have slightly negative concentrations. These are the result of the back calculation, and the errors in our model. We can establish an instrumental cutoff (i.e. anything less than X becomes a flat value). It is interesting how the PFOA and PFHxA results are substantially more negative then the others two. This may be due to the *13C4PFOA* internal standard, which was also used to calculate the *PFHxA* concentrations.

Anyways, always something to do…

Chapter 17

Non-linear Logistic Regression Modelling

Note this is from the proof-of-concept book and will be shortened/narrowed down to talk about non-l;inear regression in general while using a logistic regreassion as an example. Will be reworked shortly, just posting so people get an idea. - DH

For this tutorial we'll be using data obtained from an experiment in *CHM317*. In this experiment, students measure the fluorescence of the fluorescent dye acridine orange in the presence of sodium dodecyl sulfate (SDS). In, or near, the critical micellular region of SDS, there is a sharp change in absorbance and fluorescence of the solution. Tracking these changes in fluorescence, students are to estimate the CMC of SDS.

The setting of the fluorometer for this experiment are:

Instrumental Settings	
Instrument Name	LS50-B
Excitation Wavelength	480 nm
Emission Wavelength Range	500 to 650.5 nm
Excitation Slit Width	2.5 nm
Emission Slit Width	3 nm
Scan Speed	250 nm/min

Let's go ahead and import our data:

```
library(tidyverse)
```

```
sdsWide <- read_csv("data/CHM317/fluoro_SDSCMC.csv")

head(sdsWide)
```

```
## # A tibble: 6 x 10
##   `Wavelength (nm)` `0.001 M SDS` `0.0016 M SDS` `0.004 M SDS` `0.0048 M SDS`
##               <dbl>         <dbl>          <dbl>         <dbl>           <dbl>
## 1               500          20.0           18.6          7.02            1.12
## 2               500.         27.3           13.9          5.45            5.46
## 3               501          27.0           12.5          8.13            5.89
## 4               502.         29.7           12.5          8.17            5.81
## 5               502          32.6           15.7          4.58            6.69
## 6               502.         32.8           19.4          5.94            5.33
## # ... with 5 more variables: 0.0056 M SDS <dbl>, 0.0064 M SDS <dbl>,
## #   0.0072 M SDS <dbl>, 0.008 M SDS <dbl>, 0.012 M SDS <dbl>
```

Looking at our data headers we can see the familiar 'wide' format, with a wavelength column corresponding to the emission wavelength and the remainder accounting for the emission intensity at various concentration of SDS. Note that the intensity columns contains two pieces of information: 1) the concentration in moles per liter and 2) the identity of the chemical, SDS in this case. So when we tidy our data we'll need to split these column headers up so we get a column corresponding to the numerical value of the concentration, and another with the identity of the column.

```
sds <- sdsWide %>%
  pivot_longer(cols = !`Wavelength (nm)`, # select all columns BESIDES `Wavelength (nm
               names_to = c("conc", "conc.units", "chemical"),
               names_pattern = "(.*) (.) (.*)",
               values_to = "intensity",
               names_transform = list(conc = as.numeric)
  ) %>%
  rename(wavelength = 'Wavelength (nm)') # renaming column, less typing later on.

head(sds)
```

```
## # A tibble: 6 x 5
##   wavelength   conc conc.units chemical intensity
##        <dbl>  <dbl> <chr>      <chr>        <dbl>
## 1        500 0.001  M          SDS          20.0
## 2        500 0.0016 M          SDS          18.6
## 3        500 0.004  M          SDS           7.02
## 4        500 0.0048 M          SDS           1.12
## 5        500 0.0056 M          SDS           5.48
## 6        500 0.0064 M          SDS           7.72
```

The key bit of code here is `names_to` and `names_pattern`. The first part creates three new columns, and the second part searches and subsequently breaks up those headers. Recall our original headers looked like this: `0.001 M SDS`, where we had the concentration, a space (which is a character!), the concentration units, another space, and finally the chemical. What `names_pattern = "(.*) (.) (.*)`.is searching for three chunks of characters separated by a space. We specify the chunk of characters in the parentheses. So the first bit, `(.*)`, means look for any character (`.` in this context is a placeholder for *any* character) and the chunk of characters can be any length (as denoted by `*`). So extending this, we see our code looks for three chunks of characters, delimited by a space between them. The first can be any length, the second is 1 character long, and the third can be any length. You could have specified to look for `M` or `SDS` explicitly, but if we had different chemicals or units in our dataset these would be lost.

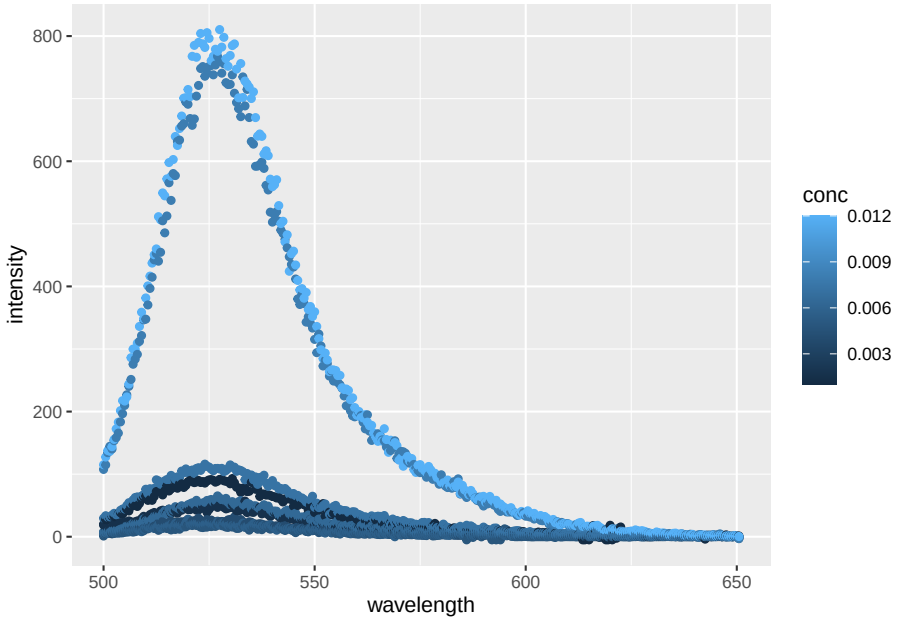
Lastly note `names_transform`. We split up our original headers to populate rows. However our original headers were stored as characters, and when w split them up we created three separate strings of characters, so R will treat out `conc` values as characters rather than numbers. By using `names_transform` we tell R to treat `conc` values as numbers.

Oh and we renamed our original `Wavelength (nm)` column to `wavelength` using the `rename` function. It's always a good idea to use the simplest column names you can (and no simpler!). A good practice is to remove any spaces (you can use `snake_case` or `camelCase` instead) as well as removing special character such as parentheses.

17.1 Visually inspecting our data

Let's make a quick plot of our fluorescence intensity data and see what we have.

```
ggplot(data = sds,
       aes(x = wavelength,
           y = intensity,
           colour = conc)) +
  geom_point()
```



Alright, alright, alright. Things are looking like we'd expect with some well behaved data. By plotting each point individually, we can really see the noise inherent with each reading. For a more robust analysis we'd typically conduct several replicates and average out the spectra for each concentration or apply some kind of model to smooth each peak. But today, we're just interested in getting the maximal fluorescence emission intensity from each reading.

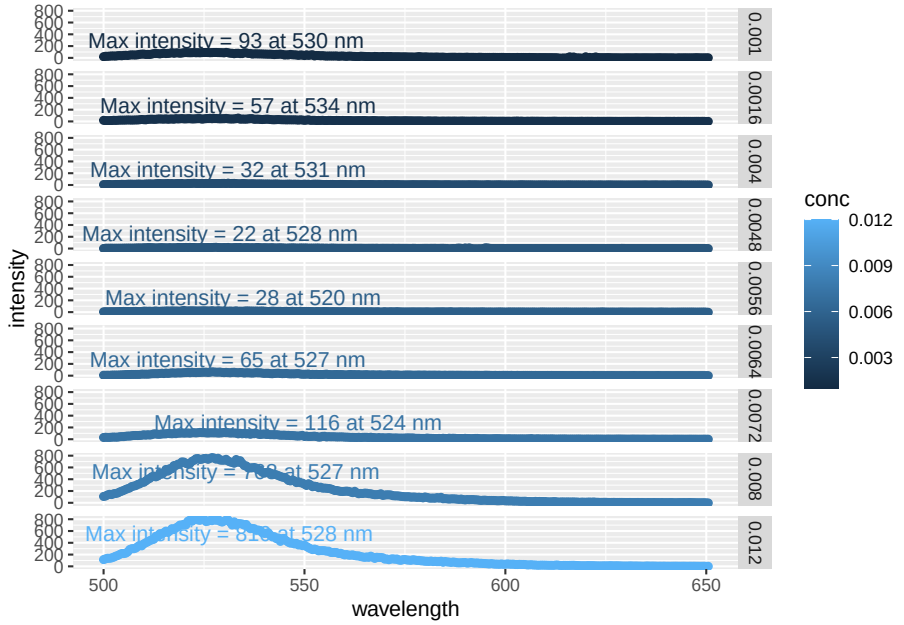
Let's first annotate our plate to find the highest point, then go about extracting our data for analysis.

17.1.1 Annotating maximal values

Annotating the maximal point on the plot will take a bit more code then actually obtaining it from the data. For this we'll need to use the `ggpmisc` package which contains miscellaneous extensions for `ggplot2`, and `ggrepel` so our labels won't overlap.

```
library(ggpmisc)
library(ggrepel)

ggplot(data = sds,
       aes(x = wavelength,
           y = intensity,
           colour = conc)) +
  geom_point() +
```

```
  ggpmisc::stat_peaks(span = NULL,
                      geom = "text_repel", # From ggrepel
                      mapping = aes(label = paste(..y.label.., ..x.label..)),
                      x.label.fmt = "at %.0f nm",
                      y.label.fmt = "Max intensity = %.0f",
                      segment.colour = "black",
                      arrow = grid::arrow(length = unit(0.1, "inches")),
                      nudge_x = 60,
                      nudge_y = 200) +
  facet_grid(rows = vars(conc))
```



By facetting the plot (i.e. arranging many smaller plots vs. one large one), we can easily see the increase in emission peak intensity as the concentration of SDS increases. Likewise, we can avoid the messy overlap of the max intensity annotations.

This is only one way to plot this data, but this is sufficient because we're simply inspecting our data at this point. And here we can see that the intensity all occur around a similar wavelength (~ 528 nm)

17.2 Extracting maximal values

The plots we made above are great for inspecting our data, but what we really want is the maximal emission intensity value to calculate the CMC of SDS. We

can see the maximal values on the plots, but there's no way we're typing those in manually. So let's go ahead and get out maximal values from our dataset:

```
sdsMax <- sds %>%
  group_by(chemical, conc.units, conc) %>%
  filter(intensity == max(intensity)) %>%
  ungroup()

head(sdsMax)
```

```
## # A tibble: 6 x 5
##   wavelength   conc conc.units chemical intensity
##        <dbl>  <dbl> <chr>      <chr>        <dbl>
## 1       520  0.0056 M          SDS           28.1
## 2       524  0.0072 M          SDS          116.
## 3       527  0.0064 M          SDS           65.3
## 4       527  0.008  M          SDS          768.
## 5       528. 0.012  M          SDS          810.
## 6       528  0.0048 M          SDS           22.0
```

All we did was tell R to take the row with the highest emission intensity value per group. We specified `chemical`, `conc.units`, and `conc`, in case we had more chemicals in our dataset.

Are maximum values match those we see in our plot above. Let's see how they stack up againt each other:

```
ggplot(data = sdsMax,
       aes(x = conc,
           y = intensity)) +
  geom_point()
```

17.3 Modelling Sigmoidal Curve

So we want to find the critical micellular concentration of SDS using the maximum fluorescence emission. The CMC is at the 'midpoint of the sigmoidal curve.' Which means we'll need to a) plot a sigmoidal curve and b) extract the midpoint.

The 'sigmoidal' or 'S-shaped' curve mentioned in the lab manual is known as a *logistic regression*. Logistic regressions are often used to model systems with a largely binary outcome. In other words, the system starts at point A, and remains there for awhile, before 'quickly' jumping up (or down) to level B and

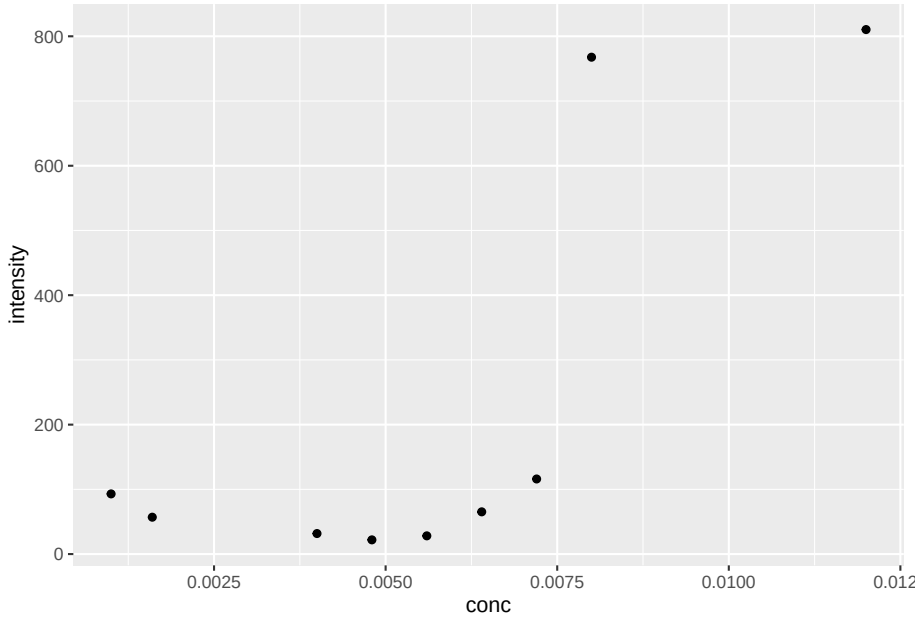


Figure 17.1: Plot of maximal fluoresence intensity at various concentrations of SDS.

remain there for the remainder. Examples include saturation and dose response curves.

For our CMC working data, the fluorescence intensity is low when the $[SDS] < CMC$, as micelles are not able to form. However once $[SDS] > CMC$, micelles form and the fluorescence intensity increases. We can see this trend in 17.1.

There are different forms of logistic regression equations. The simplest form is the 1 parameter, or sigmoid, function which looks like $f(x) = \frac{1}{1+e^{-x}}$. The outputs for this function are between 0 and 1. We could apply this formula to our model if we somehow normalized our fluoresence intensity accordingly. An alternative is to use the *four parameter logistic regression*, which looks like:

$$f(x) = \frac{a-d}{\left[1+\left(\frac{x}{c}\right)^b\right]} + d$$

where:

- **a** = the theoretical response when $x = 0$
- **b** = the slope factor
- **c** = the mid-range concentration (inflection point)

  - This is commonly referred to as the *EC50* or *LC50* in toxicology/pharmacology.

- **d** = the theoretical response when $x = \infty$

Why do we need such a complicated formula for our model? Well, looking again at 17.1 we see that the lower point is approximately 20, and not zero. Likewise, the upper limit appears to be around 825. The slope factor is necessary because the transition from the low to high steady state occurs over a small, but not immeasurable, concentration range. And lastly, by including the inflection point, we can calculate exactly for this value using `R` to get the CMC estimate.

17.3.1 Calculating Logistic Regression

A strength of `R` is it's flexibility in running various models, and logistic regression is no different. We can use a number of packages to reach these ends, specifically the `drc` package contains a plethora of functions for modelling dose response curves (hence `drc`). However, for this example well use a more generalized approach. Earlier we talked about linear regression, where we plot adjust the slope and intercept of a linear equation to best fit our data (see Calibration Curves). Recall that this optimization is based on minimizing the distance between the model and all of the experimental points (*least squares*). Well the `stats` package has a function called `nls` that expands upon the this to nonlinear models. Per the `nls` function description: "[nls] determine[s] the nonlinear (weighted) least_squared estimates of the parameters of a nonlinear model."

So we can create a formula in `R` based on the four-parameter logistic regression described above. After that, we'll need to produce some starting details from which the model can build off of. If we don't tell `nls` where to start, it can't function, as the search space is too large. Looking at @ref{fig:sdsMaxPlot}, the intensity appears to floor around 20; the intensity appears to max out around 820; the midpoint appears to be around 0.0075 M, and let's say the slope is 1. Remember, these are starting values from which `nls` starts to optimize from, and not the actual values used to construct the model.

So, let's create our model

```
logisModel <- nls(intensity ~  (a-d)/(1 + (conc /c)^b) + d,
                  data = sdsMax,
                  start = list(a = 20,       # min intensity
                               b = 1,        # slope
                               c = 0.0075,   # CMC
                               d= 820)       # max intensity
                  )
```

```
## Error in numericDeriv(form[[3L]], names(ind), env, central = nDcentral): Missing value or an i
```

… and we get an error message. Get used to these when modelling! Don't worry about understanding it completely, error messages are often written with programmers in mind so they can be a bit cryptic. You can often copy and paste these directly into any search engine to get some more information, but this one is simple enough: we either have a missing value or an infinity produced. Well we have six input parameters in our model: `a, b, c, d`, our independent variable `conc`, and our dependant variable `intensity`. We've also supplied starting values to all of them via the list we created inside the function. Therefore, one of our starting values must be too far off from a plausible start point and is causing troubles in the `nls` function. They all look good except for the slope start value `b = 1`.

The slope here is an approximation for the slope between the min value `a` and max value `d`. Looking at our data in @ref{fig:sdsMaxPlot}, that slope may be a bit shallow consider the large jump in intensity. Let's increase the value of `b` and try again:

```
logisModel <- nls(intensity ~  (a-d)/(1 + (conc /c)^b) + d,
                  data = sdsMax,
                  start = list(a = 20,        # min intensity
                               b = 10,        # new slope
                               c = 0.0075,    # CMC
                               d= 820)        # max intensity
                  )
```

Ey, no errors! Once you progress beyond simple linear regressions, modelling becomes more of a craft. If we were trying to apply this model to multiple datasets, we would probably want to shop around `cran` to find a package with self-starting models. This way we can circument having to supply starting parameters. Anyways, that's for another day.

For now, let's take a look at our model outputs which are all stored in the `logisModel` variable. To this end, we'll use the `broom` package which cleans up the default model outputs in `R`. Specifically, we'll use `tidy` to get an output of our estimated model parameters (i.e. `a,b,c`, and `d`), and `augment` for a data frame of containing the input values, and the estimated intensity values.

Let's look at our fitted values:

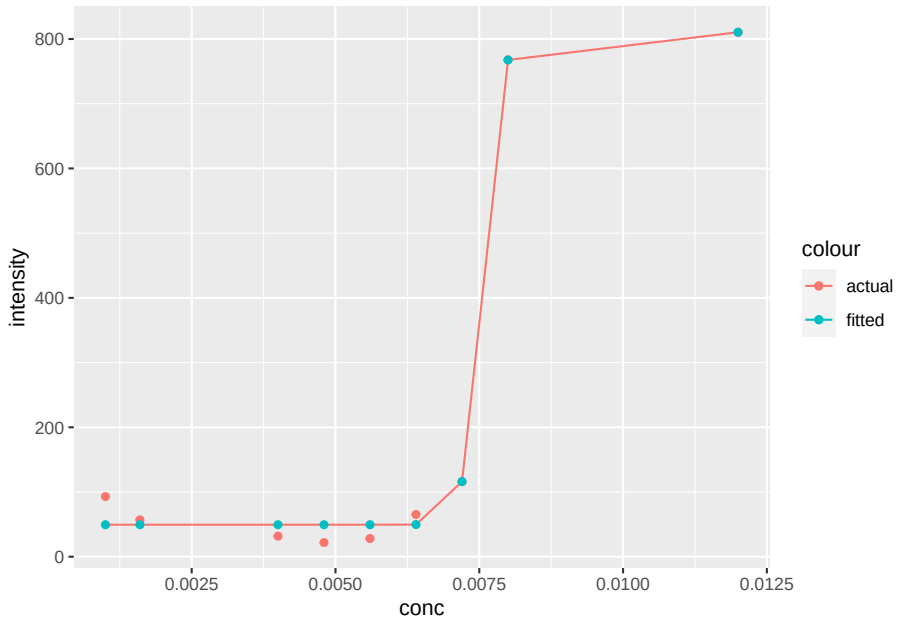
```
library(broom)

augment <- augment(logisModel)
augment
```

```
## # A tibble: 9 x 4
##   intensity   conc .fitted   .resid
##       <dbl>  <dbl>   <dbl>    <dbl>
## 1      28.1 0.0056    49.4 -21.3
## 2     116.  0.0072   116.   -0.123
## 3      65.3 0.0064    49.7  15.6
## 4     768.  0.008    768.    0.0977
## 5     810.  0.012    810.   -0.0861
## 6      22.0 0.0048    49.4 -27.5
## 7      93.0 0.001     49.4  43.5
## 8      31.7 0.004     49.4 -17.7
## 9      57.0 0.0016    49.4   7.51
```

What we can see here from `augment` are the `intensity` and `conc` values we inputted into R. `.fitted` are the intenisty values for a given concentration fitted to out model, and `.resid` is the residuals, the difference between the actual and estimated values.

Let's go ahead and plot our actual and fitted values against each other.

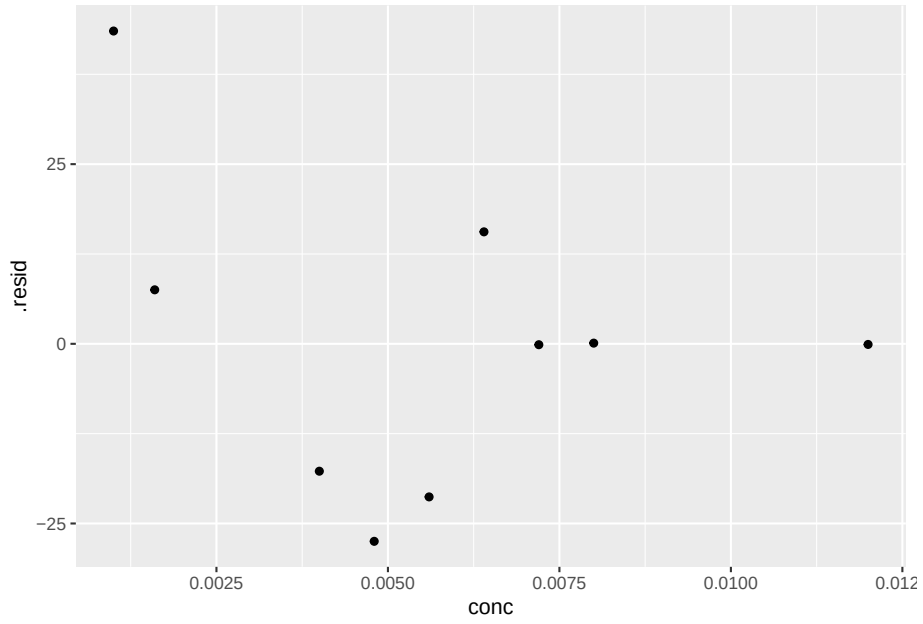
```
ggplot(augment, aes(x = conc, y = intensity, colour = "actual")) +
  geom_point() +
  geom_line(aes(y = .fitted)) +
  geom_point(aes(y = .fitted, colour = "fitted"))
```



Looks pretty good, although it's interesting how the baseline at lower concentrations doesn't plateau like the model values. You'll note that the line produced by `geom_line` will only draw a straight line between points. There's ways to address this, but we don't need to for our needs right now.

Looking again at our model results, there doesn't appear to be any gross outliers, so our model seems to have done a good job. We can verify this by checking the residuals:

```
ggplot(augment, aes(x = conc, y = .resid)) +
  geom_point()
```



We can't see any obvious patterns in the residuals (i.e. all are negative), so we can have further confidence in the fit of out model.

17.3.2 Extracting model parameters

To extrac the model parameters `a`, `b`, `c`, and `d` we can use the `tidy` function:

```
library(broom)

tidy <- tidy(logisModel)
tidy
```

```
## # A tibble: 4 x 5
##   term   estimate  std.error statistic       p.value
##   <chr>     <dbl>      <dbl>     <dbl>         <dbl>
## 1 a      49.4     11.1            4.44 0.00678
## 2 b      49.0      9.65           5.07 0.00385
## 3 c       0.00755  0.0000785     96.2  0.00000000230
## 4 d     810.      27.3           29.7  0.000000808
```

Looking past the scientific notation, our model values are pretty similar to what we estimated. Specifically, `c`, our midpoint value is 0.0076 M. Not too bad from our original estimate. And recall that the midpoint of our curve corresponds to the critical micellular concentration of SDS, which we've estimated to be 0.0076M. Not too far from the literature value of 0.0081 M.

Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 328. https://doi.org/10.1198/jcgs.2009.07098.

———. 2014. "Tidy Data." *Journal of Statistical Software* 59 (1): 1–23. https://doi.org/10.18637/jss.v059.i10.