A Short Introduction to Working With Data in R

Jonathan Whiteley

2023-09-18

Prerequisites

- Access to a copy of the R software
  - Get it from *www.r-project.org*, or ask your system administrator.
- Tidyverse packages installed on the same system as R
  - Please run this command in R *before* the workshop:

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

- Download the workshop files, including these slides, data, and scripts.
  - The workshop assumes the same file structure as in the link above.
- Knowledge of common mathematical operations: arithmetic, logarithms, etc.
- Knowledge of basic R concepts, such as *variables*, *objects*, *operators*, *functions*, *packages*, etc.
  - This is covered in the first workshop: "A Gentle Introduction to R"

Section 1

Welcome

Pop Quiz

We will review these *at the end*, so you can see how much you have learned.

- If multiple packages have functions with the same name, how can you specify which one to use?
- What are “UTF-8” and “latin1”?
- What’s the difference between a dplyr ‘verb’ and a ‘helper function’?
- Does the `filter()` verb use ‘select’ or ‘mutate’ *semantics*?
- What are the 3 rules of ‘tidy data’?
- TRUE or FALSE: If `R` reads a file without errors, there are no problems
- TRUE or FALSE: `R` has rules and conventions for naming functions
- TRUE or FALSE: if you use one package from the `tidyverse`, you have to use all of them.

Answer in the chat:

What is your favourite emoji? Why do you like to use it so much?

Introductions

- Name
- Job title, role

If you are comfortable sharing:

- Pronouns
- A hobby or activity you enjoy
- What are you hoping to learn most in today's workshop?

Learning Objectives

- Load tabular data into R
- Explore data to check that it was loaded correctly
- Export data from R to external files
- Data frames
- Clean data
  - Re-arrange & modify rows
  - Add & change columns
  - Edit values systematically
  - Change data types
- Tidy data
  - Change the *shape* of a data frame
- Re-use code, reproducible results
  - Scripts

Disclaimer

- There is often more than one way to achieve a desired result in `R`
- Some are faster in certain situations
- Some require less code, or are easier to write as code
- Some are more portable (work on multiple systems)
- But there is rarely a single 'best way'.

This workshop focuses on a coherent approach, that can be learned more easily and extended as needed to tackle bigger problems.

Feel free to take what you learn here and experiment, or explore alternatives. Find what works for *you*.

info

I may not speak to all slides during the workshop: some are mostly for reference, but may be helpful for the activities that you do on your own.

Section 2

The `tidyverse` collection of packages

The tidyverse

```
install.packages("tidyverse")
help(package="tidyverse")
```

- The tidyverse is an “opinionated” collection of packages that are designed to work together.
- All packages share an underlying design philosophy, grammar, and data structures.
    - *Unlike base R*
    - Shared naming conventions (e.g., ‘_’ instead of ‘.’ in function names)
    - Emphasis on functions that do one thing well
    - Designed to be combined together to achieve complex operations
- tidyverse is under active development.
    - New functions and features sometimes replace or supersede old ones.
    - No guarantee that functions will continue to work the same way in future versions.

Core `tidyverse` packages

Today, we will focus on a few of the core `tidyverse` packages for loading, cleaning, and manipulating data:

- readr, readxl for **loading** data
- dplyr for **manipulating** data (values)
- tidyr for **reshaping** data
- stringr for working with **strings**

Section 3

File Paths and The Working Directory

The Working Directory

- When working with external files, it helps to know the current *working directory*
  - Any paths supplied to R functions will be relative to this path.

```
getwd()
```

- You can change the working directory with this command:

```
setwd('path/to/a/directory')
```

File paths

- A *file path* is a *character string* that represents the location of a file in your system (computer and OS)
- The format of paths can depend on the operating system (OS)
  - Some use `"/"` to separate directories

    e.g., "`/dir/subdir`"
  - Windows uses `"\"`

    e.g., `"C:\\dir\subdir"`

    `R` uses this as an escape character in strings, and must be escaped itself in paths (`"\\"`)

    e.g., `"C:\\\\dir\\subdir"`

Paths in R

- R generally uses and understands "/" in paths, even on Windows.[1]
    - e.g., "C:/dir/subdir"
    - on Windows, it also understands Windows-style paths: e.g, "C:\\\\dir\\subdir"
- R also has platform-independent functions for manipulating paths, such as `file.path()`, which I will use in examples to make them as reproducible as possible.

[1] For the gory details, see section 14.2 “Filepaths” in “An Introduction to R” (`help.start()`), `?file.path`, and documentation for related functions.

My paths are not like yours

- Directory (folder) names can also vary from one computer to another — it's difficult to show a path in this document that will also work on your computer!
- Once you set a working directory *on your computer* based on the structure of the files in this project, we can use *relative paths* that should also work on your computer (assuming your downloaded the workshop files in the same structure as provided).

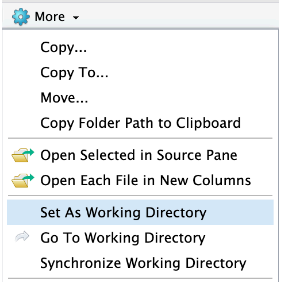
Set the working directory

- For this workshop, set the working directory to location where you downloaded this presentation and accompanying files.
  - the directory that *contains* the folder named '`data`' that you downloaded along with the files for this workshop.

- Base `R` on Mac / Linux:
  - Menu item: "Misc > Change Working Directory..."
  - `CMD+D` on Mac; `CTL+D` on Linux (or Windows)
- In RStudio, you can use the *Files* pane (default bottom-right) to navigate to a directory in your system, and click on "*More > Set As Working Directory*"
  - or "*Session > Set Working Directory > To Files Pane Location*" in the RStudio menu.

- Base `R` on Windows:

```
setwd( choose.dir() )
```


More
Copy...
Copy To...
Move...
Copy Folder Path to Clipboard
Open Selected in Source Pane
Open Each File in New Columns
Set As Working Directory
Go To Working Directory
Synchronize Working Directory

Check your working directory

- Check to see that the working directory is in the right place, by checking to see if a known file exists (from R's perspective):

```r
CSV_path <- file.path("data", "data_example.csv")
file.exists(CSV_path)
```

```
# [1] TRUE
```

- If the result of the statement above is not "TRUE" in your session, try one of the other approaches to change your working directory, and try again.

Section 4

Load Data: The `readr` & `readxl` Packages

csv files

- 'csv' = **C**omma **S**eparated **V**alues
  - files in this format have a '`.csv`' file extension.
- They are:
  - plain text files
  - used to represent tabular data, with each *row* on a line, and values in each *column* separated by commas (`,`)
  - readable by a wide variety of analysis software (highly portable)
  - simple—no embedded metadata
- We'll try to load this file into `R`:

  `example_data.csv`
  - *optional: you can try opening it in a text editor, or spreadsheet software, to see what's in the file.*

Load a csv file using read_csv()

```
?read_csv
```

```
library(readr)
try( read_csv(CSV_path) )
```

```
# Warning: One or more parsing issues, call `problems()` on your dat
# frame for details, e.g.:
#   dat <- vroom(...)
#   problems(dat)
```

Load a csv file using read_csv()

```r
?read_csv
```

```r
library(readr)
try( read_csv(CSV_path) )
```

```
# Warning: One or more parsing issues, call `problems()` on your da
# frame for details, e.g.:
#   dat <- vroom(...)
#   problems(dat)
```

- Uh oh! Something's not right.

Check the file contents

- Let's take a peek at the first few lines and see if we can identify the problem:

```
readLines(CSV_path, n = 4)
```

```
# [1] "Data from an experiment on the cold tolerance of the grass sp
# [2] "Modified from `data(CO2)`.  See `?CO2`."
# [3] "Type,Treatment,PlantNum,95,175,250,350,500,675,1000"
# [4] "Quebec,nonchilled,1,16,30.4,34.8,37.2,35.3,39.2,39.7"
```

Check the file contents

- Let's take a peek at the first few lines and see if we can identify the problem:

```
readLines(CSV_path, n = 4)
```

```
# [1] "Data from an experiment on the cold tolerance of the grass s
# [2] "Modified from `data(CO2)`.  See `?CO2`."
# [3] "Type,Treatment,PlantNum,95,175,250,350,500,675,1000"
# [4] "Quebec,nonchilled,1,16,30.4,34.8,37.2,35.3,39.2,39.7"
```

- The first **2** lines don't look like comma-separated values!
- They look like extra information that is not part of the data table *structure*.

Load a csv file into R

- We can tell R to skip the lines with no data:
  - and we'll *assign* the result to a variable so we can work on it

```
CSV <- read_csv(CSV_path, skip = 2)
```

```
# Rows: 13 Columns: 10
# -- Column specification ------------------------------------
# Delimiter: ","
# chr (3): Type, Treatment, 500
# dbl (6): PlantNum, 95, 175, 250, 350, 1000
# num (1): 675
#
# i Use `spec()` to retrieve the full column specification for this
# i Specify the column types or set `show_col_types = FALSE` to quie
```

- Just because there were no Errors or Warnings from R, doesn't mean there's nothing wrong with the data!

Section 5

Exploring Your Data

Object class: data frame

Before we explore our new data set, let's quickly review the kind of *object* we're dealing with:

```
class(CSV)
```

```
# [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
typeof(CSV)
```

```
# [1] "list"
```

Data frames

Type	Treatment	PlantNum	95	175	250	350	...
Quebec	nonchilled	1	16.0	30.4	34.8	37.2	
Quebec	NA	2	13.6	27.3	37.1	41.8	
Quebec	NA	3	16.2	32.4	40.3	42.1	
Québec	chilled	1	14.2	24.1	30.3	34.6	
Québec	NA	2	9.3	27.3	35.0	38.8	
Québec	NA	3	15.1	21.0	38.1	34.0	
Mississippi	nonchilled	1	10.6	19.2	26.2	30.0	
...	NA	NA	NA	NA	NA	NA	

- **columns** = *data variables*
- **rows** = observations, cases

Data frames: *structure*

- Each column contains data of the *same type*

```
# 'data.frame': 13 obs. of  10 variables:
#  $ Type     : chr  "Quebec" "Quebec" "Quebec" "Québec" ...
#  $ Treatment: chr  "nonchilled" NA NA "chilled" ...
#  $ PlantNum : num  1 2 3 1 2 3 1 2 3 1 ...
#  $ 95       : num  16 13.6 16.2 14.2 9.3 15.1 10.6 12 11.3 10.5 .
#  $ 175      : num  30.4 27.3 32.4 24.1 27.3 21 19.2 22 19.4 14.9
#  $ 250      : num  34.8 37.1 40.3 30.3 35 38.1 26.2 30.6 25.8 18.
#  $ 350      : num  37.2 41.8 42.1 34.6 38.8 34 30 31.8 27.9 18.9
#  $ 500      : chr  "35.3" "40.6" "42.9" "32.5 (umol/m^2 sec)" ...
#  $ 675      : num  39.2 41.4 43.9 354 375 396 32.4 31.1 28.1 22.2
#  $ 1000     : num  39.7 44.3 45.5 38.7 42.4 41.4 35.5 31.5 27.8 2
```

!

`readr` doesn't just read data into a `data.frame`, but the result is *also* a `tibble` (`"tbl_df"`).

Tibbles: `data.frames` reimagined

- A 'tibble' (`class() == "tbl_df"`) is "a modern reimagining of the data.frame".
  - See the package documentation for details.
- Many tidyverse functions produce tibbles by default.
- **Tibbles are *also* `data.frames`**, and *inherit* from that class.
  - functions that work with `data.frames` should also work with tibbles.
  - but some may behave differently (by design).
  - for example, `print()`ing a tibble includes slightly more information, and only prints a few rows and columns by default, preventing large datasets from overwhelming your console.
  - when indexing a tibble, it will *not* do partial matching on column names, making it clear if a column exists or not
- For most purposes, tibbles are interchangeable with `data.frames`.
  - A tibble can be converted to a 'plain' `data.frame` with `as.data.frame()` if necessary.

head(): peek at the first few rows

```
head(CSV)
```

```
# # A tibble: 6 x 10
#   Type   Treatment  PlantNum  `95` `175` `250` `350` `500`
#   <chr>  <chr>         <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
# 1 Quebec nonchilled        1  16    30.4  34.8  37.2 35.3
# 2 Quebec <NA>              2  13.6  27.3  37.1  41.8 40.6
# 3 Quebec <NA>              3  16.2  32.4  40.3  42.1 42.9
# 4 Québec chilled           1  14.2  24.1  30.3  34.6 32.5 (~
# 5 Québec <NA>              2   9.3  27.3  35    38.8 38.6
# 6 Québec <NA>              3  15.1  21    38.1  34   +38.9
# # i 2 more variables: `675` <dbl>, `1000` <dbl>
```

Dimensions (rows & columns)

```
dim(CSV)
```

```
# [1] 13 10
```

```
nrow(CSV)
```

```
# [1] 13
```

```
ncol(CSV)
```

```
# [1] 10
```

Names of elements (columns)

```
names(CSV)
```

```
#  [1] "Type"         "Treatment" "PlantNum"  "95"
#  [5] "175"          "250"       "350"       "500"
#  [9] "675"          "1000"
```

```
colnames(CSV)
```

```
#  [1] "Type"         "Treatment" "PlantNum"  "95"
#  [5] "175"          "250"       "350"       "500"
#  [9] "675"          "1000"
```

```
rownames(CSV)
```

```
#  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11"
# [12] "12" "13"
```

Look at a column

Remember: you can refer to elements within a data frame by *name*.

```
CSV[["Treatment"]]
```

```
#  [1] "nonchilled" NA           NA           "chilled"
#  [5] NA           NA           "nonchilled" NA
#  [9] NA           "chilled"    NA           NA
# [13] NA
```

```
unique(CSV$Type)
```

```
# [1] "Quebec"      "Québec"      "Mississippi"
```

!

Looks like there might be some missing values in the `Treatment` column, and inconsistencies in the `Type` column. We'll learn how to fix those soon, but these simple functions are already helping us understand our data.

str(): structure of an object

```
str(CSV)
```

```
# tibble [13 x 10] (S3: tbl_df/tbl/data.frame)
#  $ Type      : chr [1:13] "Quebec" "Quebec" "Quebec" "Québec" ...
#  $ Treatment: chr [1:13] "nonchilled" NA NA "chilled" ...
#  $ PlantNum : num [1:13] 1 2 3 1 2 3 1 2 3 1 ...
#  $ 95        : num [1:13] 16 13.6 16.2 14.2 9.3 15.1 10.6 12 11.3
#  $ 175       : num [1:13] 30.4 27.3 32.4 24.1 27.3 21 19.2 22 19.4
#  $ 250       : num [1:13] 34.8 37.1 40.3 30.3 35 38.1 26.2 30.6 25
#  $ 350       : num [1:13] 37.2 41.8 42.1 34.6 38.8 34 30 31.8 27.9
#  $ 500       : chr [1:13] "35.3" "40.6" "42.9" "32.5 (umol/m^2 sec
#  $ 675       : num [1:13] 39.2 41.4 43.9 354 375 396 32.4 31.1 28.
#  $ 1000      : num [1:13] 39.7 44.3 45.5 38.7 42.4 41.4 35.5 31.5
```

Tip

*The `str()` and `names()` functions can be used with **any** object.*

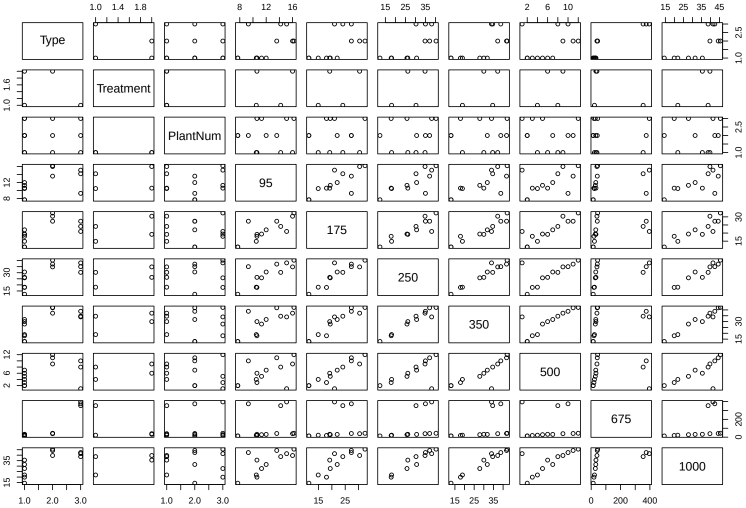
summary(): statistical summaries by column

```r
summary(CSV)
```

```
#      Type              Treatment            PlantNum
#  Length:13          Length:13          Min.   :1
#  Class :character   Class :character   1st Qu.:1
#  Mode  :character   Mode  :character   Median :2
#                                        Mean   :2
#                                        3rd Qu.:3
#                                        Max.   :3
#
#        95            175            250
#  Min.   : 7.7   Min.   :11.4   Min.   :12.3
#  1st Qu.:10.5   1st Qu.:18.0   1st Qu.:23.9
#  Median :11.3   Median :21.0   Median :30.4
#  Mean   :11.9   Mean   :21.4   Mean   :28.9
#  3rd Qu.:14.2   3rd Qu.:27.3   3rd Qu.:35.5
#  Max.   :16.2   Max.   :32.4   Max.   :40.3
#                                NA's   :1
#       350           500                675
#  Min.   :13.0   Length:13          Min.   : 13.7
```

Simple plots

```
plot(CSV)
```



Spreadsheet-like `View()`

```
View(CSV)
```

- This command opens a data frame in a spreadsheet-like view, which can be easier to navigate.
- In RStudio, you can achieve the same thing by clicking on an object name in the '*Environment*' pane (default upper-right)
  - The `View()` pane in RStudio (default upper-left; '*Source*') also allows for sorting and filtering, but these do not change the object in your session, only the view.

Know Your Data

- These functions are useful for exploring different aspects of a loaded data set
- But they won't tell you if these are *correct*.
- Ideally, you should always "Know Your Data", and use these functions to verify that the data was loaded correctly.
  - Are the number of rows and columns what you expected?
  - Are the different columns of the expected type (numeric, character, etc.)?
  - Are the values in the expected range and format?
  - Is anything missing, or different than expected?

The CO2 dataset: background

The example data file is based on the 'CO2' dataset available in R (?CO2), with a few changes added to make things interesting.

From the documentation:

> *The CO2 uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO2 concentration. Half the plants of each type were chilled overnight before the experiment was conducted.*

```r
data(CO2)
```

Activity 1: what's wrong with this data?

The original dataset has the following properties (`str(CO2)`):

- 84 rows and 5 columns

Column Name	Description
`Plant`	factor with 12 levels: Qn1, Qn2, … Mc3, Mc1
`Type`	factor with 2 levels: "Quebec" and "Mississippi"
`Treatment`	factor with 2 levels: "nonchilled" and "chilled"
`conc`	numeric: ambient carbon dioxide concentrations (mL/L)
`uptake`	numeric: carbon dioxide uptake rates (μmol/m^2·sec)

Your turn

Using the functions described in this section, can you identify some possible issues and differences with the data set you loaded?
Spoiler alert: *suggested answers on the next slide.*

Activity 1: what *is* wrong with this data

- The data we loaded has different dimensions!
  - Values from the `conc` column are shown as *column names*
  - `uptake` values are the values of these columns
  - This isn't necessarily *bad*: such a structure can be useful for presentation and interpretation by people, but it is not *tidy* and less convenient for analysis & visualization (more on this later).
- Some of the `uptake` values are *character*, but should be *numeric*
- One of the `Type` values is spelled inconsistently: "Quebec"/"Québec"
- Some values in the `Treatment` column are empty
  - The value is only included when it changes
- The `PlantNum` column does not contain a unique identifier, as in the original `Plant` column
  - The values are no longer *unique*, without also considering the `Type` and `Treatment` columns.

There are other differences you may have noticed: we'll look at ways to identify these automatically later.

Section 6

Re-using your code: scripts and other files

Re-using code

Before we practice cleaning our data, and saving it to use later …

- Imagine having to repeat the multiple steps to *load*, *clean*, and *save* a dataset.
  - How will you remember which *packages* you had to load?
  - How will you remember *all* the steps, and their order?
  - What if you need to change *one* step, but repeat all preceding steps?
  - How can you share your code with others, so that they can check your work, or replicate your results?
  - How will you write out complex operations that require multiple steps, repeated operations, or only do things under certain conditions?
- The answer to all of these questions is: a ***script***

Scripts and related files will also make it easier for you to follow along with examples as they get more complicated—copy & paste into the console less often!

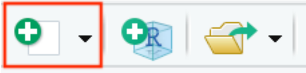
Scripts

An `R` *script* is a file that stores `R` code in *plain text*

- They have a `.R` file extension
- They are plain text files
  - so any text editor can read & write them
  - they also work well with **v**ersion **c**ontrol **s**ystems, like `git`, GitHub, and GitLab)
- All the code in a script can be run in order
  - i.e., a *program*
- They make it easy to re-use code
- Scripts provide a record of the steps in a program or analysis
  - results are more *reproducible*
  - the code is a form of documentation

Make a new script

- In your `R` interface (R GUI, RStudio, IDE, etc.), open a new R script

Application	Menu item	Keyboard shortcut (mac shortcut)
R GUI	`File > New Document`	`CTL+N (CMD+N)`
RStudio	`File > New File > R Script` 	`Shift+CTL+N (Shift+CMD+N)`

- Save it with a name like “`my_first_script.R`”
  - You can save it in the same location as the slides for this workshop (i.e., the folder *containing* the ‘`data`’ folder.)

Add some code to your script

- Paste in the following code to your script file:

```r
CSV_path <- file.path("data", "data_example.csv")
file.exists(CSV_path)

DF <- read.csv(CSV_path, skip = 2, encoding = "UTF-8")

colnames(DF)
plot(DF)
```

- and save it.

Run R code in scripts

Most IDEs have a shortcut to send portions of R code (a line or *statement* spanning multiple lines) to an R session:

- R GUI: `CTL+Return` (mac: `CMD+Return`)
- RStudio: `CTL+Return` (mac: `CMD+Return`)

You can run *all* the code in a script in different ways:

- The `source()` function, with a path to the script file as an argument
  - The code will run in the current session.
  - `?source`
  - `source("my_script.R")`
- Run R in "*batch mode*"
  - "*batch mode*" is **not** interactive (no prompt)
  - It is usually invoked from a terminal or other command-line (outside an R GUI)
  - The code in the script will run in a new session
  - You can capture output in a separate file
  - `?BATCH`

Comments

- The '#' character denotes a *comment* in R
  - *Everything on a line* after a comment character is ignored by R
  - There are no 'multi-line' comments in R

```
print("this is R code")  # this is a comment
```

- You can make an entire line a comment by putting a comment character at the beginning.
  - Divide your code into *sections*
    Shift+CTL+R (Shift+CMD+R) in RStudio

```
# SECTION ------------------------------------------------
```

  - Create 'comment headers' for your scripts:

```
##############################################################
### Title
### Project or description
### Author Name           R vX.X.X              YYYY-MM-DD
##############################################################
```

Comments in code

- You can put a comment beside a line of code (even in the middle of a mult-line statement)
  - R will ignore the rest of the line, and continue reading code on the next line

```r
DF <-         # short for "data frame"
  read.csv(   # read a csv file
    CSV_path,  # path to file
    skip = 2  # skip lines at top of file (not data)
  )
```

- Use comments to
  - organize your code (divide it into sections)
  - explain the code, where relevant
  - "comment-out" code temporarily, to stop it from running without deleting it (useful for debugging).

    `Shift+CTL+C` (`Shift+CMD+C`) comment a line in RStudio

Activity 2: add some commands to your new script

- Take a few minutes to add some commands you used in the last activity to your new script
    - **Tip:** use the up arrow to go back through your *command history*
- Be sure to add comments to explain your commands, or what you found with them.

Housekeeping

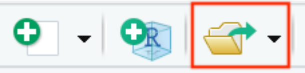
- This command removes all objects from your workspace (active session)
    - aka “clear memory” or “empty your environment”

```
rm(list = ls())
```

- Many R users find it useful to include this line near the top of their scripts, to ensure a “clean workspace” for the rest of the script.
- This can avoid problems with objects from different scripts that have the same name, or other unexpected behaviours from extra objects in your workspace.

Open a script

- All the code shown in the slides for this workshop has been collected in a script file: “`R_data_scripting.R`” (in the ‘`source`’ folder)
- Open it to follow along for the rest of the workshop.

Application	Menu item	Keyboard shortcut (mac shortcut)
R GUI	`File > Open Document...`	`CTL+O` (`CMD+O`)
RStudio	`File > Open File...`	`CTL+O` (`CMD+O`)
		

Set the Working Directory to source file location in **RStudio**

- Menu item: "`Session > Set Working Directory > To Source File Location`"
- This makes it easy to use *relative paths* in your script, relative to the location of the script file itself.

For this workshop

- All the code in this document assumes that the working directory is the *same directory* as where the script file is.

Section 7

Control how data is read: `read_csv()` options

Encoding non-English characters

- If you are running R in Windows, you may notice that some text values (character) look strange:

  “**QuÃ©bec**” or ““**Qu\xe9bec**” instead of “Québec”

- There is nothing wrong with the file — this indicates a *mismatch* between the *encoding* used to write the file, and what `R` used to read it.

- Even though ‘`.csv`’ files are plain text, letters (especially non-english characters) can be *encoded* in different ways to represent them in the computer.

- “`UTF-8`” is a character encoding standard designed to handle many non-english characters.
  - The example data file was written in “`UTF-8`”
  - Most OSes and many programs use “`UTF-8`” encoding by default.
  - But *Windows* uses “`latin1`” by default, and so does `R` ($<$ 4.2.0) when running in Windows.
  - Starting with v4.2.0, `R` uses “`UTF-8`” as the default encoding on Windows.

Read a csv file with a mismatched encoding

```
CSV_latin1 <- read_csv("data/data_example_latin1.csv")
CSV_latin1
```

Read a csv file with a mismatched encoding

```r
CSV_latin1 <- read_csv("data/data_example_latin1.csv")
CSV_latin1
```

```
# # A tibble: 6 x 10
#   Type      Treatment PlantNum  `95` `175` `250` `350` `500`
#   <chr>     <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
# 1 "Quebec"  nonchill~        1  16    30.4  34.8  37.2 35.3
# 2 "Quebec"  <NA>             2  13.6  27.3  37.1  41.8 40.6
# 3 "Quebec"  <NA>             3  16.2  32.4  40.3  42.1 42.9
# 4 "Qu\xe9b~ chilled          1  14.2  24.1  30.3  34.6 32.5~
# 5 "Qu\xe9b~ <NA>             2   9.3  27.3  35    38.8 38.6
# 6 "Qu\xe9b~ <NA>             3  15.1  21    38.1  34   +38.9
# # i 2 more variables: `675` <dbl>, `1000` <dbl>
```

read_csv(): read a file with a different encoding

- You can specify the encoding used in the file with the 'locale' argument of read_csv()

```
read_csv(CSV_path, skip = 2,
         locale = locale(encoding = "latin1")
)
```

- You might need this to read a file that was created on a Windows computer and encoded in "latin1"

Microsoft Excel on Windows

Excel in Windows is capable of saving files in .csv format, but may use "latin1" encoding by default. If it saves a .csv file using "UTF-8" encoding, it may add extra contents to the file (a "BOM") that makes it difficult to read with read.csv() (base R).
See the "Extras" document for this workshop for details.
The readr package is not affected by the BOM. :)

read_csv(): column names

- The default for `read_csv()` is to ensure all column names are *unique*, but not necessarily *syntactically valid*
- You can still refer to columns with syntactically ‘invalid’ names:
  - use functions that allow names as characters,
  - quote names with backticks (`` ` ``)

```
CSV[, "95"]  # still a `data.frame` (with 1 column)
CSV[["95"]]  # vector
CSV$`95`     # quoted name
```

read_csv(): treat column names

- Change how read_csv() treats column names with the 'name_repair' argument
  - "universal" makes names unique and syntactic

```
read_csv(
  CSV_path, skip = 2,
  name_repair = "universal" # make names unique and syntactic
)
```

- You can also pass a function by name (make.names is the same behaviour as read.csv() in base R)

```
read_csv(
  CSV_path, skip = 2,
  name_repair = make.names  # function: same as read.csv()
)
```

read_csv(): guessing column types

- By default, **read_csv**() prints a message summarizing what it did, including *guessing the data type* of each column.
  - *.csv files do not include this information as metadata*
- Control how columns are guessed with the guess_max argument:

```
# use the first 2 rows to guess column types (less successful,
  read_csv(CSV_path, skip = 2, guess_max = 2)
```

```
# Warning: One or more parsing issues, call `problems()` on your da
# frame for details, e.g.:
#   dat <- vroom(...)
#   problems(dat)
```

```
# use *all* rows to guess column types
# - slow: has to read *every row* twice.
  read_csv(CSV_path, skip = 2, guess_max = Inf)
```

read_csv(): specify column types

- If you know what the column types are (or should be), you can tell `read_csv()` what they are with the `col_types` argument.
  - for large datasets, this can be faster: read rows once
  - avoids bad guesses.

```r
## Specify column types with a compact string
read_csv(CSV_path, skip = 2, col_types = "cccdddddd")

## Or use a `column specification`
# extract specification from tibble
col_spec <- spec(CSV)
# change a column to numeric (double)
col_spec$cols[["500"]] <- col_double()
read_csv(CSV_path, skip = 2, col_types = col_spec)

# ?read_csv for more options
```

read_csv(): all columns as strings

- In extreme cases, you can read everything as ‘character’, then clean and coerce to other data types within R

```
# read all columns as character
read_csv(CSV_path, skip = 2,
         col_types = cols(.default = col_character())
         )
```

```
# # A tibble: 13 x 10
#    Type     Treatment PlantNum `95`  `175` `250` `350` `500`
#    <chr>    <chr>     <chr>    <chr> <chr> <chr> <chr> <chr>
#  1 Quebec   nonchill~ 1        16    30.4  34.8  37.2  35.3
#  2 Quebec   <NA>      2        13.6  27.3  37.1  41.8  40.6
#  3 Quebec   <NA>      3        16.2  32.4  40.3  42.1  42.9
#  4 Québec   chilled   1        14.2  24.1  30.3  34.6  32.5~
#  5 Québec   <NA>      2        9.3   27.3  35    38.8  38.6
#  6 Québec   <NA>      3        15.1  21    38.1  34    +38.9
#  7 Mississ~ nonchill~ 1        10.6  19.2  26.2  30    30.9
#  8 Mississ~ <NA>      2        12    22    30.6  31.8  32.4
#  9 Mississ~ <NA>      3        11.3  19.4  25.8  27.9  28.5
```

read_csv(): missing values

- Use the na argument to supply a list of values to replace with NA.
  - This is applied to **all** columns.

```r
read_csv(CSV_path, skip = 2,
         na = c(".", "NA")  # will not replace empty strings
         )
```

```
# # A tibble: 13 x 10
#    Type     Treatment PlantNum  `95` `175` `250` `350` `500`
#    <chr>    <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
#  1 Quebec   "nonchil~        1  16    30.4  34.8  37.2 35.3
#  2 Quebec   ""               2  13.6  27.3  37.1  41.8 40.6
#  3 Quebec   ""               3  16.2  32.4  40.3  42.1 42.9
#  4 Québec   "chilled"        1  14.2  24.1  30.3  34.6 32.5~
#  5 Québec   ""               2   9.3  27.3  35    38.8 38.6
#  6 Québec   ""               3  15.1  21    38.1  34   +38.9
#  7 Mississ~ "nonchil~        1  10.6  19.2  26.2  30   30.9
#  8 Mississ~ ""               2  12    22    30.6  31.8 32.4
#  9 Mississ~ ""               3  11.3  19.4  25.8  27.9 28.5
# 10 Mississ~ "chilled"        1  10.5  14.9  18.1  18.9 19.5
```

read_csv(): number formats

- read_csv() will attempt to read numbers with "." as the decimal, and "," between large numbers.
- But our example data includes a *mix* of these.
  - This makes it difficult to choose one approach for everything, but one option in this case is to set the 'grouping' mark to an empty string

```r
CSV_comma <- read_csv(CSV_path, skip = 2,
         locale = locale(grouping_mark = "")
)
CSV_comma[["675"]]
```

```
#  [1] "39.2" "41.4" "43.9" "35,4" "37,5" "39,6" "32.4" "31.1"
#  [9] "28.1" "22.2" "13.7" "13.7" "18.9"
```

- With this option, column '675' is now *character* instead of *numeric*, With the commas intact.
- See also the 'decimal_mark' argument of locale() (?locale)

Final read of data for workshop

- For the rest of this workshop, we will work with this version of the example data file, read with `read_csv()`:

```
DF <- read_csv(
  CSV_path, skip = 2,
  name_repair = make.names
)
```

- We use “make.names” to make *syntactically valid* names that are easier to work with later.

The `readxl` package

Provides functions for reading from (but not writing to) Microsoft Excel files (`.xls` and `.xlsx`)

```r
library(readxl)    # load the package
## Documentation: ?read_excel  help(package="readxl")
## use an example included in the package
xl_path <- readxl_example("datasets.xlsx")
excel_sheets(xl_path)  # get the names of the sheets
```

```
# [1] "iris"     "mtcars"   "chickwts" "quakes"
```

```r
## read a specified sheet from the Excel file
iris_xl <- read_excel(xl_path, "iris")
```

Activity 3: read a messy Excel file

- `read_excel()` has many of the same arguments as `read_csv()` to control how data is imported.
- Use the script file in the "`examples`" folder as a starting point:
  - "`R2_activity_3.R`"

```
XL_path <- readxl_example("deaths.xlsx")
XL <- read_excel(XL_path, ...)
```

Activity 3: read a messy Excel file

- `read_excel()` has many of the same arguments as `read_csv()` to control how data is imported.
- Use the script file in the "`examples`" folder as a starting point:
  - "`R2_activity_3.R`"

```
XL_path <- readxl_example("deaths.xlsx")
XL <- read_excel(XL_path, ...)
```

Tip

Use the '`range`' argument to read data from a specific range in a sheet, ignoring contents outside this range (rows above & below, columns before & after)

Section 8

Manipulate Data: The dplyr Package

`dplyr`: a grammar of data manipulation

- `dplyr` provides many functions that follow a coherent framework or "*grammar*"
- They are intended to help you focus on *what* you want to do, and translate your thoughts into code.
- High-level functions have active names and called "**verbs**" — they describe what they do.
- `dplyr` and `tidyselect` provide many "**helper functions**" that work *inside* verbs and other `tidyverse` functions to make common tasks easier to translate into code.
    - These functions may not work on their own, outside of `dplyr` verbs and `tidyr` functions (see `?"faq-selection-context"`).

dplyr verbs

Verbs can be grouped based on the component of the dataset that they work with[2]:

- Columns:
  - `select()` changes whether or not a column is included.
  - `relocate()` changes the order of the columns.
  - `rename()` changes the name of columns.
  - `mutate()` changes the *values* of columns and creates new columns.
- Rows:
  - `filter()` chooses rows based on column values.
  - `slice()` chooses rows based on location.
  - `arrange()` changes the order of the rows.
- Groups of rows:
  - `group_by()` defines groups of rows.
  - `summarise()` collapses a group into a single row.

[2] https://dplyr.tidyverse.org/articles/dplyr.html#single-table-verbs

Load dplyr

- When you load dplyr, you may see a message saying “objects are masked”:

```r
library(dplyr)
```

```
#
# Attaching package: 'dplyr'

# The following objects are masked from 'package:stats':
#
#     filter, lag

# The following objects are masked from 'package:base':
#
#     intersect, setdiff, setequal, union
```

- Some functions in dplyr have the same name as functions in other packages! (base, stats)

`package::object` notation

- When an "object is masked", it means the new one (`dplyr` function) will take precedence over the one being "masked" when a function of that name is called.

```r
?filter  # more than one result!
```

- To avoid confusion, you can specify which package you mean with the "`package::object`" notation (*functions* are a class of *object*):

```r
?filter  # more than one result
?dplyr::filter
?stats::filter
```

- Because `dplyr` has *masked* the names from the other packages, we can simply call these functions normally (e.g., `filter()`), but if we wanted to use a masked function, we would have to use the special notation: `stats::filter()`
  - We won't be using the masked functions today, however.

Section 9

Explore With `dplyr`: `select()` & `filter()` Parts of a Data Frame

Columns: `select()`

- Select (and rename) columns in a data frame, using a concise mini-language (`<tidy-select>`)
- Select by name: 'bare names', like regular variables; or character names

```
select(DF, Type, Treatment, PlantNum)
select(DF, "Type", "Treatment", "X95")
select(DF, Type:PlantNum)
```

- Select by position:

```
select(DF, 2:5)
```

- or both, while changing names and order along the way:

```
select(DF, c(Type, 4:6, Plant = PlantNum))
```

Columns: selection helpers

- Various **helper functions** (*selection helpers*) add ways to `select()` columns *based on criteria* (dynamically).
- Some examples (see `?select` and `?dplyr_tidy_select` for more):

```r
select(DF, starts_with("X"))
select(DF, !starts_with("X"))
select(DF, contains("m"))

select(DF, where(is.character) & starts_with("X"))

select(DF, any_of(c("Type", "Treatment", "Plant", "95")))
```

Columns: selection helpers

- Various **helper functions** (*selection helpers*) add ways to `select()` columns *based on criteria* (dynamically).

- Some examples (see `?select` and `?dplyr_tidy_select` for more):

```
select(DF, starts_with("X"))
select(DF, !starts_with("X"))
select(DF, contains("m"))

select(DF, where(is.character) & starts_with("X"))

select(DF, any_of(c("Type", "Treatment", "Plant", "95")))
```

- `starts_with()` & `contains()` match *patterns* in names
- `where()` applies a function (`is.character()`, in this case) to each column: those where the function returns `TRUE` are kept.
- `any_of()` matches names in a character vector; `all_of()` is similar, but names that don't exist cause an *error*.

Rows: filter()

- filter() retains rows that satisfy *all* the conditions specified
    - expressions must return a vector of *logical* values (TRUE or FALSE)

```r
filter(DF, X95 < 10)
filter(CO2, conc == 95, uptake < 10)  # "AND"
filter(CO2, conc == 95 | uptake < 10) # | == "OR" operator

filter(DF, Treatment != "")
filter(DF, !Type %in% c("Quebec", "Mississippi"))
filter(DF, !Type %in% unique(CO2$Type))

filter(DF, X175 > mean(X175))
```

select() & filter() work well together

- For example, you can filter on a column, then remove it with select()

```
select( filter(DF, Treatment == "chilled"),
        where(is.numeric)
)
```

```
# # A tibble: 2 x 7
#   PlantNum   X95  X175  X250  X350  X675 X1000
#      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1        1  14.2  24.1  30.3  34.6 354    38.7
# 2        1  10.5  14.9  18.1  18.9  22.2  21.9
```

- But this is clunky! This is easier to follow:

```
DF %>% filter(Treatment == "chilled") %>%
  select(where(is.numeric))
```

```
# # A tibble: 2 x 7
#   PlantNum   X95  X175  X250  X350  X675 X1000
```

A ‘pipe’ operator

- `%>%` allows you to pass results from an expression on the left-hand side (LHS) as an argument (usually the first) to a *function call* on the right-hand side (RHS).
    - you can read a pipe operator as “then” (`a %>% b()` = “*a then b*”)

This expression …	is equivalent to:
`x %>% f()`	`f(x)`
`x %>% f(y)`	`f(x, y)`
`x %>% f(y, z = .)`	`f(y, z = x)`
`x %>% f %>% g %>% h`	`h(g(f(x)))`

- This can make code easier to read, as expressions are written and evaluated from *left to right*, rather than from *inside to outside* nested parentheses.

magrittr's 'forward-pipe' operator

Figure 1: "La Trahison des Images" ("The Treachery of Images") or "Ceci n'est pas une pipe" ("This is not a pipe") by René Magritte.

%>%
Ceci n'est pas un pipe.
www.tidyverse.org

- The magrittr package (included with tidyverse) provides a "forward-pipe operator":

```
%>%     # ?magrittr::`%>%`
```

- The magrittr package is automatically loaded when loading most tidyverse packages (e.g., tidyr, dplyr, ggplot2).
    - These packages were designed to work with this operator, and use it themselves.
    - It is often unnecessary to load magrittr separately, unless you are **not** using these other packages.

R now has a 'native' pipe operator

- A pipe operator was introduced in base R in v4.1 (May 2021)[3]:

```
|>      # ?pipeOp
```

- It was inspired by the "forward pipe operator" introduced by `magrittr`, but is more streamlined. See these links for details:
  - Differences between the base R and magrittr pipes
  - "Understanding the native R pipe |>"
- Because '`|>`' is new, many examples online still use `magrittr`'s '`%>%`'.
- But '`|>`' is always available *in R >= v4.1*, without having to load additional packages.
- This document will use '`%>%`' in the examples, for consistency and because many `tidyverse` functions were designed to work with it.

[3] https://cran.r-project.org/bin/windows/base/old/4.1.0/NEWS.R-4.1.0.html

Activity : pipes

Insert a pipe in R Studio with `CTL+Shift+M` (`CMD+Shift+M`)

Re-write the expressions using pipes:

```r
# (1)
sum(1:10)

# (2)
filter(CO2, conc < 100)

# (3)
filter(
  select( DF,
    where(is.character)
  ),
  Treatment == "")

# (4)
gsub("X", "", names(DF))
```

Activity : pipes

Insert a pipe in R Studio with `CTL+Shift+M` (`CMD+Shift+M`)

Re-write the expressions using pipes:

```
# (1)
sum(1:10)

# (2)
filter(CO2, conc < 100)

# (3)
filter(
  select( DF,
    where(is.character)
  ),
  Treatment == "")

# (4)
gsub("X", "", names(DF))
```

```
# (1)
1:10 %>% sum()

# (2)
CO2 %>% filter(conc < 100)

# (3)
DF %>%
  select(
    where(is.character)
  ) %>%
  filter(Treatment == "")

# (4)
DF %>% names() %>%
gsub("X", "", .)
```

dplyr conventions

All dplyr verbs (and other related tidyverse functions) share a few things in common:

- The first argument is a data frame (or tibble).
- Other arguments describe what to do with the data frame.
- You can refer to columns in the data frame directly without using $.
- The result is a new data frame.

These features work well with the pipe operator, and can help build clear and efficient workflows.

Because all the functions have these in common, it makes it easier to extend your understanding to new functions.

Section 10

Find specific values with `dplyr`

Find unexpected data types in example data

- Character columns where numeric is expected

```r
# specify columns by pattern
# (in this case, all numeric columns start with "X")
DF %>%
  select(where(is.character) & starts_with("X")) %>% names()
```

```
# [1] "X500"
```

```r
# specify a range of known columns
DF %>%
  select(where(is.character) & X95:X1000) %>% names()
```

```
# [1] "X500"
```

Find non-numeric values in a character column

- as.numeric() replaces non-numeric values with NA, and produces a *warning*.
- suppressWarnings() suppresses these expected warnings
- is.na() checks for the NAs produced.
  - !is.na(X500) prevents existing NAs from being included.

```
DF_nonum <-
  DF %>% select(1:3, X500) %>%
  filter(!is.na(X500) &       # exclude existing NAs
           X500 %>%
           as.numeric() %>%   # non-numeric -> NA + warning
           is.na() %>%        # check for new NAs
           suppressWarnings() # suppress expected warnings
  )
DF_nonum
```

```
# # A tibble: 1 x 4
#   Type   Treatment PlantNum X500
#   <chr>  <chr>        <dbl> <chr>
```

Find numeric values outside expected range

```
DF_extreme_nums <-
  DF %>% select(1:3, X675) %>%
  filter(X675 > 100)
DF_extreme_nums
```

```
# # A tibble: 3 x 4
#   Type   Treatment PlantNum  X675
#   <chr>  <chr>        <dbl> <dbl>
# 1 Québec chilled          1   354
# 2 Québec <NA>             2   375
# 3 Québec <NA>             3   396
```

- These values had a comma (,) instead of a period (.) for the decimal.
- They are 10× what they should be.

Extras

See the “Extras” for examples of using a custom function that can be applied to multiple columns at once.

Section 11

dplyr Verbs: Modify Data Columns

Modify Columns: `mutate()`

- `mutate()` creates new columns, or modifies existing ones, as functions of existing columns.
  - it is a `dplyr` workhorse, used for many tasks, since it allows you to modify values *systematically*.

Add columns:

```
DF %>% mutate(
  Trt_n = Treatment %>% nchar(),
  Xsum = X95 + X175
  )
```

Modify columns:

```
DF %>% mutate(X95 = X95 / mean(X95))
```

mutate() helper functions

- refer to values in the previous / next row with lag() and lead

```
DF %>% mutate(
  Plant_lag  =  lag(PlantNum),
  Plant_lead = lead(PlantNum, n=2)
)
```

- refer to row numbers with row_number()

```
DF %>% mutate(
  Plant_row = PlantNum < row_number()
)
```

mutate(): conditional values

- case_when() lets you apply multiple if/else statements to a vector of values.
  - conditions go first
  - the value returned if the condition is true goes after a tilde (~)
  - multiple statements are separated by commas
  - If no default is specified (.default =), anything that does not match any condition is replaced with NA

```r
DF %>% mutate(
  Type_ab = case_when(
    Type == "Quebec"      ~ "QC",
    Type == "Mississippi" ~ "MS",
    .default = as.character(Type)
  )
)
```

dplyr *semantics*

dplyr verbs and helper functions let you refer to column names of the data frame directly in their arguments as regular variables — without having to quote them as strings. But these names have different meanings (semantics) in different verbs.

- **“select semantics”** (`<tidy-select>`): in `select()` and similar functions, a column name refers to its *position* in the data frame.
    - you can refer to a column as a quoted string in `select()`, and it is interpreted as a reference to the column.
- **“mutate semantics”** (`<data-masking>`): in `mutate()` and similar functions (`group_by()`, `summarise()`, `filter()`, etc.), a column name refers to a *vector of values*.
    - you cannot supply a column name as a string in `mutate()`, because it is treated as a vector of length 1, rather than a reference to a column of values.
- Helper functions only work in one context or the other, so knowing the difference will tell you which helper functions to use when.

Activity

Section 12

Clean Some Data With the `stringr` Package

The stringr package

- To clean character columns, we will use functions from the `stringr` library, which provides many useful functions for working with, and manipulating *strings*.

```
library(stringr)
# help(package="stringr")
# vignette("stringr")
```

Convert character column to numeric (replace characters)

- The 675 column had some values where a comma (,) was used instead of a period (.) for the decimal.
- If we loaded the data without using “,” as a grouping mark, the 675 column would be character.

```
# [1] "35,4" "37,5" "39,6"
```

- Replace ‘,’ with ‘.’ in 675 column, and convert to numeric:

```
CSV_comma %>% select(1:3, "675") %>%
  mutate(
    # Replace "," with "."
    `675` = str_replace(`675`, ",", "."),
    # convert to numeric
    `675` = as.numeric(`675`)
  )
```

Cleaning some columns in the example data

- Normalize values of the ‘Type’ column:
    - using `if_else()`

```
DF_clean1_type <- DF %>%
  mutate(
    Type = if_else(Type == "Mississippi", Type, "Quebec")
  )
```

Cleaning some columns in the example data

- Normalize values of the ‘Type’ column:
  - using `if_else()`

```
DF_clean1_type <- DF %>%
  mutate(
    Type = if_else(Type == "Mississippi", Type, "Quebec")
  )
```

- could also use `case_when()`

```
DF %>%
  mutate(
    Type = case_when(
      Type == "Québec"  ~ "Quebec",
      .default = Type
    )
  )
```

Convert character column to numeric (drop characters)

- The `X500` column had a value with extra text: "32.5 (umol/m^2 sec)"
- Remove non-numeric characters from `X500` column, and convert to numeric:
  - We can use `word()` from the `stringr` package to extract the first 'word' (separated by spaces ' ') in each row.
  - i.e., drop all text (in each row) after the first space

```
DF_clean2_500 <- DF_clean1_type %>%
  mutate(
    # drop everything after the first space:
    X500 = word(X500, 1),
    # convert to numeric
    X500 = as.numeric(X500)
  )
```

- Other `stringr` functions would also work, such as `str_split_i()`

Modify extreme values

- X675 column had some large values where a comma (,) was used instead of a period (.) for the decimal, resulting in values 10× what they should be.

Type	Treatment	PlantNum	X675
Québec	chilled	1	354
Québec	NA	2	375
Québec	NA	3	396

- Divide large values (> 100) in the X675 column by 10:

```
DF_clean3_675 <- DF_clean2_500 %>%
  mutate(
    X675 = if_else(X675 > 100, X675 / 10, X675)
  )
```

Clean 'Treatment' column

- The 'Treatment' column has some empty values

```
DF %>% distinct(Treatment) %>% pull()
```

```
# [1] "nonchilled" NA           "chilled"
```

- A value is only present when it changes
- We can use the fill() function from the tidyr package
    - without loading the package, by specifying the function with package::object notation

```
DF_clean4_trt <- DF_clean3_675 %>%
  # replace empty strings with NA (if present)
  mutate(Treatment = na_if(Treatment, "")) %>%
  # Fill Down to replace NAs (tidyr)
  tidyr::fill(Treatment, .direction = "down")
```

Activity

Section 13

dplyr Verbs: Grouped Data

Define groups: group_by()

- Group rows based on combinations of column values

```
DF %>% group_by(Type, PlantNum)  # no visible change
```

- dplyr verbs are applied to each group of rows

```
DF %>% group_by(Type, PlantNum) %>%
  filter(row_number() == 1)
DF %>% group_by(Type, PlantNum) %>%
  arrange(Type, PlantNum) %>%
  mutate(Norm95 = X95 / mean(X95))
```

- Grouping columns are excluded from the operations

```
DF %>% group_by(Type, PlantNum) %>%
  select(starts_with("X"))
```

Collapse groups: `summarise()` / `summarize()`

- Without groups specified, `summarise()` treats the data frame as a single group

```
DF_clean4_trt %>% summarise(n(), mean(X95))
```

```
# # A tibble: 1 x 2
#   `n()` `mean(X95)`
#   <int>       <dbl>
# 1    13        11.9
```

- But `summarise()` is most useful when applied to grouped data

```
DF_clean4_trt %>% group_by(Type, PlantNum) %>%
  summarise(n = n(), sum(X95))
```

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

- `summarise()` automatically drops the last level from the groups.
- `summarise()` & `summarise()` are synonyms (same function).

Activity

Locate duplicate rows in example data

- We can combine dplyr verbs, like `summarize` and `filter` to quickly locate issues

```
DF_clean4_trt %>%
  group_by(Type, Treatment, PlantNum) %>%
  summarise(n = n(), .groups = "drop") %>%
  filter(n > 1)
```

```
# # A tibble: 1 x 4
#   Type        Treatment PlantNum     n
#   <chr>       <chr>        <dbl> <int>
# 1 Mississippi chilled          2     2
```

Inspect duplicate rows in example data

- Luckily in this case, the missing values in the duplicate rows look like they are *not* missing in the other row

```
DF_duprows <- DF_clean4_trt %>%
  group_by(Type, Treatment, PlantNum) %>%
  filter(n() > 1)
DF_duprows
```

```
# # A tibble: 2 x 10
# # Groups:   Type, Treatment, PlantNum [1]
#   Type      Treatment PlantNum   X95  X175  X250  X350  X500
#   <chr>     <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 Mississi~ chilled          2   7.7  11.4  NA      13  12.5
# 2 Mississi~ chilled          2   7.7  11.4  12.3    13  12.5
# # i 2 more variables: X675 <dbl>, X1000 <dbl>
```

Combine duplicate rows in example data

- We could collapse the duplicate rows using `summarise()`,
  - but it would be tedious to write out the statements for each column
  - or figure out a more advanced method (see "Extras" for ideas).
- It will be much easier to remove duplicates after the data is *tidy*

Section 14

Reshape Data: The `tidyr` Package

Tidy data

> *"Happy families are all alike; every unhappy family is unhappy in its own way."*
> *— Leo Tolstoy*

> *"Tidy datasets are all alike but every messy dataset is messy in its own way."*
> *— Hadley Wickham (doi: 10.18637/jss.v059.i10)*

- Tidy datasets provide a standardized way to link the *structure* of a dataset (its physical layout) with its *semantics* (its meaning).
- A tidy dataset follows three interrelated rules:
  - Each variable must have its own column.
  - Each observation must have its own row.
  - Each value must have its own cell.

The tidyr package

```
library(tidyr)
```

The tidyr package provides tools for making data *tidy*:

- "**Pivoting**" converts between *long* and *wide* forms.
  - pivot_longer() and pivot_wider()
- "Rectangling" turns deeply nested lists (e.g., JSON) into tidy tibbles.
  - *Not covered here: see the vignette.*
- "Nesting" converts grouped data into nested data frames.
  - *Not covered here: see the vignette.*
- **Splitting** and **combining** character columns.
  - separate_...() a single character column into multiple columns
  - unite() multiple columns into a single character column.
- Handling **missing values**
  - make implicit missing values explicit with complete()
  - make explicit missing values implicit with drop_na()
  - replace missing values with a known value using replace_na() or fill() them with the next/previous value

Re-create the ‘Plant’ column

- Our example data is still missing unique values of ‘Plant’, made up of the first character of the ‘Type’ and ‘Treatment’ columns, and a unique number.

```
CO2 %>% pull(Plant) %>% unique() %>% as.character()
```

```
#  [1] "Qn1" "Qn2" "Qn3" "Qc1" "Qc2" "Qc3" "Mn1" "Mn2" "Mn3"
# [10] "Mc1" "Mc2" "Mc3"
```

- We can create temporary columns with mutate()

```
DF_clean4_trt %>% select(Type, Treatment) %>%
  ## Create columns with the first letter of each row
  mutate(
    Type.tmp      = str_sub(Type, 1, 1),
    Treatment.tmp = str_sub(Treatment, 1, 1),
  )
```

Combine columns

- We can create temporary columns with `mutate()`, then combine them with `unite()`

```r
DF_clean5_cols <- DF_clean4_trt %>%
  ## Create columns with the first letter of each row
  mutate(
    Type.tmp       = str_sub(Type, 1, 1),
    Treatment.tmp = str_sub(Treatment, 1, 1)
  ) %>%
  ## Combine columns, and remove them
  unite(
    Plant,                             # new column name
    Type.tmp, Treatment.tmp, PlantNum, # input columns
    sep = ""      # characters to separate each input
  ) %>%
  ## Move columns to the front (left)
  relocate(Plant, Type, Treatment)
```

Long vs wide

- Our data frame is looking clean!
  But it still has one problem: it's in 'wide' format

```
DF_clean5_cols %>% select(where(is.numeric)) %>% names()
```

```
# [1] "X95"   "X175"  "X250"  "X350"  "X500"  "X675"  "X1000"
```

- All the numeric columns are values of a hidden variable, '`uptake`'
- The names of these columns are actually *values* of another hidden variable, '`conc`' (concentration)
- This is **not** 'tidy'
  - most plotting functions will not expect column names as values
  - analysis is more complicated: relationships are between the *column names* and *values*

Pivot

- We can make this data 'tidy' by *pivoting* to a longer structure
    - then convert the former column names to numeric values.

```
DF_tidy <- DF_clean5_cols %>%
  pivot_longer(
    cols = where(is.numeric),  # columns to pivot
    names_to = "conc",         # name of new column with old c
    values_to = "uptake"       # name of new column with old v
  ) %>%
  ## Clean former column names and convert to numeric
  mutate(
    conc = str_replace(conc, "X", "") %>% as.numeric()
  )
```

Check duplicates rows

- Duplicate rows are now just duplicate values of ‘uptake’

```
DF_tidy %>%
  group_by(Plant, Type, Treatment, conc) %>%
  summarise(n = n()) %>%
  filter(n > 1)
```

```
# # A tibble: 7 x 5
# # Groups:   Plant, Type, Treatment [1]
#   Plant Type        Treatment  conc     n
#   <chr> <chr>       <chr>     <dbl> <int>
# 1 Mc2   Mississippi chilled      95     2
# 2 Mc2   Mississippi chilled     175     2
# 3 Mc2   Mississippi chilled     250     2
# 4 Mc2   Mississippi chilled     350     2
# 5 Mc2   Mississippi chilled     500     2
# 6 Mc2   Mississippi chilled     675     2
# 7 Mc2   Mississippi chilled    1000     2
```

Check duplicate values

- We can confirm *systematically* that all duplicate values are equivalent (i.e., `min() == max()`)
  - `near()` checks for equality, with a tolerance for floating-point values.

```
DF_tidy %>%
  group_by(Plant, conc) %>%
  summarise(
    n = n(),
    min_max = min(uptake, na.rm = TRUE) %>%
              near(max(uptake, na.rm = TRUE)),
    .groups = "drop"
  ) %>%
  filter(!min_max) %>% nrow()
```

```
# [1] 0
```

Combine duplicate rows in example data

- Now we only have to group by other columns, and `summarise()` the `uptake` column
  - and sort (with `arrange()`) to recover original order.

```
DF_clean <-
  DF_tidy %>%
  group_by(Plant, Type, Treatment, conc) %>%
  summarise(
    uptake = max(uptake, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(desc(Type), desc(Treatment), Plant, conc)
```

Check results

- Did we manage to re-construct the original data set?

```
all.equal(DF_clean, CO2, check.attributes = FALSE)
```

```
# [1] "Component \"Plant\": target is character, current is ordered"
# [2] "Component \"Type\": target is character, current is factor"
# [3] "Component \"Treatment\": target is character, current is fac
```

- Not quite — our character columns are not '*factors*', as in the original.
  - *factors* are a special kind of vector for categorical data
  - See `?factor`, and the `forcats` package for more information

Final steps

- Convert character columns to *factors*

```
DF_final <- DF_clean %>%
  mutate(
    Plant     = factor(Plant, levels = unique(Plant)),
    Type      = factor(Type,  levels = unique(Type)),
    Treatment = factor(Treatment, levels = unique(Treatment))
  )
```

Final steps

- Convert character columns to *factors*

```r
DF_final <- DF_clean %>%
  mutate(
    Plant     = factor(Plant, levels = unique(Plant)),
    Type      = factor(Type,  levels = unique(Type)),
    Treatment = factor(Treatment, levels = unique(Treatment))
  )
```

- Convert *all* character columns to *factors*

```r
DF_clean %>%
  mutate( across(where(is.character), factor) )
```

- Did we manage to re-construct the original data set?

```r
all.equal(DF_final, CO2, check.attributes = FALSE)
```

```
# [1] TRUE
```

Order of operations: clean, tidy

- Clean then tidy; or tidy then clean?
- It depends!
  - In this example, we cleaned columns before pivoting to combine them. This allowed us to correct *different* issues in each column, and convert them to a common type before combining.
  - If the same issue is present in multiple columns, it often makes sense to pivot first, then you have fewer columns to clean
  - In other cases, you may want to pivot *wider* to separate different groups of values, so that they can be cleaned differently.
- In many cases, you might switch between the two more than once.
  - e.g., clean → tidy → clean ...

Exercise

Section 15

Save Data Outside R

The `readr` package: writing data

readr		**base R**
`write_csv()`	← comma separated values	`write.csv()`
`write_csv2()`	← allows ‘;’ as delimiter and ‘,’ for decimals (depending on locale)	`write.csv2()` ‘,’ for decimals, ‘;’ as separator
`write_tsv()`	← tab separated values	
`write_delim()`	← (generic) files with an arbitrary delimiter	`write.table()`
`write_excel_csv()`, `write_excel_cs2v()`	← include a UTF-8 Byte order mark, which indicates to Excel the csv is UTF-8 encoded	

Save our work

- Save a data frame to a `.csv` file:
  - it will be encoded with `UTF-8` by default (on all platforms)

```
write_csv(DF_final, "data/data_clean.csv")
write_excel_csv(DF_final, "data/data_excel.csv")
```

Save our work

- Save a data frame to a .csv file:
    - it will be encoded with UTF-8 by default (on all platforms)

```
write_csv(DF_final, "data/data_clean.csv")
write_excel_csv(DF_final, "data/data_excel.csv")
```

- Read it back in to check

```
save_test <- read_csv("data/data_clean.csv")
head(save_test)
```

Section 16

Combining Data Frames

bind_rows(): stack data vertically

- Combine data with the same columns, stacking rows together
  - Columns are matched by name

```
DF_bindr <- bind_rows(DF_duprows, DF_clean5_cols)
head(DF_bindr)

# can also use it on a _list of data frames_
bind_rows(list(DF_duprows, DF_duprows))
```

bind_cols(): stack data horizontally

- Combines columns together, with rows in the order they appear
  - **No attempt to match by ID**
  - Number of rows must be a compatible

```
bind_cols(tibble(new = 1:nrow(DF_duprows)), DF_duprows)

DF_bindc <- bind_cols(DF_final, DF_clean)
head(DF_bindc)

# can also use it on a _list of data frames_
bind_cols(list("..", DF_duprows))
```

dplyr: two-table verbs

- Two-table verbs allow you to ‘join’ two data frames, matching values in common.
  - similar to SQL joins
- See `vignette("two-table")` for more information

Section 17

Sharing Code

Style

"L'enfer, c'est les autres" ("Hell is other people")
— Jean-Paul Sartre ("Huis clos" / "No Exit")

"Hell is other people's code." — programming aphorism

But it doesn't *have* to be!

- The syntax of the `R` language is strict about some things, but not others, like white space and indentation.
- As mentioned at the beginning, there is often more than one way to do things in `R`
  - different styles of *naming things*
  - different name formats: `camelCase`, `snake_case`, etc.
- Code that is written in a different style than you're used to, or with inconsistent formatting, can be difficult to read and follow.

Style Guides

- A “Style Guide” can be a useful tool to help you and your collaborators write code in a consistent style.
- It also simplifies writing code, by reducing the number of (style) decisions you have to make.
- A Style Guide is ***strongly*** *recommended* for teams collaborating on shared code.
    - Even if you are working alone, it can help you write cleaner code that's easy for your *future-self* to read and understand, and for others to help you when you get stuck.

Some popular R style guides you can use (or adapt):

- The tidyverse style guide
    - based on an earlier version of Google's style guide.
- Google's R Style Guide
    - based on the current tidyverse style guide, above.

Section 18

Review

Final **Activity**

- Use this script file in the ‘activities’ folder as a starting point: `R2_activity_final.R`
- Load *all* the data files in this directory: `data/activity`
  - They are from the “World Health Organization Global Tuberculosis Report”, with counts of new cases in each year for a subset of countries, by group (method of diagnosis, gender, and age)
- Check & clean the files, as appropriate
- Combine the data into a single data frame with:
  - The total number of new TB cases in each year and country
  - The country’s population in that year
  - The rate of new cases (total cases / population)
- *optional*: Save the data in the parent directory

Final **Activity**: Challenges

Can you answer these questions with the data?

1. Which country had the highest all-time rate of new cases?
2. In which year was it?
3. Which country has had the *lowest* average rate since 2006 (including that year)?
4. Which country had the highest & lowest *growth rate* in *cases* since 2006 (including that year)?
5. What changed in 2006?

Quiz Review

- If multiple packages have functions with the same name, how can you specify which one to use?
- What are “UTF-8” and “latin1”?
- What’s the difference between a dplyr ‘verb’ and a ‘helper function’?
- Does the `filter()` verb use ‘select’ or ‘mutate’ *semantics*?
- What are the 3 rules of ‘tidy data’?
- TRUE or FALSE: If `R` reads a file without errors, there are no problems
- TRUE or FALSE: `R` has rules and conventions for naming functions
- TRUE or FALSE: if you use one package from the `tidyverse`, you have to use all of them.

Section 19

Backmatter

Other packages to look at

- `data.table`: a high-performance version of `data.frame` with few dependencies.

Other packages in the `tidyverse`:

- `lubridate` and `hms`: for date & time values.
- `purrr`: functional programming (FP) tools for working with functions and vectors.
  - Replace `for` loops with code that is more efficient and easier to read.

References

Cheatsheets:

- readr/readxl
- Data transformation with dplyr
- Data tidying with tidyr

On the web:

- Tidyverse documentation
- R for Data Science (2e)
- Data Science in a Box (#dsbox)
- An introduction to data cleaning with R

R Documentation:

- "R Data Import/Export" (`help.start()`, under "Manuals")